בס"ד טיוטה [הגברה 2ג] אייר התשע"ד
עדכון: אלול התשע"ד. עדכון נוסף: תשרי התשע"ה. עדכון: חנוכה התשע"ה.

סיגנאל של מערכת צפני *ELS*:
דעיכה מול הגברה

מאת דורון ויצטום

תקציר
סיגנאל הנוצר באופן מלאכותי באמצעות "תפירה" של מערכת צפני *ELS* (*Equidistant Letter Sequences*) בקטע מוגדר מטקסט, צפוי לדעוך כאשר המדידות לזיהויו נעשות בקטע ארוך יותר, המכיל את הקטע המקורי.

הדבר נבדק עבור רשימת הנתונים שנמדדה בעבודת *MBBK* [1] בספר "מלחמה ושלום" [2]. תחום המדידה המקורי היה קטע בן 78064 אותיות מתחילת הספר הנ"ל. תחום המדידה החדש הינו קטע מתחילת "מלחמה ושלום", הכפול באורכו מתחום המדידה המקורי. מתברר כי בתחום החדש הסיגנאל אכן דועך באופן בולט.

לעומת זאת, בדיקה דומה בספר התורה עבור רשימת הנתונים שנמדדה בעבודת *WRR* [3] בספר בראשית, העלתה תוצאה הפוכה. תחום המדידה המקורי היה קטע בן 78064 אותיות מתחילת ספר התורה (כלומר, כל ספר בראשית). תחום המדידה החדש הינו קטע מתחילת ספר התורה, הכפול באורכו מתחום המדידה המקורי. מתברר, כי בתחום המדידה החדש הסיגנאל מתחזק באופן בולט במקום לדעוך.

מבוא
טענתם המרכזית של *MBBK* [1] היא, כי נעשתה מניפולציה ברשימת הנתונים על מנת לקבל מובהקות חזקה בניסוי *WRR* [3] בספר בראשית. טענת *MBBK* מתמקדת בתת-רשימה של הנתונים, עבורה נתקבלה המובהקות החזקה ביותר בניסוי *WRR*. נכנה תת-רשימה זו *L2* (ראה פרטים בנספח, סעיף **א**). לגופה של טענה זו כבר ענינו באופן פרטני במקום אחר (ראה [4]). כאן אנו מציעים ניסוי אשר בודק, אף מבלי להיכנס לגוף הטענה, האם *L2* מתנהגת כרשימת נתונים "תפורה".

א. הגדרות:
1. יהא T טקסט, נגדיר קטע בן n אותיות D, בטקסט T כך:

$$D \equiv ([t_{d_1}, t_{d_n}], T)$$

כאשר t_{d_1} הוא המספר הסידורי ב- T של d_1, האות הראשונה בקטע D, ו-t_{d_n} הוא המספר הסידורי ב- T של d_n, האות האחרונה בקטע D.

2. יהא T טקסט, D קטע ממנו, $LIST$ רשימה של זוגות ביטויים, ו- $ELS(LIST, D)$ קבוצת *ELSs* המייצגים את הביטויים ב- $LIST$ והמוגדרים ב- D. בניסוי המקורי של *WRR* ([3], עמ' 434-435), הוגדרה "מידת הקירבה המכוילת" $c(w, w')$ עבור כל זוג ביטויים (w, w'). "מידת הקירבה המכוילת" בין הביטויים, $c(w, w')$, היא מספר בין 0 ל- 1: ערכו קרוב ל- 0 כאשר *ELSs* מינימליים של הביטויים (w, w') נפגשים בצורה מכונסת במיוחד; ערכו קרוב ל-1 כאשר *ELSs* מינימליים של הביטויים רחוקים ומפוזרים במיוחד.

כדי לסכם את "הנטייה הכוללת לקירבה" של $\{c(w, w')| (w, w') \in LIST\}$, הוגדרו בניסוי המקורי של *WRR* ([3], עמ' 436) שתי "מידות לנטייה הכוללת לקירבה" באמצעות שני אופרטורים $\boldsymbol{P_1}$ ו- $\boldsymbol{P_2}$, המיושמים לגבי הטקסט, הקטע והרשימה הנתונים (ראה בנספח, סעיף ב). המידות המתקבלות כאן יהיו:

$$P_i \equiv \boldsymbol{P_i}(LIST, D)$$

ראה בנספח כי גם P_1 וגם P_2 מקבלת ערך הקטֵן ככל שהמפגשים בין צפני ה- *ELS* יותר "מוצלחים".

