האוניברסיטה הפתוחה
החטיבה למדעי המחשב

שימור הפרטיות בכריית מידע
Privacy in Data Mining

חיבור זה הוגש כעבודת גמר לקראת התואר

מוסמך במדעי המחשב

על ידי אורנה סופר

העבודה הוכנה בהנחיית פרופ' אהוד גודס

אוגוסט 2009

THE OPEN UNIVERSITY
Computer Science Division

Privacy in Data Mining

This work has been submitted as the final assignment towards to degree of master of art in

COMPUTER SCIENCE

By **Orna Sofer**

This study was carried out under the supervision of
Prop' Ehud Gudas

August 2009

תקציר

כריית חוקי מידע מבסיסי נתונים הוא תחום מחקר העוסק בהפקת מידע מכמויות עצומות של נתונים. התפתחות היכולת הטכנולוגית המתאימה והתקשורת העולמית הרחבה הפכה את כריית המידע לכלי רב עוצמה כך שארגונים וחברות יכולים להתקשר בניהם לצורכי כריית חוקי הקשר ושיתוף מידע למטרות שונות, כגון: מחקרים, שיתוף פעולה עסקי, התייעלות ארגונית, איתור בעיות ותקלות, הבנת מגמות שיווקיות ועוד.

יחד עם יתרונותיו המרובים, יצר שיתוף פעולה לצרכי כרייה שתי בעיות עיקריות, הראשונה, הגנה על פרטיות המידע של ישויות אינדיבידואליות (k-anonymity), והשנייה, חשיפת חוקי הקשר תדירים המג לים "סודות" פרטיים של הארגונים הבאים לידי ביטוי, אם במידת התמיכה שלהם בהתקיימותו של חוק הקשר מסוים ואם בחשיפת מידע אסטרטגי ארגוני חסוי כגון מדיניות, שאין הארגונים מעוניינים לחשוף.

עבודה זו מבקשת לבחון את הנושא הרחב של שמירת פרטיות המידע בכריית מידע מבסיסי נתונים מהיבטים שונים ומתחילה בסקירה המרכזת בקווים כלליים את מגוון סוגי הכרייה והטכניקות הקיימות לשימור פרטיות המידע הנכרה, מציגה מדדים וכלים שונים להערכת האלגוריתמים ליישום הטכניקות השונות, ודנה במגוון בעיות הנובעות מהכרייה הבטוחה. העבודה מציגה מגוון שיטות וטכניקות למצבי כרייה מגוונים, תוך התייחסות ליתרונות ולחסרונות של כל טכניקה, משווה בין הטכניקות השונות ובוחנת את יעילות הטיפול בבעיות השונות על ידי כל שיטה.

העבודה לא תתמקד בהסתרת מידע אינדיבידואלי (k-anonymity) מהסיבה שמידע המופק מכריית מידע אינו מכיל פרטי ישויות, וכן שהנתונים הראשוניים עצמם אינם מטרת העוסקים בכרית מידע.

המסקנות הנובעות מהעבודה הן, כי היכולת לכריית מידע והפקת חוקי הקשר מבסיסי נתונים היא רבת עוצמה ומכילה יתרונות עצומים לצדדים השותפים לכרייה, אך גם יוצרת סיכונים רבים הנובעים מגילוי מידע פרטי שאין הצדדים מוכנים לחשוף. ההגנה על הפרטיות מתבטאת הן בשמירה על פרטי ישויות אינדיבידואליות, הן בהסתרת חוקי הקשר חסויים, והן בהגנה על מידת התמיכה בחוקי הקשר תדירים של השותפים השונים בכרייה. מטרת הכרייה ורמת הביטחון הדרושה מהווה שיקול רב ערך בבחירת האלגוריתם המתאים.

על אף המגוון העצום של השיטות לכריית מידע תוך שמירה על הפרטיות, והשקידה על המשך הפיתוח, תוקפים פוטנציאליים משתכללים אף הם, ולכן נראה שעדיין יש מקום לחפש טכניקות לשיפור היעילות של השיטות הקיימות, לחפש פרצות באבטחה בשיטות הקיימות ולתקנן, להמשיך לחקור את יחסי הגומלין בין רמת האבטחה לבין ביצועי האלגוריתמים תוך התייחסות לערכי תמיכה שונים, וכן להתמקד באפשרויות להקטין עוד את השונות בין תוצאות הכרייה לפני הסינון ואחריו תוך שמירה אופטימאלית על פרטיות הנתונים.

Abstract

Mining association rules from databases is a field of research dealing in knowledge production from enormous amounts of data. The appropriate technological development and the ability of advanced global communication turns the data mining into a powerful tool, so that companies and organizations can join each other in order to create shared association rules and knowledge for varied goals, like: research, business cooperation, re-organization, locating problems, faults, malfunctions, understanding marketing trends etc.

Together with its great advantages, cooperation for association rule mining causes two major problems: first, protecting private information of individual entities (k-anonymity), and second, exposure of frequent association rules that reveal confidential information of the organizations which are expressed either by how frequent the existence of the association rule occurs or by revealing the confidential strategies of the organizations like "secret" policies..

This work aims to examine the issue of privacy preserving data mining from databases from different aspects beginning with an overall review of the different types of data mining and techniques for privacy preservation of the mined data. It presents measures and other tools for evaluating algorithms which apply to the different techniques and describes the variety of privacy problems that can arise as a result of data mining. The work presents some chosen methods and techniques for different mining situations, while analyzing the advantages and disadvantages of each technique, comparing between the different techniques and examining the efficiency of execution in each case. The work surveys among other techniques for privacy preserving data mining in vertically and horizontally distributed databases, and techniques for sanitizing the data to prevent the disclosure of sensitive association rules.

The work doesn't focus on protecting private information of individual entities (k-anonymity), due to the fact that information produced from data mining doesn't contain details of individual entities, and because the initial data is not the goal of the data miners involved.

The conclusions that derive from the work are, that the ability of mining data and producing association rules from databases is very powerful and contains enormous advantages for all parties that share the mining, but also creates many dangers that are connected to the disclosure of private information that the parties do not want to share. The privacy protection is expressed by protecting private information of individual entities, by hiding private association rules, and by protecting the support count of frequent association rules of the parties in the mining processing. The goal of the mining and the required confidence degree is a major consideration in selecting the appropriate algorithm.

Despite the wide variety of methods for privacy preserving data mining, and diligently furthering their development, potential attackers become more sophisticated too. Therefore, it is evident that there is still room for the additional search for techniques to improve the efficiency of the existing methods and to stem the breach in security. Moreover, it is essential to continue investigating the interaction between the confidence level and the algorithm performance within reference to different support values, and to focus on the possibilities to reduce the variance between the mining results before sanitization and after it, within optimal preservation of data privacy.

תוכן העניינים

פרק 3

Index

chapter **3**

מבוא

כריית חוקי מידע מבסיסי נתונים הוא תחום מחקר העוסק בהפקת מידע מכמויות עצומות של נתונים. היכולת הטכנולוגית המתאימה והתקשורת העולמית הרחבה הסבה את תחום המחקר של כריית מידע לנרחב יותר ולכזה שניתן לנצל לתועלתיות כללית כך ש ארגונים וחברות יכולים להתקשר בניהם לצורכי כריית חוקי הקשר ושיתוף מידע למטרות שונות, כגון: מחקרים, שיתוף פעולה עסקי, התייעלות ארגונית, איתור בעיות ותקלות, הבנת מגמות שיווקיות ועוד.

שיתוף פעולה לצרכי כרייה יצר שתי בעיות עיקריות , הראשונה, הגנה על פרטיות המידע של ישויות אינדיבידואליות, והשנייה, חשיפת חוקי הקשר תדירים המגלים "סודות" פרטיים של הארגונים הבאים לידי ביטוי אם במידת התמיכה שלהם בהתקיימותו של חוק הקשר מסוים ואם בחשיפת מידע אסטרטגי ארגוני חסוי כגון מדיניות, שאין הארגונים מעוניינים לחשוף.

הרצון להסתיר מידע ולהגן על פרטיות המידע עומד בקנה אחד עם הרצון של גורמים שונים לגלות ולחשוף את המידע המוסתר , מתחרים, גורמים עוינים ומתקיפים מסוגים שונים מהווים איום על הרצון לשתף מידע תוך שמירה על הפרטיות ונושא זה של אבטחה מחייב התייחסות רצינית.

כמות עצומה של מאמרים ומחקרים מתמקדים בתחום זה של אבטחת פרטיות המידע בכריית מידע ומציגים שיטות וטכניקות רבות ומגוונות, שחלקן נבחרו להצגה בעבודה זו.

לעבודה שלושה פרקים . הראשון שבהם סוקר ומסווג את מגוון השיטות והטכניקות שפותחו בנושא, ומציג אותן בקצרה , תוך הזכרת המאמרים הע יקריים המתייחסים לנושא . יש לזכור שמאמרים אלו ראשוניים ובודאי פותחו שיטות טובות ומתוחכמות יותר במרבית הנושאים , אם לא בכולם . אך המאמר מסווג בצורה יעילה וטובה את הרעיונות העומדים מאחורי כל טכניקה.

שני הפרקים האחרים עוסקים בשיטות עצמן ולמעשה מרחיבים חלק מהשיטות והטכניקות המוצגות בפרק הראשון.

הפרק השני עוסק בכריית מידע מבסיסי נתונים מבוזרים , ועיקר הבעיה המוצגת בו היא הסתרת מידת התמיכה בחוקי הקשר תדירים, כאשר מוצגות בו הבעיות של כרייה משותפים כנים- למחצה, כריית מידע משני אתרים, וכד', וכן התייחסות לעץ החלטות מסווג ולכריית מידע על ידי גורמים חיצוניים מבסיס נתונים מרכזי. הפרק עוסק במגוון של טכניקות, ולמעשה כל מאמר מראה טכניקה שונה לחלוטין.

פרק זה הינו המשך עבודת הסמינר באותו נושא, ומרחיב אותו.

הפרק השלישי עוסק בשיתוף פעולה עסקי של חברות וארגונים, ומתייחס למצב בו חלק מהמידע הנכרה פרטי, ויש להסתירו. בפרק מוצגות מספר שיטות לביצוע הכרייה הבטוחה, וכן קיימת התייחסות רחבה לבעיית הסקת מידע חסוי מהמידע הגלוי, ומציג בהמשך טכניקה המאפשרת להימנע מבעיה זו.

העבודה אינה עוסקת בהסתרת מידע אינדיבידואל י (k-anonymity) מהסיבה שמידע המופק מכריית מידע אינו מכיל פרטי ישו יות, וכן שהנתונים הראשוניים עצמם אינם מטרת העוסקים בכרית מידע. לתחום זה מקום מחקרי נפרד.

קריאה נעימה.

1. State-of-the-Art in Privacy Preserving Data Mining[1]

1.1 הקדמה

פרק זה המבוסס על המאמר State-of-the-art in Privacy Preserving Data Mining ומשמש כמבוא לנושא העבודה כולה. תפקידו להציג תמונה כוללת של השיטות השונות לשימור הפרטיות בכריית המידע, privacy preserving data mining. זהו תחום מחקר המתמקד בכריית מידע ובבסיסי נתונים סטטיסטיים בהם האלגוריתמים לכריית מידע מנותחים על פי כמות המידע הפרטי שהם יכולים לחשוף. השיקולים העיקריים בשימור הפרטיות בכריית המידע מתמקדים בשני תחומים. הראשון, שורות מידע רגיש כגון מזהים , שמות, כתובות וכד ' שיש לשנות או להשמיט מבסיס הנתונים המקורי כך שמשתמשים במידע לא יוכלו לזהות מידע פרטי שאינו קשור אליהם . התחום השני, מידע רגיש שנית ן לכרייה מבסיס נתונים על ידי אלגוריתמים לכריית מידע . המטרה העיקרית בשימור הפרטיות בכריית מידע היא לפתח אלגוריתמים לשינוי הנתונים המקוריים בדרך כלשהי , כך שהנתונים הפרטיים והמידע הפרטי יישארו פרטיים גם אחרי תהליך הכרייה.

בפרק זה יסווגו ויתוארו מגוון הטכניקו ת והשיטות השונות שפותחו בתחום של שימור הפרטיות בכריית מידע, תוך אזכור מאמרים בסיסיים המתייחסים לטכניקות אלו. יש לציין, כי מאמר זה פורסם ב-2004, ומכיוון שנושא אבטחת פרטיות המידע בכריית חוקי הקשר הוא נושא הנמצא בתנופת התפתחות והתחדשות מואצת , קיימים מאמרים רבים חדשים בכל אחד מהתחומים המוצגים שמביאים שיפורים תוספות ועדכונים לאלגוריתמים המוזכרים במאמר.

הפרק גם מזכיר בקצרה את הבעיה המתרחשת כאשר מידע סודי מוסק מנתונים גלויים על ידי משתמשים לא מורשים , בעיית ה- "database inference", וכן את תופעות הלוואי הנובעות מסי נון והסתרת הנתונים שיתוארו בפרוט רחב יותר בפרק 3.

בפרקים הבאים מובאים מאמרים שונים המתייחסים להיבטים שונים של שימור פרטיות המידע המוזכרים בפרק זה.

1.2 סיווג הטכניקות לשימור פרטיות

קיימות גישות רבות לשימור הפרטיות בכריית מידע. ניתן לסווג אותן באופן הבא:

- הפצת מידע Data Distribution
- שינוי מידע Data Modification
- אלגוריתמים לכריית מידע Data Mining Algorithm
- הסתרת מידע Data Or Rule Hiding
- שימור מידע Privacy Preservation

<u>הפצת מידע</u> - חלק מהגישות פותחו לנתונים מרוכזים וחלק לנתונים מפוזרים . נתונים מפוזרים ניתן לסווג לפיזור נתונים אופקי - Horizontal data distribution ולפיזור נתונים אנכי - Vertical data distribution. פיזור נתונים אופקי מתייחס למקרים בהם מספר בסיסי נתונים שונים מחזיקים רשומות שונות של אותה סכמת נתונים , בעוד פיזור נתונים אנכי מתייחס למקרים בהם כל ה ערכים המתייחסים לתכונות השונות ימצאו במקומות שונים.

<u>שינוי מידע</u> – מתייחס לסכימה לשינוי מידע. באופן כללי, שינוי מידע משמש במטרה לשנות את הנתונים המקוריים של בסיס נתונים הנחשף לציבור ובדרך זו מובטחת רמה גבוהה של אבטחת הפרטיות . חשוב ששינוי הנתונים יעשה תוך הת חשבות במדיניות האבטחה של הארגון . שיטות לשינוי נתונים כוללות:

- perturbation – בלבול נתונים המושג על ידי החלפת ערכי התכונות בערכים חדשים (למשל, החלפת ערכי 1 לערכי 0, או הוספת רעשים),
- blocking – החלפת ערכי תכונות ב "?",
- aggregation or merging – איסוף מספר ערכים לתוך קטגוריה כוללת,
- swapping – החלפת ערכים של רשומות ספציפיות,
- sampling – חשיפת מידע מתוך מדגם של אוכלוסייה.

<u>אלגוריתמים לכריית מידע</u> – מתייחס לאלגוריתמים לכריית מידע בהם נעש ה שינוי בנתונים. למעשה ללא כוונה מראש, סייעו לניתוח ולתכנון של האלגוריתמים המסתירים נתונים. הסתרת מידע ע"י קומבינציה של אלגוריתמים לכריית מידע תוצג בהמשך הפרק. כרגע, מבחר האלגוריתמים לכריית מידע יחשבו כמבודדים זה מזה. בניהם, הרעיונות החשובים ביותר פותחו לסיווג האלגוריתמים לכריית מידע, כמו decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets ו- Bayesian networks.

<u>הסתרת מידע</u> – מתייחס גם לשורת מידע וגם למידע מגובש (aggregated data) שיש להסתיר. הסיבוכיות בהסתרת מידע מגובש במבנה של חוקים גבוה יותר, ולכן, פותחו בעיקר יוריסטיקות. ההפחתה בכמות המידע הצ יבורי גרמה לכורי המידע ליצר חוקי מסקנות חלשים יותר שלא יאפשרו להסיק מסקנות מערכים סודיים. התהליך ידוע גם בשם "rule confusion".

<u>שימור מידע</u> – החשוב ביותר, מתייחס לטכניקות לשימור הפרטיות המשמשות לשינוי נתונים סלקטיבי. שינוי סלקטיבי נדרש במטרה להשיג תועלת גבוה ה יותר לנתונים המשתנים מבלי לסכן את הפרטיות. הטכניקות שבהן משתמשים למטרה זו הן:

- <u>Heuristic-based techniques</u> להתאמת השינויים כך שישונו רק ערכים נבחרים המקטינים את אבדן הניצולת במקום הערכים האפשריים.
- <u>Cryptography-based techniques</u> לחישוב מאובטח של מספר צדדים, כאשר החישוב הוא מאובטח אם בסופו אף צד לא יודע כלום מלבד נתוני הקלט שלו ותוצאת החישוב הסופית.
- <u>Reconstruction-based techniques</u> כאשר התפוצה המקורית של הנתונים נבנית מחדש מהנתונים הרנדומאליים.

חשוב לציין שתוצאת שינוי הנתונים משפיעה לרעה על ביצועי בסיס הנתוני ם. כדי לכמת את הפגיעה בנתונים משתמשים בעיקר בשתי שיטות. האחת, מדידת האבטחה בהגנה על הנתונים, והשנייה, מדידת אובדן הפונקציונאליות.

1.3 סקירת אלגוריתמים לשמירת הפרטיות

Heuristic-Based Techniques 1.3.1

פותחו מספר יוריסטיקות למספר טכנולוגיות של כריית מידע כגו ן classification, association rule discovery ו- clustering, המבוססות על כך שבחירת הנתונים לשינוי או סינון היא בעיה NP-קשה, ולכן, היוריסטיקות יכולות לשמש להערכת מידת הסיבוכיות.

Centralized Data Perturbation-Based Association Rule Confusion 1.3.1.1

הוכחה פורמאלית לכך שסינון להסתרת סטי-פריטים, גדולים ורגישים בהקשר של גילוי חוקי הקשר היא בעייה NP-קשה ניתנת ב-[2]. הבעיה הספציפית אליה נתייחס היא: יהי D בסיס הנתונים המקורי, ויהי R סט משמעותי של חוקי הקשר שניתן לכרות מ-D, ויהי R_h סט של חוקים ב- R שיש להסתיר. כיצד ניתן להמיר את בסיס הנתונים D לבסיס הנתונים שניתן לגילוי D', פרט לחוקים ב R_h. היוריסטיקה המוצעת לשינוי הנתונים התבססה על בלבול הנתונים, ובפרט הפרוצדורה המוצעת התבססה על שינוי סט נבחר של ערכי 1 לערכי 0, כך שהתמיכה בחוקים רגישים נמוכה בכך שהניצולת של הנתונים בבסיס הנתונים הגלוי נשמרת לערך מקסימאלי כלשהו. הניצולת נמדדת כ- החוקים הלא רגישים שהוסתרו כתוצאה מתופעות הלוואי של תהליך שינוי הנתונים.

עבודה עוקבת תוארה ב- [3] ומרחיבה את הסניטציה על החוקים הרגישים. הגישה היא למנוע מחוקים רגישים להיווצר על ידי הסתרת סטי-פריטים נפוצים שמהם הם נגזרים, או להפחית את הביטחון של החוקים הרגישים על ידי הבאתם לסף שיקבע על ידי משתמש. שתי גישות אלו הובילו ליצירת שלוש אסטרטגיות להסתרת חוקי הקשר רגישים. חשוב לציין כי בשלושת האסטרטגיות האלו ניתן להמיר ערכי 1 בבסיס נתונים בינארי לערכי 0, וההפך, ערכי 0 להמיר לערכי 1. לגמישות זו בשינוי הנתונים יש תופעת לוואי ש מכיוון ש חלק מחוקי ההקשר שאינם רגישים הופכים לנסתרים, חוקים שאינם נפוצים יכולים להפוך לנפוצים. חוקים אלו נק ראים "ghost rules". ואמנם חוקים רגישים מו סתרים, אך מתקבלות תופעות לוואי שהן: חוקים לא רגישים שמוסתרים, וחוקים לא נפוצים שהפכו לנפוצים (ghost rules), המורידות את התועלתיות של בסיס הנתונים הגלוי. מכיוון שכך יוריסטיקות שיובאו בהמשך

חייבות להיות יותר רגישות לנושא היעילות , אך מבלי ל התפשר על נושא האבטחה . המאמר [4] Association Rule Hiding מבוסס על רעיון זה (עוד על נושא זה בפרק 3).

Centralized Data Blocking-Based Association Rule Confusion 1.3.1.2

אחת הגישות לשינוי נתונים המשמשת לבלב ול חוקי הקשר היא "חסימת נתונים ", data[5] blocking. הגישה של חסימה מיושמת על ידי החלפת מספר תכונות של חלק מפריטי הנתונים עם סימן שאלה. לפעמים עדיף לאפליקציות מסוימות להחליף ערכים אמיתיים עם ערך לא ידוע במקום להציב ערך מדומה (למשל, באפליקציות רפואיות). גישה המיישמת חסימה לבלבול חוקי הקשר מוצגת ב- [6]. השימוש בערכים החדשים הללו בפריטי הנתונים מחייבת מספר שינויים בהגדרת ה תמיכה ורמת הביטחון של חוקי הקשר. מתוך כך , התמיכה המינימאלית ורמת הביטחון המינימאלית ישונו למרווח תמיכה מינימאלי ולמרווח רמת ביטחון מינימאלית בהתאמה. כל זמן שה תמיכה ו/או רמת הביטחון של חוקים רגישים יהיו בין שני תחומים אלו , אנו נצפה כי סודיות הנתונים לא תופר . יש לציין כי אלגוריתם המשתמש בבלבול נתונים במקרים כאלו, גם ערכי 1 וגם ערכי 0 יומרו לסימני שאלה, אחרת, יהיה ברור מהו מקורם. מאמר [7] מרחיב בנושא ההשפעה של גישה זו על בנייה מחדש של נתונים מבולבלים.

Centralized Data Blocking-Based Classification Rule Confusion 1.3.1.3

מאמר [8] מספק מבנה המשלב ניתוח סיווג חוקים וחיסכון בהחסרת נתונים. יש להבחין כאן , כי בבניית סיווג הנתונים , למנהל הנתונים יש מטרה לחסום נתונים למחלקת הסיווג . על ידי כך , מקבלי המידע לא יוכלו לבנות מודלים שניתן יהיה ללמוד מהם על הנתונים שלא הוחסרו. החיסכון בהחסרה הוא הפיכה למבנה פורמאלי את התופעה המוציאה מידע מסטים של נתונים מאינפורמציה חסרה מתו ך סביבה מאובטחת (High) לסביבה ציבורית (Low) , בהינתן הקיום של ערוץ הסקת מסקנות. בחיסכון בהחסרה אומדן העלות מיוחסת למידע החסר הפוטנציאלי שלא נשלח ל- Low. המטרה העיקרית בעבודה זו היא לגלות האם חוסר הפונקציונאליות שקשור באי הורדת המידע שווה את תוספת האבטחה . חוקי הסיווג ובמיוחד עצי החלטה משמשים בהקשר החסרת המידע לניתוח פוטנציאלי של ערוץ הסקת המסקנות בנתונים שיש להחסיר.

השיטה המשמשת להחסרה היא יצירת סט מאפייני בסיס. במיוחד, מאפיין, θ $0 \le \theta \le 1$ מחליף את הערכים החסומים . הפרמטר מייצג את ההסתברות לאחד מהערכים האפשריים שמאפיין יכול לקבל. הערך של האנטרופיה ההתחלתי ת לפני החסימה וערך האנטרופיה לאחר החסימה מחושבת . ההבדל בערכים של האנטרופיה מושווים לירידה ברמת באבטחה של החוקים שנוצרו על ידי עץ ההחלטה במטרה להחליט האם העלייה ברמת האבטחה שווה את ההפחתה בניצולת של הנתונים שה-Low יקבל.

במאמר [9] המחבר מציג תכנון של תכנה , the Rational Downgrader, המבוססת על רעיון החיסכון בהחסרה. התכנה היא הרכבה של מידע בסיסי ביצירת החלטות , לקביעת החוקים שיש להסיק, "שומר" למדידת הכמות של האינפורמציה החסרה , וחיסכון בהחסרה לשינוי ה חלטות ההחסרה הראשוניות. האלגוריתם לחיסור בנתונים מוצא אילו חוקים מאילו שמושפעים מעץ ההחלטות , יש לסווג כנתונים פרטיים. כל נתון שנמצא שאינו תומך בחוקים , מורחק בהחסרה יחד עם כל התכונות שאינן מיוצגות במחלקות הסיווג. מהנתונים הנותרים, האלגוריתם צריך להחליט אילו ערכים יש להסתיר . זאת במטרה לבצע אופטימיזציה על בלבול הנתונים. מערכת ה"שומר" קובעת מהי הרמה המתקבלת על הדעת בבלבול הנתונים.

Cryptography-Based Techniques 1.3.2

מספר גישות מבוססות הצפנה פותחו בהקשר של שימור הפרטיות באלגוריתמים לכר יית מידע, לפתרון בעיות בעלות האופי הבא : שני צדדים או יותר מעונייני ם לנהל חישובים המבוססים על נתוני הקלט הפרטיים שברשותם , אך אף צד אינו מוכן להסגיר את הנתונים בהם הוא משתמש . הסוגיה שכאן, היא כיצד לנהל חישובים משותפים תוך שימור פרטיות הקלט. הבעיה נקראת the Secure Multiparty Computation (SMC) problem. במיוחד בעיית SMS מתמודדת עם פונקציות חישובי הסתברויות על כל קלט ברשתות תקשורת נרחבות בו כל משתתף מחזיק חלק מנתוני הקלט , ומבטיח את אי התלות

של הקלט, נכונות החישוב, וכן ששום אינפורמציה נוספת לא נחשפת לשותפים בחיש וב פרט לנתוני הקלט של כל שותף ותוצאת החישוב.

נסקור שני מאמרים המכסים שטח זה בכלליותו. הראשון [10] מציע שיטת המרה המאפשרת מעבר מהעברת חישובים רגילה לביצוע חישובים מאובטחים בין מספר צדדים. בין שאר הדברים, המאמר מציג דיון לגבי המרת מגוון רחב של בעיות לכריית מידע לחישוב מאובטח בין מספר צדדים. היישומים לכריית מידע המוצגים במאמר כוללים data classification, data clustering, association rule mining, data generalization, data summarization and data characterization

המאמר השני [11] מציג ארבע שיטות לחישוב מאובטח בין מספר צד דים המבוססות על שיטות התומכות בשימור הפרטיות בכריית מידע. השיטות המתוארות כוללות, חישוב בטוח של סכום, חישוב בטוח של איחוד, חישוב בטוח של גודל סט של חיתוך וחישוב בטוח של מכפלה סקלארית. דוגמא לחישוב בטוח של סכום מדגים בפשטות חישוב מאובטח בין מספר צדדים: הנח כי הערך $u = \sum_{i=1}^{s} u_i$ שיש לחשב אותו נמצא בתחום $[0,n]$. צד אחד משמש כצד הראשי ומקבל את מספר הזיהוי 1. שאר הצדדים ממוספרים $2,\ldots,s$. צד 1 מייצר מספר אקראי R מהתחום $[0,n]$, מוסיף לו את הערך המקומי שלו u_1, ושולח את הסכום $R+u_1 \bmod m$ לאתר מס' 2. שאינו יכול לגלות כלום על הערך u_1. שאר האתרים, $2,\ldots,s$, מבצעים כל אחד: אתר i מקבל $V = R + \sum_{j=1}^{i-1} u_j \bmod n$. מכיוון שערך זה מופץ באופן אחיד בתחום $[0,n]$, i אינו לומד ממנו כלום. אתר i מחשב כעת $R + \sum_{j=1}^{i} u_j \bmod n = (u_j + V) \bmod n$ ומעביר אותו לאתר i+1. אתר s מבצע את החישוב ושו לח את תוצאתו לאתר 1. אתר 1 היודע את R, מחסר אותו לקבלת התוצאה האמיתית.

(בפרק 2 של העבודה מובאות מספר שיטות לחישוב מאובטח בין שני צדדים).

1.3.2.1 Vertically Partitioned Distributed Data Secure Association Rule Mining

כריית חוקי הקשר פרטיים מנתונים המחולקים אנכית, בהם הפריטים מפוזרים בין מספר אתרים, וכל סט-פריטים מפוצל על פני מספר אתרים, יכול להתבצע על ידי מציאת רמת התמיכה, support count, של סט-פריטים. אם ניתן לחשב בפרטיות את רמת תמיכה של סט-פריטים כזה, אז אפשר לבדוק אם התמיכה גדולה מסף (threshold) כלשהו, ולהחליט אם סט-הפריטים נפוץ. אלמנט המפתח לחישוב רמת התמיכה של סט-פריטים הוא לחשב את המכפלה הסקלארית של וקטורים המיוצגים ב תתי-סטי-פריטים, sub-itemsets, שבצדדים השונים. לכן אם ניתן לחשב בצורה בטוחה את המכפלה הסקלארית, הרי שניתן לחשב גם את רמת התמיכה בצורה בטוחה. האלגוריתם המחשב את המכפלה הסקלארית בדרך אלגברית המסתירה ערכי אמת בהחלפתם במשוואות עם ערכים אקראיים המשמשות למיסוך מתואר ב [12]. האבטחה בפרוטוקול המכפלה הסקלארית מבוססת על אי- היכולת של מי מהצדדים לפתור k משוואות עם יותר מ- k נעלמים. חלק מהנעלמים נבחרים בצורה אקראית, ויכולים להיחשב כבטוחים. דרך נוספת לחישוב רמת התמיכה היא על ידי שימוש בחישוב מאובטח של גודל סט חיתוכים מתוארת ב[10].

(פסקה 2.2 בעבודה מתארת בהרחבה שיטה זו שהובאה בהרחבה בעבודת הסמינר)

1.3.2.2 Horizontally Partitioned Distributed Data Secure Association Rule Mining

בבסיס נתונים אופקי, הטרנזקציות מפוזרות בין מספר אתרים. הספירה של התמיכה הגלובלית, global support count, של סט-פריטים הוא הסכום של כל הספירות של התמיכה המקומית, local support count. סט-פריטים X נתמך גלובלית אם ספירת התמיכה הגלובלית של X גדולה מ s% של מספר הטרנזקציות הכולל שבבסיס הנתונים. k-itemset נקרא globally large k-itemset אם הוא נתמך גלובלית. המאמר [13] מתאים את האלגוריתם המוצע לכריית נתונים מבוזרת על ידי שימוש באיחוד בטוח ובחישוב סכום בטוח באופן השומר על הפרטיות בפעולת SMC.

(פסקה 2.1 בעבודה מתארת שיטה זו שהובאה בהרחבה בעבודת הסמינר)

1.3.2.3 Vertically Partitioned Distributed Data Secure Decision Tree Induction

העבודה המתוארת ב[14] חוקרת את תהליך בניית עצי החלטה מסווגים עבור בסיס נתונים המבוזר בצורה אנכית. הפרוטוקול המוצג נבנה על פרוטוקול המכפלה הסקלארית המאובטחת על ידי שימוש בשרת של צד שלישי.

(פסקה 2.4 בעבודה מתארת בקצרה שיטה זו. האלגוריתם מובא בהרחבה בעבודת הסמינר)

1.3.2.4 Horizontally Partitioned Distributed Data Secure Decision Tree Induction

מאמר [15] מציע פתרון לבעיית סיווג שימור הפרטיות בעזרת גישה של חישוב מאובטח בין מספר צדדים, הנקראת oblivious transfer protocol for horizontally partitioned data. בהנתן שפתרון SMC כללי הוא לא מעשי, המחבר מתמקד בבעיית האינדוקציה של עץ ההחלטה, ובמיוחד האינדוקציה של ID3, אלגוריתם פופולארי לעץ החלטה הנמצא בשימוש נרחב. האלגוריתם של ID3 בוחר את התכונה ה"טובה" ביותר על ידי השוואה ת האנטרופיות הניתנות כמספרים שלמים. בכל פעם שערכי האנטרופיות של תכונות שונות קרובים אחד לשני, מצופה שתוצאות החלטת העץ מבחירת אחת האפשרויות תהיה כמעט כמו בחירה צפויה אחרת. פורמאלית, לזוג של תכונות יש δ-equivalent information gains אם ההפרש בין הפער במידע קטן מהערך δ. הגדרה זו מעלה את האומדן של ה-ID3. נסמן *ID3* כקבוצת כל העצים האפשריים הנוצרים על ידי ה- ID3, ובחירת התכונות המשותפות במקרה שהן δ-equivalent. המאמר מציע פרוטוקול לחישוב מאובטח לאלגוריתם ספציפי של $ID3_{\delta}$. הפרוטוקול מורכב מקריאות רבות של חישובים קטנים פרטיים. האלגוריתם המסובך ביותר בניהם מצטמצם להערכת הפונקציה $xlnx$.

1.3.2.5 Privacy Preserving Clustering

אלגוריתם לקיבוץ מאובטח המשתמש באלגוריתם Expectation-Maximization מוצג ב[10] ומציע אלגוריתם איטרטיבי העושה שימוש בפרוטוקול SMC לסכום מאובטח.

1.3.3 Reconstruction-Based Techniques

מספר טכניקות שהוצעו לאחרונה מתייחסות לנושא שימור הפרטיות על ידי בלבול הנתונים ושחזורם מחדש תוך שימוש ברמות קיבוץ במטרה לבצע את הכרייה. להלן מפורטים סוגי השיטות וחלק מהטכניקות.

1.3.3.1 Reconstruction-Based Techniques for Numerical Data

מאמר [16] מציג את הבעיה של בניית עץ החלטה מסווג מנתונים בהם ערכים של רשומות אינדיבידואליות מבולבלים. מכיוון שלא ניתן להעריך במדויק ערכים מקוריים של נתוני רשומות אינדיבידואליות, המאמר מציע פרוצדורה לשחזור ולהערכה מדויקת של פיזור הנתונים המקוריים. על ידי שימוש ב- reconstructed distribution ניתן לבנות מחלקות סיווג שהדיוק שלהן שווה לדיוק של מחלקות הסיווג שנבנו מהנתונים המדויקים. לסילוף הנתונים, המאמר מחשיב גישה דיסקרטית וגישה לסילוף הנתונים. לשחזור הפיזור המקורי ניתן להשתמש ב Bayesian approach וכן מוצגים שלושה אלגוריתמים לבניית עצי החלטות מדויקים המסתמכים על התפוצה המשוחזרת.

מאמר[17] מציע שיטה משופרת בהתבסס על הפרוצדורה Bayesian-based reconstruction בהתבסס על אלגוריתם Expectation Maximization (EM) לחידוש הפיזור. כאשר יש כמות גדולה של נתונים אלגוריתם ה EM מספק הערכה טובה של הפיזור המקורי. כמו כן, הערכת הפרטיות של מאמר [15] צריכה להיות נמוכה כאשר תוספת המידע שהכורה משיג מאוסף הנתונים הפזורים המורכבים מחדש נכללים בהגדרת הבעיה.

