המחלקה למערכות מידע ניהוליות

פרוייקט גמר תשף 2020

זיהוי פעולות חשודות (אנומליות) ברשת הביתית

NETWORK TRAFFIC ANOMALY DETECTOR

מסמך זה מהווה חלק מדרישות לקבלת תואר ראשון במערכות מידע BSC

מכללת אחווה

שמות המבצעים: אלמוג אליאס, קובי תורג'מן וכנרת כהן.

שם המנחה: ד"ר יצחק רוזן.

**הצהרה**

הפרויקט נעשה בהנחיית ד"ר צחי רוזן

המכללה האקדמית אחווה- מחלקת מערכות מידע ניהוליות

מכיוון שהפרויקט הנו בלמידת מכונה ושימוש בתחום איתור התקפות בתעבורת הרשת על בסיס אנומליה, הוא חדשני יחסית, דרוש מחקר רב על מנת לבצע את הפרויקט בנוסף למימוש עצמו.

הפרויקט מציג את עבודתנו האישית ומהווה חלק מקבלת תואר ראשון במערכות מידע ניהוליות.

# תיאור הארגון

## 1.1.1. רקע כללי

רשת מקומית LAN מאפשרת למחשב אחד המחובר לרשת להשתמש במשאבים ממחשבים אחרים. שימושים נוספים הם הרצת יישומים ממחשב אחד על מחשב אחר.

חסרונותיה של העבודה ברשת הם סיכון לפרטיות המידע וריבוי מקורות לתוכנות זדוניות. המחשבים שברשת יכולים להתחבר ביניהם בכבלים או בקשר אלחוטי ויש לוודא שאין פעילות חריגה על הרשת, ושרק אנשים מורשים יוכלו להשתמש במחשב שכבר מחובר אל הרשת. מעבר לזאת יש להגביל את הגישה של מי שכן מורשה כך שתוכנם של מסמכים יהיה גלוי רק לבעלי סמכות רלוונטית עבור תוכנו של כל מסמך או תוכנה.

## 1.1.2. לקוח

לקוחות המערכת הם משתמשי הבית החכם. הלקוח מגדיר לנו מה לעשות ובהתאם לזה נבנה כנדרש. המערכת תשמש בעיקר את דיירי הבית ומשתמשי הרשת המקומית LAN. הגדרנו את קובי להיות המייצג של הלקוח, אשר יהיה מסוגל להגדיר את הדרישות של הלקוח.

## 1.1.3. מומחי היישום

מומחי היישום הינם בעלי הבית אשר יטמיעו את המערכת.

ההחלטה לגבי הפיתוחים תהיה על ידי צוות היגוי שתורכב מראש צוות האפיון, ומומחה אבטחת מידע.

## 1.1.4. משתמשים

* אחראי הרשת המקומית – בעל הבית.

# 1.2. מטרות ויעדי הארגון

## 1.2.1. מטרת הארגון

הטכנולוגיה חשופה לשלל פרצות מידע, ויש צורך תמידי לשכלל ולהתמודד עם באגים, אשר הולכים והופכים למתוחכמים יותר .ישנם שיקולים שחייבים להילקח בחשבון כשעוסקים בחיבור מחשבים לרשת. הנחה היא שאי אפשר לעבוד ללא תקשורת בין מחשבים ולכן המטרה היא לזהות פעילות חריגה עויינת של תוקפים למניעת התפשטות ברשת המקומית.

## 1.2.2. עקרי ענף המערכת

הענף בו עוסק הפרויקט הינו תחום הסייבר על ידי שיטות ואלגוריתמים של למידת מכונה.  
קהל היעד של הפרויקט הנו:

1. משתמשי רשת מקומית
2. משתמשי הרשת ביתית
3. משתמשי הבית חכם

## 1.2.3. הצגת מטרות המערכת

הגדרת הפרויקט היא בניית תשתית שבאמצעותה אפשר לחקור תופעה אנומלית.

לדמות התקפות באמצעות הסימולטור ולהגדיר את ההתנהגות שלהן באמצעות אלגוריתם של למידת מכונה.

