Computer networks – Final project

GitHub Repository: https://github.com/AlonMesh/Computer-Networks-Final-Project

Group members:

1. Alon Meshulam
    a. LinkedIn: https://www.linkedin.com/in/alon-meshulam/
    b. GitHub: https://github.com/AlonMesh
2. Israel Gitler
    a. LinkedIn: https://www.linkedin.com/in/israelgitler/
    b. GitHub: https://github.com/GitlerIsrael

חלק 1 - הרעיון המרכזי של המאמר

הרעיון המרכזי של המאמר הוא לבחון את תפיסת האבטחה של אפליקציות IM (Instant messaging, הודעות מיידיות) מאובטחות לכאורה (דוגמת WhatsApp, Telegram, Signal) ולערער אותה. המחברים מסבירים שגם בהינתן הצפנה מקצה לקצה על תוכן ההודעות ניתן להפיק מידע משמעותי על משתמש באמצעות מעקב אחר תעבורת הרשת שלו. זאת, בעיקר משום שמפעילי האפליקציות הללו לא מפעילות מנגנון לערפול מאפייני התעבורה שלהן ושל המשתמשים שלהן. לבסוף, המאמר מציג מערכת לטשטוש תעבורה (IMProxy) המיועדת למנוע תקיפות נגד תעבורת הרשת של המשתמשים. המערכת כופה על תעבורת SIM (Secured IM) לעבור דרך VPN ומשלבת בתעבורה האמיתית תעבורה "שקרית" ובכך מגבילה את יכולת התקיפה הנוכחית.

עבור כל ערוץ (או קבוצה) יעד C, תוקף הרוצה להשיג מידע אמפירי (Ground truth) על תעבורת הערוץ יכול לעשות זאת בשלוש דרכים:

1. אם C הוא **ערוץ פתוח** (ציבורי), התוקף יכול להצטרף לערוץ (כחבר) ולהקליט את ההודעות שנשלחו ב-C יחד עם המטא-דאטה שלהן (לדוגמה, זמן ההודעות וגודלן).
2. אם C היא **קבוצה סגורה** שמעניקה לכל חבר את היכולת לפרסם הודעות, והתוקף הצליח להצטרף ל-C ולשלוח הודעות שם, או במקרה שהתוקף קיבל תפקיד אדמין ב-C. במצב זה, לא רק שהתוקף יכול להקליט את ההודעות שפורסמו ב-C, אלא גם הוא יכול לכתוב בעצמו הודעות ב-C עם דפוסי התעבורה ייחודיים שיכולים לגרום לשאר חברי הקבוצה להגיב, ובכך התוקף יצליח לאסוף את המידע הרצוי על הערוץ.

3. במקרה **שהתוקף אינו יכול להצטרף** ל-C (כחבר או כאדמין) אך הוא הצליח לזהות את כתובת ה-IP של אחד ממשתתפי הקבוצה C, התוקף יכול להקליט (בסתר) את תעבורת הרשת (המוצפנת) של המשתמש המזוהה ולהקליט את דפוסי התעבורה שלו.

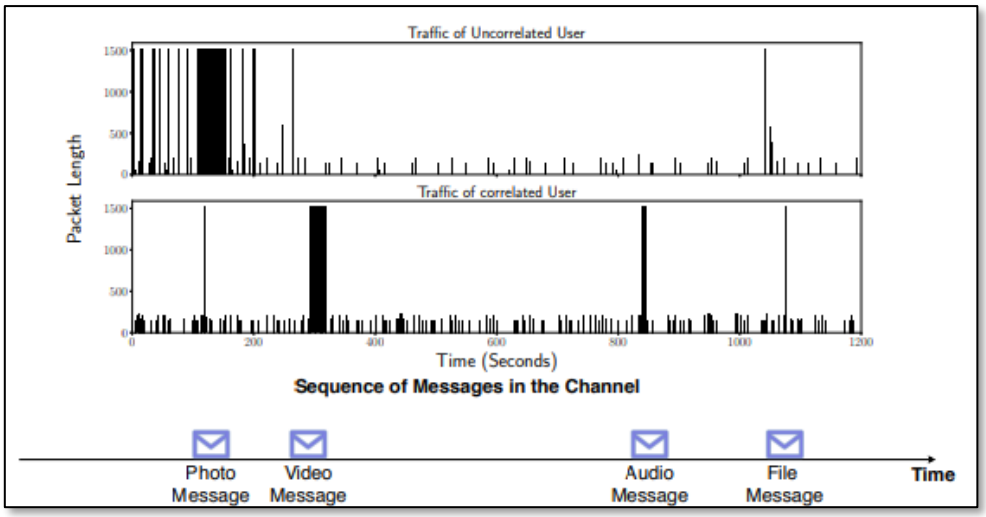
התוקף מנטר את תעבורת הרשת (המוצפנת) של משתמשי ערוץ (או קבוצה) כדי לזהות כתובות IP של חברי ומנהלי הערוץ. התוקף יכול לעשות זאת ע"י האזנת סתר לתעבורת הרשת של ה-ISP (Internet Service Provider) או ה-IXP (Internet eXchanged point) עליהם הוא יושב (שולט עליהם). אפשרות נוספת היא שהתוקף מבצע האזנת סתר לתעבורת הרשת של משתמש ספציפי (לדוגמה חשוד בעבירה) לאחר שהושג צו (מטעם בית משפט למשל) להאזנה.