1.3.3.2 Reconstruction-Based Techniques for Binary and Categorical Data

מאמרים[18] ו-[19] עוסקים בנתונים בינאריים או קטגוריים בהקשר של כריית חוקי מידע . שני המאמרים מציגים טכניקות רנדומאליות לשימור פרטיות תוך שימושיות גבוהה בסטי הנתונים.

4.1 הערכת אלגוריתמים לשמירת הפרטיות

אחד מההיבטים החשובים בפתוח ומימו ש אלגוריתמים וכלים לשימור פרטיות המידע הוא זיהוי קריטריון ההערכה המתאים והפיתוח של הערכת קריטריון זה . בדרך כלל לא קיים אלגוריתם לשימור פרטיות המידע המספק את כל הקריטריונים האפשריים . לכן מספיק שאלגוריתם יספק ביצועים טובים יותר מאחר עבור קריטריונים ספציפיים (כגון: ביצועים, שימושיות הנתונים וכד'). מכאן שחשוב לספק למשתמשים סט של מטריקות שיספקו להם את האפשרות לבחור את הטכניקה לשימור פרטיות המידע המתאימה להם ביותר עבור הנתונים תוך התייחסות לפרמטרים ספציפיים שהם מעוניינים לשמר.

להלן רשימה ראשונית להערכת פרמטרים שי ש להשתמש בהם להערכת האיכות של שימור פרטיות המידע באלגוריתמים לכריית המידע:

- Performance – ביצועי האלגוריתם במונחים של דרישות זמן, כלומר הזמן הדרוש לכל אלגוריתם להחביא את הסטים המוגדרים כמידע רגיש.
- **Data utility** – תועלתיות הנתונים לאחר יישום הטכניקה לשימור פר טיות המידע, כלומר איבוד המידע או הפונקציונאליות הקטן ביותר.
- Level of uncertainty – רמת חוסר הביטחון בה מידע רגיש שהוסתר יכול להיחשף.
- Resistance – עמידות המושגת על ידי אלגוריתמים לשימור הפרטיות עבור טכניקות שונות של כריית מידע.

1.4.1 ביצועים

גישה ראשונה לה ערכת דרישות הזמנים של האלגוריתמים לשימור פרטיות המידע היא להעריך את עלות החישוב . במקרה כזה ברור שאלגוריתם בעל סיבוכיות פולינומיאלית של $O(n^2)$ יעיל יותר מאלגוריתם בעל סיבוכיות אקספוננציאלית $O(e^n)$.

גישה אחרת היא להעריך את המספר הממוצע של הפעולות הנדרשות להפחתת התדירות של מופע ספציפי של מידע רגיש אל מתחת לסף שנקבע. ערכים אלו אולי לא מספקים אמת מידה מדויקת, אך ניתן להתחשב בהם כדי להשוות במהירות בין אלגוריתמים שונים.

עלות התקשורת , communication cost, מתרחשת במהלך החלפת מידע בין מספר אתרים המשתפים פעולה , ויש להתחשב בה גם . ברור שמחיר זה חייב להיות מינימאלי עבור אלגוריתמים לכריית חוקי מידע מבוזרים.

1.4.2 תועלתיות הנתונים

התועלתיות של הנתונים בסוף תהליך שימור הפרטיות , היא גורם חשוב , מכיוון שבמטרה להסתיר מידע רגיש בסיס הנתונים חייב להשתנות באמצעות הכנסת מידע כוזב (החלפת ערכים הוא תופעת לוואי במקרה זה) או באמצעות חסימת ערכי נתונים . יש להבחין שחלק מהטכניקות של לשימור הפרטיות אינן משנות את המידע המאוחסן בבסיס הנתונים , אך עדיין התועלתיות של הנתונים יורדת, מכיוון שהמידע אינו שלם במקרה הזה. ברור שככל שנעשים יותר שינויים בבסיס הנתונים, כך בסיס הנתונים פחות משקף את האינטרס המשותף . לכן, הפרמטר המעריך את תועלתיות הנתונים צריך להיות כמות המידע שאבד לאחר יישום תהליך שימור הפרטיות . כמובן, שסוג המדידה תלוי בטכניקה שיושמה לכריית מידע.

לדוגמא, איבוד מידע במונחי כריית חוקי מידע יימדד או במונחי מספר החוקים שנשארו או אבדו בבסיס הנתונים לאחר הסינון , או אפילו במונחי עלייה /ירידה בתמיכה וברמת הביטחון של כל החוקים. עבור המקרה של classification, ניתן להשתמש במטריקה הדומה לזו המשמשת בחוקי הקשר. עבור clustering, השונות של המרחק בין הפריטים המקובצים בבסיס הנתונים המקורי ובסיס הנתונים המסונן, יכול להיות הבסיס להערכת איבוד המידע.

1.4.3 רמת חוסר הודאות

אסטרטגיות לשימור הפרטיות מנוהלות על ידי הפחתת המידע שיש להסתיר אל מתחת לסף מסוים. למרות זאת, המידע המוסתר עדי ין יכול להיות מוסק ברמות כלשהן . לכן, את אלגוריתמי הסינון ניתן להעריך בהתבסס על רמת חוסר הודאות שהם מספקים במהלך הניסיון לשחזר את הנתונים המוסתרים. ממבט ביצועי, ראוי יהיה לקבוע למקסימום את בלבול המידע , ואז לקבוע את רמת חוסר הודאות שהושגה על ידי כל אחד מאלג וריתם הסינון תחת האילוצים שנקבעו . מצפים שהאלגוריתם שיספק את ההגנה הטובה ביותר יהיה המועדף ביותר על פני כל השאר.

תיאור נרחב יותר על הסקת מידע מובא במבוא לפרק 3.

כפי שנראה בפרק 3 קיימים אלגוריתמים המאפשרים למנהל בסיס הנתונים לכוונן את רמת האבטחה מול רמת אמי נות הנתונים המתקבלת . ובכך ליצר איזון כלשהו בין כמות המידע המוסתר ושימושיות הנתונים המופקים.

1.4.4 Endurance of Resistance to different Data Mining techniques

המטרה האולטימטיבית של האלגוריתמים להסתרת מידע היא הגנה על מידע רגיש כנגד חשיפה לא מורשית. במקרה כזה חשוב לזכור שתוקפים וגורמים בעלי כוונות זדון ינסו לחשוף את המידע על ידי שימוש במגוון אלגוריתמים לכריית מידע . כתוצאה מכך, אלגוריתמי סינון שפותחו לעמוד כנגד טכניקות מסוימות לכריית מידע במטרה להבטיח את פרטיות המידע יכולים לא לספק את אותה עמידות כנגד כל האלגוריתמים האפשריים לכריית מידע.

כדי לספק הערכה מושלמת לעמידות האלגוריתמים לסינון , יש למדוד את הסיבולת שלהם כנגד טכניקות לכריית מידע השונות מהטכניקה שאלגוריתמי הסינון פותחו בעבורם . פרמטר כזה נקרא traversal endurance. ההערכה של הפרמטר הזה צריכה להתחשב בסיווג של האלגוריתמים לכריית מידע. כמו כן, ייתכן שיהיה צורך לפתח הגדרה פורמאלית שתחת בדיקת אלגוריתם סינון כנגד בחירה מוקדמת של סטי נתונים, ניתן יהיה באופן טרנזיטיבי להוכיח פרטיות עבור כל מחלקת אלגוריתמי הסינון.

1.5 סיכום

פרק זה הציג את סוגי האלגוריתמים לשימור פרטיו ת המידע בכריית מידע , וסיווגם למחלקות וטכניקות שונות. מגוון השיטות והטכניקות שהוצגו הראו, על קצה המזלג, איך התפתחות התחום הנרחבת של תקשורת וגלובליזציה של מידע , והצורך ל הסיק מסקנות מתוכו , מגדילים באופן ניכר את הצורך למצוא שיטות חדשות , מתקדמות ועמידות יותר לשיתוף מידע, תוך הגנה על מידע רגיש ממתחרים או מתוקפים בעלי כוונות זדון.

הפרק מזכיר בקצרה חלק קטן מהבעיות הקיימות בכריית מידע , ולא מפרט חסרונות בטכניקות שונות המוצגות בפרק. הסיווג יעיל וממצה ויכול לשמש כאינדקס לשיטות השונות, אך לא בכל סיווג ניתן להבין את עקרונות השיטה המוצגת בו, ויש לעיין במאמרים עצמם כדי להבין את כוונת הרעיון.

שימור פרטיות המידע בכריית חוקי הקשר מבסיסי נתונים מבוזרים

הקדמה

פרק זה בוחן מגוון שיטות לשימור פרטיות המידע בכריית מידע מבסיסי נתונים מבוזרים כאשר מספר שותפים מעוניינים להפי ק חוקי הקשר תדירים לקבלת מידע משותף לתועלת כל הצדדים. הפרק מנסה לבחון היבטים שונים של בעיה זו ולהציג פתרונות מתאימים, כאשר הצורך בפרטיות המידע מתבטא בעיקר בצורך להסתיר את כמות התמיכה של כל שותף בחוקי ההקשר התדירים המופקים מתוצאת הכרייה.

הפרק מרחיב את עבו דת הסמינר שעסקה בנושא זה ומכיל חמישה מאמרים . המאמר הראשון Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data בסעיף 2.1, מוצג בהרחבה, ועוסק בשימור פרטיות המידע בכריית מידע על חלוקה אופקית של נתונים עלי ידי יותר משני אתרים, תוך שימוש בצפנה קומוטטיבית. המאמר השני Privacy Preserving Association Rule Mining in Vertically Partitioned Data בסעיף 2.2, מוצג בהרחבה, ועוסק בשימור פרטיות המידע בכריית מידע על חלוקה אנכית של נתונים בין שני אתרים, תוך שימוש במכפלה סקלארית . המאמר הש לישי Association Rules Mining in Vertically Partitioned Databases בסעיף 2.3 מובא בקצרה ומציג מאמר של פרופ ' גודס העוסק אף הוא בנושא שימור פרטיות המידע בכריית מידע על חלוקה אנכית של נתונים אך מכיל שני אלגוריתמים , הראשון לכרייה בין שני אתרים והשני לכרייה מרוב ת אתרים , ומשתמש בהוספת טרנזקציות מזויפות לבסיס הנתונים. המאמר הרביעי Privacy-Preserving Decision Trees Over Vertically Partitioned Data בסעיף 2.4 מוצג בקצרה בפרק זה , מציג עץ החלטה מסווג ויעיל מאד בעל רמת הבטחה גבוהה המשתמש באנטרופיה בתהליך בניית העץ . המאמר החמישי Preserving Privacy in Association Rule Mining with Bloom Filters מובא בקצרה בסעיף 2.5 שונה משאר המאמרים בפרק בכך שאינו עוסק בכריית מידע מבסיסי נתונים מבוזרים אלא בכריית מידע בשרתי קצה המרוחקים מבסיס נתונים מרכזי תוך שימוש ב Bloom Filters לצורך הסתרת המידע.

סקירה קצרה זו של תוכן הפרק מראה את המגוון הרחב של הטכניקות והשיטות הקיימות לשמירת פרטיות המידע בכריית מידע מבסיסי נתונים מבוזרים.

2.1 Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data[20]

2.1.1 הקדמה

שימור פרטיות המידע בכריית מידע על חלוקה אופקית של נתונים מתייחס למצב בו מעוניינים לבנות מאגר מידע מרכזי המבוסס על נתונים המפוזרים על פני מספר אתרים הומוגניים בעלי סכמת נתונים זהה, כאשר לכל אתר נתונים על ישויות שונות ואף אחד מהם אינו מורשה לשל וח את הנתונים שברשותו לאתרים האחרים.

המטרה היא למצוא את החוקי ההקשר הגלובליים תוך הגבלה על המידע שניתן לשיתוף מכל אתר.

לדוגמא: המכון לבקרת מגפות (Center for Disease Control - CDC) מעוניין לכרות מידע בריאותי כדי להפחית את העמידות המהירה של חיידקי ם לאנטיביוטיקה. חברות ביטוח מחזיקות מידע על מחלות החולים ומרשמים. כריית מידע זה מאפשר לחשוף חוקים כגון:
Augmentin&Summer $\Rightarrow$ Infection&fall . כלומר, אנשים הנוטלים אוגמנטין בקיץ נוטים לחלות שוב בסתיו.

הבעיה היא שחברות הביטוח יודאגו מחשיפת המי דע. לא רק מידע פרטי על חולים ייחש ף, אלא גם מידע השייך רק להן . למשל חוק המצביע על בעיה חמורה הקיימת במרכז רפואי כלשהו . אם החוק

נתמך גלובלית ניתן יהיה לזהות במדויק את הבעיה ולנסות לשפר את הטיפול . אך אם החוק ייחשף החברות יהיו חשופות לפגיעה באמינות הציבורית שלהן, מה שמפחית את הרצון שלהן להשתתף במחקר ולחלוק מידע.

חישוב התמיכה הגלובאלית ורמת הביטחון של חוק הקשר מהצורה $AB \Rightarrow C$ נעשה בקלות מבלי לגלות מידע אינדיווידואלי מכל טרנזקציה על ידי התמיכה המקומית של AB ושל ABC וגודל כל בסיסי נתונים:

$$support_{AB \Rightarrow C} = \frac{\sum_{i=1}^{sites} support\_count_{ABC}(i)}{\sum_{i=1}^{sites} database\_size(i)}$$

$$support_{AB} = \frac{\sum_{i=1}^{sites} support\_count_{AB}(i)}{\sum_{i=1}^{sites} database\_size(i)}$$

$$confidence_{AB \Rightarrow C} = \frac{support_{AB \Rightarrow C}}{support_{AB}}.$$

ניתן בקלות להרחיב אלגוריתמים לכריית מידע, למקרה המבוזר על פי הלמה הבאה:

אם לחוק תמיכה גלובלית גדולה מ-k% ,חייב להיות לפחות אתר אחד שבו התמיכה גדולה מ-k% .

האלגוריתם המבוזר יפעל באופן הבא:

כל אתר ישלח את כל החוקים המקיימים תמיכה הגדולה מ-k% . לכל חוק שנשלח, נדרש כי כל האתרים ישלחו את מספר הפעמים שהחוק נתמך אצלם , וכן את המספר הגלובלי של כל הטרנזקציות באתר. מכאן ניתן לחשב את התמיכה הגלובלית של כל חוק . על פי הלמה , בטוח שכל החוקים בעלי לפחות k תמיכה נמצאו. מאמרים [21] ו-[22] מרחיבים בנושא של כריית חוקי מידע מבוזרת.

גישה זו מגינה על פרטיות מידע אינדיווידאלי , אולם דורשת שכל אתר יסגיר את החוקים שבהם הוא תומך, וכן את אחוזי התמיכה שלו בכל חוק. מה קורה אם זהו מידע רגיש?

השיטה המוצגת כאן שומרת על חיסיון המידע והצדדים אינם לומדים כמעט כלום מעבר לחוקים הגלובליים. התוספת לעלות היא יחס לשיטה הלא בטוחה :

O(number_of_candidate_itemsets * sites) הצפנות ועלות קבועה במספר ההודעות.

ההנחה היא שקיימים לפחות 3 אתרים, מהסיבה שאם חוק נתמך גלובלית ואינו נתמך על ידי אתר אחד, הרי ברור שהוא נתמך גלובלית על ידי האתר השני , וכך נחשף מידע רגיש . וכן שאין קנוניה בין האתרים (מה שיכול להוביל למצב דומה של שני אתרים בלבד).

2.1.2 רקע

2.1.2.1 כריית חוקי הקשר

<u>הגדרה:</u>

1. יהי $\text{Let } \mathcal{I} = \{i_1, i_2, \cdots, i_m\}$ סט של ליטראלים (פריטים).
2. ויהי D סט של טרנזקציות ,כאשר כל טרנזקציה T היא סט של פריטים, כך ש $T \subseteq \mathcal{I}$.
3. הקשר עם כל טרנזקציה מוגדר כמזהה יחיד הקרוי TID.
4. טרנזקציה T מכילה את X אם X הוא סט של פריטים מ $T \subseteq \mathcal{I}$ ו $X \subseteq T$.
5. חוק הקשר הוא יחס גרירה מהצורה $X \Rightarrow Y$ כאשר $X \subset \mathcal{I}, Y \subset \mathcal{I}$ וגם $X \cap Y = \phi$.
6. החוק $X \Rightarrow Y$ מקיים את **רמת הביטחון c** ב D אם c% מהטרנזקציות ב D המכילות את X גם מכילות את Y.
7. החוק $X \Rightarrow Y$ מקיים **תמיכה s** ב D אם s% מהטרנזקציות ב D מכילות $X \cup Y$.
8. סט של k פריטים X נקרא **k-itemset.**

הבעיה של כריית חוקי מידע היא למצוא את כל החוקים שה **תמיכה** ו**רמת הביטחון** שלהם גבוהים ממינימום כלשהו שאותו מגדיר המשתמש.

על פי הגדרה זו, פריטים חסרים, ערכים שלילים או כמויות אינם נחשבים.

טרנזקציה בבסיס הנתונים יכולה להיחשב כמטריצה של 0/1, כאשר כל עמודה מייצגת פריט וכל שורה היא טרנזקציה.

2.1.2.1.1 *כריית חוקי מידע מבוזרים*

נרחיב את הבעיה לסביבה מבוזרת, בה הטרנזקציה בבסיס הנתונים DB מחולקת אופקית בין n אתרים (הנקראים $S_1, S_2, \ldots, S_n$) כאשר $DB=DB_1 \cup DB_2 \cup \ldots \cup DB_n$ ו- DB_i נמצא ב S_i *$(1 \le i \le n)$*.
לסט הפריטים X יש תמיכה מקומית (local support count) $X.sup_i$ באתר S_i אם $X.sup_i$ מהטרנזקציות מכילות X.

תמיכה גלובלית (global support count) של X מוגדרת כ - $X.\sup = \sum_{i=1}^{n} X.\sup_i$

סט פריטים X נתמך גלובלית (globally supported) אם $X.\sup \ge s \times \left(\sum_{i=1}^{n} |DB_i| \right)$.

בטחון גלובלי (global confidence) של חוק $X \Rightarrow Y$ יוגדר כ- $\{X \cup Y\}.\sup / X.\sup$.
הסט $L_{(k)}$ מוגדר כסט הפריטים בגודל k הנתמך גלובלית.
הסט $LL_{i(k)}$ מוגדר כסט הפריטים הנתמך מקומית באתר S_i.
הסט $GL_{i(k)}=L_{(k)} \cap LL_{i(k)}$ הוא סט הפריטים הגלובלי בגודל k הנתמך מקומית באתר S_i.
המטרה היא למצוא את כל הסטים $L_{(k)}$ לכל $k>1$ ואת כמות התמיכה שלהם, ומכאן לחשב את כל החוקים הנתמכים על ידי התמיכה ורמת הביטחון.

להלן אלגוריתם מהיר לכריית חוקי מידע מבוזרים [21], זהו אלגוריתם שאינו שומר על פרטיות המידע, אך מהווה את הבסיס לכריית מידע אופקית השומרת על פרטיות המידע.

A Fast Algorithm For Distributed Association Rule Mining - FDM:

1. **Candidate Sets Generation:**
   Generate candidate sets *CG*i(k) based on *GL*i(k-1) , itemsets that are supported by the Si at the (k-1)th iteration, using the classic a priori candidate generation algorithm .Each site generates candidates based on the intersection of globally large (k-1) itemsets and locally large (k-1) itemsets .
2. **Local Pruning:**
   For each *X* *CGi(k)*, scan the database *DBi* at Si to compute *X.supi*.If *X* is locally large *Si*, it is included in the *LLi*(k) set. It is clear that if *X* is supported globally, it will be supported in one site .
3. **Support Count Exchange:**
   *LL*i(k) are broadcast and each site computes the local support for the items in *UiLL*i(k) .
4. **Broadcast Mining Results:**
   Each site broadcasts the local support for itemsets in *UiLL*i(k). From this, each site is able to compute *L*(k) .

שלב 1 – מציאת מועמדים מקומיים באורך k המבוססים על החוקים הנתמכים גלובלית באתר המקומי באורך k-1.
שלב 2 – ניפוי כל המועמדים המקומיים שאינם ניתכים על ידי האתר המקומי.
שלב 3 – החלפת מידע בין האתרים לגבי החוקים הנתמכים שנמצאו.
שלב 4 – חישוב התמיכה הגלובלית של חוקים אלו, וקבלת התוצאה הסופית.

2.1.2.2 חישוב מאובטח בין מספר אתרים

2.1.2.2.1 *Security in Semihonest Model*[23]:

שותפים הכנים למחצה מקיימים את חוקי הפרוטוקול , אך חופשיים במהלך ולאחר ביצוע הפרוטוקול לנסות לחשוף מידע פרטי. הדבר תואם את העולם האמיתי בו שותפים מעוניינים לכרות מידע אמיתי לתועלת הפרטית שלהם.

על פי [23] חישוב הוא מאובטח אם כל צד תוך כדי ביצוע הפרוטוקול יכול לראות נתוני קלט ופלט מזויפים. זה לא בדיוק כמו להסתיר נתונים פרטיים . לדוגמא, אם שני צדדים משתמשים בפרוטוקול מאובטח לכריית חוקי מידע מבוזר עדיין הפרוטוקול המאובטח מגלה שאם חוק הנתמך גלובלית אינו נתמך על ידי צד אחד, הרי שבהכרח הצד השני תומך בו. אתר יכול להסיק מידע על ידי התבוננות בחוקים שלו, ובאוסף החוקים הגלובליים שהתקבלו . מצד שני לא ניתן להסיק את התמיכה המדויקת של סט פריטים מסוים מתוך אוסף החוקים הגלובליים שהתקבלו. אם קיימים 3 צדדים או יותר, ידיעה שחוק מסוים נתמך גלובלית חושף שהחוק מתקיים לפחות באתר אחד , אך האתרים השונים לא יכולים לדעת מיהו (פרט לאתר המקיים את החוק). מכאן, פרוטוקול מאובטח בין מספר אתרים לא יוביל לחשיפת מידע נוסף לצד כלשהו מעבר לנתונים המתקבלים כקלט ופלט לצדדים.

[24]Yao's General Two-Party secure Function Evaluation *2.1.2.2.2*

על פי Yao השוואה מאובטחת בין שני צדדים מבוססת על ביטוי הפונקציה f(x,y) כמעגלית והצפנת השערים להערכה מאובטחת . בעזרת הפרוטוקול , כל פונקציה בין שני צדדים יכולה להיות מוערכת בבטחה באמצעות ה מודל הכנה-למחצה, semihonest model. כדי להיות יעילים לפונקציה חייב להיות ייצוג מעגלי . ניתן להשתמש בפונקציה כדי לבדוק אם a≤b (Yao's millionaire problem). זוהי אחת מהפונקציות היעילות ביותר כדי להשוות בביטחון בין שני ערכים.

Commutative encryption 2.1.2.3

הצפנה קומוטטיבית היא כלי חשוב ומשמשת בהרבה פרוטוקולים לשמירת פרטיות . אלגוריתם הצפנה הוא קומוטטיבי אם לכל מפתחות ההצפנה K1,…,KnЄK, לכל הודעה M ולכל פרמוטציה i,j מתקיימות שתי המשוואות הבאות:

1. $$E_{K_{i_1}}(\ldots E_{K_{i_n}}(M)\ldots) = E_{K_{j_1}}(\ldots E_{K_{j_n}}(M)\ldots).$$

2. $$\forall M_1, M_2 \in M \text{ such that } M_1 \neq M_2 \text{ and for given } k,\ \epsilon < \frac{1}{2^k}$$

$$Pr(E_{K_{i_1}}(\ldots E_{K_{i_n}}(M_1)\ldots) = E_{K_{j_1}}(\ldots E_{K_{j_n}}(M_2)\ldots)) < \epsilon.$$

ההצפנה הקומוטטיבית משמשת גם לבדיקה אם שני עצמים זהים מבלי לחשוף אותם . כל צד מצפין את העצם שהוא רוצה לבדוק, שולח לצד השני שמוסיף את ההצפנה שלו . על פי משוואה (1) לשניהם אותו ערך מוצפן אם העצמים שווים, ועל פי (2) אם הם שונים נגלה זאת בסבירות גבוהה מאד.

2.1.3 כרייה בטוחה של חוקי הקשר

2.1.3.1 סקירת השיטה:

לשיטה שני שלבים המשלבים יצירת חוקים מקומיים וספירתם על ידי העברות מידע מוצפנות בין האתרים. תחילה מתגלים הסטים המועמדים, (התדירים באתר אחד או יותר), ואז נקבע מי מהסטים המועמדים נתמך גלובלית.

ראוי לציין כי המילה "חוקים" מתייחסת לחוקי הקשר , ומשמעותה סט-פריטים (itemsets) תדירים.

השלב ראשון – גילוי החוקים המועמדים.

משתמשים בהצפנה קומוטטיבית. כל צד מצפין את החוקים התדירים שלו . החוקים המוצפנים מועברים בין כל האתרים שמצפינים אותם , עד שכל החוקים מוצפנים על ידי כל האתרים . החוקים מועברים כעת לצד משותף להסרת כ פילויות ולהתחלת הפענוח. כל חוק עובר כעת בין האתרים השונים המפענחים אותו. התוצאה הסופית היא אוסף החוקים המועמדים. (ראה ציור 1) המקיימים, כל חוק תדיר לפחות באתר אחד (ע"פ הלמה מסעיף 2.1.1), ושנית, אף אתר לא יודע אילו חוקים תדירים באיזו אתר.

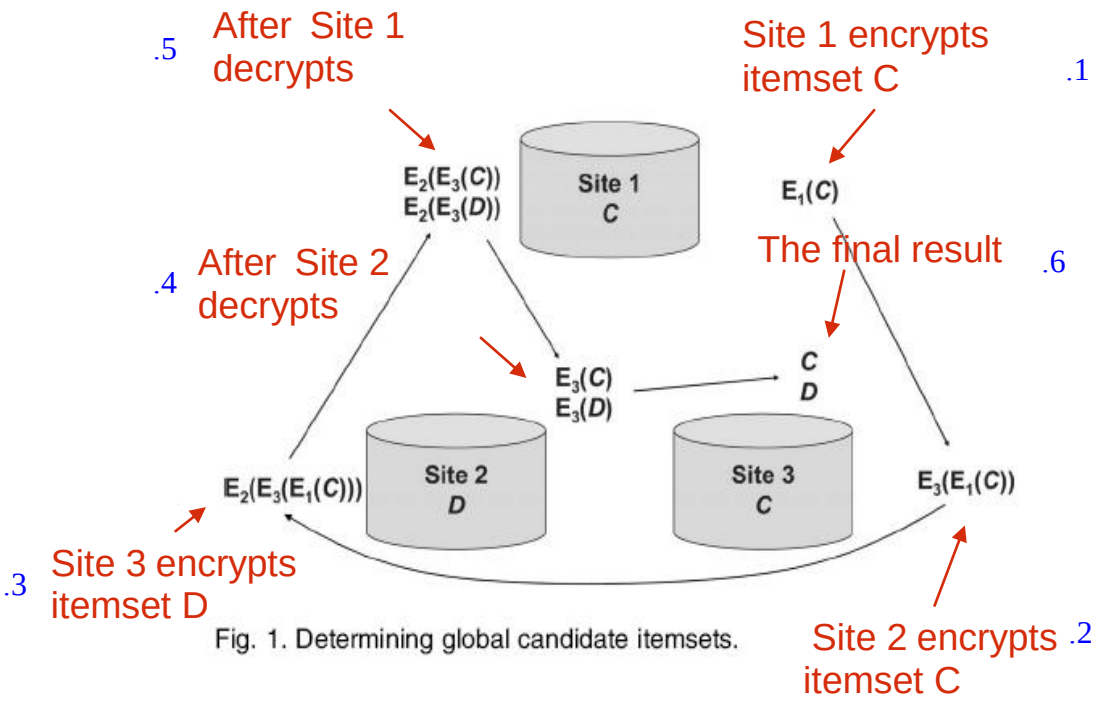


Fig. 1. Determining global candidate itemsets.

השלב השני – בדיקה אם חוק הקשר נתמך גלובלית מעל k תמיכה
כל חוק מועמד נבדק כדי לגלות אם הוא נתמך גלובלית . כל אתר בודק מהי התמיכה המקומית שלו בחוק. האתר הראשון מחשב ערך אקראי R ומוסיף אותו לכמות התמיכה שלו בחוק הנבדק. ערך זה מועבר לאתר השני שמו סיף את הערך שקיבל לכמות התמיכה שלו בחוק ומעביר את הערך שהתקבל לאתר השלישי ששוב מוסיף אותו לתמיכה שלו בחוק. הערך הסופי נבדק על ידי חישוב מאובטח אם הוא עובר את הערך R, ואם כן, אז החוק נתמך גלובלית. (ראה ציור 2)

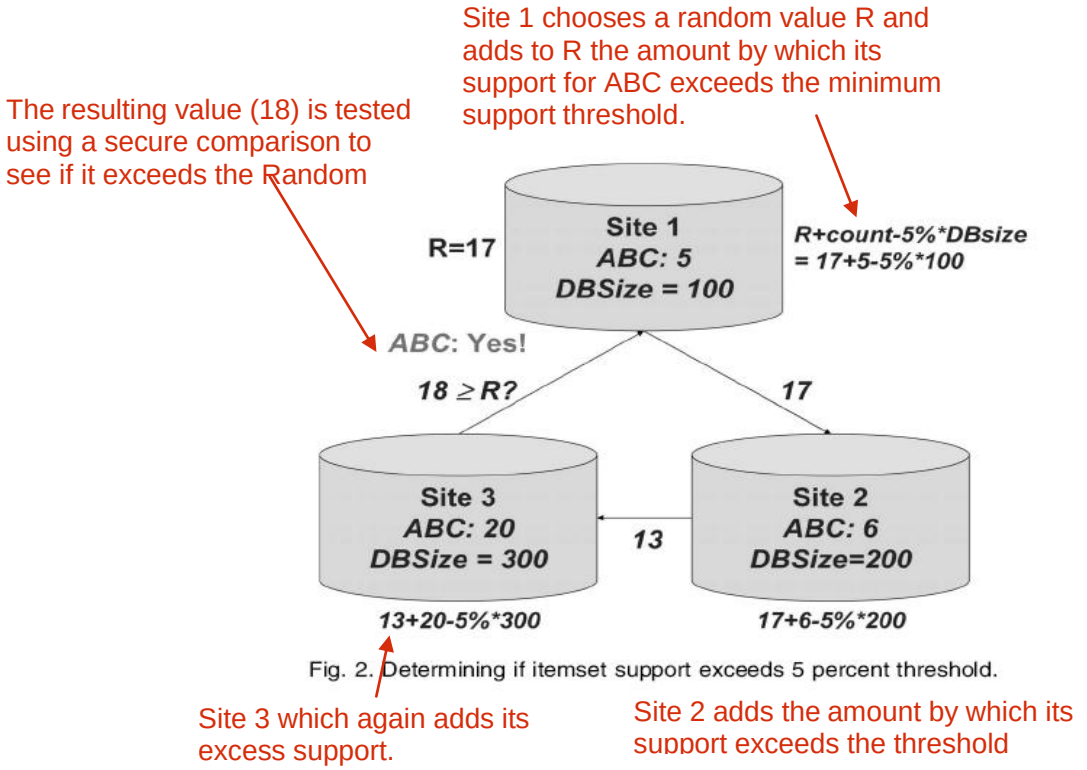


Fig. 2. Determining if itemset support exceeds 5 percent threshold.

2.1.3.2 הגדרת הבעיה:

יהי $i \geq 3$ מספר האתרים. לכל אתר טרנזקציות פרטיות על בסיס הנתונים DB_i. סף התמיכה s, support threshold s, ורמת הביטחון c נתונים. המטרה היא לגלות את כל חוקי ההקשר המספקים את

הסף, תוך כדי שמירה על הפרטיות, כלומר, אסור לאף אתר להיות מסוגל לגלות את תכולת ה טרנזקציה של אף אתר אחר, איזה חוק נתמך באיזה אתר, או את הערך המדויק של תמיכה/רמת ביטחון של חוק כלשהו באתר אחר, אלא אם כן המידע מתקבל מהחוקים הפרטים של האתר עצמו והתוצאה הסופית . ההנחה היא שאין קנוניה.

Secure Union of Locally Large Itemsets 2.1.3.3

השיטה מבו ססת על האלגוריתם FDM (שהובא לעיל), תוך שימוש בפרוטוקולים מיוחדים המחליפים את השידורים של LLi(k) ואת הרמת התמיכה של העצמים ב LL(k) (שלבים 3 ו-4).

תיאור האלגוריתם בפרוטוקול 1, איור 3 :

שלב 0 : מציאת החוקים באורך k הנתמכים על ידי כל אתר i, בדומה ל FDM. ואז, כל אתר מצפין את החוקים הנתמכים שלו בתוספת חוקים מזויפים , "fake" itemsets, כדי להסתיר את כמות החוקים שבהם הוא תומך. בפרוטוקול, F מייצג את הנתונים המשמשים כנתונים מזויפים.

שלב 1 : לאחר שכל אתר מצפין את החוקים שלו מפני האתרים האחרים הוא משנה להם את הסדר ושולח אותם לאתר הבא אחריו . כל אתר מצפין בעזרת מפתח ההצפנה שלו את החוקים המוצפנים שהגיעו מהאתר הקודם, ושוב מערבב ושולח לאתר הבא אחריו , מעין מעגל הצפנות. בסופו של התהליך, כל החוקים המועמדים באורך k מוצפנים על ידי כל אחד מהאתרים השונים. הסימון |LLei(k)| מייצג את הסטים המוצפנים באורך k באתר i. Ei מייצג את תהליך ההצפנה באתר i ו Di מייצג את הפענוח.

שלב 2 : איחוד כל החוקים המוצפנים בשני אתרים, כל אתר זוגי שולח את החוקים המוצפנים שאצלו לאתר S_1, וכל אתר אי-זוגי שולח את החוקים המוצפנים שאצלו לאתר S_0.

שלב 3 : אתר S_1 שולח את איחוד החוקים שאצלו לאתר S_0, שמבצע איחוד לכל חוקי ההקשר המועמדים, המוצפנים.

שלב 4 : שלב הפענוח. האתר S_0 מפענח את החוקים שברשותו ושולח אותם לאתר הבא אחריו, שמפענח ושולח לבא אחריו. האתר האחרון, לאחר הפענוח מסיר את החוקים המזויפים ושולח לכל האתרים האחרים את החוקים המועמדים שהתקבלו.