גבולות המערכת-

1. הקמת מעבדה (חומרה) – סקלביליות
2. סימולטור- מדמה סביבה רגילה וגם סביבה שמדמה התקפות.
3. אוספים את התנועות ברשת - גשש שאוסף את התנועות ברשת בזמן אמת ומתעד אותם - ושליפה של ערכים להגדרת מודל לאנליזה עליו נעבוד.
4. הקמת db.
5. ויזואליזציה- הצגת הנתונים בdb
6. אנליזה של הנתונים באמצעות אלגוריתם של למידת מכונה.

## 1.2.4. יעדי המערכת

פרויקט זה מתחלק לשלושה חלקים עיקריים:

1.איסוף נתונים

1. צורך בדאטה סט עדכני וגדול על מנת לספק בסיס לאימון המודל ולכן יש ליצור דאטה סט מתאים.
2. הבנת הנתונים ומחקר על יצירת מודל עבורם.

2.בניית המודל

1. חיפוש אחר מחקר קיים
2. למידת המודל
3. אפיון חריגות בתעבורה ברשת הביתית באמצעות למידת מכונה.

3.(בונוס )

ניתור הרשת בזמן אמת

1. בניית תוכנה המנתרת את הרשת משייכת מנות לחיבורים מסווגת את החיבורים החריגים והתקינים.
2. מתריעה במקרה של חיבור חריג.
3. אינטראקציית אדם מחשב - חסימת תעבורה חשודה בעקבות הפעולות החשודות שנמצאו.

פרויקט זה יעשה ע"י סביבת מעבדה המורכבת משני עמדות קצה ושרת אשר יוקצה למפתחים כיוון שנדרשת חומרה עבור כמות המידע ופעילה 24/7 לצורך האימון.

## תיאור הבעיה

כמות המידע שגוברת עם הזמן והתממשקות מול רכיבים נוספים גורמת לסיכון גבוה יותר לתקיפת סייבר בעולם וברשת הפנימית בפרט.

כיום ישנם חברות קיימות אשר משלבות מוצר חומרתי (נתב) או חוצץ בין הרשתות כדי לזהות אנומליות אך לא בהכרח כל התעבורה עוברת דרך הנתב.

אף על פי כן, מוצרים אלו מאתרים איומים והתקפות בתעבורה ע"י חיפוש תבניות ספציפיות או ע"י הכרת דפוס הפעולה של הקובץ הזדוני, כך שלכל מתקפה קיימת תבנית ייחודית עבורה.

כאשר נחשף איום חדש יוצרים תבנית תואמת לאיום ולפי התבנית המערכת מזהה את האיום לאחר מכן.

ההנחה היא שהתוקף כבר נמצא ברשת הביתית וע"י למידת וחקירת הרשת הוא מזהה מטרות תקיפה.

**הבעיה** היא זיהוי פעילות בשלב  Active Reconnaissance של תוקף ובנוסף זמן התגובה לאיום זה מהרגע שבו הוא מזוהה עד לבניית התבנית עבורו, כיוון שכבר נוצרה פגיעה במערכת.

היום בעולם המודרני ישנה עלייה בכמות המידע העוברת ברשת ובעקבות כך נוצר ריבוי בסכנות וסיכוי גבוה יותר לפריצות ע"י תוקפים פוטנציאלים.

מדובר על **מערכת מידע מורכבת** שלא ניתנת לחיזוי מתקפות סייבר עתידיות.

## סקירת ספרות

הוחלט להתבסס על הספר "Anomaly Detection Principles and Algorithms.pdf"

סקירת ספרות נמצאת במאגר מסמכים.

## תיאור המערכת העתידית

**איסוף המידע**

1. הסנפת חבילות בעמדת קצה.
2. שליחת המידע למאגר המידע ב – DB.

**עיצוב מודל נתונים**

1. בחירת Features למידע הקיים.
2. יצירת מודל לFeatures שנבחרו.
3. כתיבת התוצאות לטבלה
4. אפיון תעבורה חריגה באמצעות למידת מכונה.