3. נגדיר סיגנאל S_i באופן הבא:

$$S_i(LIST, D) \equiv 1 / P_i(LIST, D)$$

לכן, הסיגנאל יקבל ערך הגדל ככל שהמפגשים בין צפני ה- ELS יותר "מוצלחים".
(הגדרת הסיגנאלים S_i אינה מחדשת דבר מהותי לעומת ההגדרה המקורית של P_i, אבל עשויה לסייע לתפישה אינטואיטיבית של הנדון כאן, לקוראים מתחומי מדע מגוונים.)
את מובהקות הסיגנאל, כלומר הסיכוי "לקבל סיגנאל כה גבוה במקרה", ניתן לקבוע באמצעות מבחן רנדומיזציה כמו ב[3], או במבחנים עדיפים כמו במאמר הנוכחי (ראה בנספח, סעיף ג).

4. יהא D' קטע ב- T, ארוך מ-D והמכיל אותו:

$$D' \equiv ([t_{d'_1}, t_{d'_{n'}}], T)$$

כאשר: $t_{d'_1} \le t_{d_1}$, $t_{d'_{n'}} \ge t_{d_n}$, $n' > n$.
הסיגנאלים המתקבלים עבור $LIST$ ב- D' יסומנו $S_i(LIST, D')$. אם הסיגנאל של $LIST$ ב- D נתקבל באופן מלאכותי, על ידי מניפולציה של נתוני $LIST$ – אזי במדידה חדשה בקטע D', אנו מצפים לדעיכת הסיגנאל:

$$S_i(LIST, D') < S_i(LIST, D)$$

הסבר: בהגדלת הקטע מ-D ל- D' מקבלים במקום $ELS(LIST, D)$ עליה בוצעה המניפולציה, קבוצה גדולה יותר $ELS(LIST, D')$, שבה $ELSs$ חדשים המופיעים באקראי ויוצרים "רעש" אקראי המקטין את הסיגנאל.

5. נגדיר את מדד ההגברה Q_i: $Q_i \equiv S_i(LIST, D') / S_i(LIST, D)$
$Q_i < 1$ מצביע על דעיכה, היחלשות S_i, ואילו $Q_i > 1$ מצביע על הגברת S_i.
יתכן כי עבור i שונים תתקבל מידת הגברה שונה. אף ייתכן ש- Q_1 יצביע על הגברה, בעוד Q_2 מצביע על דעיכה, או להיפך. אנו מעוניינים במדד ההגברה המירבית Q:

$$Q \equiv \max\{Q_1, Q_2\}$$

ב. דוגמא:
רשימה הנתונים שנמדדה בעבודת $MBBK$ [1] בספר "מלחמה ושלום" הוכנה באופן מוצהר על ידי מניפולציה של נתוני תת-רשימה, אותה נכנה $BM2$ (ראה בנספח, סעיף א). תחום המדידה המקורי, D, היה הקטע [1, 78064] מתחילת הספר הנ"ל. תחום המדידה החדש, D', הינו הקטע [1, 156128] מתחילת "מלחמה ושלום".
לפי מדידה:

$S_1(BM2, D) = 4.9293E+4$, $S_2(BM2, D) = 3.4723E+4$.

לעומת זאת, בקטע המוגדל מקבלים:

$S_1(BM2, D') = 7.1855$, $S_2(BM2, D') = 3.4247E+1$.

מתברר, כי בתחום המוגדל הסיגנאל אכן דועך באופן בולט:

$Q_1 = 1.4577E-4$, $Q_2 = 9.8629E-4$,

ולכן:

$$Q = \max\{Q_1, Q_2\} = 9.8629E-4.$$

ג. מה יקרה אם קיימת הצפנה אמיתית של $LIST$ ב-T?
ישנן שתי אפשרויות:
1. אם תחום ההצפנה של $LIST$ ב-T מוכל בקטע D, אזי הרחבת הקטע מוסיפה רק "רעש" אקראי, ולכן הסיגנאל ייחלש וידעך.
2. אם תחום ההצפנה של $LIST$ ב-T גדול מן הקטע D, אזי הרחבת הקטע לא תחליש את הסיגנאל, וייתכן שאף יתגבר.

ד. לפי נסיון שהצטבר [5] מאז פורסמה עבודת WRR [3] נראה הדבר, כי התופעה לה טענו WRR אינה מוגבלת לספר בראשית, אלא מצוייה בתורה כולה. אנו רואים את כל התורה כחטיבה אחת לעניין התופעה

שלפנינו [6], כאשר נושאים שונים מוצפנים באזורים שונים בספר. בראיה זו, ספר בראשית בו נמדדה רשימה *L2*, אינו אלא הקטע [1, 78064] מספר התורה. נבדוק את *L2* בקטע גדול מספר התורה, המכיל את ספר בראשית. לפי טענת *MBBK* כי *L2* אינה אלא רשימת נתונים "תפורה", אנו מצפים אך ורק לדעיכת הסיגנאל המקורי.

