"טוניק": איסוף מודיעיני באמצעות שיטוט ברשתות ו-SNA

ד"ר רמי פוזיס, לירון קצ'קו, ברק חג'בי, ד"ר רוני שטרן ופרופ' אריאל פלנר[1]

ליאב סלע[2]

1. הקדמה

המאמר "Target oriented network intelligence collection: effective exploration of social networks", פורסם ביולי 2019 בגיליון ה- 22(4) של כתב העת World Wide Web ונכתב על ידי ד"ר רמי פוזיס, לירון קצ'קו, ברק חג'בי, ד"ר רוני שטרן ופרופ' אריאל פלנר. במאמר, הכותבים מציגים שיטה חדשה לאקספלורציה (שיטוט) ברשתות חברתיות אינטרנטיות לשם השגת מידע על מטרה מוגדרת.

בעידן הנוכחי יש שימוש נרחב במיצוי מידע על בסיס ווביגט, ובשימוש נרחב בכלים של ניתוח רשתות חברתיות (SNA). אי לכך, ההגיונות כמו גם הכלים הפרקטיים המוצעים במאמר, מקבלים משנה חשיבות.

השיטה, המכונה בפי הכותבים TONIC (Target oriented network intelligence collection- איסוף מודיעין ברשת מונחה מטרה), באה לענות על שלוש שאלות-

א. באיזה פרופילים מועיל לשוטט?
ב. איך לתעדף את סדר השיטוט?
ג. מתי שיטוט נוסף אינו כדאי עוד?

המחברים מחלקים את השיטה לשני אפיקים מרכזיים- חיפוש מוגבל המבוסס על איתור פרופילים מועילים הקשורים זה לזה במישרין, או חיפוש מורחב שיוצא מנקודת הנחה כי ייתכן שיש פרופיל מועיל, הגם שהוא לא נמצא במרחק של קפיצה אחת מהפרופיל המועיל האחרון שנמצא.

המחברים מגדירים מספר מטרות לשיטה, כשהשתיים המרכזיות הן איתור מספר גבוה ככל הניתן של פרופילים מועילים; ומקסום התועלת מהשיטה, קרי סך כל התמורות פחות סך כל העלויות של השיטוט.

הזרקור ייפתח בהסדרה מושגית, ולאחר מכן יוצגו הנחות היוריסטיות לחיפוש המוגבל; התאמת ההנחות ההיוריסטיות לחיפוש המורחב; שקלול על בסיס תועלת (המטרה השנייה); ולבסוף סיכום ומסקנות.

---

[1] הכותבים הם אנשי המחלקה להנדסת מערכות תוכנה ומידע באוניברסיטת בן גוריון בנגב.
[2] עוזר מחקר במכון לחקר המתודולוגיה של המודיעין.

2. **<u>הסדרה מושגית</u>**

להלן המושגים המרכזיים בהם נעשה שימוש במאמר:

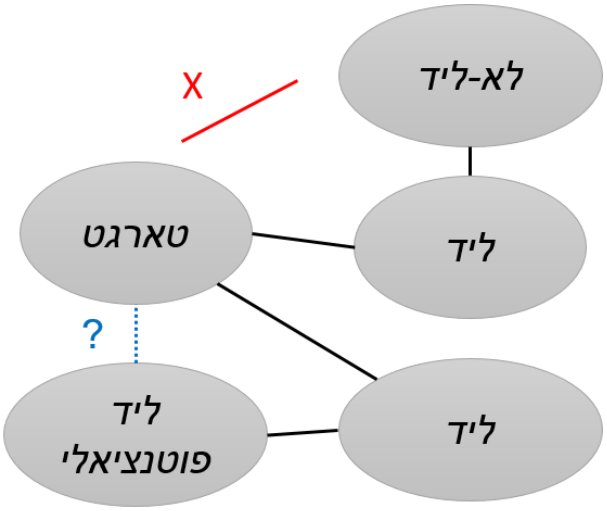
א. TONIC (*טוניק*) - שם השיטה שמשמעו איסוף מודיעין ברשת מונחה מטרה.

ב. מטרה (Target) - מטרת החיפוש. יכולה להיות איש, חברה או קבוצה, לה פרופיל ברשת החברתית.

ג. Lead - פרופיל המקושר *למטרה*, קרי בעל ערך מבחינת השיטוט.

