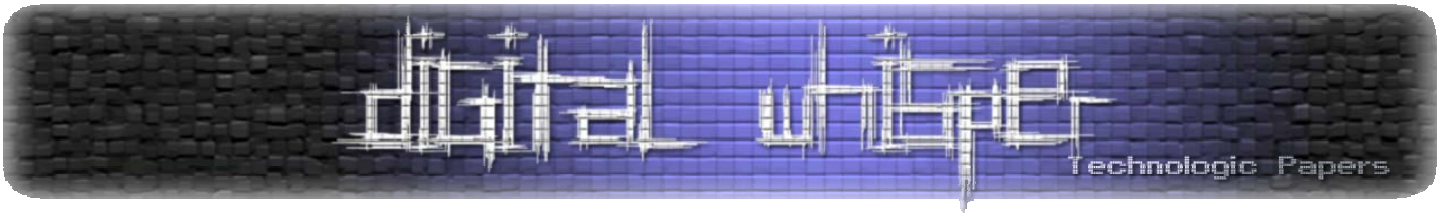


אנונימיות בעידן הדיגיטלי

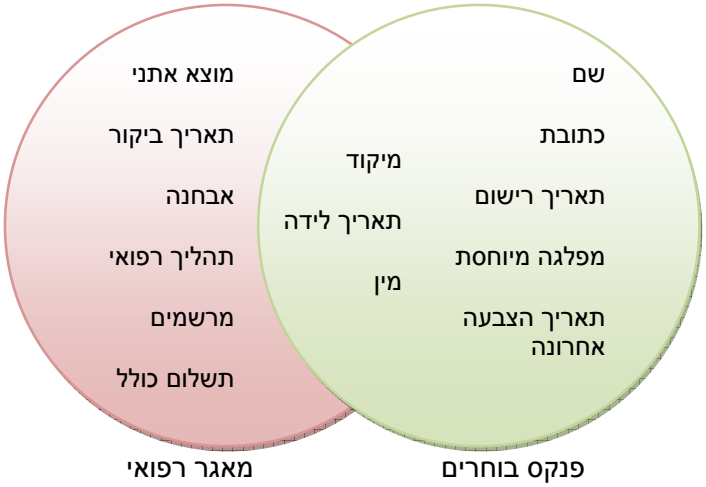
מאת אריק פרידמן

במהלך שני העשורים האחרונים חווינו התפתחות עצומה בטכנולוגיית המידע. בעבר היכולת לאסוף ולשמור כמויות גדולות של נתונים אודות אנשים פרטיים הייתה שמורה אך ורק לממשלות או לארגונים גדולים כגון בנקים. בעקבות מהפכת האינטרנט, בד בבד עם ירידות חדות בעלויות חומרה, יכולות אלה זמינות לכל גוף המעוניין להקים אתר ולהציע שירות כלשהו באינטרנט – אתר מכירות האוסף נתונים על לקוחותיו; אתר תוכן (חדשות, בלוגים) הבוחן את מידת העניין של המשתמשים בתכנים השונים וכן הלאה. במקרים רבים, לגופים אלה יש תמריץ כלכלי לעשות שימוש במידע המצטבר, לרוב על ידי שיתופו או מכירתו לגורמים אחרים. עם זאת, משיקולי פרטיות, הנתונים אינם משותפים עם גורמים אלה בצורתם הגולמית אלא הם עוברים תחילה עיבוד כלשהו לצורך הסתרת זהות הלקוחות ורק אז מועברים הלאה. אף על פי כן, מספר מקרים מהשנים האחרונות מצביעים על בעיתיות בגישה זו. נראה כי שיטות "מסורתיות" להבטחת אנונימיות, הבאות בדרך כלל לידי ביטוי בהסרה של מספר פרטים מזהים, אינן עומדות במבחן המציאות. במאמר זה נסקור את המקרים הבולטים מהשנים האחרונות על מנת להמחיש עד כמה קשה להשיג אנונימיות אמיתית בעידן הדיגיטלי.

