Samuel Neaman Institute
for National Policy Research

# Profiling Online User Behavior via Triangulated Datasets: Evidence from Survey Data and Digital Trace Data


Researchers:

Sheizaf Rafaeli
Eran Leck
Yael Albo
Yael Oppenheim



Haifa, December 2020

# ABOUT THE SAMUEL NEAMAN INSTITUTE

The Samuel Neaman Institute was established in 1978 in the Technion at Mr. Samuel Neaman's initiative. It is an independent multi-disciplinary national policy research institute. The activity of the institute is focused on issues in science and technology, education, economy and industry, physical infrastructure and social development which determine Israel's national resilience.

National policy research and surveys are executed at the Samuel Neaman Institute and their conclusions and recommendations serve the decision makers at various levels. The policy research is conducted by the faculty and staff of the Technion and scientists from other institutions in Israel and abroad and specialist from the industry.

The research team is chosen according to their professional qualifications and life achievements. In many cases the research is conducted by cooperation with governmental offices and in some cases at the initiative of the Samuel Neaman institute and without direct participation of governmental offices.

So far, the Samuel Neaman Institute has performed hundreds of exploratory national policy research projects and surveys that serve decision makers and professionals in economy and government. In particular the institute plays an important leading role in outlining Israel's national policies in science, technology and higher education.

The Samuel Neaman Institute's various projects and activities can be viewed at the Institute website.

The chairman of Samuel Neaman Institute is Professor **Zehev Tadmor** and the director is Professor **Irad Yavneh**. The institute operates within the framework of a budget funded by Mr. Samuel Neaman in order to incorporate Israel's scientific technological economic and social advancement.

# Table of Contents

# List of Figures

# List of Tables

## Acknowledgments

This research was supported by the Ministry of Science and Technology (MOST). We thank MOST for their generous financial support which made this research possible.

We thank our research partners from the University of Ljubljana, Slovenia – Professor Vasja Vehovar, Dr. Gregor Cheovin and Dr. Nejc Berzelak for the very constructive, pleasant, and fruitful cooperation in the past two years. Their willingness to share their vast knowledge on web surveys and to allow us to use their sophisticated methodological tools (1KA platform) have greatly contributed to the success of this research.

We would like to thank Mr. Alon Talmod, Project Manager at iPanel, Mr. Alon Shlomkovich (CEO) and Didi Grimberg-Zehavi (Development and Information System manager at ipanel) for their professional, technical and methodological assistance of the panel data.

We would like to express our deep gratitude to Dr. Einat Orr, CTO at SimilarWeb, who opened for us the doors of SimilarWeb's fascinating world of tools and data. We would also like to thank Mr. Felix Vaisman, Senior Team Lead at SimilarWeb, for his kind help with problem solving. We would like to thank the Buzzilla team for their technical assistance with the Buzzilla tool.

We thank Ms. Efrat Yaskil, Senior Consultant at the Statistics Consulting Unit -of the University of Haifa for her methodological insights and assistance with the sampling procedure.

We would like to convey our gratitude to Ms. Amal Haddad and Mr. Nasim Khoury for their kind help with the translation of the online surveys to Arabic.

Finally, we thank Mr. Golan Tamir, Information Systems Manager at the Samuel Neaman Institute for his important technical support.

# Hebrew Executive Summary

מחקר זה מציג גישה חדשנית לניתוח ואפיון גורמים סוציו-דמוגרפיים והתנהגותיים של משתמשים ברשת על ידי שילוב של שיטות שאינן פולשניות (ניתוח עקבות דיגיטליים ומדיה חברתית) עם שיטות פולשניות (סקרים מקוונים). המחקר הנו חלק מפרויקט ישראלי-סלובני: "התמרה דיגיטלית באיסוף נתונים כמותיים במחקר אמפירי במדעי החברה: שילוב נתוני סקרים עם ביג דאטה ופארא דאטה לשם זיהוי התנהגות מקוונת" במימון משרד המדע והטכנולוגיה וסוכנות המחקר הסלובנית. המחקר נערך בין התאריכים 1 באוקטובר 2018 ל-30 בספטמבר 2020, בהשתתפות המרכז למידע חברתי באוניברסיטת ליובליאנה שבסלובניה (CSI) ומוסד שמואל נאמן לחקר מדיניות לאומית (SNI). משימות הפרויקט, הן המשותפות והן הנפרדות, עסקו בהיבטים מתודולוגיים ומעשיים של איסוף נתונים מסקרים מקוונים תוך הרחבה ושילוב (טריאנגולציה) עם סוגי נתונים נוספים. קבוצת המחקר ב-CSI התמקדה בעיקר בהיבטים המתודולוגיים של עיצוב סקר מקוון, בפיתוחים בתחום של איסוף פארה-דאטה וכן בהבניית מדדים מורכבים מנתוני פארה-דאטה המיועדים לחקר איכות נתוני סקרים. קבוצת המחקר ב-SNI התמקדה במיפוי פרופיל התנהגות המשתמשים ברשת באמצעות גישת טריאנגולציה אשר כללה שילוב של נתונים הנאספים בשיטות פולשניות (כדוגמת סקר מקוון) ובשיטות בלתי פולשניות (בעזרת כלים המנטרים שימוש ברשת).

בדו"ח זה מפורטים ממצאי המחקר הישראלי, שמטרתו הייתה לתאר, לאפיין, להסביר ולחזות (באמצעות סימולציות נומריות) התנהגות משתמשים ברשת במגוון תחומים כגון קניות, נסיעות ותיירות, ניהול כספי, שימוש ברשתות חברתיות, ופעילות חיפוש ברשת. שילוב נתוני הסקר עם נתוני העקבות הדיגיטליים תרם להעמקת ההבנה של ההתנהגות הנחקרת והבניית מדדים יציבים. גישת הטריאנגולציה ושילוב נתוני הסקר, נתוני עקבות דיגיטליים ונתוני שיח במדיה החברתית הודגמה בחקר-מקרה שבחן תפיסות והתנהגות משתמשים בנושאי פרטיות והגנה על נתונים אישיים. תוצר נוסף של המחקר הנו כלי ויזואלי אינטראקטיבי גנרי לתצוגה וניתוח של נתוני שאלונים. תהליך הפיתוח של כלי זה הניב הצעה לקווים מנחים לעיצוב ויזואליזציה של נתוני סקר.

שני שאלונים גובשו לצורך למידת היבטים התנהגותיים ואפיון של משתמשי אינטרנט. הסקר הראשון ("סקר דו-לאומי") כלל קבוצות משיבים ישראליות וסלובניות והתמקד בתחום הקניות המקוונות וכן בתפיסות של פרטיות ואבטחת מידע ברשת. הסקר השני ("סקר לאומי") כלל משיבים ישראלים בלבד ועסק בהיבטים נוספים של התנהגות מקוונת: בריאות, נסיעות ותיירות, אמון (trust) בטכנולוגיה, ניהול כספי, מאפייני חיפוש ברשת וכן שימוש בטכנולוגיות מידע ותקשורת. שני הסקרים התבססו על מדגם מייצג של אוכלוסיות ישראל וסלובניה בגילאי +18. הנתונים נאספו באמצעות פאנלים של משתמשי אינטרנט ע"י פלטפורמת הסקר הדיגיטלי 1KA בין התאריכים 23/1/2020 ל- 16/2/2020. שתי גרסאות של סקרים הופצו בקרב האוכלוסייה הישראלית תוך שימוש בשני פאנלים ייעודיים נפרדים: גרסה עברית וגרסה ערבית. שימוש במכסות הבטיח ייצוג מספק של תתי אוכלוסיות. מדגם הסקר הדו-לאומי כלל 1283 משיבים ישראלים (1083 דוברי עברית ו -246 דוברי ערבית) ו -4058 משיבים סלובניים. מדגם הסקר הלאומי כלל 1270

משיבים ישראלים (1001 דוברי עברית ו -269 דוברי ערבית). טעות הדגימה המרבית עבור שני הסקרים ברמת ביטחון של 95% היא ± 2.7% .

איסוף נתוני העקבות הדיגיטליים התבצע באמצעות שני כלים עיקריים: SimilarWeb ו- Buzzilla. פלטפורמת SimilarWeb מנטרת נתוני גלישה אנונימיים ממגוון מקורות ומעריכה באמצעות אלגוריתמים ייחודיים מדדי שימוש באתרי אינטרנט ובאפליקציות הכוללים: סך כל הביקורים, נתח התנועה (מחשב, טלפון נייד), דירוג בתוך המדינה ומחוצה לה, משך ביקור ממוצע, מספר עמודים ממוצע לביקור, נתח תנועה לפי מדינה ואזור, סך ביקורים לפי מגדר ולפי קבוצות גיל וכו'. Buzzilla היא פלטפורמה דיגיטלית המנטרת את מרחב המדיה החברתית, תגובות לכתבות, הודעות בפורומים, בלוגים וכו'. מאגר נתונים זה משמש לביצוע מחקרי מדיה חברתית על נושאי השיח, הזירות הפעילות ביותר, אפיון קהילות ומשתתפים, ניתוח סנטימנט חיובי ושלילי, ומדידת נפח פעילות לאורך זמן. הנתונים אשר הופקו משני כלי ניטור העקבות הדיגיטליים הללו עוסקים באוכלוסיית מחקר זהה (משתמשי אינטרנט בוגרים) ובתקופת זמן דומה להפקת נתוני הסקרים (ינואר-פברואר 2020).

המחקר עשה שימוש במגוון רחב של שיטות סטטיסטיות איכותניות וכמותיות, כולל סטטיסטיקה תיאורית והסקה סטטיסטית על מנת לתאר, להסביר ולחזות (באמצעות סימולציה נומרית) התנהגות משתמשים ברשת. ממצאי המחקר מצביעים על הבדלים ופערים דיגיטליים בהתנהגות מקוונת במגוון פעילויות ותחומי תוכן.

**קניות ברשת:**

סימולציות נומריות שבוצעו לבחינת הנטייה לקנות בתכיפות מצביעות כי המאפיינים ההתנהגותיים והסוציו-דמוגרפיים המשפיעים על הסיכוי לערוך קניות תכופות ברשת הם:

• דאגה לפרטיות ולהגנה על מידע אישי ברשת היא המנבא החזק ביותר לתכיפות הקניות המקוונות. אנשים בעלי חששות כבדים לפרטיותם המוטרדים מדליפת הנתונים האישיים שלהם, הם בעלי סיכוי נמוך ב -34% להיות קונים תכופים מאשר אנשים נטולי דאגה לפרטיות או לאבטחת נתונים.

• מיומנות דיגיטלית - משתמשים התופסים את עצמם כחסרי כישורים דיגיטליים הם בעלי סיכוי נמוך ב -17% לערוך קניות תכופות ברשת בהשוואה לאנשים התופסים את עצמם כבעלי כישורים אלה.

• התנהגות אימפולסיבית - משתמשים אשר דיווחו על נטייה חזקה להתנהגות אימפולסיבית היו בעלי סיכוי גבוה ב- 15% לערוך קניות תכופות ברשת לעומת משתמשים ששקלו בקפידה את הוצאותיהם.

• השתתפות פעילה ברשת (לדוגמא ע"י כתיבת תגובות והמלצות) – משתמשים פעילים ברשת הם בעלי סיכוי גבוה בשיעור של 12% לערוך קניות תכופות ברשת מאשר משתתפים שאינם פעילים.

• הכנסה - בעלי הכנסת משק בית גבוהה מהממוצע הם בעלי סיכוי גבוה ב -16% מבעלי הכנסת משק בית נמוכה לערוך קניות תכופות ברשת.

השכלה - בעלי תואר ראשון הם בעלי סיכוי גבוה ב -12% לערוך קניות תכופות ברשת בהשוואה לבוגרי תיכון.

מגדר - לגברים סיכוי גבוה ב- 11% מאשר נשים לערוך קניות תכופות ברשת.

גיל - לקבוצות צעירות יחסית (25-43) יש סיכוי גבוה ב-10% לערוך קניות תכופות ברשת בהשוואה לקבוצות גיל מבוגרות (+65);

גורמים נוספים המשפיעים על קניות ברשת:

• נמצא מתאם גבוה בין סוג המכשיר המשמש ברכישות מקוונות למחיר המוצר או השירות. עבור מוצרים ושירותים שמחירם נמוך מ- 100 ₪, ב -58% מהמקרים בוצעה הרכישה באמצעות הטלפון הנייד (נתח השימוש במחשב האישי לשם ביצוע הרכישה עמד על 42%). בעוד שנתון זה ירד ל -33% (נתח של 67% לשימוש במחשב האישי) כאשר מחיר המוצר או השירות עולה על 1000 ₪.

• גם נתוני הסקר וגם ניתוח נתוני העקבות הדיגיטליים העלו כי ימי קניות מיוחדים כמו "Black Friday" משפיעים באופן ניכר על נטיית המשתמשים לערוך קניות באינטרנט.

**נסיעות ותיירות מקוונות:**

מממצאי המחקר עולה כי:

• השימוש בפלטפורמות דיגיטליות להזמנות נסיעות על ידי אנשים חילוניים גבוה באופן ניכר משימוש בפלטפורמות אלה על ידי האוכלוסיות הדתיות והחרדיות אשר מבצעות נתח גבוה יחסית של הזמנות (~40%) באמצעות סוכני נסיעות.

• נמצא כי גיל המשתמש קשור קשר הדוק להעדפות ההזמנה. הזמנת טיסות ומלונות דרך האינטרנט נפוצה הרבה יותר בקרב קבוצות גיל צעירות יותר מאשר בקרב קבוצות גיל מבוגרות (76% בקרב קבוצת 35-44 לעומת 60% בקרב קבוצת גיל +65).

• ניתן להבחין בפער גדול בהעדפות ביצוע ההזמנות בהקשר של הרקע האתני. ישנו שימוש תכוף משמעותית בפלטפורמות מקוונות בקרב דוברי העברית (74%) לעומת דוברי ערבית (45%).

הגורמים המובילים שנמצאו קשורים להחלטה להזמין טיסות, מלונות או חבילות נסיעות **באינטרנט** היו:

• היכולת לערוך חיפוש מקיף ברשת (94% מהנשאלים מסכימים בהחלט או מסכימים עם הצהרה זו).

• היכולת להשוות עלויות (הסכמה של 88%).

• היכולת להתאים טיסה גמישה המותאמת לצרכי המטייל (הסכמה של 87%).

• היכולת לקבל מידע נוסף אודות הטיסה (הסכמה של 85%).

• עלות נמוכה יותר של מוצרי נסיעות מקוונים (80% הסכמה).

הגורמים המובילים שנמצאו קשורים להחלטת הפרט להזמין טיסות, מלונות או חבילות נסיעות באמצעות **סוכן נסיעות** היו:

• הצורך באינטראקציה אנושית עם אדם שיענה על שאלות ויפתור בעיות (86% מהנשאלים מסכימים בהחלט או מסכימים עם הצהרה זו).

• חששות בנוגע לפרטיות ואבטחת מידע (הסכמה של 41%).

• מיומנויות דיגיטליות נמוכות - הימנעות מטכנולוגיה וחשש לטעות בהזמנה באינטרנט (41% הסכמה).

כ- 54% מהנשאלים ציינו כי הדירוגים וחוות הדעת באתרי הזמנת נסיעות כמו booking.com, trivago, Airbnb ו- TripAdvisor משפיעים על החלטתם בנוגע להזמנת מקום לינה. מבחינה הזו, קבוצות גיל צעירות יותר (18-24; 25-34; 35-44) מושפעות במידה רבה יותר מדירוגי נסיעות בהשוואה לקבוצות גיל מבוגרות.

**ניהול כספי מקוון:**

ביחס לבנקאות אלקטרונית ועסקאות פיננסיות מקוונות, ממצאי המחקר מראים כי חלקן של פעילויות פיננסיות על ידי גברים גבוה יותר משיעורן בקרב נשים כמעט בכל קטגוריות הפעילויות:

- בדיקת יתרת החשבון (95% בקרב גברים לעומת 94% בקרב נשים),
- תשלום חשבונות (59% בקרב גברים לעומת 46% בקרב נשים),
- צפייה בפרטי קופות הגמל והפנסיה (39% בקרב גברים לעומת 31% בקרב נשים),
- קנייה ומכירה של מניות ואג"ח (19% בקרב גברים לעומת 9% בקרב נשים).

**בריאות ברשת:**

מניתוח נתוני הדיווח העצמי ונתוני המעקב הדיגיטליים עולה כי קיימים הבדלים מגדריים בהתנהגות החיפוש של מידע בריאותי ובשימוש בשירותי בריאות מקוונים, כאשר חלקן של הנשים גבוה יותר לעומת הגברים:

- קביעת תורים לרופא משפחה (88% נשים לעומת 87% גברים),
- צפייה בבדיקות מעבדה (80% נשים לעומת 73% גברים),
- הגשת בקשות מקוונות לבדיקות / בחינות (54% נשים לעומת 47% גברים),
- בקשות לחופשת מחלה (41% נשים לעומת 34% גברים).