הסטים של הפריטים המוצפנים מתמזגים ב שלבם 2, 3 באלגוריתם. מכיוון שמשוואה (1) בהצפנה קומוטטיבית מתקיימת, כפילויות בחוקים הנתמכים מקומית נמחקים . הסיבה לכך שפעולת המיזוג נעשית בשני שלבים שונים (2, 3) נובעת מכך שאם אתר יודע מהם כל הסטים המוצפנים שהגיעו מכל האתרים הוא יכול לחשב את גודל החיתוך בין כל קבוצה של אתרים ואז לנחש מהם העצמים הנתמכים בכל אתר. שינוי הסדר לאחר ההצפנה בכל אתר ב שלב 1 מונע את האפשרות לדעת בוודאו ת איזה סט תואם . המיזוג הנפרד מאתרים זוגיים ואי זוגיים ב שלב 2 מונע מכל אתר לדעת את מלוא הערכים המוצפנים של עצמו . ב- שלב 4 מפענחים את הסטים הממוזגים התדירים . הקומוטטיביות של ההצפנה מאפשרת לפענח את כל הסטים באותו הסדר, מבלי להתחשב בסדר ההצפנה.

Protocol 1 Finding secure union of large itemsets of size k

Require: $N \geq 3$ sites numbered $0..N-1$, set F of non-itemsets.

Phase 0: Encryption of all the rules by all sites
for each site i **do**
  generate $LL_{i(k)}$ as in steps 1 and 2 of the FDM algorithm
  $LLe_{i(k)} = \emptyset$
  for each $X \in LL_{i(k)}$ **do**
    $LLe_{i(k)} = LLe_{i(k)} \cup \{E_i(X)\}$
  end for
  for $j = |LLe_{i(k)}| + 1$ to $|CG_{(k)}|$ **do**
    $LLe_{i(k)} = LLe_{i(k)} \cup \{E_i(\text{random selection from } F)\}$
  end for
end for

Phase 1: Encryption by all sites
for Round j = 0 to $N-1$ **do**
  if Round j= 0 **then**
    Each site i sends permuted $LLe_{i(k)}$ to site $(i+1) \bmod N$
  else
    Each site i encrypts all items in $LLe_{(i-j \bmod N)(k)}$ with E_i, permutes, and sends it to site $(i+1) \bmod N$
  end if
end for{At the end of Phase 1, site i has the itemsets of site $(i+1) \bmod N$ encrypted by every site}

Phase 2: Merge odd/even itemsets
Each site i sends $LLe_{i+1 \bmod N}$ to site $1 - (((i+1) \bmod N) \bmod 2)$
Site 0 sets $RuleSet_1 = \cup_{j=1}^{\lceil (N-1)/2 \rceil} LLe_{(2j-1)(k)}$
Site 1 sets $RuleSet_0 = \cup_{j=0}^{\lceil (N-1)/2 \rceil} LLe_{(2j)(k)}$

Phase 3: Merge all itemsets
Site 1 sends permuted $RuleSet_1$ to site 0
Site 0 sets $RuleSet = RuleSet_0 \cup RuleSet_1$

Phase 4: Decryption
for i = 0 to $N-1$ **do**
  Site i decrypts items in $RuleSet$ using D_i
  Site i sends permuted $RuleSet$ to site $i+1 \bmod N$
end for
Site $N-1$ decrypts items in $RuleSet$ using D_{N-1}
$RuleSet_{(k)} = RuleSet - F$
Site $N-1$ broadcasts $RuleSet_{(k)}$ to sites $0..N-2$

Fig. 3. Protocol 1: Finding secure union of large itemsets of size k.

<u>טענה</u>: פרוטוקול1 מחשב בפרטיות את האיחוד של הפריטים הגדולים ה מקומיים, בהנחה שאין קנוניה, וחושף לכל היותר את תוצאת $\cup_{i=1}^{N} LL_{i(k)}$ ואת:

1. גודל החיתוך של סט הפריטים הנתמכים מקומית של כל תת קבוצה של אתרים אי-זוגיים.
2. גודל החיתוך של סט הפריטים הנתמכים מקומית של כל תת קבוצה של אתרים זוגיים.
3. מספר הסטים הנתמכים לפחות על ידי אתר אחד אי-זוגי ולפחות אתר אחד זוגי.

(ההוכחה מופיעה במאמר ומבוססת על כל אחד מהשלבים בפרוטוקול).

Testing Support Threshold without Revealing Support Count 2.1.3.3

כדי לחשב מי מהקבוצה של פריטים גדולים מקומיים $LL_{(k)}$ נתמכים גלובלית מבלי לחשוף מהי התמיכה של כל אתר בחוק, יש לבדוק לכל $X \epsilon LL_{(k)}$ אם $X.sup \geq s\% \times |DB|$?
נהפוך את הבעיה למקומית באופן הבא:

$$X.sup \geq s * |DB| = s * \left(\sum_{i=1}^{n} |DB_i| \right)$$

$$\sum_{i=1}^{n} X.sup_i \geq s * \left(\sum_{i=1}^{n} |DB_i| \right)$$

$$\sum_{i=1}^{n} (X.sup_i - s * |DB_i|) \geq 0.$$

מכאן שמספיק לבדוק אם הנוסחא האחרונה מתקיימת . הבעיה לבצע זאת מבלי לחשוף את $X.sup_i$ או את $|DB_i|$.

פרוטוקול 2 איור 4 מתאר את האלגוריתם:

Protocol 2 Finding the global support counts securely

Require: $N \geq 3$ sites numbered $0..N-1$, $m \geq 2 * |DB|$
$rule\_set = \emptyset$
at site 0:
for each $r \in candidate\_set$ **do**
choose random integer x_r from a uniform distribution over $0..m-1$;
$t = r.sup_i - s * |DB_i| + x_r \pmod m$;
$rule\_set = rule\_set \cup \{(r,t)\}$;
end for
send $rule\_set$ to site 1 ;
for $i = 1$ to $N-2$ **do**
for each $(r,t) \in rule\_set$ **do**
$\bar{t} = r.sup_i - s * |DB_i| + t \pmod m$;
$rule\_set = rule\_set - \{(r,t)\} \cup \{(r,\bar{t})\}$;
end for
send $rule\_set$ to site $i+1$;
end for
at site N-1:
for each $(r,t) \in rule\_set$ **do**
$\bar{t} = r.sup_i - s * |DB_i| + t \pmod m$;
securely compute if $(\bar{t} - x_r) \pmod m < m/2$ with the site 0; { Site 0 knows x_r }
if $(\bar{t} - x_r) \pmod m < m/2$ **then**
multi-cast r as a globally large itemset.
end if
end for

Fig. 4. Protocol 4: Finding the global support counts securely.

האתר הראשון יוצר מספר רנדומאלי x_r לכל סט פריטים X, מוסיף אותו לחישוב $(X.sup_i - s * |DB_i|)$ ושולח אותו לאתר הבא , שמוסיף את המספר שקיבל מהאתר הקודם לחישוב שלו ושוב שולח אותו הלאה . המספר שנשלח לכל אתר ממוסך על ידי המספר הרנדומאלי שהתקבל מהאתר הראשון , כך שגם החישוב הבא ממוסך , והאתר אינו יכו ל ללמוד ממנו כלום . לאתר האחרון נותר החישוב הבא:

$$\sum_{i=1}^{n} (X.sup_i - s * |DB_i|) + x_r \pmod m$$

(כל החישובים הם מודולו m, כאשר m≥2*|DB|).
האתר האחרון צריך לבדוק אם כאשר מפחיתים x_r מהמספר שהתקבל התוצאה קטנה מ- m/2.
מכיוון שהאתר הראשון יודע מהו x_r ניתן לבצע בדיקה באמצעות האלגוריתם של Yao להשוואה מאובטחת בין שני הצדדים.

פרוטוקול 2 מחשב בפרטיות את התמיכה הגלובלית ב semihonest model. (הוכח במאמר).

Securely Finding Confidence of a Rule 2.1.3.4

כדי למצוא אם חוק $X \Rightarrow Y$ גבוה מסף רמת הביטחון c (confidence threshold c) הנתון, יש לבדוק אם $\frac{\{X \cup Y\}.\sup}{Y.\sup} \geq c$. המשוואה הבאה מראה כיצד לחשב בצורה מאובטחת אם רמת הביטחון עבר את הסף הדרוש. נסמן $\{X \cup Y\}.\sup_i$ כ- $XY.\sup_i$:

$$\frac{\{X \cup Y\}.sup}{Y.sup} \geq c \Rightarrow \frac{\sum_{i=1}^{i=n} XY.sup_i}{\sum_{i=1}^{i=n} X.sup_i} \geq c$$

$$\Rightarrow \sum_{i=1}^{i=n} XY.sup_i \geq c * \left(\sum_{i=1}^{i=n} X.sup_i\right)$$

$$\Rightarrow \sum_{i=1}^{i=n} (XY.sup_i - c * X.sup_i) \geq 0.$$

מכיוון שכל אתר יודע מהו $XY.\sup_i$ ו- $X.\sup_i$, הרי שניתן לחשב בקלות את רמת הביטחון של החוק באמצעות פרוטוקול 2.

2.1.4 אבטחה כנגד קנוניה

בפרוטוקול 1, ניתן באמצעות קנוניה לחשוף את הסטים התדירים של אתר מסוים אחרי ההצפנה על ידי כל הצדדים האחרים . בכך ניתן ללמוד את גודל החיתוך בין הסטים שלו לבין האתרים האחרים . במיוחד אם אתר i קושר קשר עם אתר i-1 , הוא יכול ללמוד את גודל החיתוך שלו עם אתר i+1. קנוניה בין אתר 0 ואתר 1 מחריף את הבעיה מכיוון שהם מכירים את כל הערכים המוצפנים מהאתרים הזוגיים/אי-זוגיים. בכך הם יכולים לחשוף את הסטים עצמם. אם $\left|LL_{i(k)} \cap LL_{i+1(k)}\right| = LL_{i(k)}$ אזי אתר i יכול ללמוד תת סט של אתר i+1.

גם בפרוטוקול 2 קנוניה יכולה להיות בעיה מכיוון שאם אתר i-1 ואתר i+1 ישתפו פעולה הם יכולים לחשוף את התוספת של i לערך התמיכה שלו . ניתן לחסום את האפשרות לקנוניה בפרוטוקול . הרעיון הוא שכל צד יחלק את הקלט שלו ל n חלקים וישלח n-1 חלקים לאתרים השונים . כעת כדי לחשוף קלט של אתר כלשהו, n-1 אתרים צריכים לקשור קשר לקנוניה.
הפרוטוקול הבא המתואר בהרחבה ב[25] מתאר את השיטה:

1. Each site i randomly chooses n elements such that $x_i = \sum_{j=1}^{n} z_{i,j} \bmod m$, where x_i is the input of site i. Site i sends $z_{i,j}$ to site j.
2. Every site i computes $w_i = \sum_{j=1}^{n} z_{j,i} \bmod m$ and sends w_i to site n.
3. Site n computes the final result $\sum_{i=1}^{n} w_i \bmod m$.

2.1.5 הקושי במקרה של שני אתרים

קיימת בעיה באבטחת המידע במקרה שיש רק שני אתרים . ראשית, חוק גלובלי שאינו נתמך באתר אחד, ידוע שהוא נתמך באתר השני. על פי פרוטוקול 1 המצב גרוע יותר, חוק שנתמך מקומית גם אם הוא לא נתמך גלובלית ידוע לצד השני . כדי לקבל מידת מה של אבטחה לא נבצע pruning מקומי, אלא נחשב תחילה את כל המועמדים האפשריים CG(k) מ- L(k-1) (שלבים 1 ו-2 מהאלגוריתם FDM) ונחשב תמיכה לכל המועמדים. בשלב הבא (פרוטוקול 2) לא ניתן לוותר על השוואה מאובטחת, אחרת תיחשף התמיכה המקומית של הצדדים.

אולם הבעיה הראשונה שהוצגה עדיין קיימת ללא קשר למידת האבטחה של החישוב, לכן שיטה זו לכריית מידע בין שני צדדים אינה מעשית.

2.1.6 עלות החישוב והתקשורת:

מספר האתרים הוא N, המספר הכללי של המועמדים בכל אתר הוא |CGi(k)| ומספר המועמדים שיכולים להיווצר מהפריטים הגלובליים ה-(k-1) גדולים הוא |CG(k)| (=apriory_gen(L(k-1))). התוספת לתמיכה X.supi-|DBi| של סט פריטים X יכול להיות מיוצג כ $m = \left\lceil \log_2 \left(2 * |DB| \right) \right\rceil$ סיביות. יהי t מספר הסיביות בפלט המוצפן של סט פריטים . חסם תחתון על t הוא $\log_2 \left(\left| CG_{(k)} \right| \right)$. בהתבסס על סטנדרט הצפנה נוכחי t=512.

העלות המוחלטת של תקשורת סיביות עבור פרוטוקול 1 היא $O\left(t * \left| CG_{(k)} \right| * N^2\right)$.

כאשר התקשורת מקבילית ניתן לחלק ב N לקבלת הערכת זמן.

פרוטוקול 2 דורש $O\left(m * \left| \cup_i LL_{i(k)} \right| * \left(N + t\right)\right)$ סיביות של תקשורת . גורם ה t הוא עבור ההערכה המאובטחת בין אתר 0 לביו אתר N-1 כדי להחליט עבור כל סט פריטים אם הוא נתמך . ההשוואה הסופית דורשת עלות חישוב של $O\left(\left| \cup_i LL_{i(k)} \right| * m * t^3\right)$.

2.1.7 מסקנות:

לדעתי השיטה שהוצגה לכריית חוקי מידע מבוזרים אופקית יעילה ובעלת הנחות אבטחה מציאותיות. השימוש בכלים קריפטוגרפים מאפשר לבצע כריית מידע תוך שמירה על פרטיות , ובכך להשיג שיתוף פעולה מלא של הצדדים השותפים לכרייה.

2.2 Privacy Preserving Association Rule mining In Vertically Partitioned Data[26]

2.2.1 הקדמה

שימור פרטיות המידע בכריית מידע על חלוקה אנכית של נתונים מתייחס למצב בו טרנזקציות פועלות על פני מספר מקורות. כל אתר מחזיק מספר מאפיינים של כל טרנזקצייה וכולם מעוניינים לשתף פעולה על מנת לגלות חוקי הקשר משותפים. הצדדים אינם מעוניינים לחשוף נתונים ומידע פרטי.

לדוגמא: אתר אחד המנהל מכירות מצרכי מכולת , ואתר אחר מכירות בגדים . על ידי הצלבת מידע בניהם, למשל על פי מספר כרטיס האשראי ניתן לגלות הרגלי צריכה של לקוחות , או אפילו של לקוח מסוים, אך קיימת כאן חדירה לפרטיות אותו לקוח והפרת האמון בינו לבין בית העסק.

דוגמא נוספת: מפעלי FORD רכשו צמיגים מיצרן מסוים . למרות שהתגלו ליקויים מסוימים בכל אחת מהחברות בנפרד, לא נחשפו מימדי הבעיה. הפעלת חוקי הקשר בניהם בזמן המתאים היה חוסך החלפת 1.4 מיליון צמיגים. נזק כלכלי לא מבוטל.

ניתן כמובן לנסות לנקוט בגישה הפשוטה , לבצע כריית מידע על כל בסיס נתונים בנפרד ולהצליב את התוצאות. גישה זו נכשלת בדרך כלל במציאת תוצאות מעשיות. הסיבות לכך הן:

1. נתוני ישות אחת יכולה להיות מפוצלים במספר בסיסי נתונים .כריית מידע בכל בסיס נתונים בנפרד אינה יכולה לחשוף את הקשר בין הנתונים –הבעיה המרכזית .
2. פריט מסוים יכול להימצא במקביל במספר בסיסי נתונים , כך שיכול להיות לו משקל יתר בתוצאות.
3. נתונים באתר מסוים בדרך כלל מיצגים אוכלוסיה הומוגנית , גיאוגרפית או דמוגרפית . תוצאת כריית המידע מסתירה עובדה זו.

בכריית מידע ומציאת חוקי הקשר בין שני בסיסי נתונים האחד נחשב כ primary והוא זה שמפעיל את הפרוטוקול והשני נחשב כ responder. קיים join key המצוי בכל אחד מהם . שאר התכונות מצויות בכל אחד מבסיסי הנתונים, אך לא בשניהם ביחד. המטרה היא למצוא חוקי הקשר שאינם קשורים דווקא למפתח המשותף. כמובן תוך כדי שמירה על פרטיות המידע.
מאמר זה הוצג בעבודת הסמינר.

2.2.2 הגדרת הבעיה:

קיימת חלוקה אנכית של בסיס נתונים בין שני שותפים A ו- B. בעיית כריית חוקי ההקשר מוגדרת כלהלן :

1. יהי $\text{Let } \mathcal{I} = \{i_1, i_2, \cdots, i_m\}$ סט של ליטראלים (פריטים).
2. ויהי D סט של טרנזקציות ,כאשר כל טרנזקציה T היא סט של פריטים, כך ש $T \subseteq \mathcal{I}$.
3. הקשר עם כל טרנזקציה מוגדר כמזהה יחיד הקרוי TID.
4. טרנזקציה T מכילה את X אם X הוא סט של פריטים מ $T \subseteq \mathcal{I}$ ו $X \subseteq T$.
5. חוק הקשר הוא יחס גרירה מהצורה $X \Rightarrow Y$ כאשר $X \subset \mathcal{I}, Y \subset \mathcal{I}$ וגם $X \cap Y = \phi$.
6. החוק $X \Rightarrow Y$ מקיים את רמת הביטחון c ב D אם c% מהטרנזקציות ב D המכילות את X גם מכילות את Y.
7. החוק $X \Rightarrow Y$ מקיים תמיכה s ב D אם s% מהטרנזקציות ב D מכילות $X \cup Y$.

אנו מניחים חוקי כריית מידע בוליאניים , בהם קיום או העדר תכונה מסוימת מיוצגים על ידי 1 או 0 בהתאמה . טרנזקציות הינן מחרוזות של 0 ו -1.

כדי לגלות אם קבוצת פריטים (Itemset) מסוימת היא תדירה יש לספור את מספר הרשומות שבהן הערכים עבור כל התכונות בקבוצת הפריטים היא 1. זה מתורגם לבעיה מתמטית פשוטה על פי ההגדרה שלהלן:

1. יהי מספר התכונות $l+m$, כאשר ל l-תכונות A עד A_l, ול m-התכונות B הנותרות B_1 עד B_m.
2. טרנזקציות/רשומות הינן רצף של $l+m$ 1-ות או 0-ים .
3. יהי k סף התמיכה הנדרש ויהי n המספר הסופי של הטרנזקציות/הרשומות .
4. יהיו $\vec{X}$ ו $\vec{Y}$ -מיצגים עמודות בבסיסי הנתונים , כך ש $x_i=1$ אם"ם לשורה i יש את הערך 1 לתכונה X.
5. המכפלה הסקלארית של שני וקטורים ראשיים באורך n מוגדרת כ -

$$\vec{X} \cdot \vec{Y} = \sum_{i=1}^{n} x_i * y_i$$

6. כדי לקבוע אם שני פריטים $\langle XY \rangle$ הם תדירים ,מספיק לבדוק אם

$$\vec{X} \cdot \vec{Y} \geq k$$

<u>הרחבת הפרוטוקול ל -w פריטים:</u>

נניח כי ל A p-תכונות ול B q-תכונות. רוצים לחשב את התדירות של $w = p + q - \text{itemset} \langle a_1,...,a_p,b_1,...,b_q \rangle$. כל פריט מ $\vec{X}$-ו $\vec{Y}$ -מורכב ממכפלת הפריטים האינדיבידואליים בכל אתר, מבלי לשתף מידע מ-A ומ-B, כלומר

$$x_i = \prod_{j=1}^{p} a_j \text{ and } y_i = \prod_{j=1}^{q} b_j$$

.

כעת יש לחשב באופן בטוח את המכפלה הסקלארית בין שני האתרים כדי לגלות את התדירות של כל w -הפריטים.

לדוגמא: נניח כי נרצה לגלות אם סט של 5פריטים הוא תדיר ,כאשר ל- A 2 תכונות ול-B 3 התכונות הנוספות .כלומר A- B רוצים לגלות אם הסט l = <Aa,Ab,Ba,Bb,Bc> הוא תדיר.

A ייצור וקטור חדש $\vec{X}$ באורך n כאשר $\vec{X} = \vec{A}_a * \vec{A}_b$ ו- B ייצור וקטור חדש $\vec{Y}$ באורך n כאשר $\vec{Y} = \vec{B}_a * \vec{B}_b * \vec{B}_c$. כעת חישוב המכפלה הסקלארית של $\vec{X}$ ו $\vec{Y}$ -יחזיר את התדירות של סט הפריטים.

A and B want to know if the itemset I=<3, 7, 10,12, 16> is frequent

Database A

TID	Items
100	1, 3, 4, 5, 8
300	2, 3, 7, 8
150	1, 2, 3, 5, 7
400	2, 4, 5, 7, 8
500	1, 3, 4, 7

Database B

TID	Items
100	9,10,13,16
300	10,12, 13, 16
150	9, 15, 16
400	12, 13, 15
500	10,12,14,16

$\vec{X}$	$\vec{A}_a * \vec{A}_b$
0	1*0
1	1*1
1	1*1
0	0*1
1	1*1

$\vec{Y}$	$\vec{B}_a * \vec{B}_b * \vec{B}_c$
0	1*0*1
1	1*1*1
0	0*0*1
0	0*1*0
1	1*1*1

$$\vec{X} \cdot \vec{Y} = 0+1+0+0+1 = 2$$

2.2.3 תיאור האלגוריתם למציאת כל סטי-הפריטים התדירים:

1. $L_1 = \{\text{large 1-itemsets}\}$
2. for (k=2; $L_{k-1} \neq \phi$; k++) do begin
3. C_k = apriori-gen(L_{k-1});
4. for all candidates $c \in C_k$ do begin
5. if all the attributes in c are entirely at A or B
6. that party independently calculates $c.count$
7. else
8. let A have l of the attributes and B have the remaining m attributes
9. construct $\vec{X}$ on A's side and $\vec{Y}$ on B's side where $\vec{X} = \prod_{i=1}^{l} \vec{A}_i$ and $\vec{Y} = \prod_{i=1}^{m} \vec{B}_i$
10. compute $c.count = \vec{X} \cdot \vec{Y} = \sum_{i=1}^{n} x_i * y_i$
11. endif
12. $L_k = L_k \cup c | c.count \geq minsup$
13. end
14. end
15. Answer = $\cup_k L_k$

שלב 1,מציאת כל סטי הפריטים באורך 1. בשלב 3 באלגוריתם, הפונקציה apriori-gen לוקחת כקלט את סט הפריטים באורך L_{k-1} שמצאנו במעבר (k-1) ויוצרת ממנו סטים מועמדים ב C_k באמצעות יצירת סט-על המכיל את כל המועמדים האפשריים, וביצוע pruning על הסט שנוצר [27].

בהינתן כמות ותדירות המופעים של כל סט פריטים ניתן לחשב את כל חוקי ההקשר המקיימים support ≥ minsup.

שלבים 1, 3, 10 ו-12 הם היחידים שדורשים שיתוף מידע. מכיוון שהתוצאה הסופית $U_k L_k$ ידועה לשני הצדדים, הרי ששלבים 1, 3, 12 אינם חושפים מידע נוסף לצדדים. שלב 10 בו נדרש לבצע מכפלה סקלארית בין הצדדים הוא היחיד שעלול לחשוף מידע, ולכן יש למצוא דרך לבצע אותו תוך שמירה על הפרטיות.

2.2.4 חישוב בטוח של מכפלה סקלארית:

החישוב המוצע להלן הינו חישוב אלגברי המחביא ערכי אמת במשוואות מיסוך המכילות ערכים רנדומאליים. המידע שנחשף על ידי משוואות אלו מאפשר לבצע חישובים על מידע פרטי המגיע ממקור אחר.

נניח מבלי הגבלת כלליות כי n הוא זוגי.

<u>שלב 1</u>: A מייצר n מספרים רנדומאליים R1…Rn. מתוך כך $\vec{X}$, ומטריצה C יוצרים ביעילות סט של משוואות ליניאריות בלתי תלויות, $\vec{X}$'שאותן Aשולח לB:

$$\langle x_1 + c_{1,1} * R_1 + c_{1,2} * R_2 + \cdots + c_{1,n} * R_n \rangle$$
$$\langle x_2 + c_{2,1} * R_1 + c_{2,2} * R_2 + \cdots + c_{2,n} * R_n \rangle$$
$$\vdots$$
$$\langle x_n + c_{n,1} * R_1 + c_{n,2} * R_2 + \cdots + c_{n,n} * R_n \rangle$$

<u>שלב 2</u>: B מחשב $S = \vec{X}' \cdot \vec{Y}$ ומחשב גם את nהערכים הבאים:

$$\langle c_{1,1} * y_1 + c_{2,1} * y_2 + \cdots + c_{n,1} * y_n \rangle$$
$$\langle c_{1,2} * y_1 + c_{2,2} * y_2 + \cdots + c_{n,2} * y_n \rangle$$
$$\vdots$$
$$\langle c_{1,n} * y_1 + c_{2,n} * y_2 + \cdots + c_{n,n} * y_n \rangle$$

B אינו יכול לשלוח ערכים אלו ל- A מכיוון של- A יהיו כעת n משוואות בלתי תלויות עם n נעלמים (y1 … yn), שיחשפו את ערכי ה-y. לכן B מייצר r ערכים רנדומאליים משלו : R'1…R'r.

B מחלק את ערכי ה n שיצר ל r קבוצות כאשר ערכי ה R' משמשים להסתרת המשוואות שנוצרו:

$$\langle c_{1,1} * y_1 + c_{2,1} * y_2 + \cdots + c_{n,1} * y_n + R'_1 \rangle$$
$$\vdots$$
$$\langle c_{1,n/r} * y_1 + c_{2,n/r} * y_2 + \cdots + c_{n,n/r} * y_n + R'_1 \rangle$$
$$\langle c_{1,(n/r+1)} * y_1 + c_{2,(n/r+1)} * y_2 + \cdots + c_{n,(n/r+1)} * y_n + R'_2 \rangle$$
$$\vdots$$
$$\langle c_{1,2n/r} * y_1 + c_{2,2n/r} * y_2 + \cdots + c_{n,2n/r} * y_n + R'_2 \rangle$$
$$\vdots$$
$$\langle c_{1,((r-1)n/r+1)} * y_1 + c_{2,((r-1)n/r+1)} * y_2 + \cdots + c_{n,((r-1)n/r+1)} * y_n + R'_r \rangle$$
$$\vdots$$
$$\langle c_{1,n} * y_1 + c_{2,n} * y_2 + \cdots + c_{n,n} * y_n + R'_r \rangle$$

B שולח את S ואת n ל- A שמבצע:

$$\begin{aligned} S = {} & (x_1 + c_{1,1} * R_1 + c_{1,2} * R_2 + \cdots + c_{1,n} * R_n) * y_1 \\ & + (x_2 + c_{2,1} * R_1 + c_{2,2} * R_2 + \cdots + c_{2,n} * R_n) * y_2 \\ & \vdots \\ & + (x_n + c_{n,1} * R_1 + c_{n,2} * R_2 + \cdots + c_{n,n} * R_n) * y_n \end{aligned}$$

פישוט נוסף וריכוז הרכיבים x_i*y_i נותן:

$$\begin{aligned} S = {} & (x_1 * y_1 + x_2 * y_2 + \cdots + x_n * y_n) \\ & + (y_1 * c_{1,1} * R_1 + y_1 * c_{1,2} * R_2 + \cdots + y_1 * c_{1,n} * R_n) \\ & + (y_2 * c_{2,1} * R_1 + y_2 * c_{2,2} * R_2 + \cdots + y_2 * c_{2,n} * R_n) \\ & \vdots \\ & + (y_n * c_{n,1} * R_1 + y_n * c_{n,2} * R_2 + \cdots + y_n * c_{n,n} * R_n) \end{aligned}$$

כאשר השורה הראשונה נותנת את התוצאה הרצויה . $\sum_{i=1}^{n} x_i * y_i$

בשלבים הבאים מסדרים מחדש את כל גורמי המ כפלה האנכיים, מסדרים את המשוואה מחדש להוצאת כל ה R_i:

$$
\begin{aligned}
S = &\sum_{i=1}^{n} x_i * y_i \\
&+R_1 * (c_{1,1} * y_1 + c_{2,1} * y_2 + \cdots + c_{n,1} * y_n) \\
&+R_2 * (c_{1,2} * y_1 + c_{2,2} * y_2 + \cdots + c_{n,2} * y_n) \\
&\vdots \\
&+R_n * (c_{1,n} * y_1 + c_{2,n} * y_2 + \cdots + c_{n,n} * y_n)
\end{aligned}
$$

כעת נבצע מניפולציה נוספת על המשוואה ונקבל :

$$
\begin{aligned}
S = &\sum_{i=1}^{n} x_i * y_i \\
&+\{R_1 * (c_{1,1} * y_1 + c_{2,1} * y_2 + \cdots + c_{n,1} * y_n) \\
&\quad +R_1 * R'_1 - R_1 * R'_1\} \\
&\vdots \\
&+\{R_{n/r} * (c_{1,n/r} * y_{n/r} + c_{2,n/r} * y_2 + \cdots + c_{n,n/r} * y_n) \\
&\quad +R_{n/r} * R'_1 - R_{n/r} * R'_1\} \\
&+\{R_{n/r+1} * (c_{1,n/r+1} * y_{n/r+1} + c_{2,n/r+1} * y_2 + \\
&\qquad \cdots + c_{n,n/r+1} * y_n) \\
&\quad +R_{n/r+1} * R'_2 - R_{n/r+1} * R'_2\} \\
&\vdots \\
&+\{R_{2n/r} * (c_{1,2n/r} * y_{2n/r} + c_{2,2n/r} * y_2 + \\
&\qquad \cdots + c_{n,2n/r} * y_n) \\
&\quad +R_{2n/r} * R'_2 - R_{2n/r} * R'_2\} \\
&\vdots \\
&\vdots \\
&+\{R_{(r-1)n/r+1} * (c_{1,(r-1)n/r+1} * y_{(r-1)n/r+1} + \\
&\qquad c_{2,(r-1)n/r+1} * y_2 + \cdots + c_{n,(r-1)n/r+1} * y_n) \\
&\quad +R_{(r-1)n/r+1} * R'_r - R_{(r-1)n/r+1} * R'_r\} \\
&\vdots \\
&+\{R_n * (c_{1,n} * y_1 + c_{2,n} * y_2 + \cdots + c_{n,n} * y_n) \\
&\quad +R_n * R'_r - R_n * R'_r\}
\end{aligned}
$$

כעת A מוציא את ה- R'_1 החוצה, ונקבל:

$$
\begin{aligned}
S &= \sum_{i=1}^{n} x_i * y_i \\
&+R_1 * (c_{1,1} * y_1 + c_{2,1} * y_2 + \cdots + c_{n,1} * y_n + R'_1) \\
&\vdots \\
&+R_{n/r} * (c_{1,n/r} * y_{n/r} + c_{2,n/r} * y_2 + \\
&\qquad \cdots + c_{n,n/r} * y_n + R'_1) \\
&+R_{n/r+1} * (c_{1,n/r+1} * y_{n/r+1} + c_{2,n/r+1} * y_2 + \\
&\qquad \cdots + c_{n,n/r+1} * y_n + R'_2) \\
&\vdots \\
&+R_{2n/r} * (c_{1,2n/r} * y_{2n/r} + c_{2,2n/r} * y_2 + \\
&\qquad \cdots + c_{n,2n/r} * y_n + R'_2) \\
&\vdots \\
&+R_{(r-1)n/r+1} * (c_{1,(r-1)n/r+1} * y_{(r-1)n/r+1} + \\
&\qquad c_{2,(r-1)n/r+1} * y_2 + \cdots + c_{n,(r-1)n/r+1} * y_n + R'_r) \\
&\vdots \\
&+R_n * (c_{1,n} * y_1 + c_{2,n} * y_2 + \cdots + c_{n,n} * y_n + R'_r) \\
&-R_1 * R'_1 - \cdots - R_{n/r} * R'_1 \\
&-R_{n/r+1} * R'_2 - \cdots - R_{2n/r} * R'_2 \\
&\vdots \\
&-R_{(r-1)n/r+1} * R'_r - \cdots - R_n * R'_r
\end{aligned}
$$

A מכפיל את n הערכים שקיבל מ B עם ערכי ה R_i ומחסר את הסכום מ- S לקבלת:

$$\begin{aligned} Temp &= \sum_{i=1}^{n} x_i * y_i \\ &\quad -R_1 * R'_1 - \cdots - R_{n/r} * R'_1 \\ &\quad -R_{n/r+1} * R'_2 - \cdots - R_{2n/r} * R'_2 \\ &\quad \vdots \\ &\quad -R_{(r-1)n/r+1} * R'_r - \cdots - R_n * R'_r \end{aligned}$$

הוצאת ה R'_i תיתן:

$$\begin{aligned} &Temp = \\ &\quad \sum_{i=1}^{n} x_i * y_i \\ &\quad -(R_1 + R_2 + \cdots + R_{n/r}) * R'_1 \\ &\quad -(R_{n/r+1} + R_{n/r+2} + \cdots + R_{2n/r}) * R'_2 \\ &\quad \vdots \\ &\quad -(R_{((r-1)n/r)+1} + R_{((r-1)n/r)+2} + \cdots + R_n) * R'_r \end{aligned}$$

לקבלת התוצאה הסופית המבוקשת A יוסיף את הסכום של r גורמי המכפלה ל Temp.

<u>שלב 3:</u> A שולח את r הערכים שהתקבלו ל- B ו- B שמכיר את ערכי R'_i מחשב את התוצאה הסופית. לבסוף, B שולח ל- A את התוצאה.

2.2.4.1 בחירת המטריצה C:

ערכי המטריצה C לפרוטוקול שלעיל דורשים מקדמים ליניאריים למשוואות בלתי תלויות. דרישה זו נובעת מהעובדה שהמשוואות משמשות להסתרת ערכי נתונים.

אם משוואה כלשהי יכולה להיזרק כאשר משתמשים בפחות ממחצית המשוואות האחרות הרי שייווצר קשר לחלק מ- n/2 מהנתונים הנסתרים.

מקדמי מטריצה הנוצרים על ידי pseudo-random function יספקו בסבירות גבוהה משוואות ליניאריות בלתי תלויות. השיטה מאפשרת בניה של מטריצה $c_{i,j}$ על ידי שיתוף של גרעין ופונקציה יוצרת.

2.2.5 מה חושפת השיטה

מטרת השיטה היא ליצור שיטה יעילה לחישוב חוקי הקשר מבלי לחשוף את ערכי הישויות, כאשר לא נדרש אפס מוחלט של מידע.

2.2.5.1 מה מתגלה

כל צד מכיר את הנתונים שלו, ולומד את הנתונים הגלובליים המתקבלים מכריית המידע. נתונים אלו חושפים מידע מסוים. למשל אם סף התמיכה עומד על 5% והחוק $A_1 \rightarrow B_1$ מתקיים, ובדיוק ב- 5% מהפריטים ב- A יש את הפריט A_1 , A יודע שלפחות אותם עצמים ב B מקיימים את התכונה

B_1.