ד. Potential Lead (*"ליד פוטנציאלי"*) - או בקיצור PL. פרופיל שטרם נבדק אך יכול להיות Lead.

איור 1:



ה. RTF- ראשי תיבות של Restricted TONIC Framework- הוא החיפוש המוגבל שמתבסס על קשר של קפיצה אחת בין *ליד* אחד לבין משנהו.

ו. ETF- ראשי תיבות של Extended TONIC Framework- הוא החיפוש המורחב, שלא מחויב בקפיצה אחת מ*ליד* אחד למשנהו.

ז. ETF(n)- במטרה ליצור שיטת ביניים בין ה-RTF לבין ה-ETF, הוגדרה ההגדרה הזו. משמעות ה- n, היא מספר הקפיצות שמותר לעשות בין *ליד* אחד למשנהו. לצורך העניין, ETF(0) אומר שאין אפשרות שיהיה אפילו פרופיל אחד לא רלוונטי בין *ליד* אחד לבין משנהו, שזה בעצם RTF. ETF(1) אם כך, הוא ההגבלה על פרופיל אחד שאיננו *ליד*, בין שני *לידים*. באופן הזה, ETF(∞) הוא בפשטות ETF בלי הגבלה על מספר הקפיצות.

ח. היוריסטיקה- כלל אצבע בעלות נמוכה לקבלת החלטות. במאמר הוא משמש כקו מנחה לביצוע החיפוש.

ט. isLead (*אזליד*)- פעולה בעלות נמוכה שבודקת האם הפרופיל הנוכחי הוא *ליד*, כלומר מקושר *למטרה*.

י. Acquire (*אקווייר*)- פעולה בעלות גבוהה שמרכישה את רשימת החברים של הפרופיל הנוכחי, במטרה לבדוק *אזליד*.

3. הנחות היוריסטיות לחיפוש המוגבל- RTF

במסגרת RTF הפעולה *"אקווייר"* תבוצע רק על פרופיל שאנחנו יודעים בוודאות שהוא *ליד*. השיטה היסודית ביותר למציאת *"לידים"* נוספים, היא בפשטות לבדוק אותם אחד-אחד בתור, לפי שיטת מי שנמצא ראשון נבדק ראשון וכן הלאה (ביצוע *איזליד* על כל *ליד פוטנציאלי* חדש שנמצא, על פי הסדר שהם נמצאו. אם אכן מדובר ב*ליד*, ביצוע *אקווייר*).

עם זאת, מטרת השיטה היא להיות יעילה ככל הניתן. אי לכך אפשר להגדיר מספר הנחות היוריסטיות שיעזרו לבדוק איזה *ליד פוטנציאלי* עדיף לבדוק קודם.

<u>מקדם האשכול (CC)</u>

הנחת המוצא היא שפרופילים לא מקושרים באופן אקראי זה לזה, אלא שהם שייכים לאשכולות מסוימים. לצורך העניין, סביר שאם ה*מטרה* עובדת בחברה מסוימת, תיווצר מעין רשת של פרופילים סביב העובדים בחברה הזו. אי לכך, ניתן לחשב באופן פשוט למדי את מידת הקרבה של *ליד פוטנציאלי* לאשכול שאנחנו מעוניינים בו (חישוב סך כל ה*לידים* המוכרים לנו שהוא קשור אליהם, חלקי סך כל הפרופילים שאליהם הוא מקושר). ככל שהתוצאה גבוהה יותר, כך הסיכוי ש*ליד הפוטנציאלי* מועיל למחקר שלנו עולה.