מגמה מגדרית דומה עולה מניתוח נתוני עקבות דיגיטליים של התנועה באתרי קופות החולים (לדוגמא מכבי, כללית, מאוחדת), בהם חלקן של הנשים הוא 59% מהתנועה. מתברר כי בנוסף להבדלים בשימוש בשירותי בריאות מקוונים, גם ביחס לחיפוש מידע באינטרנט הקשור לבריאות (למשל מחלות ותסמינים; פענוח תוצאות בדיקות מעבדה ובדיקות, מידע על תרופות וטיפול תרופתי וכו') ניכר פער משמעותי על רגע מגדרי וכי נשים פעילות יותר מגברים בחיפושים בנושאי בריאות.

המחקר מצא פערים ניכרים בשימוש בשירותי בריאות מקוונים ובהתנהגות החיפוש של מידע הקשור לבריאות בין יהודים לערבים:

- כ -83% ממשתמשי הרשת היהודים הצהירו כי הם בוחנים את תוצאות בדיקות המעבדה, לעומת 54% בלבד מאוכלוסיית המשתמשים המקוונים הערבים. במקביל, 70% מהמשתמשים היהודים מחפשים באופן פעיל הסברים ופענוחים אפשריים של תוצאות המעבדה שלהם ברשת, לעומת 41% בלבד בקרב משתמשים ערבים.

**פרטיות והגנת מידע ברשת:**

המחקר מצא כי הצעדים השכיחים ביותר שמשתמשים נוקטים בהגנה או שמירה על פרטיותם הם:

• סירוב לאפשר את השימוש בנתונים האישיים שלהם למטרות פרסום (65% מהנשאלים משתמשים בהם לעתים קרובות או לעתים קרובות מאוד),
• שימוש בסיסמאות שאינן זהות בכניסה לאפליקציות ושירותי אינטרנט שונים (52%),
• הגבלה או סירוב להעניק גישה לנתוני מיקומם הגאוגרפי (GPS) (41%).

אמצעי הזהירות הננקטים בשכיחות הנמוכה ביותר בהקשר של הגנה על פרטיות או אבטחת נתונים ברשת הם:
• שימוש בתוכנה ייעודית לניהול סיסמאות (18% משתמשים בה לעיתים קרובות או לעיתים קרובות מאוד)
• שימוש בכלים מקוונים כגון VPN (10%) או דפדפן Tor (4%).

פרוצדורה של ניתוח גורמים שבוצעה לצמצום 13 משתנים הקשורים למאפייני הפרטיות ואבטחת מידע ברשת זיהתה שלושה פקטורים מובהקים, שתוייגו באופן הבא:
o **פרטיות כללית** - קריאת הצהרות פרטיות ומודעות לשימוש במידע אישי על ידי צד שלישי; הגבלת הגישה לנתונים אישיים.
o **פרטיות טכנית "רכה"** - נקיטת אמצעים פשוטים ושגרתיים לשמירה / אבטחת אנונימיות ופרטיות ברשת, לדוגמא מחיקת "עוגיות" והיסטוריית גלישה.
o **פרטיות טכנית "קשה"** - שימוש בכלים מורכבים וייעודים, טכנולוגיות ותוכנות במטרה להגן על פרטיות, דליפת מידע ואנונימיות, לדוגמא VPN, TOR.

ניתוח סטטיסטי של שלושת הגורמים או המדדים הללו מצא כי:
• מגדר נמצא במתאם חיובי מובהק עם כל שלושת מדדי הפרטיות. מתברר כי בקרב גברים מדדים אלו גבוהים יותר לעומת נשים. מגמה דומה נצפתה גם מניתוח נתוני העקבות הדיגיטליים שהעידו על מדדי פרטיות טכנית קשה גבוהים יותר בקרב גברים.
• מדד הפרטיות הכללית גבוה יותר בקרב קבוצות גיל המבוגרות, בעוד שבקבוצות הגיל הצעירות ממוצע מדד הפרטיות הטכנית הקשה גבוה יותר מהקבוצות המבוגרות. מגמה דומה נצפתה מניתוח נתוני העקבות הדיגיטליים שהראו אותות גבוהים יותר למיומנויות טכניות קשות בקרב משתמשים מקוונים צעירים יותר.
• רמת ההשכלה נמצאה במתאם חיובי הן עם פרטיות כללית והן עם כישורים טכניים רכים.
• השימוש ברשתות החברתיות נמצא במתאם חיובי מובהק הן עם מדד הפרטיות הכללית והן עם מדד הפרטיות הטכנית הקשה.
• שתי תכונות התנהגותיות הקשורות לתפיסה עצמית בנושא סדר וארגון נמצאו במתאם חיובי מובהק עם מדד הפרטיות הכללית.

ניתוח המדיה החברתית סביב פרטיות ברשת העלה כי:
• שיח הפרטיות מתמקד בשלוש תת-קטגוריות עיקריות: היעדר מודעות לפרטיות ברשת בקרב בני נוער, מציצנות ואי- כיבוד פרטיות, וכן שימוש בנתונים אישיים על ידי תאגידים. השיח היה לרוב שלילי במהותו וכלל ביטויים של חשש מפני פגיעה בפרטיות ושיפוט מוסרי של האשמים בהפרתו.

• השיח סביב היבטים טכניים קשים של פרטיות מקוונת (דיונים שקשורים למונחים "גלישה בסתר" ו"דפדפן Tor ") היה בולט ביותר בקרב פורומים של בני נוער ופורומים של קהילות דתיות וחרדיות. ניכר כי מטרות השיח היו לספק למשתמשים כלים כדי להגן על הנתונים שלהם וכן כדי לקבל "מבצעים" טובים יותר לטיסות ולקניות ברשת.

• ניתוח התוכן של המדיה החברתית מראה כי בעוד שהשיח סביב המונחים "פרטיות מקוונת" מתמקד בחששות חברתיים ובשיפוט מוסרי, השיח סביב המונחים "היסטוריית גלישה", "דפדפן Tor" ו- "גלישה בסתר" (רכיב הפרטיות הטכנית הקשה) הוא בעל אופי טכני / אינסטרומנטלי.

במסגרת המחקר פותח כלי גנרי אינטראקטיבי לוויזואליזציה של נתוני הסקר. פיתוח הכלי סייע בתהליך ניתוח נתוני השאלונים ותרם לגיבוש הצעה לקווים מנחים לעיצוב ויזואלי של נתוני שאלונים לצורך הפקת תובנות ו"סיפורי נתונים". ההמחשה החזותית של הנתונים עשויה לסייע בעיבוד תובנות אינטגרטיבי ובשילוב נתוני הסקר עם סוגי נתונים נוספים (למשל עקבות דיגיטליים).

ממצאיו ותוצאותיו של מחקר זה יכולים לספק למשרדי הממשלה בישראל, לגורמים עסקיים ולקהילה המחקרית מספר תובנות אשר עשויות לתרום לגיבוש מדיניות ציבורית בתחום הפער הדיגיטלי ופרטיות ברשת, לשפר ממשקי משתמשים בתחום הצריכה באינטרנט, כמו גם לקחים מתודולוגיים ופרוצדוראליים אשר יכולים לשמש למחקר מתקדם בשילוב סקרים עם נתוני עקבות דיגיטליים.

**ההמלצות הן כדלקמן:**

**אנו ממליצים לבעלי עניין ומקבלי החלטות מהסקטור הציבורי לפעול ליצירת פרוטוקול אשר יסדיר ויגדיר את השימוש בנתוני עקבות דיגיטליים.** על הפרוטוקול להגדיר הנחיות ברורות עבור: איסוף, ניטור וכריית נתונים ממקורות מקוונים; אנונימיזציה של מידע אישי מטעם בעל הנתונים; נהלים לעיבוד נתונים, איחוד וקישור של נתוני עקבות דיגיטליים ממקורות מרובים; הנחיות לגבי הצגת הנתונים (מטעם החוקר); בנייה ותחזוקה של מאגרי עקבות דיגיטליים (עם או באמצעות גופים כגון הספרייה הלאומית או ארכיון המדינה); שימוש של צד שלישי; הקנסות שיוטלו על החוקר במקרה של הפרת תנאי החוזה.

**אנו ממליצים למשרדי הממשלה ולעמותות להנגשת מידע ציבורי :**

- להעלות מודעות לגבי ההשלכות של התנהגות אימפולסיבית בהקשר של ביצוע קניות מיותרות ברשת.
- להעלות מודעות, בעיקר בקרב נשים, לחשיבות של רכישת ידע בתחום הבנקאות האלקטרונית וביצוע עסקאות פיננסיות מקוונות.
- להעלות מודעות, בעיקר בקרב גברים והאוכלוסייה הערבית באשר ליתרונות בביצוע פעולות מקוונות בתחום הבריאות.
- להעלות מודעות, בעיקר בקרב בני נוער, לנושא הפרטיות ברשת, לדוגמא קריאת הצהרות פרטיות ומודעות לשימוש במידע אישי על ידי צד שלישי, והגבלת הגישה לנתונים אישיים. בנוסף,

העלאת המודעות, בעיקר בקרב נשים, לחשיבות וליתרונות הקיימים בשימוש בכלים מורכבים וייעודים, טכנולוגיות ותוכנות במטרה להגן על פרטיות, דליפת מידע ואנונימיות.

- לבצע ניטור וניתוח שוטפים של השיח בנושא פרטיות ברשת בתקשורת המרכזית ובפורומים השונים. זאת על מנת להעמיק את הבנת הצרכים של הקהלים הפעילים בשיח זה כמו גם על מנת לזהות את הקהלים שאינם פעילים בשיח.

**אנו ממליצים לבעלי עסקים :**

- להעלות מודעות, בעיקר בקרב האוכלוסיות הדתיות והחרדיות, מבוגרים ודוברי ערבית באשר ליתרונות של שימוש בשרותי נסיעות ותיירות ברשת.
- לשפר את ידידותיות אתרי האינטרנט והאפליקציות בתחום ממשקי פעולות הרכישה בכל סוגי המכשירים (טלפונים ניידים, טאבלטים ומחשבים שולחניים למיניהם).

**המלצותינו לקהילת החוקרים:**

- קידום ופיתוח של מתודולוגיות לטריאנגולציה של נתונים וכלים שיתרמו לשיפור מהימנות הנתונים ולהבנת התופעה הנחקרת (לדוגמא פער דיגיטלי).
- פיתוח ושיפור מתודולוגיות מחקר לאיחוד סקרים אינטרנטיים עם נתוני עקבות דיגיטליים (שיפור ייצוגיות הדגימה, התוכן וכיו"ב), על מנת להעמיק את ההבנה של התנהגות מקוונת סמויה וגלויה של משתמשים. אחד האמצעים עשוי להיות פיתוח של רכיבים ויזואליים כחלק אינטגרלי ומובנה בפלטפורמות הסקרים.

# Executive Summary

The research is a part of the Israel-Slovenia bilateral project “Digital transformation of quantitative data collection in social science research: Integrating survey data collection with big data and paradata for identifying social behavior". The research was funded by the Ministry of Science and Technology and the Slovenian Research Agency and took place from October 1, 2018 to September 30, 2020, with the participation of the Centre for Social Informatics (CSI) at the University of Ljubljana, Slovenia and the Samuel Neaman Institute for National Policy Research (SNI). The project shared both mutual and separate tasks relating to the methodological and practical aspects of data collection from online surveys and the augmentation and triangulation of various types of data. The CSI research group has mainly focused on the methodological aspects of online survey design, developments in the field of paradata collection and in the formulation of composite paradata indices intended for the study of survey data quality, whereas the SNI research group centered on profiling online user behavior via triangulated data, using obtrusive and unobtrusive methods.

This manuscript reports the findings of the Israeli study, which aims at investigating the socio-economic and personal trait characteristics of online behavior, pertaining to various activities such as e-shopping, e-travel, e-finance, the use of social networks, search activity and the perception of privacy and personal data security. This examination is carried out by a triangulated approach which fuses together evidence from survey data, digital trace data and social media data.

In order to tackle the research objectives at hand, two comprehensive questionnaires aimed at investigating and profiling behavioral aspects of online Internet users were formulated. The first survey (“***Bi-national online behavior survey***”) included both Israeli and Slovenian cohorts and focused on particular aspects of online user behavior – the perception of privacy and information security online and the behavioral characteristics of online shopping. The second survey (***“National online behavior survey”***) included only Israeli respondents and centered on wider aspects of online behavior: e-health, e-travel and tourism, trust in technology, e-finance, search behavior and the use of communication and information technologies. The two surveys were based on a “representative sample” of Israeli and Slovenian population, aged 18+. The data was collected using Internet panels via the 1KA digital survey platform between 23/1/2020 and 16/2/2020. For the

Israeli population, the surveys were distributed in two versions: Hebrew and Arabic using two separate, designated panels. A quota/stratified sampling approach was used to ensure sufficient representation of sub-populations. The Binational Survey sample included 1283 Israeli respondents (1083 Hebrew speakers and 246 Arabic speakers) and 4058 Slovenian respondents, and the National Survey included 1270 Israeli respondents (1001 Hebrew speakers and 269 Arabic speakers). The maximal sampling error at the 95% confidence level for both the Binational (Israeli cohort) and National Surveys samples is ±2.7%.

The digital trace data for the research was collected and analyzed via two main online tools (SimilarWeb and Buzzilla). SimilarWeb collects anonymous clickstream data from a diverse panel of users and employs algorithms to estimate overall metrics for web and apps. Available metrics include: total visits, traffic share (desktop, mobile), global and country rank, average visit duration, pages per visit, traffic share by country and region, visits by gender and by age groups etc. Buzzilla is a digital platform for monitoring and tracking social media and information from forums, groups and message boards. This data pool is used for conducting social media research on themes such as conversation topics. Both of these digital trace sources relate to the same research population (adult on-line Internet users) and represents the same time period (the year 2019) as the self-report data (surveys).

The research employed a wide range of qualitative and quantitative research methods including descriptive statistics and inferential statistics in order to describe, explain and predict (via numerical simulation) online user behavior.

Numerical simulations that were held with respect to the effect of socio-demographic and behavioral factors on the propensity of being a **frequent shopper** show that:

- The strongest predictor of online shopping behaviour was the individual's concern for privacy and data security online. Individuals who have very strong concerns for their privacy and fear for the leak of their personal data were 34% less likely to be frequent online shoppers than individuals who have no privacy or data security concerns.
- Online users who lack digital skills are 17% less likely to be frequent shoppers as compared to individuals who possess these skills.
- Online users who reported strong tendency towards impulsive behavior were 15% more likely to be frequent shoppers than users who carefully weighted their expenses.

- Individuals who exercise active participation online are 12% more likely to be frequent shoppers than non-active participants.
- Online Internet users with well above average household income are 16% more likely to be frequent shoppers than individuals with well below average household income; Education also exerts a large effect, with individuals holding a Bachelor level degree or equivalent are 12% more likely to conduct frequent shopping online than high school graduates. Male online users are 11% more likely to be frequent online shoppers than female users and younger age groups (25-43) are 10% more probable to shop on a frequent basis than older age groups (65+).

Additionally in this regard:

- Strong correlation was found between the type of device used in online purchases and the price of the good or service. For products or services costing less than 100 NIS, smartphone was the device of choice in 58% of the cases (PC share was 42%). This figure drops to 33% (67% PC share) when the price of the good or service is greater than 1000 NIS.
- Both survey data and digital trace data analysis revealed that special shopping days such “Black Friday” exert a strong influence on the propensity of users to conduct shopping online.

With respect to online **travel behaviour** the research findings show that:

- The use of digital platforms for travel bookings by secular individuals is much higher than the use of these platforms by the religious and ultra-orthodox populations which are characterized by relatively high share (~40%) of bookings made by travel agents.
- Age was found to be to be closely related to booking preferences. Online travel bookings are much more prevalent among younger age groups than among older age groups (76% in the 35-44 group as compared to 60% among in the 65+ age group).
- A large gap in booking preferences can be observed with respect to ethnic background, showing much more frequent use of online platforms among Jewish online users (74%), as compared to Arab online users (45%).

The leading factors that were found to be associated with the individual’s decision to book flights, hotels or travel packages online were:

- The ability to conduct a comprehensive search (94% of the respondents definitely agree or agree with this statement).
- The ability to compare costs (88% agreement).
- The ability to tailor a flexible flight that suits the traveler's needs (87% agreement).
- The ability to receive more information about the flight (85% agreement).
- Lower cost of online travel products (80% agreement).

The leading factors that were found to be associated with the individual's decision to book flights, hotels or travel packages via a travel agent were:

- The need to interact with a person who will answer questions and solve problems (86% of the respondents definitely agree or agree with this statement).
- Online privacy and data security concerns (41% agreement).
- Low digital skills - avoiding technology and the fear of making mistakes when booking online (41% agreement).

Nearly 54% of the respondents indicated that ratings and opinions on travel bookings websites such as booking.com, trivago, Airbnb and TripAdvisor affect their decision to either book or not book a particular accommodation. In this respect, younger age groups (18-24; 25-34; 35-44) were found to be influenced to a greater degree from travel ratings than older age cohorts (65+; 55-64).