הפרוטוקול מאפשר חשיפת מידע שיכול להתקבל מהנתונים הגלובליים והנתונים הפרטיים של כל בסיס נתונים.

השיטה חושפת יותר מאשר אחוזים או העדר חוק הנתמך מעבר לסף הנדרש, צד A יכול לגלות את התמיכה המדויקת של סט כלשהו. זה מגביר את ההסתברות ש A ילמד שסט מסוים ב- B מקיים תכונה מסוימת. הדבר מתרחש כאשר התמיכה הגלובלית של חוק כלשהו שווה לתמיכה של A באותו חוק. במקרה כזה, A מחשב את ההסתברות שעצם בקבוצה שנתמכת ב A מקיים תכונה ב B על פי השיעור של התמיכה הגלובלית בתמיכה של A.

לא סביר שמידע ספציפי אינדיווידואלי ייחשף בוודאות באמצעות השיטה.

2.2.5.2 הבעיה עם {0, 1}

כאשר כורים נתונים בינאריים הערכים המתקבלים מוגבלים ל-0 ו-1, כך שקיים סיכון בחשיפה בפרוטוקול זה ובאחרים [27, 28]. A יוצר n+r משוואות עם 2n נעלמים. B יכול לקחת r משוואות רק עם xi. אם למשוואות פתרון יח יד, B יכול לנסות את כל האפשרויות ש ל 0 ו-1 לכל yi ולקבל את הפתרון הנכון. כך גם ההפך, A יכול לקבל n-r משוואות רק עם yi.

פתרון אחד לבעיה הוא להבטיח שיש מספר פתרונות למשוואות על ידי בחירה חכמה של ערכי Ci,j כך שלא ניתן לגלות בוודאות מי מ xi /yi הוא 1.

להלן אחת מהשיטות:

נניח כי הצורה של משוואה הנשלחת על ידי B היא $C_{1,1} * y_1 + C_{2,1} * y_2 + \cdots + C_{n,1} * y_n + R'_1$ B. יכול לרכז את ה- yi לזוגות של 0 ו- 1 ולבחור באופן סלקטיבי סטים של מקדמים של חלק מהזוגות זהים. אז, גם אם A ימצא פתרון למשוואה הוא לא יהיה הפתרון היחיד ולכן לא מסגיר מי מה- yi הוא 0 ו-1. כך גם ההפך, A ישתמש בערכים כפולים כך ש B לא יוכל לוודא את ערכי ה xi.

2.2.6 ניתוח מידת האבטחה והתקשורת

2.2.6.1 ניתוח מידת האבטחה

האבטחה בחישוב המכפלה הסקלארית מבוססת על חוסר היכולת של כל אחד מהצדדים לפתור k משוואות עם יותר מ- k נעלמים. חלק מהנעלמים הם ערכים רנדומאליים שניתן להתייחס אליהם כפרטיים. בכל מקרה, אם אחד מהצדדים יודע מספיק ערכי נתונים ניתן לפתור את המשוואות כדי לגלות את כל הערכים. לכן, הסיכון במידת הגילוי של הנתונים בשיטה זו מבוסס על מספר ערכי הנתונים שאחד מהצדדים יכול לקבל ממקור חיצוני.

טבלה 1 מציגה את מספר הנעלמים ביצירת המשוואות ומראה מהי כמות הנתונים שהצד השני חייב שיהיו ברשותו כדי להגיע לחשיפה מלאה:

Table 1: Security Analysis of Protocol

	Protected values	*Number of randoms generated*	*Total number of unknowns*	*Number of equations revealed*
A	$x_1 \cdots x_n$	n	$2n$	$n+r$
B	$y_1 \cdots y_n$	r	$n+r$	n

פרוטוקול המכפלה הסקלארית משמש בכל פעם למועמד אחר מסטי-הפריטים, מה שיכול ליצור כמות גדולה של משוואות. כאשר לסטים המועמדים תכונות רבות מכל צד אין החלשות ברמת האבטחה.

ייתכן מצב בו w-itemset הוא מועמד המפוצל ל- 1 בצד אחד ול- w-1 בצד השני. למשל שני מועמדים אפשריים, הסט: A1, B1, B2, B5 והסט: A1, B2, B3, B6. אם A משתמש במשוואות שונות לכל סט מועמדים האבטחה של A1 גדלה. לעומת זאת, B יכול לשלוח את הערכים שהתקבלו בפעם הראשונה. B יכול להשתמש שוב באותה קומבינציה עבור Bi ולשלוח סכום חדש. מספר הפעמים

ש B יכול להשתמש באותו פרוטוקול מוגבל על יד י r, ואם מספר הישו יות גבוה לא סביר שניתן יהיה לחשוף את זה.

2.2.6.2 ניתוח התקשורת

עלות התקשורת תלויה במספר הסטים המועמדים וניתנת לביטוי כמכפלת עלות ה - I/O של ה-apriori algorithm. בדיקת התמיכה של כל מועמד דורשת הרצה אחת של פרוטוקול המכפלה הסקלארית.

העלות של כל הרצה, מבוססת על מספר העצמים n מחושבת:

Aשולח הודעה אחת עם n ערכים. B משיב עם n+1 ערכים. A שולח הודעה עם r ערכים , לבסוף B שולח את התוצאה, האורכת 4 סבבי תקשורת.

עלות ה- bitwise communication היא O(n) עם שיעור של שני קבועים (הנח ש r קבוע). מודגם בטבלה 2:

Table 2: Communication Cost

Rounds	*Bitwise cost*	
4	$2 * n * MaxValSz$	$O(n)$

*MaxValSz = Maximum bits to represent any input value

קיימת גם עלות ריבועית של שליחת ערכי ci,j. אולם עלות זו יכולה להפוך לקבוע על ידי הסכמה על פונקציה וגרעין שיחוללו את הערכים.

2.2.7 מסקנות

לדעתי, שיטה זו מציגה פרוטוקול יעיל לכריית מידע מאובטחת באמצעות מכפלה סקלארית המשמרת את הפרטיות של פריטים אינדיווידואלי ים בין שני אתרים. ומראה שניתן להשיג הגנה טובה על פרטיות הישויות עם עלות תקשורת השווה לנדרש לבניית מאגר מידע.

קיימת הגנה על פרטי הישויות, אך עדיין ניתן להסיק מסקנות שונות מנתוני הכרייה המתקבלים , כך שפרוטוקול זה אינו מתאים למצבים בהם נדרש אפס מידע.

הרחבת ה פרוטוקול למספר אתרים אינו טריביאלי , במיוחד אם צריך להתחשב גם באפשרות לקנוניה בין הצדדים .

הפרוטוקול מניח ערכים בוליאניים ואינו מתייחס לכמויות או לקטגוריות אפשריות של ערכי הנתונים .

2.3 *Association Rules Mining in Vertically Partitioned Databases*[29]

המאמר מציג שני אלגורית מים למציאת סטים נפוצים ולחישוב רמת הביטחון של החוקים בכריית מידע ממספר בסיסי נתונים תוך שמירה על הפרטיות, כאשר כל אתר מחזיק מספר תכונות של כל טרנזקציה, והאתרים מעוניינים לשתף פעולה ביניהם. האלגוריתמים מבוססים על ה-Apriori algorithm עבור מספר אתרים המוצג ב- [30]. הרעיון הוא, שקיים שלב מקדים לפני ביצוע האלגוריתם , שבו מוסיפים לכל בסיס נתונים טרנזקציות מזויפות כפי שמתואר ב [31]. לאחר שלב זה , כל צד יכול לשלוח לצדדים האחרים את בסיס הנתונים שנוצר כדי לחשב א ת סטי הפריטים התדירים בבסיסי הנתונים החדשים . מכיוון שהטרנזקציות המזויפות לא שינו את הנתונים האמיתיים , אלא רק הוספו עליהם , כל הפריטים בעלי התמיכה הגבוה יתגלו, וכעת נותר להיפטר מהתוספת השקרית.

באלגוריתם הראשון המיועד למציאת חוקי הקשר באופן מאובטח בין שני אתרים, צד אחד נחשב כאדון, master, והוא זה שיוזם את התהליך ומחשב את הפריטים הנתמכים גלובלית תוך שימוש בסיוע צד שלישי . הצד השני הוא עבד, slave, ותפקידו לשלוח ל אדון את בסיס הנתונים שברשותו לאחר שנוספו לו טרנזקציות מזויפות. האדון בונה בסיס נתונים רחב המכיל את כל התכונות הקיימות, בהסתמך על התכונות שלו ושל העבד . שני הצדדים שולחים לצד השלישי את מספרי הזיהוי של הטרנזקציות האמיתיות שברשותם. האדון מוצא את החוקים המועמדים תוך שימוש באלגוריתם apriori-gen ושולח

עבור כל אחד מהחוקים המועמדים את מספרי הטרנזקציות המקיימות אות ו, לצד השלישי, שתפקידו הוא לספור את מספר הטרנזקציות המקיימות חוק כלשהו ושנמצאות בשני בסיסי הנתונים , ולהשיב לאדון אם חוק כלשהו נתמך גלובלית או לא . הצד השלישי אינו מכיר את מבנה בסיס הנתונים של שני הצדדים או את פרטי החוקים שנמצאו ולכן מידת האמון הנדרשת ממנו נמוכה.

לחילופין, אפשר להשתמש בפרוטוקול מאובטח לחישוב מכפלה סקלארית , בו האתרים ייצרו וקטורים שיכילו סימון בינארי להתקיימות טרנזקציה.

האלגוריתם השני, המיועד לחישוב מאובטח בין- n צדדים, משתמש ברעיון דומה , בו יש אדון אחד ו n-1 עבדים. העבדים שולחים ל אדון את בסי ס הנתונים שלהם המכיל טרנזקציות מזויפות כדי שיבנה בסיס נתונים ר חב המאחד את כל האתרים . האדון מוצא את כל המועמדים האפשריים ומבצע פעולות של חישוב מאובטח בין האתרים השונים למציאת החוקים המקיימים את התמיכה הגלובאלית.

האלגוריתם לחישוב רמת הביטחון בין שני אתרים, משתמש בטכניקה כמעט זהה , האדון מיצר שתי קבוצות של מספרי טרנזקציות TIDx ו-TIDxy (האחת לחוקים המכילים את X והשנייה לחוקים המכילים את XY, לחוק אפשרי מהצורה $X \Rightarrow Y$) ושולח לצד השלישי המחשב האם $|\frac{TIDxy \cap STID}{TIDx \cap STID}| \geq c$ (כאשר STID הוא הסט של מספרי הטרנזקציות האמיתיות של ה עבד) ושולח לאדון תשובה מתאימה. במקרה של מספר אתרים, האדון בונה קבוצה של אתרים המכילים פריט מהחוק בבסיס הנתונים שלהם , ושולח לאתר הראשון את TIDx ו-TIDxy שמחשב את $TIDx \cap STID, TIDxy \cap STID$ ושולח לאתר הבא אחריו . האתר האחרון בקבוצה בודק אם $|\frac{TIDxy \cap STID}{TIDx \cap STID}| \geq c$ ושולח לאדון את התוצאה.

המאמר מציג ביצועים טובים יותר מאלגוריתמים אחרים שאותם הוא מנסה לשפר . המאמר מסכם את נושא האבטחה ומסביר כי מכיוון שהצדדים יודעים כי חוק מסוים נתמך גלובלית , אך אינם יודעים במדויק מהו ה תמיכה, מסייע למידת האבטחה המתקבלת במיוחד במקרה של שני צדדים , אך מסביר כי לא ניתן למנוע חשיפת מידע חלקי ולפעמים אפילו מלא כתוצאה מהסקת מסקנות , או מאופיים של הנתונים, וכן מתייחס לבעיה אם האדון מנסה לרמות כדי לחשוף את הטרנזקציות האמיתיות.

2.4 *Privacy-Preserving Decision Trees over Vertically Partitioned Data*[32]

מאמר זה (הובא בהרחבה בסמינר) מציג זווית שונה לצורך בשיתוף או בקבלת מידע מנתונים המחולקים בצורה וורטיקאלית על פני מספר אתרים, ומתיחס לסיווג שהיא אחת מהבעיות הקיימות בכריית מידע בחיי היומיום . עץ החלטת סיווג, Decision Tree Classification, הוא אחת מהגישות הטובות ביותר לשימור פרטיות המידע , ובמיוחד ID3 שהוא מהפתרונות האלגנטיים והאינטואיטיביים ביותר.

המאמר מציג אלגוריתם הבונה עץ החלטה ID3, תוך כדי שמירה על פרטיות , לנתונים המחולקים בצורה אנכית , בכל אתר נמצא חלק מהם, ואין אתר המחזיק מידע מלא של מופע כלשהו . השיטה מתאימה לכל מספר של אתרים, אך תכונות הסיווג ידועות רק לצד אחד מתוכם.

רמת האבטחה גבוהה במיוחד : לא רק שישויות אינדיבידואליות מוגנות, אלא גם סכמת הנתונים של כל אתר מוגנת מגילוי (תכונות וערכים אפשריים לתכונות). המטרה היא שכל אתר יחשוף מעט ככל האפשר, תוך כדי בנייה יעילה של העץ.

כל המידע שנחשף הוא המבנה הבסיסי של העץ (כלומר, מספר ההסתעפויות בכל צומת התואמות למספר הערכים השונים האפשריים לתכונה, והעומק של כל תת עץ) וכן איזה אתר אחראי להחלטה הנעשית בכל צומת . כדי להשתמש בעץ בצורה יעילה יש לסווג את האובייקט . בנוסף, כל צד לומד את מספר הסיווגים בחלק מהצמתים הפנימיים, אולם רק האתר המכיל את מחלקות הסיווג מכיר את המיפוי לסיווג המתאים. שאר האתרים לא יכולים אפילו לדעת אם סיווג מסוים בצומת עם 30% תפוצה הוא אותו סיווג בצומת נמוכה יותר עם 60% תפוצה אלא אם כן הוא יכול להסיק את המסקנה מהעץ ומהנתונים של עצמו . בעלים, ניתן לדעת זאת . אם ידוע מספר הטרנזקציות לכל צומת עלה ניתן לחשב תפוצה מהעץ, אך בכך לא נחשף מידע חדש.

הצמתים מחולקים בין האתרים השונים על פי התכונות המייצגו ת כל צומת . האתר הדורש החלטה שנחשב כאתר הראשי מכיר את שורש העץ . הרעיון העיקרי הוא בהעברת השליטה מאתר לאתר בהתבסס על ההחלטה הנעשית . כל צד מכיר את ערכי הטרנזקציה של הצמתים השייכים לו ולכן יכול

להעריך את מהי הצומת הבאה, אולם אינו יודע כלום על ערכי תכונות אחרות .

לסיכום, התועלת העיקרית משיטה זו היא : יצירת פרוטוקול לבניית עץ החלטה על נתונים המחולקים אנכית על פני מספר אתרים תוך שמירה על פרטיות המידע של השותפים , כאשר אתר אחד בלבד מכיר את מחלקות הסיווג. רמת האבטחה גבוהה מאד ומתבססת על יחסי כנות למחצה semihonest בין הצדדים. מרגע בניית העץ, ביצוע ההחלטות מהיר ויעיל. ניתן להרחיב בקלות את הפרוטוקול למקרה שבו כל השותפים מכירים את מחלקות הסיווג ובכך להעלות בצורה ניכרת את יעילות הפרוטוקול . ניתן בקלות להרחיב את הפרוטוקול כך שאת הסיווג של כל מופע רק האתר המסווג יכיר.

2.5 *Preserving privacy in association rule mining with bloom filters*[33]

מאמר זה הינו ייחודי בפרק זה בכך שאינו מתייחס לבסיסי נתונים שונים המצ ויים בידי ארגונים שונים באתרים שונים המעוניינים לבצע כריית מידע למציאת חוקי הקשר המשותפים לכולם , אלא מציג כלי יעיל לביצוע שאילתות מחוץ לבסיס נתונים המרכזי תוך הרשאה חלקית לביצוע כריית מידע בשרתי קצה השייכים לארגון אך מרוחקים ממנו ובכך למנוע עומס יתר על המערכת ופגיעה בביצועים.

איור 1 מדגים את סכמת השימוש בשיטה:

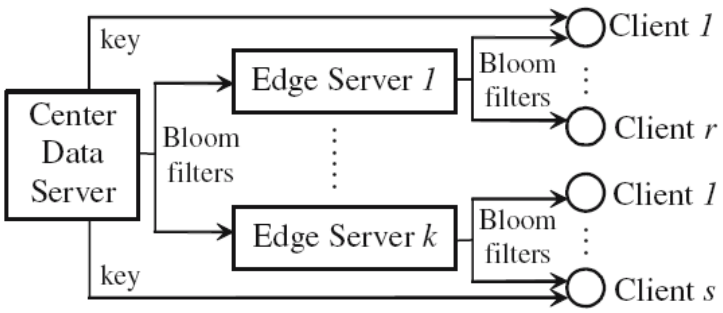


Fig. 1 The client-server-based architecture

לשמירה על הפרטיות, הנתונים המקוריים מומרים תחילה לאוסף keyed Bloom filters תוך שימוש במפתח סודי על ידי השרת המרכז את הנתונים , ונשלחים למספר שרתי קצה המבצעים כריית חוקי מידע נדרשים עבור המשתמשים . המשתמשים מקבלים אוסף של Bloom filters המייצגים סטי פריטים תדירים ומשחזרים אותם באמצעות מפתח משותף.

ההנחה היא שאין קנוניה בין שרתי הקצה והלקוחות. לכן, מבלי לדעת את המפתח הסודי ואת כל פונקציות ה hash הממפות ל-Bloom filter, שרת הקצה אינו מסוגל לפרש את תוצאות הכרייה. כמו כן, מניחים כי המטרה העיקרית היא למנוע השגת מידע רגיש מהנתונים הגלויים בזמן ביצוע כריית המידע המרוחק.

2.5.1 הגדרת Bloom filter

Bloom filter היא שיטה פשוטה , יעילה בנפח, מארגנת בסדר אקראי רשומות נתונים לייצוג סט של אובייקטים וכן גם לתמוך בשאילתות שייכות.

עקרון ה Bloom filter הוא המרת כל פריט לרצף של סיביות באורך קבוע m באמצעות k פונקציות hash רנדומאליות הידועות רק לשרת המרכז את הנתונים. סט-פריטים יורכב מאיחוד הסיביות באורך m של כל הפריטים המומרים המוכלים בו.

<u>הגדרה 2.1</u> – בהינתן סט n-אלמנטים $S=\{s_1,\ldots,s_n\}$ ו-k פונקציות hash בעלות $h_1,\ldots,h_k$ הטווח m, ה- Bloom filter של S, המסומן כ- B(S), הוא וקטור בינארי באורך m שהורכב באמצעות הצעדים הבאים: (א) כל ביט מאותחל לאפס ; (ב) כל אלמנט $s \in S$ מקוצץ לתוך וקטור הסיביות דרך ה- k פונקציות hash, והביטים התואמים $h_i(s)$ נקבעים ל-1. פונקצית Bloom filter, המסומנת B(.), היא מיפוי של סט אלמנטים ל Bloom filter שלו.

עבור שאילתות שייכות האם פריט $x \in S$, קוצצים את x ל- Bloom filter של S ובודקים האם כל ה- $h_i(x)$ הם 1. אם לא, אז x אינו שייך ל S. אם כן, אז אומרים כי x ב-S למרות שיתכן בהסתברות מסוימת כי זו טעות.

<u>הגדרה 2.2</u> – לאלמנט s וסט S, נגדיר $s \in_B S$ אם s קוצץ לכל ה 1-ים ב Bloom filter של S,

אחרת $s \notin_B S$. שיעור השגיאה החיובי של ה-Bloom filter של S מוגדר כהסתברות של $s \in_B S$ כאשר $s \notin S$, או $\Pr(s \in_B S \mid s \notin S)$. הרעיון הוא להמיר את בסיס הנתונים של הטרנזקציות לאוסף של Bloom filters כדי לשמר את הפרטיות בכריית הסטים התדירים. כל טרנזקציה $T_i \subseteq T$ מומרת ל- Bloom filter $B(T_i)$ באורך m תוך שימוש ב- k hash functions. לשימור הפרטיות של העצמים I_i מניחים שתהליך הכרייה נעשה על ה- (keyed) Bloom filter $B(I_i)$ של הפריטים מאשר על הפריטים עצמם.

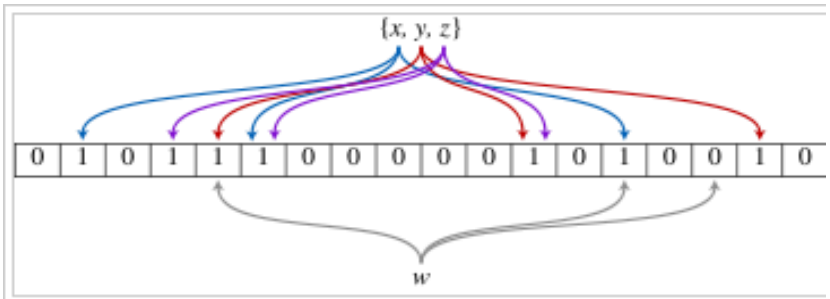
ע"פ דוגמא הבאה באיור 2, פריטים x, y, z ממופים כל אחד על ידי k פונקציות hash לרצף סיביות באורך m=18:

B(x) = 010001000000010000
B(y) = 0 0 0 00 10 00000 1 0 0 0 0 1 0
B(z) = 0 0 0 1 0 1 00000 1 0 0 0000

איחוד הסיביות של שלושת הפריטים מייצג את סט הפריטים {x, y, z} כפי שמוצג באיור. על רצף סיביות זה ניתן לבצע שאילתת שייכות (לדוגמא, האם פריט w המיוצג על ידי ה- Bloom filter, B(w), שייך לסט-הפריטים), וחשוב יותר, ניתן לבצע כריית נתונים על סטי-הפריטים הממופים לסיביות.

איור 2:

B({x,y,z})=



An example of a Bloom filter, representing the set {x, y, z}. The colored arrows show the positions in the bit array that each set element is mapped to. The element w, not in the set, is detected as a nonmember as it is mapped to a position containing a 0.

משימת כריית הנתונים יכולה להתבצע על שרת קצה כאשר בסיס הנתונים הוא מקור חיצוני. שרת הקצה מסופק עם Bloom filters לביצוע משימת כרית הנתונים. המטרה העיקרית היא למנוע גילוי נתונים רגישים ממידע ציבורי. ניתן להשיג זאת בעזרת Bloom filters המספקים בו זמנית שלושה תנאים: תחילה, טרנזקציות המכילות מספר שונה של פריטים ממופות ל Bloom filters באורך אחיד. בכך נמנעת אפשרות פענוח הרכב הטרנזקציה על ידי ניתוח אורך ה- Bloom filters על ידי גורם עוין. שנית, Bloom filters תומכים בשאילתות שייכות. הדבר מאפשר לגורמים חיצוניים מורשים (שרתי הקצה) לבצע כריית מידע רק בעזרת Bloom filters. שלישית, מבלי לדעת את כל פריטי היחיד, קשה לזהות כל פריט מה Bloom filter טרנזקציה על ידי ספירת מספר ה- 1-ים או ה 0-ים. וזאת מכיוון שההסתברות של ביט להיות 0 או 1 היא 0.5 (הסבר לכך מופיע במאמר המקורי) כאשר הפרמטרים של Bloom filter מספקים $k = \frac{m}{n} \ln 2$.

חלק מהטרנזקציות מכילות פריט אחד בלבד. המספר של 1-ים ב Bloom filters האלה קרוב מאד ל-k אך לא יותר ממנו. לכן בסיס הנתונים החיצוני יכול לחשוף פריטים אישיים באופן חלקי. כדי להמנע מחשיפה כזו ניתן להוסיף מספר פריטים וירטואליים כ- white noise לטרנזקציות אלו כך שמספר הפריטים לא יעלה על הסף.

כאשר לקוח שולח בק שת כרייה, יש לשלוח מועמדים תדירים באורך k לשרת הקצה באותו הזמן. מועמדים תדירים באורך 1 הם בדיוק פריט בודד. כדי למנוע את החשיפה שלהם לשרת קצה מומלץ לא לייצא משימות כרייה של פריטים תדירים באורך 1. אלטרנטיבה אחרת היא להסתיר את הפריטים הבודדים על ידי הרחבת השימוש במפתח הסודי כפי שצוין בחלק הקודם. תחילה מכניסים מספר פריטים וירטואליים, מסומנים $k_1, k_2, \ldots k_i$, לתוך כל הטרנזקציות ורק אז מייצאים אותן לשרת הקצה. בצד

הלקוח, כל פריט מועמד באורך 1 מצורף לפריט שנבחר רנדומאלית מתוך k_1 עד k_i. לכן שרת הקצה אינו מזהה בקלות מועמדים תדירים באורך 1 או 2 מכיוון שהם דומים. שיטה זו ניתן ליישם על כל אורך k של פריטים תדירים. בנוסף ניתן ליישם זאת לפני או אחרי הוספת הרעשים שתוארו קודם.

2.5.2 הצגת שיטת הכרייה באמצעות Bloom filters

אלגוריתם 1 מציג את המסגרת לשיטה של כריית נתונים תדירים עם Bloom filters.

Algorithm 1 Mining frequent itemsets with Bloom filters

1: $C_1 = \{B(I_1), \ldots, B(I_d)\}$ // $B(I_i)$ is the Bloom filter of item I_i
2: **for** $(\ell = 1; C_\ell \neq \varnothing; \ell ++)$ **do**
3: **for** each $B(S) \in C_\ell$ and each transaction filter $B(T_i)$ **do**
4: **if** $B(S) \sqsubseteq B(T_i)$ **then** Bsupport(S) ++ // $B(S) \sqsubseteq B(T_i)$ iff $B(S) \wedge B(T_i) = B(S)$
5: **end for**
6: **for** each $B(S) \in C_\ell$ **do**
7: $\bar{f} = 0.5^{||B(S)||}$
8: $min\_sup' = min\_sup + \mu$ // $\mu = \left(N - Bsupport(S)\right)\frac{\bar{f}}{1-\bar{f}}$, see below
9: **if** $Bsupport(S) < min\_sup'$ **then** delete $B(S)$ from C_ℓ
10: **end for**
11: $F_\ell = C_\ell$ // F_ℓ is the collection of Bloom filters of all "frequent" itemsets with length ℓ
12: $C_{\ell+1} = can\_gen(F_\ell)$ // generate filters of candidate itemsets for the next round
13: **end for**
14: Answer = $\bigcup_\ell F_\ell$ // all filters of frequent itemsets

ניתן לחלק את האלגוריתם לשלושה חלקים:

- שלב הספירה, Counting phase (שורות 3-5)
- שלב הקיצוץ, Pruning phase (שורות 6-10)
- שלב יצירת המועמדים, Candidates generating phase (שורות 11-12).

2.5.2.1 שלב הספירה (שורות 3-5)

כל מועמד מסונן נבדק מול כל הטרנזקציות המסוננות וספירת המועמדים מתעדכנת . השיטה הנאיבית היא לבדוק כל קומבינציה מסוננת של מועמדים וטרנזקציה מסוננת , אך שיטה זו כמובן אינה יעילה. כדי לשמור על יעילות נארגן את המועמדים המסוננים באמצעות Bloom filters בעץ היררכיות וכל q סיביות נחלק אותם לרמות שונות, כאשר q הוא פרמטר. לדוגמא, ברמת השורש, החלוקה מובילה ל- 2^q תתי צמתים; ה Bloom filters בכל צומת משתף את אותם q הביטים הראשונים. צומת מתפצלת אם היא מכילה יותר מ- c Bloom filters, כאשר c הוא פרמטר נוסף . בסוף החלוקה, כל צומת עלה מכילה מספר מוגבל של Bloom filters, כאשר כל צומת שאינה עלה (פרט לשורש) משויכת למקטע של q- ביטים שממנה התפצלה הצומת.

בגלל הרנדומאליות של מפתחות פונקציות הhash, הפיזור של ה Bloom filters הוא אחיד, מה שגורם לעץ להיות מאוזן היטב . לכן, עץ בעל L-level יכול לשמש כאינדקס עבור עד $c \cdot 2^{qL}$ מועמדי Bloom filters. לדוגמא, אם q=5 ו c=20, עץ 4-levels יכול לשמש כאינדקס ל- 20M Bloom filters.

2.5.2.2 שלב הקיצוץ (שורות 6-10)

בשלב ה קיצוץ, pruning, כל bloom filter מסולק מסט המועמדים אם הספירה שלו (Bsupport) פחות מ הסף המבוקש . $min\_sup'=min\_sup+\mu$, מכיוון ש $support(S)$ אינו ידוע בבסיס הנתונים, פותרים זאת על ידי החלפתו ב $Bsupport(S)-\mu$, כך- $\mu=(N-Bsupport(S))\frac{\bar{f}}{1-\bar{f}}$.

הסף החלופי מחושב לכל Bloom filter מועמד.

2.5.2.3 שלב יצירת המועמדים (שורות 11-12)

על ידי ספירה וקיצוץ בלבד , האלגוריתם יכול לגלות את כל ה Bloom filters של הסטים התדירים מכל סט של Bloom filters. זה יכול להיעשות על ידי צעד יחיד המקשר בין לקוח ושרת.
כדי לתמוך בכריית מידע אינטראקטיבי ת לגילוי כל הסטים התדירים, יש לנהל הדברות מרובת

שלבים בין הלקוח לשרת. הלקוח מספק לשרת סט של מועמדים מסוננים C_ℓ; אחרי שהשרת מחזיר את תוצאת הכרייה F_ℓ הלקוח יוצר סט נוסף $C_{\ell+1}$ של מועמדים מסוננים מתוך F_ℓ ושולח אותם לשרת לכרייה. וכן הלאה.

יצירת סט המועמדים $C_{\ell+1}=can\_gen(F_\ell)$ בצד הלקוח מנוהל בצעדים הבאים מטעמי בטח ון. תחילה ה Bloom filters שב F_ℓ מומרים חזרה לסטים. ניתן לבצע זאת על ידי מיפוי חד- חד ערכי בין כל Bloom filter שב C_ℓ והסט התואם לו, (F_ℓ הוא תת קבוצה של C_ℓ). מאוסף כל הסטים, הלקוח מיצר סט חדש של מועמדים כל ידי שימוש ב apriori_gen (אלגוריתם של Agrawal & Strikant 1994). הרעיון העקרי של ה-apriori_gen הוא שמועמדים באורך $\ell+1$ נוצרים רק אם כל כל תת הסטים באורך ℓ מופיעים באוסף של הפריטים. הלקוח יכול גם לערוך את סט המועמדים של סטי הפריטים בהתאם לדרישות היישום והאילוצים. לבסוף, הלקוח ממיר את הסטים המועמדים ל - Bloom filters ושולח אותם כ- $C_{\ell+1}$ לשרת. כל הניסויים המוצגים בהמשך מבוססים על תרחיש זה.

בהתאם לחיובים השקריים, סט פריטים S המומר מ Bloom filter B(S) יכול להיות סט תדיר אמיתי (כלומר נתמך בלא פחות מה min_sup) התואם ל- Bloom filter B(S). למרבה המזל, S אינו תת סט התואם את הסט הפריטים התדיר האמיתי. לכן, שום סט תדיר אמיתי לא אובד ביצירת המועמדים של הסיבוב הבא.

בחירה נוספת ביצור המועמדים נעשית בצד השרת. לשרת אין מפתח סודי לביצוע פונקציות ה hash, לכן הוא אינו להמיר בין ה- Bloom filter לסטי הפריטים וחזרה. לכן פתרון אפשרי הוא להשתמש ב- $C_{\ell+1}=\{B(S_1)\vee B(S_2): B(S_1),B(S_2)\in F_\ell\}$ כסט מועמד לסיבוב הבא, כאשר הסימן $\vee$ מציין OR בסיביות. ניתן בקלות לאמת ש $B(S_1)\vee B(S_2)$ הוא Bloom filter של סטי הפריטים $S_1\cup S_2$. לכן, פתרון זה מייצר את כל ה- Bloom filters של הסטים שהם איחוד של כל שני סטי פריטים תדירים (ברור שלא קיים פריט תדיר שאובד בת הליך). החיסרון של פתרון זה הוא שהשרת אינו יכול למצוא Apriori ברמה של הסטים.

2.5.3 תרומת המאמר:

המאמר מציע שימוש ב- keyed Bloom filters לייצוג מידע, ביצוע כריית חוקי הקשר בעזרת אלגוריתם שהותאם ל- Bloom filters ובתוך כך גם לשמר את פרטיות המידע.

המאמר מציג -

- גישה חדשה לשימור הפרטיות בכריית חוקי מידע והצגת ניתוח תיאורטי על יעילותה.
- הגדרת מבנה לכריית חוקי מידע בעזרת Bloom filters בעל סף משתנה.
- שיטה מעשית וגמישה בפועל והוכחה כעובדת היטב.
- טכניקת δ-folding ליעילות רבה באחסון מבלי לפגוע ברמת דיוק הכרייה ובביצועים.

ניתן לסכם כי Bloom filter[34] הוא כלי תכנותי יעיל המבוסס על סכמת hash הסתברותית היכולה לייצג קבוצה של עצמים באמצעות דרישת זיכרון מינימאלית. ניתן להשתמש בו כדי לבצע שאילתות שייכות עם zero false negatives ו-low false positives. Bloom filters משמשים במגוון רחב של אפליקציות במערכות מרושתות ובבסיסי נתונים.

בהשוואה לשיטות לבלבול הנתונים, השיטה המוצגת במאמר יכולה להשיג דיוק גבוה בתוצאת כריית המידע תוך שמירה מלאה על פרטיות רשומות נתונים אינדיבידואליו ת וזאת מבלי להזדקק לעלויות גבוהות של הצפנה. בנוסף, השיטה מספקת גמישות בדרישות האחסון תוך התייחסו ת לדיוק הכרייה הנדרש.

2.6 סיכום

פרק זה הציג מספר מאמרים הבוחנים צרכים שונים ל שימור פרטיות המידע בכריית מידע מבסיסי נתונים מבוזרים, אם בארגון שונה של הנתונים בכל אתר - כרייה אופקית או אנכית, אם למטרת כרייה שונה כגון קבלת החלטה מתוצאת הכרייה (מאמר 2.4), או אם כריית מידע באתרים שונים מבסיס נתונים מרכזי (מאמר 2.5). כל מאמר עוסק בצורך אחר ומשתמש בטכניקה שונה המקשה להשוות בין השיטות. כמו כן, מידת האבטחה המתקבלת ודרישות האבטחה שונות בכל שיטה.

