מסחר אלקטרוני

תרגיל בית 1

תאריך הגשה: 15.05.22

מגישים:
יונתן קויפמן 212984801
דניאל מריח 316111442

תוכן עניינים

הקדמה..... 2
ייבוא נתונים ובניית אובייקטים ראשוניים 2
חלק א – חקירת השינויים בין הגרפים הנתונים..... 3
חלק ב'- מציאת ההסתברות להיווצרות קשתות..... 8
הקדמה 8
איסוף מידע ובחירת דרך פעולה 8
בניית הסתברות בהסתמך על מדד אחד 10
Logistic Regression 11
דרך הפעולה הסופית – מדד RA משופר 12
חלק ג'- בחירת המשפיענים..... 14
ביבליוגרפיה 19

הקדמה

ייבוא נתונים ובניית אובייקטים ראשוניים

הצעד הראשון שלנו היה קריאת הנתונים.
הנתונים ההתחלתיים שלנו היו:

1. קובץ csv שמתאר את הרשת החברתית instaglam בזמן 1-
2. קובץ csv שמתאר את הרשת החברתית instaglam בזמן 0
3. קובץ csv שמתאר לכל אומן במאגר הנתונים את המשתמשים ששומעים אותו באפליקציה spotifly וכמות ההשמעות שלו אצל כל משתמש

כדי להתחיל לעבוד על תרגיל הבית היה עלינו להמיר נתונים אלה לכדי אובייקטים שנוכל לנתח. מכל קובץ שתיאר את הרשת החברתית, בנינו גרף מסוג nx באמצעות הספרייה networkX. בגרף זה, כל משתמש הינו צומת וקשת מעידה על חברות בין 2 משתמשים. שמנו לב כי ה csv שניתן לנו לא כולל כפילויות:

לדוגמא, אם צומת 122111 מופיע בעמודה ימין ולצידו מופיע צומת 333111, הצמד לא יופיע שוב בהיפוך כאשר נסתכל על החברויות של 333111. מניעת כפילות זאת נובעת מהעובדה כי חברות איננה קשר בעל כיוון כפי שהיה נתון בתרגיל הבית. אז הגרפים שנבנו מקבצי ה csv אלה גרפים לא מכוונים.

לאחר בניית הגרפים, רצינו לסדר את כל המידע, כולל המידע אודות השמעות האומנים, באובייקטים שקלים לקריאה ושליפת נתונים והם אובייקטים מסוג מילון. לכל גרף בנינו את המילון הבא:

מילון שכל מפתח בו הוא מזהה של צומת והערך הינו רשימת הצמתים שמחוברים לצומת זה באמצעות קשת, דהיינו חברים של הצומת ברשת החברתית.

עבור קובץ ה csv שמתאר את ההשמעות של כל אומן בנינו את המילון הבא:

מילון שכל מפתח בו הוא מזהה של צומת והערך הוא רשימה של tuples מהצורה הבאה: (מס' השמעות של האומן, מזהה של אומן).

לאחר שקראנו את כל הנתונים ההתחלתיים הללו, נוכל להתחיל לנתח את השינוי בין הגרף הנתון בזמן 1- לבין הגרף בזמן 0.

חלק א – חקירת השינויים בין הגרפים הנתונים

כעת שהנתונים הגולמיים בידינו, נוכל להתחיל לחקור את השינויים בין הגרף בזמן -1 לבין הגרף בזמן 0. לשם נוחות, מעתה ואילך נקרא לגרף בזמן 0 G0 ולגרף בזמן -1 נקרא G_before.

ראשית, רצינו לראות מה סדר הגודל של הגרפים בידינו. הדפסנו את המידע עליהם וקיבלנו:

```
G in time -1:
Graph with 1892 nodes and 12717 edges
G in time 0:
Graph with 1892 nodes and 14324 edges
```

ממצא א': נוצרו 1607 קשתות חדשות

לאחר מכן רצינו ללמוד אילו קשתות חדשות נוצרו. לכן בנינו פונקציה (findNewEdges) שמקבלת את 2 הגרפים, ומחזירה את כל הקשתות שנמצאות ב G0 ולא נמצאות ב G_before.

עם הקשתות החדשות בידינו רצינו ללמוד מה מאפיין אותן. בנינו לשם כך פונקציה (dataOnConnections) שתקבל את רשימת הקשתות החדשות והגרף G_before ותחקור מה המאפיינים של הצמתים שנוצרה ביניהם קשת.

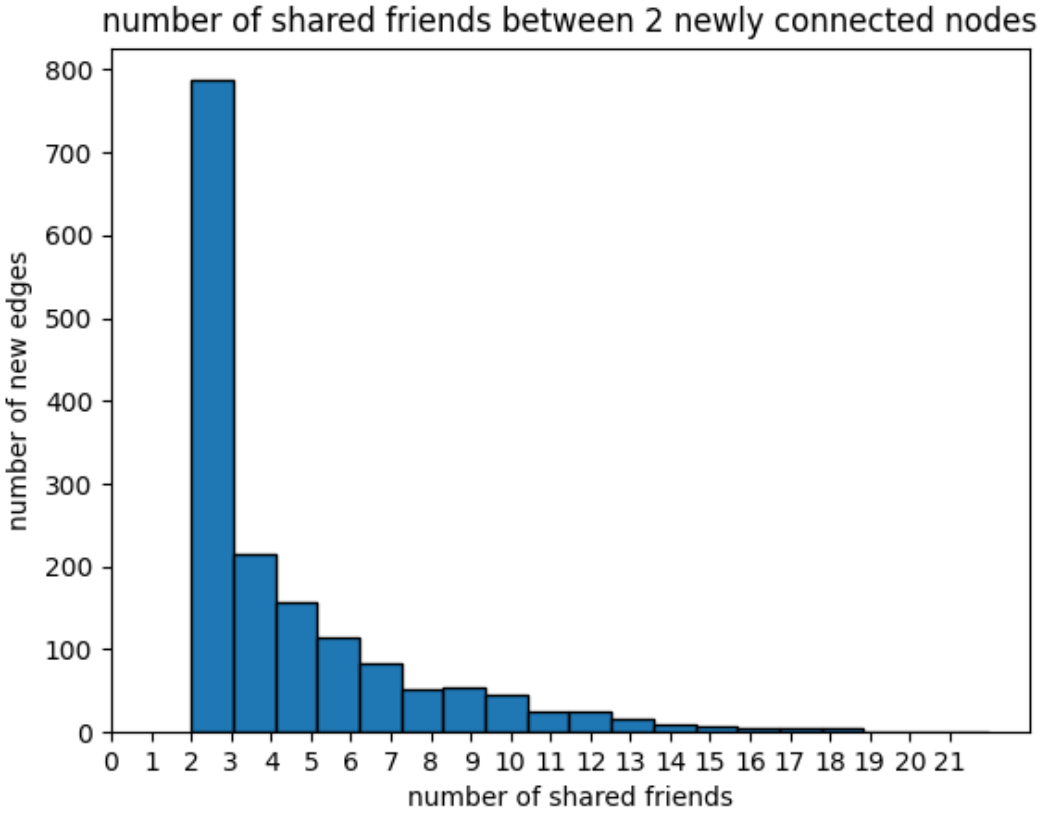
המידע שבחרנו לחקור הוא:

1. כמות השכנים המשותפים של 2 צמתים, שנוצרה עבורם קשת , בזמן -1
2. כמות החברים היו לכל צומת בזמן -1, מבין הצמתים שנוצרו איתם קשרים חדשים
3. מי הם הצמתים שהיוו [1]קצה משולש שנסגר וכמה משולשים הם סגרו

בחרנו להסתכל על מדדים אלה שכן לטעמנו במציאות חברויות ברשת חברתית בין אנשים נוצרות על סמך כמות החברים המשותפים שיש להם. בנוסף יתר המדדים יוכלו אולי להעיד אילו צמתים משתתפים ביצירת קשתות חדשות, בין אם זה ככאלו שנוצרה עוברם קשת חדשה ובין אם ככאלו שעזרו לקשת אחרת להיווצר כיוון שהיוו שכן משותף בין 2 צמתים אחרים

[1] עבור שלישיית צמתים a,b,c צומת a ייקרא קצה משולש ⇔ הקשתות (a,b) , (a,c) קיימות בגרף ו (b,c) לא קיימת

1. **<u>כמות שכנים משותפים</u>**

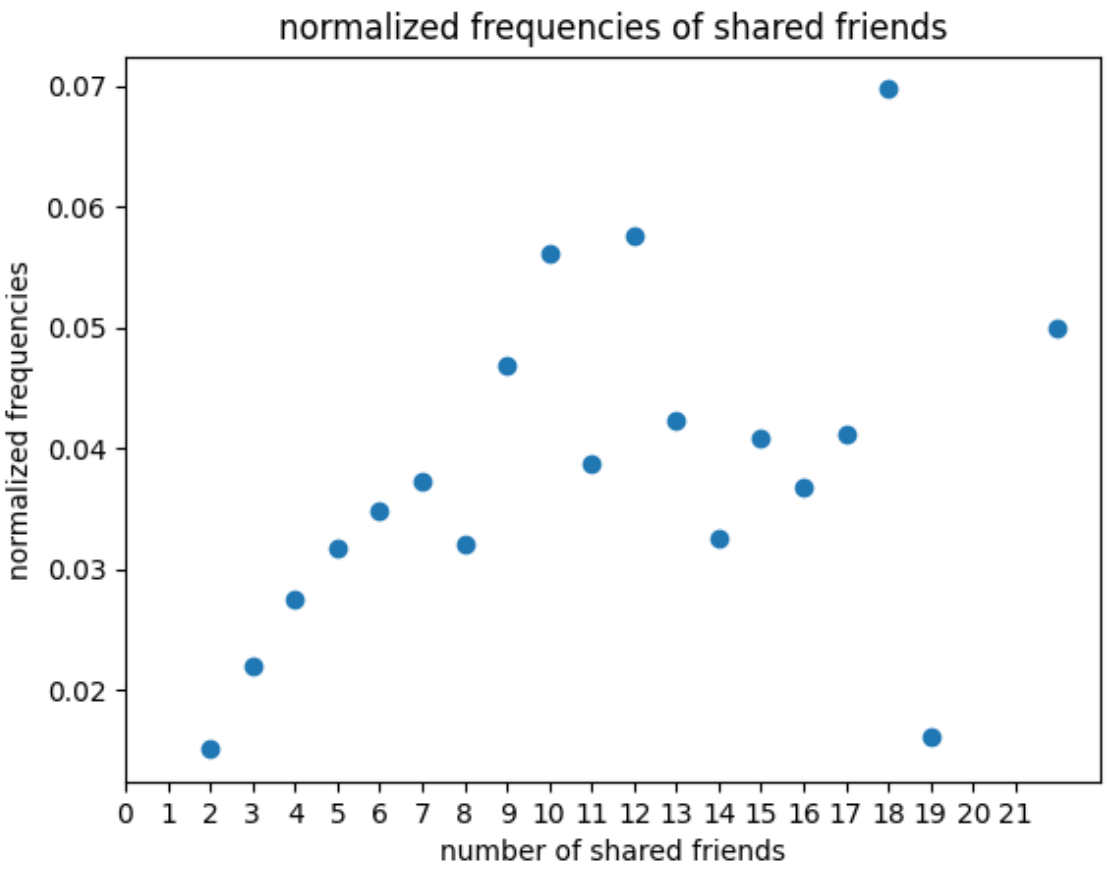


את המידע הזה ייצגנו באמצעות מילון. כל מפתח הוא מס׳ שכנים משותפים ולו ערך של כמות הקשתות שנוצרו של2 הצמתים שלהן יש כמות שכנים משותפים שתואמת למפתח. בנוסף, הצגנו מידע זה בצורת היסטוגרמה:

ממצא ב׳- גרסה ראשונית: רוב הקשתות נוצרו בין צמתים עם מעט שכנים משותפים.

הנתונים הללו הפתיעו אותנו מאוד. בניגוד לציפיות שלנו, שמס׳ גבוה יותר של שכנים מעיד על קרבה גדולה יותר של 2 צמתים, כאן ההיפך הוא הנכון. תוצאה זו גרמה לנו לחשוד שיש הטייה של הנתונים. כלומר, ייתכן שהסיבה מאחורי ממצא זה היא שמראש לא היו הרבה צמתים עם מס׳ גבוה של שכנים לכן כמות הקשתות שיחברו צמתים עם הרבה שכנים קטנה מלכתחילה. לכן יצרנו היסטוגרמה חדשה עם הנתונים מנורמלים. את הנתונים שמרנו במילון באופן דומה לנתונים הקודמים. כלומר, נסתכל על כמות הקשתות שנוצרו בין צמתים עם כמות x של שכנים, ונחלק בכמות הקשתות שהיו יכולות להיווצר בין צמתים עם אותה כמות של שכנים.

לאחר הנרמול קיבלנו את ההיסטוגרמה הבאה:



ממצא ב׳-גרסה סופית: כמות שכנים גדולה מגדילה את ההסתברות שקשת תיווצר, אם כי ישנם גורמים נוספים שמשפיעים על כך.