**עיבוד הנתונים**

1. רישום הנתונים לתוך מאגרי Packets - Features.
2. עיבוד חבילות ל- Flows – שיחות.
3. רישום ה Flows למאגרי Features – Flows
4. מאגר חריגות בהתאם למודל שנבנה.

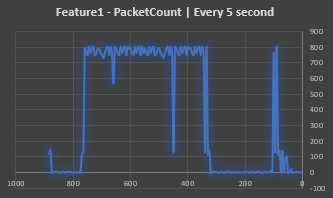
**ויזואליזציה**

1. יצירת ממשק WEB ויזואלי לגראפים על סמך מאגרי הFeatures.
2. יצירת טבלה המציגה את החריגות שנמצאו על ידע המודל.

## בחירת Features

* + - 1. כמות חבילות ליחידת זמן   
         כמות גדולה של חבילות ליחידת זמן עשויה להעיד על פעילות חריגה.

חיזוי טבלאי:



**מודל – נתונים נאספים:**

* רשימה של זמני החבילות.
* כמויות של חבילות לפי פרק זמן.

**סימולציה – אוסף נתונים:**

הרצת תעבורה בפרוטוקול: Microsoft – ds אשר משתמש בפורט 445.

פרוטוקול זה משמש להעברת קבצים בין עמדות קצה

סימולציה נורמלית:

העברת קבצים בזמן רנדומלי בין 0.1-10 שניות באופן תמידי.

כל קובץ ינוע בגודל שבין 10-100 Bytes.

סימולציות אנומליות:

1. העברת קובץ יחיד בנפח שגדול פי 1000 מהסימולציה הנורמאלית.
2. העברת 100 קבצים בגודל של 10 Bytes בפרק זמן בין 0.1-2 שניות.

**למה זו יכולה להיחשב חריגה**

* כמות של חבילות למשך זמן ארוך מהממוצע (מכיוון שהקובץ גדול אז תהינה כמות רבה של חבילות לאורך זמן) נחשבת לחריגה מכיוון שזמן הפעילות גבוה מהזמן הממוצע.
* כמות גבוהה של חבילות בזמן קצר נחשבת לחריגה מכיוון שהכמות גבוהה מהממוצע.

**על מה היא מעידה?**

* מעידה על עומס ברשת – שיכול לגרום לאיטיות – בדרך כלל אין פעילות מסיבית, ולכן זאת תופעה פוטנציאלית שעלולה לקרות בזמן של חדירה.
  + - 1. כמות הודעות ל- Flow

כמות קטנה/גדולה של חבילות בשיחה ליחידת זמן עשויה להעיד על פעילות חריגה.

**מודל – נתונים נאספים**זמן התחלת השיחה, יום התחלת השיחה, כתובת יוזם השיחה, כתובת יעד השיחה, פורט מקור, פור יעד, פרוטוקול שירות (Service), כמות החבילות הכללי בשיחה, כמות ה-Syn Flag בשיחה, כמות ה-Ack Flag בשיחה, כמות ה-Payload בשיחה, כמות ה - Fin בשיחה.

**סימולציה – אוסף נתונים**

סימולציה נורמלית:

הרצת תעבורה בפרוטוקולים : 445,3389,21,23.

העברת רצפי מחרוזות בגודל 2-10 מילים אשר בגודל 1-15 אותיות בזמן רנדומלי בין 1-10 שניות באופן תמידי.

סימולציות אנומליות:

1. הרצת תעבורה בפורט שלא היה קיים קודם לכן לדוגמה פורט מספר 4444.
2. העברת רצף מחרוזות יחיד בגודל של פי 10 מהסימולציה הנורמלית.
3. העברת 100 רצפי מחרוזות בגודל של מילה יחידה בפרק זמן בין 0.1-2 שניות.
4. שליחת רצף מחרוזת יחיד בגודל של אות בודדת.