מאמר 2.4 ייחודי מסוגו בפרק , ועוסק ב עץ קבלת החלטות מסווג בעל רמת אבטחה גבוהה ביותר, בו גם הצדדים השותפים אינם חלק מתהליך קבלת ההחלטות ונחשפים למידע מועט במיוחד . בשאר השיטות מקבלים הצדדים את תוצאות הכרייה המלאות , כאשר הם יכולים לנסות לנחש מהו אחוז התמיכה של כל חוק באתרים השונים.

המאמרים היחידים שניתן להשוות ב יניהם הם המאמרים המתייחסים לכרייה אנכית של נתונים (2.2, 2.3) כאשר מאמר 2.2 מציג אלגוריתם המוגבל לשני צדדים בלבד ואינו ניתן להרחבה סבירה עבור יותר משני צדדים , כאשר קיים חיסרון מהותי בכמות המידע שהצדדים יכולים לנחש אחד על השני . האלגוריתם המ וצע נראה מסורבל מעט , וכבר קיימות שיטות חדשניות ויעילות יותר לביצוע החישוב המאובטח בין האתרים. לעומתו, מאמר 2.3 מציג שיטה שונה לחלוטין הניתנת להרחבה גם למספר גדול יותר של אתרים ובכך למעשה פותר את המגבלה העיקרית של האלגוריתם הקודם . כמובן שבעיית כמות המידע ששני צדדים יכולים לנחש אחד על השני אינה תלויה בסוג האלגוריתם והיא נשארת בעיה מהותית ברמת האבטחה, כך שרק שני אתרים המוכנים להתפשר בצורה ניכרת על כמות המידע שהם מוכנים לחשוף יכולים לבצע כריית מידע משותפת שכזו.

Hiding Sensitive Rules

3.1 הקדמה

תהליך כריית המידע מבסיסי נתונים רחבים התגלה כיעיל מאד לחברות עסקיות וניתן להפיק ממנו תועלת בהרחבת העסק, שיפור רווחיות, הקטנת עלויות ותמיכה בשווקים ביעילות רבה יותר.

חברות מתחרות נאלצות לשתף פעולה ב יניהן כדי לשפר את יעילותן העסקית ולהציע ללקוחותיהן מגוון מוצרים . שיתוף פעולה זה יכול לספק למתחרים מידע עסקי או ארגוני שישמש את המתחרים כנגדן . לכן, החברות יכולות להחליט לחשוף רק חלק מסוים מהמידע שברשותן , ולהסתיר תבניות אסטרטגיות , הנקראות חוקים רגישים , החייבים להיות מוגנים בפני תהליך השיתוף . חלק מהחברות מעדיפות לשתף נתונים, בעוד אחרות מעדיפות לשתף תבניות שנוצרו מאותם נתונים.

הבעיה העיקרית שיש לפתור היא : *כיצד חברות יכולות להעביר את הנתונים שברשותן על מנת לתמוך בשיתוף פעולה עסקי מבלי לחשוף מידע רגיש ומבלי לאבד את התועלתיות של תהליך הכרייה ?* בהמשך הפרק מוצגות מספר טכניקות המספקות מענה לשאלה זו.

הפתרון הפשוט הוא ליישם פילטר לאחר תהליך הכרייה ולסלק או להסתיר את החוקים הרגישים שהתגלו. הבעיה בשיטה זו היא שסילוק חוקים מסוימים אינה מבטיחה הגנה מושלמת. יש לדאוג לסלק סט של חוקים רגישים כך שלא יישארו עקבות שניתן יהיה לעקוב אחריהם על ידי מתקיפים.

שיטה יעילה יותר להסתרת חוקים רגישים היא על ידי המרת בסיס הנתונים של הטרנזקציות לבסיס נתונים חדש המסתיר את החוקים הרגישים, אך מציג את מרבית החוקים שאינם רגישים. במקרים מסוימים מספר פריטים נמחקים מקבוצת החוקים הרגישים במטרה להסתיר את החוקים הנגזרים מפריטים אלו, ובכך התמיכה בחוקים רגישים אלו יורד מתחת לסף הגילוי שנקבע ψ. שיטה נוספת היא דווקא להוסיף פריטים לטרנזקציות מסוימות במטרה להפחית את רמת הביטחון של חוקים רגישים. לדוגמא אם קיים חוק רגיש $X \rightarrow Y$, אם מוסיפים פריטים לחלק הראשון X בטרנזקציות התומכות ב X אך לא תומכות ב Y, הרמת הביטחון של חוק כזה יורד. בכל מקרה יש לשנות חלק מהנתונים בתהליך הסינון.

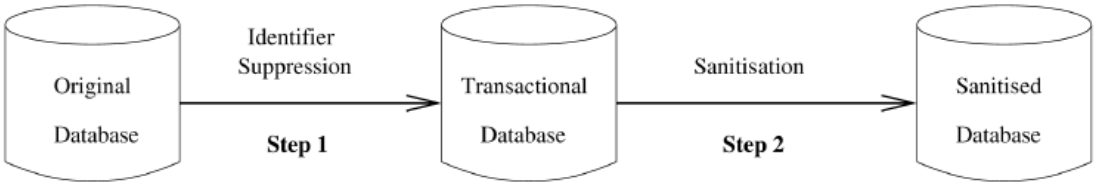
3.1.1 הצגת עקרונות הסתרת נתונים רגישים

תהליך ההגנה על חוקים רגישים נחלק לשני שלבים עיקריים כמתואר באיור 1, כאשר השלב הראשון Identifier Suppression נועד להרחיק את המאפיינים המזהים מבסיס הנתונים (ת.ז., שמות וכד') שנועד לשיתוף, ובכך לאפשר חשיפה של התנהגויות של לקוחות מבלי לחשוף את זהותם . לאחר שלב זה, הנתונים שנותרו יאוחסנו בטבלה בודדת הנקראת בסיס נתונים של טרנזקציות, Transactional Database, שאינה מכילה פרטים אישיים על לקוחות, אלא רק פעילויות עסקיות שאותן בצעו . שיטה זו יעילה מאד ומבטיחה פרטיות מלאה . במקרים רבים, קשה מאד להרחיב על ישות מסוימת או יותר מתוך בסיס הנתונים של הטרנזקציות גם בשילוב טרנזקציות נוספות עם נתונים אחרים . אולם, טרנזקציה מסוימת יכולה להכיל מספר פריטים שניתן לקשר עם בסיסי נתונים אחרים ולהגיע לזיהוי ישות מסוימת.

השלב השני נועד להסתיר בצורה יעילה מידע המיוצג על ידי החוקים הרגישים . במרבית המקרים, ההבחנה מהו המידע הרגיש אינה נצפית מראש, וזו הסיבה שתהליך הזיהוי של מידע רגיש דורש התערבות אנושית בשלב ביני ים לפני שיתוף הנתונים לצורך הכרייה . בהקשר זה, מידע רגיש מיוצג על ידי קבוצה מיוחדת של חוקים הקשורים לחוקי ההקשר הרגישים.

הנתונים מסוננים לתוך בסיס נתונים חדש המכיל את הנתונים המופצים . או כנתונים שיש לכרות מהם מידע, או כאוסף חוקי הקשר שיש לסננן (DSA).

Figure 1 Major steps of the process of protecting sensitive knowledge



3.1.2 בעיות בתהליכי סינון נתונים רגישים

תהליך הסינון נועד להסתיר חוקים רגישים בלבד , אך לתהליך מספר תופעות לוואי שיש להתחשב בהשפעה שלהן על תוצאות הכרייה המתקבלות כאשר מיישמים את התהליך.
שתי התופעות הראשונות שיש לשים אליהן לב הן, הצורך לדאוג מפני הסתרת יתר של מידע, כך שגם מידע שאינו רגיש יוסתר כתוצאה מתהליך הסינון , והבעיה השנייה , היא יצירת חוקי הקשר מלאכותיים כתוצאה משינוי והסתרה של הנתונים . שתי תופעות אלו אמנם אינן חושפות מידע רגיש אך פוגעות באמינות המידע המתקבל ובשימושיות שלו כתוצאה מכך . ברור כי המטרה עיקר ית של שיתוף הנתונים והפעלת אלגוריתמי הכרייה הוא לקבל מידע רלוונטי ואמיתי עד כמה שניתן.
תופעת לוואי שלישית, היא היכולת להסתיר בצורה מלאה את החוקים הרגישים.

בנוסף לתופעות הלוואי שפורטו לעיל , קיימת בעיה מהותית של יכולת הסקת תבניות מידע רגישות מהמידע המופץ הנקראת ערוצי-הסקה, inference channels, המתוארת בפסקה הבאה.

3.1.2.1 Inference Channels

Foreword-Inference Attack[35] הנקרא גם Foreword-Inference Channel[36] הוא היכולת להסיק מסקנות לגבי המידע הרגיש מתוך הנתונים המסוננים הגלויים . על פי המהות, רק המניעה של תבניות רגישות מקבוצת התבניות התדירות אינה בטוחה מספיק כתוצאה מערוצי-ההסקה.
קיימים שלושה ערוצי הסקה[37]:

- the superset inference channel פירושו שאם תבנית מוסתרת , אך סט העל שלה מוכל בתבניות התדירות הגלויות , אזי על פי עקרון ה anti-monotone principle, ניתן לגלות את התבניות המוסתרות.
- the subset inference channel פירושו שאם תבנית מוסתרת אך תתי התבניות שלה מוכלות בתבניות התדירות הגלויות , אזי על פי עקרון ה inclusion-exclusion principle, ניתן לחשב את התמיכה של התבנית הנסתרת תוך שימוש בתמיכה של תת התבניות שלה, ובכך לגרום לתבניות הרגישות להיחשף.
- the chain inference channel הוא indirect channel. כלומר, בהתאם לשיטת הסקת נתונים זו התוקפים תחילה מסיקים תבניות אחרות , ואז מסיקים את התבניות הרגישות מהתבניות שנחשפו תוך שימוש בשני העקרונות ב inclusion-exclusion וב anti-monotone.

במציאות, ערוצי הסקת מסקנות אלו אינם רלוונטיים בגלל the supports of the frequent patterns must be affected לאחר ביצוע הסינון.

בשונה מערוצי-ההסקה, ניתן להחשיב את התקפות-ההסקה-קדימה, Foreword-Inference Attacks, כקומבינציה של ערוצי ה subset inference וה chain inference.
נגדיר את התקפות-ההסקה-קדימה: יהי e אחד מהתבניות הרגישות באורך k בבסיס הנתונים המקורי D. נניח שתחת אותה תמיכה מינימאלית s, e שונה על ידי שיטת סינון ספציפית להסתרת כל התבניות הרגישות ואינו תדיר ב D'. בהתאם ל apriori property[38], תבנית באורך k תהיה תבנית מועמדת אם כל תתי התבניות של ה באורך k-1 הן תדירות. אם מתקיפים מאמינים שקיבלו בסיס נתונים ששונה , בהתאם ל apriori property, הם יכולים להסיק ש e הוחבא בקפדנות על ידי מנהל בסיס הנתונים.
הגיוני להניח שמתקיפים ינסו כל שיטת הסקה כדי לקבל יותר תבניות ממה שרוצים שידעו. לכן, ניתן להתייחס ל התקפות-ההסקה-קדימה כפרמטר שיש להתייחס אליו כאשר באים להעריך את יכולתו של תהליך הסתר ת נתונים רגישים . כדי למנוע התקפות-ההסקה-קדימה בבסיס נתונים מסונן , יש לדאוג לכך שלפחות תת- תבנית אחת באורך של שתיים תוסתר או שהתבנית הרגישה תתגלה על ידי recursive inference.
לדוגמא באיור 1: נניח ש {1,2,3,4} היא תבנית רגישה המוסתרת בבסיס הנתונים המסונן, אך כל תת י התבניות שלה פרט ל {1,2,3} נכרו מבסיס הנתונים המסונן . המתקיפים יכולים להסיק את {1,2,3} מ {1,2}, {1,3}, {2,3} ואז, לאחר קבלת {1,2,3} הם יכולים להסיק רקורסיבית את {1,2,3,4} מ- {1,2,4}, {1,3,4} ו- {2,3,4} ובכך התבנית {1,2,3,4} הופכת לתבנית לא בטוחה.

התוקפים יכולים להסיק בדרך זו גם תבניות שאינן תדירות , אך אם התבנית רגישה והצליחו להסיק אותה, קיימת בעיה.

כדי לה ימנע מ התקפות-ההסקה-קדימה, לכל תבנית רגישה יש להסתיר לפחות אחת מתת התבניות שלה באורך 2 ולא תבנית באורך 1, מכיוון שהתבניות התדירות באורך של אחד (כלומר הפריטים התדירים) יכולים לתאר רק אילו פריטים תדירים קיימים בבסיסי הנתונים אבל אינם מכילים מידע על הקשרים בין הפריטים.

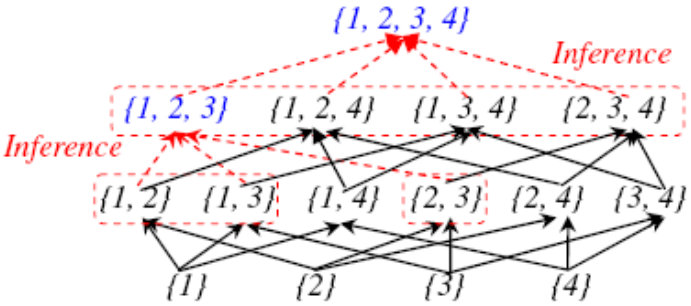


Fig. 1. An example of the Forward-Inference Attacks.

נניח שהתבנית {1,2,3,4} היא תבנית רגישה והטרנזקציות המכילות בו זמנית {1,2,3,4} נבחרו להיות טרנזקציות רגישות . קיימות מספר גישות המתייחסות לשינוי הטרנזקציות הרגישות בלבד , אך לא משנה איזה פריט נבחר להיות victim item, הפריט הקורבן, ומסולק מהטרנזקציות הרגישות , עדיין {1,2,3}, {1,2,4}, {1,3,4} ו- {2,3,4} יכולים להיות תדירים מכיוון שהם נתמכים לא רק על ידי הטרנזקציות הרגישות המכילות אותם, אלא גם על ידי טרנזקציות אחרות המכילות רק אותם ופריטים אחרים. כך, ששיטות אלו נשארות חשופות להתקפות-ההסקה-קדימה.

השיטות במאמר המוצג בפסקה 3.2 חשופות בחלקן למתקפת התקפות-ההסקה-קדימה, אך המאמר המוצג בפסקה 3.3 מכיל שיטה המרחיבה את השיטה המוצגת במאמר[39] מיישם את קשרי הגומלין בין התבניות הרגישות והתבניות שאינן רגישות באמצעות מטריצת סינון. הכפלת בסיס הנתונים במטריצת הסינון יוצר בסיס נתונים מסונן שעמיד בפני התקפות-ההסקה-קדימה.

3.2 A unified framework for protecting sensitive association rules in business collaboration[44]

מאמר זה מתמקד בצורך לתת מענה לצורך של חברות מסחריות לשיתוף פעולה עסקי על ידי כריית חוקי הקשר מבלי לחשוף מידע רגיש כגון זיהוי הלקוחות. המאמר מתייחס למבנה שרת (מחשב מארח המכיל תוכנה מומלצת עבור ה- e-commerce) ומספר לקוחות (חברות עסקיות), שלכל אחד מהם סט של פריטים שנמכרו. הלקוחות מעוניינים שהשרת יאסוף מידע סטטיסטי על הקשרים בין הפריטים מבלי לחשוף מידע על הלקוחות וכן שהשרת לא יוכל לחשוף חוקי הקשר רגישים.

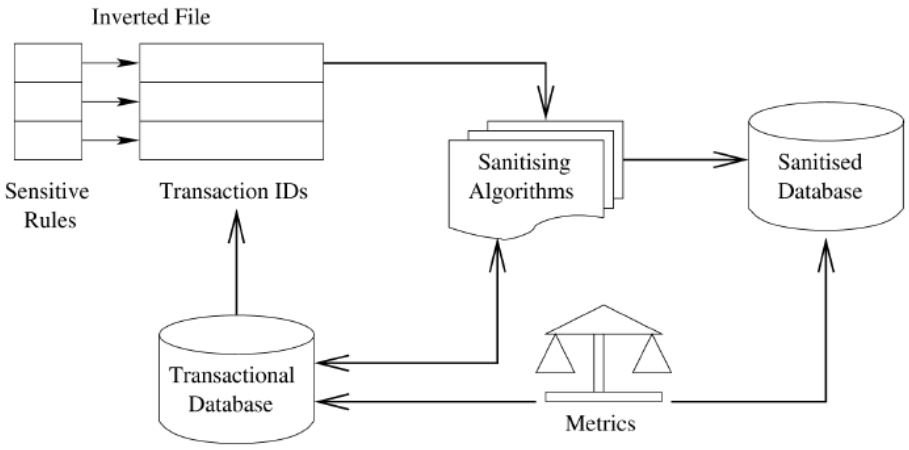
המאמר משלב מספר שיטות להסתיר ביעילות תבניות רגישות: קבוצה של אלגוריתמים להגנה על המידע הרגיש ואחזור יכולות להאיץ את תהליך ההגנה, וקבוצה של מטריצות להערכת האפקטיביות של האלגוריתמים המוצעים במונחים של אבדן מידע, והערכת כמויות המידע הפרטי שנחשף.

האלגוריתמים דורשים שתי סריקות המתייחסות לגודל בסיס הנתונים ולמספר החוקים שיש להגן עליהם. הראשונה מיועדת לבניית אינדקס כדי להאיץ את תהליך הסינון, והשנייה משמשת להסרת החוקים הרגישים מבסיס הנתונים הגלוי.

3.2.1 הצגת מבנה השיטה

איור 2 המוצג להלן מציג את השיטה. Inverted File, קובץ מהופך, משמש להאצת תהליך הסינון, ספרייה של אלגוריתמי סינון המשמשים להסתרת חוקי הקשר רגישים מבסיס הנתונים וקבוצה של מטריצות להערכת האינפורמציה שנחשפה ולהערכת מידת ההשפעה של אלגוריתמי הסינון על בסיס הנתונים המסונן ועל תוצאות הכרייה.

Figure 2 The framework to protect sensitive knowledge in association rule mining



3.2.1.1 Inverted file

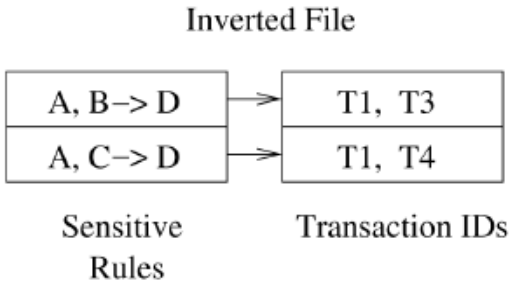
הסריקה של הטרנזקציות של בסיס הנתונים נעשית פעם אחת בלבד, ובמהלכה בונים את היכולת לשלוף מידע (Inverted file). אוצר המילים של ה- Inverted file מורכב מכל חוקי ההקשר שיש להסתיר, ולכל חוק רגיש קיימת רשימה תואמת של מספרי זיהוי של טרנזקציות שבהן החוק מוכל.

איור 3(b) מציג דוגמא של Inverted file המתייחס לבסיס הנתונים המוצג באיור 3(a). נניח כי חוקי ההקשר הרגישים הם: $A, B \rightarrow D$ ו- $A, C \rightarrow D$

Figure 3 (a) A sample transactional database and (b) the corresponding inverted file

TID	Items
T1	A B C D
T2	A B C
T3	A B D
T4	A C D
T5	A B C
T6	B D

(a)



(b)

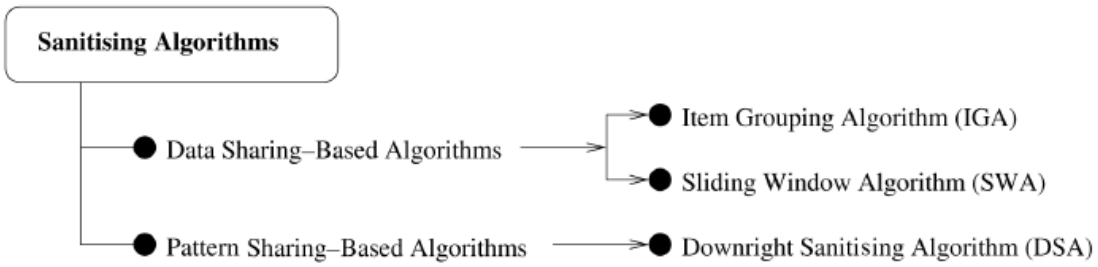
היתרונות בידיעת הטרנזקציות המכילות את החוקים הרגישים הן זמני CPU קצרים בתהליך הסינון. המידע הנשמר בזכרון הראשי הוא מינימאלי ומכיל רק את החוקים הרגישים , וכן יש לבצע רק סריקה אחת נוספת כדי לסנן את הנתונים.

3.2.1.2 ספריית אלגוריתמי הסינון

תהליך הסינון משנה מספר טרנזקציות כדי להסתיר חוקים רגישים כאשר סף רמת החשיפה נשלט בידי ה בעלים. סף זה משפיע בעקיפין על האיזון בין חשיפת המידע והגנה על המידע . וקובע את אחוז הטרנזקציות המסוננות.

האלגוריתמים מסווגים לשתי קבוצות עיקריות : data sharing-based algorithms ו- patterns sharing-based algorithms כפי שמוצג באיור 4.

Figure 4 A taxonomy of sanitising algorithms



תהליך הסינון מסי ר או מסתיר את חוקי ההקשר הרגישים . ובנוסף מספר קטן של טרנזקציות המשתתפות ביצירת החוקים הרגישים עוברות שינוי על ידי מחיקת אחד או יותר פריטים מתוכן . בכך האלגוריתם מסתיר חוקים רגישים על ידי הקטנת התמיכה או רמת הביטחון שלהם מתחת לסף פרטי.

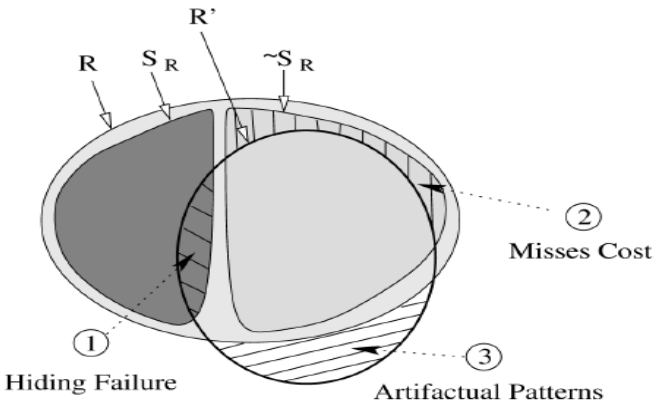
3.2.1.3 קבוצת המטריצות

קבוצת המטריצות משמשת לא רק למדידת כמות המידע רגיש נחשף , אלא גם למדידת השפעת האלגוריתם המוצע במונחים של איבוד מידע ובמונחים של חוקים לא רגישים המוסתרים כתוצאת לוואי של תהליך הטרנספורמציה. קבוצת המטריצות מתחלקת לשתי קבוצות עיקריות : Data sharing – based metrics, ו- Pattern sharing – based metrics.

3.2.1.3.1 Data sharing – based metrics

קבוצת מטריצות אלו קשורה לבעיות המוצגות באיור 5, המראה את היחס בין הקבוצה R של כל חוקי ההקשר בבסיס הנתונים D, החוקים הרגישים S_R, חוקי ההקשר הלא רגישים $\sim S_R$, וכן הקבוצה R' של כ ל החוקים המתגלים מבסיס הנתונים המסונן D'. המספרים 1, 2, 3, מיצגים את הבעיות הפוטנציאליות המתרחשות בתהליך ההסתרה.

Figure 5 Data sharing–based sanitisation problems



בעיה 1 - Hiding Failure (HF) מתרחשת כאשר מספר חוקי הקשר רגישים נחשפים כתוצאת לוואי של תהליך הסינון . הבעיה נמדדת במונחים של אחוזי החוקים הרג ישים שנחשפים מ 'D. כמובן שבאופן אידיאלי אחוז ה- HF צריך להיות 0%.

ה-HF נמדד באופן הבא: $HF = \frac{\#S_R(D')}{\#S_R(D)}$ (# מציין מספר)

בעיה 2 – Miss Cost (MC) מתרחשת כאשר חוקי הקשר לגיטימיים מוסתרים כתוצאת לוואי של תהליך הסינון. הבעיה מתרחשת כאשר מספר חוקי הקשר לא רגישים מאבדים את ה תמיכה כתוצאה מתהליך הסינון. הבעיה נמדדת באחוזים של חוקי ההקשר הלגיטימיים שלא התגלו ב 'D. באופן אידיאלי MC צריך לעמוד על 0%.

ה- MC נמדד באופן הבא: $MC = \frac{\#\sim S_R(D) - \#\sim S_R(D')}{\#\sim S_R(D)}$

ככל שיש יותר חוקים להסתיר כך בעיית ה MC גדלה, ולכן על מנת להשיג פשרה בין ה- MC לבין ה- HF יש להגדיר סף כלשהו ψ המאפשר לכוונן ולהשיג את האיזון הרצוי , אם היישום מאפשר זאת.

בעיה 3 – Artifactual Patterns (AP) מתרחשת כאשר נוצרים מספר חוקי הקשר מלאכותיים מ 'D כתוצאת לוואי של תהליך הסינון . הבעייה נמדדת במונחים של אחוזי החו קים שאינם מצויים בבסיס הנתונים המקורי . חוקים כאלו נוצרים אשר פריטים חדשים מוספים לחלק מהטרנזקציות כדי לשנות את מידת האבטחה של חוקים רגישים.

ה- AP נמדד באופן הבא: $AP = \frac{|R'| - |R \cap R'|}{|R'|}$, כאשר |X| הוא עצמת הקבוצה.

מידת השונות נמדדת על ידי השונות בין בסיס ה נתונים המקורי והמסונן באופן הבא :

$Dif(D,D') = \frac{1}{\sum_{i=1}^{n} f_D(i)} \times \sum_{i=1}^{n} [f_D(i) - f_{D'}(i)]$, כאשר $f_X(i)$ מייצג את התדירות של הפריט ה- i בבסיס הנתונים X, ו-n הוא מספר הפריטים המובחנים בבסיס הנתונים המקורי.

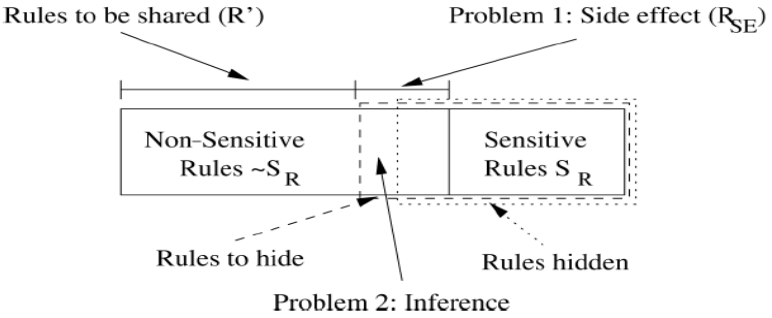
Pattern sharing – based metrics 3.2.1.3.2

מטריצות אלו מיוחסות לבעיות המוצגות באיור 6.

בעיה 1 – Side Effect מתייחסת לחוקים הלא רגישים S_R~, שהוסרו כתוצאת לוואי של תהליך הסינון R_{SE}, והיא קשורה לבעיית ה- miss cost שלעיל.

בעיה 2 – recovery factor מתרחשת כאשר משתמשים במספר חוקים לא רגישים. המתקיפים יכולים לשחזר חלק מהחוקים הרגישים על ידי הסקת מסקנות.

Figure 6 Pattern sharing – based sanitisation problems



Side-Effect Factor (SEF) מודד את מספר החוקים הלא רגישים שהוסרו כתוצאת לוואי של תהליך הסינון והוא מחושב באופן הבא : $SEF = \frac{(|R| - (|R'| + |S_R|))}{(|R| - |S_R|)}$, כאשר R היא קבוצת החוקים שנכרו מבסיס הנתונים, R' קבוצת החוקים המסוננים ו- S_R קבוצת החוקים הרגישים, בהתאמה. |S| הוא גודל הקבוצה S.

השאיפה היא שה- SEF יהיה קטן ככל האפשר . המקרה האידאל י הוא כמובן המצב בו |R|=|R'|+| S_R|, ואז SEF=0. במקרה הגרוע ביותר מתקיים |R'|=0, ואז SEF=1.

Recovery Factor (RF) מבטא את האפשרות של המתקיפים לשחזר חוקים רגישים בהתבסס על הלא-רגישים. ה- RF של תבנית אחת לוקח בחשבון את הקיום של תת הקבוצות שלו . ההיגיון הוא שכל תת קבוצה לא ריקה של פריטים תדירים חייב להיות תדיר.

בגישה של Pattern sharing – based metrics, סט החוקים המסונן המשותף R' מוגדר כ- $R'=R-(S_R+R_{SE})$, כאשר R היא קב וצת החוקים שנכרו מבסיס הנתונים המקורי , S_R הוא קבוצת החוקים הרגישים ו- R_{SE} הוא קבוצת החוקים שהוסרו כתוצאת לוואי של תהליך הסינון.

3.2.2 יוריסטיקות להגנה על חוקי הקשר רגישים

בסעיף זה מוצגות 3 יוריסטיקות להסתרת חוקים רגישים ב בסיס הנתונים של הטרנזקציות . השתים הראשונות פועלות על הנתונים להגן או להסתיר קבוצה של חוקי הקשר רגישים . לאחר הסינון, בסיס הנתונים ניתן לכרייה משותפת. ומתייחסים לשיטות אלו כאל Data sharing – based heuristic. היוריסטיקה השלישית נקראת Pattern sharing – based heuristic , ובה תהליך הסניטציה פועל על החוקים שנכרו מבסיס הנתונים . במקרה זה, הבעלים כורים את הנתונים של עצמם ומשתפים תבניות גלויות, ואז כמובן, הסינון מסיר רק את החוקים הרגישים עצמם וכן נחסמים חוקים אחרים שניתן היה להשתמש בהם כדי לנסות ולגלות חוקים רגישים נסתרים.

3.2.2.1 Heuristic 1: sanitization based on the degree of sensitive transaction

היוריסטיקה מבוססת על העובדה שבהרבה מקרים טרנזקציות רגישות משתתפות ביצירת חוקי הקשר רגישים שמוסתרים . נתייחס למספר חוקי ההקשר שנתמכים על ידי טרנזקציות רגישות כדרגת הטרנזקציה הרגישה, המוגדרת להלן:

<u>הגדרה 1</u> - (Degree of a sensitive transaction): יהי D בסיס הנתונים של הטרנזקציות ו S_T סט כל החוקים הרגישים ב- D. דרגת הטרנזקציה הרגישה t, מסומנת כ degree(t), כך ש- $t \in S_T$, מוגדר כמספר חוקי ההקשר הרגישים שניתן למצוא ב- t.

ליוריסטיקה 4 שלבים עיקריים:

- שלב 1: סריקת בסיס הנתונים ומציאת הטרנזקציות הרגישות לכל חוק הקשר . בסיום שלב זה ה-inverted file בנוי.
- שלב 2: מבוסס על סף החשיפה ψ, מחשב לכל חוק הקשר רגיש את מספר הטרנזקציות הרגישות שצריכות להיות מסוננות ומסמן אותן . הטרנזקציות הרגישות נבחרות בהתבסס על הרמה שלהן (בסדר יורד).
- שלב 3: לכל חוק הקשר רגיש יש לזהות פריט מועמד שיש לסלק מהטרנזקציות הרגישות. מועמד זה נקרא פריט קורבן.
- שלב 4: יש לבצע סריקה נוספת של בסיס הנתונים , לזהות את הטרנזקציות הרגישות המסומנות שיש לסנן אותן ולהסיר את הפריט הקורבן מהן.

להדגמת פעולת היוריסטיקה נתבונן בבסיס הנתונים באיור 7(a). נניח שקיים סט של חוקי הקשר רגישים S_R={A,B→D; A,C→D}, מתרחשות התוצאות הבאות:

Figure 7 (a) A copy of the sample transactional database in Figure 3(a) and (b) the sanitised database using Heuristic 1

TID	Items
T1	A B C D
T2	A B C
T3	A B D
T4	A C D
T5	A B C
T6	B D

(a)

TID	Items
T1	A B C
T2	A B C
T3	A B D
T4	A C D
T5	A B C
T6	B D

(b)

- שלב 1: תחילה נבצע סריקה של בסיס הנתונים לזיהוי טרנזקציות רגשיות . בדוגמא זו, הטרנזקציות הרגישות S_R המכילות את חוקי ההקשר הרגישים הן: {T1,T3, T4}. הדרגה של הטרנזקציות T1, T3 ו T4 היא 2, 1, ו-1 בהתאמה. כך שהחוק A,B→D יכול להכרות מהטרנזקציות T1 ו- T3, והחוק A,C→D יכול להכרות מהטרנזקציות T1 ו- T4.
- שלב 2: נניח שסף החשיפה ψ (disclosure threshold) נקבע ל- 50%. נסרוק את הטרנזקציות הרגישות בדרגה בסדר יורד . כתוצאה מכך, נסנן מחצית מהטרנזקציות הרגישות לכל חוק הקשר רגיש. במקרה כזה, רק הטרנזקציה T1 תסונן.
- שלב 3: יש לבחור את הקורבנות. מקבצים את החוקים הרגישים החולקים פריט משותף. שני החוקים חולקים את פריט A ו- D. אך רק פריט אחד יבחר , נניח D. על ידי סילוק D מ T1, החוקים הרגישים יוסתרו מ T1 בשלב אחד, וסף החשיפה יסופק.
- שלב 4: בצוע הסינון על ידי לקיחה בחשבון שה קורבנות נבחרו בשלב הקודם, ובסיס הנתונים המסונן נראה באיור 7(b).

3.2.2.2 Heuristic 2: sanitization based on the size of sensitive transactions

הרעיון ביוריסטיקה זו הוא לסנן את הטרנזקציה הרגישה הקצרה ביותר . ההיגיון הוא שעל ידי סילוק פריטים מהטרנזקציה הקצרה ביותר , נוכל להקטין את ההשפעה על בסיס הנתונים המסונן מכיוון שלטרנזקציות הקצרות ביותר יש את מספר הקומבינציות הקטן ביותר של חוק הקשר . בכך תקטן תוצאת הלוואי של תהליך הסינון על החוקים הלא-רגישים.