TABLE II: Distribution of various message types

Type	Count	Volume (MB)	Size range	Avg. size
Text	12539 (29.4%)	3.85 (0.016%)	1B-4095B	306.61B
Photo	20471 (48%)	1869.57 (0.765%)	2.40Kb-378.68Kb	91.33KB
Video	6564 (15.4%)	232955.19 (95.3%)	10.16Kb-1.56Gb	35.49MB
File	903 (2.1%)	47.46 (0.019%)	2.54Kb-1.88Mg	52.56KB
Audio	2161 (5.1%)	9587.36 (3.92%)	2.83Kb-98.07Mg	4.44MB

הטבלה מציגה נתונים (כמות, נפח, טווח וגודל ממוצע) על חמישה סוגים מרכזיים של תעבורת IM – הודעות, תמונות, סרטונים, קבצים וקבצי שמע. המסקנות העולות מהתעמקות בטבלה הן:

1. ישנה תעבורה משמעותית של תמונות (48%). עם זאת, הנפח שהן תופסות הוא קטן יחסית (פחות מאחוז). כמו כן, ניתן לראות שטווח הגודל של התמונות הוא יחסית קטן (ביחס לסרטונים ולקבצי השמע), וניתן להסיק מכך שאנשים נוטים לשתף תמונות באיכות נמוכה או שלחלופין תוכנת ה-IM מורידה את איכות התמונה.
2. לעומת זאת, הסרטונים מהווים 15% מהתעבורה אך הם תופסים נפח של 95% מהתעבורה. טווח הגודל של הסרטונים הוא רחב מאוד (KB10 עד GB1.5) מבחינה אבסולוטית לעומת שאר הסוגים, בהם סדרי הגודל נשארים קטנים. ניתן להסיק מכך שמשתמשים שולחים סרטונים באיכויות מגוונות ובאורכים שונים.
3. ישנה תעבורה גדולה של הודעות טקסט (30%) אך הנפח שהן צורכות הוא בסביבות מאית האחוז, מה שמעיד על תפיסת מקום מינימלית ויעילה בציר תקשורת זה.
4. סוג התקשורת הכי פחות נפוץ הוא העברת קבצים (2%) ולאחר מכן קבצי שמע (5%). ניתן להסיק שאנשים נוטים פחות להשתמש בסוגי תקשורת אלו.
5. טווח הגודל של קבצי השמע רחב (KB3 עד MB98) מבחינה אבסולוטית, וניתן להסיק מכך שנשלחים קבצי שמע באורכים שונים ובאיכויות שונות.



באיור מופיעים שני תיעודים של תעבורת רשת, המציגים את גודל התעבורה כפונקציה של זמן. התיעוד העליון מעיד על תעבורת רשת של משתמש שאינו חבר בערוץ (Uncorrelated User) והתיעוד התחתון מעיד על תעבורת רשת של משתמש הרשום לערוץ (Correlated User). כמו כן, בתחתית האיור מופיע ציר זמן עם "מופעים", כאשר כל מופע מעיד על שליחת תעבורה בערוץ ומפרט את סוגה.

