פרויקט סוף - אלגוריתמים ברשתות

מגישים:

**עומר שחורי - 207196304**

**עמרי ריינדורף - 302141874**

נושא - **עבודת מחקר על סופות טורנדו בארצות הברית בין השנים 1950 - 2021**

מקור הנתונים - [**https://www.kaggle.com/datasets/danbraswell/us-tornado-dataset-1950-2021**](https://www.kaggle.com/datasets/danbraswell/us-tornado-dataset-1950-2021)

לינק לפרויקט בDRIVE-

https://drive.google.com/drive/folders/1xt6rwL4YCzl\_jH-7r5YhxALDyKqbgvyh

לינק לפרויקט בGITHUB-

https://github.com/OmerShchory/Network\_Algorithms

מבוא –

סופות טורנדו היא תופעה יחסית נפוצה בארצות הברית. בכל שנה נצפות מספר סופות במקומות שונים ברחבי ארה"ב. חלקן לא מזיקות וחלקן מסוכנות וגורמות להרס רב. באופן כללי כל תופעת טבע ובפרט סופות טורנדו הוא נושא שכבר חקרו עשרות אלפי פעמים וממשיכים לחקור אך לאחר סקירה ב sciencedirect ו ב pubmed ראינו שאין כלל מחקר על סופות טורנדו ע"י רשת חברתית. בעבודה זו ננסה לראות או להבין האם יש קשר כלשהו בין נתוני הטורנדו למדדי הרשת המוכרים ואילו מסקנות ניתן להסיק ממבט על רשת של טורנדואים הקשורים למציאות. בנוסף, נחקור את הקהילות ובכל רשת שנבנה ונסיק מסקנות לגבי המגמות במהלך השנים על המפה, על עוצמות הטורנדו וכו'. נעשה זאת ע"י שימוש בספריית networkx ונבנה דרכה רשת דינאמית – כל תת רשת תהיה עשור.

הכנת הנתונים **–**

כבר במבט ראשוני על הדאטה, גילינו כי הוא מסודר מאוד, בעל תכונות מעניינות וללא ערכים ריקים. ביצענו עליו מספר פעולות על מנת שיהיה ברור ונוח יותר לעבודה. ראשית סיננו את הדאטה והשארנו סופות טורנדו שאורך מסלולן מעל 5 ק"מ ובכך ירדנו מ 67 אלף רשומות ל 17 אלף רשומות. בנוסף, שינינו את שמות העמודות לשמות ברורים יותר, המרנו יחידות מרחק אמריקאיות ליחידות שמשתמשים בהן בישראל והשתמשנו ב API על מנת לקבל כתובת טקסטואלית מקואורדינטות המוצא והסיום של כל טורנדו. מצורף קובץ "data preparation.ipynb" המכיל את הקוד להכנת הדאטה הכולל הסברים בתוכו.

קשיים ומגבלות –

* הקמת הרשת - כפי שכבר הזכרנו במבוא, לא מצאנו מחקר על סופות טורנדו ע"י שימוש ברשת וכנראה שזה לא מקרי. ארה"ב היא מדינה ענקית. בניית רשת של סופות טורנדו בה כל צומת הוא מיקום על המפה (התחלתי או סופי), תעניק לנו זוגות של צמתים המחוברים בקשת מכוונת. על מנת להתגבר על הבעיה ויתרנו על דיוק המיקומים ובחרנו לעגל את הקואורדינטות ל- 0 ספרות אחרי הנקודה. בכך הפכנו את הגרף מזוגות צמתים לרשת.
* ויזואליזציה של הרשת - למרות שהרשת שלנו דינמית ומחולקת לעשורים, קשה להציג את הרשת ע"י networkx. זאת מכיוון שהרשת עדיין גדולה מאוד ואין מספיק מקום בפלט של הגרף על מנת להציג את הרשת בצורה ברורה. ניסינו להתגבר על בעיה זו ונעזרנו בספריית pandas המשתמשת בעיקר לניתוח דאטה כדי להציג את נתוני הקשתות המייצגות סופות טורנדו, ונציג את הצמתים על גבי מפה.
* הצלבה עם נתונים אחרים- רצינו לנסות לשלב מידע הקשור למזג האוויר במדינות של ארה"ב על מנת לקשר בין מדדי מזג האוויר לסופות הטורנדו אך התקשנו למצוא דאטה מתאים. למרות שאנחנו מוגבלים להשתמש אך ורק בנתונים הקשורים לסופות טורנדו והשלכותיהן, למזלנו הנתונים מספקים לשאלת מחקר. בנוסף, למרות שאין ברשותנו נתונים להצלבה אי אפשר להזניח את נושא ההתחממות הגלובאלית. ניתן להסיק שבטווח השנים של המחקר הטמפרטורה הממוצעת הולכת וגדלה בכדור הארץ (בפרט בארה"ב), נוכל לקשר את התנהגות הרשת לאורך השנים עם ההתחממות הגלובאלית.