- שלב 1: Distinguishing the sensitive transactions from the none-sensitive ones. לכל טרנזקציה שנקראה מבסיס הנתונים D, מזהים אם הטרנזקציה מעורבת ביצירת חוקי הקשר רגישים . אם כן, הטרנזקציה מועתקת ישיר ות לבסיס הנתונים המסונן 'D. אחרת הטרנזקציה הזו רגישה ויש לסנן אותה.
- שלב 2: Selecting the Victim Item. תחילה יש לחשב את התדירויות של כל הפריטים שבחוקי הקשר הרגישים הנמצאים בטרנזקציה הרגישה הנוכחית . הפריט עם התדירות הגבוהה ביו תר יבחר להיות הפריט הקורבן מכיוון שהוא משותף לקבוצה של חוקים רגישים . אם לחוק הקשר רגיש אין פריט משותף עם חוקים אחרים, התדירויות של הפריטים שלו הם אחידים (freq=1). במקרה כזה, יבחר הפריט הקורבן באופן אקראי . ההיגיון בכך הוא שסילוק פריטים שונים מחוקי ההקשר הרגישים יקטין רק במקצת את התמיכה בחוקי הקשר לגיטימיים שיוכלו להכרות מבסיס הנתונים המסונן 'D.
- שלב 3: Computing the number of sensitive transactions to be sanitized. בהינתן סף החשיפ ה ψ, שנקבע על ידי בעלי בסיס הנתונים , מחשבים את מספר הטרנזקציות שיש לסנן. לכל חוק רגיש תהיה רשימה של מספרי ז יהוי של טרנזקציות הקשורות אליו . בשלב זה הטרנזקציות הרגישות ממוינות לפי החוקים הרגישים

שחושבו קודם. הטרנזקציות הרגישות מסוננות בסדר יורד של גודל , כך שמסננים תחילה את הטרנזקציות הקצרות ביותר.

- שלב 4: Sanitizing a sensitive transaction. לכל חוק הקשר רגיש יש כעת רשימה של מספרי זיהוי של טרנזקציות רגישות עם הפריט הקורבן הנבחר שלהן. בכל פעם שמסלקים פריט קורבן מהטרנזקציה הרגישה מבצעים פרוצדורת look-ahead כדי לבדוק האם טרנזקציה זו נבחרה כטרנזקציה רגישה עבור ח וקי הקשר רגישים אחרים. אם כן , הפריט הקורבן שהסרנו ממנה הוא גם חלק מאותם חוקי הקשר רגישים אחרים, מסירים את אותה טרנזקציה מרשימת מספרי הזיהוי של הטרנזקציות המסומנות בחוקים אחרים . בכך, הטרנזקציה תסונן ותועתק לבסיס הנתונים המסונן D'. פרוצדורת look-ahead זו נעשית רק כאשר סף החשיפה הוא 0%, וזאת מכיוון שהפרוצדורה משפרת את העלות של ה miss cost אבל יכולה להנמיך באופן משמעותי את רמת ה hiding failure. אם ψ=0(כל החוקים הרגישים מוסתרים), אז אין hiding failure, וניתן לשפר את ה miss cost.

להדגמת פעולת היוריסטיקה נתבונן בבסיס הנתונים באיור 8(a). נניח שקיים סט של חוקי הקשר רגישים S_R={A,B→D; A,C→D}, ונקבע ש ψ=50%, מתרחשות התוצאות הבאות:

Figure 8 (a) A copy of the sample transactional database in Figure 3(a); (b) an example of partial sanitisation and (c) an example of full sanitisation

TID	Items
T1	A B C D
T2	A B C
T3	A B D
T4	A C D
T5	A B C
T6	B D

(a)

TID	Items
T1	A B C D
T2	A B C
T3	A D
T4	C D
T5	A B C
T6	B D

(b)

TID	Items
T1	A B C
T2	A B C
T3	A D
T4	C D
T5	A B C
T6	B D

(c)

- שלב 1: הטרנזקציות הרגישות מזוהות. במקרה זה, הטרנזקציות הרגישות של החוקים A,B→D ו- A,C→D הן {T1,T3} ו- {T1, T4} בהתאמה.
- שלב 2: בוחרים פריט קורבן. הפריט הקורבן עבור טרנזקציה T1 יכול להיות A או D מכיוון שפריטים אלו משתתפים בחוקי ההקשר וכתוצאה מכך , התדירות שלהם שווה ל-2. אולם הפריט הקורבן עבור חוק ההקשר הרגיש A,B→D ב T3 יבחר אקראית מכיוון שהפריטים A, B ו- D בעלי תדירויות שוות ל- 1. נניח ש B נבחר כפריט קורבן . באופן דומה נבחר פריט קורבן לחוק הקשר הרגיש A,C→D, בטרנזקציה T4 יבחר באופן אקראי הפריט A.
- שלב 3: מחשבים את מספר הטרנזקציות הרגישות שיש לסנן עבור כל חוק הקשר רגיש על פי סף החשיפה, שנקבע כ- ψ=50% לשני החוקים (ניתן לקבוע סף שונה לכל אחד בנפרד). והטרנזקציות הרגישות ממוינות בסדר יורד של גודל לפני ביצוע הסינון.
- שלב 4: שלב הסינון. מחצית מהטרנזקציות של כל חוק יישאר ו ללא שינוי (הסף נקבע ל- ψ=50%). תחלה יסוננו הטרנזקציות הקצרות יותר , לכן, טרנזקציות T3 ו- T4 יסוננו. בסיס הנתונים המסונן מוצג באיור 8(b). החוקים הרגישים מוצגים בבסיס הנתונים המסונן, אך עם תמיכה נמוך. זהו סינון חלקי. אם יקבע ψ=0%, יהיה סינון מלא מכיוון שחוקי ההקשר הרגישים לא יחשפו כלל . נניח כי ה פריט קורבן בטרנזקציה T1 הוא D, מכיוון שפריט זה משתתף בשני חוקי ההקשר הרגישים. איור 8(c) מציג את בסיס הנתונים לאחר סינון מלא . ניתן לראות כי הבעלים של בסיס הנתונים יכול לכוונן את סף החשיפה ולאזן בין הגנה על חוקי הקשר רגישים וכמות המידע שניתן לכרות.

Heuristic 3: Rule sanitization with block inference channels 3.2.2.3

שתי היוריסטיקות הקודמות סווגו כ- data sharing - based heuristics ובצעו סינון על בסיס הנתונים לפני שיתוף הנתונים . יוריסטיקה זו מסננת את חוקי ההקשר הרגישים מתוך קבוצת החוקים שנכרו מבסיס הנתונים, תוך חסימת אפיקי הסקת מסקנות . והיא מסווגת כ pattern sharing - based heuristics.

נשתמש בטרמינולוגיה מתוך תורת הגרפים , כך שפריטי בסיס הנתונ ים יוצגו כגרף מכוון . גרף כזה יקרא גרף סט-פריטים תדירים, a frequent itemset graph, והוא מוגדר להלן:

<u>הגדרה 1</u> – Frequent itemset graph : נגדיר גרף תדיר של סט פריטים G=(C, E), כגרף מכוון המכיל סט לא ריק של סטי פריטים תדירים C, קבוצה של קשתות E הנתונות כזוגות סדורים של אלמנטים של C, כך שלכל $u, v \in C$ קיימת קשת מ- u ל v אם $|v|-|u|=1$, $u \cap v = u \quad and$

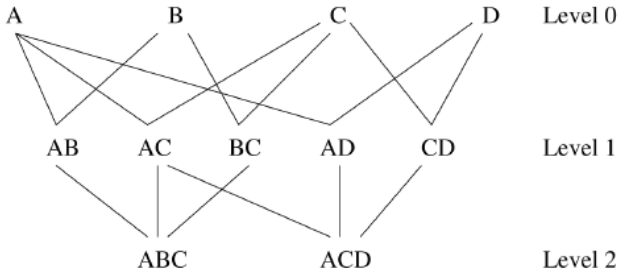
כאשר |x| הוא גודל סט הפריטים x.

איור 9(b) מראה את גרף סטי הפריטים התדירים המותאם ל בסיס הנתונים של הטרנזקציות המוצג באיור 9(a).

Figure 9 (a) A transactional database and (b) the corresponding frequent itemset graph

TID	List of Items
T1	A B C D
T2	A B C
T3	A C D
T4	A B C
T5	A B

(a)



(b)

בגרף סטי הפריטים התדירים G, קיים סדר לכל סט פריטים. שנתייחס אליו כרמת סט-הפריטים והוא מוגדר להלן:

<u>הגדרה 2</u> – The itemset level: יהי G=(C,E) גרף פריטים תדירים. רמת סט פריטים u, כך ש $u \in C$, האורך של המסלול המקשר בין 1-itemset ל-u.

בהתבסס על הגדרה זו, נגדיר את רמת גרף סט הפריטים התדירים G כ:

<u>הגדרה 3</u> – Frequent itemset graph level: יהי G=(C,E) גרף פריטים תדירים. הרמה של G היא האורך של המסלול המקסימאלי המקשר בין 1-itemset u לכל סט פריטים אחר v, כך ש- $u,v \in C$ וגם $v \subset u$.

באופן כללי, הגילוי של סט פריטים ב- G הוא תוצאת המעבר top-down על G תחת האילוץ של סף התמיכה המינימאלית σ. תהליך הגילוי מבצע גישה רקורסיבית שבה k-itemsets משמשים לחקור (k+1)-itemsets.

ליוריסטיקה 3 שלבים עיקריים המבוצעים לאחר תהליך הכרייה . כלומר הגרף G נבנה. סט כל סטי-הפריטים שניתן לכרות מ G, המבוסס על סף התמיכה המינימאלית σ הוא C.

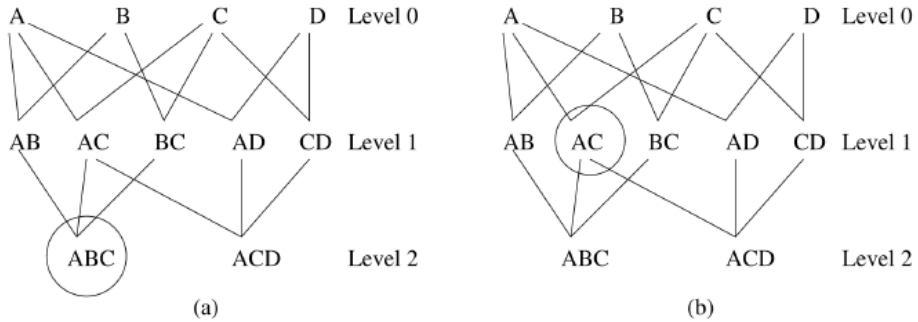
- שלב 1: Identifying the sensitive itemsets – כל חוק רגיש $sr_i \in SR$, יומר לסט פריטים רגיש $c_i \in C$.
- שלב 2: Selecting subset to sanitise: לכל סט פריטים c_i שיש לסנן, מחשבים את זוגות הפריטים שלו מדרגה 1 ב-G, תת סטים של c_i. אם אף אחד מהם לא מסומן , בוחרים באופן אקראי אחד מהם ומסמנים אותו כפריט למחיקה.
- שלב 3: Sanitizing the set of supersets of marked pairs in level 1: הסינון של סטי הפריטים הרגישים הוא הסרת הסט של ה- supersets של כל סטי הפריטים ברמה 1 של G שמסומנים להסרה. תהליך זה חוסם את האפשרויות של הסקת החוקים הרגישים. אפשרויות אלו נקראות inference channels ומתוארות בהמשך.

ביוריסטיקה זו ה- inference channels קורה כאשר כורים סט חוק ים מסונן ובהתבסס על חוקים לא רגישים מסיקים חוקים רגישים. ניתן לסווג חלק מההסקות כנגד החוקים הרגישים באופן הבא:

Forward-inference channel – נתון גרף פריטים תדירים באיור 10(a) . נניח כי רוצים לסנן את החוקים הנגזרים מסט הפריטים ABC. בגישה הנאיבית יסולק סט ה פריטים ABC. אולם אם AB, AC ו- BC הם תדירים, ניתן יהיה לנחש מתוצאת הכרייה שהסט ABC הוא תדיר וחסוי. יש להניח תמיד כי תוקפים ינסו כל ערוץ של הסקת מסקנות כדי ללמוד יותר מחוקי ההקשר המותרים . ייתכן כי התוקפים יסיקו גם סטים שאינם תדירים, אך יחד עם זאת הם יכולים ל הסיק מידע רגיש. הסקה זו נקראת Forward-inference channel. כדי להתמודד עם ערוץ ההסקה יש להסיר לפחות תת סט אחד של ABC (באקראי) מרמה 1 של הגרף. בגרף עמוק יותר ההסרה נעשית בצורה רקורסיבית עד רמה 1. לכן, פריטים ברמה 0 של הגרף אינם משתתפים עם פריט נוסף . ניתן להס יר גם את הסטים של ABC רקורסיבית עד רמה 0. במקרה כזה יש להתחשב באיזון בין הגנה על המידע וגילוי מידע בחוקים שנחשפים, מכיוון שתבניות תדירות אובדות בתהליך הסינון.

Backward-Inference Channel – סוג נוסף של הסקה קורה במהלך סינון סט פריטים לא סופי. בהתבסס על איור 10(b), נניח כי רוצים לסנן חוק הנגזר מהסט AC, אם נסיר את AC, קל להסיק את החוקים שנכרו מ AC, מכוון שגם ABC וגם ACD תדירים. הסקה זו נקראת Backward-Inference Channel, וכדי לחסום אותה יש להסיר כל סט המכיל את AC. במקרה זה, יש להסיר את ABC ו ACD.

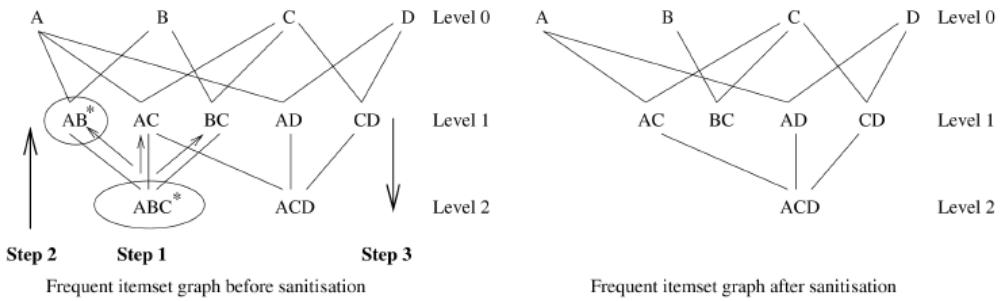
Figure 10 (a) An example of forward-inference and (b) an example of backward-inference



הדגמת יורי סטיקה 3 – נתבונן בגרף סט הפריטים התדירים שבאיור 9(b) המתיחס לבסיס הנתונים המתואר באיור 9(a), כאשר התמיכה המינימאלית σ=2, ויש לסנן את החוק הרגיש A,B → C לפני שיתוף הגרף:

- שלב 1: כל חוק רגיש מומר לסט הפריטים התואם לו . כך שהחוק A,B → C מומר ל ABC.
- שלב 2: תתי הסטים של כל חוק נבחרים. באופן כללי, תתי הסטים נבחרים מרמה 1 של הגרף. בדוגמא, בהנחה שאף תת סט לא סומן יבחר תת סט באופן אקראי AB, והוא יתוסף לרשימה המכילה את כל הזוגות המסומנים לסינון. אם כבר סומן תת סט, נבחר בו לאופטימיזציה של תהליך הסינון.
- שלב 3: שלב הסינון בו מסירים את כל הסטים המכילים את הזוגות המסומנים לסינון . בדוגמא, כל הסטים המכילים את AB יסולקו. איור 11 מדגים את הגרף לפני ואחרי סינון החוקים.

Figure 11 An example of a frequent itemset graphs before and after sanitisation



3.2.3 אלגוריתמי הסינון

אלגוריתמי הסינון נועדו להגן על המידע הרגיש ב בסיס הנתונים של הטרנזקציות והם מתייחסים ליוריסטיקות שהוצגו בסעיף הקודם והם מחולקים ל- 2 קבוצות: data–sharing based ול- pattern–sharing based.

3.2.3.1 The item-grouping algorithm (IGA)

אלגוריתם זה מבוסס על יוריסטיקה 1. הרעיון העיקרי הוא לקבץ חוקים רגישים בקבוצות של חוקים החולקים את אותם סטי הפריטים. אם שני חוקים רגישים נחתכים על ידי אותה טרנזקציה רגישה, הרי ששניהם יוסתרו בבת אחת, ובכך תקטן ההשפעה על בסיס הנתונים שייחשף. מצד שני, קיבוץ החוקים הרגישים בהתבסס על החיתוך בין הפריטים מוביל לקבוצות חופפות מכיוון שהחיתוך אינו טרנזיטיבי. על ידי חישוב החפיפה בין המקבצים ובידוד הקבוצות ניתן לבחור את הסטים המקשרים את החוקים הרגישים בכל קבוצה כקורבנות לכל החוקים המופיעים בקבוצה.

<u>בשלב 1</u>, האלגוריתם בונה אינדקס מהופך, inverted index, המבוסס על הטרנזקציות ב- D, במעבר יחיד. האינדקס מכיל את כל החוקים הרגישים, ולכל חוק רגיש קיימת רשימה של מספרי זיהוי של טרנזקציות המכילות את החוק. בשורות 4-5 ה IGA מחשב את התדירויות של כל הפריטים בבסיס הנתונים D(תדירויות אלו ישמשו בשלב 3 לבחירת הפריט קורבן). משורות 7 – 11 ה IGA בונה את האינדקס המהופך.

<u>בשלב 2</u>, האלגוריתם ממיין את כל הטרנזקציות הרגישות המזוהות עם החוקים הרגישים בסדר יורד של t_{ID}, שורה 15. בשורה 16, מספר הטרנזקציות הרגישות שיש לסנן, בכל חוק רגיש sr_i, נבחר בהתבסס על סף החשיפה ψ.

<u>בשלב 3</u>, המטרה היא לזהות את הפריט הקורבן לכל חוק רגיש. מהרגע שנקבע הפריט הקורבן לחוק רגיש sr_i, יש לסלק אותו מכל הטרנזקציות הרגישות הקשורות עם החוק sr_i. הבחירה של הפריט קורבן נעשית על ידי קיבוץ כל החוקים הרגישים בסט של קבוצות חופפות GP (step3.1) כך שכל החוקים הרגישים באותה קבוצה G יחלקו את אותם הפריטים. הקבוצות של החוקים הרגישים ממוינות בסדר יורד של פריטים משותפים (step 3.2). הפריטים המשותפים הם המזהים של המחלקות של הקבוצות. כך שלמשל התבניות ABC ו ABD יהיו באותה קבוצה המסומנת כ A או B, תלוי בתמיכה של A ו B (step 3.3). ABC יכול לה ימצא בקבוצה נוספת אם קיימת כזו שבה החוקים הרגישים חולקים 'C'. משורות 26-32, ה IGA מזהה חפיפות בין קבוצות ומונע אותם על ידי העדפת קבוצות גדולות או קבוצות המתויגות כבעלות תמיכה נמוכה בבסיס הנתונים. ברגע שסודרו הקבוצות נקבע לכל חוק רגיש הפריט קורבן שהוא למעשה מזהה הקבוצה שאליה הוא שייך.

ההיגיון העומד מאחורי בחירת הפריט קורבן ב IGA הוא שמכיוון שהפריט קורבן מייצג סט של חוקים רגישים (מאותה קבוצה), סינון טרנזקציה רגישה יכולה לאפשר לחוקים רגישים רבים להיות מטופלים בבת אחת. אסטרטגיה זו מבטיחה הפחתה של תוצאות הלוואי של הכרייה על החוקים שאינם רגישים בבסיסי הנתונים המסונן.

<u>בשלב 4</u>, סורקים תחילה את הווקטו ר Victim בסדר עולה של t_{ID}, שורה 40. אז נסרק בסיס הנתונים שוב (בפעם השנייה והאחרונה) בלולאה, שורות 42-47. אם הטרנזקציה הנוכחית (t_{ID}) נבחרת להיות מסוננת, הפריט קורבן הקשור לטרנזקציה זו t מסולק ממנה. טרנזקציות שאין צורך לסנן מועתקות ישירות מ D ל- 'D.

(דוגמא לאופן ביצוע האלגוריתם ניתן לראות בסעיף 3.2.2.1, איור 7, המציג את יוריסטיקה 1).

זמן הריצה של האלגוריתם IGA הוא $O(n_1 \times N \times \log N)$, כאשר n_1 הוא מספר החוקים הרגישים, ו N הוא מספר הטרנזקציות בבסיס הנתונים. (ההוכחה מובאת במאמר של Oliveira 2005[40]).

Algorithm 1: Item_Grouping_Algorithm

input: D, S_R, ψ
output: D'

```
1  begin
2      // Step 1:  Identifying sensitive transactions and building index T
3      foreach transaction t ∈ D do
4          for k = 1 to size(t) do
5              Sup(item_k, D) ← Sup(item_k, D) + 1;  //Update support of each item_k in t;
6          Sort the items in t in alphabetic order;
7          foreach sensitive association rule sr_i ∈ S_R do
8              if items(sr_i) ⊆ t then
9                  T[sr_i].tid_list ← T[sr_i].tid_list ∪ TID_of(t);
10             end
11         end
12     end
13     // Step 2:  Selecting the number of sensitive transactions
14     foreach sensitive association rule sr_i ∈ S_R do
15         Sort the vector T[sr_i].tid_list in descending order of degree;
16         NumbTrans_sr_i ← |T[sr_i]| × (1 − ψ);
17         // |T[sr_i]| is the number of sensitive transactions for sr_i
18     end
19     // Step 3:  Identifying victim items for each sensitive transaction
20     3.1 Group sensitive rules in a set of groups GP such that ∀ G ∈ GP,
21         ∀ sr_i, sr_j ∈ G, sr_i and sr_j share the same itemset I. Give the class label
22         α to G such that α ∈ I and ∀β ∈ I, sup(α, D) ≤ sup(β, D);
23     3.2 Order the groups in GP by size in terms of number of sensitive rules
24         in the group;
25     // Compare groups pairwise G_i and G_j starting with the largest
26     3.3 forall sr_k ∈ G_i ∩ G_j do
27         if size(G_i) ≠ size(G_j) then
28             remove sr_k from smallest(G_i, G_j);
29         else
30             remove sr_k from group with class label α such that sup(α, D) ≥ sup(β, D) and α, β are class
               labels of either G_i or G_j;
31         end
32     end
33     3.4 foreach sensitive association rule sr_i ∈ S_R do
34         for j = 1 to NumbTrans_sr_i do
35             ChosenItem ← α such that α is the class label of G and sr_i ∈ G;
               Victims[T[sr_i, j]].item_list ← Victims[T[sr_i, j]].item_list ∪ ChosenItem;
36         end
37
38     end
39     // Step 4:  D' ← D
40     Sort the vector Victims in ascending order of t_ID;
41     j ← 1;
42     foreach transaction t ∈ D do
43         if t_ID == Victims[j].t_ID then
44             t ← (t − Victims[j].item_list);
45             j ← j + 1;
46         end
47     end
48 end
```

(SWA) The sliding-window algorithm 3.2.3.2

אלגוריתם זה מב וסס על יוריסטיקה 2. האינטואיציה היא , שה SWA סורק קבוצה של K טרנזקציות (גודל החלון) בכל פעם , ואז מסנן את קבוצת הטרנזקציות הרגישות , המסומנות כ S_T, בהתבסס על סף החשיפה ψ. כל השלבים של יוריסטיקה 2 מיושמות על כל קבוצה של K טרנזקציות שנקראו מבסיס הנתונים המקורי D.

שלא כמו אלגוריתם ה IGA, שבו לכל החוקים הרגישים אותו סף חשיפה, לSWA יש סף חשיפה השייך לכל חוק הקשר רגיש. הסט המקשר בין חוקי ההקשר הרגישים ל בין סף החשיפה המתאים

נקרא the set of mining permissions, והוא מסומן כ- M_P, כאשר כל mining permission, mp, מאופיין על ידי זוג סדור המסומן $mp=<sr_i, \psi_i>$ כאשר לכל $sr_i \in S_R$,i ו $\psi_i \in [0..1]$.

הקלט עבור ה SWA הוא בסיס הנתונים של הטרנזקציות D, הסט של ה mining permissions M_P וגודל החלון K. הפלט הוא בסיס הנתונים המסונן 'D.

ל SWA ארבעה שלבים עיקריים:

<u>בשלב 1</u>, האלגוריתם סורק K טרנזקציות ושומר חלק מהמידע במבנה נתונים T המכיל:

(1) רשימה של מספרי זיהוי של טרנזקציות רגישות לכל חוק רגיש

(2) רשימה עם גודל הטרנזקציות הרגישות התואמות.

(3) רשימה נוספת עם הפריט קורבן לכל טרנזקציה רגישה תואמת.

הטרנזקציה t היא רגישה אם היא מכילה את כל הפריטים של לפחות חוק רגיש אחד. ה SWA מחשב את התדירויות של הפריטים של החוקים הרגישים המיוצגים בכל טרנזקציה רגישה. חישוב זה יתמוך בבחירת הקורבנות בשלב הבא. בשורה 11, הווקטור v_transact מאחסן את הטרנזקציות הרגישות בזיכרון הראשי.

<u>בשלב 2</u>, הווקטור עם התדירויות שחושב בשלב הקודם, ממוין בסדר יורד. כתוצאה מכך, הפריט קורבן נבחר לכל טרנזקציה רגישה. הפריט עם התדירות הגבוהה ביותר הוא פריט קורבן והוא מסומן כמועמד לסילוק מהטרנזקציה. אם התדירויות של הפריטים שוות ל- 1, ניתן לבחור כל אחד מהפרטים כפריט קורבן, והוא יבחר באופן אקראי.

<u>בשלב 3</u>, מספר הטרנזקציות שיש לסנן לכל חוק רגיש נבחר. בשורה 31 מראה ש ψ_i משמש לחישוב מספר זה, $NumTrans_{sri}$. ה SWA ממיין את הטרנזקציות הרגישות לכל חוק רגיש בסדר עולה. מיון זה הוא הבסיס ליוריסטיקה 2.

<u>בשלב 4</u>, הטרנזקציות הרגישות מסוננות בלולאה בשורות 35-42. אם סף החשיפה הוא 0 (כלומר יש לסנן את כל החוקים הרגישים), יש לבדוק את ה mining permission (M_P) כדי להחליט אם טרנזקציה רגיש ה אינה צריכה להיות מסוננת יותר מפעם אחת. כך משפרים את ה miss cost. הפונקציה look_ahead() מתבוננת ב M_P מתוך sr_i כדי להחליט אם טרנזקציה נתונה t נבחרה כטרנזקציה רגישה עבור חוק רגיש אחר r. אם כן, ואם $T[sr_i].tid\_list[j]$ ו- $T[sr_i].victim[j]$ הם חלק מהחוק הרגיש r, אז הטרנזקציה t מסולקת מהרשימה מכיוון שהיא כבר סוננה.

(דוגמא לאופן ביצוע האלגוריתם ניתן לראות בסעיף 3.2.2.2, איור 8, המציג את יוריסטיקה 2).

זמן הריצה של האלגוריתם WSA הוא $O\left(n_1 \times N \times \log K\right)$, כאשר $\psi \neq 0$ ו- $\psi=0$ כאשר $O\left(n_1^2 \times N \times K\right)$, n_1 הוא מספר החוקים הרגישים ב D, N הוא מספר הטרנזקציות ב D, ו K הוא גודל החלון שנבחר. (ההוכחה מובאת במאמר של Oliveira 2005[1]).

Algorithm 2: Sliding_Window_Algorithm

input: D, M_P, K
output: D'

```
1  begin
2      foreach K transactions in D do
3          // Step 1:  Identifying sensitive transactions & building index T
4          foreach transaction t ∈ K do
5              Sort the items in t in alphabetic order;
6              foreach sensitive association rule sr_i ∈ M_P do
7                  if items(sr_i) ⊆ t then
8                      T[sr_i].tid_list ← T[sr_i].tid_list ∪ TID_of(t);    //t is sensitive
9                      T[sr_i].size_list ← T[sr_i].size_list ∪ size(t);
10                     freq[item_j[← freq[item_j] + 1;
11                     v_transac ← v_transac ∪ t;   //Sensitive transactions in memory
12                 end
13             end
14             // Step 2:  Identifying the victim items
15             if t is sensitive then
16                 Sort vector freq in descending order;
17                 foreach sensitive association rule sr_i ∈ M_P do
18                     Select item_v such that item_v ∈ sr_i and ∀ item_k ∈ sr_i,
19                             freq[item_v] ≥ freq[item_k];
20                     if freq[item_v] > 1 then
21                         T[sr_i].victim ← T[sr_i].victim ∪ item_v;
22                     else
23                         T[sr_i].victim ← T[sr_i].victim ∪ RandomItem(sr_i);
24                     end
25                 end
26             end
27         end
28     end
29     // Step 3:  Selecting the number of sensitive transactions
30     foreach sensitive association rule sr_i ∈ M_P do
31         NumTrans_sr_i ← |T[sr_i]| × (1 − ψ_i);
32         Sort the vector T in ascending order of size;
33     end
34     // Step 4:   D' ← D
35     foreach sensitive association rule sr_i ∈ M_P do
36         for j = 1 to NumbTrans_sr_i do
37             remove(v_transac[T[sr_i].tid_list[j], T[sr_i].victim[j]]);
38             if ψ_i = 0 then
39                 do look_ahead(sr_i, T[sr_i].tid_list[j], T[sr_i].victim[j]);
40             end
41         end
42     end
43 end
```

(DSA) The downright-sanitizing algorithm 3.2.3.3

אלגוריתם זה מבוסס על יוריסטיקה 3. הרעיון הוא לסנן מספר חוקים רגישים תוך חסימת inference channels. ה DSA מסלק לפחות תת סט אחד של כל סט פריטים רגיש ברמה 1 של גרף סטי הפריטים התדירים . הסילוק נעשה בצורה רקורסיבית עד רמה 1. הDSA מתחיל לסלק מרמה 1 מכיוון שמניחים שחוקי הקשר המשוחזרים מהסטים (המשותפים) המסוננים מכילים לפחות שני פריטים. ניתן להתחיל לסלק מרמה 0, אבל אפשרות זו תפחית את השימושיות של המידע הנחשף מכיוון שיותר

סטי פריטים יסוננו, מה שיגביר את תופעת הלוואי של ה miss cost. לכן, הפריטים ברמה 0 של סטי הפריטים התדירים לא ישותפו כלל. בכך, מקטינים את ה inference channels ואת תופעת הלוואי על החוקים הלא רגישים שנכרו מגרף סטי הפריטים המסונן.

הקלט לאלגוריתם הוא גרף הפריטים התדירים G, קבוצת החוקים הרגישים שיש לסנן S_R. הפלט הוא גרף הפריטים התדירים המסונן 'G.

Algorithm 3: Downright_Sanitising_Algorithm

input: G, S_R
output: G'

```
1  begin
2      // Step 1:  Identifying the sensitive itemsets
3      foreach sensitive association rule sr_i ∈ S_R do
4          c_i ← sr_i;   //Convert each sr_i into a frequent itemset c_i
5      end
6      // Step 2:  Selecting subsets to sanitize
7      foreach c_i in the level k of G, where k ≥ 1 do
8          Pairs(c_i);   //Compute all the item pairs of c_i
9          if (Pairs(c_i) ∩ MarkedPair = ∅) then
10             p_i ← random(Pairs(c_i));   //Select randomly a pair p_i ∈ c_i;
11             MarkedPair ← MarkedPair ∪ p_i;   //Update the list MarkedPair
12         end
13     end
14     // Step 3:  Sanitizing the set of supersets of marked pairs
15     //          in level 1 (R' ← R)
16     foreach itemset c_j ∈ G do
17         Sort the items in c_j in alphabetic order;
18     end
19     foreach itemset c_j ∈ G do
20         if ∃ a marked pair p, such that p ∈ MarkedPair and p ⊂ c_j then
21             Remove(c_j) from R';   //c_j belongs to the set of supersets of p;
22         end
23     end
24 end
```

לאלגוריתם 3 שלבים:

<u>בשלב 1</u>, מגדירים את סטי הפריטים התדירים על ידי המרת חוקי ההקשר לאוסף של פריטים c_i.

<u>בשלב 2</u>, בוחרים תתי סטים לסינון. לכל c_i ברמה הגדולה מ-0 מחשבים את קבוצת כל זוגות הפריטים של ה- c_i. אם לא קיים חיתוך בין קבוצה זו לבין אוסף הזוגות הקיים MarkedPair מוסיפים לאוסף זוג פריטים השייך ל c_i שנבחר באופן אקראי.

<u>בשלב 3</u>, מסננים את קבוצת סטי הע ל של ה Marked Pair ברמה 1. תחילה ממיינים את הפריטים בכל סט פריטים בסדר מילוני. לאחר מכן בודקים לכל סט פריטים אם קיים זוג השייך ל MarkedPair המוכל בו. אם כן, סט הפריטים יוסר.

(דוגמא לאופן ביצוע האלגוריתם ניתן לראות בסעיף 3.2.2.3, איור 9, המציג את יוריסטיקה 3).

זמן הריצה של ה DSA הוא $O\left(n\times\left(k^2+m\times\log k\right)\right)$, n הוא מספר החוקים הרגישים שיש לסנן, m הוא מספר סטי הפריטים התדירים בגרף G, ו k הוא המספר המקסימאלי של פריטים ב G. (ההוכחה מובאת במאמר של Oliveira 2005[1]).

3.2.4 מסקנות וסיכום

3.2.4.1 הערכת data sharing - based algorithms

סט נרחב של ניסויים נערך על מנת להעריך את היתרונות והחסרונות שבכל אחת מהשיטות של ה- data sharing - based algorithms. התוצאות העיקריות הן:

- ככל שבסיס הנתונים גדול יותר , כך מתקבלות תוצאות טובות יותר גם מבחינת ה miss cost וגם מבחינת ה hiding failure.
- ה IGA מצריך שני מעברים על בסיס הנתונים , כאשר לעומתו ה SWA מצריך מעבר יחיד.
- ה IGA מציג ביצועים טובים מאד. כמעט בכל המדדים IGA הניב את התוצאות הטובות יותר גם מבחינת ה miss cost וגם מבחינת ה hiding failure. חריגות אירעו במקרים בהם חוקים רגישים הכילו פריטים עם תמיכה גבוה מאד. במקרה זה, האלגוריתם SWA הראה תוצאות טובות יותר עבור ה miss cost.
- ה SWA הניב את התוצאות הטובות יותר בהבדל בין בסיס הנתונים המקורי והמסונן , אך לא השיג את התוצאות הטובות ב miss cost. מה שממחיש את פרדוקס סינון הנתונים . הקטנת הה שפעה על בסיס הנתונים המסונן אינה מבטיחה את התוצאות הטובות ביותר במונחים של miss cost. הסיבה לכך היא שבחירת הפריט קורבן ב SWA היא דינאמית. בכל כל פעם בוחרים פריט קורבן חדש. בכך מקטינים את ה תמיכה של כל פריט בחוק רגיש מבלי להתחשב במידת התמיכה של הפריט . הפחתת הפריטים עם תמיכה גבוהה תקצץ את יצירת המועמדים של חוקים גלויים תוך פשרה עם ערכי ה miss cost. ולהפך, הפריט קורבן הנבחרים על ידי IGA עבור החוקים הרגישים קבוע לכל הטרנזקציות הרגישות. ובנוסף, הIGA תמיד בוחר את המועמדים בעלי התמיכה המינימאלית לכל חוק. מה שמבטיח שיפור בערכי ה miss cost.