1. ניתן לראות בבירור כי ישנה קורלציה גבוהה בין שליחת תעבורה "כבדה" (שאינה טקסט) בערוץ לבין גדילה משמעותית בתעבורת הרשת של המשתמש הרשום אליו מיד לאחר מכן.
2. ניתן להסיק שהקורלציה בין שליחת תעבורה "כבדה" בערוץ לבין תעבורת הרשת של המשתמש שאינו רשום היא נמוכה. ישנה תעבורה משמעותית ב-200 השניות הראשונות, אך לא סביר שהיא כתוצאה משליחת התמונה בערוץ (בערך אחרי 100 שניות). כמו כן, קל לראות שאין השפעה משמעותית על תעבורת הרשת לאחר שליחת הסרטון, קובץ השמע והקובץ.
3. תובנה נוספת, שרק מחזקת את התובנות הקודמות היא שיש קורלציה נמוכה בין המשתמש הרשום למשתמש שאינו רשום. תעבורות הרשת שלהן שונות זו מזו ואין דפוסים זהים.

חלק 2 – הקלטת תעבורה של משתמש WhatsApp

*כדי להפעיל את הקוד במחשב שלך, יש לעקוב אחר ההוראות בקובץ README.MD.

כדי לבצע את החלק הרטוב של הפרויקט, כתבנו קוד המקבל לכל קבוצה שנעקוב אחריה קובץ pcap וקובץ csv תואם המכילים את תעבורת הרשת של אחד ממשתתפי הקבוצה.

כדי להפעיל את התוכנה, יש להקליט את תעבורת הרשת של משתמש מסוים החבר בארבע קבוצות, כאשר בכל קבוצה נשלחת תעבורה מסוג ייחודי (במקרה שלנו: טקסט, אודיו, תמונות ווידאו). עבור כל קבוצה יש לשמור קובץ pcap בשם *"<name>s_record"* ולשמור בתיקייה resources.

לאחר הקלטת תעבורת הרשת של אחד ממשתתפי הקבוצה, פילטרנו את ההקלטה באופן הבא:

1. ה-TCP פורט צריך להיות 443. לפי מחקר שערכנו ברשת, זהו הפורט המקובל להעברת מידע מאובטח באמצעות HTTPS והוא בשימוש של WhatsApp.
2. כתובת ה-IP השולחת / מקבלת היא 157.240.214.60. מבדיקה עולה כי כתובת זו שייכת לחברת Meta (המפעילה של WhatsApp). מלכתחילה כתובת זו נחשדה כמקושרת ל-WhatsApp לאור הדומיננטיות שלה בתעבורת הרשת של המשתמש.
3. הפרוטוקול הוא TLS, מכיוון שהוא משמש להעברת החבילות המכילות את המידע המוצפן.

כעת, נביט בפונקציית ה-main שלנו המציגה בבירור כיצד התוכנה שלנו עובדת:

```python
def main():
    # Modify this list to specify different WhatsApp group types for analysis
    GROUPS = ['Message', 'Photo', 'Audio', 'Video']
    for group in GROUPS:
        pcap_file_path = f"resources/{group}s_record.pcap"
        csv_file_path = f"resources/{group}s_record.csv"
        if not os.path.exists(pcap_file_path):
            raise FileNotFoundError(f"{pcap_file_path} does not exist.")
        try:
            print(f"Converting {pcap_file_path} → {csv_file_path} ...")
            convert_pcap_to_csv(pcap_file_path, csv_file_path)
            print("Conversion complete.")
        except Exception as e:
            print(f"Error occurred during conversion: {e}")
        try:
            print(f"Starting plots for {csv_file_path} ...")
            whatsapp_analysis.creating_plots(csv_file_path, group)
            print("Plots generated.")
        except Exception as e:
            print(f"Error occurred during plot generation: {e}")
```

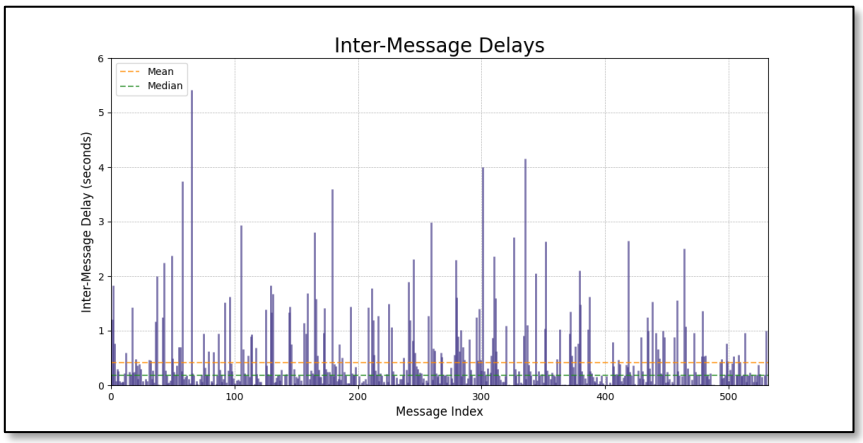
(כאשר המשתנה GROUPS הוא רשימה שערכיה נקבעו ידנית על ידינו).