**שאלת המחקר –**

במחקר זה נרצה לבדוק האם קיים קשר בין מדדי הרשת לנתוני סופות הטורנדו. בנוסף, אילו מסקנות מהמציאות ניתן להסיק מהרשת, נרצה להבין האם כדאי להשתמש במדדי רשת ובאלגוריתמים בשביל לחקור את הנושא לעומק. כמו כן, נחקור את הקהילות השונות בין הרשתות, ונבחין במגמות בין כלל העשורים.

**מהלך הפרויקט –**

שם הקובץ המכיל את קוד המחקר: "Tornados\_Network\_Analysis.ipynb" .

בחלק זה ננסה להסביר את מהלך המחקר שלנו ומה עומד מאחורי קטעי הקוד. ניגשנו לכתיבת הפרויקט בגישת "נרה את החץ ונשרטט את המטרה מסביבו". כשהתחלנו לעבוד על הפרויקט היה קשה להתחיל כי לא ראינו את המטרה מה שהקשה עלינו לקבל כיוון. בתחילת הדרך התחלנו בניסוי וטעיה ולכן לא כל קטעי הקוד נועדו להסקת מסקנות אך עזרו לנו להגיע לתוצאה סופית.

ראשית כל בנינו רשת לסופות הטורנדו כך שכל צומת מהווה קואורדינטות מעוגלות, וסופת טורנדו תוצג בתור קשת מכוונת מנקודת התחלה לנקודת סיום. מכיוון שמה שהכי מעניין אותנו ברשת הן סופות הטורנדו שהן קשתות בגרף רוב המדדים שנסתכל עליהם יהיו קשורים לקשתות. מעתה והילך נתייחס לסופת טורנדו בתור קשת ולמיקום בתור צומת. לפני בניית הגרף ביצענו סינון לדאטה על מנת לבנות רשת בין השנים 1950-1961. קוד הפרויקט וגם המסקנות על מדדי רשת/הקהילות בנויים תחילה על העשור הזה. לאחר מכן נריץ את הקוד מחדש על כל עשור שלאחר מכן כדי להסיק מסקנות על התנהגויות ומגמות שונות במהלך העשורים.

לאחר סינון הדאטה בנינו את הגרף בעזרת פונקציית עזר שבנינו "initGraph" המקבלת data frame בתור פרמטר ויוצרת ממנו גרף מכוון ע"י שימוש בפונקציה מובנת מספריית networkx ובניית צמתי וקשתות הגרף מהנתונים בטבלה. כפי שכבר ציינו, מה שמעניין אותנו הן קשתות הגרף לכן כל המידע מהטבלה יישב על הקשתות. ע"פ הטבלה קשת יכולה להופיע יותר מפעם אחת, ולכן אנחנו נאחד כל קשת באופן הבא:

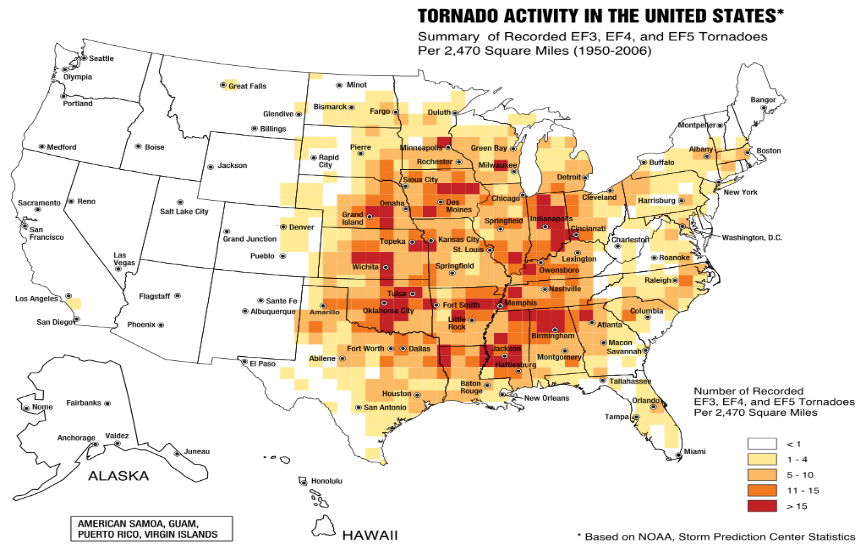
* לכל קשת תהיה שדה edgeCount שתסכום את מספר הרשומות מהטבלה שיוצרים את הקשת.
* אורך המסלול, מספר הפצועים ומספר ההרוגים יסכם.
* מדד ה EF ו רוחב הטורנדו יחושבו בתור ממוצע.

לאחר בניית הגרף רצינו לבדוק את המדדים הפשוטים ביותר in/out degree של הצמתים. השתמשנו בפונקציה המובנית של הספרייה networkx והוצאנו את חמשת הצמתים שקיבלו את המדדים הגבוהים ביותר. לאחר מכן, בדקנו שוב את ה in/out degree אבל הפעם לפי אורכי המסלול ולא לפי כמות הקשתות. רצינו לבדוק האם המקומות יוצאים או נכנסים מהם/אליהם סופות טורנדו, האם אותם סופות עברו מסלול ארוך. מהסתכלות על התוצאות ראינו שני דברים:

א. יש רשימה של כמה מדינות ממרכז ארה"ב שחוזרות על עצמן.

ב. יש חיתוך בין שתי הרשימות.

לכן על פי הבדיקה הפשוטה שעשינו קיבלנו תחושה שבאזורים הללו פופולאריים לא במקרה. ניגשנו לאינטרנט על מנת להבין יותר לעומק את התופעה ומצאנו כי יש רשימה של 17 מדינות בארה"ב שמהוות קבוצה שנקראת Tornado Alley שלפי ויקיפדיה יש בהן את השכיחות הגבוהה ביותר להיווצרות סופות טורנדו - <https://he.wikipedia.org/wiki/%D7%A1%D7%9E%D7%98%D7%AA_%D7%94%D7%98%D7%95%D7%A8%D7%A0%D7%93%D7%95> .



כדי להמשיך, רצינו למצוא דרך נוחה לגשת למידע שקיים על קשתות הגרף בצורה נוחה וויזואלית. על מנת לגשת למידע שעל הקשתות בצורה נוחה עברנו על כל קשתות הגרף ויצרנו טבלה כאשר המידע ששמור עליהן יהפוך לעמודות בטבלה. עברנו שוב על הטבלה על מנת לחשב לכל קשת את מדדי ה: edge degree, closeness, betweenness ו eigenvector שיהפכו לעמודות בטבלה. סיימנו בכך שחילצנו מטבלת סופות הטורנדו את מדינת המוצא ומדינת הסיום של הטורנדו, הוספנו עמודה בינארית שנקראת "Is Tornado Alley" שתקבל ערך true רק אם מדינת המוצא והסיום שייכות לרשימת המדינות ב tornado alley.

ע"י הטבלה שיצרנו נוכל לחפש בקלות קשתות מעניינות. ראשית כל רצינו לעשות מהלך דומה למה שעשינו בתחילת בפרויקט על מנת למצוא צמתים פופולאריים. יצרנו 4 רשימות של צמתים, כל רשימה תכלול את 10 הצמתים בעלי הציון הכי גבוהה לכל מדד מרכזיות edge degree, closeness, betweenness ו eigenvector. לאחר מכן הצגנו את החיתוכים בין כל זוג קבוצות ולא כל כך הופתענו שכמות החיתוכים הייתה קטנה מאוד מהסיבה הפשוטה שארבעת המדדים שונים במשמעותם ובחישוב שלהם. דווקא בזכות התוצאה שקיבלנו אנחנו יכולים להיות בטוחים שהרשת מתנהגת באופן תקני ואכן לא מדובר בזוגות של צמתים.