With respect to **e-banking and online financial transactions**, the research findings show that the share of carrying financial activities by male users is higher than its comparable share among female users in almost all transaction categories:

- Checking account balance (94% women, 95% male)
- Payment of bills (59% among men and 46% among women)
- Viewing details of provident funds and pensions (39% among men and 31% among women).
- Buying and selling stocks and bonds (19% among men and 9% among women).

The examination of self-report data and digital trace data has revealed gender-based differences in the **search behaviour of health information and in the use of online health services**:

- Making appointments to a family doctor (88% women, 87% male)
- Viewing laboratory tests (80% women, 73% male)

- Making online requests for tests/examinations (54% women, 47% male)
- Sick leave requests (41% women, 34% male)

Similar trend with respect to gender can be observed from digital trace data where women account for 59% of the traffic in the various sick-fund (Kupot-Holim) websites (e.g. Maccabi, Clalit, Meuhedet).

In addition to differences in the use of online health services, substantial gaps can be also observed between female online users and male online users with respect to the search of health related information, with female users exercising higher search activity.

Stark gaps between Jewish and Arab online users were observed in the use of online health services and in the search behavior of health-related information.

- About 83% of Jewish online users stated that they review the results of laboratory tests, as compared to only 54% of the Arab online users' population. Concurrently, 70% of the Jewish online users actively search for possible explanations and deciphering of their laboratory results online, as compared to only 41% among Arab online users.

With respect to **privacy and data security behavior** of online users, the research found that the most frequent measures that online users exercise in protecting or maintaining their privacy are:

- Refusing to allow the use of their personal data for advertising purposes (65% of the respondents exercise it often or very often)
- Using nonidentical passwords to login to various apps and web services (52%)
- Restricting or refusing access to their geographical (GPS) location (41%).

The least frequent precaution in the protection of privacy or data security online are:

- Using designated software for password management browser (18% use it often or very often)
- Using online tools such as VPN (10%) and the Tor Browser (4%).

Factor analysis procedure has identified three factors or underlying variables describing online privacy and data security, which were labeled as follows:

  - **General Privacy** - reading privacy statements and being aware of the use of personal information by third parties; restricting access to personal data.

- **Soft Technical Privacy -** carrying out simple, routine measures to maintain/secure user anonymity and privacy online, e.g. deleting cookies and browsing history.
- **Hard Technical Privacy -** using complex and designated tools, technologies and software in order to protect privacy, data security and anonymity online, e.g. VPN, TOR.

Further analysis of these three aggregated indices has found that:

- Gender is positively and significantly correlated with all three privacy indices, showing higher perception of privacy and data security among the male population. Similar trend was observed from the analysis of digital trace data which showed higher signals for hard technical skills among male users.
- General privacy skills are high among older age cohorts, whereas younger age cohorts display high rates of hard technical skills. Similar trend was observed from the analysis of digital trace data which showed higher signals for hard technical skills among younger online users.
- Education level is positively correlated both with general privacy and with soft technical skills.
- The use of social networks is positively and significantly correlated with the general and hard technical indices.
- Two “Big Five” behavioral attributes pertaining to self-perception of order are positively and significantly correlated with general privacy attributes.

The analysis of public social media surrounding online privacy revealed that:

- The privacy discourse focuses on three main sub-categories: Teenagers’ (lack of) awareness to online privacy, Voyeurism and disrespect for privacy and corporations’ use of personal data. The discourse was mostly negative in its nature and included expressions of concerns about privacy and moral judgement of those who are blamed for breaching it.
- The discourse around hard technical aspects of online privacy (discussions which were related to the terms “Incognito browsing” and “Tor Browser”) was most prominent among teenagers’ forums and religious Jewish communities forums, and its purpose was to provide users with tools to protect their data and receive better “deals” for flights and shopping.
- The content analysis of public social media shows that while the discourse surrounding the terms “online privacy” focuses on societal concerns and moral

judgement, the discourse surrounding the terms “browsing history”, “Tor Browser” and “Incognito browsing” (“hard privacy”) is of technical/instrumental nature.

An interactive generic tool for the **visualization and analysis of survey data** was developed in the framework of the research. This tool has highlighted the importance of following sequential steps and guidelines in facilitating the understanding of data stories which could be compared to or used in conjunction with other types of data (e.g. digital traces).

**Our recommendations to government and public policy makers are:**

• Raise awareness about the consequences of **impulsive and addictive shopping behavior**.

• Raise awareness and enhance education, especially among women, of the importance of acquiring knowledge in the field of **e-banking and online financial transactions**.

• Raise awareness, especially among men and the Arab population regarding the **benefits and importance of online health services.**

• Raise awareness, especially among teenagers, regarding the issue of **online privacy**. In addition, raise awareness, especially among women as to the importance and advantages of using designated tools, technologies and software in order to protect privacy, data security and anonymity online.

**Our recommendations to the business sector are:**

• Raise awareness, especially among the religious and ultra-Orthodox populations, adults and Arab speakers regarding the benefits of using **online travel and tourism services**.

• Improve the friendliness of websites and applications especially in purchasing transactions interfaces on all types of devices (mobile phones, tablets and desktops of all types).

**Our recommendations to the research community are:**

• Promote and develop data triangulation methodologies and tools for the purpose of enhancing data reliability and understanding online behavior.

• Develop and improve existing methodologies for consolidating online surveys with digital traces for the purpose of deepening understanding of hidden and visible online behavior of users. This could be achieved through the development of **visual components** as an integral and built-in part of survey platforms.

# Introduction

In recent years, empirical methods in the Social Sciences have been witnessing radical change due to the emergence of novel digital technologies. The transition from traditional surveys to web based surveys on the one hand and the introduction of web harvesting tools of "digital traces" on the other hand, created a new research environment based on multi-source and large scale data. Within this broader framework, one of the key challenges relates to the interaction between the collection, utilization and harmonization of passive data collection (e.g. digital traces), which is non-invasive and non-intrusive in its nature, with active data collection (e.g. surveys), where subjects are involved participants. A few studies have shown that passive data may actually replace active data, while many others accentuate complementary aspects of integrating these two types of data.

This research sets out to profile and investigate the socio-economic and personal trait characteristics of online behavior, pertaining to various activities such as e-shopping, e-travel, e-finance, the use of social networks, search activity and the perception of privacy and personal data security. This examination is carried out by a triangulated approach which fuses together evidence from survey data, digital trace data and social media data. The research focuses on the following theoretical, methodological and practical aspects of this approach: (1) laying the methodological foundations for augmenting and triangulating different digital data sources; (2) establishing specific sets of survey questions to complement digital trace data; (3) creating standardized sets of composite indices for investigating online behavior using the two types of data; (4) designing practical guidelines on using the new types of datasets for policy decision-makers and (5) expanding visualization techniques for evaluating online user behavior based on rich datasets.

The report is organized as follows: **Chapter 1** provides the literature overview for this work. **Chapter 2** reviews the methodological framework of the research, including the research goals, the research questions and the research population. It also provides a description of research data and discusses the motivation and novelty of the research. **Chapter 3** reports the main research findings pertaining to online user behavior in four main content usage themes: online shopping, e-travel, e-finance and e-health. A specific attention is given to both socio-demographic factors and personal or behavioral attributes as well as

to consumer related factors in explaining and predicting online user behavior. **Chapter 4** attempts at deepening our understanding of online user behavior by triangulating survey data, digital trace data and social media data. The triangulation methodology is demonstrated by focusing on **online privacy as a case study**. In the framework of **Chapter 5,** a generic interactive visualization tool for survey data in the context of online user behavior is developed and demonstrated. **Chapter 6** concludes and summarizes the research findings and provides recommendations for policy makers.

# Chapter 1: Literature Review

Over the past two decades, vast and rapid changes have been witnessed in the use and diffusion of information technologies. The introduction and growing use of the Internet has exerted a substantial impact on everyday life, changing the way humans interact, consume information and conduct their daily activities. During this time span many activities that once required physical interaction such as shopping, banking and finances, local and state government services and access to medical services have met suitable digital alternatives. However, the behavioral and the social attributes and determinants of online usage vastly differs across users and are characterized by gaps in access, skills and the type of on-line content consumed.

## Socio-demographic and behavioral attributes of online usage

Socio-demographic differences in online behavior is one of the most studied themes in the empirical literature relating to the study of information and communication technology (ICT). These socio-demographic gaps with respect to ICT usage were coined in the early 1990's by the term "digital divide" (Vehovar et al., 2006; Cruz-Jesus, 2012;). The phrase generally refers to the gap between individuals, households, businesses and geographic areas at different socio-economic levels, with regards to their access to information and communication technologies and to their use of the Internet for a wide variety of activities (OECD, 2001). The mitigation of digital gaps is seen by many countries as a moral and social interest (raising personal welfare, alleviation of social gaps, promotion of equal opportunities among various population groups), as an economic interest (as means for achieving a competitive advantage) and as a political interest (a strategy for promoting and safeguarding national resilience) [Rafaeli et al., 2013].

The literature shows that the type of content people use differs by gender. Studies reveal that women, on the one hand, prefer religious content, health related information, online games and are more likely to use the Internet's communication tools. On the other hand, adult males are more likely to use the Internet for information, entertainment, commerce (Jackson et al., 2001; Subrahmanyam et al., 2001; Peter and Valkenburg, 2007; Park, Kim and Na, 2007; Zillien and Hargittai, 2009), online gaming (Schumacher and Morahan-Martin, 2001) and dating (Rudder, 2014). Age also appears to be one of the most significant variables that influence Internet use (Bonfadelli, 2002; Fox and Madden, 2005; Zillien and Hargittai, 2009). Studies show that young adults extensively use

communication tools, such as instant messaging (IM) and chatting, and are more likely to pursue entertainment and leisure activities, such as gaming, downloading files or music (Howard et al., 2001; Dutton et al., 2011; Fox and Madden, 2005; Jones and Fox, 2009).

Socio-economic status indicators were found to have a significant impact on Internet use (e.g. Zillien and Hargittai, 2009). DiMaggio et al. (2004) found that that persons of higher socio-economic status employ the Internet more productively and to greater economic gain than their less privileged, but nonetheless connected, peers. There is evidence to suggest that people with lower levels of socio-economic status tend to use the Internet in more general and superficial ways (Van Dijk, 2005).

A few studies suggest that education is the most important predictor in explaining the types of online activities a person will pursue (Robinson et al., 2003; Van Dijk, 2005). People with higher levels of education use the Internet for health information, financial transactions and research, while people with lower levels of education use the Internet for casual browsing, playing games or gambling online (Howard et al., 2001). Hargittai and Hinnant (2008) found that those with higher levels of education use the Internet for 'capital-enhancing' activities, which include seeking political or government information, exploring career opportunities and consulting information about financial and health services. Helsper and Galacz (2009) show that the lower educated are least likely to use the Internet for educational and economic purposes, even when they have similar levels of Internet access and skills (Van Deursen and Van Dijk, 2014).

The past two decades has seen exponential growth in online use in a vast number content usage categories and digital services. A few notable ones are online shopping, online finance, online health and online travel.

**Online shopping** is a form of electronic commerce which allows consumers to directly buy goods or services from a vendor over the Internet using a web browser. Consumers find a product of interest by visiting the website of the retailer directly or by searching among alternative vendors using a shopping search engine, which displays the same product's availability and pricing at different e-retailers (Lim et al., 2016). Studying the factors influencing online shopping behavior is interesting as it can shade light on its triggers and barriers. Online shopping can be explained by behavioral theories such as the theory of reasoned action (TRA) proposed by Fishbein and Ajzen (1977), the theory of planned behavior (TPB) proposed by Ajzen (1991) or Triandis' (1979) model. Socio-

demographic differences might explain differences in online shopping behavior, as well as personality traits (Tsao and Chang, 2010), for example personal innovativeness (Limayem et all., 2000). Other aspects that might influence online shopping behavior include price and product selection, payment, product delivery, website design, customers review, privacy concerns and the device in use. For example, purchasing on a mobile device might be challenging as consumers are required to search for extensive information from multiple intermediaries, compare prices, and book properly (Law and Leung, 2000).

**The travel and tourism** industry thrives on information. A traveler needs to manage huge amount of data such as scores of messages, itineraries, schedules, payment information, destination and product information (Benckendorff et al., 2019). Information technology (IT) has dramatically transformed the travel and tourism industry (Sheldon, 1997; Werthner and Klein, 1999) and it continues to evolve and impact the way travelers gain access to information (Xiang et al, 2015). Various tools such as search engines have become a dominant force that influence travelers' access to tourism products (Xiang et al., 2008). Developments in mobile computing, particularly with the adoption of smartphones and their apps for travel, creates new venues and opportunities for information search and use whereby the contextually defined needs of on-the-go travelers become increasingly prominent in guiding travel decisions (Wang et al., 2012). There are differences in online travel behavior that are rooted in consumer characteristics as well as in perceived channel characteristics (Amaro and Duarte, 2013). For example, travelers might experience sense of risk while using travel technology (Park and Tussyadiah, 2017). Furthermore, there are evidence that demographic characteristics are involved in differences in travelers perceived risks (e.g. air-ticket purchases, Kim et al., 2019).

**Electronic finance**, especially online banking, has significantly reshaped the financial landscape and transformed the activities of people and corporations (Claessens et al., 2002; Dandapani, 2017). Information technology enabled electronic channels to perform many banking functions that would traditionally be carried out over the counter (Giannakoudi, 1999). The evolution of electronic banking, such as Internet banking from e-commerce, has altered the nature of personal-customer banking relationships and has many advantages over traditional banking delivery channels. This includes an increased customer base, cost savings, mass customization and product innovation, marketing and communications, development of non-core businesses and the offering of services regardless of geographic area and time (Giannakoudi, 1999; Gan and Clemes, 2006).

Polatoglu and Ekin (2001) identified instant feedback, quick transactions and easy access, as important attributes in electronic banking. Furthermore, Liao and Cheung (2002) and Gerrard and Cunningham (2003) found that the transaction speed and the fast access to electronic banking accounts were important attributes for consumers that used electronic banking (Gan and Clemes, 2006). Consumers who were more financially innovative had a higher probability of adopting electronic banking than less financially innovative consumers (Gerrard and Cunningham; 2003). In terms of socio-demographic factors, education qualification was found to have significance for the choice of digital payment (Singh and Dutta, 2019). Gan and Clemes (2006) note that both financial risk (financial loss that is caused in the use of electronic banking as result of making a mistake) and physical risk (breach of privacy and accessing personal information by a third party) may deter the adoption and the use of online banking.

**eHealth** is defined as the "ability to seek, find, understand and appraise health information from electronic sources and apply knowledge gained to addressing or solving health problem (Norman and Skinner, 2006). Access to health information and knowledge resources is highlighted to be crucial for health care and public health and is an extremely important motivator for ICT use[1]. A recent comparative study in 28 European countries regarding the persistence of digital divides in the use of health information found significant differences in the use of the Internet for health information with regards to gender, age, education, long-term illness and health-related knowledge (Alvarez-Galvez et al., 2020). Regarding the gender gap, they found that females search health-related information on the web more frequently than males. They point out a possible explanation for the gendered difference is that women are more caregiving-oriented (e.g. for children).

While ICT enhances our lives in many ways, it also raises new concerns with regards to online **privacy and data security** which also bears **profound impact on user behavior**. When online users communicate and interact, they leave digital footprints behind them, generating information about their lives and daily activities. This information is accessed, stored, manipulated, data mined, shared, bought and sold, analyzed, stolen or misused by government, corporate, public and private entities, often without the user's awareness or consent. Online privacy, as being highly complex in nature, is often defined through

[1] https://www.un.org/en/chronicle/article/bridging-digital-divide-health

various dimensions – informational, accessibility and expressive. Informational privacy relates to an individual's right to determine how, when, and to what extent information about the self will be released to another person or organization (Burgoon et al., 1989). Accessibility privacy relates to "attempted acquisition of information that involves gaining access to an individual" (DeCew, 1997). This dimension includes physical access (e.g. spam mail, computer virus and personal contact details). Expressive privacy "protects a realm for expressing one's self-identity or personhood through speech or activity, shielded from interference, pressure and coercion from government or from other individuals" (DeCew, 1997).

Finn et al. (2013) mentions seven types of privacy: Privacy of the person, privacy of behavior and action, privacy of communication, privacy of data and image, privacy of thoughts and feelings, privacy of location space and finally privacy of association. Central to these dimensions is the aim to keep personal information out of the hands of others. Studies show that the level of privacy concerns and perceptions of privacy vary from person to person and are related to culture, experience in online use, lifestyle, gender and age (Christofides et al., 2012).

Buchanan et al. (2007) developed and validated a set of three scales which were found to be a robust and reliable measure of privacy concerns and behavior suitable for administration via the Internet. Two scales address different aspects of things people do (i.e. reflect behavior) to protect their privacy: exercising general caution, and technical protection. The third scale, privacy concern, is attitudinal rather than behavioral, and reflects general concerns about privacy on the Internet.

