

**Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data**

מנחה: פרופ' תמיר טסה

מגיש העבודה: נדב שור

מס' סטודנט: 200954675

**סמינר בקריפטוגרפיה**

**20374**

**תוכן עניינים**

1. הקדמה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_3
2. מבוא \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_3
   1. סקירת כרייה פרטית של חוקי הקשר \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_4
3. רקע ועבודות קשורות \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5
   1. כריית חוקי הקשר \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5
      1. כרייה מבוזרת של חוקי הקשר \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5
   2. חישוב מבוזר מאובטח \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6
      1. בטיחות במודל ישרים-למחצה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6
      2. הפרוטוקול של Yao לחישוב מבוזר מאובטח של פונקציה בעלת שני קלטים \_\_\_\_\_\_\_\_\_\_\_6
   3. הצפנה קומוטטיבית \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6
4. כרייה בטוחה של חוקי הקשר \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_7
   1. הגדרת הבעיה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_7
   2. שיטה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_7
      1. איחוד בטוח של סלי-פריטים שנתמכים מקומית \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_7
      2. בדיקת סף התמיכה ללא חשיפה של מספר התומכים \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_9
   3. מציאת ביטחון בחוק הקשר בצורה בטוחה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10
5. חולשות הפרוטוקול \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10
   1. בטיחות כנגד קנוניה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10
   2. קשיים במקרה של שני שותפים \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10
6. עלויות תקשורת וחישוב \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10
   1. אופטימיזציות ודיון נוסף \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11
7. מסקנות ועבודה נוספת \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11

רשימת ביבליוגרפיה \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12

סיכום המאמר

Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data [1]

מאת

Murat Kantarcioglu and Chris Clifton, Senior Member, IEEE

1. הקדמה

המאמר, שמסוכם בהמשך, עוסק בביצוע כריית מידע מבוזרת של חוקי הקשר בצורה בטוחה. ההגדרה של המושג "כריית מידע" ע"פ הערך בויקיפדיה – "כריית מידע ... היא הפעלת אלגוריתם או תוכנית מחשב לצורך גילוי מידע הטמון בבסיסי נתונים קיימים, והסקת מסקנות מהצלבתו" [2].

אוספי המידע העצומים שקיימים היום גורמים לכך שכריית מידע היא אחד מהתחומים המתפתחים היום הן בתעשייה והן באקדמיה. ניתן למצוא את כריית המידע בתחום זיהוי פנים, סיווג מקרים חדשים על בסיס מקרי עבר, למידת חוקי הקשר ועוד.

הצורך בכריית מידע בטוחה קיים גם הוא בתחומים רבים. לדוגמה: שיתוף מידע בין סוכנויות ביטחון, שימוש במידע של בתי-חולים על החולים שבהם כדי למצוא קשר בין תרופות ותופעות לוואי וכדומה. לכן, דרוש פרוטוקול יעיל לביצוע כריית מידע מבוזרת של חוקי הקשר בצורה בטוחה.

הסיכום שמובא בהמשך מחולק לפרקים, כאשר בפרקים 1 ו-2 מוצגים הרקע והצורך לפרוטוקול שמוצג במאמר. בפרק 3 מוצג הפרוטוקול עצמו ובפרק 4 מוצגות החולשות של הפרוטוקול. לבסוף, בפרק 5 מבוצע ניתוח של עלות התקשורת והחישוב של הפרוטוקול ופרק 6 מסכם.

1. מבוא

מרבית הכלים לכריית מידע פועלים ע"י איסוף כלל הנתונים למאגר מידע יחיד והרצת אלגוריתם שמוצא חוקי הקשר. הבעיה שמאמר זה מנסה לפתור היא מציאת חוקי הקשר כאשר אין אפשרות לאגד את כל המידע תחת מאגר יחיד מפאת החשיבות של פרטיות המידע.

כדי להסביר את המושגים שנשתמש בהם בהמשך נמשיל את הבעיה לשיתוף פעולה בין חברות סופר-מרקט שונות. לכל סופר-מרקט יש מאגר מידע שמכיל *רשומות*. כל רשומה מכילה את המצרכים שנקנו ע"י לקוח ברכישה ספציפית. *סל-פריטים* מוגדר כקבוצה כלשהי, לא ריקה, של מצרכים שנמכרים בסופר-מרקט. יש לשים לב שגם רשומה וגם סל-פריטים מכילים אוסף כלשהו של מצרכים.

אנו מניחים *אחידות מידע*, שמשמעה בדוגמה היא שכל הפריטים שנמכרים בסופרים זהים. כלומר, לא קיים פריט שנמצא רק בחלק מהסופרים. בהתבסס על מאגרי המידע של כל הסופרים, חברות הסופר-מרקט מעוניינות להבין אילו פריטים נקנים בשכיחות גבוהה.

כדי להבין האם סל-פריטים כלשהו נקנה בשכיחות גבוהה, חברות הסופר-מרקט מונות את מספר הרשומות שמכילות סל פריטים מסויים – זוהי ה*תמיכה של סל-הפריטים*. *שכיחות סל-הפריטים* היא אחוז הרשומות שמכילות את סל-הפריטים הזה, כלומר התמיכה של סל-הפריטים חלקי גודל מאגר המידע.