התיק הרפואי של מושל מסצ'וסטס

לפני קצת יותר מעשור חוקרת בשם לטניה סוויני הדגימה כיצד בהינתן רשומות מידע שהן לכאורה אנונימיות ניתן לשחזר את זהות בעלי הרשומות. סוויני קיבלה לרשותה מאגר נתונים מידי הועדה לביטוח קבוצתי (GIC – Group Insurance Commission) של מסצ'וסטס. ארגון זה אחראי לרכישת ביטוח רפואי עבור עובדי המדינה. במסגרת עבודתו, הארגון אסף מידע רפואי עבור כ-135,000 עובדי מדינה ומשפחותיהם. כדי לשמור על פרטיות המטופלים, הוסרו ממאגר המידע פרטים מזהים כגון שם, מספר טלפון וכתובת. מכיוון שהמידע הנותר נחשב אנונימי, הארגון ספק עותק של המידע לחוקרים וכן מכר עותקים שלו לתעשיה. למרות הסרת הנתונים המזהים, סוויני גילתה כי ניתן לשחזר את זהות המטופלים בקלות יחסית. בעלות של 20$, רכשה סוויני מידי המדינה את פנקס הבוחרים של קמברידג', מסצ'וסטס (בניגוד לארץ, בארה"ב מידע מסוג זה זמין לציבור לשימושים לא מסחריים). בין השאר, רשימת הבוחרים

הכילה עבור כל בוחר נתוני תאריך לידה, מין ומיקוד. מאחר ונתונים אלו לא הוסרו מהמאגר של GIC, ניתן היה להצליב בין שני המאגרים וכך לקשור זהות של עובד מדינה המופיע ברשימת הבוחרים לנתוניו הרפואיים במאגר של GIC. התברר כי הצלבה זו יעילה מאוד. ויליאם ולד, מושל מסצ'וסטס באותו זמן, התגורר בקמברידג', ולכן רשומותיו נכללו ברשימת הבוחרים שסוויני רכשה. על-פי רשימת הבוחרים, ששה אנשים בלבד במאגר של GIC היו בעלי אותו מיקוד כמו המושל; רק שלושה מהם היו גברים; מבין השלושה, המושל ולד היה היחיד בעל המיקוד המתאים. באופן כללי, סוויני העריכה כי כ-87% מתושבי ארצות הברית ניתנים לזיהוי ייחודי על-בסיס תאריך הלידה, המין והמיקוד שלהם. בעבודה מאוחרת יותר, חוקר נוסף בשם פיליפ גול נקט בהערכה זהירה יותר על-פיה "רק" 67% מתושבי ארצות הברית חשופים לזיהוי מסוג זה. יש לציין שעבור הנותרים, גם אם לא ניתן להגיע לזיהוי ייחודי, ניתן לצמצם משמעותית את רשימת ה"חשודים", לרוב לרשימה של לכל היותר חמישה אנשים.



על מנת להתמודד עם הבעיה, סוויני הציעה גישה לעיבוד מידע לפני הפצתו (להלן, אנונימיזציה של הנתונים) באופן שימנע את הצלחתן של הצלבות נתונים כפי שביצעה היא עצמה. כדי לוודא כי כל הצלבה כזו תותיר ברשימת החשודים לפחות k אנשים, על מפרסם הנתונים לדאוג שכל רשומה במאגר תהיה זהה לפחות ל-k-1 רשומות אחרות במאגר. מודל זה כונה בשם k-anonymity. המחקר של סוויני הכה גלים בעולם האקדמי וגרם אושר למדעני מחשב רבים שחיככו ידיהם בהנאה וניגשו להתמודד עם בעיית האנונימיזציה, שהיא לגמרי לא טריוויאלית. חלקם הציעו הצעות כיצד ניתן לבצע אנונימיזציה יעילה, אחרים הסבירו לראשונים שבעצם הגישה שלהם לא עובדת וצריך לנקוט בגישה אחרת לחלוטין ובתורם גילו שהפתרון לא מושלם כשגם שיטתם נשברה. בתהליך זה, הוצע להחליף את k-anonymity ב-l-diversity, שבתורה הוצע להחליפה ב-t-closeness ואולי m-invariance. המרוץ לכלות את אותיות האלפבית האנגלי נגדע באיבו כאשר גישה אחרת לבעיית הפרטיות הצביעה על נקודת תורפה הנוגעת לכל קו המחקרים המדוברים, אך מדובר בנושא לדיון נפרד.