בשלב הבא רצינו לבדוק את המתאם בין מדדי המרכזיות של הקשתות לנתונים מהמציאות. השתמשנו בפונקציית corr של ספריית pandas שמשתמשת בקורלציית ספירמן מאחורי הקלעים ומחזירה את מטריצת הקורלציה כפלט. לצערנו הקורלציות היו גבוהות יותר בין נתוני הטורנדו (מספר פצועים והרוגים לדוגמה) ובין מדדי הרשת אך לא בין מדד רשת לנתון על הטורנדו. בכל זאת ניסינו לראות אילו מדדים באופן יחסי מקבלות מתאם גבוהה יותר במדד ביחד לקבוצה וראינו שה edge degree וה edge closenessמקבלים את המתאמים הגבוהים יותר גם אם הם נמוכים ביחס לכל המטריצה. בנוסף ראינו שלפי העמודה "Is Tornado Alley" שיצרנו מאוחר יותר דווקא היא קיבלה מתאם של כמעט 50% עם ה edge closeness שזה גבוה באופן משמעותי משאר התוצאות שקיבלנו. לכן אנחנו יכולים להסיק מכך שלמדדים אלו יהיה משמעותיים יותר בהסתכלות על הרשת כולה. המשכנו בכך שאיחדנו ללא כפילויות את רשימות עשרת הקשתות שקיבלו את התוצאות הגבוהות ביותר במדד ה degree ו closeness למען הסקת מסקנות שהן:

1. כל סופות הטורנדו התחילו והסתיימו ב Tornado Alley.
2. רוב סופות הטורנדו היו ברוחב של עשרות או מאות מטרים ונחשבות ליחסית גדולות.
3. מדד ה EF הממוצע של כל סופות הטורנדו הוא 2 ומעלה.
4. רוב אורכי המסלול של סופות הטורנדו הרבה מעל 5 ק"מ, חלקם עברו מעל 100 ק"מ.
5. יש כמה סופות טורנדו בעלות מספר גבוהה של פצועים והרוגים. מכיוון שארה"ב היא מדינה ענקית אנחנו יכולים רק להסיק שהרבה סופות מסוכנות לא בהכרח עוברות דרך אזור מגורים אז גם סופות מסוכנות לא בהכרח נגמרות עם עשרות פצועים או הרוגים.

לסיכום לפי הממצאים אנחנו יכולים להסיק סופות הטורנדו ברשת בעלי מדדי degree או closeness גבוהים יכולות להיות מסוכנות.

**קהילות –**

לאחר מחקר באינטרנט וחזרה על החומר מהקורס, התחלנו להשתמש באלגוריתמים הבאים למציאת קהילות ברשת:

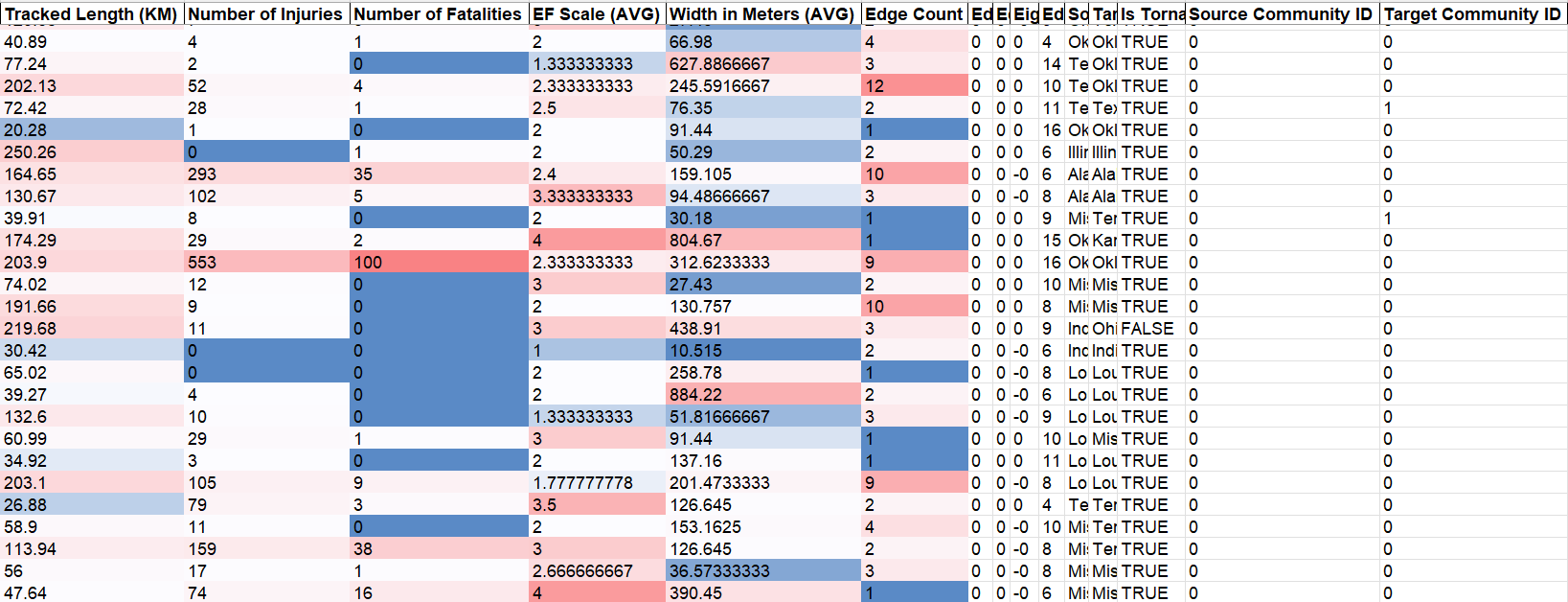
1. אלגוריתם louvaine – האלגוריתם מהיר וסקיילאבילי. מתאים למציאת קהילות. מוצא מדד modularity אופטימלי על מנת לזהות קהילות ברשת. ידוע בכך שהוא יעיל בהתמודדות ברשתות גדולות.
2. אלגוריתם greedy modularity – אלגוריתם פופולרי למציאת קהילות ברשת. המטרה שלו היא לחלק את הרשת לקהילות על ידי מקסום של מדד הmodularity. מדד זה מודד את חוזק החלוקה של הרשת לקהילות על בסיס הצפיפות של הקשרים בתוך הקהילות בהשוואה לצפיפות המצופה מרשת רנדומלית.
3. אלגוריתם girvan newman – אלגוריתם פופולרי נוסף למציאת קהילות. מסיר באופן פרוגרסיבי את הקשתות ברשת על מנת לחלק אותה לקהילות. מבוסס על הקונספט של מדד ה- edge betweenness centrality.

לאחר שניסינו לחלק את הרשת בעזרת שלושת האלגוריתמים שמצוינים מעלה, הסקנו כי האלגוריתם של greedy modularity חילק את הרשת לקהילות בצורה ההגיונית והנוחה ביותר לעבודה. נתנו לכל קהילה מספר ונוצרו 167 קהילות אך עדיין זה מספר גדול מדי יחסית לגודל הרשת. לכן רצינו לבדוק אם נוכל בדרך פשוטה לאחד בין קהילות. ראשית כל ניתן לראות כי כל קהילה קיבלה מספר מזהה ורצינו לראות אם יש קשר בין המספרים. בשביל הבדיקה ביצענו סינון לדאטה על מנת לקבל את רשימת הקשתות שמחברות בין קהילות. היה ניתן לראות יחסית בבירור שההפרש בין מזהה הקהילה של מוצת המוצא ליעד לא היה גדול במיוחד. לפי כך יכלנו להסיק ש שני קהילות שערך המזהה שלהם קרוב שייכים ל cluster גדול יותר. ראינו לנכון לעבוד עם 11 קהילות לכן קיבצנו את כל מזהי הקהילות בקבוצות של מספר הקהילות המקורי חלקי 10. הסבר- במקרה שלנו מספר הקהילות הוא 167 אז 167/10 = 16.7 לכן כל המזהים מ 0 עד 16.7 יקבלו את הערך 1, המזהים מ 16.7 עד 33.4 יקבלו את הערך 2 ו באופן כללי המזהים מ 16.7n עד 16.7(n+1)כאשר n הוא מספר טבעי בין 0 ל 10 כולל, יקבלו את הערך n. מציינים שהמספר 16.7 כנראה ישתנה בכל תת רשת של הרשת הדינאמית.