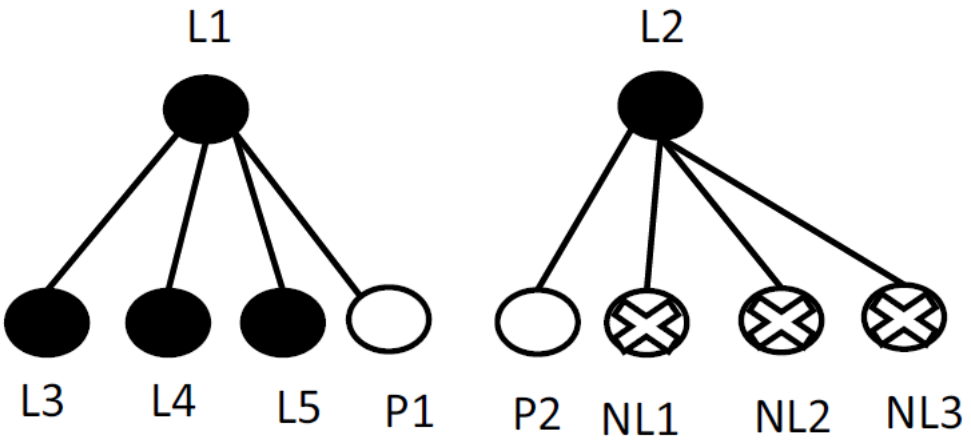
הבעיה בחישוב הזה, היא שהוא מתעלם מפופולריות של פרופיל. לצורך העניין, פרופיל שיש לו רק חבר אחד, שהוא במקרה *ליד* מוכר, יקבל את הציון המרבי (100%) במדד ה-CC, בעוד שפרופיל שיש לו 200 חברים, מתוכם 199 הם *לידים*, יקבל ציון נמוך יותר, חרף העובדה שהוא בסבירות גבוהה יועיל יותר. לכן, מוגדרת ההיוריסטיקה השנייה- KD.

<u>מידת המוכרות (KD)</u>

מידת המוכרות היא בפשטות המכנה של ה-CC, כלומר לכמה פרופילים ה*ליד הפוטנציאלי* מקושר. שקלול של ה-CC, יחד עם ה-KD, ינבא בצורה טובה יותר את ערכיות ה*ליד*.

<u>מדרג הסיכויים הטובים (PL)</u>

איור 2:



באיור 2, ניתן לראות כי ל-P1 ול-P2 (שניהם *לידים פוטנציאליים*) יש CC שווה (שניהם מקושרים *לליד* אחד מוכר כל אחד) וכך גם KD שווה (שניהם מקושרים בדיוק לפרופיל אחד נוסף, שבמקרה הזה הוא *ליד*). עם זאת, בבדיקה של *הליד* המוכר, ניתן לראות ש-L1 מקושר להרבה יותר *לידים* מוכרים בהשוואה ל- L2. אי לכך, כדאי לבדוק קודם את P1. זה למעשה מדד ה-PL.

כשמדובר ב*ליד* אחד, החישוב הוא טריוויאלי. עם זאת, כשכל *ליד פוטנציאלי* קשור למספר *לידים*, יש לדעת כיצד לשקלל את ציון ה-PL. שיטה אחת יכולה להיות ממוצע של ה-PL עבור כל *הלידים* אליהם *הליד הפוטנציאלי* קשור (ובדיקת *הליד הפוטנציאלי* בעל הממוצע הגבוה יותר). שיטה שנייה יכולה להיות בחינה של ה-PL המקסימלי עבור כל אחד מ*הלידים* אליהם קשורים *הלידים הפוטנציאליים*, ובדיקת *הליד הפוטנציאלי* אליו הוא קשור.

עם זאת, במאמר מוצעת שיטה שלישית, והיא הערכית ביותר כפי שיוצג בהמשך, והיא השיטה הבייסיאנית. באופן כללי, השיטה הזו מחשבת את ה-PL של כל *ליד*, כתלות ב-PL של *הליד* שלפניו (על בסיס נוסחת בייס). שיטה זו מכונה במאמר BysP, והיא מלווה בנוסחה מתמטית.

<u>איכות החברים (FM)</u>

הרכיב ההיוריסטי האחרון בא לבדוק מה הסיכוי *שליד פוטנציאלי* יהיה קשור *לליד* אחר (שבעצמו קשור למטרה). הרעיון מתכתב עם הגיון הקהילתיות שהוצג קודם. במאמר אף מוצג חישוב לבדיקת מדד זה, שמתבסס באופן כללי על כמות הקפיצות שנעשו עד כה כדי להגיע *לליד הפוטנציאלי* (היות שהרשת אינה ממופה, אי אפשר לדעת זו בדרך ישירה יותר כמו הצלבה של רשימות חברים).

<u>סיכום RTF</u>

מבחינת כלל הפרמטרים, הניסוי הראה כי BysP, KD ו-FM, הם הפרמטרים המועילים ביותר לשיטוט. עם זאת, כלל הפרמטרים ב-RTF מגיעים לגבול מסוים (כ- 70%), משום שהם לא מחפשים פרופילים רחוקים יותר שיכולים להיות קשורים למטרה. כאן, נכנס למערכה החיפוש המורחב- ETF.