פרטיות בשאילתות חיפוש

אחד המקרים המשמעותיים בהם בעיית האנונימיות נחשפה לציבור הרחב התרחש באוגוסט 2006. שני עובדים ב-AOL (America Online), אחת מספקיות האינטרנט הגדולות בארה"ב, החליטו לגלות יוזמה ולהעלות לאינטרנט, לטובת המין האנושי, כ-20 מליון שאילתות חיפוש. שאילתות אלה בוצעו על ידי 657,000 מלקוחותיה של AOL בתקופה של שלושה חודשים. יוזמה זו זיכתה את AOL במקום 57 ברשימה 101 הרגעים הטפשיים בעסקים של CNN מ-2007, ונתנה השראה למחזה ("AOL user 927"), לסרט ("I love Alaska") ולתביעה ייצוגית.

ראשית, אין להפחית בערכה של הכוונה הנעלה שמאחורי צעד אמיץ זה. רשומות חיפוש הן מידע הקיים בנפח משמעותי אך ורק בידיהן של חברות פרטיות, ולציבור הרחב אין גישה למידע מסוג זה. מדובר במידע יקר ערך עבור חוקרים מתחומים שונים, במיוחד כאלה שאינם עובדים עבור החברות הפרטיות הנכונות או עבור ארגון ממשלתי. ניתן להפיק מרשומות החיפוש תובנות רבות לגבי סוג המידע שאנשים מחפשים ברשת וכיצד הם מאתרים אותו, תהליכי קבלת ההחלטות של אנשים, כיצד הם מסתגלים לאורך זמן לעבודה עם מנועי החיפוש וכן הלאה.

עם זאת, לערך הרב שניתן להפיק משיתוף המידע יש גם מחיר – רשומות החיפוש לפרסום הן שאילתות חיפוש אמיתיות השייכות לאנשים אמיתיים. כאשר אנשים משתמשים במנוע חיפוש הם יוצאים מנקודת הנחה ששאילתות אלה הן מידע פרטי שאינו נחשף כלפי חוץ. חשיפת שאילתות החיפוש ללא אישורם של המשתמשים שביצעו אותן מהווה פגיעה באמונם של המשתמשים והפרת מחויבות מצד החברה לשמור על חסיון הנתונים. עובדי AOL היו מודעים להשלכה זו ולכן דאגו להסיר סימנים מזהים משאילתות החיפוש: שמות המשתמש הוסרו ובמקומם נעשה שימוש במספרים אקראיים כדי להתייחס למשתמשים.

הרגישות של הנתונים התבררה זמן קצר לאחר פרסום רשומות החיפוש, בלוגרים רבים החלו לדווח על המציאות שמצאו במאגר המידע (cnet news ריכזו מספר דוגמאות מעניינות). המאגר המחיש כי שאילתות החיפוש המצטברות במנועי החיפוש עשויות להיות בעלות אופי פלילי ("פורנו ילדים", "להרוג את המאהב של אשתי"), רגישות ואישיות ("אשתי לא אוהבת אותי יותר", "דכאון וחופשת מחלה") או סתם מביכות ("הראל סקעת שולת"). אין שום ספק שזה לא סוג המידע שאנשים יהיו מעוניינים שייקשר לזהותם. בעקבות התגובות בבלוגים, חברת AOL נוכחה בטעותה והזדרזה להסיר את המידע מעל אתר האינטרנט. למרות זאת, בשלב זה רבים כבר הורידו אליהם את המאגר, כך שלא ניתן היה באמת להחזיר את השד לבקבוק, והמידע עדיין זמין באינטרנט.

מבחן המציאות הראה כי עובדי AOL לא היו זהירים מספיק. שני עיתונאים מה-New York Times, מיכאל בארברו ותום זלר, הריחו סיפור מעניין והחליטו לבדוק מה אפשר ללמוד משאילתות החיפוש. הוחלט להתמקד באוסף שאילתות חיפוש של משתמש 4417749 ולבדוק מה יוכלו ללמוד על אותה ישות אנונימית. לאורך שלושת החודשים מהם נאספו השאילתות, משתמש 4417749 ביצע שאילתות כגון "numb fingers", "60 single men", " dog that urinates on everything".