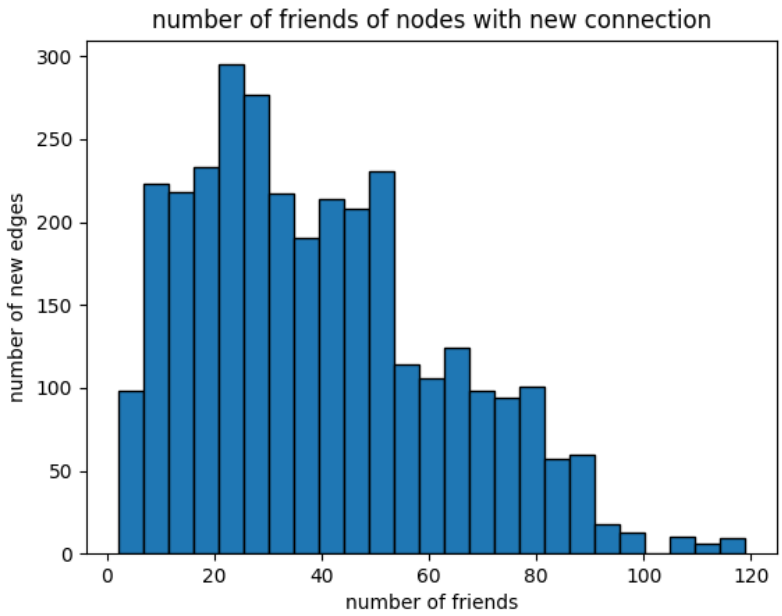
ההיסטוגרמה לעיל מראה שהגרף אכן מוטה. ככל שמס׳ השכנים המשותפים בין 2 צמתים גדל, ניתן להכליל ולומר שההסתברות שתיווצר ביניהם קשת תגדל, אם כי לא מקבלים מגמה לינארית וברורה.
את המילונים המתוארים טרם לכן בנינו באמצעות הפעולות (addToHist) ו (BuildNormalizedDict).

ממצא ג׳: לא נוצרו קשתות בין צמתים ללא שכנים משותפים

ממצא זה הגיוני, שכן גם במציאות אם ל2 אנשים אין חברים משותפים כלל, הסיכוי שיהיו חברים ברשת חברתית קטן מאוד.

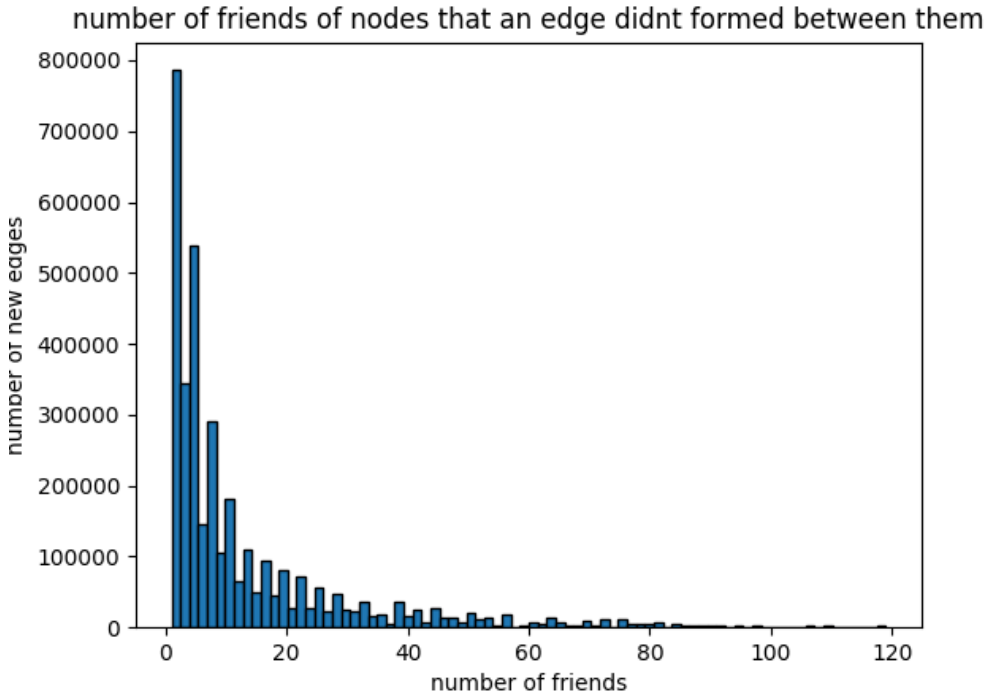
כעת נעבור לפריט המידע השני אותו חקרנו והוא כמות החברים שהיו לכל צומת מבין הצמתים שנוצרה עבורם קשת חדשה.

2. <u>כמות חברים שהיו לכל צומת בזמן 1- עבור צמתים שנוצרה ביניהם קשת</u>



ממצא ד': רוב הצמתים שהתחברו היו עם כמות חברים מסוימת טרם היווצרות הקשת ביניהם לצומת אחר.

כמות החברים של אדם ברשת חברתית לרוב מעידה על כמה הבן אדם חברתי, ופתוח ליצור קשרים חברתיים נוספים. לפיכך, הגיוני שהקשתות שייווצרו יהיו בין צמתים שהיו מלכתחילה חברותיים. כדי לחזק מסקנה זו, נוסיף היסטוגרמה שתציג את המדד אך עבור הקשתות שהיו יכולות להיווצר ולא נוצרו:



ניתן לראות בבירור כי המסקנה מתקיימת.
המאפיין השלישי ברשימה אמנם נחקר אך לא תרם לנו מבחינת מציאת מגמות בגרף. את המידע מסעיף 2 וסעיף 3 שמרנו במילונים באמצעות הפעולות (addToConnectors) ו (addToHistFriends).

לאחר שסיימנו לאסוף את כל המידע המצוין לעיל, שמנו לב לעוד פרט חשוב אודות אחוז הקשתות שנוצרו. להלן הפלט של החישוב:

```
percentage of edges created:
0.000905575573014636 9
```

ממצא ה': נוצרו מעט מאוד קשתות, 0.001% בקירוב

משמעות הדבר היא שבשלב מציאת המשפיענים, אולי כדאי שנחפש אותם תחילה בצורה "סטטית" לפני הרצת הסימולציה. כלומר, אם בכל שלב הגרף אינו גדל בצורה משמעותית, אזי משפיענים שידביקו בצורה די טובה את הצמתים בגרף קבוע, ללא היווצרות קשתות חדשות, יניבו תוצאה די טובה בסימולציה. נבחן זאת בחלק חיפוש המשפיענים בעבודה זו.