המדידות

מדידה 1: בדיקת *L2* בקטע גדול מספר התורה.

הקטע המקורי, D, היה ספר בראשית (G). אנו צריכים להרחיב את D בקטע "גדול דיו" כדי לצפות לדעיכת הסיגנאל. כ"כלל אצבע" בחרנו ב-D' הכפול באורכו מ- D.
נתוני הקטעים:

- הקטע המקורי D, שהוא ספר בראשית (G), הוא "הקטע [1, 78064] מן התורה" (1 הוא המספר הסדורי של האות הראשונה בתורה, ו- 78064 הוא המספר הסדורי של האות האחרונה בספר בראשית).
- הקטע D' הוא הקטע [1, 156128] מן התורה: $D' = ([1, 156128], TORAH)$

א. ערכנו שתי מדידות:

- בספר בראשית (G).
- בקטע D'.

1. התוצאות בספר בראשית (G):

$S_1(L2, G) = 1.1923E+6, \quad S_2(L2, G) = 1.078E+8.$

2. התוצאות בקטע D':

$S_1(L2, D') = 2.7504E+9, \quad S_2(L2, D') = 2.268E+7.$

מכאן:

$Q_1 = 2.307E+3, \quad Q_2 = 2.104E-1, \quad Q = \max\{Q_1, Q_2\} = 2.307E+3.$

ב. צריך להעריך מהי המובהקות של התוצאה הנ"ל, דהיינו: "מה הסיכוי שיתקבל במקרה Q כה גדול?"

1. לשם כך בנינו אוסף גדול של N טקסטים, T_j, "דומים" לקטע D' מספר התורה, המוגדר בסעיף **א**. הקטע D' בנוי משני חלקים:

- מחלקו הראשון, $D = ([1, 78064], TORAH)$, הוא ספר בראשית בדיוק,
- ומתוספת – הקטע D'' מן התורה: $D'' = ([78065, 156128], TORAH)$.

כל טקסט T_j אף הוא בנוי משני חלקים:

- מחלקו הראשון, $D = ([1, 78064], TORAH)$, הוא ספר בראשית בדיוק,
- ומתוספת – הקטע D''_j - שהוא ערבוב אקראי של הקטע D'' מן התורה (ערבוב אקראי של מלים בתוך פסוקים. הדבר נעשה בדיוק כפי שנעשה ב[3] לגבי טקסט U שם).

2. כאשר מודדים את *L2* בטקסט T_j, התוצאה הסופית "נהנית" מכל היתרונות (והחסרונות) שהיו במדידה המקורית של *L2* בספר בראשית. במלים אחרות: כל רץ T_j נהנה בדיוק מאותה "רוח גבית".
בודקים עבור N טקסטים T_j בכמה טקסטים n מתקבל ערך Q גדול או שווה ל-2,307.
הסיכוי לקבל כזו תוצאה במקרה הוא $p = n/N$.

3. תוצאות:
מתוך 50,000 טקסטים T_j היו רק 48 טקסטים בהם נתקבל Q כה גדול. לפיכך:

$$p < 0.001.$$

מדידה 2: בדיקות ביקורת.

א. היפוך סדר הקטעים:
יצרנו טקסט נוסף, D^*, על ידי הפיכת הסדר של הקטעים D ו–D''. דהיינו:

- החלק הראשון (D^*, [1, 78064]) הוא D'', דהיינו הקטע [78065, 156128] מן התורה.
- החלק הנוסף, (D^*, [78065, 156128]), הוא ספר בראשית, דהיינו הקטע [1, 78064] מן התורה.