## Novel practices in joint data collection for the analysis of online behavior

In the past decades, empirical methods in the Social Sciences have vastly changed due to the development of IT technologies. The process has started in 1970's with the introduction of computers into survey data collection. Further acceleration of the process was experienced in 1990's with the rise of the Internet and with the introduction and advancement of various interactivity features. These developments created an entirely new environment for social science research (Vehovar and Lozar Manfreda, 2008). In addition to changes in data collection, the intensive progress in computer science and informatics, including artificial intelligence, has also revealed important new potentials. These advancements have created a paradigm shift in the way we collect and use social

science data. Digital technologies have revolutionized the entire research cycle - from conceptualization, analyses, collaboration, and research management to practices of publishing and dissemination. This new digital environment is sometimes labelled as e-Social Science (Vehovar, Petrovčič and Slavec, 2015).

Within this broader framework, one of key challenge relates to the interaction between the collection, utilization and harmonization of nonreactive (passive) data collection (e.g. digital traces), which is non-invasive and non-intrusive for a subject, with reactive (active) data collection, where subjects are active participants (e.g. surveys). A few studies have shown that passive data may actually replace active data (Vehovar and Slavec, 2016), while many others accentuate complementary aspects of integrating these two types of data with other auxiliary data (e.g. administrative datasets, socio-economic datasets, geographic data).

### Survey data

For more than a century, surveys have been used as the main method for obtaining data in social sciences. In recent years, web-based surveys are rapidly replacing traditional survey methods of data collection (telephone surveys, face to face interviews, mail surveys), as they are cost and time-efficient, easy to set-up and implement, and flexible for inclusion of advanced interface features and multimedia elements (Evans and Mathur 2005; Callegaro, Lozar-Manfreda and Vehovar, 2015). Web surveys also have some notable drawbacks. They have a much higher potential for non-coverage and nonresponse bias (Callegaro et al., 2015; Cho et al., 2013; Lozar Manfreda et al., 2008; Shih and Fan, 2008; Yarger et al., 2013; Bosnjak et al., 2005). In addition, they suffer from many of the same “illnesses” as traditional data collection methods, such as reliability and validity (e.g. self-report bias, honesty of response), introspective ability (the ability to provide an accurate response to the question), the degree of understating and interpreting the question, and difficulty in providing “accurate” measure in rating questions (Graham et al., 1993; Donaldson et al., 2002; Hoskin, 2012).

### Big data

The technological revolution witnessed in the past two decades, characterized by exponential computation growth and advancement in software, hardware, cloud and information technologies has produced enormous opportunities, as well as challenges in the production and utilization of complex data. This can be especially observed in the context known as “Big Data”. The definition of “Big Data” is complex and constantly

changing. However, there is some consensus in the literature regarding its main characteristics, relating to three dimensions: **volume**: vast data that cannot be handled by traditional analytical tools; **velocity of production**: the recording of real-time events; and **Variety**: complex datasets including numerous sources of digital traces or footprints, such as unstructured text, images, videos and logs (Beyer and Laney 2012).

The third dimension of Big Data (variety) particularly relates to “digital traces” or “digital footprints” which are defined as “records of activity undertaken through online information systems”. They are marks left as a sign of passage, a recorded evidence that something has occurred in the past” (Howison et al., 2011). Jones and Rafaeli (2000) used archaeology as an analogous field for describing the role of digital artefacts on society and human behavior: “Like archaeological tells, the remains of digital traces can supply evidence on human behavior and interaction”. O'Brien (2010) has upgraded this idea by describing the information age as an “archaeology site of modern existence waiting for excavation”.

Big data can be either human generated or machine produced data. The latter is information produced by mechanical or digital devices without the active intervention of a human (e.g. process logs, traffic bandwidth, location data such GPS system output, Internet clickstream data and sensor readings). These human and machine generated methods provide useful basis for ‘data-mining’, digital trace studies and cultural analytics to better understand the huge amount of data that exists and the evolution of social behavior and communication on digital platforms (O'Brien, 2010).

Callegaro and Yang (2018) have created a typology of the main sources and subclasses of digital traces and “Big data” as follows: ***Internet data*** *(Online text and multimedia),* ***Website data*** *(logs, cookies, transactions, and website analytics),* ***The Internet of Things data*** (traces from any device using the Internet as communication transmission protocol), ***Behavioral data*** (a specific subset of the IOT based devices such as smartphones and wearables, recording locations, movements etc.), ***Transaction data*** (records of orders, shipments, payments, returns, billing, and credit card activities), ***Administrative data*** (national health records, taxes, benefits, pensions etc.), ***Commercial data*** (tracks from companies, businesses, consumers, users), *and* ***Social media data.***

The latter subclass of digital trace data, **Social data** refers to data that is generated on online spaces which enable shared public interpersonal communications (Jones, Ravid

and Rafaeli, 2004). These spaces include social media platforms (e.g. Twitter, Facebook), blogs, forums, new websites, web and mobile applications and other online social spaces. Social data is “an umbrella concept for all kind of digital traces produced by or about users, with an emphasis on content explicitly written with the intent of communicating or interacting with others” (Olteanu et al., 2016). The availability of social data, combined with powerful computational resources provides researchers with unprecedented access to public discourse and to social interactions (Gandomi and Haider, 2015; Hampton, 2017; Ruths and Pfeffer, 2014).

In the past few years there is a growing trend of using various methods of harnessing and harvesting social and digital data for social and psychological research, as well as for other research disciplines (Chan et al., 2017; Gandomi and Haider, 2015; Stieglitz et al., 2018). Scholars have also been applying traditional methods of analysis such as ethnography (Belk and Kozinetz, 2016) and conversation analysis (Giles et al., 2014), adapting them to the digital spaces. The growing use of social data analysis stems, in part, from the advantages it offers, in comparison to traditional social research methods such as surveys, focus groups and interviews. While surveys, interviews and focus groups depend on participants’ memory retrieval, which tends to worsen over time, social data is generated organically by users who wish to share their thoughts and experiences at their own will, a process which does not rely on memory retrieval (Schober et al., 2016). Similarly, while surveys, interviews and focus groups are structured and based on questions created by researchers in advance, according to their pre-defined research goals (Schober et al., 2016), social data analysis is unstructured and based on an indirect observation of people’s natural online conversations (Gandomi & Haider, 2015) and therefore has a potential for surfacing meaningful discoveries, that were not part of the researcher’s hypothesis or questions. One of the biggest strengths of unobtrusive research is the documentation of actual rather than self-reported behavior. Other advantages include repeatable results, easier access to data, continuity and the fact that permission from subjects is not necessary (Kellehear,1993; Webb 2000).

Alongside its advantages, social data has a few limitations as well. First, there is the self-selection bias, which stems from the fact that users decide whether or not to participate on social media platform, what to comment about and in what frequency (Olteanu et al., 2016; Schober et al., 2016). The majority of users are actually “lurkers” – people who passively consume web content in a read-only mode (Bronstein et al., 2016; Rafaeli, et

al., 2004). This is related to another shortcoming, which is the lack of generalization ability. Social data analysis does not provide researchers access to users' demographics nor can it be assumed to match population's characteristics. In addition, different platforms attract different types of populations (Olteanu et al., 2016; Schober et al., 2016).

## The state of the art: Integrating survey data with digital trace data

Survey data enhance our ability as social scientists to understand the research questions at hand in greater depth and in a specifically designed manner. This is due to the fact that surveys collect attitudes and opinion data which cannot be readily covered by Big Data. However, studying contemporary human behavior with survey methods has several drawbacks. The most important one is the limited reliability of self-reported behavioral measures. On the other hand, big data approaches also have limitations. Importantly, most studies relying exclusively on digital trace data lack relevant information on individuals' attributes (e.g., sociodemographic characteristics or personality traits) or attitudes and motivations behind the actions (Stier et al., 2019, Mishkin, 2014). Moreover, the data are most often based on biased samples, making difficulties to link online behavior to microlevel theories from the social sciences (Jungherr, 2018). Therefore, many big data studies remain descriptive as the nature of their data offers very limited opportunities for theory-driven analyses. Hence, these data alone cannot answer questions about individual-level determinants of human behavior. To sum up, Big Data can accentuate behaviors and tell us the "what" while surveys can reflect on attitudes and opinions and tell us the "why."

A commonly shared view among researchers is that combining data and methods from surveys and Big Data can and should be used together to maximize the value of each other (Japec et al. 2015). Integrating traditional research methods with Big Data analytics provides an exceptional opportunity to understand what human subjects are doing, why they are doing it and what can be done to change their behavior (Mishkin, 2014).

While both Big Data and survey research have a lot to offer, relatively little work has been conducted up to date to see how these two types of data can be used together to provide richer datasets (Callegaro and Yang Y, 2018). Some notable examples for the use of joint data are described in three studies conducted by Google research (Müller and Sedley 2014) on Happiness Tracking Surveys (HaTS) and by Mastrandrea et al. (2015) and Hitachi Ltd. on measuring happiness and social interaction using wearable technology (Yano et al. 2015). Happiness Tracking Surveys (HaTS) were developed by Google for

collecting large-scale in-product measurement of user attitudes and experiences. HaTS has been deployed successfully across dozens of Google's products to measure progress towards product goals and to inform product decisions (Müller and Sedley, 2014). Mastrandrea et al. (2015) compared diaries and surveys to wearable sensors and online social media to study social interactions among students in a high school in France. In another application of wearable sensors, Hitachi collected more than a million days' worth of data on employees' activities over the span of nine years (Yano et al., 2015). The authors were able to correlate the sensor data with happiness measured via questionnaires (in Callegaro and Yang Y, 2018). Some other examples of combining surveys with Big Data can be found also in various specific case studies, such as Martin (2016), Wells and Thorson (2017) and Buntain et al. (2016).

Stier et al. (2019) in their recent overview of integrating survey data and digital trace data highlight three key issues regarding the collection and analysis of such hybrid data sources: "(1) data linking including informed consent for individual-level studies, (2) methodological and ethical issues impeding the scientific (re)analysis of linked survey and digital trace data sets, and (3) developing conceptual and theoretical frameworks tailored toward the multidimensionality of such data".

### Visualizing online survey data

Data visualizations are highly important for raising stakeholders' interest and for strengthening the understanding and trust in the data (Cherchye et al. 2007). Design choices of visualization can influence the interpretation of various metrics and are therefore critical. Sharing data and key insights among researchers or between researchers and non-academic audiences are often requisitioned. However, the rate of data sharing is relatively low in the social sciences (Jones et al., 2016). Visualization tools use in general, and with survey data in particular, make data and insights sharing more approachable (Wexler,2016).

Visualization is not a trivial issue (Nardo et al. 2005). Its complexity is derived from the data characteristics (digital trace data or survey data), as well from its goals and tasks. While in our previous research the focus was on visualization of digital trace data in the context of digital divide (Rafaeli et al., 2018), in this research we focus on visualization of survey data. As Wexler (2016) says: "All too often, the best stories in the survey data remain hidden behind canned reports that are too difficult…". In chapter 5 we discuss the

issue of visualization of survey data and introduce a visualization tool that we developed to illustrate the data that was collected in the surveys.

The literature review has clearly demonstrated the advantages and potentials in fusing survey data with Big Data to produce richer datasets which significantly enhance our abilities as social scientists to understand, explain and analyze human behavior and vastly improve the methodological aspects related to research validity. However, as the literature review reveals, the research in this domain is still in its infancy and substantial knowledge gaps remain.This research thus seeks to fill in these gaps and provide theoretical and empirical underpinnings for the integrated use of survey data and digital trace data.

# Chapter 2: Methodology

This research sets out to profile and investigate the socio-economic and personal trait characteristics of online behavior, pertaining to various activities such as e-shopping, e-travel, e-finance, the use of social networks, search activity and the perception of privacy and personal data security. This examination is carried out by a triangulated approach which fuses together evidence from survey data, digital trace data and social media data.

The research employs a wide range of qualitative (e.g. social discourse analysis) and quantitative research methods and tools including descriptive statistics (e.g. graphs, two-dimensional tables) and inferential statistics (t-test, ANOVA, Non-parametric methods, Post-hoc tests, OLS and binary regression models, factor analysis (used in the composition of normalized index for online privacy), in order to describe, characterize, explain and predict (via simulation games) online user behavior.

## Research goals

The main goals and objectives of the research are as follows:

- To study and analyze online user behavior and specific personal traits with respect to socio-demographic attributes and the content usage consumed.
- To construct more robust measurements and indices for on-line behavior.
- To consolidate and triangulate digital trace data with web survey data to better understand and predict online user behavior, digital divide, literacy and skills.
- To deepen our research on the topic of survey data visualization, with focus on abstraction of online behavior.
- To achieve improved understanding of “online privacy” perceptions in social discourse contexts.
- To identify, classify and map the discourse surrounding the concept of “online privacy” utilizing social media analytics tools.

## Research questions

- Which socio-demographic factors best explain on-line user behavior? Are there any significant differences between the various socio-demographic groups? What kind of patterns and digital gaps can be observed?
- What type of behavioral traits best explain on-line user behavior? Could significant differences or patterns can be observed?

- How could the triangulation of self-report data and digital trace data can be used to deepen and broaden our understanding of online user behavior in general and online privacy in particular?

## Research population and data

The research population is composed of Israeli and Slovenian on-line Internet users. In the framework of the research, four different type of data sources were used to profile and analyze online users, with the specific aim of reflecting on both stated and actual (revealed) online-user behavior. A triangulated approach, fusing self-report data and digital trace data, is demonstrated on a specific case study aimed at analyzing and explaining online privacy behavior at the macro and micro levels.

The main methodological tool used in the framework of this research to investigate online user behavior is based on self-report methods in the form of online web surveys. In addition, three data sources are based on digital trace data - either at the aggregated level (SimilarWeb online, Google Trends) or at the disaggregated user-level (Buzzila). The following paragraphs present a short description of each tool or data source.

### Self-report data: online web surveys

In order to tackle the research questions at hand, two comprehensive questionnaires aimed at investigating and profiling behavioral aspects of online Internet users were formulated. The first survey (dubbed as "***Bi-national online behavior survey***") included both Israeli and Slovenian cohorts and focused on particular aspects of online user behavior - the perception of privacy and information security online and the behavioral characteristics of online shopping. The second survey (labeled as ***"National online behavior survey"***) included only Israeli respondents and centered on wider aspects of online behavior. In addition to online privacy and information security themes, the National Survey covered the following themes: e-health, e-travel and tourism, trust in technology, e-finance, search behavior and the use of communication and information technologies. Both surveys included identical socio-demographic and "Big Five" (personality traits taxonomy) questions. The survey response scales that were used in both surveys are the five category Likert type scale (either agreement level or rating) and the dichotomous scale (e.g. "yes" and "no" type questions).

The two surveys are based on a "representative sample" of Israeli and Slovenian population, aged 18+. The data was collected using Internet panels (Israeli and Slovenian

professional panelists) via an online digital survey platform (1KA) between 23/1/2020 and 16/2/2020. The respondents were asked questions about their online behavior during the past year (the year 2019). For the Israeli cohort, iPanel Ltd. provided the panel service (distributing the survey links to panelists by specified pre-defined socio-demographic quotas) and the system interface between its own system and the Client system (1KA online digital survey platform). 1KA is an open source application that enables services for online surveys. The application was developed by the Centre for Social Informatics, at the Faculty of Social Sciences, University of Ljubljana. It can be used unlimitedly and free of charge for the purposes of online surveys, under certain terms of use. 1KA basic guideline is to minimize the number of clicks - hence the designation EnKlikAnketa (1KA) (which is translated as 'One click survey'). All operations are therefore carried out with a minimum number of clicks or pressures on the keyboard (keystrokes). 1KA application can be installed on any server and can be linked to other programs via the API. The online service supports the following functionalities: Development, design and technical creation of an online questionnaire; The implementation of online survey: support for invitations, publication and distribution of data; and compiling and analyzing data and paradata[2].

For the Israeli population, the surveys were distributed in two versions: Hebrew and Arabic using two separate, designated panels (four surveys ran simultaneously). All questions in the Arabic and Hebrew versions were identical. A quota/stratified sampling used was used to ensure sufficient representation of sub-populations (e.g. Arab and ultra-orthodox population) which are important for making statistical inference on the differences in user behavior and digital gaps - within and between groups. Different quotas were used for both versions of the Israeli sample: gender, age group and religiousness level for the Hebrew language version and gender, age and religion for the Arabic language version. The Slovenian sample (Binational Survey) included only two quotas - age group and gender. The Binational Survey sample included 1283 Israeli respondents (1083 Hebrew speakers and 246 Arabic speakers) and 4058 Slovenian respondents and the National Survey included 1270 Israeli respondents (1001 Hebrew speakers and 269 Arabic speakers). The maximal sampling error at the 95% confidence level for both the Binational (Israeli cohort) and National Surveys samples is ±2.7%. Cronbach's alpha was used to assess the reliability of the two surveys (multiple Likert-type scales questions). The results

[22] https://www.1ka.si/d/en/about/general-description

of the procedure found the two questionnaires to be reliable (see Annex 1), showing high internal consistency between the items. Table 1 below summarizes the differences and commonalities between the two surveys.

