תיאור המשימה

בהינתן שני מאגרי נתונים הכוללים חברויות ברשת חברתית ו-נתונים מאפליקציית מוסיקה ובהינתן רשימת אמנים, עלינו לבחור 5 אמנים שיקדמו את המוצר של החברה.

מאגרי הנתונים

- <u>חברויות ברשת חברתית</u>
מכיל רשימה של זוגות לפי ת"ז, כאשר הופעת כל זוג מצביעה על חברות בין שני המשתמשים ברשת החברתית ברגע מסוים.
קיימים שני קבצים מסוג זה, המתארים ימים שונים בהפרש של שבוע.

- <u>רשימת השמעות</u>
מכיל רשימה של (ת"ז, מזהה אומן, ומספר השמעות)

אופן בחירת המשפיעים

הנחות להסבר - לאורך כל ההסבר נשתמש במונחים הבאים:

גרף $G(V, E_t)$ יציין את מצב החברויות ברשת החברתית בזמן t, כאשר:

- V – מספר המשתמשים ברשת החבר
- E_t – קבוצת כל החברויות הקיימות בין כל שני משתמשים
- קודקוד i – המשתמש עם תעודת זהות i
- קשת בין קודקוד i ל-j – המשתמשים עם תעודות הזהות i, j חברים ברשת החברתית
- דרגה של קודקוד i – מספר החברים של i ברשת החברתית
- $t = -1,0,1, \dots ,6$

<u>שלב ראשון – חיזוי מצב החברויות ברשת החברתית בכל תקופת זמן</u>

על מנת למצוא את המשפיעים שיצליחו לפרסם את המוצר בצורה הטובה ביותר, ראשית רצינו להעריך כיצד יראה הגרף בכל נקודת זמן נתונה.

על מנת לבצע זאת, ברצוננו למצוא אלגוריתם חיזוי להסתברות יצירת קשת (חברות) בין כל שני משתמשים של הרשת החברתית, שבעזרתו נוכל לקבל תחזית למצב הגרף בכל זמן נתון ובכך לבחור את המשפיעים בצורה מושכלת.

- **בחירת האלגוריתם:**

ראשית, הנחנו שהשפעת מספר ההאזנות לאומן מסוים של שני קודקודים על ההסתברות ליצירת קשת ביניהם זניחה.
ולכן, המידע הנתון לנו על כל קודקוד הוא מספר החברים שלו וזהותם.

יהיו i, j שני קודקודים כלשהם בגרף, ונניח שאנחנו נמצאים בזמן t. נסמן:

$k_1 :=$ מספר השכנים של הקודקוד i
$k_2 :=$ מספר השכנים של הקודקוד j
$k_0 :=$ מספר השכנים המשותפים של קודקודים i ו- j

מטרה:

בהינתן שבזמן t, לא הייתה קיימת קשת בין שני קודקודים נרצה למצוא את ההסתברות ליצירת קשת בזמן $t + 1$.
עבור המקרה שקיימת ביניהם קשת בזמן t, לפי הנחת הבעיה הקשת תישאר בכל זמן $\bar{t} > t$.

כלומר, בהינתן שני קודקודים שלא קיימת ביניהם קשת, מספר השכנים של כל קודקוד ומספר השכנים המשותפים, נרצה למצוא את

$$p\big((i,j) \in E_{t+1} \big| k_0, k_1, k_2\big)$$

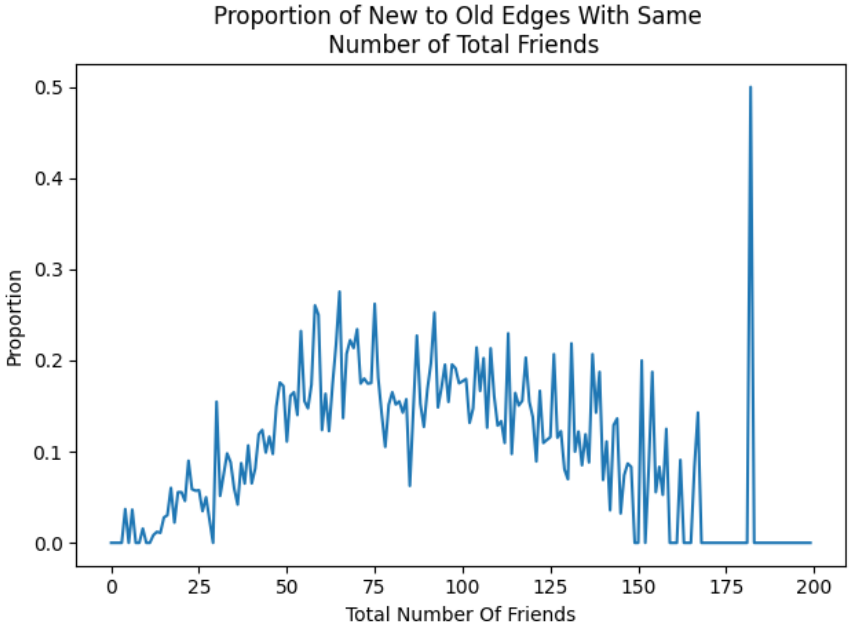
ההשערה שהצענו:

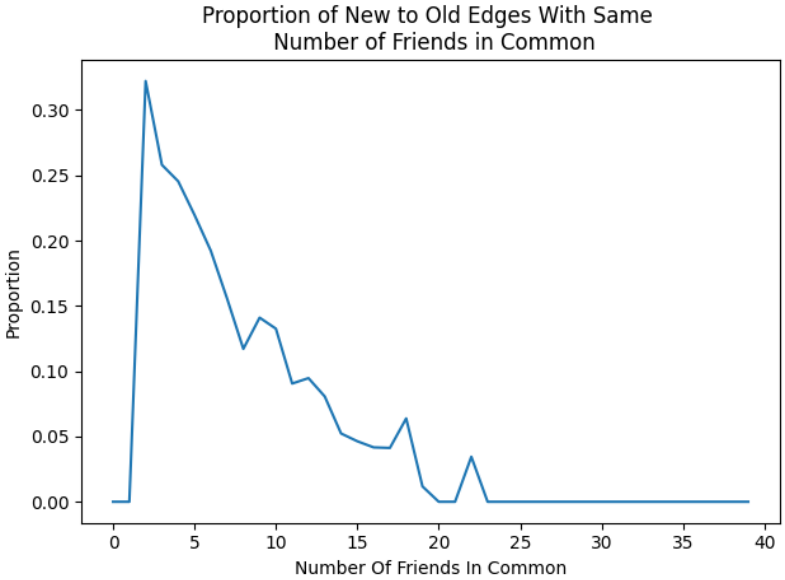
קיים טווח ממוצע של כמות קשתות שרוב הקודקודים ברשת יגיעו אליו בשלב מסוים, כאשר הגידול בכמות הקשתות יהיה גדול ביותר ככל שהקודקוד חדש יותר בגרף (כלומר, מספר הקשתות שלו עדיין נמוך משמעותית מהממוצע)

בנוסף, אופן יצירת הקשתות עבור קודקוד חדש יהיה תלוי באופן חזק במספר השכנים המשותפים שיש לו עם קודקוד אחר בגרף. כלומר, ככל שמספר השכנים של קודקוד i נמוך יותר, כך תיתנן חשיבות גדולה יותר למספר השכנים המשותפים שיש לו עם קודקוד j בקביעת ההסתברות ליצירת קשת בין שני הקודקודים.

בחינת ההשערה:

על מנת לבחון את ההשערות שלנו, ביצענו ויזואליזציה של מצב הגרף בזמנים $t = 0, -1$. להלן התוצאות:





- מהגרף הראשון ניתן לראות שאכן מספר השכנים של כל קודקוד ברשת החברתית מזכיר התפלגות נורמלית, כאשר לרוב הקודקודים מספר שכנים שנע באותו הטווח

- מהגרף השני ניתן לראות שאכן קצב הגידול בכמות השכנים החדשים של כל קודקוד קטן ככל שמספר השכנים ההתחלתי גדול יותר

הצעת הסתברות ליצירת קשתות:

לאור התוצאות שקיבלנו, הצענו את הנוסחה הבאה:

$$p\big((i,j) \in E_{t+1} \big| k_0, k_1, k_2\big) = \min\left\{1, \left(\frac{1}{k_1} + \frac{1}{k_2}\right) * (\lambda k_0 + c)\right\}$$

כאשר $\lambda > 0$ מתפקד בתור hyper parameter, והקבוע c נועד לתת מענה לכך שבפועל גם לקודקודים ללא חברים משותפים בכלל קיימת הסתברות גדולה מ-0 ליצירת קשת ביניהם

בחינת ביצועי האלגוריתם:

ראשית בחרנו $\lambda = 1$.

על מנת לבחון את דיוק פונקציית ההסתברות בחיזוי השתמשנו במצב הגרף בזמן $t = -1$ בתור סט האימון, ובמצב הגרף בזמן $t = 0$ בתור סט המבחן ובדקנו את התוצאות שקיבלנו בעזרת מציאת:

- ערכי TP – אחוז חיזוים נכונים ליצירת קשתות מתוך כלל הקשתות שנוצרו
- ערכי TN – אחוז חיזויים נכונים שקשת לא תיווצר מתוך כלל השקתות שנוצרו

את תוצאות אלה בחנו מול פונקציות הסתברות שונות ליצירת קשתות, ומאחר וקיבלנו את התוצאות הטובות ביותר עבור הפונקציה הנ"ל, בחרנו בה.

על מנת למצוא את הפרמטרים λ, c המתאימים ביותר, ביצענו מספר הרצות שונות עם ערכים שונים, ובדקנו שוב מהם ערכי ה-TP וה-TN, מצאנו שהערך שהמביא לחיזוי הטוב ביותר הוא $\lambda = 10$, $c = 0.05$

תוצאות דיוק החיזוי עבור פונקציית ההסתברות והפרמטרים הנ"ל:

$$TP = 99.75$$
$$TN = 88.09$$

חיזוי הגרפים בכל שלב

לאחר בחירת פונקציית ההסתברות ליצירת קשתות, הפעלנו את האלגוריתם הנתון בקובץ ה-py על מנת לחזות את מצב הגרף בכל זמן.