ניתן לראות זאת גם בקובץ האקסל שאליו ייצאנו את הדאטה שנקרא – output.xlsx. בקובץ זה ניתן להבחין כי עולה קשר בין מספר הקהילה לבין נתוני הטורנדו. ככל שערכי מזהי הקהילות של צמתי התחלה וסוף נמוכים יותר יהיה ניתן לראות שיש ערכים גבוהים יותר של כל נתוני הטורנדו ולהיפך (מתאם שלילי). למען ההמחשה השתמשנו בצביעה של הערכים לכל עמודה כך של ככל שהערך נמוך יותר יצבע בגוון ככול וככל שיהיה גבוה יותר ייצבע בגוון באדום.

לדוגמה:

ערך מזהה הקהילה נמוך:

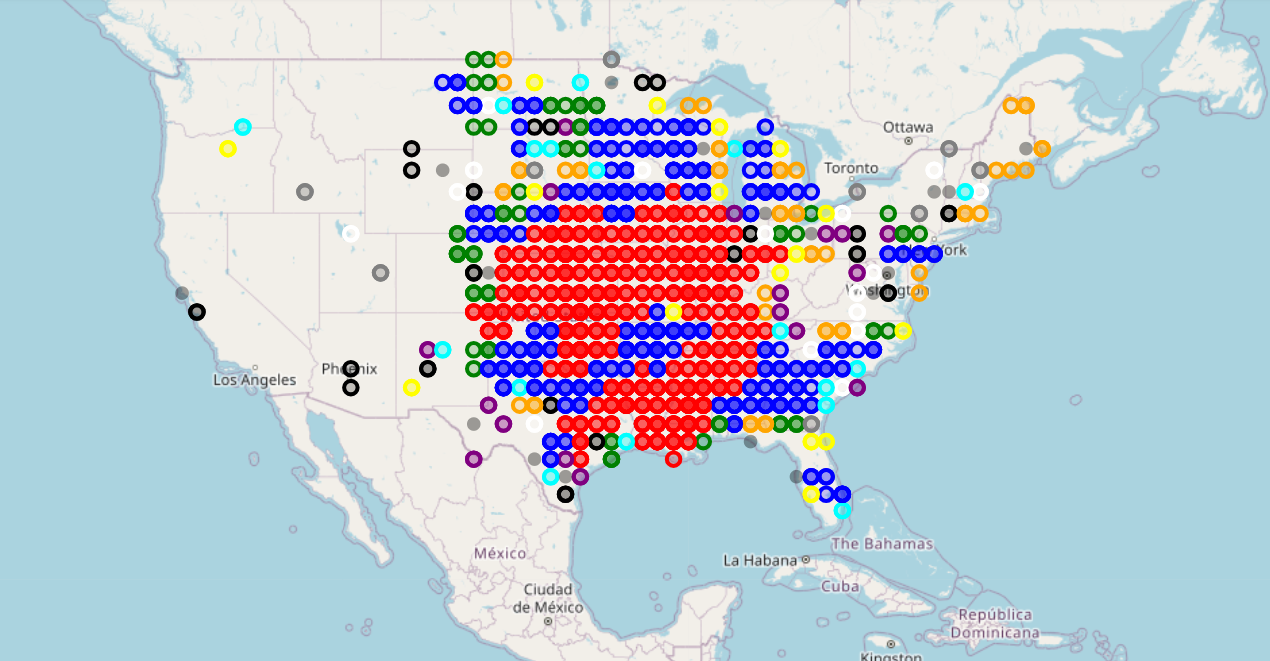


ערך מזהה הקהילה גבוהה:

תמונה שמכילה טקסט, צילום מסך, מקביל, מספר

התיאור נוצר באופן אוטומטי

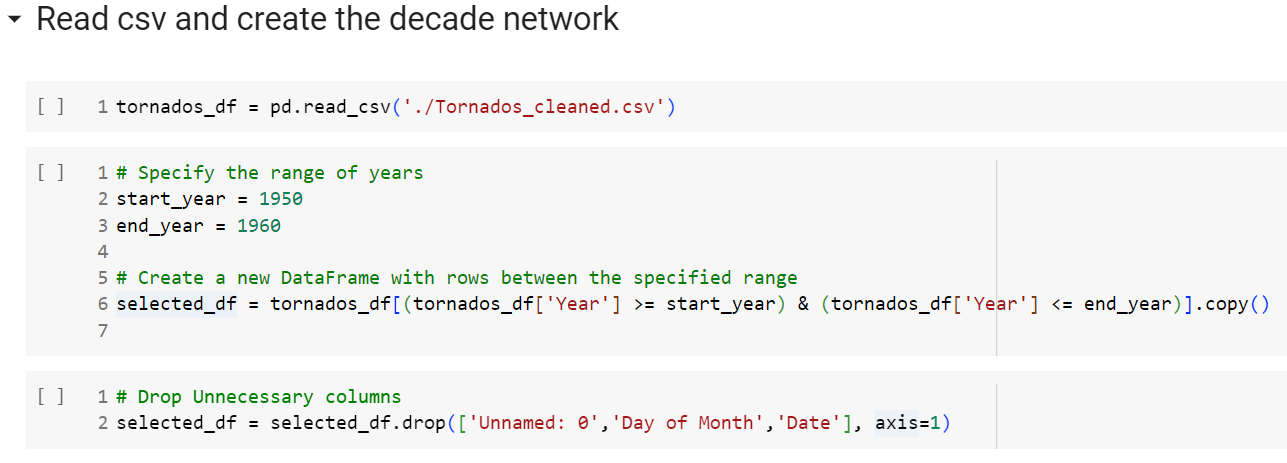
לאחר הרצת הקורלציה של edges\_df שמנו לב כי יש קורלציה שלילית בין מספר הקהילה לבין אורך הטורנדו (-0.28) כפי שציפינו. בנוסף, שמנו לב כי קיים קשר (אמנם חלש) בין סקאלת הEF לבין מספר הקהילה (-0.15).

על מנת להבין לעומק את הממצאים של הקהילות, לנתח את הקהילה המשמעותית ביותר, ולהבין האם הממצאים שלנו הם מבוססים, הדפסנו את אחד עשר הקהילות המובילות על מפת ארה"ב, כל אחת בצבע שונה עבור העשור הנוכחי:

כעת הבנו כי קהילה מספר 0 (באדום) וקהילה מספר 1 (כחול) בגודלן אכן מהוות את השטח tornado valley. הממצאים שלנו התבססו. שאר הקהילות מפוזרות יותר במפה ומקיפות את האזור המשמעותי באמצעי במפה. כמו כן, הן בעלות מספר קטן יותר של סופות טורנדו.

**ניתוח הרשת הדינאמית (רשת לכל עשור של דאטה) –**

על מנת לנתח את הרשת הדינאמית הרצנו את התוכנית המקורית ע"י הגדרה מחדש של שנת התחלה וסיום בקטע הקוד הבא ושמנו דגש על הבלוקים החשובים: (שורות 2 ו 3 בבלוק השני)



ניתוח הרשת הדינאמית הוא חלק פשוט בפרויקט שלנו מכיוון שכל העבודה הקשה נעשתה כבר כאשר חקרנו את הרשת הראשונה. לשמחתנו רוב הדברים שציינו בחלק הקודם לא השתנו יותר מדי. אז ניתן לומר יותר בביטחון את הדברים שציינו עד כה במחקר: הקשר החיובי בין ממדי ה edge closeness וה degree לבין מדדי הטורנדו, שימוש באלגוריתם greedy modularity, חילוק לקהילות ולאחד אותן לפי ערכי מזהי הקהילות ועל הקשר השלילי בין ערך מזהה הקהילה לבין מדדי הטורנדו.

בנוסף הצלחנו למצוא כי:

* יש מגמת עלייה במיוחד ב 40 שנה האחרונות במתאם בין מדדי ה edge closeness/degree למספר הפצועים וההרוגים בסופות טורנדו.
* בעשורים האחרונים חלה עלייה משמעותית במספר הצמתים שמקבלים את מספרי 0, 1 ו 2 במזהי הקהילות ובגלל הקשר השלילי של הערכים הללו למדדי הטורנדו ניתן להסיק שבתקופה האחרונה סופות הטורנדו שנצפו וכנראה יתרחשו בעתיד בארה"ב יהיו מסוכנות יותר מכל בחינה.
* מזכירים כי למרות שאין לנו קובץ נתונים ידוע שב 20 שנה האחרונות הטמפרטורה בכדור הארץ הולכת ועולה. לכן ייתכן שמשפיעה בפרט על סופות הטורנדו בארה"ב ובגלל זה ניתן לראות את זה במגמות במהלך השנים כאשר משווים בין הרשתות.