התוצאות בקטע D^*:

$S_1(L2, D^*) = 1.0655E+3, \quad S_2(L2, D^*) = 4.00E+4.$

מכאן:

$Q_1 = 8.937E-4, \quad Q_2 = 3.7106E-4, \quad Q = \max\{Q_1, Q_2\} = 8.937E-4.$

ב. מדידה נוספת:
בניסוי המקורי של *WRR* [3] נמדדה המובהקות של "המידות הכוללות לקירבה" P_i באמצעות מבחן פרמוטציות. בכל פרמוטציה מוצמדים באקראי התאריכים של אישיות i מרשימת הנתונים לאישיות j מרשימה זו, וכך לכל $i, j = 1,2,3,\ldots,32$. לכן כל הביטויים הנמצאים ברשימת הפרמוטציה הם שמות, כינויים ותאריכים הנמצאים ברשימת הנתונים המקורית.
התוצאה הסופית נקבעה למעשה על ידי הדירוג של P_4 במבחן הפרמוטציות, שהיה 4 מתוך מיליון. כלומר, היו שלוש פרמוטציות אשר "נצחו" את הרשימה המקורית: הסיגנאל S_4 עבורן היה חזק יותר מן הסיגנאל המקורי. מעניין לבדוק מה יקרה לסיגנאלים אלה כאשר נרחיב את תחום המדידה מספר בראשית לקטע D' (הנ"ל). להלן נשתמש בסימון של המאמר הנוכחי בו האינדקסים 3 ו-4 הוחלפו ל-1 ו-2 בהתאמה.

1. הפרמוטציה שהיתה ראשונה מתוך מיליון פרמוטציות (PER1).

מתברר שהסיגנאל נחלש מאד:

$$Q = \max\{Q_1, Q_2\} = 9.3927E-7.$$

פרטי PER1 והחישובים נמצאים בנספח, חלק ד. כאן רק נציין, כי תוצאה זו מפתיעה. לא זו בלבד שהביטויים ב- PER1 הם השמות, הכינויים והתאריכים הנמצאים ברשימת הביטויים המקורית, אלא אף שני אישים "שהצליחו" בעבודה המקורית במפגשים עם תאריכיהם (מס' 1 ומס 31), נמצאים עם תאריכיהם, ללא כל שינוי, גם ב- PER1.

2. הפרמוטציה שהיתה שניה מתוך מיליון פרמוטציות (PER2).

גם כאן מתברר שהסיגנאל נחלש מאד:

$$Q = \max\{Q_1, Q_2\} = 1.7945E-5.$$

פרטי PER2 והחישובים נמצאים בנספח, חלק ד. מה שנכתב לעיל לגבי PER1 רלוונטי גם כאן. במקרה זה מדובר על ארבעה אישים "שהצליחו" בעבודה המקורית במפגשים עם תאריכיהם (מס' 1 מס' 14 מס' 22 ומס' 23), הנמצאים עם תאריכיהם, ללא כל שינוי, גם ב- PER2.

3. הפרמוטציה שהיתה שלישית מתוך מיליון פרמוטציות (PER3).

מתברר שהסיגנאל נחלש:

$$Q = \max\{Q_1, Q_2\} = 1.95E-3.$$

פרטי PER3 והחישובים נמצאים בנספח, חלק ד. מה שנכתב לעיל לגבי PER1 ו- PER2 רלוונטי גם כאן. במקרה זה מדובר על שלשה אישים "שהצליחו" בעבודה המקורית במפגשים עם תאריכיהם (מס' 14 מס' 23 ומס' 30), הנמצאים עם תאריכיהם, ללא כל שינוי, גם ב- PER3.

ג. מדידה בספר מלחמה ושלום:
נחזור לדוגמה שבסעיף ב של המבוא. כאמור שם, רשימה הנתונים שנמדדה בעבודת *MBBK* [1] בספר "מלחמה ושלום", הוכנה באופן מוצהר על ידי מניפולציה של נתוני תת-רשימה, אותה נכנה *BM2*. תחום המדידה המקורי, D, היה הקטע [1, 78064] מתחילת הספר הנ"ל. תחום המדידה החדש, D', הינו הקטע [1, 156128] מתחילת "מלחמה ושלום".
לפי מדידה מתברר, כי בתחום המוגדל הסיגנאל אכן דועך באופן בולט:

$$Q_1 = 1.4577E\text{-}4, \quad Q_2 = 9.8629E\text{-}4,$$

ולכן:

$$Q = \max\{Q_1, Q_2\} = 9.8629E\text{-}4.$$

לפי בקשת אחד הקוראים, מדדנו מה הסיכוי לקבל מדד הגברה חלש כמו Q_2 או חזק ממנו (כלומר, דעיכה כזו, של הסיגנאל S_2 או קטנה ממנה).
השתמשנו באוסף של 1000 טקסטים. מחציתו הראשונה של כל טקסט מן האוסף הוא הקטע [1, 78064] מתחילת הספר הנ"ל, ומחציתו השניה הוא הקטע [78065, 156128] ממנו, המעורבב באופן אקראי.
מתברר, כי ב- 396 טקסטים נתקבל $Q'_2 \geq Q_2$.