עם כל המידע שבידינו, החלטנו שהצעד הבא הוא לפתח את נוסחת ההסתברות של היווצרות הקשתות. זאת משום שכך נוכל לבחון בצורה עמוקה יותר את התנהגות הגרף ובנוסף, ברגע שתהיה לנו נוסחה מהימנה נוכל לחזות את התנהגות הגרף בצורה טובה ותוצאות הסימולציה יהיו מדויקות יותר.

חלק ב'- מציאת ההסתברות להיווצרות קשתות

הקדמה

החלק הזה היה המאתגר ביותר עבורנו. לקח לנו הרבה זמן להבין כיצד לבנות את ההסתברות שלנו ועל סמך מה. שקלנו בהתחלה לממש את ההצעה הנאיבית שהוצגה בהרצאה, והיא לתת הסתברות לקשת על סמך triadic closure. אך לאחר לבטים רבים הגענו למסקנה שזה לא אפקטיבי שכן עלינו לעבור כל זמן על כל המשולשים שעתידים להיסגר מתוך המון משולשים. בנוסף, גישה זו לא מביאה לידי ביטוי את הידע שיש לנו לאחר המחקר שנעשה בחלק הקודם.

רצינו למצוא הסתברות שגם תביא תוצאות שדומות הן מבחינה מספרית והן מבחינת הקשתות שחוזרות על עצמן למדגם הקשתות שהובא לנו, ובנוסף שיהיה מאחוריה הסבר לוגי שיבטא מגמה שמתרחשת גם במציאות.

למדנו כי נושא זה, של חיזוי היווצרויות של קשתות, נקרא link prediction analysis ושיש המון מחקרים, מאמרים ומידע עליו באינטרנט. במשך כמה ימים, לא כתבנו קוד ורק קראנו חומר אודות הנושא במגוון מאמרים ואתרים.

כל המאמרים שמצאנו אותם כרלוונטיים לעבודה שלנו ימצאו בביבליוגרפיה.

לכן כל החלק של מציאת ההסתברות להיווצרות קשת מחולק למעשה ל3 חלקים:

1. איסוף מידע על מדדים לחיזוי היווצרות קשתות ובחירת דרך פעולה
2. בניית קוד לחיזוי הקשתות שייווצרו בזמן 0 בהסתמך המדדים והשוואת התוצאה עם המדגם הנתון

איסוף מידע ובחירת דרך פעולה

תחילה נפרוש את כל המדדים שנתקלנו בהם ונתאר בקצרה כל מדד. כל המדדים שיצוינו הם מדדים **לוקאליים**[2]:

1. **<u>Common neighbours</u>**:

$$s_{xy} = \left|N_x \cap N_y\right|, N_y \text{ is number of neighbours of } y$$

מדד זה מסתכל על כמות השכנים המשותפים בין 2 שכנים משותפים. זהו איננו מדד הסתברותי שכן גודלו אינו נע בין 0 ל 1.

[2] מדד לוקאלי: בהינתן קשת, מדד לוקאלי יסתכל על 2 הצמתים בקצה הקשת ועל לכל היותר כל הצמתים שמחוברים אליהם בקשת אחת. כלומר 2 הצמתים והשכנים שלהם

2. **Salton Index (cosine similarity)**:

$$s_{xy} = \frac{|N_x \cap N_y|}{\sqrt{\deg(x) * \deg(y)}}$$

קוסינוס הזווית בין הוקטורים שמייצגים את הצמתים במטריצת השכנויות הכללית של הגרף. מדד זה בוחן את הקרבה של 2 צמתים מבחינת מעבר המידע ביניהם

3. **Jaccard Index (= Neighbourhood overlap)**:

$$s_{xy} = \frac{|N_x \cap N_y|}{|N_y \cup N_x|}$$

מודד את פרופורציית השכנים המשותפים של 2 צמתים מבין כלל השכנים של שניהם.

4. **Hub Promoted Index**:

$$s_{xy} = \frac{|N_x \cap N_y|}{\min\{\deg(x), \deg(y)\}}$$

מדד זה מעניק ציון גבוה יותר לקשרים שנוצרים עם צמתים בעלי דרגה גבוהה בגרף, קרי מרכזיים בו, שכן עם נוצרה קשת בין צומת לצומת מרכזי, ניקח את הדרגה של הצומת הלא מרכזי ואז הציון יגדל

5. **Hub Depressed Index**:

$$s_{xy} = \frac{|N_x \cap N_y|}{\max\{\deg(x), \deg(y)\}}$$

מדד שהפוך במשמעותו מן המדד הקודם. מדד זה "מעניש" קרבה לצומת מרכזה ובכך קשתות שנוצרות עבור צומת מרכזי מקבלות ציון נמוך

6. **Sorensen Index**:

$$s_{xy} = \frac{2 * |N_x \cap N_y|}{\deg(x) + \deg(y)}$$

דומה במשמעותו למדד Jaccard

7. **Leicht-Holme-Newman Index**:

$$s_{xy} = \frac{|N_x \cap N_y|}{\deg(x) * \deg(y)}$$

דומה במשמעותו ל Salton index והוא מעין וריאציה של Jaccard.

8. **Preferential Attachment**:

$$s_{xy} = \deg(x) * \deg(y)$$

מדד זה מודד חברותיות של 2 צמתים במשותף.

דרגה של צומת מייצגת כמה הוא חברותי

9. **Adamic-Adar Index**:

$$s_{xy} = \sum_{z \in N_x \cap N_y} \frac{1}{\log(deg(z))}$$

בשונה ממדדים קודמים, כאן הציון שיוענק לקשת מתבסס רק על התכונות של השכנים המשותפים של הצמתים ולא התכונות של הצמתים עצמם.

הדרגה של כל שכן משפיעה על הציון, וציון גבוה יותר יוענק לקשת של2 הצמתים שלה יש שכנים עם דרגה קטנה

10. **Resource Allocation Index**

$$s_{xy} = \sum_{z \in N_x \cap N_y} \frac{1}{\deg(z)}$$

דומה למדד הקודם, אך לוקח את ההופכי של דרגת הצומת ולא את ההופכי של log דרגת הצומת.