4. <u>התאמת ההנחות ההיוריסטיות לחיפוש המורחב- ETF</u>

באופן כללי, כשניגשים לעסוק ב-ETF, יש שני רכיבים שכדאי להגדיר- מספר *הלידים* ההתחלתיים שאנחנו מכירים (על בסיס חיפוש איכותי למשל), ומספר הקפיצות שאנחנו מעוניינים לאפשר (n, כפי שהוזכר בהסדרה המושגית). בדיקות של הניסוי הראו שהמספר האידיאלי *ללידים* התחלתיים הוא 3 (כל אחד נוסף הוא בבחינת תפוקה שולית פוחתת), ובהתאם מספר הקפיצות האידיאלי הוא 1 (כלומר, מותר שיהיה *לא-ליד* אחד בין שני *לידים*).

בניסוי, אפשר לראות כי הרחבת חיפוש על בסיס תור (כפי שהוזכרה בתחילת הסעיף הקודם) היא יסודית יותר ומגיעה לתוצאות טובות יותר בהשוואה לשימוש בהיוריסטיקות, אך משמעותית פחות יעילה משימוש בהיוריסטיקות. אי לכך, כדאי לבחון מודל משולב, במסגרתו יבוצע שימוש בהיוריסטיקות כמו הBysP (או בגרסתו המורחבת הEBysP), עד לנקודה בה מספר *הלא-לידים* אליהם הגענו הגיע לרף מסוים, ובה נעבור לבחון את הנתונים בתור.

לצורך העניין, אם *המטרה* היא רופא, שימוש בשיטה הבייסיאנית ישיג במהירות גבוהה את מקביליו המקצועיים ואת משפחתו (היות שיש קשר סיבתי בין החברות שלו עמם, ובין החברות שלהם אחד עם השני). עם זאת, בשלב מסוים, כשמרבית הפרופילים האלה יימצאו, השיטה הבייסיאנית תהפוך ליעילה פחות, וזה יהיה השלב לעבור לשיטת התור (לבדוק אותם אחד-אחד), על מנת לאתר פרופילים בודדים שאינם קשורים כקבוצה בקשר סיבתי לרופא, כמו למשל מטופליו.

5. <u>שקלול על בסיס תועלת</u>

עד כה עסקנו במציאת מספר רב ככל הניתן של*לידים*. עם זאת, נשאלת השאלה מתי כל*ליד* נוסף שנמצא יהיה לא אפקטיבי בהשוואה למידת ההשקעה שהושקעה במציאתו. לשם כך, יש להגדיר תנאי עצירה, כלומר תנאי מסוים שאם הוא ממומש, המערכת מפסיקה את החיפוש. במאמר, הוצגו שתי אופציות לתנאי עצירה-

<u>אופציה ראשונה לתנאי עצירה- ZeroR</u>

במסגרת אופציה זו, מוגדר מראש מספר מסוים של בדיקות שהמערכת מבצעת עד שהיא עוצרת. המספר הזה, יהיה מדויק יותר ככל שתבוצענה יותר הרצות באותם פרמטרים התחלתיים, והדבר ניתן לכיול על ידי בניית כלי פשוט של למידת מכונה.

<u>אופציה שנייה לתנאי עצירה- LearnDynamic</u>

במסגרת אופציה זו, כותבים כלי של למידת מכונה שבוחן עבור כל בדיקה האם היא כדאית או לא. על הכלי להתבסס על שלושה פרמטרים- מספר *הלידים* שהורכשו עד כה בהרצה הנוכחית; היחס שבין *הלידים* לבין *הלא-לידים* עד כה בהרצה הנוכחית; והיחס בין *הלידים* שהורכשו לבין *הלידים הפוטנציאליים*.

גם בהיעדר כלי ללמידת מכונה, המתודה הזו מהווה קו מחשבתי מעניין. לצורך העניין, אם עוסקים בניתוח רשתות חברתיות (SNA) על בסיס מודיעין ללא תוכן (מדל״ת), מעקב אחר הפרמטרים שהוזכרו, יכול ללמד רבות על מידת היעילות של המשך מיצוי המדל״ת לעומת סיכום המסקנות שהושגו עד כה.

ישנן שתי בעיות באופציה הזו, ושתיהן מתייחסות למסד נתונים (DB) לא מאוזן. הראשונה, היא מצב בו מסד הנתונים כולו ערכי, ואז המערכת לא תעצור לעולם. השנייה היא מצב בו מסד הנתונים לא ערכי ככלל, ואז המערכת תתייאש מוקדם מדי.