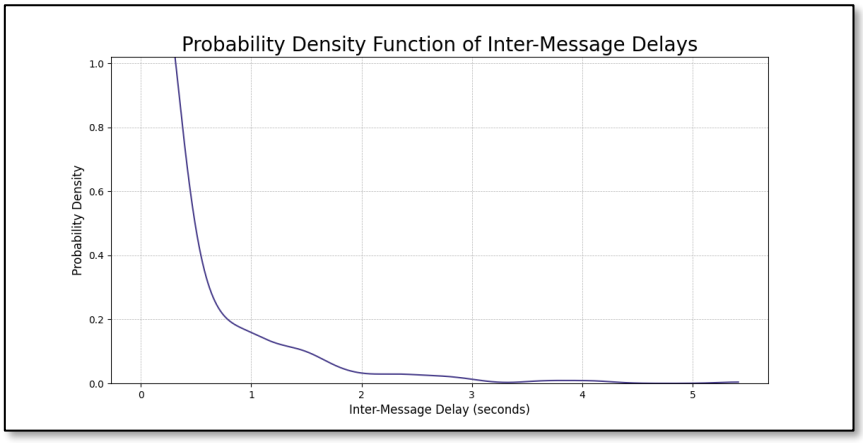
עבור כל סוג תעבורה ייחודי, התוכנה ממירה קובץ pcap נתון לקובץ csv ולאחר מכן שולחת את הקובץ המומר לניתוח בפונקציה נפרדת (creating_plots(...)), היוצרת שלושה גרפים על בסיס ניתוח תעבורת הרשת של משתתף הקבוצה (ושומרת כל גרף כקובץ png בתיקייה res):

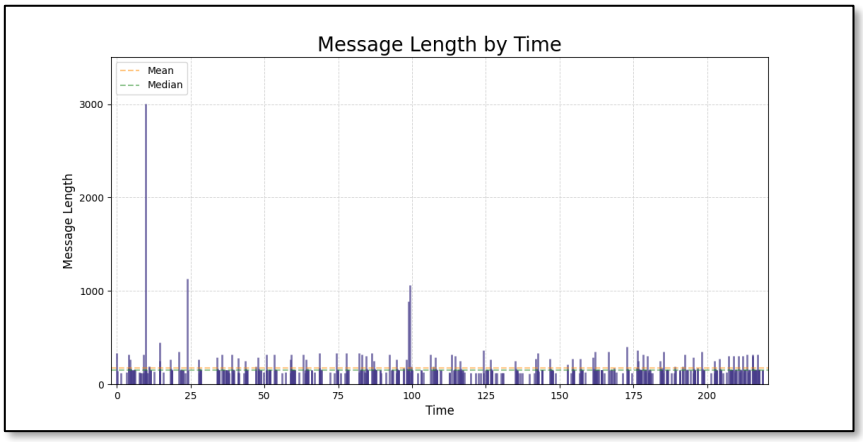
1. הפרשי הזמנים בין ההודעות
2. פונקציית הצפיפות של הפרשי הזמנים
3. גדלי ההודעות שנשלחו כפונקציה של זמן

להלן פירוט התוצאות על בסיס הגרפים שנוצרו:

1. הודעות טקסט:

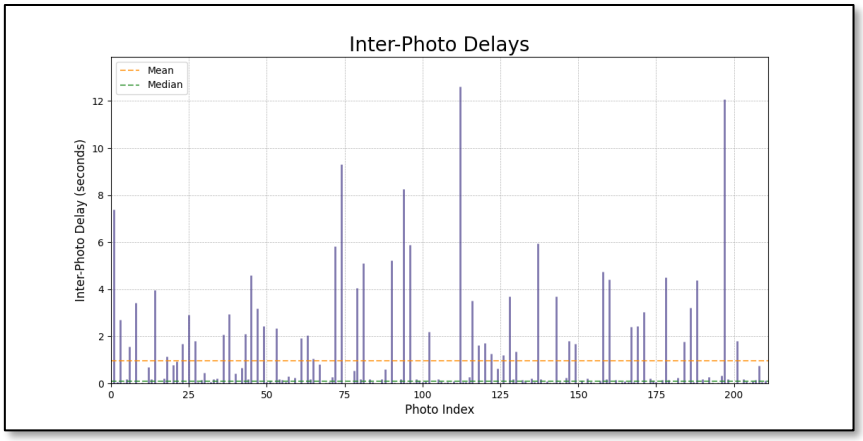


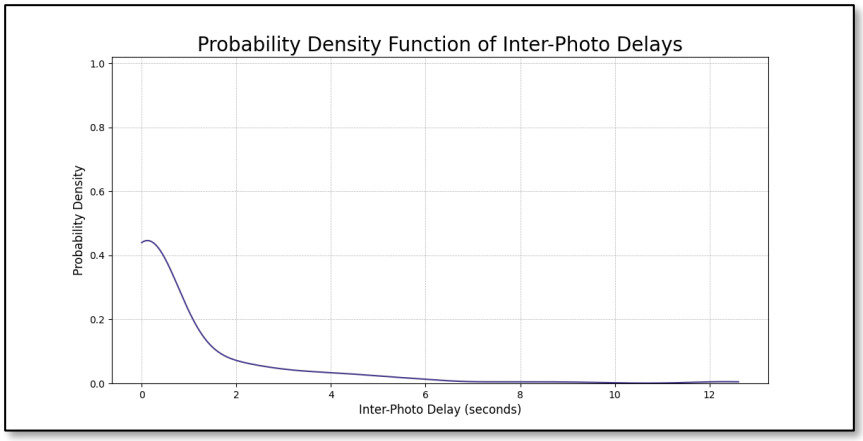


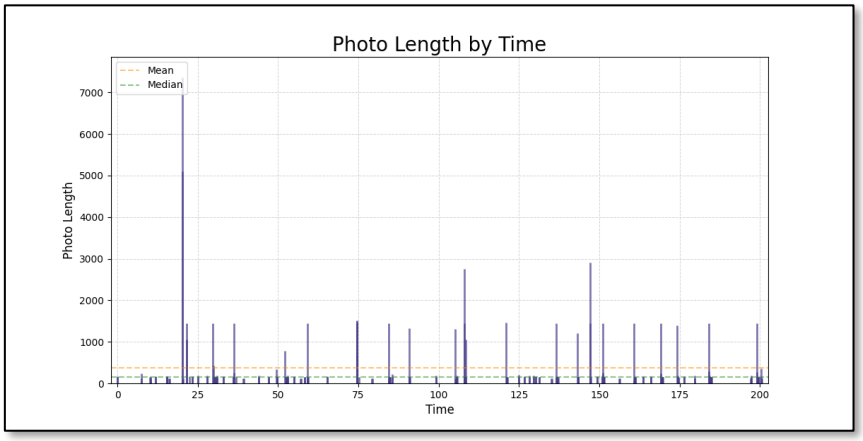


ניתן לראות כי הפרשי הזמנים (Delays) בין ההודעות משתנים בטווח רחב יחסית (0-5.5 שניות), ובנוסף ניתן להסיק כי מדובר בתעבורת רשת צפופה וכי הקבוצה שהוקלטה פעילה מאוד (לכל הפחות בעת ההקלטה). ניתן לראות כי רוב התעבורה שנשלחה היא בגודל 100-300 בייטס בהערכה גסה, ופרט לכך יש גם ארבע חריגות (שלוש בגודל 1000~ בייטס ואחת בגודל 3000~ בייטס). להערכתנו, סביר להניח שמדובר ב"רעש", כאשר ככל הנראה אלו עדויות בתעבורת הרשת של סוגי תקשורת שונים (תמונות, סרטונים) שנשלחו במקורות אחרים בזמן ההקלטה.

2. תמונות:

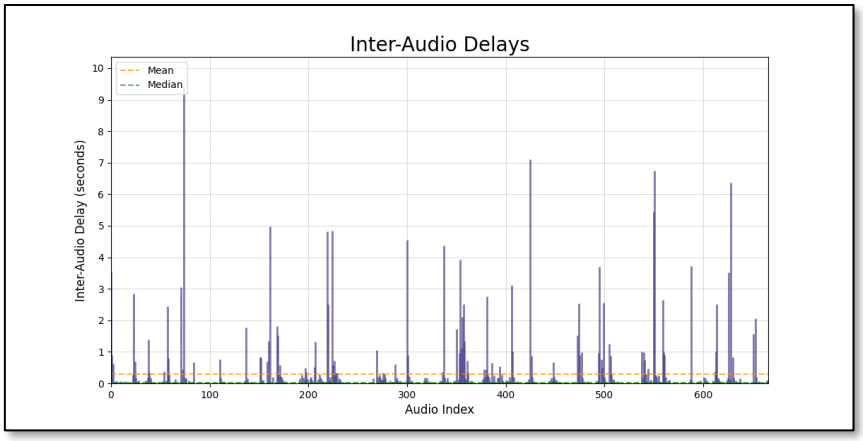


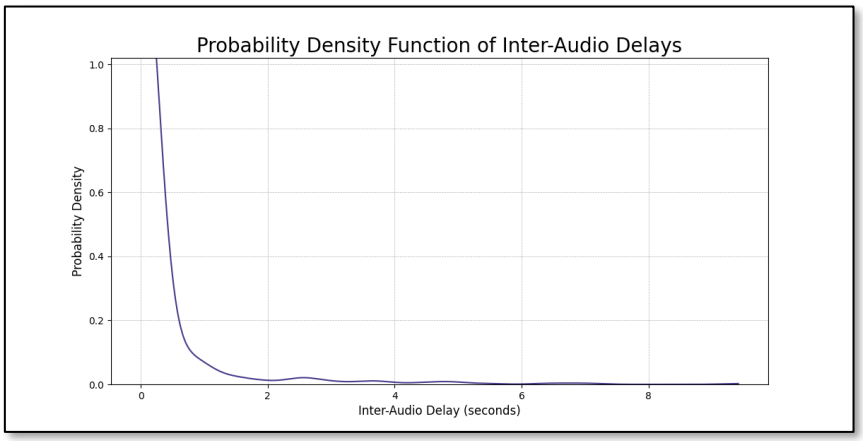


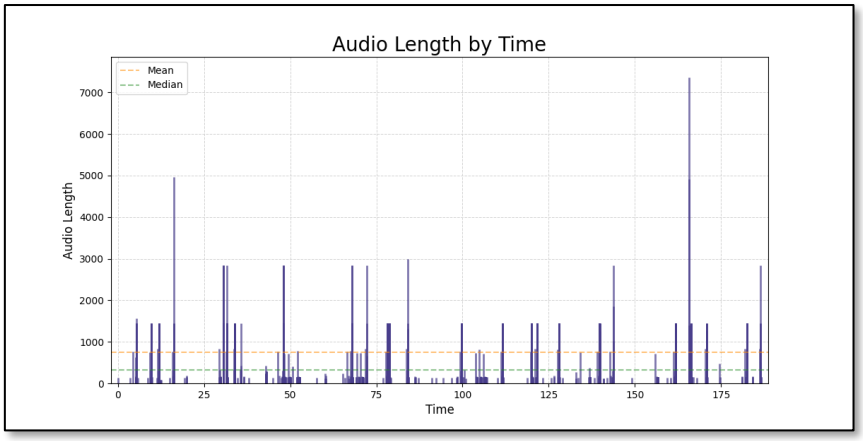


ראשית, נשים לב כי החבילות בתעבורה הנוכחית גדולות יותר (1500~ בייטס) לעומת ההודעות (200~ בייטס). בנוסף, ניתן לראות כי הפרשי הזמנים של העברת התמונות (1-13 שניות) ארוכים יותר מאשר הפרשי הזמנים של ההודעות (0-5 שניות). להבנתנו, תופעה זו מתרחשת מכיוון שבאופן טבעי – תמונות "שוקלות" יותר מהודעות טקסט (בייטס בודדים), במיוחד תמונות עם איכות גבוהה, ולכן גם זמן ההעברה ארוך יותר, מה שגורם להפרשי זמנים ארוכים יותר. נציין כי ישנה חריגה אחת יוצאת דופן (20~ שניות, בגודל 7,000 בייטס) שככל הנראה נובעת כתוצאה מתמונה שנשלחה באיכות גבוהה מאוד או "רעש" כפי שהוסבר בפסקה הקודמת.

3. קבצי אודיו:

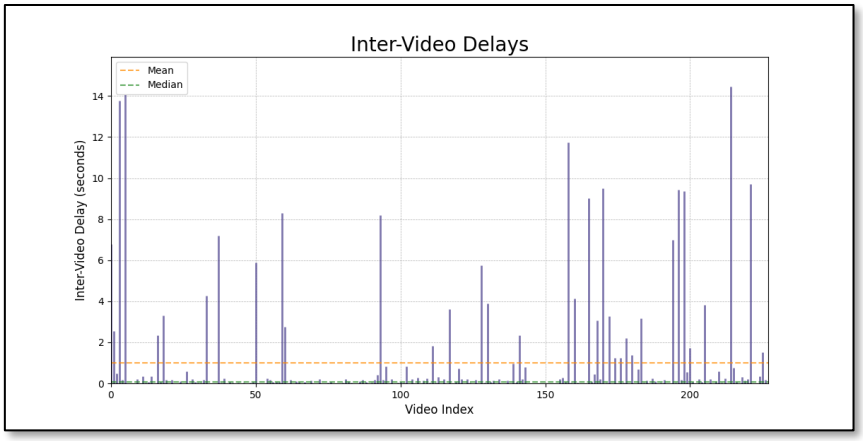


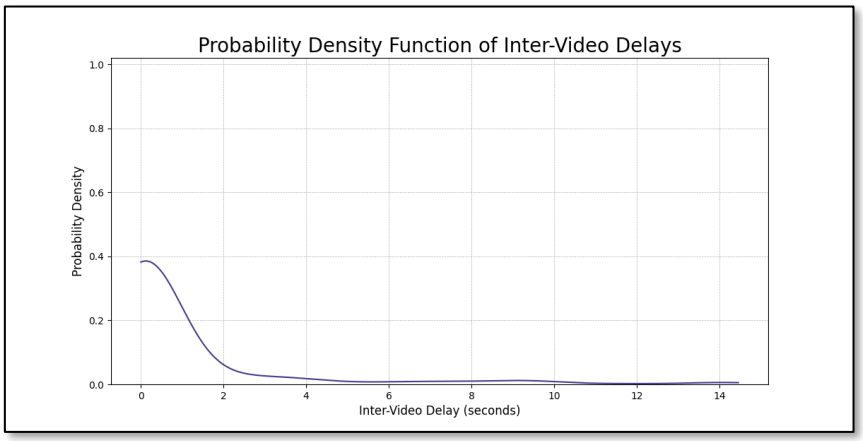


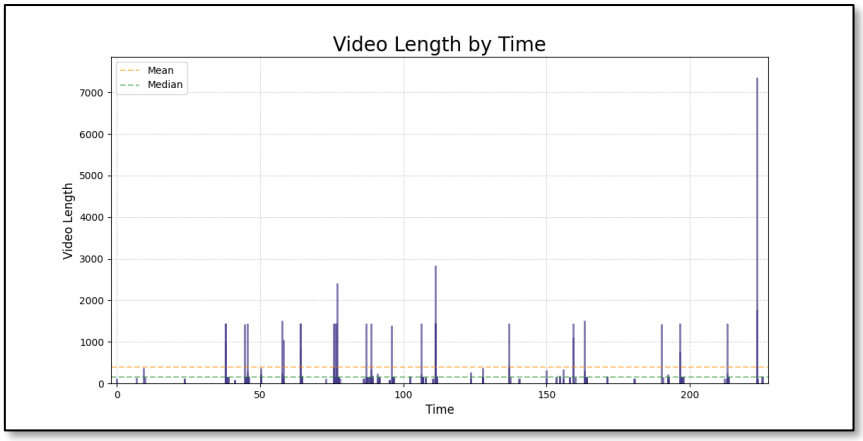


נשים לב כי ישנו גיוון גדול בגרף המציג את הפרשי הזמנים בין החבילות וכן בגרף המציג את גודל החבילות. להבנתנו, שינוי זה נובע מהעבודה הפשוטה שמשתתף המקליט הקלטה לא יכול לשלוח הודעות בזמן שהוא מקליט, ולכן למשל אם הקליט הקלטה באורך s שניות אזי במקרה ואף משתמש אחר לא שלח הודעה נוצר בהכרח הפרש של s שניות. כמו כן, גודלי החבילות מגוונים מאוד משום שטווח הזמנים של קבצי האודיו שנשלחו הוא רחב (1-80 שניות), ובהינתן שכל קבצי האודיו בוצעו מאותם מכשירים (באותה איכות) – אזי משך ההקלטה הוא הפרמטר המרכזי לגודלה.