תלמה ארנולד (משתמשת מספר 4417749) והכלב דאדלי

ע"י בחינת שאילתות החיפוש הם ליקטו פיסות מידע אחת לאחת, ובחנו מה כל שאילתה יכולה ללמד על מבצעה. למשל, שאילתות כגון "landscapers in Lilburn, Ga" הסגירו מקום מגורים, ומספר שאילתות על אנשים ששם משפחתם "Arnold" רמזו על זהות המשתמש. העיתונאים לא נדרשו למאמצים רבים עד שנקשו על דלתה של תלמה ארנולד, אלמנה בת 62 מלילבורן, ג'ורג'יה. גב' ארנולד המופתעת אישרה שאכן שאילתות החיפוש המדוברות הן שלה.

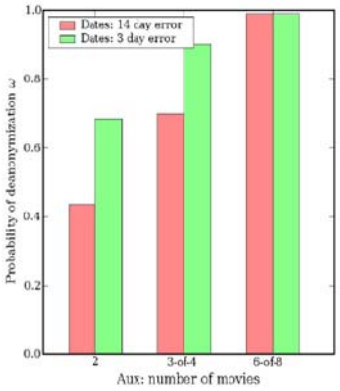
המאמר של הניו-יורק טיימס שחשף את מקרה זה עורר מהומה לא קטנה. הש.ג. כמובן שלם את המחיר – החברה טענה שפרסום המאגר היה יוזמה אישית של אחד העובדים וללא אישור. החוקר האחראי על פרסום המידע פוטר לאלתר ביחד עם המנהל שלו. עם זאת, כעבור שבועיים המנהל הטכנולוגי הראשי של החברה התפטר מתפקידו.

נטפליקס ותיק ברוקבק

במקרה של ה-GIC, לצורך שחזור הזהויות נעשה שימוש בנתונים "מזהים למחצה" כגון תאריך לידה ומין. במקרה של רשומות החיפוש של AOL, שאילתות החיפוש עצמן הכילו מידע רב שאפשר ללמוד על מבצע החיפושים ולהסגיר לבסוף את זהותו. אחת הדוגמאות המפתיעות לגבי שבריריותה של האנונימיות הגיעה כאשר חברת נטפליקס פרסמה מאגר המכיל דירוגי סרטים, חודשיים בלבד לאחר השערוריה של AOL.

נטפליקס היא חברה ידועה בארה"ב העוסקת בהשכרת סרטים. בין השאר, חברה זו מאפשרת לקהילת המשתתפים לדרג את הסרטים הנצפים. דירוגים אלה משמשים את נטפליקס כדי להמליץ ללקוחותיה על סרטים נוספים (בדומה להמלצות הספרים שאמזון מספקת ללקוחותיה, למשל). באוקטובר 2006 נטפליקס הכריזה על תחרות (Netflix Prize) שמטרתה לשפר את אלגוריתם ההמלצות שלה. הובטח פרס של מליון דולר לקבוצה שתשפר את אלגוריתם ההמלצות של נטפליקס בלפחות 10%. התחרות הייתה פתוחה לכל המעוניין (למעט עובדי החברה ומקורביהם). מי שנרשם לתחרות קיבל אפשרות להוריד מאגר נתונים שהכיל כ-100 מליון דירוגים על כ-18,000 סרטים. דירוגים אלה נעשו על ידי בערך 500,000 מלקוחות החברה לאורך 7 שנים. בממוצע, כל לקוח במאגר דירג כ-200 סרטים, וכל סרט דורג ע"י מעל ל-5000 לקוחות. לכל אחד מהסרטים, הדירוגים ניתנו כשלשות מהצורה <מזהה משתמש, דירוג, תאריך>, כאשר דירוג הוא מספר שלם בין 1 ל-5.