**למה זו יכולה להיחשב חריגה**

1. שימוש בפורט שלא היה בו שימוש קודם לכן עלול להיחשב כחריג מכיוון שהוא חדש.
2. העברת מחרוזת גדולה נחשבת לחריגה מכיוון שתכיל חבילה בגודל אשר גדול מהממוצע.
3. כמות גדולה של מחרוזות ליחידת זמן עשויה להעיד על פעילות חריגה מכיוון ששיחה תכיל מספר רב של חבילות.
4. כמות קטנה של חבילות עם Payload בשיחה נחשבת לחריגה מכיוון שכמות החבילות שמכילות תוכן קטנה מהממוצע.

**על מה היא מעידה**

תוקף מנסה להסוות את פעילותו על ידי שימוש מופחת/מופרז בשיחות.

* + - 1. גודל (bytes) חבילה ליחידת זמן:  
         מספר חבילות גדולות לאורך זמן שגדול מהממוצע עשויה להעיד על פעילות חריגה.

**מודל – נתונים נאספים**גודל החבילה, מספר זיהוי שיחה.

**סימולציה – אוסף נתונים**

הרצת תעבורה בפרוטוקולים : 445

בין השעות 11:00-20:00 ירוצו סימולציות בכל שעה

בכל 5 דקות ירוץ סימולטור בחירה רנדומלית בין 1-50 חבילות.

העברת קובץ בפורט 445 בגודל 100 bytes.

**למה זו יכולה להיחשב חריגה**

* כמות גבוהה של חבילות למשך 5 שניות נחשבת לחריגה מכיוון שהכמות גבוהה מהממוצע.
* כמות של חבילות למשך זמן ארוך מהממוצע נחשבת לחריגה מכיוון שזמן הפעילות גבוה מהזמן הממוצע.

**על מה היא מעידה?**

* מעידה על עומס ברשת – שיכול לגרום לאיטיות – בדרך כלל אין פעילות מסיבית, ולכן זאת תופעה פוטנציאלית שעלולה לקרות בזמן של חדירה.
* מעיד על תקשורת רציפה כגון סריקת רשת של תוקף והעברת מידע בין מחשבים.
  + - 1. הפרש הזמנים בין יצירת הקובץ לשימוש בקובץ.  
         הפרש זמנים קטן בין יצירת הקובץ עלול להעיד על קובץ חדש שנכנס למערכת ויזם תקשורת עם רכיב נוסף ברשת המקומית, לצורך הדוגמה: תוקף או נוזקה שמנסה להשתמש בשיטת Lateral movement כדי להגיע לעוד רכיבים שונים.

**מודל – נתונים נאספים**הפרש הזמנים בין יצירת הקובץ לשימושו (בזמן מנורמל), למספר זיהוי חבילה.

**סימולציה – אוסף נתונים**

הרצת תעבורה בתוכנות: ,Putty.exe,FTP.exe

בין השעות 11:00-20:00 ירוצו סימולציות בכל שעה

בכל 10 דקות ירוץ סימולטור בחירה רנדומלית בין 1-50 חבילות.

**למה זו יכולה להיחשב חריגה**

* הפרש הזמנים בין יצירת הקובץ לשימוש בקובץ יכול להיחשב כחריג על סמך קטריון זמן אשר נמוך מהממוצע.

**על מה היא מעידה?**

* מעיד על קובץ חדש שנוצר במחשב ויצר תקשורת עם מחשבים אחרים ברשת הביתית, דבר המעיד על ניסיון התפשטות של נוזקה.

## דרישות פונקציונאליות

### איסוף המידע

התחלת/הפסקת ניטור המידע

**תיאור פונקציונאלי כללי:** התהליך מתרחש באופן ייזום על ידי משתמש המערכת בעת לחיצה על לחצן "ניטור המידע". בעת לחיצה על כפתור " "Start Captureאו "Stop Capture" תשלח הודעה מתאימה לסוכן שיושב בעמדת קצה שיתחיל בניטור הרשת.

ניטור המידע יהווה האזנה לרכיב התקשורת הפועל ברשת המקומית, על ידי בדיקת חיבורים (Local Area Network).