מדידה 3: בדיקת *L2M* בקטע גדול מספר התורה.

במאמרם [1] העלו *MBBK* טענות אחדות בנוגע לתאריכים ולצורות התאריך שננקטו ברשימות הנתונים של *WRR* [3]. מטענות אלה רלוונטית רק טענה אחת ויחידה לגבי תת הרשימה *L2*: כי עבור 4 אישים יש להוסיף או לשנות תאריך (הפרטים בנספח, חלק ה). נגדיר את הרשימה *L2M* כתת הרשימה *L2* עם השינויים הנ"ל בתאריכים (כל שאר התאריכים והפרטים, בפרט השמות והכינויים – נשארים כפי שהם, ללא שינוי).

נחזור על **מדידה 1** כאשר אנו מודדים את *L2M* במקום *L2*.

א. ערכנו שתי מדידות:

- בספר בראשית (G).
- בקטע D'.

1. התוצאות בספר בראשית (G):

$$S_1(L2M, G) = 1.1837E+7, \qquad S_2(L2M, G) = 1.2739E+9.$$

(שים לב כי הסיגנאל עבור *L2M* בספר בראשית חזק פי 10 מאשר הסיגנאל עבור *L2*!)

2. התוצאות בקטע D':

$$S_1(L2M, D') = 2.5023E+12, \qquad S_2(L2M, D') = 2.9411E+9.$$

מכאן:

$$Q_1 = 2.114E+5, \qquad Q_2 = 2.309, \qquad Q = \max\{Q_1, Q_2\} = 2.114E+5.$$

ב. מובהקות התוצאה:
מדידה זו נעשתה בדיוק כמו במדידה 1 סעיף ב.

התוצאות:
מתוך 50,000 טקסטים T_j היו רק 13 טקסטים בהם נתקבל Q כה גדול. לפיכך:

$$p = 0.00026.$$

מסקנות

א. תוצאת הניסוי בתורה מורה, כי הסיגנאל מתחזק בקטע המורחב D'. עובדה זו מפריכה את טענת *MBBK* כי הסיגנאל שנתקבל בקטע D המקורי היה מלאכותי ונוצר ממניפולציה בנתוני $L2$.

ב. תוצאה א של מדידה 2 מראה, כי התגברות הסיגנאל אינה תולדה של סכום ההצפנות בקטעים D בפני עצמו ו–D'' בפני עצמו. סדר החיבור הנכון בין הקטעים הוא קריטי להתגברות הסיגנאל.

ג. נראה, שתחום ההצפנה של $L2$ קרוב יותר לתחום המורחב, D', מאשר לתחום המקורי, D (ספר בראשית). לכן, יש עניין רב לערוך ניסויים נוספים בנושא $L2$ בתחום המורחב. [יש להדגיש, כי אין אנו טוענים כי בתחום D' מוצפנת $L2$ באופן אופטימאלי. דבר זה כלל לא נבדק.]

נספח

א. על תת-הרשימה $L2$:

1. רשימת הנתונים בניסוי *WRR* [3] בנויה מזוגות ביטויים. כל זוג מוגדר משם אישיות (או כינויה) ותאריך הלידה או הפטירה.
2. השמות והכינויים של כל אישיות הם משני סוגים: שמות וכינויים המיוחדים לאותה אישיות, והכינוי הסטנדרטי רבי "פלוני" (כאשר "פלוני" הוא שמו העברי הפרטי).
לדוגמא: לאישיות #1 ברשימה, כינוי "אישי" הראב"י, וכינוי סטנדרטי רבי אברהם. הכינוי רבי אברהם ניתן לכל חכם ששמו הפרטי אברהם, ולכן הוא משותף לארבעת האישים הראשונים ברשימת הנתונים.
3. לכן, ניתן להציג את רשימת הנתונים השלמה, *LIST 2*, כאיחוד של שתי תת-רשימות:
   - תת-הרשימה $L2$ הבנויה מאותם זוגות ביטויים שבהם השמות והכינויים המיוחדים.
   - תת-הרשימה $L'2$ הבנויה מאותם זוגות ביטויים שבהם הכינוי הוא רבי פלוני.