לאחר שקראנו על כל המדדים הללו ועל המשמעות שלהם, חשבנו כיצד להשתמש בהם. רוב המדדים אינם הסתברותיים ועל כן אם נבחר מדד כזה, עלינו לנרמל אותו. דרכי הפעולה שלנו היו:

בניית הסתברות בהסתמך על מדד אחד

דרך הפעולה הראשונית שלנו הייתה לקיחת כל מדד, אם לא הסתברותי אז מנורמל, ולשחזר את בניית G0 מ G_before באמצעות בניית פונקציית הסתברות שבנויה מהמדד הנבחר. כלומר אם בחרנו את מדד Jaccard, אזי ההסתברות לכל קשת תהיה הציון שלה במדד זה.

להבנתנו, משום שהביאו לנו מדגם מייצג של קשתות שנוצרו בסימולציה, עלינו ליצור מספר קרוב של קשתות ובנוסף לקבל בממוצע 2% של קשתות זהות.

לצערנו הרב כל המדדים לעיל לא הביאו תוצאות טובות דיו. להלן פלט של הרצת 3 סימולציות של בניית קשתות מ G_before ל G0 עם מדד Jaccard בתור פונקציית ההסתברות:

```
*****************************************************
created: 9293
wanted edges in common: 108
edge with highest Prob:  ((445584, 497872), 1.0)
edge with smallest Prob:  ((308470, 702370), 0.007751937984496124)
*****************************************************
created: 9356
wanted edges in common: 99
edge with highest Prob:  ((445584, 497872), 1.0)
edge with smallest Prob:  ((548221, 654577), 0.005952380952380952)
*****************************************************
created: 9490
wanted edges in common: 117
edge with highest Prob:  ((445584, 497872), 1.0)
edge with smallest Prob:  ((411093, 827484), 0.006711409395973154)
*****************************************************
result: 0.0672059738643435
```

ניתן לראות שנוצרו הרבה יותר קשתות מ1607. אמנם יש הצלחה של 6% אך זה משום שנוצרו יותר קשתות לכן הסיכוי שנקלע לקשתות שהובאו לנו גדל.

לכן הבנו שהכיוון הזה לא טוב מספיק ועלינו למצוא גישה מתוחכמת יותר

Logistic Regression

בחלק מהמאמרים שקראנו ראינו ששיטת Logistic regression היא גישה מקובלת בענף ובתחום של Link prediction analysis.
משום שלא ניתן היה לייבא את הספרייה skLearn, מימשנו בעצמנו את המחלקה.

בתור מטריצה ה Data בנינו מטריצה בגודל כמות הקשתות שיכולות להיווצר מזמן 1- לזמן 0 ולכל קשת נתנו 2 פיצ'רים. בתור פיצ'רים ניסינו כל פעם שילוב של 2 מדדים מבין אלו שמצוינים לעיל.
בנוסף לכל קשת נתנו label 1 אם היא באמת נוצרה בזמן 0 ו 0 אחרת.
כאשר הרצנו את הרגרסיה ההסתברויות שחזרו היו יחסית זהות וקטנות מאוד.
ניסינו להריץ את הרגרסיה עם גדלי צעד שונים ומספר איטרציות שונה אך כל הניסיונות עלו בתוהו ולא צלחו. לכן החלטנו לזנוח את הרגרסיה ולחזור למציאת נוסחה הסתברותית מבין המדדים. נציין כי את מימוש הרגרסיה עשינו באמצעות המחלקה (Logistic Regression) שבנינו ובאמצעות הפעולות (buildLabelsAndFeatures), (runLogiReg).

דרך הפעולה הסופית – מדד RA משופר

לאחר ש2 הניסיונות הקודמים לא הביאו תוצאות טובות, כלומר לא השיגו את מספר הקשתות הרצוי (בין 1600 ל 1700) או שלא השיגו התאמה של 2 אחוז בממוצע, החלטנו לחשוב שוב ולהמשיך לחקור באינטרנט ובמאמרים.

זכרנו כי יש לשכנים המשותפים ש ל2 צמתים חשיבות בהסתברות להיווצרות קשת לכן הפעם התמקדנו במדדים שמתייחסים לשכנים בלבד. לאחר סקירה רחבה, הגענו ל [3]מאמר שעוסק במדדי דמיון עבור חיזוי היווצרות קשת באמצעות שימוש ב Power Law Degree Distribution. משום שאכן עסקנו בהתפלגות זו בהרצאות, התמקדנו במאמר זה ובתוצאותיו. במאמר עלתה המסקנה הבאה:

עבור 2 צמתים שאין עבורם קשת, הסיכוי שתיווצר ביניהם קשת גדל ככל של2 הצמתים יש יותר שכנים עם דרגה נמוכה מאשר גבוהה.

תהינו מדוע הדבר מתקיים והגענו למסקנה כי הדבר גם עומד במבחן המציאות. נסתכל על 2 אנשים a,b שיש להם חבר משותף ברשת החברתית, שנסמנו ב c. אם c בן אדם מאוד חברותי ובעל הרבה חברים, הסיכוי שנתונים יזרמו מאדם a ל c ומאדם c לאדם b קטן, שכן c יכול להעביר את המידע להמון אנשים אחרים פרט לאדם b. לעומת זאת, אם c אדם בעל 2 חברים בלבד a ו b, אם אדם a יעביר נתונים לאדם c, אין לו למי להעביר פרט לאדם b. אזי החלטנו להיצמד לגישה זו בתקווה שתניב נתונים שיתאימו למדגם האימון שלנו. להלן המדד, שהוא וריאציה של Resource Allocation Index רק שבהתאם למסקנה הכתובה לעיל, "נעניש" צמתים שיש להם שכנים משותפים עם דרגה גבוהה ו "ניתן פרס" לצמתים שיש להם שכנים עם דרגה נמוכה:

$$RA_{Improved}(x,y) = \sum_{z \in N_x \cap N_y \ \wedge \deg(z) < T} \frac{1}{\deg(z)} + \sum_{z \in N_x \cap N_y \ \wedge \deg(z) \geq T} \frac{1}{\deg(z)^2}$$

T מסמן דרגת סף אותה אנו בוחרים.

במאמר מופיע מדד זה עם שורש במכנה של הביטוי השמאלי אך כאשר ניסינו להריץ סימולציה של היווצרות הקשתות עם המדד המקורי לא הגענו לתוצאות מספקות, עם בערך 8000 קשתות שנוצרות. לכן החלטנו להוריד את השורש במכנה הביטוי השמאלי והגענו למדד הזה. כמובן שמדד זה אינו הסתברותי (בין 0 ל1 כולל) לכן אנו מנרמלים אותו בערך המקסימלי של המדד בגרף בזמן t של הסימולציה.