כל פאקט שמנוטר מפורסר לדאטה המתאים למערכת, נרשם לרשימה המונה את החבילות, ובכל 5 שניות נשלחת הרשומה למאגר המידע "Packets" בשרת ה-DB ונרשמים לקובץ לוג file בשם ייחודי (Logfile[Unix-Time]) בעמדת הקצה. בסיום כל כתיבה מאפסים את הרשומה הישנה.

הקושי במימוש : הינו יצירת Flow מתוך איסוף החבילות.

### עיבוד הנתונים

**תיאור פונקציונאלי כללי:** התהליך מתרחש כאשר מתווספת רשומה חדשה למאגר הנתונים "Packet", כתוצאה מכך מעבדים את הנתונים לתוך מאגרי הFeatures בהתאם ב4 פיצ'רים שנבחרו לעיל.

הקושי במימוש: הינו הפקת DATASET מתאים בהתאם למאגר ה"Packet" וה-"Flow".

### ויזואליזציה

**תיאור פונקציונאלי כללי:** התהליך מתרחש לאחר לחיצה על כפתור המסכים "Dashboard", אשר במהלכו מאגרי הFeatures מוצגים בצורה גראפית במסך.

הקושי במימוש: יצירת ממשק Web שמושך את המידע מהמאגרים.

### זיהוי חריגות

**תיאור פונקציונאלי כללי:** התהליך מתרחש לאחר הזנת הנתונים למאגרי ה – Features. שימוש באלוריתם אשר יבדיל בין הפעולות הנורמליות לפעולות החריגות. רישום במאגר "חריגים" את הנתונים שהתקבלו מהאלגוריתם לזיהוי החריגות.

הקושי המימוש : בחירת מודל מתאים בהתאם לFeatures.

## חלוקת תפקידים

**הקמת מעבדה**

אלמוג, קובי וכנרת בעזרת פלד מנהל מערכות מידע במכללה.

**בחירת Features**

1. בחירת Features תוך שימוש במחקרים קיימים: קובי אלמוג וד"ר צחי.

**יצירת מודל**

1. יצירת מודל המבוסס על בחירת ה Features: אלמוג וד"ר צחי.

**צד לקוח**

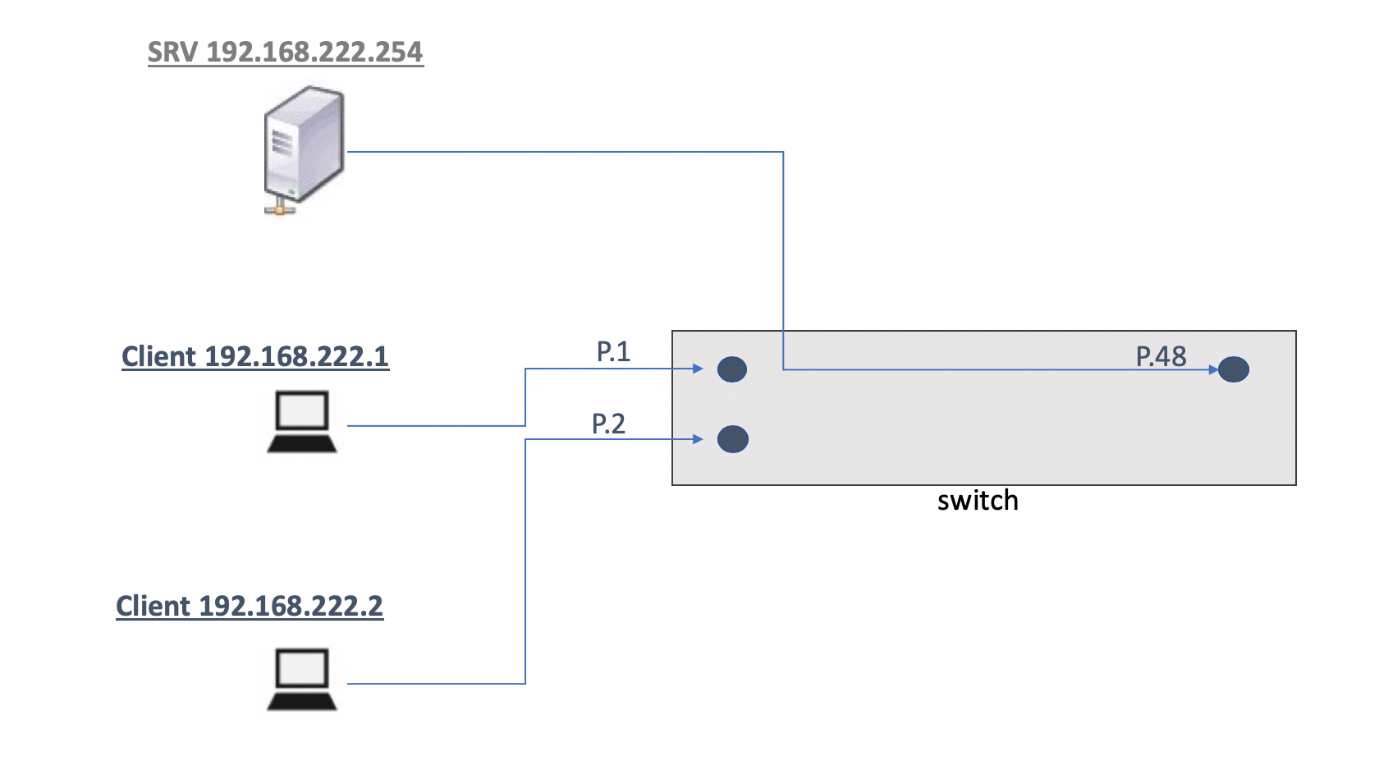
* פיתוח סוכן מנהל את התהליך כולו- קובי
* פיתוח הסנפת/האזנה פקטות-אלמוג
* פיתוח כתיבה לDB – כנרת