   אבחנה זו נעשתה כבר לפני ביצוע מבחן הפרמוטציות ע"י *WRR* [3].
4. הצלחת הניסוי של *WRR* נקבעה ע"י תת-הרשימה $L2$.
5. טענתם העיקרית של *MBBK* [1] היא, כי תת-הרשימה $L2$ ניתנת "לתפירה" ע"י בחירה/השמטה מכוונת של שמות וכינויים (בניגוד לכינויים הסטנדרטיים רבי פלוני שהם קבועים). לכן, כאשר הכינו באופן מוצהר רשימה "תפורה" של נתונים כדי שתצליח בספר "מלחמה ושלום", רק תת-הרשימה (ללא הכינויים הסטנדרטיים) אותה נכנה *BM2*, תרמה להצלחה.
6. הניסוי הנוכחי נועד לבדוק אם תת-הרשימה $L2$ מתנהגת כרשימה "תפורה".

ב. המידות הכוללות לקירבה:

1. בניסוי המקורי של *WRR* ([3], עמ' 436) הוגדרו ארבע "מידות כוללות לקירבה" P_1, P_2, P_3 ו-P_4. שם ([3], עמ' 431) מבואר כי:
   - מידות P_1 ו-P_2 הוגדרו כשני סטטיסטים שונים לגבי *LIST 2*.
   - כאשר מידות P_1 ו-P_2 מיושמות לגבי $L2$, הן מסומנות כ- P_3 ו- P_4.

   במלים אחרות: ישנם שני אופראטורים, $\boldsymbol{P_1}$ ו- $\boldsymbol{P_2}$.
   היישום שלהם על רשימת הנתונים השלמה נותן את P_1 ו- P_2:

$$P_1 \equiv \boldsymbol{P_1}(LIST\ 2),\quad P_2 \equiv \boldsymbol{P_2}(LIST\ 2)$$

   והיישום שלהם על תת-הרשימה $L2$ נותן את P_3 ו- P_4:

$$P_3 \equiv \boldsymbol{P_1}(L2),\qquad P_4 \equiv \boldsymbol{P_2}(L2)$$

   בניסוי הנוכחי קיימת רק רשימת נתונים אחת $L2$, לכן איננו זקוקים עוד לאינדקסים 3 ו-4, ונסמן כאן:

$$P_1 \equiv \boldsymbol{P_1}(L2),\qquad P_2 \equiv \boldsymbol{P_2}(L2)$$

   וכן לגבי *BM2*.
2. להקל על הקורא נביא כאן את הגדרת P_1 ו- P_2 מ[3].

(א) הגדרת מידת "הנטייה הכוללת לקרבה" P_1.

לפי מידה זו מונים את מספר התוצאות ב"אזור ההצלחה", אשר הוגדר (שרירותית) כמרווח בין 0 ל- 0.2, ומחשבים מה הסיכוי לקבל באקראי את הערך המתקבל. המדגם העומד לבדיקה הוא קבוצה של זוגות ביטויים. "מידת הקרבה המכוילת" של כל זוג ביטויים (w, w') ניתנת לחישוב על ידי $c(w, w')$. כך

מקבלים N מספרים, שכל אחד מהם הוא בין 0 ל- 1. נניח שמספר הזוגות (w,w') עבורם $c(w,w')\le 1/5$ הוא k. נגדיר

$$P_1 \equiv \sum_{j=k}^{N} \binom{N}{j} \left(\frac{1}{5}\right)^j \left(\frac{4}{5}\right)^{N-j}$$

כדי להבין הגדרה זאת, נשים לב לכך, ש**אם** המספרים $c(w,w')$ הם משתנים אקראיים בלתי-תלויים המתפלגים בצורה אחידה (אוניפורמית) בין 0 ל- 1, אזי P_1 היא ההסתברות, כי לפחות k מ-N המספרים - קטנים או שווים ל- 0.2. אומנם, איננו עושים כאן כל שימוש בהנחה של אחידות ואי-תלות. לכן, למרות ש-P_1 מכויל כהסתברות, הוא משמש רק כמדד סידורי. נשים לב, כי המידה P_1 מתעלמת מכל ערכי $c(w,w')$ הגדולים מ- 0.2, ומעניקה אותו המשקל לכל ערכי $c(w,w')$ הקטנים מ- 0.2. כלומר, אנו מתמקדים במפגשים המצליחים ללא הבחנה באיכותם, ואיננו מתעניינים לדעת באיזו מידה נכשלו אלה שלא הצליחו.

(ב) הגדרת מידת "הנטייה הכוללת לקִרבה" P_2.

המידה P_2 נבנתה כך, שהיא רגישה לגודלם של **כל** המספרים $c(w,w')$. המשמעות של P_2 היא, ש**אם** המספרים $c(w,w')$ הם משתנים אקראיים בלתי-תלויים המתפלגים בצורה אחידה בין 0 ל- 1, אזי P_2 היא ההסתברות, שמכפלת ערכי $c(w,w')$ תהיה קטנה כפי שהיא, או קטנה מזה.