בהינתן שכיחות מבוקשת s, *סל-פריטים שכיח* מוגדר כסל-פריטים שהשכיחות שלו גדולה מ-s. אם לסל-פריטים נתון יש שכיחות גדולה מ-s בסופר-מרקט i הוא נקרא *נתמך מקומית באתר i*. סל-פריטים נקרא *נתמך גלובלית* אם סכום התמיכה שלו בכל הסופרים חלקי סכום גדלי מאגרי המידע גדול מ-s.

מושג נוסף שנתייחס אליו הוא *חוק-הקשר.* בהינתן שני פריטים A ו-B, חוק ההקשר A⇒B משמעו שאם פריט Aנמצא בסל-פריטים (או רשומה) מסויים, אז גם פריט B נמצא בסל-פריטים זה.

*אומדן הביטחון של חוק-ההקשר* "A⇒B" הוא c, אם c% מהרשומות שמכילות את הפריט A מכילות גם את הפריט B. ניתן להבדיל בין אומדן הביטחון המקומי לאומדן הביטחון הגלובלי, אך מכיוון שבמאמר יש התייחסות רק לאומדן הביטחון הגלובלי הוא יקרא גם כאן "אומדן הביטחון" בלבד. *התמיכה המקומית של חוק-ההקשר* "A⇒B" מוגדרת כתמיכה המקומית של סל הפריטים {B,A} ובדומה מוגדרת התמיכה הגלובלית של חוק-ההקשר. יש להדגיש שההבדל בין חוק-ההקשר "A⇒B" לחוק-ההקשר "B⇒A" מתבטא אך ורק באומדן הביטחון של חוק-ההקשר ולא בתמיכה המקומית/גלובלית של חוק-ההקשר.

חברות הסופר-מרקט מחפשות את כל חוקי-ההקשר שאומדן הביטחון שלהם גדול מ-c נתון, ושהתמיכה הגלובלית שלהם גדולה מ-s נתון. כפי שניתן לראות מההגדרות, בדיקת התנאי הנ"ל על חוקי-ההקשר שקול לבדיקתו על סלי-הפריטים המתאימים – התמיכה הגלובלית של חוק-ההקשר "X⇒Y" שקולה לתמיכה הגלובלית של סל-הפריטים {X,Y} ואומדן הביטחון של חוק-ההקשר שקול לתמיכה הגלובלית של סל-הפריטים {X,Y} חלקי התמיכה הגלובלית של סל-הפריטים {X}. לכן, מעתה והלאה נתייחס לסלי-פריטים בלבד, ולא לחוקי הקשר (פרט להתייחסות קצרה בסעיף 2.1 בו נגדיר את הבעיה בצורה מתמטית).

הנוסחאות הבאות מסכמות בצורה מתמטית את המושגים שהוגדרו מעלה, כאשר מייצג את התמיכה המקומית של אתר i בסל-הפריטים {A,B}, מייצג את גודל מאגר המידע של אתר i, מייצג את אומדן הביטחון של חוק-ההקשר "AB⇒C" ו-n מייצג את מספר השותפים.

הבעייה שמנסה לפתור המאמר היא מציאת כל סלי הפריטים הנתמכים גלובלית באחוז גבוה מ-k נתון כלשהו, תחת המגבלה שלא ניתן ליצר מאגר יחיד. המאמר מניח את שתי ההנחות הבאות: אחידות המידע וקיומם של לפחות שלושה משתתפים. אחידות המידע משמעה שלכל האתרים יש סכמה זהה לשמירת המידע, אך לכל אתר יש רשומות שונות שאת המידע שלהן הוא שומר אצלו.

מטרת הדרישה לשלושה משתתפים לפחות היא למנוע מצב בו יש חוק הקשר שנתמך בצורה גלובלית ולא נתמך אצל משתתף 1 ולכן משתתף 1 יכול להסיק כי הוא נתמך אצל משתתף 2.

כדי למצוא את כל סלי-הפריטים שנתמכים גלובלית באחוז גבוה מ-k, כל אתר שולח את סלי-הפריטים שהתמיכה המקומית שלו בהם גדולה מ-k%. עבור כל סל-פריטים כזה, בעזרת התמיכה המקומית של כל האתרים בסל-הפריטים, מתבצע חישוב התמיכה הגלובלית ואומדן הביטחון (אופן החישוב מפורט בסעיפים 3.2.2 ו-3.3).

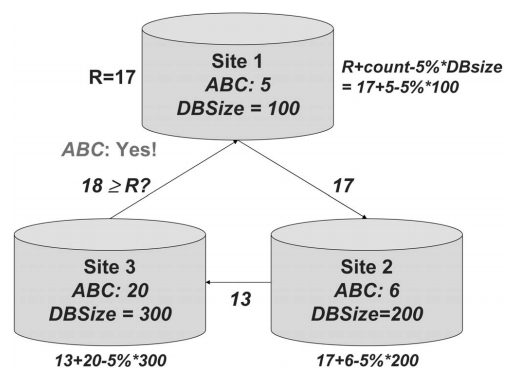
המאמר מציג פתרון שמאפשר לצדדים ללמוד את התוצאות הכלליות בלי ללמוד (כמעט) כלום על הנתונים של השותפים האחרים. העלות הנוספת של הפתרון לינארית ביחס למכפלת מספר המשתתפים וסכום סלי-הפריטים שנתמכים מקומית (בכל המאגרים).