שתי השיטות פועלות היטב כאשר קיימת ערבות הדדית בין הפריטים שבחוקי ההקשר שיש לסנן. הכוונה היא, שאם לא קיים חיתוך בין הפריטים בכל חוקי ההקשר , לא מומלץ להעדיף את השיטות הללו המבוססות על היתרון שבחיתוך בין החוקים.

3.2.4.2 הערכת pattern sharing – based algorithm

אלגוריתם DSA מסיר חוקי הקשר רגישים לפני תהליך שיתוף התבניות . תהליך סינון זה חוסם אפשרויות של הסקת חוקי הקשר רגישים, inference channels.

כפי הנראה לא קיימים אלגוריתמים מסוג pattern sharing – based algorithm לסינון חוקי הקשר בספרות. מסיבה זו תיערך השוואה בין אלגוריתם ה- DSA לבין ה- IGA שהניב את התוצאות הטובות ביותר. בגלל אופיים השונה של האלגוריתמים , יש לבצע שלבים שונים לכל אחד לפני ביצוע ההשוואה:

שלבים עבור IGA:

א. הפעלת IGA כדי לסנן את הסטים של החוקים הרגישים בבסיסי הנתונים הנדרשים.
ב. ביצוע האלגוריתם לכריית חוקי ההקשר על בסיסי הנתונים המסוננים כדי לקבלת התבניות המורחבות לשיתוף.

שלבים עבור DSA:

א. ביצוע האלגוריתם לכריית חוקי ההקשר על בסיסי הנתונים לקבלת התבניות המורחבות לשיתוף.
ב. הפעלת DSA לסינון החוקים הרגישים לפני שיתוף החוקים שהתקבלו.

המטרה היא לבדוק מתי עדיף להשתמש בכל אחד מהאלגוריתמים כדי להגן על חוקי הקשר רגישים. המסקנות שהתקבלו מהניסויים שנערכו מראות של DSA ערכי side-effect נמוכים וכך כמובן גם עבור ה miss-cost. כן התברר כי יכולת השחזור בו נמוכה , כך ש DSA מבטיח הגנה כנגד inference channel לפני שיתוף חוקי ההקשר. ניתן לסכם את יתרונותיו:

- שימוש ב DSA מאפשר לבעלים לשתף תבניות (תוצאות) ולא נתונים ממש.

- קיימת ירידה משמעותית ביכולת לבצע inference channel מכיוון שסף התמיכה ואלגוריתם הכרייה נבחרים על ידי מנהל בסיס הנתונים.
- סינון החוקים במקום תו צאות הנתונים אינו שינוי של ה תמיכה ורמת הביטחון של החוקים הלא רגישים , כלומר, לחוקים המופצים יש את התמיכה ורמת הביטחון המקוריים. כך שהחוקים המופצים הם בעלי משמעות טובה יותר עבור יישומים תכליתיים. (השיטות האחרות מפחיתות את התמיכה ורמת הביטחון כתוצאת לוואי של תהליך הסינון).

החיסרון המשמעותי של ה DSA הוא בהפחתת גמישות המידע המשותף מכיוון שבכל פעם שצד שלישי רוצה לנסות רמות שונות של תמיכה ורמת הביטחון, עליו לבקש זאת ממנהל הנתונים.

3.2.4.3 סיכום

המאמר הציג שיטות עבודה אחידות לסינון חוקי הקשר רגישים בכריית מידע. המאמר כלל:

- Retrieval facilities. להאצת התהליך נבנה אינדקס , ובכך נדרשות שתי סריקות בלבד כדי להסתיר חוקי הקשר רגישים ללא קשר למספרם . סריקה אחת לבניי ת אינדקס, והשנייה להסתיר את החוקים הרגישים . בספרות הקיימת עד הצגת טכניקה זו (2006), האלגוריתמים דורשים מספר מעברים.
- A library of sanitizing algorithms. האלגוריתמים מחולקים לשתי קבוצות עיקריות: data sharing - based algorithms, ו pattern sharing – based algorithm. בשיטה הראשונה הסינון פועל על הנתונים כדי להסתיר את קבוצת חוקי ההקשר הרגישים המכילים מידע רגיש, ובשנייה הסינון פועל על חוקי ההקשר עצמם שנכרו מבסיס הנתונים במקום על הנתונים . שלושת האלגוריתמים הפועלים על שלושת היוריסטיקות שהוצגו פועלים על ידי הקטנת התמיכה ורמת הביטחון.
- A set of metrics. במאמר הוצעו מטריצות שונות (אותן לא הצגתי בעבודה זו) שתוכננו להעריך את כמו ת המידע הרגיש שנחשף , וכן למדוד את היעילות של אלגוריתמי הסינון במונחים של איבוד מידע של חוקים לא רגישים שהוסרו כתוצאה מהשפעות הלוואי של תהליך הסינון. המטריצות חולקו אף הן לשתי קבוצות עיקריות: data sharing - based metrics, ו pattern sharing – based metrics.

מהתבוננות בשיטות שהוצגו המאמר , ניתן להסיק כי שני האלגוריתמים הראשונים , IGA ו-SWA, יפעלו ביעילות רבה יותר ככל שהקשר בין התבניות הרגישות יגדל . במקרה כזה, סילוק פריט קורבן מתאים יסנן מספר רב של טרנזקציות רגישות.

לאלגוריתם SWA יש שני יתרונות בולטים וחשובים ע ל פני ה IGA. הראשון שבהם הוא היכולת לבחור סף חשיפה שונה לכל חוק הקשר רגיש . שלא כמו באלגוריתם ה IGA, שבו לכל החוקים הרגישים אותו סף חשיפה. היתרון השני הוא שה- SWA הניב את התוצאות הטובות יותר בשונות בין בסיס הנתונים המקורי והמסונן . יתרון זה משמעותי מאד אם מס תכלים על המטרה המרכזית של כריית הנתונים. ואכן, המאמר הבא המוצג בהמשך העבודה , משתמש ב SWA לביצוע ההשוואות והערכות הביצועים, וכן למידת האבטחה.

חיסרון מהותי מאד באלגוריתמים אלו היא עובדת היותם חשופים להתקפת inference channel, מה שאולי יכול לגרום למנהל הנתונים לבחור שיטות אחרות שיספקו מידת הגנה גבוהה יותר.

האלגוריתם DSA מציג ביצועים טובים יותר בכל המדדים לעומת האלגוריתמים הקודמים , ובפרט, הוא מספק מידת הגנה גבוהה יותר כנגד inference channel, ובכך יש לו שיפור גדול לעומתם. בנוסף, לחוקים המופצים יש את ה תמיכה ורמת הביטחון המקוריים. כך שהחוקים הם בעלי משמעות טובה יותר עבור יישומים תכליתיים.

יתרון נוסף שיש ל DSA הוא העובדה כי הסינון נעשה לאחר כריית הנתונים . הדבר מבטיח כריית נתונים מקסימאלית , שרק לאחריה סינון מידע רגיש , כך שלדעתי כמות המידע המופקת מתהליך כזה תהיה בד"כ גבוהה ואיכותית יותר מאשר ביצוע כרייה על טרנזקציות שסוננו מראש.

3.3 *An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining* [41]

המאמר מציג אלגוריתם סינון יעיל לאיזון בין פרטיות המידע וגילוי המידע בהקשר של כריית תבניות מידע.

מטרת האלגוריתם היא לקבוע איזון מתאים בין הצורך בפרטיות וגילוי המידע בתבניות תדירות, ומציע שיטה מקורית לשינוי בסיס הנתונים כדי להסתיר תבניות רגישות. הכפלת בסיס הנתונים המקורי על ידי מטריצת סינון ליצירת בסיס נתונים מסונן המספק פרטיות. בנוסף, שתי הסתברויות מוצגות כדי למנוע שחזור של תבניות רגישות וכדי להפחית את כמות התבניות המוסתרות שאינן רגישות בבסיס הנתונים המסונן. המאמר מספק גם ניתוח סיבוכיות ודיון על מידת האבטחה בתהליך הסינון המוצע ומוכיח את היעילות שבשיטה.

3.3.1 הקדמה

המאמר מתייחס למצב בו קיימת מערכת המורכבת משרת וקבוצת לקוחות. לכל אחד מהלקוחות בסיס נתונים של טרנזקציות. הלקוחות מעוניינים שהשרת ירכז מידע סטטיסטי על קשרים בין הפריטים בבסיסי הנתונים המאוחדים ויספק המלצות על הצרכנים. אולם, הלקוחות אינם מעוניינים שהשרת ישיג תבניות רגישות שיתגלו מבסיסי הנתונים שלהם. כאשר תבנית רגישה היא תבנית תדירה עם מידע רגיש, כגון סודות מסחריים. לכן, לפני שליחת בסיסי הנתונים לשרת הלקוח חייב להחביא את התבניות הרגישות על פי מדיניות אבטחה כלשהי הקשורה רק לבסיס הנתונים ששונה.

המאמר מציג שיטה לשינוי בסיסי הנתונים כדי להסתיר תבניות רגישות. הקשר בין התבניות הרגישות והלא-רגישות קיים בהגדרת מטריצת הסינון (sanitized matrix). הכפלת בסיס נתוני הטרנזקציות המקורי על ידי מטריצת סינון יכולה לצור בסיס נתונים מסונן (sanitized database) להסתרת תבניות רגישות שיכולות להתגלות על ידי Forward-inference Attacks. שתי מדיניות הסתברותיות עם רמה של אבטחה המוגדרת על ידי המשתמש מוצעות כדי להגן מפני שחזור תבניות רגישות וכדי להפחית את רמת התבניות הלא-רגישות המוסתרות בבסיס הנתונים המסונן. בהתייחס ל (1) רמת השוני בין בסיס הנתונים ששונה לבסיס הנתונים המקורי ו- (2) הטולרנס של הכישלון בהסתרה, הרמה של האבטחה יכולה להינתן על ידי המשתמש.

ניתן לסכם את תרומת המאמר כדלקמן: תחילה, במקום לבדוק את כל בסיס הנתונים כדי לזהות את הטרנזקציות הרגישות, שיטה מבוססת מטריצה משמשת לתהליך הסינון המקורי. לכן, הבעיה בהסתרת התבניות הרגישות מומרת ל-'איך להגדיר את מטריצת הסינון'. שנית, המדיניות הסתברותית עם רמת האבטחה המוגדרת על ידי המשתמש מאפשרת בפעם הראשונה לפתור את בעיית המידע הפרטי. שלישית, ניתן להמנע מבעית ה- Forward-inference Attacks בבסיס הנתונים המסונן שנוצר בתהליך הסינון.

3.3.2 רקע

3.3.2.1 הגדרת הבעיה

הבעיה של גילוי חוקי הקשר מוגדרת כמציאת קשרים בין מופעי פריטים לטרנזקציות. התמיכה בסטי פריטים נמדדת בהסתברות שטרנזקציה תכיל את התבנית שלהם. עקב הכמות העצומה של הקומבינציות של הפריטים, המשתמש יכול להגדיר את התמיכה המינימאלית כסף תחתון שבו חוקי הקשר יהיו משמעותיים. נגדיר frequent pattern (תבנית תדירה) כחוק הקשר המספק את התמיכה המינימאלית.

בגישה זו, הטרנזקציות של בסיס הנתונים מיוצגות על ידי מטריצה בינארית D שבה השורות מייצגות טרנזקציות, והעמודות מייצגות פריטים. כניסה D_{ij} היא 1 אם פריט j נמצא בטרנזקציה i, ו-0 אם לא. הבעייה של הסתרת תבניות רגישות מוגדרת כלהלן: תהי P קבוצת כל התבניות התדירות שנכרו מ D פרט לתבניות הרגישות באורך 1 (פריטים תדירים), יהי P_H סט של תבניות הרגישות ו- $\sim P_H$ סט של כל שאר התבניות התדירות (הלא רגישות), כך ש- $P_H \cup \sim P_H = P$. הרעיון הוא להמיר את D לבסיס נתונים מסונן D', כך שרק התבניות השייכות ל $\sim P_H$ יוכלו להכרות מ D'. בנוסף, התבניות

הרגישות לא יוכלו להחשף ל- Forward-inference Attacks . מכאן, עבור תבניות רגישות , לפחות אחת מתת התבניות בעלות אורך הגדול מ- 2 יש להסתיר. מהפריטים התדירים הבודדים מתעלמים . רק בתבניות פריטים באורך גדול מ- 2 מתחשבים. ברור, כי מרבית הפריטים הבודדים הם תדירים ברוב בסיסי הנתונים. רק על ידי פריטים בודדים תדירים כמעט ולא ניתן לייצר חוקי הקשר . אם התוקפים משתמשים בפריטים בודדים תדירים כדי להסיק רקורסיבית את ההסתברות לתבניות רגישות אפשריות בעלות אורך גדול מ-2, הם יכולים לקבל כמעט כל תבנית מבסיס הנתונים.

Basic concepts of the sanitization matrix 3.3.2.2

כפי שמוצג באיור 2, בגישה זו, תהליך הסינון של הטרנזקציות שולח את בסיס הנתונים המקורי D לבסיס הנתונים המסונן D' על ידי מטריצת הסינון S. כך ש- $D'_{n\times m} = D_{n\times m} \times S_{n\times m}$. אם S היא מטריצת הזהות (כלומר $S_{ij}=1$ אם i=j, אחרת $S_{ij}=0$), D' יהיה זהה ל- D. לכן, קביעת רכיבים שאינם באלכסון של S לערכים המתאימים יכולה להוביל לבסיס הנתונים המסונן.

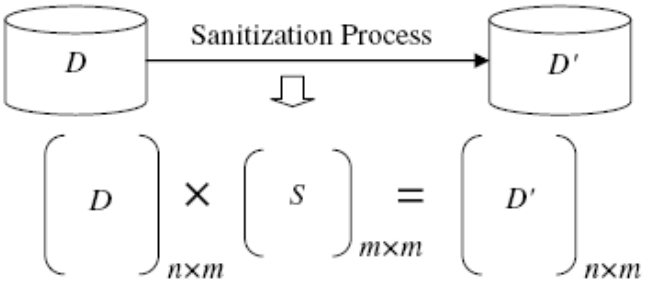


Fig. 2. Concept of the sanitization process.

הגדרה חדשה להכפלת מטריצות 1.2.2.3.3

כדי לספק את המאפיינים של המטריצה הבינארית המייצגת את טרנזקציות בסיס הנתונים , ניתנות ההגדרות הבאות לכפל מטריצות:

1. אם $D_{ij}=0$ אז D'_{ij} יקבע לערך 0. מכיוון שהמטרה היא להסתיר תבניות רגישות על ידי הקטנת התמיכה, יש לדאוג רק למקרים שבהם יש להמיר את הכניסה ב D עם הערך 1 ל-0 ב-D'. יתר על כן, הדבר יבטיח שלא ייווצרו תבניות מלאכותיות לאחר תהליך הסינון.
2. אם ערך התוצאה $\sum_{k=1}^{m} D_{ik} \cdot S_{kj} \geq 1$, אזי D'_{ij} יקבע ל-1.
3. אם ערך התוצאה $\sum_{k=1}^{m} D_{ik} \cdot S_{kj} \leq 0$, אזי D'_{ij} יקבע ל-0.

הבחנה 1: קביעת כניסות ל -1 2.2.2.3.3

במטרה להחביא תבנית {i, j}, התמיכה צריכה לקטון. למשל, אם D_{ki} ו- D_{kj} קבועים ל-1 עבור טרנזקציה k, הערך של D_{ki} או D_{kj} יכול להקבע ל-0 כדי להקטין את התמיכה של {i, j}. אם מספר סופי של כניסות כאלו יכול להיות מוחלף ב-0, {i, j} לא יהיו יותר תדירים. בהתייחס לאיור 3, אם S_{21} יקבע ל -1, D'_{21} ו- D'_{41} יהפכו ל-0. (והתמיכה של פריט 1 תקטן); כאשר אם S_{12} יקבע ל -1, D'_{22} ו- D'_{42} יהפכו ל-0. (והתמיכה של פריט 2 תקטן).לכן התמיכה של {1, 2} יכולה לקטון על ידי קביעת S_{12} ו- S_{21} ל -1. יתר על כן, אם S_{ij} יקבע ל -1, עבור עמודה t שבה D_{ti} ו-D_{tj} שניהם 1, D'_{ij} יהפוך ל-0 (והתמיכה של פריט j תקטן).

$$\underset{D}{\begin{pmatrix}0&0&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}} \times \underset{S}{\begin{pmatrix}1&0&0\\-1&1&0\\0&0&1\end{pmatrix}_{3\times3}} = \underset{D'}{\begin{pmatrix}0&0&1\\0&1&0\\0&0&1\\0&1&1\end{pmatrix}_{4\times3}} \quad \text{and} \quad \underset{D}{\begin{pmatrix}0&0&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}} \times \underset{S}{\begin{pmatrix}1&-1&0\\0&1&0\\0&0&1\end{pmatrix}_{3\times3}} = \underset{D'}{\begin{pmatrix}0&0&1\\1&0&0\\0&0&1\\1&0&1\end{pmatrix}_{4\times3}}$$

Fig. 3. Effect of setting entries in S to -1.

3.2.2.3.3 *הבחנה2: קביעת כניסות ל 1*

קביעת כניסות ב- S ל- 1- יכולה להקטין את התמיכה של תבניות רגישות. אולם, הדבר יכול גם להסתיר תבניות שאינן רגישות ב D'. וזאת מכיוון שאין התי יחסות לקשר בין תבניות רגישות ולא-רגישות. למרבה המזל, ניתן לפתור את הבעיה על ידי קביעת הכניסות המתאימות ב- S ל-1. בהתייחס לאיור 4, תהי התמיכה המינימאלית 50%, ו- {1, 2} ו- {1, 3} תבניות רגישות ולא-רגישות בהתאמה. בהתייחס לשאילתא השמאלית באיור 4, {1, 2} ו- {1, 3} מוסתרים שניהם ב D'. אולם אם S_{31} תקבע ל-1 (כמו בשאילתא הימנית), לכניסות D_{t1} ו- D_{t3} שערכן 1, D'_{t1} תשמור את הערך של D_{t1}. לכן, קביעת S_{ij} ל-1 יכולה לשמר את הקשר בין הפריטים i ו-j על ידי חיזוק פריט j.

$$\underset{D}{\begin{pmatrix}1&0&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}} \times \underset{S}{\begin{pmatrix}1&0&0\\-1&1&0\\0&0&1\end{pmatrix}_{3\times3}} = \underset{D'}{\begin{pmatrix}1&0&1\\0&1&0\\0&0&1\\0&1&1\end{pmatrix}_{4\times3}} \quad \text{and} \quad \underset{D}{\begin{pmatrix}1&0&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}} \times \underset{S}{\begin{pmatrix}1&0&0\\-1&1&0\\1&0&1\end{pmatrix}_{3\times3}} = \underset{D'}{\begin{pmatrix}1&0&1\\0&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}}$$

Fig. 4. Effect of setting entries in S to 1.

3.3.3 תהליך הסינון

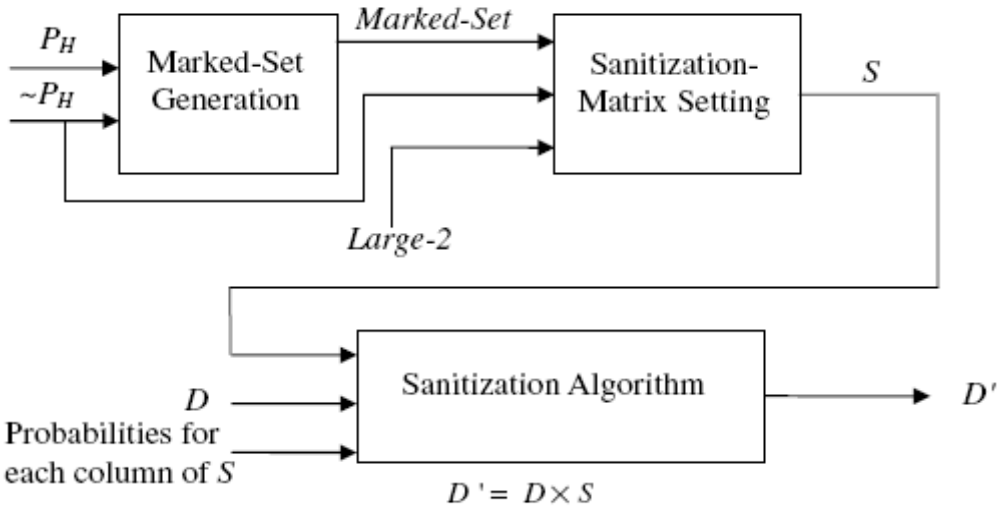


Fig. 5. Flowchart of the sanitization process.

איור 5 מציג את תרשים הזרימה של כל תהליך הסינון המוצע במאמר, כל שלב יוסבר בהמשך.

1.3.3.3 קביעת מטריצת הסינון

<u>הגדרה 1</u> – תהי F תבנית תדירה. ה- pair-subpattern (זוג-תת-תבנית), של F היא תת תבנית של F באורך 2. הקבוצה המכילה את כל זוגות-תת-התבנית של F נקראת the pair-subset של F.

כדי להתגונן בפני Forward-Inference Attacks, לכל תבנית רגישה P ב- P_H, לפחות תבנית אחת המשתייכת ל pair-subset של P צריכה להיות מוסתרת, או שהמתקיפים יוכלו להסיק באופן רקורסיבי את התבניות הנסתרות.

לדוגמא, אם {1,2,3,4} הוא סט פריטים תדיר, {1,2} הוא זוג-תת-תבנית של {1,2,3,4}. בנוסף, {{1,2},{1,3},{1,4},{2,3},{2,4},{3,4}} הוא ה- pair-subset של {1,2,3,4}. אם {1,2,3,4} הוא רגיש, התגוננות בפני Forward-Inference Attacks דורשת שלפחות אחד מהאלמנטים של {{1,2},{1,3},{1,4},{2,3},{2,4},{3,4}} יהיה חבוי. בגישה זו, קבוצה זמנית הנקראת קבוצת-הסימון, קבוצת-סימון, משמשת לאחסון זוגות-תת-התבנית-הקורבנות שהוא זוג-תת-תבנית שנבחר מתוך ה- pair-subset של התבנית הרגישה שיש להסתיר. לאחר קבלת קבוצת-הסימון, כל התבניות שבו משמשות לקבוע את הכניסות המתאימות ב- S ל 1-. האלגוריתם ליצירת קבוצת-הסימון מתואר באיור 6.

Marked-Set Generation
Input: $\sim P_H$ and P_H
Output: Marked-Set
Initial: Marked-Set = ϕ

1.$\forall$ $P \in P_H$ do
    If (the length of P is equal to two)
        Position P into Marked-Set.
2.$\forall$ Remaining $P \in P_H$ do
    If (P has no pair-subpatterns included in Marked-Set)
        Generate k groups. k equals the number of pair-subpatterns of P. A class label of each group is named by each pair-subpattern of P. P is stored in each group.
3. Merge the groups that have the same class label.
4. $\forall$ $NP \in \sim P_H$ do
    Generate all pair-subpatterns of NP. Count the frequency of each pair-subpattern in $\sim P_H$.
5. $\forall$ Groups do
    Set the frequency of the group to the frequency of the pair-subpattern that equals the class label of the group.
6. Sort the groups into an increasing order of frequency.
7. For i = 1 to the number of groups − 1
    For $j = i + 1$ to the number of groups
        Compare two groups G_i, G_j.
            If the number of patterns stored in G_i is not equal to the number of patterns stored in G_j
                Remove all the patterns both stored in G_i and G_j from the smaller one.
            Else if the frequency of G_i is not equal to the frequency of G_j
                Remove all the patterns both stored in G_i and G_j from the larger one.
            Else
                Remove all the patterns both stored in G_i and G_j from a randomly chosen group.
8. $\forall$ Groups do
    If the number of patterns stored in a group is more than 0, position its class label into Marked-Set.

Fig. 6. Marked-Set generation.

בהתייחס לאיור 6, בשלב 1 התבניות הרגישות באורך 2 מנותבות ישירות ל קבוצת-הסימון מכיוון שה- זוגות-תת-התבנית-הקורבנות זה הם עצמם . בשלב 2, התבניות שנותרו שה זוגות-תת-תבנית לא נכללות בקבוצת-הסימון משמשות ליצירת קבוצות למציאת זוגות-תת-התבנית-הקורבנות המתאימים. מכיוון שהתבניות השייכות ל $\sim P_H$ צריכות להיות מושפעות מעט ככל האפשר , יש להתחשב בהן כאשר בוחרים את זוגות-תת-התבנית-הקורבנות. התדירויות של זוגות-תת-התבנית של התבניות הלא- רגישות מחושבות בשלב 4.

לדוגמא, אם $\sim P_H$={{1,2,3},{1,3,5}} זוגות-תת-תבניות של {1,2,3} ושל {1,3,5} הם {1,2},{1,3},{2,3} ו- {1,3},{1,5},{3,5} בהתאמה. התדירויות הן 1,1,1,1 2, בהתאמה.

אם תדירות של זוג-תת-תבנית שנוצר מ- $\sim P_H$ נמוך, מעט התבניות שב- $\sim P_H$ יושפעו על ידי הסתרתו. לכן, בשלב 7, זוגות-תת-התבניות בעלי התדירות הנמוכה נבחרים להיות מוצבים ב קבוצת-הסימון. שלב 6 ממיין את הקבוצות בסדר עולה של תדירויות כדי לסווג את התבניות בקבוצות עם תדירויות נמוכות. לכן, מעט התבניות ב $\sim P_H$ יכולות להיות מושפעות. בשלבים 3-7, הקבוצות ממוזגות בהתאם לתבניות המאוחסנות בתוכן , כך שהתבניות הרגישות יכולות להיות מוסתרות בזמנית על ידי הסתרת זוג-תת-תבנית השכיח. יתר על כן , השונות בין D ו D' יכולה גם לקטון . דוגמא לאלגוריתם ליצירת קבוצת-הסימון ניתנת בטבלאות 1-3.

Table 1
An example of setting the sanitization matrix

P_H	$\sim P_H$
{3, 5, 6, 7} {4, 5, 7} {3, 8} {2, 5} {1, 5} {1, 2} {4, 6} {4, 6, 7}	{3, 5, 6} {3, 5, 7} {3, 6, 7} {5, 6, 7} {1, 7} {1, 3} {1, 6} {2, 6} {2, 7} {3, 5} {3, 6} {3, 7} {4, 5} {4, 7} {5, 6} {5, 7} {5, 8} {6, 7} {6, 8} {7, 8}

The items of D are {1,2…,8}. Initially, P is decomposed into P_H and $\sim P_H$.

Table 2
An example of Marked-Set generation: Part 1

Group

{3, 5}	3	{3, 6}	3	{3, 7}	3	{5, 6}	3	{5, 7}	3
{3, 5, 6, 7}		{3, 5, 6, 7}		{3, 5, 6, 7}		{3, 5, 6, 7}		{3, 5, 6, 7} {4, 5, 7}	

Class label
Frequency
Patterns stored in the group

{6, 7}	3	{4, 5}	1	{4, 7}	1
{3, 5, 6, 7}		{4, 5, 7}		{4, 5, 7}	

After processing Steps 1 to 5 of the algorithm for generating Marked-Set, it contains the following pair-subpatterns: {3, 8} {2, 5} {1, 5} {1, 2} {4, 6}.

Table 3
An example of Marked-Set generation: Part 2

Group

{4, 5}	1	{4, 7}	1	{3, 5}	3	{3, 6}	3

{3, 7}	3	{5, 6}	3	{5, 7}	3	{6, 7}	3
				{3, 5, 6, 7} {4, 5, 7}			

After processing the entire algorithm, Marked-Set contains the following pair-subpatterns: {3, 8} {2, 5} {1, 5} {1, 2} {4, 6} {5, 7}.

לאחר קבלת קבוצת-הסימון, ניתן לקבוע את מטריצת הסינון על ידי שימוש באלגוריתם המוצג באיור 7. בשלב 2 של האלגוריתם, מכיוון שכל התב ניות {i,j} המאוחסנות ב קבוצת-הסימון צריכות להיות מוסתרות, הפריט בעל ההשפעה הנמוכה ביותר על התבניות שב $\sim P_H$ ייבחר להיות הפריט הקורבן שהתמיכה בו מצומצמת להסתיר את התבנית המאוחסנת ב קבוצת-הסימון. אם ההשפעה על התבניות ב- $\sim P_H$ הן זהות, יבחר פריט קורבן בהתאם למספר המופעים שלו ב קבוצת-הסימון. זאת מכיוון שבחירת פריטים תדירים יותר יכולה להפחית את התמיכה של יותר תבניות המשתייכות לקבוצת-הסימון וכן להפחית את ההבדלים בין D ובין D'. בשלב 3, חוקי ההקשר של התבניות הקשורות ל $\sim P_H$ מוגברים על ידי קביעת הכניסות המתאימות של S ל-1. למרות שכל כניסה במטריצת ההכפלה תלויה בתהליך, הקשרים בין התבניות הרגישות והלא- רגישות נשמרים על ידי יצירת קבוצת-הסימון והאלגוריתם ליצירת מטריצת הסינון. דוגמא ליישום שיטת קביעת מטריצת הסינון מוצגת בטבלה 4.

Sanitization-Matrix Setting
Input: Marked-Set, $\sim P_H$, large-2 (the set of frequent patterns with length equaling two)
Output: Sanitization Matrix

1. $\forall i, 1 \le i \le m$ do
   Set S_{ii} to 1 // diagonal entries
2. $\forall \{i, j\} \in$ Marked-Set do
   If the number of patterns containing i in $\sim P_H <$ the number of patterns containing j in $\sim P_H$
      Set S_{ji} to –1. // Decreasing the support of item ***i***
   Else if the number of patterns containing i in $\sim P_H >$ the number of patterns containing j in $\sim P_H$
      Set S_{ij} to –1. // Decreasing the support of item ***j***
   Else
      If the number of patterns containing i in Marked-Set > the number of patterns containing j in Marked-Set
         Set S_{ji} to –1. // Decreasing the support of item ***i***
      Else if the number of patterns containing i in Marked-Set < the number of patterns containing j in Marked-Set
         Set S_{ij} to –1. // Decreasing the support of item ***j***
      Else
         Set S_{ij} or S_{ji} to –1 randomly.
3. $\forall \{i, j\} \in \{large\text{-}2 - Marked\text{-}Set\}$ do
   Set S_{ij} and S_{ji} to 1. // simultaneously enhance the strengths of i and j
4. $S_{ij} = 0$, otherwise.

Fig. 7. Setting the sanitization matrix.

Marked-Set	*Large-2 – Marked-Set*	$\sim P_H$
{3, 8} {2, 5} {1, 5} {1, 2} {4, 6} {5, 7}	{1, 3} {1, 6} {1, 7} {2, 6} {2, 7} {3, 5} {3, 6} {3, 7} {4, 5} {4, 7} {5, 6} {5, 8} {6, 7} {6, 8} {7, 8}	{3, 5, 6} {3, 5, 7} {3, 6, 7} {5, 6, 7} {1, 7} {1, 3} {1, 6} {2, 6} {2, 7}{3, 5} {3, 6} {3, 7} {4, 5} {4, 7} {5, 6}{5, 7} {5, 8} {6, 7} {6, 8} {7, 8}

$$S = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ -1 & -1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}_{8\times 8}$$

S_{38} ← (row 3) S_{51} → (row 5)

Table 4
Setting the sanitization matrix

Probability policies 2.3.3.3

מדיניות הסתברותית נקבעה כד י להתגבר על שתי בעיות , האחת, הסתרת יתר (overhiding problem), והשנייה בעיית מיעוט הסתרה (underhiding problem). המחזקות את יכולת התוקפים לנחש את מטריצת הסינון . תפקיד מדיניות הסתברותית זו הוא לחזק את מידת הביטחון שמטריצת הסינון מספקת.

Distortion Probability ρ 3.3.3.2.1

קביעת הכניסות במטריצת הסינון ל -1 יכולה להפחית את התמיכה של זוגות-תת-תבניות המשתייכים ל P_H. איור 8 מציין שתוצאות אלו יכולות לגרום לבעי ה הנקראת overhiding problem, הסתרת יתר. משני איברי המשוואה באיור 8, התמיכה של {1,2} ב 'D היא 0. זאת מכיוון שפריט 1 ו-2 לא מופיעים יחד כלל. הם mutually exclusive! מצב זה כמעט ולא קורה בבסיס נתונים שאינו מסתיר מידע. התוקפים יכולים לאתר מצב זה ולהסיק ש {1,2} הוסתרו בקפדנות.

$$\underbrace{\begin{pmatrix}0&0&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}}_{D}\times\underbrace{\begin{pmatrix}1&0&0\\-1&1&0\\0&0&1\end{pmatrix}_{3\times3}}_{S}=\underbrace{\begin{pmatrix}0&0&1\\0&1&0\\0&0&1\\0&1&1\end{pmatrix}_{4\times3}}_{D'}\quad\text{and}\quad\underbrace{\begin{pmatrix}0&0&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}_{4\times3}}_{D}\times\underbrace{\begin{pmatrix}1&-1&0\\0&1&0\\0&0&1\end{pmatrix}_{3\times3}}_{S}=\underbrace{\begin{pmatrix}0&0&1\\1&0&0\\0&0&1\\1&0&1\end{pmatrix}_{4\times3}}_{D'}$$

Fig. 8. Overhiding problem occurs while setting entries in S to -1.

במציאות, כדי להחביא תבניות רגישות, יש להקטין את התמיכה אל מתחת לתמיכה המינימאלית, כך שאין צורך להקטין את התמיכה שלהם ל- 0. לכן, ההסתברות ρ הנקראת Distortion Probability, נועדה לפתור את בעיית הסתרת יתר. ה-Distortion Probability משמשת כאשר עמודה j של S מכילה רק כניסה אחת שערכה 1 (כלומר $S_{jj}=1$) וזה פועל באופן הבא :

$\forall i,j \quad 1\le i\le n, 1\le j\le m \;,\; \sum_{k=1}^{m} D_{ik}\cdot S_{kj}\le 0$, ההסתברות של D'_{ij} להיות 1 או 0 היא ρ_j ו $1-\rho_j$ בהתאמה.

<u>למה 1</u> – יהי פרמטר σ התמיכה המינימאלית, והפרמטר c רמת האבטחה. תהי {i,j} תבנית בקבוצת-הסימון ו n_{ij} רמת התמיכה של {i,j}. ρ הוא ה- Distortion Probability של עמודה j. ללא הגבלת הכלליות מניחים כי $S_{ij}=-1$. אם ρ מספק $n_{ij}\times\rho<\sigma\times|D|$ וגם $\sum_{x=0}^{\lceil\sigma\times|D|\rceil-1}\binom{n_{ij}}{x}\rho^{x}(1-\rho)^{n_{ij}-x}\ge c$, כאשר |D| הוא מספר הטרנזקציות ב-D, ניתן להבטיח בהסתברות c ש- {i,j} הוא לא תדיר ב 'D. (ההוכחה מובאת במאמר המקורי).