**Table 1: The Binational and National Surveys**

| | Bi-national | National |
|---|---|---|
| Population | Representative sample of the Israeli and Slovenian adult (18+) population | Representative sample of the Israeli adult (18+) population |
| Sample size | Israeli sample: n=1283<br>Slovenian sample: n=4058 | Israeli sample: n=1270 |
| Survey method | Online web survey based on professional panelists. Quota/stratified sampling used. | Online web survey based on professional panelists. Quota/stratified sampling used. |
| Data collection date | 23/1/2020 to 16/2/2020 | 23/1/2020 to 16/2/2020 |
| Survey quotas | Israeli sample: ethnic background (Jewish/Arab), gender, age group.<br>Slovenian sample: gender, age group. | Israeli sample: ethnic background (Jewish/Arab), gender, age group. |
| Survey language | Hebrew, Arabic, Slovenian | Hebrew and Arabic |
| Scaling approach | Likert scale – five measurement categories | Likert scale – five measurement categories |
| Content usage/variables covered (with respect to online behavior) | Privacy and information security, e-shopping, trust, big-five, socio-demographic variables. | Privacy and information security, e-health, e-travel and tourism, trust in technology, e-finance, use of communication and information technologies, search behavior, big five themes, socio-demographic variables. |

### Digital trace data

The digital trace data for the research was collected and analyzed via three online tools (SimilarWeb, Buzzilla and Google Trends). The digital trace data relates to the same research population (adult on-line Internet users) and represents the same time period (the year 2019) as the self-report data (surveys). The following paragraphs present a short description of the various digital trace data sources:

**SimilarWeb On-line platform**: A digital platform based on data extracted from four main sources: 1. A panel of web surfers made of millions of anonymous users equipped with a portfolio of apps, browser plugins, desktop extensions and software. 2. Global and Local Internet Service Providers. 3. Web traffic directly measured from a learning set of selected websites and apps intended for specialized estimation algorithms. 4. A colony of web crawlers that scan the entire Web and apps stores. SimilarWeb collects anonymous clickstream data from a diverse panel of users and employs algorithms to estimate overall

metrics for web and apps. Available metrics include: total visits, traffic share (desktop, mobile), global and country rank, average visit duration, pages per visit, bounce rate, traffic share by country and region, visits by gender and by age groups etc. The platform, including various web tools, covers the last 24-month period of the on-line activity (SimilarWeb, 2016).

**Buzzilla:** A digital platform for monitoring and tracking social media and information from forums, groups and message boards, collecting millions of responses (talkbacks) to articles, forum posts, and blogs in various fields. This data pool is used for conducting social media research on themes such as conversation topics. The platform allows to perform segmentation of communities and participants and to measure the volume of activity.

**Google Trends:** An online search tool that allows the user to see how often specific keywords, subjects and phrases have been queried over a specific period of time. This tool works by analyzing a portion of Google searches to compute how many searches have been done for the terms entered, relative to the total number of searches conducted on Google over the same time. The service provides information on the search query volumes of its users since January 2004 and allows researchers to select searches by geographical region (provinces, states, countries), categories and sub-categories (e.g., travel, finance, food), and frequency (daily, weekly, monthly). Results are displayed in a graph that Google calls "Search Volume Index". The data in the graph can be exported to a csv file and edited in Excel or other spreadsheet applications (Siliverstovs and Wochner, 2018).

## Research motivation, novelty, and expected contribution of the research

The literature review shows consensus among researchers that augmenting and triangulating various data sources, such as survey data and digital trace data, can lead to enhanced understanding of human behaviour (e.g. Callegaro and Yang, 2018, Japec et al., 2015). Most digital trace data relevant to social science research are ‘organic’ data, collected for some other primary purpose or generated automatically as a by-product of the main data collection. Survey data, by contrast, are designed for a specific research purpose. Within this context, a tailored set of survey question items were designed and tested to complement the key types of Big Data collection. This included a set of basic ‘webographic’ questions for evaluating the results of Big Data analysis which provided the

basis for creating the augmented datasets required for effective substantive analysis of various online behaviours.

Fusing these rich types of datasets allow us to better understand the large amounts of data (as opposed to merely describing the data), particularly these relating to the evolution of social behaviour on digital platforms. Within this context, the project has achieved some novel methodological, theoretical and practical contributions. To the best of our knowledge, no studies so far have offered a comprehensive methodological framework for augmenting, consolidating and integrating web survey data and digital trace data to describe and analyse online user behaviour. Thus, laying the conceptual foundations and outlining standard techniques for fusing these types of data constitutes a clear and significant methodological contribution to the digitalisation process in social science research.

The project also brings several important practical novelties of relevance to policy research and decision-makers, particular the ability to capture data ‘on demand’, integrate them properly with survey data, as well as to present the relevant CIs at a detailed level. This all provides decision-makers and stakeholders with new and high-resolution data.

# Chapter 3: Analyzing Online User Behaviour via Digital Trace Data Analysis and Self-report Examination

In this chapter, we use self-report data derived from online web surveys and digital trace data obtained from online platforms to shed light on content usage and behavioral attributes of online users.

The SimilarWeb platform was used as the main instrument for digital trace analysis. This unique platform includes several tools for the analysis of “big data” - website analysis, category analysis and keywords search analysis. The website analysis tool is aimed at analyzing website traffic. The analysis of categories can be performed either by “ready-to-use” taxonomy (e.g. “Shopping” category) or by specially tailored user customization (aggregation of several websites to a single category - e.g. Amazon and eBay and AliExpress).

The website visits frequency measure was one of the key metrics used in the analysis of digital trace data. We explored the transformation and change in this metric over time (12 months period, in concordance with the online surveys’ timeline) and parsed it with the socio-demographic profile of its audience (gender and age), as well as with other attributes such as the type of device used (PC, mobile). The analysis of keywords facilitated our understanding of how traffic flows across websites or usage content categories. The general keyword analysis tool of SimilarWeb was used to determine which websites receive the most traffic share from a specific keyword. Comparisons were conducted by the Google Trends tool which was used to zoom on specific search terms. This tool uses a normalized index to represent the popularity of searches over time and space.

Our analysis centers on four main content usage themes: online shopping, e-travel, e-finance and e-health. We apply and demonstrate a triangulation methodology which fuses together digital trace data and survey data in these specific content usage categories.

## Online shopping

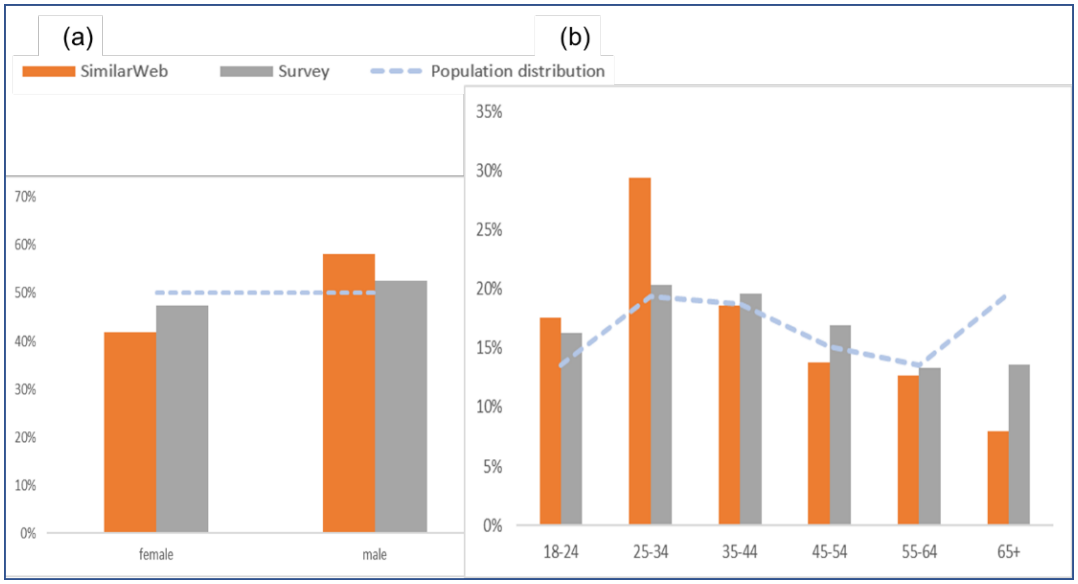
Over the past decade, online shopping has grown in exponential rate and significantly changed consumer behavior worldwide. Figure 1 presents the leading online shopping websites visited in Israel in 2019, as reported by stated behavior data (online surveys) and digital trace data (SimilarWeb’s category analysis for “E-commerce and Shopping”). As can be clearly seen from the figure, both data sources show that AliExpress, Amazon and eBay were the top three most visited international shopping websites for Israeli online

users in 2019. The signals for both sources also indicate that AliExpress is the leading website. Figure 2 presents online shopping distribution[3] parsed by socio-demographic attributes. As can be observed from the data, both data sources indicate higher visit rates in online shopping websites by male users and younger age cohorts (especially the 25-34 age group).

**Figure 1: Leading online shopping websites visited in Israel 2019: comparison between self-report data (a) and digital trace data (b)**


Source: Binational Survey data and SimilarWeb website analysis report.

**Figure 2: Online shopping distribution parsed by gender (a) and age (b)**

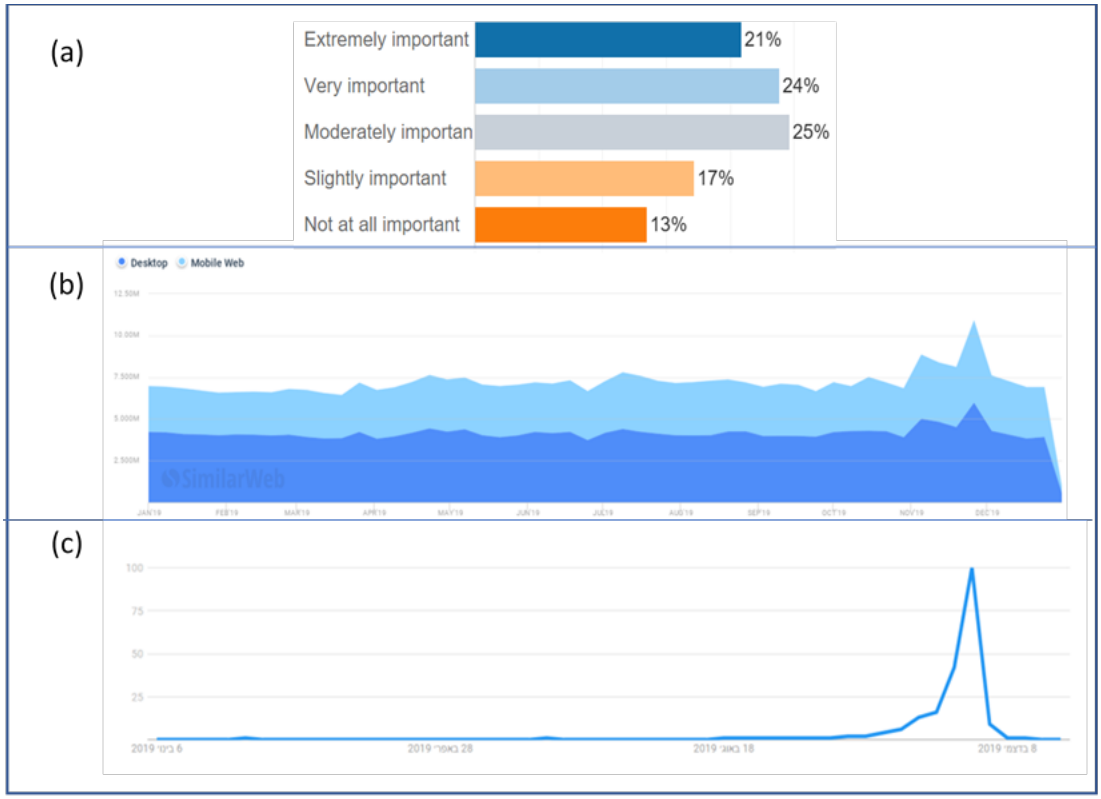



Source: Special processing of Binational Survey data and SimilarWeb category analysis data

[3] AliExpress.com visits (in the survey – reported as visits frequency in the range of less than once a month to several times a day)

As for the timing of the online shopping, it seems that special shopping days such “Black Friday” exert a strong influence on the propensity of users to conduct shopping online. About 45% of Israeli online shoppers indicated that special shopping events constitute an important or extremely important factor in their decision to shop online. It is evident from the analysis of digital trace data (Figure 3b – SimilarWeb and Figure 3c Google Trends) that the frequency of visits in online shopping websites significantly rises during the “Black Friday” shopping event (end of November).

**Figure 3: Impact of “Black Friday” shopping event on online shopping frequency**



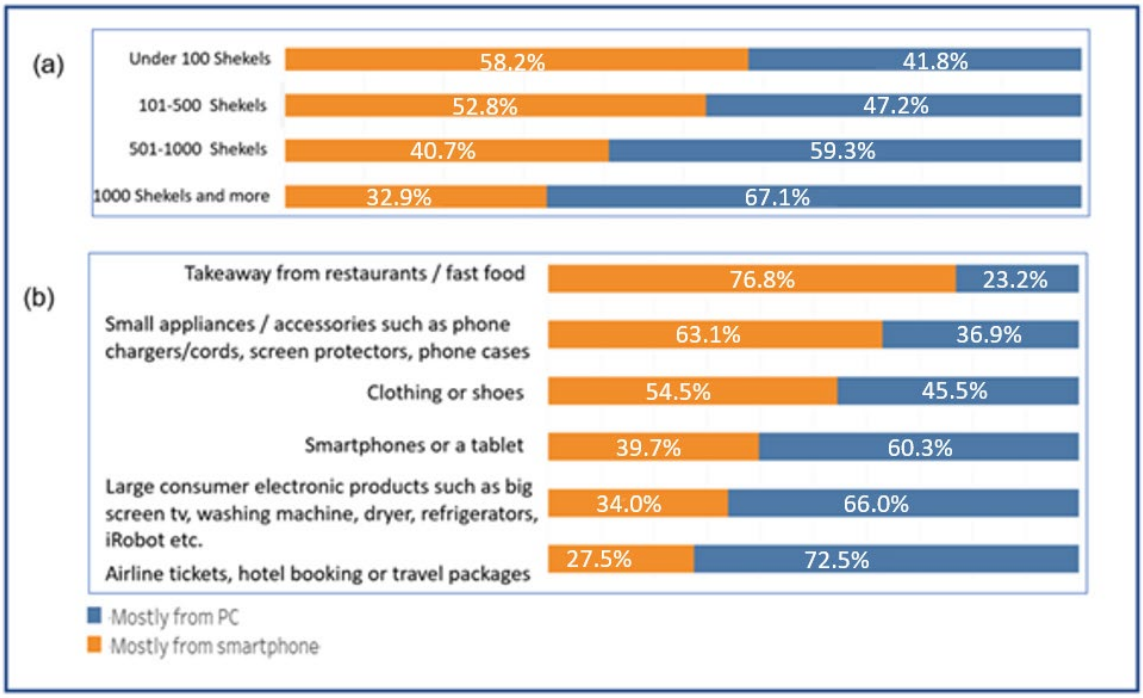

(a)-survey data (Importance of special shopping days); (b)-SimilarWeb data (category analysis-sites visits); (c)-Google Trends data (interest in “Black Friday” search term)

Source: Binational Survey data; SimilarWeb data (category analysis); Google Trends report

Another important factor that impacts the user’s decision to shop online is the cost of the ordered good or service which is also directly linked to device selection. As can be seen from the data, smartphone share use significantly diminishes as the cost of the ordered good or service rises. For products or services costing less than 100 NIS, about 58% of online users stated that they used smartphones as their means of order. This figure drops to about 33% smartphone share use when the price of the good or service is more than

1000 NIS (Figure 4a). Similarly, there is a much higher propensity to use PCs over smartphones when making either high risk, rare or expensive orders (e.g. Airline tickets, hotels, travel packages, large consumer electronic products) than daily or frequent (e.g. ordering food from restaurants, ordering cheap small electrical appliances) online transactions (Figure 4b). A Spearman's rank-order correlation shows a statistically significant correlation between product or service cost and pc use ($r$ = .19, $p$ < .01, n=3570).

**Figure 4: The relationship between product/service characteristics and device selection (smartphone/PC), shown by price intervals (a) and by category type (b)**




Source: Binational Survey data

## Binary regression and simulation model for explaining and predicting shopping behavior

In order to test the effect of various socio-demographic characteristics as well as behavioral attributes on the propensity to shop online, a binary logistic model for a “frequent online shopper” was fitted based on the National Survey data (for Israeli online users only).

### Model formulation

The binary logistic regression model is used when the dependent variable is dichotomous (e.g. "occurrence or non-occurrence"). The independent variables may be nominal,

dichotomous, ordinal or interval. The model predicts the probability of event Z (being "a frequent online shopper") to occur (1) by matching the data (2) to a logistic curve:

1. $P(Z) = \frac{1}{1+e^{-Z}}$

where z is a linear combination of the coefficients:

2. $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$

It is important to note that in our case the dependent variable is not a "true" dichotomous variable. It is rather a "threshold" variable based on aggregation of Likert scale. Online users who indicated that they either shop on a weekly or a monthly basis (frequently or very frequently) were assigned the value "1" and those who indicated that they shop occasionally, rarely or very rarely (less than once a month) were assigned the value "0". Users who indicated that they do not shop online were excluded from the model as they did not answer the questions pertaining to online shopping.