* 1. סקירת כרייה פרטית של חוקי הקשר

השיטה שמוצגת במאמר מורכבת משני שלבים – הראשון, גילוי סלי-הפריטים שנתמכים מקומית. בשלב השני מוודאים אילו מסלי-הפריטים שהועלו בשלב הראשון נתמכים גלובלית.

השלב הראשון משתמש בהצפנה קומוטטיבית (מוסבר בסעיף 2.3). כל אתר מצפין את סלי-הפריטים בהם הוא תומך מקומית. המידע המוצפן מועבר בין כל האתרים כך שכל אתר מצפין פעם אחת את כל המידע. לבסוף, מתקבל מאגר של מידע מוצפן שכולל את כל סלי-הפריטים שנתמכים מקומית. מבוצעת הסרה של כפילויות, כך שנשארים סלי-הפריטים המוצפנים שמועמדים לבחינה עבור תמיכה גלובלית. סלי-הפריטים האלו עוברים, שוב, בין כל האתרים שהפעם מפענחים את המידע. בסוף התהליך מתקבלת קבוצת סלי-פריטים לבדיקה.

בשלב השני, כל סל-פריטים שמועמד להיות סל-פריטים עם תמיכה גלובלית נבדק לתמיכה גלובלית. דוגמה לתהליך הבדיקה ניתן לראות באיור 1. עבור על כל סל-פריטים שנבדק, כל אתר מחשב את התמיכה שלו בסל-הפריטים. לאחר מכן, בוחר האתר הראשון ערך אקראי וסודי כלשהו R ומוסיף ל-R את ההפרש בין ערך התמיכה המקומית בסל-פריטים לסף שהוגדר עבור תמיכה גלובלית. האתר הראשון מעביר את הערך שיצא לאתר השני שמבצע חישוב זהה ומעביר לאתר הבא (וכך הלאה לאתרים הבאים). כשהאתר האחרון מסיים לבצע את החישוב, מתבצעת השוואה בין ה-R שבחר האתר הראשון לבין התוצאה שהתקבלה. אם הערך שהתקבל קטן מ-R אז סל-הפריטים לא נתמך גלובלית, אחרת, סל-הפריטים נתמך גלובלית.



**איור 1**: בדיקה לתמיכה גלובלית של סל-פריטים עבור סף של 5%

1. רקע ועבודות קשורות

ישנן מספר עבודות קודמות בתחום כריית מידע תוך שמירה על פרטיות המידע [3]. בעבודות אלו נקטו החוקרים בשיטה של שיבוש המידע, תוך הנחה כי שיבוש המידע לא חושף דבר על המידע המקורי. המידע המשובש ביחד עם המידע על התחלקות המידע האקראי ששימש לשיבוש המידע, מאפשר להשיג קירובים לחלוקת המידע המקורי ללא חשיפת הערכים המקוריים. ההנחה בעבודות האלו היא שהמידע חייב להישאר חסוי כך שהגורם שכורה את המידע לא יכול לקבל את הערכים המקוריים. לעומת זאת, במאמר הנוכחי, מאפשרים החוקרים להראות חלק מהמידע לחלק מהמשתתפים כך שאף אחד מהמשתתפים לא רואה את כל המידע, מה שמאפשר קבלת תוצאות מדוייקות ולא קירובים.

* 1. כריית חוקי הקשר

בעיית כריית חוקי הקשר מוגדרת כך [4]: יהי - קבוצת פריטים ויהי DB, מאגר מידע המורכב מסט של רשומות, כך שכל רשומה T מורכבת ממספר פריטים (). בהינתן סל-פריטים X (), נגיד ש-T מכיל את X אמ"מ . חוק הקשר הוא מהצורה כאשר , ו- . לחוק יש תמיכה (support) s במאגר המידע אם s% מהרשומות במאגר מכילות גם את X וגם את Y. לחוק יש אומדן ביטחון (confidence) c, אם c% מתוך הרשומות במאגר המידע שמכילות את X מכילות גם את Y. בעיית כריית חוקי הקשר היא מציאת כל החוקים שהתמיכה שלהם והביטחון שלהם עוברים סף מסויים שהוגדר מראש.

* + 1. כרייה מבוזרת של חוקי הקשר

בעיית כריית חוקי הקשר שהוצגה מעלה יכולה להתרחב לסביבה מבוזרת. יהי DB מאגר נתונים מחולק אופקית (הטורים בכל תתי-המאגרים זהים אך הרשומות שונות) בין n אתרים כך ש-

וכל מאוחסן באתר i. לסל-הפריטים X יש תמיכה מקומית של אם מהרשומות ב- מכילות את X. התמיכה הגלובלית של X מחושבת ע"י סכימת התמיכה המקומית של X בכל האתרים, . סל-פריטים נחשב כנתמך גלובלית אם

. הביטחון הכללי של חוק הקשר מחושב ע"י .

מכיל את כל סלי-הפריטים שנתמכים גלובלית ומכילים k פריטים. מכיל את כל סלי-הפריטים שנתמכים מקומית באתר i ומכילים k פריטים. הסל מכיל את כל סלי-הפריטים שנתמכים גלובלית וגם נתמכים מקומית באתר i ומכילים k פריטים (). מטרת הכרייה המבוזרת של חוקי הקשר היא למצוא את החוקים שעוברים את סף התמיכה והביטחון שהוגדר ע"י מציאת עבור כל .