בנוסף, אם מספר כניסות בעמודה j של S שוות ל -1, כך ש- $S_{ij}=-1$, $S_{kj}=-1$ ו- $S_{mj}=-1$, מספר מועמדים Distortion Probability (כלומר ρ_i, ρ_k, ρ_m) מתקבלים. ה-Distortion Probability ρ_j של עמודה j, נקבעת למועמד Distortion Probability המינימאלי כדי להבטיח בהסתברות לפחות c שכל הזוגות-תת-תבניות מוחבאים ב-'D.

Conformity Probability μ 3.3.3.2.2

קביעת הכניסות במטריצת הסינון ל 1 מגביר את ההקשר של תבניות לא רג ישות שיכול להביא לתוצאה שמספר תבניות רגישות לא יוסתרו . הבעיה נקראת underhiding problem, בעיית מיעוט הסתרה. בהתייחס לאיור 9 תהי התמיכה המינימאלית 50%, {1,2} תבניות רגישה, ו- {1,3}, {2,3} הן תבניות לא רגישות . S_{23}, S_{31}, S_{13} ו- S_{32} נקבעו ל-1 להגביר את הקשר של {2,3}, {1,3}. הדבר גרם לכישלון בהסתרת {1,2}. כדי להימנע ממצב כזה, משתמשים ב- Conformity Probability כאשר עמודה j של S מכילה לפחות שתי כניסות השוות ל- 1. זה עובד באופן הבא : אם $\sum_{k=1}^{m} D_{ik}\cdot S_{kj}\ge 1$ ולפחות כניסה אחת בעמודה j השווה ל -1- מוכפלת בכניסה ע"י כניסה עם ערך 1 ב D, אזי D'_{ij} יקבע ל-1 בהסתברות μ_j ויקבע ל-0 בהסתברות $1-\mu_j$. יש להבחין שכאשר כל הכניסות השוות ל -1- בעמודה j

של S מוכפלות בדיוק בכל הכניסות בעלות הערך 0 בשורה i של D, הכוונה היא ששורה I של D אינה מכילה אף כניסה המשפיעה על עמודה j ב S, ורק חוקי מטריצת המכפלה שנידונו בפרק 2 נחוצים.

$$\underbrace{\begin{pmatrix}1&1&1\\1&1&0\\0&0&1\\1&1&1\end{pmatrix}}_{4\times3}_{D} \times \underbrace{\begin{pmatrix}1&-1&1\\0&1&1\\1&1&1\end{pmatrix}}_{3\times3}_{S} = \underbrace{\begin{pmatrix}1&1&1\\1&0&0\\0&0&1\\1&1&1\end{pmatrix}}_{4\times3}_{D'}$$

Fig. 9. Underhiding problem while setting entries in S to 1.

<u>למה 2</u> – יהי פרמטר σ התמיכה המינימאלית ופרמטר c רמת הביטחון (confidence). יהי {i,j} תבנית בקבוצת-הסימון, {k,j} תבנית ב- {large-2 – Marked-set}, ו- n_{ikj} רמת התמיכה של {i,k,j}. ללא הגבלת כלליות, נניח כי S_{ij}=-1. μ הוא ה- Conformity Probability של עמודה j. אם μ מתאים לכלל הבא:

$$\mu \begin{cases} =1, & n_{ikj} < \lceil \sigma \times |D| \rceil \\ \text{should satisfy } n_{ikj} \times \mu < \sigma \times |D| \text{ and } \sum_{y=0}^{\lceil \sigma \times |D| \rceil - 1} \binom{n_{ikj}}{y} \mu^{y} (1-\mu)^{n_{ikj}-y} \geqslant c, & \text{otherwise} \end{cases}$$

אזי ניתן להבטיח בהסתברות c ש- {i,j} אינם תדירים ב D'. (ההוכחה מובאת במאמר המקורי).

בנוסף, אם מספר כניסות בעמודה j של S שוות ל -1, ניתן לקבל מספר ערכים עבור מועמדי ה Conformity Probability μ. על ידי הקביעה של ה- Conformity Probability של j, μ_j, לערך המינימאלי של המועמד, ניתן להבטיח שלפחות בסבירות c שכל הזוגות-תת-תבניות מוחבאים ב-D'.

Discussion on Distortion probability and Conformity Probability 3.3.3.2.3

מבלי להתחשב במדיניות ההסתברות, בעיית הסתרת יתר (overhiding problem) הגורמת לזוגות-תת-תבניות נסתרות להתגלות עלולה להתרחש. ניתן להימנע מכך אם ניקח בחשבון את ה- Probability Distortion. בנוסף, במצבים של בעיית מיעוט הסתרה (underhiding problem), ה- Conformity Probability יכול לסייע במניעת בעיית הכישלון בהסתרה, hiding failure problems הנגרמת כתוצאה מקביעת 1 באלמנטים שמחוץ לאלכסון במטריצת הסינון. שתי הסתברויות אלו מחזקות את מידת הביטחו ן של מטריצת הסינון כך שהמנסים לנחש את מט ריצת הסינון נתקלים בחומה אטומה. למרות שמטרות שתי ההסתברויות נראות מנוגדות, הרי שבמציאות שתיהן משמשות בתנאים שונים של העמודות של מטריצת הסינון. כך, שכל עמודה במטריצת הסינון S יכולה להכיל לכל היותר הסתברות אחת (או Distortion probability או Conformity Probability), מכיוון שעמודה מספקת רק אחד מהתנאים שנידונו לעיל. בנוסף, בהתאם לרמת השוני בין בסיס הנתונים המסונן ובסיס הנתונים המקורי ומידת הגמישות של הכישלון בהסתרות, המשתמשים יכולים להגדיר לעצמם את רמת הביטחון כדי לאפנן את מדיניות ההסתברויות ולמצוא כזו המתאימה לצרכיהם.

Sanitization algorithm 3.3.3.3

בהתאם למדיניות ההסתברויות שנידונה בסעיף קודם , אלגוריתם הסינון מספק $D'_{n\times m} = D_{n\times m} \times S_{m\times m}$ כפי שניתן לראות באיור 10. על פי האלגוריתם, ערך 0 בבסיס הנתונים המקורי מועתק מיידית לבסיס הנתונים המסונן, אחרת, האלגוריתם שוקל על פי ה ערכים בעמודה המתאימה במטריצת הסינון ועל פי הערכים ההסתברותיים שהתקבלו , מה יהיה הערך שיקבע בבסיס הנתונים המסונן. יש לשים לב שאם תהליך הסינון גורם לכל הכניסות בשורה i של D' להיות 0, כניסה k כלשהי שבו D_{ik}=1 נבחר רנדומאלית ו D'_{ik} נקבע ל-1. מהלך זה מבטיח ש D' ו D יהיו שווים בגדלם.

Sanitization Algorithm
Input: $D_{n\times m}$, $S_{m\times m}$, **probabilities for each column of** S
Output: $D'_{n\times m}$

For $i = 1$ to n // i is a row of D
  For $j = 1$ to m // j is a column of S
    If ($D_{ij} = 0$)
      $D'_{ij} = 0$.
    Else
      $\text{temp} = \sum_{k=1}^{m} D_{ik} \cdot S_{kj}$.
      If (j in S contains some entries = –1, only $S_{jj} = 1$ and temp ≤ 0)
        D'_{ij} is set to 1 with probability ρ_j and 0 with probability $1-\rho_j$.
      Else if (j in S contains some entries = –1, more than one entry = 1, temp ≥ 1 and at least one entry in j with value –1 is multiplied by the entry in D with value 1)
        D'_{ij} is set to 1 with probability μ_j and 0 with probability $1-\mu_j$.
      Else if (temp ≤ 0)
        D'_{ij} is set to 0.
      Else if (temp ≥ 1)
        D'_{ij} is set to 1.
  If (all entries in row i of D' are 0)
    Randomly choose an entry from some k where $D_{ik} = 1$ and set D'_{ik} to 1.

Fig. 10. Sanitization algorithm.

3.3.4 ניתוח הסיבוכיות ומידת האבטחה

3.3.4.1 סיבוכיות הזמן

ניתוח סיבוכיות הזמנים מתרכז בתהליך הפיכת בסיס הנתונים המקורי לבסס הנתונים המסונן ומתעלם מזמני התהליך למציאת תבניות תדירות.

<u>תאורמה 1</u> – זמן הריצה של תהליך הסינון הוא $O\left(n_{item}^2 n_{trans} + n_{p_H} n_{\sim p_H} L^4 + n_{p_H}^4 L^4 + n_2\right)$ כאשר n_{item} הוא מספר הפריטים ב- D, n_{terms} הוא מספר הטרנזקציות ב- D, n_{p_H} הוא מספר התבניות הרגישות, $n_{\sim p_H}$ הוא מספר התבניות שאינן רגישות, n_2 הוא מספר התבניות התדירות באורך 2, ו L הוא האורך המקסימאלי של התבניות התדירות. (ההוכחה מובאת במאמר המקורי).

3.3.4.2 מידת האבטחה

<u>למה 3</u> – בהינתן בסיס נתונים מסונן D', תבניות רגישות מוסתרות באורך r יכולות להתגלות לכל היותר בהסתברות $p^{2^{r-2}}$, כאשר p היא ההסתברות שהמתקיפים ינחשו שתבנית לא תדירה ב D' היא רגישה ב D.

<u>הוכחה</u> – מכיוון שהתבניות הרגישות מוסתרות על ידי הסתרת ה זוגות-תת-תבניות שלהן, נניח שהמתקיפים ינחשו שתבנית מסוימת באורך 2 היא hidden victim pattern בהסתברות p. נניח גם שכל ה immediate subpatterns של התבניות הרגשות המוסתרות יופיעו , המתקיפים ינחשו גם ש זוהי תבנית רגישה מוסתרת בהסתברות p. ה- immediate subpatterns של תבנית e באורך r הן תת התבניות של e באורך r-1.

נניח שתבנית רגישה e באורך r מוסתרת רק על ידי הסתרת אחד מה זוגות-תת-תבניות, *sube*, שישוחזר בהסתברות p. בנוסף, מכיוון שגילוי *sube* ישפיע על $\binom{r-2}{1}$ תת התבניות של e באורך 3,

ההסתברות לשחזור כל תת התבניות של e באורך 3 היא $p \times p\binom{r-2}{1}$. מכיוון שגילוי *sube* ישפיע על $\binom{r-2}{2}$ תת התבניות של e באורך 4, ההסתברות לשחזור כל תת התבניות של e באורך 4 היא $p \times p\binom{r-2}{1} \times p\binom{r-2}{2}$. מכאן, ניתן להסיק כי ההסתברות לשחזור e היא $p \times p\binom{r-2}{1} \times p\binom{r-2}{2} \times p\binom{r-2}{3} \times \ldots \times p\binom{r-2}{r-2} = p^{2^{r-2}}$. התנאי שלעיל הוא המצב הטוב ביותר עבור המתקיפים, מכיוון שרק זוג-תת-תבנית אחד של e מוסתר. בכל מקרה, עקב ההסתרה של תבניות רגישות אחרות, ל- e כנראה יותר מ זוג-תת-תבנית אחד מוסתר, מה שמקטין את ההסתברות לגילוי e.□

באופן כללי, ההסתברות לשחזור תבניות רגישות משתנה בהתאם ל- p שניתן. אם p קטן, קשה למצוא את התבניות הרגישות הארוכות. במציאות, ככל ש p גדול יותר, כך המתחרה יכול למצוא יותר תבניות רגישות. בכל מקרה, p גבוה יכול לגרום ליותר תבניות שאינן תדירות ב D להתגלות, וכך להקטין את המשמעותיות של התבניות הרגישות. במקרה הקיצוני, אם המתחרה רוצה לגלות את כל התבניות הרגישות, p יקבע ל- 1. מה שעלול לגרום לכל הצרופים האפשריים של העצמים להתגלות, ובכך להקשות על מציאת התבניות שהן אכן הרגישות.

שיטה אחרת להתקפה היא ניחוש מטריצת הסינון. אם המתקיפים משיגים את בסיס הנתונים המשוחזר 'D שמפורסם, הם יכולים בוודאות להשיג את כל התבניות התדירות שנכרו מ 'D. מכיוון ש 'D שונה, ומספר תבניות שאינן רגישות הושפעו כתוצאה מבצוע תהליך הסינון, התבניות שנכרו הן רק חלק מהתבניות הלא רגשיות שב D. המטריצה S מוגנת היטב בגלל תהליך האקראיות, כלומר, תהליך סינון עם המדיניות ההסתברותית המונעת את כל הבעיות האפשריות של חשיפת אלמנטים מסוימים ב קבוצת-הסימון. נניח שמספר התבניות התדירות באורך 2 ב 'D הוא n'_2. מכיוון שבהליך קביעת 1 במטריצה S הוא למקם 1 בשתי כניסות סימטריות, לפיכך, קימות $n^2_{item} - n_{item} - 2 \times n'_2$ כניסות לא ידועות ב S. בנוסף, בהתאים לתהליך קביעת S, מצבים אפשריים של ערכים של שתי כניסות סימטריות $\forall i,j \left(S_{ij}, S_{ji}\right)$ ב S יכולים להיות מסווגים ל- 4 סוגים : (1,1), (-1,0), (0,-1) ו- (0,0). לכן, ניתן לנחש כל שתי כניסות סימטריות בהסתברות של 1/4. כפי שניתן לראות, מספר הכניסות הסימטריות ב- S הוא $\frac{n^2_{item} - n_{item} - 2 \times n'_2}{2}$ מה שמקשה על ניחוש S. הכוונה שהמתקיפים יכולים להשיג את S האמיתית בהסתברות של $1/4^{\frac{n^2_{item} - n_{item} - 2 \times n'_2}{2}}$. גם אם המתקיפים ישיגו את S האמיתית, עדיין יהיה להם קושי להשיג את התבניות הרגישות. מכיוון שהתמיכה של התבניות הוא מעין מדד סטטיסטי, להיות תבנית רגישה מחייב תמיכה מספקת המסתמכת על נתונים משמעותיים. מדיניות ההסתברות מספקת הגנה מפני שחזור D המקורי, גורמת למידת התמיכה של התבניות הרגישות להי ות לא משמעותית. הבעיה המטרידה יותר למתקיפים היא כיצד יזהו את הניחוש הנכון של S כ- S האמיתית. כך, שהם אינם יכולים לדעת אם S שניחשו נכונה או לא אפילו אם ניחשו נכון. בנוסף, תהליך הסינון מסתיר את התבניות הרגישות על ידי הסתרת ה pair-patterns שלהם המאוחסן בקבוצת-הסימון בביטחון של c, מאפיין זה של תהליך הסינון מספק את הדרוש כדי להימנע מה- Forward-Inference Attacks.

3.3.5 הערכת הביצועים

שלוש טעויות יכולות להתרחש בבעיית הסתרת תבניות רגישות תחת תמיכה מינימאלית קבועה. הראשונה, חלק מהתבניות הרגישות לא יוסתרו בהצלחה. כלומר חלק מהתבניות הרגישות עדיין יכולות להכרות מ 'D. השנייה, חלק מהתבניות הלא- רגישות לא יוכלו להכרות מ 'D. והשלישית, תבניות מלאכותיות יכולות להווצר לאחר המרת D ל-'D. בנוסף קיימים שני קריטריונים, שונות, dissimilarity, וחולשה, weakness. הקריטריונים נידונים להלן:

קריטריון 1: מספר תבניות רגישות עדיין תדירות ב 'D. מצב זה נקרא Hiding Failure[42] והוא נמדד על ידי $HF = \frac{|P_H(D')|}{|P_H(D)|}$, כאשר $|P_H(X)|$ מייצג את מספר התבניות המוכלות ב P_H שנכרו מבסיס הנתונים X.

קריטריון 2: מספר תבניות שאינן רגישות מוסתרות ב 'D. מצב זה נקרא Misses Cost[42] והוא נמדד על ידי $HMC\frac{|\sim P_H(D)|-|\sim P_H(D')|}{|\sim P_H(D)|}$, כאשר $|\sim P_H(X)|$ מייצג את מספר התבניות המוכלות ב $\sim P_H$ שנכרו מבסיס הנתונים X.

קריטריון 3: חלק מהתבניות המלאכותיות נוצרו לאחר תהליך הסינון. הטעות מכונה Artificial Patterns[2] והיא נמדדת על ידי $AP = \frac{|P(D')|-|P(D)\cap P(D')|}{|P(D')|}$, כאשר |P(X)| מייצג את מספר התבניות שנכרו מבסיס הנתונים X. AP בגישה הנוכחית ו SWA (Sliding Window Algorithm) (שיטה שתוצג בסעיף 3.2 בהמשך העבודה) הם תמיד 0%, מכיוון ששני תהליכי סינון אלו מנסים למחוק את הנתונים המקוריים בטרנזקציות מבלי להכניס נתונים אחרים במקומם.

קריטריון 4: השונות בין בסיס הנתונים המקורי ובסיס הנתונים המסונן נמדדת על ידי $Dis = \left(\sum_{i=1}^{n}\sum_{j=1}^{m}(D_{ij} - D'_{ij})\right) / \left(\sum_{i=1}^{n}\sum_{j=1}^{m}D_{ij}\right)$.

קריטריון 5: בהתאם לסעיף הקודם, ה- Forward-Inference Attacks נמנע כל זמן שלפחות זוג-תת-תבנית אחד של ה תבנית הרגישה מוסתר . לכן, אם כל ה זוגות-תת-תבניות של תבניות רגישות נכרים ב-'D אבל התבניות הרגישות מוסתרות ב 'D, בעיית ה- Forward-Inference Attacks תתרחש. לכן, ה- Forward-Inference Attacks יכולה להימדד על ידי Weakness $(WK) = \frac{|(P_H(D)-P_H(D'))\cap PairS(D')|}{|P_H(D)-P_H(D')|}$, כאשר PairS(D') הוא הסט של התבניות הרגישות שתתי הזוגות שלהן יכול להכרת במלואו על ידי תהליך הכרייה ב'D.

במציאות, הערכים הכמותיים של הקריטריונים שלעיל יכולים להתנגש . למשל, הערכים הנמוכים של HF יכולים לגרום לערכי MC גבוהים. וזאת כתוצאה של אסוציאציות בין התבניות הרגישות והלא-רגישות. דוגמא נוספת, הערכים הנמוכים של HF יכולים לגרום לערכי Dis גבוהים. וזאת מכיוון שבסיס הנתונים המקורי שונה באופן ניכר כדי להגן על התבניות הרגישות.

בהתאם לקריטריונים שלעיל המשתמש יכול להגדיר את רמת הביטחון כדי לאפנן את מדיניות ההסתברות ולמצוא את האיזון המתאים עבורו.

במאמר הוצגו שתי סדרות של ניסויים . הראשונה היא לחקור את האיזון בין רמת האבטחה והקריטריונים האחרים בתהליך הסינון . השנייה, להשוות בין גישה זו לגישה של SWA[43,44] שעד כה הציגה את האלגוריתם בעל הביצועים הטובים ביותר.

התוצאות הראו את השפעת רמת האבטחה c על ה- hiding failure, miss cost, weakness, וה- dissimilarity. של תהליך הסינון. ככל ש c גדול יותר, כך גדלות ההסתברויות של ה- distortion וה- conformity. כתוצאה, יותר כניסות ב D נבחרו להשתנות. לכן, הכשלון בהסתרה יורד ככל ש c עולה. בנוסף, ה- miss cost והשונות עולים ככל ש c עולה.

3.3.6 מסקנות

המאמר מציג תהליך סינון חדשני לשיפור היחס בין הגנה על מידע רגיש וגילוי תבניות תדירות . בסיס נתונים מסונן מתקבל על ידי קביעת הכניסות במטריצת הסינון לערכים מתאימים והכפלת בסיס הנתונים המקורי במטריצת הסינון עם מדיניות הסתב רות מוגדרת. בנוסף, טכניקה חדשנית זו יכולה להגן בצורה מוחלטת בפני Forward Inference Attacks כאשר רמת האבטחה – הנשלטת על ידי מנהל בסיס הנתונים – היא בערך 1. תוצאות הניסויים מראים שלמרות שהעלות של ה misses והשונות בין בסיס הנתונים המקורי ובין בסיס הנתונים המס ונן המתקבל מתהליך סינון זה גדול במעט מאשר אלו של SWA תחת אותם התנאים, הוא בטוח יותר מאשר ה SWA. שלא כמו ב SWA, תהליך סינון זה מחוסן

מפני ה- Forward Inference Attacks. ועוד, מכיוון שמדיניות ההסתברות בגישה הזו גם מתחשבת בתמיכה המינימאלית, המשתמש צריך רק להח ליט מהי רמת האבטחה מבלי להתחשב ביחס בין סף החשיפה והתמיכה המינימאלית.

3.4 סיכום

פרק זה עסק בהיבט שונה של פרטיות המידע בכריית מידע בין מספר שותפים. הפרק התמקד בצורך להסתיר חלק מהמידע המופק בתהליך הכרייה והציג מספר שיטות לסינון המידע החסוי. הבעיות העקרוניות הקיימות בהסתרת חוקי מידע חסויים שנכרו הוצגו בתחילת הפרק והראו שהמשימה אינה קלה, תוקפים פוטנציאליים עלולים לנצל את המידע הגלוי כדי לגלות חוקי הקשר פרטיים שהוסתרו, וכן, ייתכן שהמידע הגלוי אינו שלם ויכול להכיל נתונים מוספים שגויים, חלק מהמידע החסוי או חוסר במידע גלוי.

השיטות העוסקות במלאכה ניסו ליצור איזון בין איכות הנתונים המפורסמים לבין אבטחת הפרטיות, לא תמיד בהצלחה מרובה כפי שניתן להבין לאחר עיון במאמר השני 3.3 שפורסם בפרק.

כדי להבין את גודל הקושי, פתח הפרק בדיון בבעיות מהותיות אלו, ורק אח"כ התייחס לשיטות הסינון המובאות במאמרים. במאמר הראשון מובאות שלוש יוריסטיקות שונות ל סינון ול שמירת הפרטיות בכריית מידע, ובהן ה- SWA. קיימת השוואה בין השיטות השונות וכן הערכות ביצועים של כל אחת מהן. המאמר מתייחס בקצרה לבעיית ניחוש המידע, מתוך מטרה להתמודד עם הבעיה. המאמר מפורט מאד ומעמיק אך למעשה לא מציע פתרון טוב דיו ל בעיית ניחוש המידע. המאמר השני, מאמר חדשני, מביא שיטה שונה הפותרת את הבעיה של היכולת לנחש מידע מהמידע המופק. המאמר מתייחס לשיטת ה SWA שהוצגה במאמר הראשון לצורך השוואה עם השיטה שלו, מתוך הבנה כי שיטה זו הי יתה היעילה ביותר שפורסמה עד כה, מבחינת אמינות הנתונים המסוננים וקירבתם לנתונים המקוריים. השיטה שהציג משופרת ופתרה את בעיית ניחוש המידע תוך שמירה על יעילות האלגוריתם ומבלי לפגוע בהיבטים האחרים.

כל השיטות שהוצגו הראו את הצורך לשמור על איכות המידע המתקבל תוך כוונון מידת האבטחה הרצויה. היכולת לשלוט על האיזון בין הגורמים השונים ולקבל תוצאות אופטימאליות קובעות את איכות האלגוריתם.

סיכום

שימור פרטיות המידע בכריית מידע הוא נושא מרתק ה נמצא בעיצומו של המחקר על מגוון שיטותיו ונראה שעדיין לא הגיע לרוויה. הצורך בשיתוף מידע ובכריית מידע הולך וגובר ויחד איתו גובר גם הצורך באמצ עי אבטחה מתקדמים. התוקפים משתכללים יחד עם מפתחי האלגוריתמים להגנה על הנתונים, ונראה שכל נושא הקשור לאבטחה דורש השתכללות מתמדת.

חיפוש אחר מי מהנושאים הקשורים לשימור פרטיות המידע באבטחת מידע באינטרנט יחשוף אלפים רבים של מאמרים ומחקרים הקשורים בנושא אם בהצגת שיטות ואלגוריתמים אם במחקרים על היכולת לחשוף מידע ושיטות למדידת יכולת הגילוי ואם בדרכי התגוננות בפני תוקפים.

עבודה זו ביקשה לבחון את הנושא של שמירת פרטיות המידע בכריית מידע מבסיסי נתונים מזוויות שונות, דבר שהגביר את הצורך להעמיק בתחום, והביא לבחירת מאמרים שנועדו להציג מגוון שיטות וטכניקות למספר מצבי כרייה, ולהציג בעיות ופתרונות לחלק מהבעיות הקיימות. הפרק הראשון כולל מאמר שנבחר כדי לתת רקע כללי, להציג את סוגי הטכניקות הקיימות לשימור פרטיות המידע, תוך סיווגן לסוגים שונים, ולהראות את הכלים המשמשים להערכת האלגוריתמים. הפרק השני בוחן את שימור פרטיות המידע בכריית מידע מבסיסי נתונים מבוזרים, בהם כל אתר רוצה להסתיר את מידת התמיכה שלו בחוק הקשר מסוים, תוך שימוש בטכניקות לחישוב מאובטח בין מספר אתרים. שלושת המאמרים הראשונים המופיעים בפרק נבחרו כדי לבחון את הנושא משני היבטים עיקריים של שימור הפרטיות בחלוקה אופקית ובחלוקה אנכית של הנתונים. המאמר הרביעי העניק זווית שונה של כריית מידע בחלוקה אנכית של נתונים, והיא קבלת החלטות, תוך שמירה מקסימאלית על הפרטיות. המאמר החמישי נועד להציג טכניקה חדשנית המשתמשת ב- Bloom filters המאפשרת ל שרתי קצה לכרות נתונים מבסיס נתונים מרכזי, תוך שמירה על הפרטיות.

הפרק השלישי עוסק ב טכניקות לסינון חוקי הקשר ובבעיות הנלוות לטכניקות אלו. המאמר הראשון המופיע בפרק נבחר מכיוון שהוא מכיל שלוש גישות מעניינות לסינון חוקי הקשר, כולל שיטת ה SWA שנחשבת כשיטה יעילה ביותר מבחינת תועלתיות הנתונים שהיא מספקת, עד הופעת מטריצת הסינון, שיטה מעניינת וחכמה שהוצגה במאמר השני תוך ביצוע השוואה עם אלגוריתם ה SWA.

היכולת לכריית מידע והפקת חוקי הקשר מבסיסי נתונים היא רבת עוצמה ומכילה יתרונות עצומים לצדדים השותפים לכרייה, אך יוצרת סיכונים רבים הנובעים מגילוי מידע פרטי שאין הצדדים מוכנים לחשוף. ההגנה על הפרטיות מתבטאת הן בשמירה על פרטי ישויות אינדיבידואליו ת, הן בהסתרת חוקי הקשר חסויים, והן בהגנה על מידת התמיכה בחוקי הקשר תדירים של השותפים השונים בכרייה, ומהווה שיקול רב ערך בבחירת האלגוריתם המתאים. על אף המגוון הרחב של האלגוריתמים הקיימים נראה שעדיין יש מקום לשיפור היעילות של השיטות הקיימות. בעתיד, יש מקום להמשיך לחקור את יחסי הגומלין בין רמת האבטחה לבין ביצועי האלגור יתמים תוך התייחסות לערכי תמיכה שונים. ניתן להתמקד באפשרויות להקטין עוד את השונות בין תוצאות הכרייה לפני הסינון ואחריו וכן להקטין את חיסרון הנתונים תוך שמירה אופטימאלית על פרטיות הנתונים.

References:

1 Vassilios S. Verykios and Elisa Bertino and Igor Nai Fovino and Loredana Parasiliti Provenza and Yucel Saygin and Yannis Theodoridis, State-of-the-art in privacy preserving data mining, In Proceedings of the ACM SIGMOD,Volume 33 , Issue 1 Pages: 50 - 57 (March 2004)

2 Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, Disclosure Limitation of Sensitive Rules, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999), 45–52.

3 Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino, Hiding Association Rules by using Confidence and Support, In Proceedings of the 4th Information Hiding Workshop (2001), 369–383.

4 Vassilios S. Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin, and Dasseni Elena, Association Rule Hiding, IEEE Transactions on Knowledge and Data Engineering (2003). SIGMOD Record, Vol. 33, No. 1, March 2004.

5 LiWu Chang and Ira S. Moskowitz, An integrated framework for database inference and privacy protection, Data and Applications Security ,172–161 ,(2000) Kluwer, IFIP WG 11.3, The Netherlands.

6 Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, Privacy preserving association rule mining, In Proceedings of the 12th International Workshop on Research Issues in Data Engineering (2002), 151–158.

7 Yucel Saygin, Vassilios Verykios, and Chris Clifton, Using unknowns to prevent discovery of association rules, SIGMOD Record 30 (2001), no. 4, 45–54.

8 LiWu Chang and Ira S. Moskowitz, Parsimonious downgrading and decision trees applied to the inference problem, In Proceedings of the 1998 New Security Paradigms Workshop (1998), 82-89.

9 Ira S. Moskowitz and LiWu Chang, A decision theoretical based system for information downgrading, In Proceedings of the 5th Joint Conference on Information Sciences (2000).

10 Wenliang Du and Mikhail J. Attalah, Secure multi-problem computation problems and their applications: A review and open problems, Tech .Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001.

11 Chris Clifton, Murat Kantarcioglou, Xiadong Lin, and Michael Y. Zhu, Tools for privacy preserving distributed data mining, SIGKDD Explorations 4 (2002), no. 2.

12 Jaideep Vaidya and Chris Clifton, Privacy preserving association rule mining in vertically partitioned data, In the 8thACMSIGKDD International Conference on Knowledge Discovery and Data Mining (2002), 639–644.

13 Murat Kantarcioglou and Chris Clifton ,Privacy-preserving distributed mining of association rules on horizontally partitioned data, In Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2002), 24–31.

14 Wenliang Du and Zhijun Zhan, Building decision tree classifier on private data, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002).

15 Yehuda Lindell and Benny Pinkas, Privacy preserving data mining, In Advances in Cryptology -CRYPTO 2000 (2000), 36–54.

16 Rakesh Agrawal and Ramakrishnan Srikant ,Privacy-preserving data mining, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), 439–450.

17 Dakshi Agrawal and Charu C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms, In Proceedings of the 20th ACMSymposium on Principles of Database Systems (2001), 247–255.

18 Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, Privacy preserving mining of association rules, In Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining (2002).

19 Shariq J. Rizvi and Jayant R. Haritsa, Maintaing data privacy in association rule mining, In Proceedings of the 28th International Conference on Very Large Databases (2002).

20 Murat Kantarcioglu, Chris Clifton: Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE Trans. Knowl. Data Eng. 16(9): 1026-1037 (2004)

21 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.

22 D.W.-L. Cheung, V. Ng, A.W.-C. Fu, and Y. Fu: Efficient Mining of Association Rules in Distributed Databases. IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, pp. 911-922, Dec. 1996.

23 O. Goldreich: Secure Multiparty Computation. (working draft), Sept. 1998, available: http://www.wisdom.weizmann.ac.il/ ~oded/pp.html.

24 W. Du and M. J. Atallah. Secure multi-party computational geometry. In Proceedings of the Seventh International Workshop on Algorithms and Data Structures. Providence, Rhode Island, Aug. 8-10 2001.

25 J.C. Benaloh: Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret. Advances in Cryptography (CRYPTO86): Proc., A. Odlyzko, ed., pp. 251-260, 1986, available: http://spring erlink.metapress.com/openurl.asp ?genre=article&issn=03029%743&volume=263&spage=251.

26 Jaideep Vaidya, Chris Clifton: Privacy preserving association rule mining in vertically partitioned data. KDD 2002: 639-644

27 R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, Sept. 12-15 1994. VLDB.

28 Quinlan, J.R.: Induction of decision trees. Machine Learning 1 (1986) 81-106

29 Boris Rozenberg, Ehud Gudes: Association rules mining in vertically partitioned databases. Data Knowl. Eng. 59(2): 378-396 (2006)

30 J.Vaidya, C.Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proceedings of SIGKDD 2002, Edmonton, Alberta, Canada.

31 Y.Elovici, B.Shapira, A.Maschiah. A New Privacy Model for Web Surfing. In Proceedings of NGITS 2002 Zichron_Jaacov, Israel, pp. 45-57, 2002.

32 Jaideep Vaidya, Chris Clifton: Privacy-Preserving Decision Trees over Vertically Partitioned Data. DBSec 2005: 139-152

33 Ling Qiu, Yingjiu Li, Xintao Wu: Preserving privacy in association rule mining with bloom filters. Springer Science + Business Media, LLC 2007

34 Bloom, B. (1970) Space time tradeoffs in hash coding with allowable errors. Communications of the ACM, 13(7), 422–426.

35 S.R.M. Oliveira, O.R. Zaı ¨ane, Y. Saygin, Secure association rule sharing, advances in knowledge discovery and data mining, in: Proceedings of the 8th Pacific-Asia Conference (PAKDD2004), Sydney, Australia, 2004, pp. 74–85.

36 S.R.M. Oliveira, O.R. Zaı ¨ane, A unified framework for protecting sensitive association rules in business collaboration, International Journal of Business Intelligence and Data Mining l (3) (2006) 247–287.

37 Z. Wang, W. Wang, B. Shi, Blocking inference channels in frequent pattern sharing, in: Proceedings of the 23rd International Conference on Data Engineering (ICDE'07), Istanbul, Turkey, 2007, pp. 1425–1429.

38 R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), Santiago, Chile, 1994, pp. 487–499.

39 E.T. Wang, G. Lee, Y.T. Lin, A novel method for protecting sensitive knowledge in association rules mining, in: Proceedings of the 29th IEEE Annual International Computer Software and Applications Conference (COMPSAC'05), Edinburgh, Scotland, 2005, pp. 511–516.

40 Oliveira, S.R.M. (2005) Data Transformation For Privacy-Preserving Data Mining, PhD Thesis, Department of Computing Science, University of Alberta, Edmonton, AB, Canada, June.

41 En Tzu Wang , Guanling Lee, An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining: Data & Knowledge Engineering Volume 65, Issue 3, June 2008, Pages 463-484

42 S.R.M. Oliveira, O.R. Za.¨ ane, Privacy preserving frequent itemset mining, in: Proceedings of the IEEE ICDM Workshop on Privacy, Security, and Data Mining, Maebashi City, Japan, 2002, pp. 43–54.

43 S.R.M. Oliveira, O.R. Za.¨ane, Protecting sensitive knowledge by data sanitization, in: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03), Melbourne, Florida, USA, 2003, pp. 613–616.

44 S.R.M. Oliveira, O.R. Za.¨ane, A unified framework for protecting sensitive association rules in business collaboration, International Journal of Business Intelligence and Data Mining l (3) (2006) 247–287.