In our model, Z is a combination of socio-demographic characteristics (gender, age, education, household income), personal or behavioral attributes of the online user (impulsive behavior, active behavior, passive behavior, concern for privacy, digital literacy) and consumer related factors such as the cost of the good or service (cost) and the need to tangibly "feel" it prior to making a purchase.

The key assumptions of the model are:

- P(Z=1) of the dependent variable represents the desired (occurring) outcome.
- Error terms are independent.
- All explanatory variables are independent from each other (no multicollinearity).
- Linearity of independent variables and log odds.
- To satisfy maximum likelihood estimation, sample size is "large enough" (larger than 30 observations per each independent variable estimated in the analysis).

### Estimation results

Table 2 presents the estimation results for the binary logistic online shopping model. The dependent variable is a dichotomous dummy variable for a user who shops frequently online (within the last week or month=1, else=0). As can be seen from the table, the parameter estimation for gender (male dummy) is positive, suggesting that men shop online more frequently than women. The coefficient for age is negative, indicating that young individuals shop more frequently online than older individuals. The coefficients for

education and household income are also positive and significant, suggesting that individuals with higher education levels and higher income levels are more prone to be frequent shoppers. As for the impact of personal or behavioral attributes of the online user on shopping behavior, it seems that impulsive (making unnecessary purchases frequently), active (submitting reviews for products frequently) and passive/lurking (reading reviews for products frequently without participation) behaviors are significantly and positively correlated with frequent online shopping. Other behavioral attributes such as the "lack of digital skills" and "having privacy concerns with regards to the leak of personal data when browsing" were found to be negatively associated with frequent shopping. Finally, consumer related factors such the cost of the product and the "need to physically feel or test the product" were also found to significantly impact shopping behavior. Low product price was found to be positively associated with frequent online shopping, whereas individuals who indicated that they need to tangibly feel the product they buy were less likely to be frequent shoppers.

**Table 2: Online shopping model estimation**

| | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| Gender (male dummy variable) | 0.439 | 0.149 | 8.728 | 1 | 0.003 | 1.552 |
| Age | -0.105 | 0.044 | 5.549 | 1 | 0.018 | 0.901 |
| Education level | 0.158 | 0.059 | 7.075 | 1 | 0.008 | 1.171 |
| Household income | 0.166 | 0.063 | 6.905 | 1 | 0.009 | 1.181 |
| Impulsive behavior (I make unnecessary purchases) | 0.152 | 0.065 | 5.492 | 1 | 0.019 | 1.165 |
| Active behavior (submitting reviews for products) | 0.124 | 0.068 | 3.283 | 1 | 0.070 | 1.132 |
| Low product cost | 0.194 | 0.071 | 7.460 | 1 | 0.006 | 1.214 |
| Passive behavior (Reading reviews for products) | 0.219 | 0.063 | 11.943 | 1 | 0.001 | 1.245 |
| Need to tangibly test the product | -0.228 | 0.080 | 8.107 | 1 | 0.004 | 0.796 |
| Lack of digital skills | -0.172 | 0.077 | 4.998 | 1 | 0.025 | 0.842 |
| Privacy concerns regarding leak of personal data | -0.354 | 0.072 | 24.020 | 1 | 0.000 | 0.702 |
| Constant | -1.251 | 0.775 | 2.602 | 1 | 0.107 | 0.286 |

Dependent variable: frequent shopper

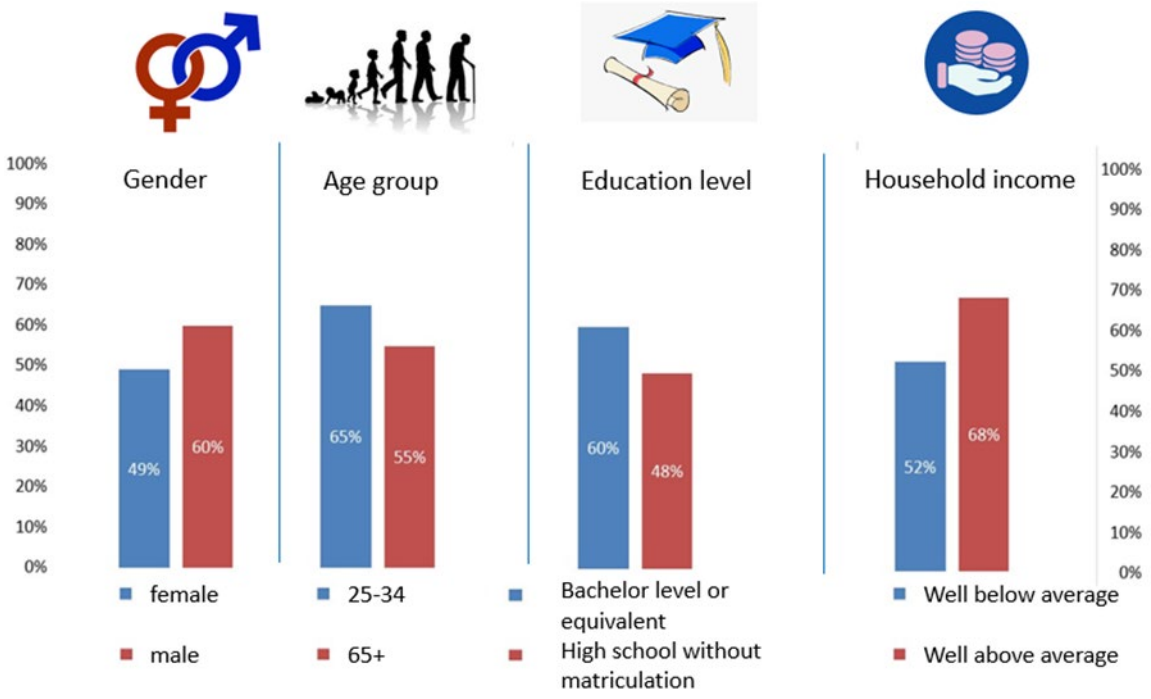
### Simulation forecasts

Based on the model estimation results (betas) and given specific user profiles/attributes, numerical simulation scenarios were played out for predicting the probability of a "particular online user" to be a "frequent online shopper". Scenarios were played out incrementally, changing the value of one variable at a time, with the values of all other variables held constant.

Figure 5 demonstrates the impact of socio-demographic attributes on the probability of being a "frequent shopper" in two opposing scenarios. As can be seen from figure, male

online users are 11 percentage points more likely to be frequent online shoppers than female users (60% vs 49%) and younger age groups (25-43) are 10% more probable to shop on a frequent basis than older age cohorts (65+ group). Education also has quite a large effect, with individuals holding a Bachelor level degree or equivalent are 12% more likely to conduct frequent shopping online than high school graduates without a matriculation diploma. The largest socio-demographic gap in the probability of being a "frequent online shopper" is observed with respect to the household income. Online Internet users with well above average household income are 16% more likely to be frequent shoppers than individuals with well below average household income.

**Figure 5: Simulation results – the impact of socio-demographic attributes on the probability of being a "frequent online shopper"**

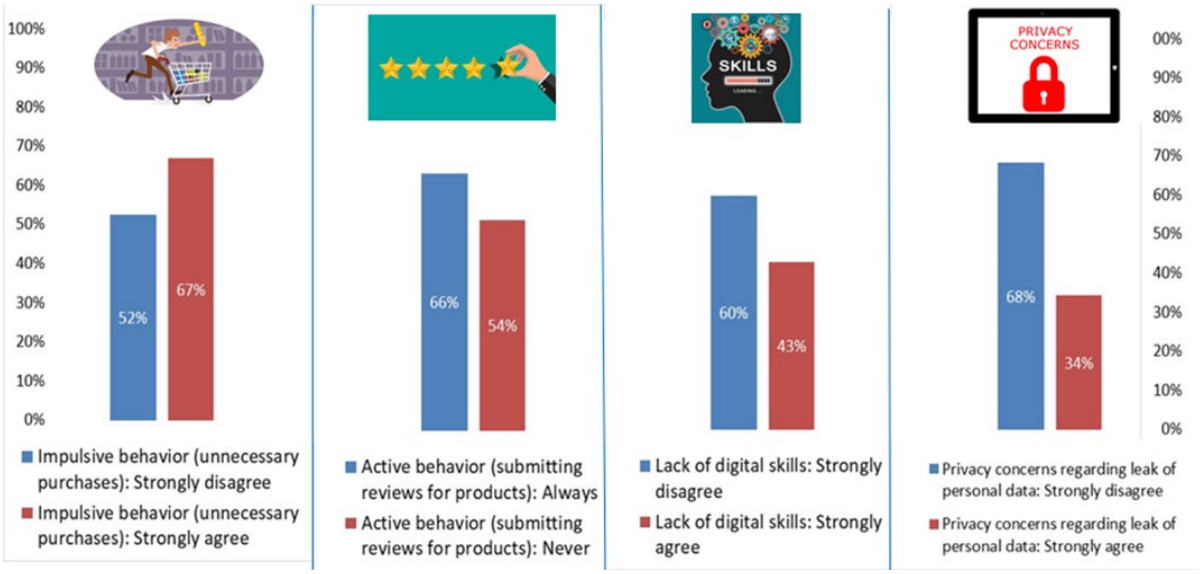



Source: Special data processing of the Binational Survey data

Figure 6 demonstrates the impact of the behavioral attributes on the probability of being a "frequent shopper". As can be seen from the illustration below online users who reported strong tendency towards impulsive behavior (making unnecessary purchases often) were 15% more likely to be frequent shoppers than users who carefully weighted their expenses. Individuals who exercise active participation online (e.g. regularly submit reviews for products) are 12% more likely to be frequent shoppers than non-active

participants. The lack of digital skills was found to strongly impact shopping behaviour, as individuals who lack these skills are 17% less likely to be frequent shoppers as compared to individuals who possess these skills. By far, the strongest predictor of online shopping behaviour was the individual's concern for privacy and data security online. Individuals who have very strong concerns for their privacy and fear for the leak of their personal data were 34% less likely to be frequent online shoppers than individuals who have no privacy or data security concerns.

**Figure 6: Simulation results – the impact of behavioral attributes on the probability of being a "frequent online shopper"**




Source: Special data processing of the Binational Survey data

## Online travel

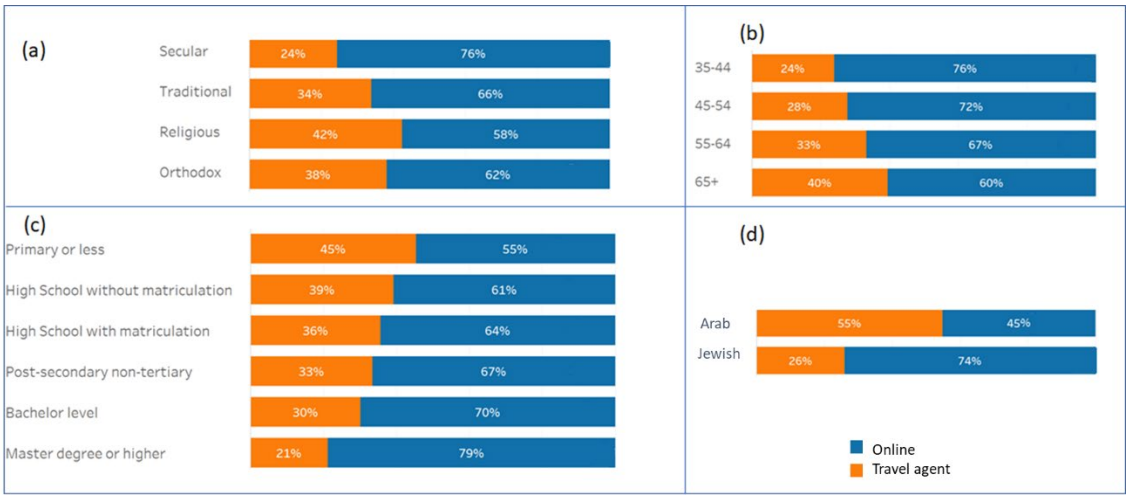
Digital trace data obtained from the SimilarWeb platform and data extracted from the National Survey were used to study and analyze Israeli online user behavior relating to travel. In our analysis, we focus on four main themes: the impact of socio-demographic attributes on booking preferences (online booking versus booking by a travel agent), the relationship between online travel search behavior and actual booking, the main motivations and preferences by individuals for making online and face-to-face (travel agent) bookings and the impact of online user rating on actual booking.

### Booking preferences parsed by socio-demographic attributes

Figure 7 presents the booking preferences of online users parsed by socio-demographic attributes - religiosity level, age, education, and ethnicity (Jewish and Arab). The

respondents were asked “how do you usually make the actual booking of the airline ticket or vacation package that you purchased? Two selection options were given: “Most often through travel agents” (face to face or over the phone) and “Mostly through online purchase”. As can be seen from the illustration below, the use of digital platforms for travel bookings by secular and traditional (Mesortim) populations (Figure 7a) is much higher than the use of these platforms by the religious and ultra-orthodox populations which are characterized by relatively high share (~40%) of bookings made by travel agents. Both age and education (Figure 7b and Figure 7c) seem to be closely related to booking preferences. As can be seen from the data, online travel bookings are much higher among younger age groups than older age groups (e.g. 76% in the 35-44 group as compared to 60% among in the 65+ age group). Online bookings are also much more frequent among individuals holding higher education degrees (70% for individuals holding a Bachelor’s degree, 79% for individuals holding a Master’s degree as compared to 61% among high school graduates and about 55% for individuals with a primary education). A large gap in booking preferences can be observed with respect to ethnic background (Figure 7d), showing much more frequent use of online platforms among the Jewish population (74%), as compared to the Arab population (45%).

**Figure 7: Booking preferences (online vs. travel agent) as function of socio-demographic attributes**




Source: National survey data

The data also shows direct relationship between the source of travel information (search for flights, hotels, travel packages via the Internet or by a travel agent) and the actual booking venue (via OTA – online travel agents. e.g. Booking.com, Expedia or by travel agents). As can be seen from Table 3, 91% of the survey respondents who indicated that

travel agents are their main source for travel information also stated that they use their services for making the physical bookings. Similarly, 82% of respondents who reported that online platforms constitute their main source of information for flights, hotels and travel packages indicated that they complete the bookings by themselves online.

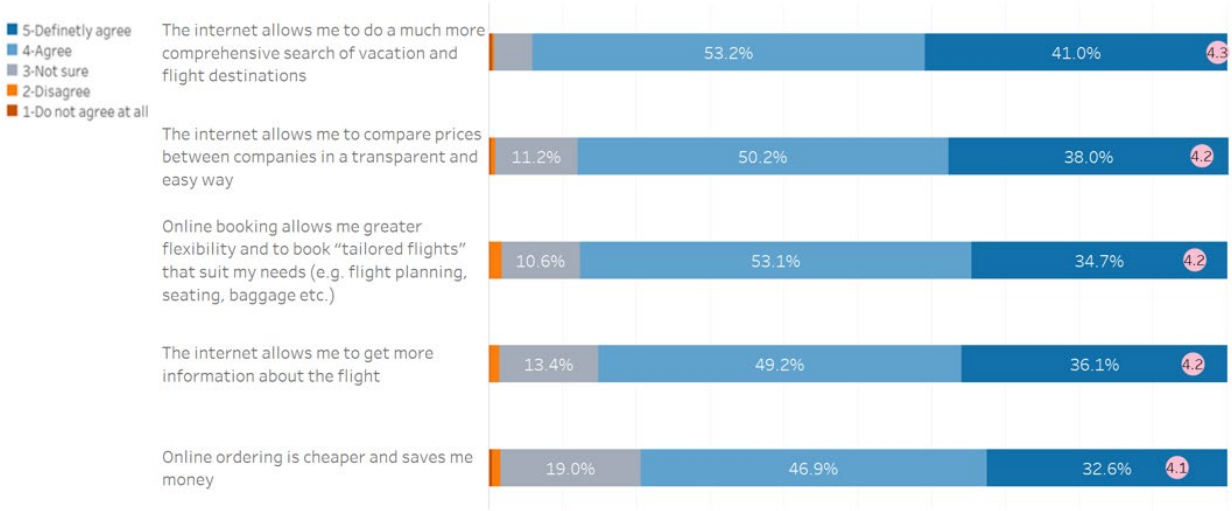
**Table 3: Search for travel information and booking matrix**

| Information source \ Actual booking | | How do you usually make the actual booking (purchase) of a flight or a vacation package? | | | |
|---|---|---|---|---|---|
| | | *Most often through travel agents* | | *Most over the Internet* | |
| | | Count | Row N % | Count | Row N % |
| Where or how do you usually look or get information about a flight or a vacation package? | *Most often through travel agents* | 121 | 91.0% | 12 | 9.0% |
| | *Most over the Internet* | 138 | 17.6% | 644 | 82.4% |

Source: National survey data

The data also shows that 69% of the survey's respondents prefer to make the actual booking of their airline ticket or vacation package via online platforms versus 31% who prefer to involve a travel agent in the process. An interesting question in this regard is what are the main reasons or factors for choosing a human interaction (travel agent) in the booking process on the one hand and what are the main factors for choosing online platforms (self-bookings) on the other hand. Figure 8 presents the main reasons for self-booking of travel related products (e.g. airline tickets, travel packages, hotels) over the Internet for respondents who perform the actual booking via online platforms. As can be seen from the figure, the ability to conduct a comprehensive search is the leading factor in the decision to book online (94% of the respondents definitely agree or agree with this statement), followed by the ability to compare costs (88% agreement), the ability to tailor a flexible flight that suits the traveler's needs (87% agreement), the ability to receive more information about the flight (85% agreement) and the lower cost of online travel products (80% agreement). Please note that the mean values for each item are presented in the pink circles.