קיים אלגוריתם מהיר לכרייה מבוזרת של חוקי הקשר בשם FDM [5]. זהו אלגוריתם שעליו מבוססת השיטה במאמר זה. סיכום האלגוריתם:

1. יצירת סט מועמדים על בסיס תוך שימוש באלגוריתם א-פריורי (מוסבר בפסקה הבאה).
2. לכל יש לבדוק האם הוא נתמך מקומית ב-i. אם כן, להוסיפו ל-.
3. כל אתר i משדר את ה- שלו לכל שאר האתרים. כל אתר מחשב את התמיכה המקומית שלו בכל סל-פריטים ששיך ל- .
4. כל אתר משדר את תוצאות החישוב שלו מהשלב הקודם לכל שאר האתרים. כעת כל אתר בנפרד יכול לחשב את .

אלגוריתם א-פריורי הוא אלגוריתם שבהינתן מאגר מידע עם רשומות וסף תמיכה, מוצא חוקי הקשר במאגר שעוברים את סף התמיכה המבוקש. בכל שלב, האלגוריתם מחפש סלי-פריטים הגדולים ב-1 מסלי הפריטים שנמצאו בשלב הקודם. לשם כך הוא מתבסס על קבוצת המועמדים שהוא מצא בשלב הקודם, מכיוון שאם סל-פריטים כלשהו נמצא לא מתאים בשלב הקודם (מבחינת תמיכה) אז צירוף של עוד פריטים אליו יכול רק להרע את התמיכה ולכן בכל מקרה לא יעבור את הסף הדרוש.

* 1. חישוב מבוזר מאובטח
     1. בטיחות במודל ישרים-למחצה

צד ישר-למחצה מבצע את הפרוטוקול ע"פ חוקי הפרוטוקול תוך שימוש בקלטים נכונים, אך הוא חופשי לבצע שימוש מאוחר בכל המידע שהוא נחשף אליו תו"כ ביצוע הפרוטוקול. מודל כזה מדמה מצב אמת בו הצדדים מעוניינים בתוצאות לשימוש משותף ולכן יבצעו את הפרוטוקול עם הקלטים הנכונים.

הבטיחות של הפרוטוקול לא מאפשרת להגיד שהמידע הפרטי מוגן מכיוון שהסתכלות על קבוצת סלי-הפריטים שנתמכים מקומית וקבוצת סלי-הפריטים שנתמכים גלובלית מאפשרת להסיק אילו סלי-הפריטים נתמכים ע"י לפחות אתר אחד אחר. כמובן שככל שכמות המשתתפים גדלה המידע הנ"ל פחות ספציפי (לפחות אתר אחד מתוך n האתרים האחרים תומך ב-X). לעומת זאת, הפרוטוקול לא מאפשר הסקת מידע ספציפי על סל-פריטים כלשהו. לדוגמה, לא ניתן להסיק את כמות התומכים בסל-פריטים מסויים ע"י הסתכלות על קבוצת סלי-הפריטים שנתמכים מקומית וקבוצת סלי-הפריטים שנתמכים גלובלית.

* + 1. הפרוטוקול של Yao לחישוב מבוזר מאובטח של פונקציה בעלת שני קלטים

הפרוטוקול מאפשר לבצע חישוב בטוח של הפונקציה כמעגל עם שערים מוצפנים במצבים ישירם-למחצה [6]. הפרוטוקול מאפשר לפתור את בעיית המיליונרים של Yao ביעילות. בעיית המיליונרים של Yao מדברת על שני מליונרים, אליס ובוב, שמעוניינים לחשב למי יש הון גדול יותר מבלי לחשוף את ההון עצמו. כלומר, ביצוע חישוב בטוח עם שני-משתתפים עבור .

* 1. הצפנה קומוטטיבית

אלגוריתם הצפנה נקרא קומוטטיבי אם הוא מקיים את המשוואה הבאה עבור כל מפתח הצפנה אפשרי , עבור כל הודעה אפשרית M ועבור כל פרמוטציה של i, j:

תכונה זו של הצפנה קומוטטיבית מאפשרת לבדוק האם שני פריטים זהים מבלי לחשוף את הפריטים עצמם. הבדיקה נעשית כך: כל צד מצפין את הפריט שלו - , ושולח את הערך המוצפן לצד השני שמצפין את הערך שקיבל - . נוסחא (1) מבטיחה שאם אז גם .

בנוסף לתכונה שמתוארת מעלה, ההצפנה נדרשת להיות בטוחה – בהינתן סל-פריטים מוצפן לא ניתן להסיק דבר על סל-הפריטים. כלומר, אם ניתן שני סלי-פריטים מוצפנים לתוקף פולינומיאלי, ביחד עם סלי-הפריטים המקוריים, התוקף לא ידע לשייך את הערך המוצפן לערך המקורי.

1. כרייה בטוחה של חוקי הקשר

בפרק זה מוצג האלגוריתם שמציעים כותבי המאמר. האלגוריתם, שנסמך על הכלים שתוארו קודם לכן, משמש לכרייה מבוזרת של חוקי הקשר בצורה שמשמרת את פרטיות התוצאות באתרים עבור שלושה משתתפים ומעלה.