|  |  |  |  |
| --- | --- | --- | --- |
| 1981-1990 | 1991-2000 | 2001-2010 | 2011-2021 |
|  | תמונה שמכילה צילום מסך, מפה, טקסט  התיאור נוצר באופן אוטומטי | תמונה שמכילה צילום מסך, מפה, תרשים, טקסט  התיאור נוצר באופן אוטומטי | תמונה שמכילה מפה, תרשים, צילום מסך, טקסט  התיאור נוצר באופן אוטומטי |
| 1971-1980 | 1961-1970 | 1950-1960 |  |
|  |  |  |  |

**דיון –**

בפרויקט זה התמקדנו בעיקר בנתוני רשת ומטרתו בעצם לפתוח דלתות למחקרים עתידיים על הנושא של סופות טורנדו באמצעות אלגוריתמים של רשתות. לכן מיקדנו את מירב המאמצים בניתוח המידע שברשותנו ע"י מדדי רשת וקהילות. מציינים כי בהחלט יש מקום לשלב מודלים הקשורים למדעי נתוני עתק על מנת לקבל תוצאות מדויקות יותר. בנוסף רוצים לציין כי ייתכן שיהיו דרכים אחרות הקשורות לאלגוריתמים ברשתות ממה שנעשה בפרויקט שיתרמו להבנת טובה ומקיפה יותר של התהליכים של סופות טורנדו.

**סיכום –**

בפרויקט זה רצינו לבחון את האפשרות של התבססות על מדדי רשת ואלגוריתמים ברשתות בשביל מחקר בנושא של סופות טורנדו. השתמשנו במידע איכותי על סופות טורנדו שנצפו בארה"ב בין השנים 1950-2021. יצרנו ממנו רשת דינאמית לפי עשורים כך שמיקום מעוגל יהיה צומת, וקשת מכוונת בין מיקום התחלתי לסופי תהווה סופת טורנדו. כל המידע הרלוונטי יהיה שמור על הקשת. לאחר ניתוח על בסיס מדדי רשת גילינו שיש קשר בין ה closeness וה degree של הקשת לבין הנתונים של הטורנדו. לדוגמה, ככל שהמדדים הללו גבוהים יותר ייתכן כי הסופה תעבור מסלול ארוך יותר, רוחב ומדד ה EF יהיו גדולים יותר וייתכן שיהיו יותר פצועים או הרוגים מסופות אחרות ברשת. לניתוח קהילות השתמשנו באלגוריתם greedy modularity שהראה תוצאות טובות יותר מ Girvan Newman ומ Louvaine. קיבצנו מחדש את הקהילות על מנת לקבל מספר קהילות נוח יותר לעבודה לאחר שגילינו כי ניתן לקבץ ערכים קרובים של מזהי קהילות. בניתוח הקהילות גילינו שעל פי האלגוריתם שהשתמשנו יש מתאם שלילי בין ערך מזהה הקהילה לנתוני הטורנדו. בניתוח הרשת הדינאמית גילינו שיש מגמת עליה של מספר הסופות שהאלגוריתם משייך אותם לקהילה שערכה נמוך מה שאומר שיש מגמת עליה במיוחד בשנים האחרונות שהטמפרטורה הממוצעת על כדור הארץ עולה של סופות מסוכנות יותר בעיקר באזור ה tornado alley.

**תרומה של כל אחד מחברי הצוות לפרויקט –**

עמרי ריינדורף – מציאת דאטה, ניקוי הדאטה, ניתוח דאטה ראשוני, כתיבת הדוח, הרצת אלגוריתמים, כתיבת תוצאות, קידוד פונקציות, קידוד מדדי רשת, קידוד מפות התחלה וסיום.

עומר שחורי – הרצת הקוד על כלל העשורים, כתיבת הדוח, כתיבת הסברים והערות, ניסוח הבעיה, פיתוח אלגוריתמים, כתיבת מסקנות, קידוד קהילות, כתיבת דיון, קידוד מפות לקהילות, סידור מסמכים והעלאה לgithub.