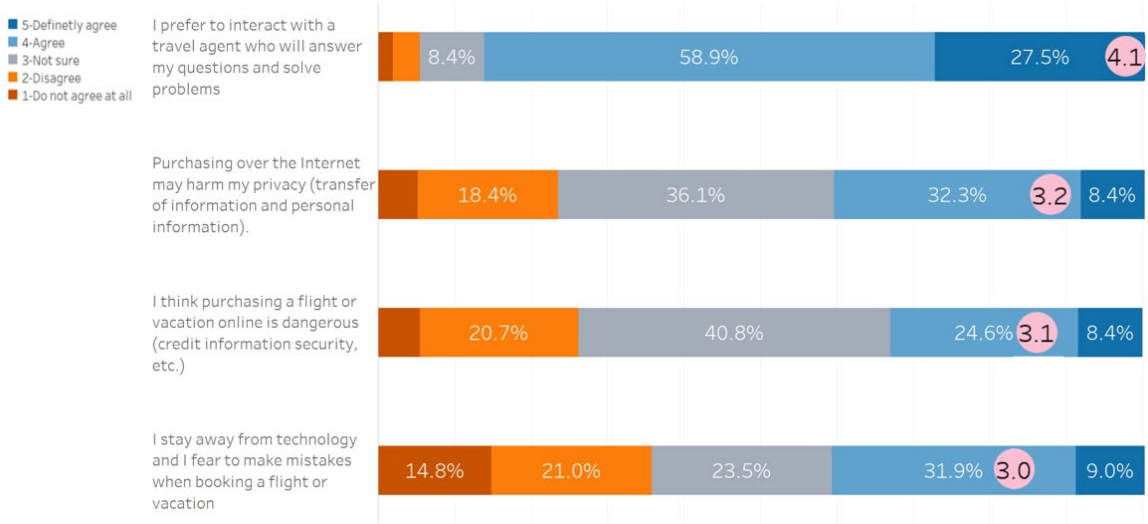
## Figure 8: Main reasons for booking flights online




Source: National Survey data

Figure 9 presents the main factors for using the services of a travel agent for travel bookings (for individuals who do not use online platforms). As can be seen from the figure, the leading factor for choosing a travel agent is rooted in the need to interact with a person who will answer questions and solve problems (86% of the respondents definitely agree or agree with this statement), followed by online privacy and data security concerns (41% agreement) and low digital skills - avoiding technology and the fear of making mistakes when booking online (41% agreement).

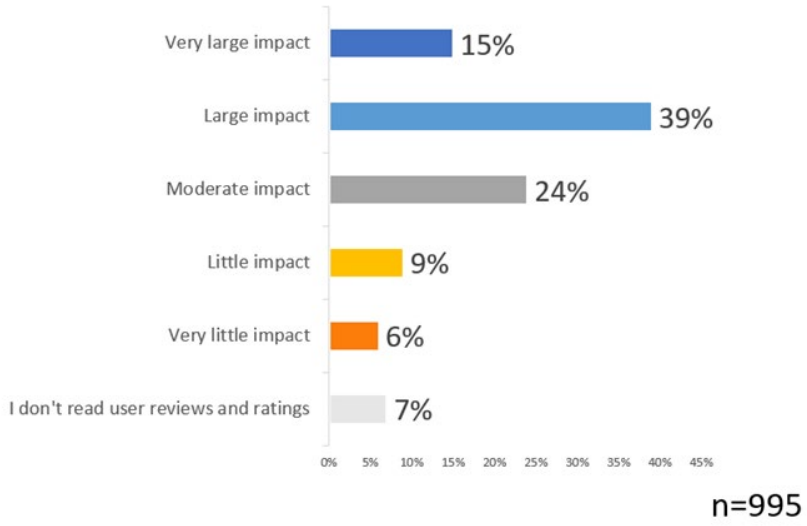
## Figure 9: Main reasons for booking flights by a travel agent




Source: National Survey data

Recent studies have found that online reviews and “ratings” have a significant impact on consumers’ behavior toward hotel selection and booking considerations (Neirotti et al., 2016; Gavilan et al., 2017). The findings of the National Survey show that user reviews and user rating have quite a large effect on the decision to book particular accommodations, with nearly 54% of the respondents indicating that grades, ratings, and opinions appearing on websites such as booking.com, trivago, Airbnb, TripAdvisor etc. affect their decision to either book or not book a particular accommodation (Figure 10).

**Figure 10: The effect of user ratings and reviews on the decision to book hotel accommodations**




Source: Special data processing of the National Survey data

**Table 4: Post-hoc tests (LSD) between age groups, accounting for differences in pair of means (effect of user rating on booking decision)**

| Age (I) | Age (J) | Mean Difference (I-J) | Std. Error | Sig. |
|---|---|---|---|---|
| 55-64 | 18-24 | -.409* | .158 | .010 |
| | 25-34 | -.515* | .152 | .001 |
| | 35-44 | -.382* | .154 | .013 |
| | 45-54 | -.236 | .163 | .149 |
| | 65+ | .426* | .154 | .006 |
| 65+ | 18-24 | -.836* | .145 | .000 |
| | 25-34 | -.942* | .138 | .000 |
| | 35-44 | -.808* | .140 | .000 |
| | 45-54 | -.662* | .151 | .000 |
| | 55-64 | -.426* | .154 | .006 |

Source: Special data processing of the National Survey data

With respect to socio-demographic attributes, we found that in younger age groups, user ratings and reviews had a larger effect (Table 4) on the decision to book accommodations (18-24; 25-34; 35-44) than in older age cohorts (65+; 55-64).

## Online finance

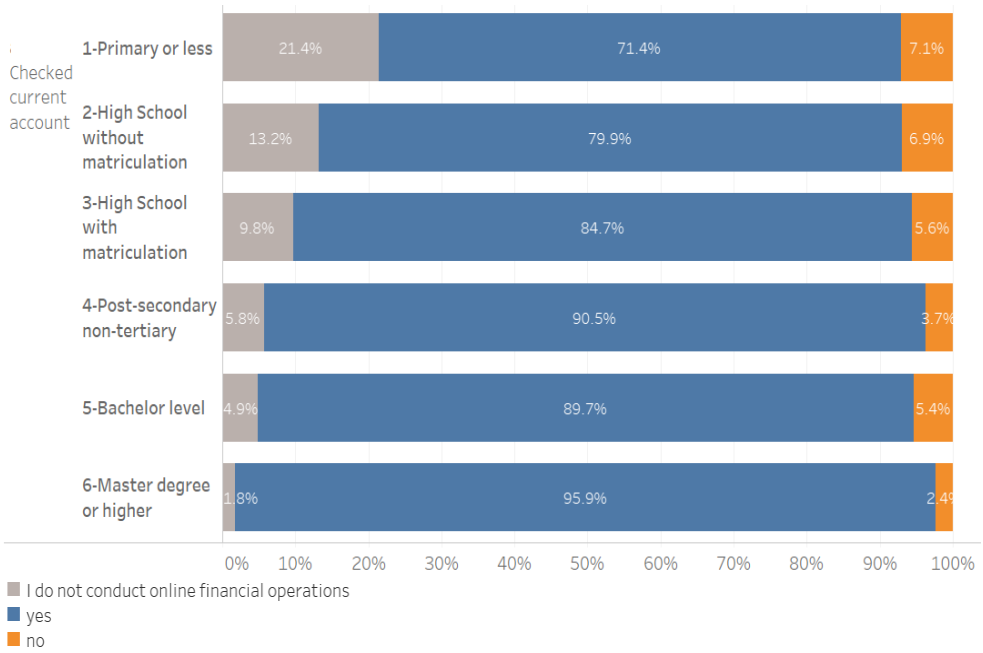
Online banking has grown in an exponential rate in the past decade and is rapidly becoming the prime source for conducting financial transactions by modern-day consumers. Table 5 presents the share of online Internet users who conducted various financial transactions in 2019, parsed by gender. As can be seen from the figure, checking account balance is the most frequent transaction both by men (94%) and women (95%), followed by the payment of bills (59% among men and 46% among women) and cash transfer (60% among men and 62% among women). It is interesting to see that the share of carrying financial activities by male users is higher than its comparable share among female users in almost all transaction categories (with the exception of checking account balance and transferring cash/funds) and the gap is statistically significant. In addition to the gender differences, we can identify a clear linkage between the education level of the user and the scope of online financial transactions (Figure 11), with the share of online use rising with the education level.

**Table 5: Online financial transactions parsed by gender**

| Transaction type | N Male | % | N Female | % |
|---|---|---|---|---|
| Checking account balance | 544 | 94.4% | 580 | 95.1% |
| Payment of bills** | 338 | 58.7% | 282 | 46.2% |
| Cash transfer | 346 | 60.1% | 381 | 62.5% |
| Ordering checkbooks* | 260 | 45.1% | 238 | 39.0% |
| Viewing details of provident funds and pensions** | 223 | 38.7% | 189 | 31.0% |
| Deposit and withdrawal of digital checks* | 177 | 30.7% | 153 | 25.1% |
| Ordering or renewing credit cards** | 171 | 29.7% | 127 | 20.8% |
| Buying and selling stocks and bonds** | 107 | 18.6% | 55 | 9.0% |
| Generating quarterly and annual reports** | 104 | 18.1% | 61 | 10.0% |
| Taking a loan* | 88 | 15.3% | 67 | 11.0% |
| Generating financial documents** | 87 | 15.1% | 55 | 9.0% |
| Changing credit limit | 68 | 11.8% | 46 | 9.2% |
| Buying and selling foreign currency* | 53 | 9.2% | 35 | 5.7% |
| Applying for a mortgage* | 23 | 4.0% | 11 | 1.8% |
| * Difference is significant at the 0.05 level<br>** Difference is significant at the 0.01 level | | | | |

Source: Special data processing of the National Survey data

**Figure 11: Share of online users checking their account balance, parsed by education**




Source: Special data processing of the National Survey data

## Online health

Over the past two decades, the Internet has become a preferred source for finding health information. It is estimated that worldwide, about 4.5% of all Internet searches are for health-related information. Most users of online health information are looking for information about specific health conditions because they or someone they know was diagnosed with a medical condition (Morahan-Martin, 2004). In addition to searching health related data, there is also a constant and steady rise in the use of online health platforms and digital tools which offer patients many online health services such scheduling appointments for physicians, consulting with their family doctor or specialists by video conference, viewing the results of laboratory tests, requesting digital prescriptions etc.

In Israel these digital health services are usually given by the national sick-funds organizations (Kupot-Holim). Previous studies in Israel (e.g. Shahrabani and Mizrachi, 2016; Mizrachi, 2020) have found that significant share (78%) of the general online Internet population uses digital health services. However this share was found to be much lower for older age groups and the Arab population. Our own examination of self-report data and digital trace data has also revealed both gender-based differences and ethnic

gaps (Figure 12 and Figure 13) in the search behaviour of health information and in the use of online health services. As can be seen from Figure 12a, the most frequent digital health activities conducted by Israeli online users are making appointments to a family doctor, followed by viewing laboratory tests and searching for doctors. Findings from the National Survey show clear differences between males and females with respect to conducting online health activities, with women exercising higher online presence in all of the surveyed digital health activities (Figure 12a). Similar trend with respect to gender can be observed from digital trace data (Figure 12b), where women account for 59% of the traffic in the various sick-fund (Kupot-Holim) websites (e.g. Maccabi, Clalit, Meuhedet).

In addition to differences in the use of online health services, substantial gaps can be also observed between female online users and male online users with respect to the search of health related information (e.g. diseases and symptoms; deciphering the results of laboratory test and examinations, information about medicines and drug treatment etc.), with female users exercising higher search activity. Similar result can be observed from the analysis of digital trace data (Figure 12d), which shows parsing of web traffic of popular health information websites by gender. Here too, the majority of traffic (52%) is generated by women. Alvarez-Galvez et al., 2020 suggests that the fact that women are more likely than man to be caregivers, contributes to their higher tendency to access the Internet for health-related purposes.

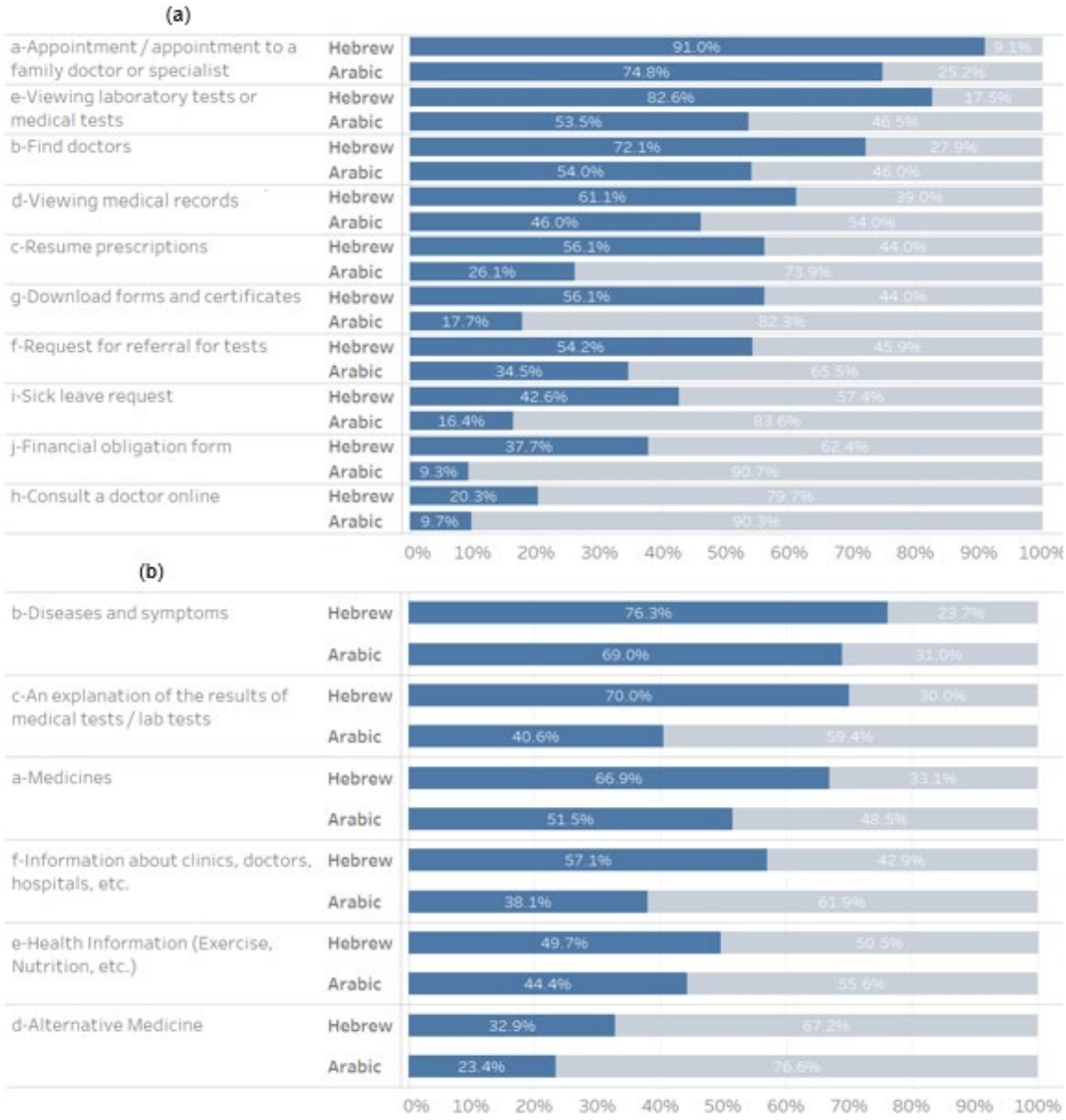
The findings of the National Survey reveal stark and consistent gaps in the use of online health services and in the search behavior of health-related information between Jewish and Arab online users (Figure 13) in almost all of the surveyed items. For example, about 83% of Jewish online users stated that they review the results of laboratory tests, as compared to only 54% of the Arab online users population. Concurrently, 70% of the Jewish online users actively search for possible explanations and deciphering of their laboratory results online, as compared to only 41% among Arab online users. This finding stands in line with previous research (Gamliel, 2017) and highlights the need to keep on the efforts to narrow this sectorial divide in Israel.