* 1. הגדרת הבעיה

יהי , מספר האתרים. כל אתר מחזיק מאגר מידע פרטי של רשומות . בהינתן אחוז סף של תמיכה s ואומדן ביטחון c, יש למצוא את כל חוקי ההקשר שמקיימים את הסף שהוגדר (כמו שהוגדר בסעיף 2.1.1). כמו כן, אתר כלשהו i לא יכול ללמוד שום מידע על אתר אחר j, פרט למידע שניתן להסיק מהמידע המקומי של האתר i והתוצאה הסופית של החישוב.

* 1. שיטה

השיטה שמוצגת במאמר מתבססת בכלליות על אלגוריתם FDM שהוצג בסעיף 2.1.1, כאשר החלק של שידור ו- מוחלף בפרוטוקול מיוחד. בהמשך יוצגו שני חלקי השיטה שמחליפים את שלבים 3 ו-4 באלגוריתם FDM: איחוד בטוח של ובדיקה האם סל-פריטים מסויים עומד בסף שהוגדר.

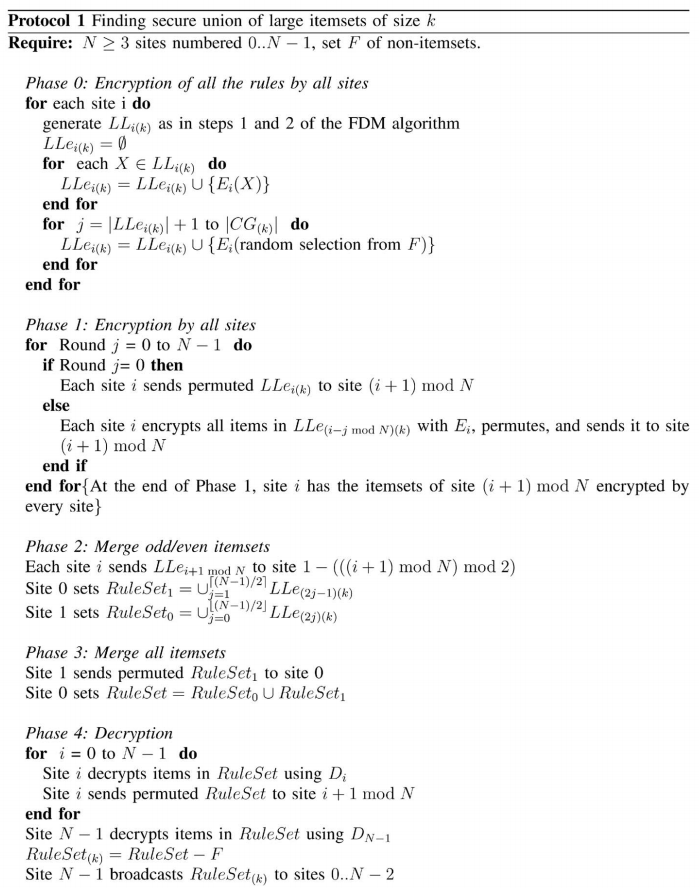
* + 1. איחוד בטוח של סלי-פריטים שנתמכים מקומית

בשלב 3 באלגורים FDM כל אתר חושף את שלו. כדי לאפשר את החלפת ה- ללא חשיפת התמיכה של כל אתר, מבוצעת החלפה בדרך שמטשטשת את המקור של כל סל-פריטים. במאמר מניחים שימוש באלגוריתם הצפנה קומוטטיבית עם הסתברות זניחה להתנגשויות (ע"פ סעיף 2.3).

הרעיון הכללי הוא שכל אתר יצפין את סלי-הפריטים שנתמכים מקומית ביחד עם מספיק סלי-פריטים "מזוייפים", שיחביאו את מספר הסלים האמיתיים בהם האתר תומך. בשלב הבא, כל אתר יצפין את סלי-הפריטים של האתרים האחרים. לאחר מכן, בשלבים 2 ו-3, מבוצע מיזוג של כל ה- ומנפים את כל הכפילויות (אם יש כפילויות בסלי-הפריטים המקוריים, אז (1) מבטיח שיהיו כפילויות גם בסלי-הפריטים המוצפנים). הסיבה שתהליך זה מתרחש בשני שלבים הוא למנוע מצב בו אתר מסויים יודע לשייך איזה סלי-פריטים מוצפנים הגיעו מאיזה אתר. האתר יכול להשתמש בחיתוכים בין כל מספר של אתרים שונים, שבשילוב עם ידיעת ההצפנה של סלי-הפריטים שלו, יכול לאפשר לו לנחש באילו סל-פריטים תומכים האתרים השונים. הפרמוטציה שמבוצעת בשלב 1 מונעת את ידיעת ההצפנה של כל סל-פריטים. בנוסף, המיזוג שמבוצע בשלב 2 (הפרדת המיזוג של האתרים הזוגיים מהאתרים האי-זוגיים) מונעת מכל אתר לקבל את הערכים המוצפנים של כל סלי-הפריטים שלו. בשלב 4 מבוצע הפענוח של סלי-הפריטים הממוזגים. ההצפנה הקומוטטיבית מבטיחה שלא משנה מה יהיה סדר הפענוח, התוצאה שתתקבל תהיה זהה.