נטפליקס התחשבה כמובן בפרטיות לקוחותיה- לא רק שנעשה שימוש במספרים אקראיים כמזהים במקום שמות המשתמשים אלא, לטענת החברה, גם הוכנסו שגיאות מכוונות בחלק מהנתונים כדי למנוע דמיון מושלם לנתונים האמיתיים. על-פניו, נראה כי מאגר הנתונים שפורסם אכן מבטיח אנונימיות. הרי דירוג של סרט הוא נתון "ניטרלי" שאין דבר בינו לבין זהות המשתמש. גם תאריך אותו הדירוג אינו מעיד על המשתמש. למרות זאת, שני חוקרים מאוניברסיטת טקסס באוסטין, ויטאלי שמטיקוב וארווינד נריאנאן, הראו כיצד המידע התמים הזה יכול לשמש כדי לזהות את הלקוחות המדרגים. החוקרים שמו לב כי דירוגי הסרטים הינם נתונים "דלילים" מאוד. מתוך מאגר של 18,000 סרטים, כל משתמש בודד רואה מספר מצומצם יחסית של סרטים. למרות שישנם מספר סרטים שהינם פופולריים ביותר (אוואטר, גבעת חלפון), בפועל עבור כל זוג אנשים תהיה חפיפה מאוד קטנה בין אוספי הסרטים שבהם צפו. למשל, עבור הרוב המוחלט של הלקוחות במאגר נטפליקס לא ניתן למצוא אפילו לקוח אחר יחיד מבין חצי מליון הלקוחות, עם חפיפה של 50% או יותר בסרטים הנצפים.



הסיכוי לנחש את זהות המשתמש כאשר ידועים הדירוגים המדויקים שנתן למספר סרטים, ותאריך מתן הדירוג ידוע בטווח שגיאה של שלושה או 14 יום.

מצויידים באבחנה זו, ניסו החוקרים להעריך עד כמה קל או קשה יהיה לזהות משתמש נטפליקס בהינתן מידע מוקדם מתאים על העדפותיו בסרטים. לצורך זה בחנו שיטות שונות באמצעותן ניתן להתאים אוסף דירוגים משוערים למשתמש נטפליקס. בהינתן אוסף דירוגים משוער, חישבו עד כמה הדירוגים המשוערים דומים לדירוגים של כל אחד מהלקוחות המופיעים במאגר. על מנת לקבל תוצאות טובות יותר, הסרטים הנצפים פחות קיבלו משקל גדול יותר בחישוב הדמיון. דמיון רב הוא אינדיקציה להתאמה בין הדירוג המשוער לבין לקוח נטפליקס. על מנת להימנע מהתאמות כוזבות, דרשו כי עבור המועמד הבא בתור להתאמה (כלומר משתמש נטפליקס שציון ההתאמה שלו לדירוגים המשוערים הינו השני בגודלו) החישוב של הדמיון יתן תוצאה שהיא גרועה יותר משמעותית. הסיכוי לנחש נכון את זהות המשתמש יכול להשתנות כתלות בשיטה המדויקת בה משתמשים, וכתלות ברמת הדיוק של הדירוגים המשוערים. אולם המסקנה שעלתה מהניסויים שערכו החוקרים הייתה שבאמצעות ידע מוקדם מועט ניתן לזהות באופן חד-משמעי את המשתמש הממוצע במאגר נטפליקס. לדוגמה, בהינתן 8 דירוגי סרטים (מתוכם שניים עשויים להיות מוטעים לחלוטין) ותאריכי דירוג בטווח שגיאה של 14 יום, ניתן לזהות באופן ייחודי 99% מהרשומות.