[3] Similarity Measures for Link PredictionUsing Power Law Degree Distribution by Srinivas Virinchiand Pabitra Mitra

כמו כן, הגדרנו T=10, משום שבהרצת סימולציה עם כמה ערכי דרגה שונים, דרגה זו הייתה המיטבית. לאחר הנרמול וקיבוע ערך דרגת השכנים קיבלנו תוצאות הרבה יותר טובות מהמדדים הקודמים אך עדיין לא תוצאות מספקות דיו. לפיכך, ניסינו לבצע סימולציה, שבה אנו כופלים את פונקציית ההסתברות שתיארנו בגורם α בין 0 ל1 כדי לשפר את הביצועים.

לאחר הרצת סימולציה עם כמה ערכים שונים, הגענו לערך שמביא את מספר הקשתות ל1600 בקירוב וגם דיוק של 2 אחוז בקירוב כאשר אנו מסתכלים על כמות הקשתות הזהות בין מה שאנחנו בנינו לבין הקשתות שנתונות לנו שנוצרו. ערך זה הינו $\alpha = 0.19$.

לפיכך פונקציית ההסתברות שבנינו, שמקבלת 2 צמתים שאין ביניהם קשת בגרף G ומחזירה את ההסתברות שקשת ביניהם תיווצר הינה:

$$P(x,y,G) = \frac{RA_{improved}(x,y)}{\max\left\{RA_{improved}(x,y) \,\middle|\, x,y \in G\right\}} * 0.19$$

כדי לבדוק את הנכונות שלנו, הרצנו סימולציה של קשתות בין G_before לבין G0 5 פעמים ובדקנו את כמות הקשתות שנוצרות ולבסוף, מהו אחוז הדיוק הממוצע. להלן התוצאות:

```
*******************************************************
created: 1618
wanted edges in common: 31
edge with highest Prob:  ((212644, 748485), 0.14004192114762484)  | edge with smallest Prob:  ((693769, 916590), 0.0015352993136142049)
*******************************************************
created: 1641
wanted edges in common: 32
edge with highest Prob:  ((308470, 468812), 0.19)  | edge with smallest Prob:  ((878825, 916590), 0.0015352993136142049)
*******************************************************
created: 1651
wanted edges in common: 31
edge with highest Prob:  ((494923, 525454), 0.18198024201560345)  | edge with smallest Prob:  ((709535, 960673), 0.0015352993136142049)
*******************************************************
created: 1584
wanted edges in common: 27
edge with highest Prob:  ((212644, 748485), 0.14004192114762484)  | edge with smallest Prob:  ((804434, 847991), 0.0015352993136142049)
*******************************************************
created: 1674
wanted edges in common: 31
edge with highest Prob:  ((308470, 468812), 0.19)  | edge with smallest Prob:  ((693769, 921349), 0.0015352993136142049)
*******************************************************
result: 0.01891723708774113
```

קיבלנו דיוק של 1.9% בממוצע וכמות הקשתות שנוצרו הייתה קרובה לכמות הקשתות שהביאו לנו, לכן החלטנו לדבוק במדד זה שכן גם מייצג נאמנה את המציאות בעינינו וגם מחזיר תוצאות טובות. את בניית המדד והרצת הסימולציה שמתוארת לעיל עשינו באמצעות הפעולות (RA_improved_score), (findRA_improved_max), (RA_improvedN), (simulateEdges) ו (simulation).

כעת עם פונקציית ההסתברות בידינו, נוכל להתחיל בתהליך מציאת המשפיענים שלנו

חלק ג'- בחירת המשפיענים

שלב זה היה מאתגר ומעניין עבורנו יותר ממה שציפינו.

הנחת הבסיס שלנו הייתה כי משפיען "חזק" הוא משפיען שנמצא בשכונה שבה קל יותר להדביק צמתים שכנים, וציפינו בהתחלה ששכונות כאלה יהיו שכונות אשר מכילות כמה שיותר צמתים ששמעו את האמן הרלוונטי, תוך התחשבות במספר ההשמעות. בהתאם לכך, קבענו מדד השפעה המורכב מ: סך מספר ההשמעות שיש לשכנים של צומת כלשהו מנורמל בסך מספר השכנים ועוד אחד (פרופורציית ההשמעות ביחס למספר השכנים משקפת נאמנה את כמות ההשמעות שיש לשכנים של אותו צומת, שכן ללא הנרמול ייתכן שנמצא צומת אשר לכאורה כמות ההשמעות של שכניו גבוהה מאוד, אך בפועל יש לו כמות גדולה של שכנים ולכל שכן כמות השמעות נמוכה) כפול קבוע אלפא, שאנו בוחרים את ערכו להיות בין 0 ל-1, ועוד מספר השכנים שיש לצומת כפול אחד פחות אלפא. קבוע אלפא שימש אותנו במשקול וקביעת החשיבות של כל אחד משני המחוברים הללו, וחישובו של מדד זה התבצע במתודה calc_influence. כמו כן, מתוך ההערכה שמדד זה לבדו לא יניב תוצאות מוצלחות, החלטנו לשלב בין המדד הנ"ל לבין מדד מרכזיות אשר למדנו בכיתה, PageRank – ביצענו תעדוף ראשוני של הצמתים לפי המדד שמחושב במתודה לעיל ולקחנו את 100 הצמתים המובילים, ואז בחרנו את המשפיענים לפי ציוני ה-PageRank הגבוהים ביותר. השיקול המרכזי שלנו להשתמש ב-PageRankהיה שבגרף לא מכוון, הציון שכל צומת מקבל לפי מדד זה מחושב בפועל לפי הדרגה שלו, ולכן ציפינו שציונים אלו ישקפו בצורה לא מבוטלת את מרכזיותו של צומת כלשהו בגרף.

אולם לצערנו, אחוזי ההדבקה שקיבלנו היו נמוכים ולא מספקים. ניסינו בתור חלופה לעשות תעדוף ראשוני לפי ציוני ה-PageRank ולאחר מכן לתעדף לפי המדד אשר הצגנו לעיל, אך עדיין תוצאות ההדבקה לא השתנו הרבה.
תוצאות אלו הובילו אותנו לביצוע חיפוש מעמיק באינטרנט אודות מדדי מרכזיות שונים שהחבילה networkX יכולה לספק, ובנוסף לכך, לנסות לאפיין יותר את הגרף.
במסגרת החיפושים שלנו באינטרנט, מצאנו המון מדדי מרכזיות שונים, כגון dispersion, voterank, load, affiliation ועוד.
מתוך כלל המדדים שמצאנו, החלטנו להשתמש בחמישה מדדים:

1. **<u>Degree Centrality</u>:** הדרגה של כל צומת. ככל שצומת מקושר ליותר צמתים אחרים ניתן להניח כי הוא מרכזי יותר ותהיה לו יותר השפעה בהדבקה. מחושב ע"י: $\frac{d_v}{|N|-1}$, כאשר d_v הוא הדרגה של צומת V ו- $|N|$ מספר הצמתים בגרף.
2. **<u>Closeness Centrality</u>:** כפי שראינו בהרצאה, כמה הצומת קרוב לצמתים אחרים. מחושב ע"י הנוסחה שראינו בהרצאה.
3. **<u>Betweenness Centrality</u>:** כפי שראינו בהרצאה, מודד כמה צומת מסוים נמצא במסלולים קצרים ביותר בין צמתים אחרים. מחושב ע"י הנוסחה שראינו בהרצאה.
4. **<u>Eigenvector Centrality</u>:** מחשב את המרכזיות של כל צומת בהתבסס על המרכזיות של שכניו. מחושב ע"י: $Ax = \lambda x$, כאשר A היא מטריצת השכנויות, ו- λ הוא הערך עצמי הגדול ביותר של A. ה- Eigenvector centrality של הצומת ה-i הוא הכניסה ה-i בווקטור x שהתקבל.
5. **<u>Harmonic Centrality</u>:** וריאציה של Closeness Centrality, שמטרתה לפתור את הבעייתיות של מדד זה במקרים שהגרף לא קשיר. מחושב ע"י:

$$\sum_{v \neq u} \frac{1}{d(v,u)},$$

כאשר $d(v,u)$ הוא המרחק הקצר ביותר בין u ו- v.

בתור ניסיון, הסתכלנו על חמשת הצמתים המובילים בכל אחד מן המדדים לעיל ובחרנו מתוכם בצורה אקראית חמישה צמתים שיהיו משפיענים עבור כל האמנים. להפתעתנו, אחוזי ההדבקה שהתקבלו היו גבוהים משמעותית ממה שקיבלנו עד כה עבור כל אחד מהאמנים, ותוצאה זו הייתה תמוהה בעינינו לאור העובדה שבחירת המשפיענים הייתה אקראית וללא כל התחשבות באמן. בו-בעת, תוצאות אלו גרמו לנו להבין שמדדי המרכזיות מסייעים רבות במציאת משפיענים חזקים, ובהתאם לכך, החיפושים שלנו התמקדו בצמתים מובילים לפי המדדים הנ"ל, תוך רצון להתאים חמישייה לכל אמן.

על מנת להתאים את המשפיענים לאמנים, שוב הסתכלנו על שכנים שמאזינים לאמן, אולם הפעם התעדוף הראשוני של משפיען התבסס על מדדי המרכזיות, והתעדוף המשני התבסס על כמות השכנים שמאזינים לאמן. לצערנו, בחירת משפיענים לפי שיטה זו קרסה במבחן התוצאה, שכן המשפיענים הניבו אחוזי הדבקה נמוכים יותר ביחס לחמישייה אשר נבחרה באופן אקראי לכל האמנים.

בכדי לנסות למצוא הסבר הגיוני לתוצאות, החלטנו לבדוק את כמות רכיבי הקשירות וכמה מאזינים פר אמן יש בכל רכיב קשירות.

במסגרת הבדיקה גילינו כי ישנם 20 רכיבי קשירות, אחד מרכזי עם 1843 צמתים (כ-97% מצמתי הגרף) ועוד 19 רכיבי קשירות עם 3 צמתים בממוצע בכל אחד מהם. כמו כן, כמות המאזינים עבור רוב האמנים הייתה שולית בצורה מגוחכת – עבור אחד האמנים היה מאזין אחד (!) בכל הגרף. אמן יחיד בלט עם כמות יחסית משמעותית של מאזינים, אם כי רובם (504 מאזינים) היו ברכיב קשירות המרכזי. ממצאים אלו גרמו לנו לשער שהנתונים אודות ההסתברות כתלות במספר ההשמעות מהווים מסיח דעת יותר מאשר מידע מועיל, והגענו להחלטה שצריך להתמקד בהדבקה ברכיב קשירות המרכזי ולהזניח את השוליים. במסגרת אפיון רכיב הקשירות המרכזי, מצאנו כי הקוטר שלו באורך 9 קשתות, וכי הצמתים המובילים במדדי המרכזיות אכן נמצאים בתוכו (מה שמסביר את העלייה באחוזי ההדבקה כאשר בחרנו חמישייה אקראית מהמובילים הללו).

רכיבי הקשירות ומספר הצמתים בכל אחד מהם:

```
[1843, 2, 2, 3, 2, 2, 2, 2, 7, 2, 2, 2, 2, 3, 3, 2, 2, 4, 2, 3]
```

מספר מאזינים בכל רכיב קשירות, פר אמן:

```
389445: [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
390392: [554, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
511147: [220, 0, 0, 0, 0, 0, 0, 2, 4, 0, 1, 0, 0, 2, 0, 0, 0, 0, 0, 0]
532992: [25, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

כמו כן, שמנו לב כי הצמתים המובילים במדדי המרכזיות לרוב שכנים, או לחלופין במרחק קשת אחת אחד מהשני, ועל כן בחנו את האפשרות של "פיזור" המשפיענים בכמה "מוקדים" ברכיב הקשירות המרכזי מתוך השיקול שכמה מוקדי הדבקה קטנים עשויים להניב תוצאות הדבקה טובות יותר מאשר מוקד הדבקה מרכזי גדול. הקו המנחה לשיקול זה היה שברמת התיאוריה, הימצאותו של צומת נגוע אחד מספיקה בכדי שהדבקת שכניו תקרה בהסתברות חיובית כלשהי, וכי הוספת צומת נגוע (או יותר) לאותה שכונה אמנם תעלה את ההסתברות להדבקה של הסובבים אותם, אך תקטין את הסיכוי להגיע לצמתים שרחוקים מהמוקד המרכזי. אולם, גם פיזור מוקדי ההדבקה נכשל במבחן התוצאה והניב אחוזי הדבקה נמוכים יותר.

לאחר מכן, התבוננו שוב בהסתברות ההדבקה של צומת אשר מאזין לאמן מסוים והבחנו בנתון חשוב – לפי נוסחת ההסתברות הנתונה, ההסתברות להדביק צומת

שיש לו מתחת ל-1000 השמעות עבור אמן כלשהו, נמוכה יותר מההסתברות של צומת אשר מעולם לא שמע את האמן (כופלים את ההסתברות במספר ההשמעות ומחלקים באלף). הבחנה זו גרמה לנו לבדוק שוב את כמות המאזינים בכל רכיב קשירות פר אמן, אולם כעת ספרנו כמה מאזינים יש עם מעל ל-1000 השמעות. התוצאות הראו שבפועל רק לחצי מהמאזינים בערך יש יותר מאלף האזנות פר אמן, וכתוצאה מכך עלה החשד שקיימות בגרף "שכונות קשות להדבקה" – שכונות שקיימים בהן הרבה צמתים כאשר כל אחד מהצמתים הללו עם מספר השמעות שנמוך מ-1000 ומהווים מעין "מחסום" בשרשרת ההדבקה, ולכן הגיוני להניח שצומת בשכונה כזו ככל הנראה יתקשה להדביק דה-פקטו, גם אם הוא מוביל במדדי מרכזיות, ובהתאם לכך יניב אחוזי הדבקו נמוכים יחסית.

מספר מאזינים בכל רכיב קשירות עם מעל ל-1000 השמעות, פר אמן:

```
389445: [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
390392: [209, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
511147: [31, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
532992: [6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

אי-לכך, ניסינו בשלב הזה לחפש צמתים שמובילים במדדי המרכזיות, ולעשות תעדוף משני לפי פרופורציית מספר השכנים הקשים להדבקה (קרי, עם פחות מ-1000 השמעות) לחלק במספר השכנים הכולל הנמוכה ביותר (ככל שהפרופורציה נמוכה יותר, משמע יש יותר צמתים שניתן להדביק ביחס לכמות הצמתים שקשה להדביק). לצערנו, גם גישה זו לא שיפרה את אחוזי ההדבקה, ועבור אמנים מסוימים אף הורידה את אחוזי ההדבקה.

לסיכום, בהעדר היכולת לזהות מגמה כלשהי או תנאי כלשהו אשר מבטיח אחוזי הדבקה גבוהים, יכולנו להסתפק בבחירת החמישייה האקראית אשר הניבה אחוזי הדבקה טובים ללא התחשבות באמן. אולם, בהתחשב בתחרות שמתבססת על כמות הדבקה מרבית, ומתוך רצון להבטיח לעצמנו בחירה אולטימטיבית של משפיענים אשר יניבו את אחוזי ההדבקה הגבוהים ביותר (תוך התחשבות באלמנט האקראיות של היווצרות הקשתות וההדבקה), בחרנו את המשפיענים בסופיים באופן הבא –

לקחנו את חמשת הצמתים המובילים בכל אחד מחמשת מדדי המרכזיות שהזכרנו לעיל, ואיחדנו בין החמישיות הללו – התקבלה קבוצה של 13 צמתים. מתוך 13 הצמתים הנ"ל, הבחנו בשני צמתים שהיו מוביל ב-3 מתוך חמישה המדדים – 548221 ו-874459, ולכן בחרנו אותם כמשפיענים עבור כל האמנים. מתוך 11 הנותרים, סיננו כאלה שפחות בלטו במדדי המרכזיות או לחלופין שבמסגרת

סימולציות בהן הם השתתפו התקבלו אחוזי הדבקה נמוכים יותר, עד שנשארנו עם חמישה. מהחמישה הנותרים בנינו את כל הפרמוטציות (5 בחר 3 – סה"כ 10 פרמוטציות) ושילבנו אותם יחד עם שני הצמתים המובילים שהזכרנו לעיל. התקבלו 10 חמישיות, עבור כל אחת הרצנו 3 סימולציות, ועשינו ממוצע לאחוזי ההדבקה. בחרנו את החמישיות אשר החזירו את הממוצע הגבוה ביותר תוך שימת דגש על שונות נמוכה בתוצאות ההדבקה אשר התקבלו ב-3 הסימולציות. להלן טבלת חישובי הממוצעים:

אינדקס	1	2	3	4	5	6	7	8	9	10
389445 :	58.5	66.49	58.61	59.35	62.89	56.02	50.73	60.30	47.46	55.49
	54.28	56.60	51.69	56.23	63.9	63	58.98	59.25	43.18	60.94
	57.82	57.66	50.21	59.46	58.24	67.07	62.1	50.31	58.35	51.21
ממוצע :	56.87	60.25	53.5	58.35	61.67	62.03	57.27	56.62	49.66	55.88
390392 :	70.03	60.88	69.23	63.84	64.32	64.05	63.58	70.08	67.75	60.99
	64.85	68.71	64.37	63.42	66.27	69	65.32	57.92	66.06	64.43
	64.69	64.80	64.48	62.94	67.6	59.03	59.93	62.52	59.77	60.62
ממוצע :	66.52	64.79	66.03	63.4	66.06	64.02	62.94	63.5	64.52	62.01
511147 :	51.74	52.16	46.09	50.95	51.42	48.94	53.64	46.72	47.72	52.11
	43.49	58.4	50.52	46.24	48.83	51	44.71	43.81	47.56	53.27
	46.72	48.89	59.46	51.1	49.31	44.5	50.52	44.55	48.46	50.73
ממוצע :	47.32	53.15	52.02	49.43	49.85	48.15	49.62	45.02	47.91	52.04
532992 :	44.97	72.04	64.58	60.83	63.42	58.56	55.32	53.64	46.77	52.37
	53.91	59.03	72.78	50.42	48.73	58	52.32	55.49	58.87	61.89
	60.30	69.45	57.66	72.83	60.67	63.68	47.25	50.63	53.22	59.19
ממוצע :	53.06	66.84	65	61.36	57.6	60.08	51.63	53.25	52.95	57.8

– 3 הממוצעים הגבוהים ביותר.

ביבליוגרפיה

חלק ב' – הסתברות להיווצרות קשתות

- https://cse.iitkgp.ac.in/~pabitra/paper/iconip13.pdf
- https://www.analyticsvidhya.com/blog/2020/01/link-prediction-how-to-predict-your-future-connections-on-facebook/#h2_5
- https://cs.stanford.edu/~jure/pubs/node2vec-kdd16.pdf
- https://cran.r-project.org/web/packages/linkprediction/vignettes/proxfun.html
- https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=wright1621899961924795&disposition=inline

חלק ג' – מציאת משפיענים

- https://www.geeksforgeeks.org/network-centrality-measures-in-a-/graph-using-networkx-python
- https://www.turing.com/kb/graph-centrality-measures
- https://www.analyticsvidhya.com/blog/2020/03/using-graphs-to-/identify-social-media-influencers
- https://medium.com/neo4j/finding-influencers-and-communities-in-the-graph-community-e3d691296325
- https://www.pulsarplatform.com/blog/2014/identifying-/influencers-with-social-network-analysis
- https://towardsdatascience.com/notable-nodes-identifying-influencers-with-network-analysis-2f51f1d8fec4