האלגוריתם המפורט מופיע בפרוטוקול 1 (איור 2). בפרוטוקול, F מייצג את המידע שיכול לשמש כסלי-פריטים מזוייפים. מייצג את ההצפנה של באתר i. מסמן את ההצפנה של אתר i ו- מסמן את הפענוח.

כפי שניתן לראות, פרוטוקול 1 באיור 2 מוצא את מבלי לחשוף אילו סלי-פריטים שייכים לאיזה אתר. למרות זאת, הפרוטוקול לא נחשב בטוח תחת ההגדרות של חישוב מבוזר מאובטח מכיוון שהוא חושף את מספר סלי-הפריטים המשותפים בין אתרים, אך לא אילו סלי-פריטים בדיוק. ההקלה הזו שמבוצעת באלגוריתם מאפשרת ביטחון מספק עם עלות נמוכה יותר מאשר גישה בטוחה לחלוטין.



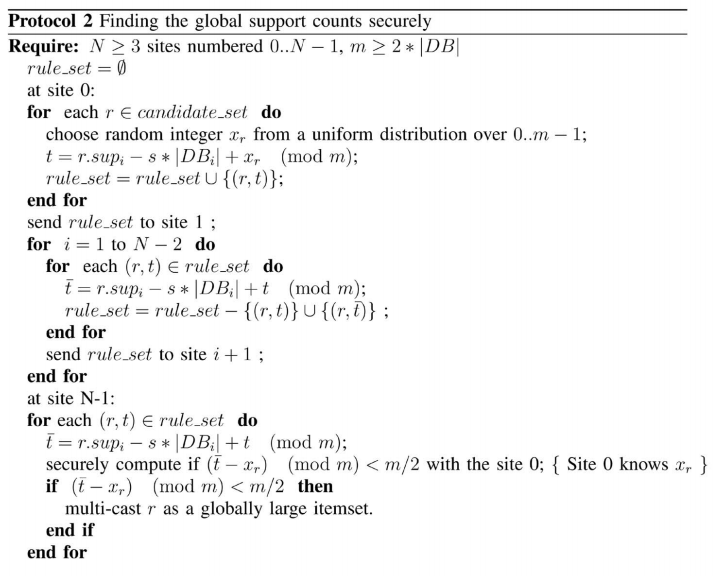
**איור 2**: פרוטוקול 1 – מציאת איחוד בטוח של סלי-פריטים בגודל k

* + 1. בדיקת סף התמיכה ללא חשיפה של מספר התומכים

פרוטוקול 1 מניב את , כך שנותר לבדוק רק אילו מסל-הפריטים ב- נתמך גלובלית. שלב 4 באלגוריתם FDM מחייב כל אתר לפרסם את מספר התומכים בכל סל-פריטים. עבור שימוש בנוסחא יאפשר לנו לדעת האם סל-הפריטים נתמך גלובלית או לא. כדי לחשב זאת מבלי לחשוף את המידע הפרטי של כל אתר נשתמש בתובנה הבאה:

כלומר, בדיקה האם סל-פריטים נתמך גלובלית שקול לבדיקה האם .

הרעיון הכללי הוא שהאתר הראשון יגריל מספר אקראי כלשהו לכל סל-פריטים , יוסיף את ל- וישלח אותו לאתר הבא. המספר הרנדומלי ממסך את נתוני התמיכה של האתר הראשון ולכן האתר השני לא יכול ללמוד את אחוזי התמיכה של האתר הראשון או את גודל מאגר הנתונים שלו. האתר השני מבצע את החישוב בדומה לאתר הראשון, מוסיף את הערך שקיבל לערך שהאתר הראשון שלח, שולח לאתר השלישי וכך הלאה עד לאתר האחרון. כל החישובים מבוצעים מודולו m כאשר , m כזה מאפשר למפות את הערכים השליליים לערכים שגדולים או שווים ל- (). לאתר האחרון ברצף יש את כאשר הוא צריך לבדוק האם הסכום פחות קטן מ- . החישוב הזה יכול להתבצע תוך שימוש בפרוטוקול של Yao שהוזכר בסעיף 2.2.2. אפשרות נוספת היא שהאתר הראשון ישלח לאתר האחרון את הערך , מה שיאפשר לאתר האחרון ללמוד את כמות התומכים הגלובלית אך לא את כמות התומכים בכל אתר. בנוסף, אם כמות התמיכה הגלובלית היא חלק מהתוצאות הכלליות שמוצגות לכולם, אז השיטה עדיין בטוחה. האלגוריתם המלא , פרוטוקול 2, מובא באיור 3.



**איור 3**: פרוטוקול 2 – מציאת התמיכה הגלובלית באופן בטוח

* 1. מציאת ביטחון בחוק הקשר בצורה בטוחה

כפי שהצגנו בסעיף 2.1.1, חישוב הביטחון הכללי של חוק-ההקשר מבוצע ע"י הנוסחא . כדי לבדוק האם חוק מסוים עובר סף ביטחון של c עלינו לבדוק האם . נפתח את המשוואה, ונקבל משוואה שדומה למשוואה שקיבלנו לפני הכניסה לפרוטוקול 2:

ניתן להשתמש בפרוטוקול 2 לביצוע חישוב בטוח של ביטחון החוק, מכיוון שכל אתר יודע את ואת .