4. סרטונים:



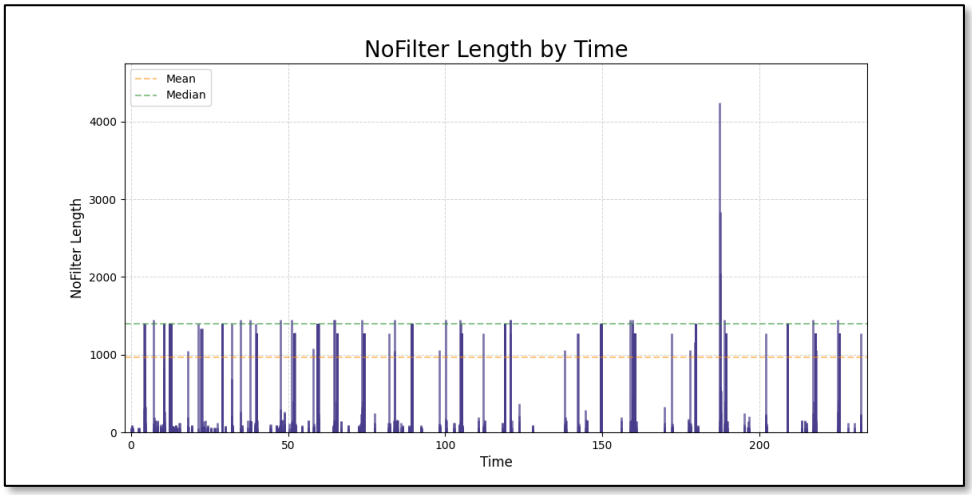


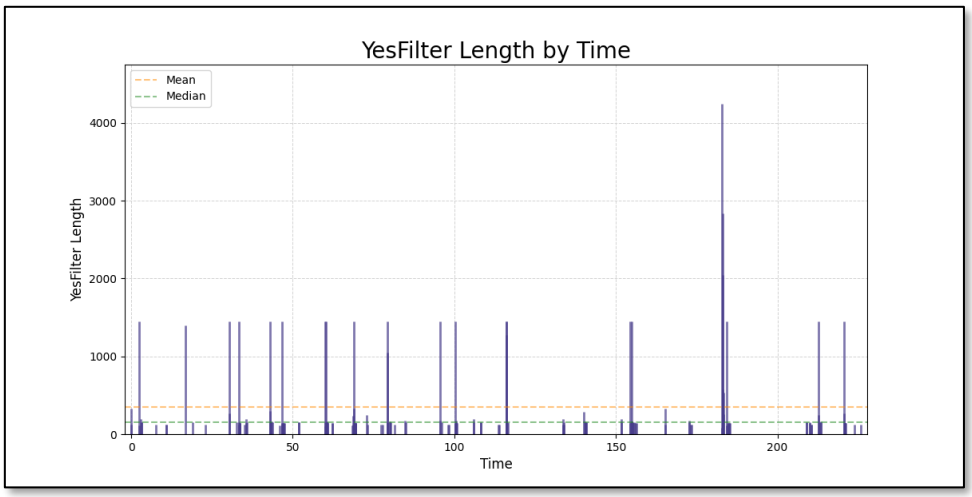


ראשית, נשים לב כי הפרשי ההודעות מאוד גדולים (בהינתן סינון רעשים, רוב ההפרשים נמשכים 2-14 שניות). כמו כן, גדלי החבילות הם גדולים יחסית - בהינתן סינון רעשים, כל חבילה שוקלת ~1400 בייטס לכל הפחות. להערכתנו, ההפרשים הגדולים נובעים מכך שנדרש לכל סרטון כמה שניות כדי להישלח מהמשתמש לשרת וכמה שניות נוספות מהשרת ליתר המשתמשים בקבוצה.

כעת נבחין כי בהינתן שהקלטנו את כל תעבורת הרשת של המשתמש נוכל להעריך – גם ללא פילטור – אם הוא משתתף בקבוצה מסוימת או לא. לצורך הניסוי, הקלטנו את כל תעבורת הרשת של משתמש בקבוצה מסוימת ושלחנו תמונות בקבוצה זו. במקביל, המשתמש הפעיל מחיקה דרך הדפדפן (YouTube Music) והשתמש ברשת כרגיל.

להלן שתי תמונות, הראשונה מציגה את תעבורת הרשת ללא סינון והשנייה מציגה את תעבורת הרשת עם סינון בהתאם לאמור בתחילת חלק 2 (פורט 443, שרת של Meta ופרוטוקול TLS).





ניתן לראות שאמנם אין קורלציה חד משמעית בין התעבורות, עם זאת ניתן לראות בקלות כי חבילה משמעותית המופיעה בתעבורה המסוננת מופיעה בהכרח בתעבורה שאינה מסוננת. כלומר, תוקף שברשותו יכולת לשלוח הודעות בקבוצה ובנוסף נגיש לתעבורה של המשתמש יכול לבדוק התאמה בקלות – האם חבילה שהוא שלח בזמן t ובגודל l מופיעה גם בתעבורת הרשת של המשתמש בזמן t ובגודל l.