אנו מחשבים את המכפלה $\prod c(w,w')$ ואחר כך אנו מגדירים

$$P_2 \equiv F^N\left(\prod c(w,w')\right)$$

כאשר
$$F^N(X) \equiv X\left(1-\ln X+\frac{(-\ln X)^2}{2!}+\cdots+\frac{(-\ln X)^{N-1}}{(N-1)!}\right)$$

כדי להבין הגדרה זו, נשים לב כי אם $x_1, x_2, \ldots, x_N$ הם משתנים אקראיים בלתי-תלויים המתפלגים בצורה אחידה בין 0 ל- 1, אזי ההתפלגות של מכפלתם $X \equiv x_1 x_2 \cdots x_N$ ניתנת על ידי

$$\Pr(X \le X_0) = F^N(X_0)$$

[הדבר נובע מתוצאה (3.5) של פלר [7] מכיוון ש-$-\ln x_i$ מתפלגים באופן אקספוננציאלי, וגם $-\ln X = \sum_i (-\ln x_i)$.]

אומנם, איננו עושים כאן כל שימוש בהנחה של אחידות ואי-תלות. לכן, למרות ש-P_2 מכויל כהסתברות, הוא משמש רק כמדד סידורי.

ג. מדידת מובהקות הסיגנאל

כאשר *WRR* [3] ביצעו בספר בראשית מדידות של רשימות הנתונים, הם מדדו את האפקט הכולל באמצעות מידות כוללות P_i. כדי למדוד את מובהקות התוצאות היה צריך לחזור על אותן מדידות בטקסטים רבים "דומים". דבר זה לא היה בהישג ידם לפני 25 שנים (מבחינת כמות החישובים), ולכן הוצע מבחן הפרמוטציות (להלן - *PT*), אשר סוכם בין פרסי דיאקוניס וישראל אומן, ובוצע ב-[3].

למעשה, כחמש וחצי שנים לאחר ביצוע ניסוי הרנדומיזציה המתואר ב[3], סבר אליהו ריפס שהתקדמותם המהירה של אמצעי החישוב מאפשרת לשוב לרעיון הפשוט ביותר של מדידת המובהקות: באמצעות השוואה לטקסטים רבים "דומים".

במכתב [8] לדוד קשדן הציע אליהו ריפס את העקרונות הבאים לגבי מדידת מובהקות של רשימה בספר בראשית:

- להשתמש ב- 1000 (או יותר) טקסטים T_j, כל T_j נוצר ע"י ערבוב אקראי של מלים בתוך פסוקים (הדבר נעשה בדיוק כפי שנעשה ב[3] לגבי טקסט U).
- להעריך את מובהקות הסטטיסטים P_i בספר בראשית, באמצעות השוואתם לערכי P'_i המתקבלים בטקסטים T_j.

שיטה זו של קביעת המובהקות עדיפה על שיטת *PT* מכמה בחינות. בין השאר לא רלוונטיות לגביה ההסתייגויות שהועלו במשך השנים כנגד שיטת *PT* [1].

ד. פרטי המדידה הנוספת:

1. פרמוטציה PER1:

זו היתה פרמוטציה מס' 808836 במרוץ המקורי:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	17	8	13	25	19	30	29	23	14	5	18	11	10	4	32

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
15	9	2	26	28	12	3	16	24	6	21	22	20	7	31	27

(א) התוצאות בספר בראשית (G):

$S_1(PER1, G) = 1.2680E+7, \quad S_2(PER1, G) = 1.8485E+9.$

(ב) התוצאות בקטע D':

$S_1(PER1, D') = 1.1910E+1, \quad S_2(PER1, D') = 1.9802E+2.$

מכאן:

$Q_1 = 9.3927E-7, \quad Q_2 = 1.0713E-7, \quad Q = \max\{Q_1, Q_2\} = 9.3927E-7.$

2. פרמוטציה PER2:

זו היתה פרמוטציה מס' 788884 במרוץ המקורי:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	20	32	11	17	19	8	29	7	9	15	18	26	14	6	28

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
10	16	2	3	27	22	23	21	4	31	25	30	13	24	12	5

(א) התוצאות בספר בראשית (G):

$S_1(PER2, G) = 3.380E+7, \quad S_2(PER2, G) = 6.8970E+8.$

(ב) התוצאות בקטע D':

$S_1(PER2, D') = 3.6614E+2, \quad S_2(PER2, D') = 1.2376E+4.$

מכאן:

$Q_1 = 1.0833E-5, \quad Q_2 = 1.7945E-5, \quad Q = \max\{Q_1, Q_2\} = 1.7945E-5.$

3. פרמוטציה PER3:

זו היתה פרמוטציה מס' 777442 במרוץ המקורי:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
24	20	17	12	28	19	31	18	25	8	7	27	9	14	32	10

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
22	15	2	4	3	6	23	1	29	11	21	5	13	30	16	26

(א) התוצאות בספר בראשית (G):

$S_1(PER3, G) = 1.1392E+8, \quad S_2(PER3, G) = 1.47E+8.$

(ב) התוצאות בקטע D':

$S_1(PER3, D') = 3.2854E+3, \quad S_2(PER3, D') = 2.86E+5.$

מכאן:

Q_1 = 2.8838E-5, Q_2 = 1.95E-3, $Q = \max\{Q_1, Q_2\}$ = 1.95E-3.

ה. פרטים בנוגע ל-*L2M*:

במאמרם [1] העלו *MBBK* טענות אחדות בנוגע לתאריכים ולצורות התאריך שננקטו ברשימות הנתונים של *WRR* [3]. הם טענו כי ל-*WRR* היה חופש פעולה בתחום זה, וכי *WRR* נקטו תמיד בבחירה שנתנה אופטימיזציה של התוצאות.

במאמרנו "על בחירת התאריכים למדגמים של *WRR*" [9], דחינו את טענות *MBBK*. לעניינו, הראינו כי מטענותיהם רלוונטית אך ורק טענה אחת ויחידה לגבי תת הרשימה *L2*: הטענה כי עבור 4 אישים (מתוך 32 האישים ברשימת הנתונים) יש להוסיף או לשנות תאריך:

1. תיקון תאריך עבור אישיות מס' 20.
2. תיקון תאריך עבור אישיות מס' 21.
3. הוספה של תאריך לידה עבור אישיות מס' 24.
4. הוספה של תאריך לידה עבור אישיות מס' 30.

נגדיר את הרשימה *L2M* כתת הרשימה *L2* עם השינויים הנ"ל בתאריכים (כל שאר הפרטים, בפרט השמות והכינויים – נשארים כפי שהם, ללא שינוי). אגב, בניגוד לטענת *MBBK*, מתברר מן התוצאות כי השינויים בתאריכים אינם גורמים לירידה בסיגנאל.

הכרת תודה

החישובים נעשו באמצעות תוכנה של יעקב רוזנברג. רוברט הראליק עורר אותי לחשב את המובהקות באמצעות השוואה לטקסטים "דומים". יִשרון יצחק לוי ומיכל לוי עזרו בהכנת הקטע שנוסף ל"מלחמה ושלום".

מקורות והערות

1. עבודת *MBBK*:
McKay, B. D., Bar-Natan, D., Bar-Hillel, M. and Kalai, G. (1999). *Solving the Bible Code puzzle*. Statist. Sci. 14 No. 2 150-173.
2. מלחמה ושלום מאת ל. נ. טולסטוי. תרגום לעברית: לאה גולדברג (1953), ספריית פועלים, מרחביה.
3. עבודת *WRR*:
Witztum, D., Rips, E. and Rosenberg, Y. (1994). *Equidistant letter sequences in the Book of Genesis*. Statist. Sci. 9 No. 3 429-438.
4. טענה זו הופרכה כבר בדרכים אחרות, ראה באתר צופן בראשית במדור ההתנגדות למחקר: http://www.torahcode.co.il/oppose_heb.htm, ובמיוחד במאמר הסקירה: http://www.torahcode.co.il/rev1_heb.htm.

5. הנה כמה מן העבודות שנעשו בספר התורה (גם מחוץ לספר בראשית): באתר הנ"ל, במדור "פרסומים מדעיים", http://www.torahcode.co.il/pub_index_heb.htm, מאמרים 5, 8, 9. וכן במאמרים http://www.torahcode.co.il/kriah1_heb.htm, http://www.torahcode.co.il/weis2_heb.htm.

6. זאת ניתן לראות בצפנים המחברים בין החומשים (למשל, בשני הקישורים האחרונים).
7. הספר:
Feller, W. (1966). *An Introduction to Probability Theory and Its Applications* 2. Wiley, New York.
8. בדואר אלקטרוני מיום 15 במאי שעה 16:40:13, שנת 1997 (למניינם).
9. ד. ויצטום (התשס"א): על בחירת התאריכים למדגמים של *WRR*. ראה באתר הנ"ל בקישור: http://www.torahcode.co.il/date_heb.htm.