1. חולשות הפרוטוקול
   1. בטיחות כנגד קנוניה

שיתוף פעולה בפרוטוקול הראשון יכולה לאפשר לאתר מסויים לדעת את ההצפנות של סלי-הפריטים שלו ובעזרתם לבצע חיתוכים מול סלי-פריטים מוצפנים של אתרים אחרים וללמוד את מספר סלי-הפריטים המשותפים לו ולאתר האחר. בפרט, אם אתר i משתף פעולה עם אתר i-1, אז הוא יכול ללמוד את גודל החיתוך עם אתר i+1. שיתוף פעולה בין אתר 0 לאתר 1 מחמיר את הבעיה מכיוון שהם יודעים את ההצפנות של כל האתרים האי-זוגיים והזוגיים בהתאמה.

שיתוף פעולה בפרוטוקול השני מהווה בעיה, מכיוון שאם אתר i-1 ואתר i+1 חוברים, הם יכולים לדעת את התמיכה הנוספת של אתר i. הפרוטוקול יכול להיות עמיד לקנוניות מהסוג הזה ע"י שילוב האלגוריתם הבא [7] [8]:

1. כל אתר i בוחר בצורה רנדומלית n ערכים כך ש- הוא הקלט של אתר i. אתר i שולח את לאתר j.
2. כל אתר i מחשב את ושולח את לאתר n.
3. אתר n מחשב את התוצאות הסופיות .

האלגוריתם שהוצג מעלה יכול בקלות להשתלב בפרוטוקול 2 כך: נניח שאתר 0 הוא האתר הראשון בפרוטוקול ואתר N-1 הוא האחרון. נבחר m כך ש- . נקבע את להיות ואת להיות , . מנקודה זו, ניתן להשתמש באלגוריתם שמתואר למעלה כדי למצוא את . בסיום האלגוריתם, מבוצע חיסור אחד בטוח והשוואה אחת ע"פ הפרוטוקול כדי לבדוק האם X נתמך גלובלית.

* 1. קשיים במקרה של שני שותפים

דבר ראשון, סלי-פריטים שנתמכים גלובלית אך לא נתמכים מקומית ע"י אתר אחד, ידועים כנתמכים מקומית ע"י האתר השני – זוהי תולדה ישירה מהתוצאה. פרוטוקול 1 מחריף את הבעיה: סלי-פריטים שנתמכים מקומית ע"י אתר אחד ולא נתמכים גלובלית יוודעו לאתר השני. כדי לשמור על איזו שהיא פרטיות יש להוריד את שלבים 1 ו-2 מפרוטוקול FDM ולחשב את התמיכה של כל המועמדים ב- . שנית, החישוב הבטוח שמבוצע בסוף פרוטוקול 2 חייב להישאר, אחרת התמיכה של אתר אחד חשופה לאתר השני.

1. עלויות תקשורת וחישוב

נבצע הערכת עלויות לשיטה שהוצגה במאמר. יהיו N - מספר האתרים, - המספר הכולל של סלי-הפריטים שמועמדים לבדיקה האם הם נתמכים מקומית ו-- מספר המועמדים שיכולים להיווצר ישירות מסלי-הפריטים בגודל k-1 שנתמכים גלובלית. התמיכה הנוספת של סל-פריטים X יכולה להיות מיוצגת ע"י ביטים. יהי t – מספר הביטים בתוצאת ההצפנה של סל-פריטים (הגבול התחתון של t הוא ).

עלות סך כל הביטים שעוברים בתקשורת בפרוטוקול 1 היא . אך מכיוון שרוב הפעולות קורות במקביל, ניתן לחלק את התוצאה ב-N כדי לקבל את הערכת זמן התקשורת. לשם השוואה, אלגוריתם FDM דורש, עבור השלבים המתאימים, . כפי שניתן לראות, פרוטוקול 1 מגיע לסדר גודל של אותה סיבוכיות בגלל המקבילות שנובעת משידור ההודעות (broadcast). העלות הנוספת של פרוטוקול 1 נובעת מריפוד ה- שמטרתו להסתיר את כמות סלי-הפריטים בהם תומך כל אתר, וממספר הביטים שנדרשים כדי לייצג את סלי-הפריטים המוצפנים. במקרה הגרוע ביותר , אך בפועל, בגלל המבנה של האלגוריתם הא-פריורי עבור , העלות תהיה יותר קרובה ל- . העלות החישובית גדלה בעקבות ההצפנות ועומדת על , כאשר t הוא מספר הביטים במפתח ההצפנה. נובע מעלות החישוב, בביטים, של העלאה בחזקה מודולרית.

עלות סך כל הביטים שעוברים בתקשורת בפרוטוקול 2 היא , כאשר העלות הנוספת של t היא לחישוב הבטוח בין אתר 0 לאתר N-1. העלות החישובית של החישוב הבטוח בסוף הפרוטוקול היא .