<u>סיכום תנאי עצירה</u>

תנאי העצירה מדבר על ההפרש בין כמות ההצלחות לבין כמות הניסיונות. אי לכך, שניהם רגישים מאוד לציון הניתן עבור כל הצלחה ועבור כל ניסיון. לצורך העניין, אם כל*ליד* חדש הוא בגדר ידיעת זהב, אזי כדאי לתת ערך גבוה למציאת *ליד*, וערך נמוך עבור כל ניסיון. ראו שתי דוגמאות בהמשך הפרק.

הניסוי הראה שאם הוגדרה עלות נמוכה עבור כל ניסיון (קרי שכל*אקוויירר* קיבל ערך נמוך יחסית), עדיפה השיטה השנייה (LearnDynamic). בהתאם לכך, אם הוגדרה עלות גבוהה לכל

ניסיון, אזי עדיפה השיטה הראשונה (ZeroR). מעניין לראות שמסקנה זו אינה תלויה בערך שהוגדר למציאת *לידים* (קרי הצלחה).

לשם המחשה, אפשר לתאר שני מושאי מחקר (שלמעשה מהווים את *המטרות*)-

מושא המחקר הראשון הוא בכיר בארגון טרור קטן וחדש שעלה לראש הציי"ח ושהמידע עליו מועט מאוד. במקרה כזה, בו יש מעט קצות חוט, כל *ליד* חדש שיימצא יהווה ערך של ממש. במקביל, כל ניסיון יהיה בעלות נמוכה, משום שהוא לא לוקח משאבים ממושאי מחקר דחופים יותר (כאמור, הארגון הזה נמצא בראש הציי"ח). במקרה הזה, עדיף לפי המחקר להגדיר את תנאי העצירה לפי LearnDynamic.

מושא מחקר שני הוא גורם זוטר בממשלה כלשהי שנמצאת במרחב ההתעניינות אך לא במרחב האחריות. במקרה זה, מציאת *לידים* חדשים, אומנם מועילה, אך פחות מהמקרה הראשון. באופן דומה, כל משאב שיושקע בחיפוש שלה, הוא משאב שבא על חשבון נושאים אחרים דחופים יותר, ועל כן עלות כל חיפוש גבוהה יותר. במקרה הזה, עדיף לפי המחקר להגדיר את תנאי העצירה לפי ZeroR.

6. <u>סיכום והשלכות אתיות</u>

בבחינה של השיטות שהוזכרו על כלל הפרמטרים שלהן במספר רשתות חברתיות שונות, נראה שהוסקו שלוש מסקנות מרכזיות-

א. ל-RTF יש סף אליו הוא יכול להגיע אך לא לעבור, בשונה מ-ETF. אי לכך, מבחינת המטרה הראשונה (כמות *לידים*), ETF עדיף על RTF.
ב. BysP ו-EBysP הם ההנחות ההיוריסטיות המועילות ביותר בכלל המצבים שנבדקו.
ג. EBysP אפקטיבית יותר מ-BysP, גם אם מתעלמים מהסף העליון של RTF באופן כללי (קרי גם מתחתיו, הראשון יעיל יותר מהאחרון).

הכותבים מצאו לנכון להזכיר את הבעייתיות האתית שבשימוש בשיטת *הטוניק*. כך למשל, השיטה מתעלמת לחלוטין מהגבלות הפרטיות של המשתמשים ברשתות החברתיות ולמעשה עוקפת אותן, ובמקביל השיטה יכולה להיות בשימוש גורמים עוינים.

המאמר מופנה לשיטוט על בסיס רשתות חברתיות אינטרנטיות, ומסקנותיו עשויות לשפר את המתודה האיסופית הזו. עם זאת, מסקנות המחקר יכולות לשמש גם מחקרים על בסיס פלטפורמות אחרות.

כך למשל, ניתוחי מדל"ת למיניהם, יכולים בהחלט לשאוב השראה הן מההיוריסטיקות שהוצגו במאמר (למשל, שימוש בנוסחה מתמטית לחישוב תלות על מנת לייעל את החיפוש אחר קצ"חים נוספים) כמו גם מתנאי העצירה שהוצגו במאמר (כפי שהוצג בדוגמה שבסוף הפרק הקודם).