בונוס- פיתוח הריגת תהליך וחסימה של תעבורה ספציפית.

**צד שרת**

* אפיון DB בטבלאות- כנרת ואלמוג
* יצירת פיתוח תקשורת מול סוכנים-קובי
* אינטראקציית אדם מחשב- אלמוג, קובי וכנרת.

## תיאור ארכיטקטורה הכללית של המערכת ואילוצים חשובים

צד לקוח:

תהליך איסוף המידע יתחיל בצד הלקוח על ידי הסנפת חבילות בזמן אמת.

יצירת סימולציות (נורמליות, לא נורמליות) יבוצעו בצד הלקוח.

תהליך איסוף המידע יבוצע בשפת Python על ידי ספרייה שנקראתScapy .

צד שרת:

יכיל DB עם מאגרי מידע בהתאם לדרישות הפונקציונאליות ויאכלס בתוכו את המידע שנאסף על ידי הלקוח.

בנוסף, השרת יאפשר שירות WEB לויזואליות של המערכת.

שפת התכנות שבה נשתמש בשרת היא C# בטכנולוגית MVC , והDB שנשתמש בו יהיה MSSQL 2014.

אילוצים חשובים:

* 2 עמדות לצורך ביצוע הסימולציות.
* שרת אשר מסוגל לעבד נתונים בצורה מהירה.
* רכיב תקשורת בכדי לאפשר תעבורה בין כלל הרכיבים של המערכת.

# לוחות זמנים

## הספקת גרסאת אלפא למערכת

על הסטודנטים לעמוד בלוח זמנים צפוף אשר תהא תואמת את לוחות הזמנים הסימסטריאלים והלו"ז שנקבע על ידי מנחה הפרוייקט והינם כ-שלושה חודשים מתחילת העבודה   
וכן לעמוד בלוז הפורט להל"ן

## 

|  |  |
| --- | --- |
| שלב | זמן |
| מסמך דרישות מפורט | 9 שבעות |
| מסמך עיצוב תכנון | שבוע 12-13 |
| מסמך מפורט | תחילת סמסטר ב |
| PILOT | שבוע רביעי סמסטר שני |
| מוצר סופי | סוף שנת הלימודים. |