במקרה של שני שותפים, כפי שתואר בסעיף 5, יבוצע שימוש בטוח בפרוטוקול 2 ישירות על . עלות התקשורת הופכת להיות , אך מכיוון שפרוטוקול 2 מבוצע בצורה סדרתית, זמן התקשורת, בקירוב, זהה לפרוטוקול המלא. העלות החישובית של החישוב הבטוח בסוף הפרוטוקול היא . בנוסף, קיימת תוספת עלות משמעותית מכיוון שנדרש לחשב את התמיכה עבור כל סלי-פריטים – גדול משמעותית מהחישוב של שנדרש באלגוריתם המלא [5].

* 1. אופטימיזציות ודיון נוסף

ריפוד בסלי-פריטים מזוייפים מ- F כדי להימנע מחשיפת מספר הסלי-הפריטים הנתמכים מקומית יכול להוסיף עלות משמעותית לעלות של התקשורת וההצפנה. בפועל, עבור k>1, יהיה בגודל סביר, אך יכול להיות גדול מאוד מכיוון שהוא אינו מוגבל ע"י מספר סלי-הפריטים התדירים שנמצאו ותלוי רק בכמות הפריטים. אם המשתתפים יגדירו גבול עליון עבור מספר סלי-הפריטים שנתמכים בכל אתר, ניתן להשיג שיפור משמעותי בעלות ללא פגיעה בבטיחות. לחלופין, אם אנו מוכנים "להדליף" את הגבול העליון של כל אתר, כל אתר ירפד רק עד הגבול העליון שלו. גישה כזו תשיג רמת בטיחות מקובלת ותשנה את עלות הגורם בתקשורת ובהצפנה בפרוטוקול 1 ל- , בדומה ל-FDM.

אפשרות נוספת היא להקטין את עלות ההצפנה ע"י בחירת ערכי הריפוד מתחום ערכי הפלט של ההצפנה במקום להצפין ערכים מ-F. בהנחה ש- , ההסתברות של ריפוד עם ערכים שמתפענחים לסלי-פריטים אמיתיים הוא נמוך, וגם אז זה כן קורה, זה רק יוסיף סלי-פריטים נוספים לבדיקה בפרוטוקול 2, שיתגלו כלא נתמכים גלובלית.

ניתן להשמיט את שלב ההשוואה בסוף פרוטוקול 2, תוך הפחתה של ביטים בעלות התקשורת ו- בעלות ההצפנה. בפועל, תיתכן דרישה לגילוי ערכי התמיכה של כל אתר עבור פריטים שנתמכים גלובלית ולכן המידע היחיד שדולף הוא התמיכה ב- . מכיוון שלא ניתן לשייך נתונים אלו לאתרים ספציפיים, זה בדרך כלל נחשב מקובל.

1. מסקנות ועבודה נוספת

כלים קריפטוגרפיים מאפשרים כריית מידע שאחרת לא הייתה מתאפשרת עקב בעיות ביטחון. במאמר הוצגה שיטה לכרייה מבוזרת של חוקי הקשר על מידע מחולק אופקית. השיטה מאפשרת את ביצוע הכרייה תוך מספר הנחות ביטחון סבירות.

ההגדרות של חישוב מבוזר מאובטח מתחום הקריפטוגרפיה עלולות להיות מגבילות מידי בתחומים כמו זה. ניתן לראות דוגמה בפרוטוקול 1, כאשר מרחב הריפוד F מוגדר אינסופי, כך שהסיכוי להתנגשות בין אברי F אפסי – דבר לא מעשי בפועל.

נדרשות הגדרות בטיחות יותר מתאימות שמאפשרות לבחור את הרמה הרצוי של הבטיחות ובכך לאפשר פתרונות יותר יעילים תוך שמירה על רמת הבטיחות הנדרשת.

לסיכום, ניתן לבצע כרייה גלובלית של תוצאות נכונות ממידע מבוזר מבלי לחשוף מידע שיכול לסכן את הפרטיות של המקורות. כריית מידע שומרת-פרטיות שכזו יכולה להתבצע תוך הוספת עלות סבירה לשיטה שלא שומרת על פרטיות.

**רשימת ביבליוגרפיה**

1. Murat Kantarcioglu, Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", *IEEE Transactions on Knowledge & Data Engineering*, vol.16, no. 9, pp. 1026-1037, September 2004.
2. Wikipedia, כריית מידע, URL: <https://he.wikipedia.org/wiki/כריית_מידע>.
3. R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” Proc. 2000 ACM SIGMOD Conf. Management of Data, pp. 439-450,2000.
4. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th Int’l Conf. Very Large Data Bases, pp. 487-499, 1994.
5. D.W.-L. Cheung, J. Han, V. Ng, A.W.-C. Fu, and Y. Fu, “A Fast Distributed Algorithm for Mining Association Rules,” Proc. 1996 Int’l Conf. Parallel and Distributed Information Systems (PDIS ’96), pp. 31-42, 1996.
6. A.C. Yao, “How to Generate and Exchange Secrets,” Proc. 27th IEEE Symp. Foundations of Computer Science, pp. 162-167, 1986.
7. J.C. Benaloh, “Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret,” Advances in Cryptography (CRYPTO86): Proc., A. Odlyzko, ed., pp. 251-260, 1986.
8. B. Chor and E. Kushilevitz, “A Communication-Privacy Tradeoff for Modular Addition,” Information Processing Letters, vol. 45, no. 4, pp. 205-210, 1993.