שלב שני – צמצום האפשרויות למשפיעים אפשריים:

כעת, לאחר שנתונות לנו התחזיות למצב הגרף בכל שלב, ברצוננו לבנות סימולציה להתפשטות פרסום המוצרים.

מכיוון שמספר האפשרויות לבחירת כל החמישיות האפשריות של משפיעים נתונה ע"י (בערך) $2 * 10^{16}$, החלטנו למצוא דרך לצמצם את מספר האפשרויות.

בחירת משפיעים פוטנציאלים

אימצנו את הרעיון הכללי של *PageRank* ויצרנו דירוג חדש (*rank*) לכל משתמש ברשת, ניתן דירוג *rank* ביחס לאמן מסוים נתון. האלגוריתם לחישוב הדירוג פועל כך שאם לחבר של אדם כלשהו יש מספר השמעות חיובי, האלגוריתם מגדיל את הדירוג של האמן ומוסיף לו את היחס של מספר השמעות לאלף כפול מספר החברים שיש לחבר עצמו.

קל לראות כי עבור מספר השמעות שקטן מאלף עבור משתמש שנמצא ברשימת שכנויות של אדם (זהו מספר השמעות שנמוך וכנראה מעיד שהוא לא אוהב את האמן שכן הסיכוי שיקנה את המוצר קטן יותר) האלגוריתם מעניש את הדירוג של האדם, ועבור מספר השמעות שגדול מאלף האלגוריתם מתגמל את הדירוג במחשבה שמספר השמעות גדול מאלף מעיד על סיכוי גבוהה יחסית של קנייה בשלבים הראשונים של הריצה.

באופן אינטואיטיבי למשפיען טוב יהיו הרבה חברים, והרבה חברים שאוהבים במיוחד את האמן הנתון. וגם לחברים של החברים של המשפיען יהיו הרבה חברים ("מעגל שני"). כך שמעגל ההדבקה יתפשט מהר ככל הניתן.

זהו אלגוריתם "גרידי" שמנסה למקסם את מספר הקניות בתחילת ריצת הסימולציה כדי שהן בתורן יהיו יכולות להשפיע לטובה על קניה של משתמשים אחרים במשך זמן רב ככל האפשר ובכך להגדיל סטטיסטית את הסיכוי לקניה של אדם כלשהו. כלומר פונקציית המטרה היא יותר ניסיונות הדבקה לכל אדם.

בסוף הפעלת הפעולה הנ"ל, לכל אדם יהיה ציון בתכונה rank עבור אמן כלשהו. ניקח את 20 האנשים שדורגו הכי גבוה ונמשיך עם הרשימה המצומצמת של משפיעים פוטנציאל לאמן מסוים.

שלב שלישי – הרצת סימולציה לבחירת 5 משפיעים לכל אומן

כעת, לאחר שעבור כל אומן, נתונים לנו *20* הקודקודים עם פוטנציאל ההדבקה הגדול ביותר, הרצנו סימולציה שמשתמשת בתחזית שלנו למצב הגרף בכל שלב על מנת לבחון את כמות הקודקודים שקודקוד נתון (משפיען) הצליח "להדביק"

תוצאות ביניים של הסימולציה

עבור כל קודקוד מתוך הקודקודים הפוטנציאלים, שמרנו את רשימת כל הקודקודים שהצליח "להדביק".

לאחר מכן, לכל חמישייה אפשרית של קודקודים (משפיעים) מצאנו את גודל קבוצת האנשים שהצליחו להדביק יחד (כלומר, אם קודקוד מסוים נדבק ע"י שני קודקודים התחלתיים שונים, קודקוד זה ייספר בסה"כ פעם אחת)

בסופו של דבר, לכל אומן החזרנו את הקודקודים עם אחוז ההפצה הגדול ביותר.

להלן התוצאות (הממוצעות) הצפויות לכל אומן:

- עבור האומן *144882*:
  - המלצה לבחירת משפיעים: [441435, 999659, 249667, 473614, 338687]
  - אחוז הדבקה משוער: 94%

- עבור האומן 194647:
  - המלצה לבחירת משפיעים: [249667, 473614, 705172, 798855, 555524]
  - אחוז הדבקה משוער: 93%

- עבור האומן 511147:
  - המלצה לבחירת משפיעים: [249667, 473614, 705172, 555524, 189144]
  - אחוז הדבקה משוער: 87%

- עבור האומן 532992:
  - המלצה לבחירת משפיעים: [672617, 999659, 175764, 798855, 852394]
  - אחוז הדבקה משוער: 94%

- הערה:
  מכיוון שהאלגוריתם הסתברותי (יצירת הקשתות והדבקת קודקוד) תוצאות האלגוריתם משתנות הן באחוז ההדבקה המשוער והן בבחירת האומנים. עם זאת, עבור כל אומן קיים משפיען אחד או יותר אשר תמיד נבחרים, בנוסף אחוז ההדבקה משוער זהה עד לסטייה של אחוזים בודדים.