בשלב הבא, רצו החוקרים להעמיד את שיטתם למבחן המציאות, כל שנדרש לחוקרים לטובת המשימה היה למצוא מקור מידע טוב שניתן להצליב עם המאגר של נטפליקס. בנסיבות רגילות, ניתן ללמוד מידע כזה בשיחות מסדרון שגרתיות עם הקורבן אותו מעוניינים לזהות במאגר ("ראית איזה סרט טוב לאחרונה?"). לצורך הוכחת יכולת, החוקרים החליטו להשתמש במאגר אחר שהיה זמין, ואספו באקראי מאתר IMDb (Internet Movie Database) 50 רשומות של משתמשים שהזדהו בשמם (תנאיי השימוש של IMDb מנעו מהם לאסוף מספר רב של רשומות באופן אוטומטי). ההנחה שלהם הייתה שלפחות חלק מהמשתמשים של נטפליקס מחזיקים גם חשבון ב-IMDb. לכל אחד מהמשתמשים שבחרו מ-IMDb הייתה בידם רשימת הסרטים עליהם המליץ ב-IMDb. למרות הפער הגדול הצפוי בין שני המאגרים (לא ברור מה מידת החפיפה בין לקוחות האתרים, לא ברור מה מידת הדמיון בין הדירוגים שלקוח מספק בשני האתרים), עבור שתיים מהרשומות שנדגמו החוקרים מצאו התאמה משמעותית בין הדירוגים ב-IMDb לבין דירוגיו של לקוח במאגר נטפליקס.

על-פניו, לא נראה כי נזק כלשהו יכול להיגרם מחשיפתם של דירוגי סרטים. עם זאת, לקוח המספק דירוגים תחת שמו באתר IMDb, עשוי להימנע מדירוג פומבי של סרטים מסויימים המעידים על אמונות או העדפות מסוגים שונים (כגון סרטי דת, סרטים פוליטיים, סרטים המזוהים עם הקהילה הגאה וכן הלאה). אותו לקוח עשוי היה לדרג את אותם סרטים באופן פרטי בחשבונו בנטפליקס מתוך אמונה כיחשבונו חסוי. לפיכך פרסום המאגר עשוי להביא לחשיפה לא רצויה של הלקוח המדרג. על-בסיס עקרון זה, שזכה לכינוי The Brokeback mountain factor, הוגשה לאחרונה תביעה ייצוגית, מאחוריה עומדת לסבית בארון, אם לשני ילדים, החוששת להמשך קיום אורח חייה לאור ממצאים אלה. על-פי הבלוג הרשמי של נטפליקס התביעה יושבה אך בעקבות התביעה החברה גנזה לעת עתה את תכניותיה לתחרות המשך.

הערת שוליים: בספטמבר 2009 הוכרז על הקבוצה המנצחת שגרפה את הפרס הגדול, הכוללת בין השאר את הישראלי יהודה קורן.

מילות סיכום

עוד בינואר 1999, מנכ"ל Sun לשעבר, סקוט מקנילי, טען " You have zero privacy anyway... get over it!". לאחרונה נשמעו טענות ברוח זו גם ממנכ"ל פייסבוק, מארק צוקרברג, וממנכ"ל גוגל, אריק שמידט, האחרון טען כי אם למישהו יש משהו שאינו מעוניין לגלות, כנראה שעליו להימנע מלעשות את אותו משהו מלכתחילה. אמירות אלה מדאיגות למדי, בהתחשב בכך שאנשים אלה אחראים על מאגרי מידע עצומים שרבים מאיתנו משתמשים בהם לאחסון מידע אישי ופרטי.

ולמרות זאת נראה כי הקרב אינו אבוד- מחאות המוניות המגיעות בתגובה למקרים כגון פרוייקט Beacon של פייסבוק או בתגובה ל-Google Buzz של גוגל מלמדות כי עדיין יש ערך לפרטיות וכי ליחידים יש יכולת להשפיע על המדיניות שנוקטות החברות הגדולות ביחסן למידע פרטי. למרות התפשטותן של הרשתות החברתיות והנטייה הגוברת של אנשים לחשוף על עצמם מידע באופן פומבי, אין פירוש הדבר כי אנשים ויתרו לחלוטין על פרטיותם. גם בתחום האקדמי, חוקרים מקהילות שונות – מדעני מחשב, סטטיסטיקאים, משפטנים, כלכלנים וסוציולוגים, ממשיכים ומרחיבים את ההבנה לגבי משמעות הפרטיות וכיצד ניתן לשמר אותה בעולם שבו המידע נעשה זמין כל כך. אף על פי כן, הצעד הראשון להבטחת הפרטיות ברשת מתחיל אצל כל אחד ואחד מאיתנו – באמצעות מודעות לערכו של מידע נוכל לקבל החלטות מושכלות יותר לגבי המידע שאנו בוחרים לשתף ברשת.