## Figure 12: The use of online health services and other health related activities, parsed by gender: comparison between survey data and digital trace data, 2019.


(a) Using online health services – survey data. (b) Traffic share in Israeli sick fund (Kupot Holim) websites – digital trace data. (c) popular search of health-related information – survey data. (d) Traffic share in health information websites - digital trace data.

Sources: National Survey data; SimilarWeb demographic category analysis

## Figure 13: The use of online health services and other health related activities, parsed by ethnic background: comparison between survey data and digital trace data, 2019




(a) Using online health services – survey data. (b) popular search of health-related information – survey data.
Sources: National Survey data; SimilarWeb demographic category analysis

# Chapter 4 - Online Privacy Case Study

The objective of this chapter is to deepen our understanding of online user behavior by triangulating various types of data sources which allow to examine a phenomenon from different angles and resolution levels. Triangulation is a commonly used approach, both in case studies and mixed methods research. In this approach, findings from one method are cross validated by those in another with the aim of achieving greater validity in the research. Denzin (1978), who advocated a multi-source approach, defined triangulation as "the combination of methodologies in the study of the same phenomenon". We demonstrate the triangulation methodology by focusing on ***online privacy as a case study***.

Three types of data sources were used in order to study the phenomenon of online privacy from different angles:

- **Self-report data:** designated **survey questions** and **aggregated indices** pertaining to online privacy and data security.
- **Digital trace data:** clickstream data; websites, keywords and phrases analysis pertaining to online privacy. The data was extracted using **designated online platforms** (SimilarWeb).
- **Social media data:** obtained from micro analysis of the discourse surrounding the concept of "online privacy" using **social media analytics tools**.

## The self-report privacy items

A detailed account of the online surveys is provided in the methodological part of this report (Chapter 2). With respect to online privacy, the national and Binational Surveys included a set of 13 Likert-scale questions (1-5 agreement scale) which examined the attitudes of the respondents towards privacy and data security issues (Table 6). These variables were later parsed by socio-demographic and "big five" (personality traits) variables in order to examine self-perception of online privacy among users.

**Table 6: Privacy and data security variables included in the online surveys**

| # | Privacy and data security variable |
|---|---|
| 1 | I use passwords that are not identical |
| 2 | I use dedicated password management software |
| 3 | I read the privacy regulations before I disclose personal information |
| 4 | I restrict or refuse access to my geographical location (GPS authorization) |
| 5 | I refuse to allow personal information to be used for advertising purposes |
| 6 | When providing personal information, I check that the website uses secure protocol (using HTTPS) |
| 7 | I ask companies or private or public organizations why they need my personal information |
| 8 | I delete browsing history |
| 9 | I use the private / incognito option (In private / in Cognito mode) in my browser |
| 10 | I delete cookies |
| 11 | I use 2-step verification |
| 12 | I use the Tor Browser |
| 13 | I use a VPN (Virtual Private Network) |

## Digital trace analysis

Digital traces on privacy behavior were extracted and analyzed via the SimilarWeb platform. Three "off-the-shelf" tools were used: "Keyword Search Analysis" (limited to desktop use only), "Website Analysis" and "Category Analysis". The data time range was set to 2019 and the location was set to Israel. The Keyword Search Analysis tool was used to search "technical privacy" related terms such as "VPN" and "Tor", with the specific aim of analyzing website traffic share. The Website Analysis and the Category Analysis tools facilitated the understanding of the socio-demographic attributes of privacy. This was done by using the tools' estimations for the distribution of websites visits (in a user-defined category for hard/technical privacy composed of leading websites dealing with VPN use and the TOR Browser), which were parsed by gender and age.

## Social media discourse analysis

The social data analysis was comprised of both computational and manual procedures and involved several steps. The first step involved the creation of "queries" based on term and keywords related to the subject of online privacy. These queries enabled to identify basic elements in social discourse "behavior": Volume; temporal trends and prevalent websites in which the discourse was taking place. Second, a content analysis was performed in order to identify and classify the main discourse themes that were most prominent in the discourse. Data was collected via the **Buzzilla platform**, a social media

analytic tool that crawls the web and collects publicly available conversations and comments from forums, blogs, news websites and social media platforms like Facebook, Twitter and YouTube. Buzzilla was used to search and extract discussions in Hebrew, that were published in a one-year period – from January 2019 to December 2019 (the same time period covered by the online surveys).

The focus was placed on analyzing public comments that referred specifically to the concept of online privacy. Therefore, there was a need to define appropriate queries that reflect public perceptions surrounding this issue. The following queries (in Hebrew) were used***: Online privacy, Tor Browser, Browsing history and Incognito browsing***. It is important to note that even though the discourse surrounding online privacy includes many other issues such as online photo sharing, cyber hacking and etc., due to the limited scope of this study, we decided to focus on terms which corresponded to the survey questions on the one hand, and gained enough volume in the social media platforms on the other hand.

Several aspects of the public discourse surrounding online privacy were analyzed:

- Volume analysis – how popular is the discourse about online privacy
- Trend analysis – when are people discussing online privacy
- Discourse/theme analysis – which issues are in the focus of interest and which are not
- Audience analysis – who participates in the conversation about online privacy

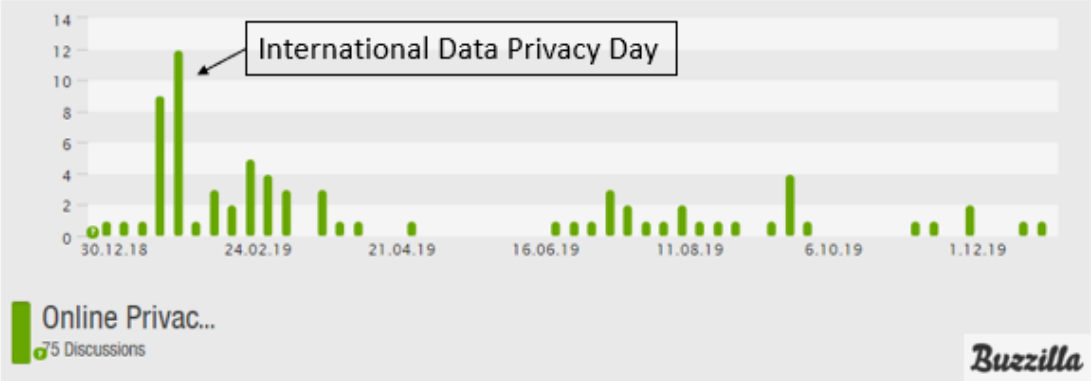
Although social data analysis is based on computational approach in its core, human involvement is still needed when the context needs to be taken into account. Interweaving computational and manual approaches can improve the overall analysis process by enabling us to simplify the entire procedure and also to magnify the results of a small data set (Lewis, Zamith, & Hermida, 2013). Thus, we used theme analyses to examine the online discourse surrounding “online privacy” and we categorized it according to prominent themes that were discovered during the analysis.

## General perception of “online privacy”

The availability of public social data originating from conversations and comments from forums, blogs, news websites and social media platforms (e.g. Facebook) allowed to obtain a glimpse of the perception of online privacy and data security among the Israeli population.

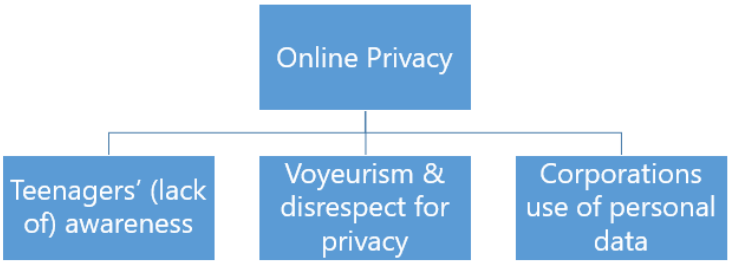
The term “online privacy” was mentioned in 75 discussions during 2019. About 20 of these discussions were taking place at the time of International Data Privacy Day (Figure 14). Analysis of the discussions revealed three prevailing themes: Concerns regarding teenagers’ (lack of) privacy awareness; Moral judgement and concerns regarding people’s disrespect for others’ privacy; and concerns regarding corporations use of personal data (Figure 15).

**Figure 14: Online privacy discourse volume**




Source: query and data are powered by Buzzilla

**Figure 15: Online privacy discourse themes**




### Concerns regarding teenagers’ (lack of) privacy awareness

The first theme surrounding online privacy dealt with the lack of privacy awareness among teenagers. This theme was comprised by several articles (Figure 16) which covered a survey conducted by Israel’s ministry of Justice, that indicates that half of teenagers do not apply privacy considerations when using apps. This finding received exposure in several media outlets and reflected the perceived public concerns around online behavior among youth. Notably, these articles have not produced comments on the news websites, which may reflect the lack of wide public interest in this issue.

**Figure 16: Examples of lack of online privacy awareness among youth in online articles**

הרשות להגנת הפרטיות: 48% מבני הנוער מאשרים אנשים שאינם מכירים כחברים ברשתות החברתיות

מסקר שערכה הרשות לרגל יום הגנת הפרטיות הבינלאומי עולה עוד, כי בני נוער מורידים אפליקציות מבלי להתייחס למדיניות ולשיקולי פרטיות וכי נשים אחראיות יותר מגברים באישור אנשים שאינן מכירות

משה גורלי 06:55 28.01.19

חדשות סכנולוגיה, סכנולוגי

סקר חושף: כמה בני נוער מודעים לפרטיות ברשת?

סקר חדש של משרד המשפטים חושף נתונים מדאיגים על המודעות של בני נוער לנושא שמירה על פרטיות ברשת. לפי הממצאים, כמעט חצי מבני הנוער מאשרים אנשים שהם לא מכירים כחברים בפייסבוק, ורובם לא בוחנים הרשאות באפליקציות

אוריאל באר, חדשות סרונים 10:06 28.01.19 כ"ב בשבט תשעט

TheMarker | TechNation

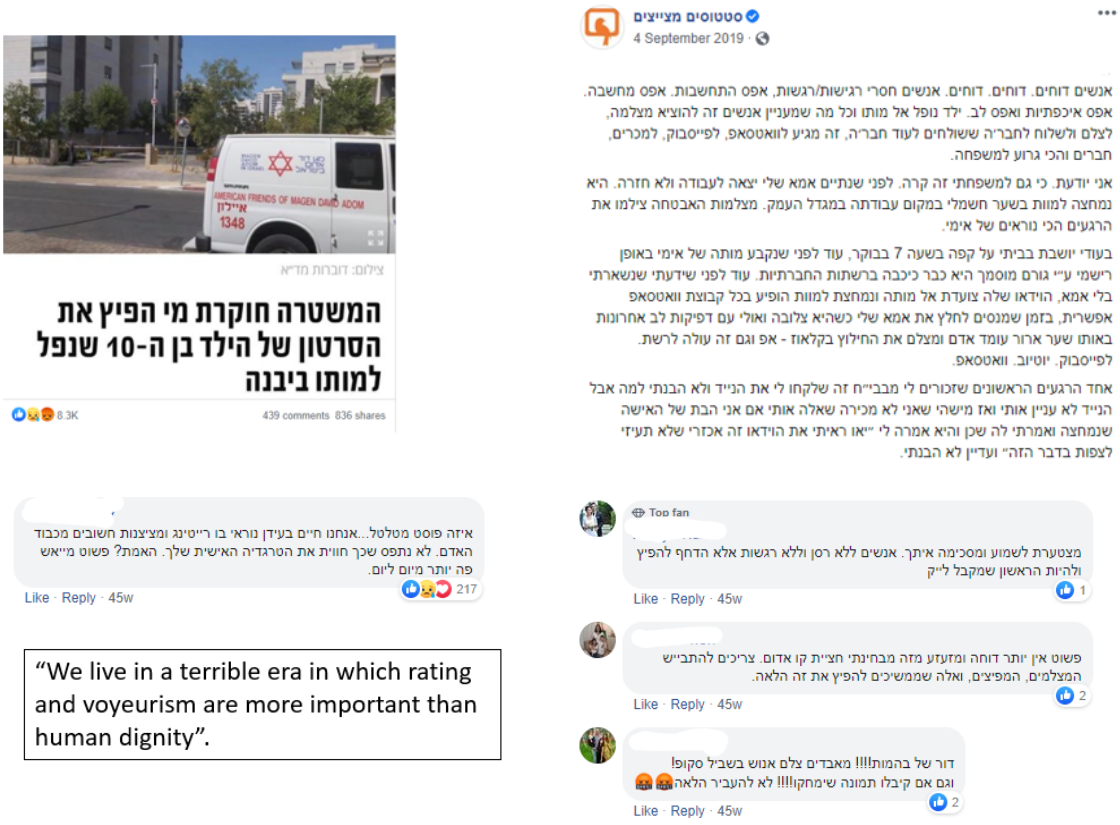
מי צריך פרטיות? הצעירים מוכנים שיידעו עליהם הכל ברשת

סקר שערכה הרשות להגנת הפרטיות במשרד המשפטים חושף התנהלות מדאיגה של בני נוער ברשת ■ קרוב ל-50% מהם מאשרים אנשים זרים ברשתות חברתיות ■ רק 3% מעדכנים סיסמאות בתדירות גבוהה

### Moral judgement and concerns regarding disrespect for privacy

This theme focused on individuals' dismay of the ease of sharing other people's private moments and private information online, without asking for their consent. This theme evolves mainly around moral judgement. For example, the most commented post in 2019 that included the term "online privacy" was a testimony of a woman who found out about her mother's death in an accident from social media posts (Figure 17). The comments to this post included sympathizing expressions and moral judgement of people who share photos of other people's private moments (who happen to be in public sphere) which is perceived as nothing more than voyeurism. For example, one person commented: "We live in a terrible era in which rating and voyeurism are more important than human dignity". This comment gained 217 "likes". Another one commented: "There are people with no boundaries and no emotions, who are only motivated by the urge to get as many "likes" as possible".

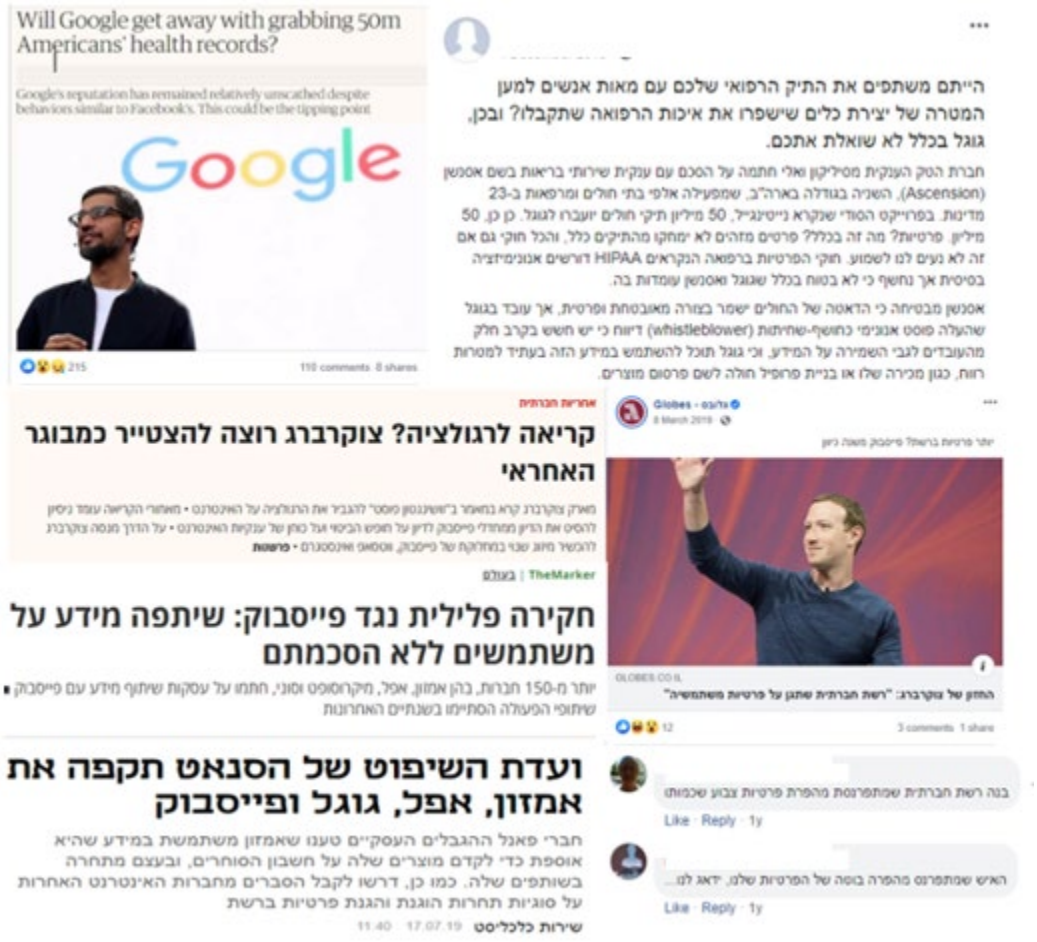
**Figure 17: Moral judgement of people who share photos of private events**




## Concerns regarding corporations' use of personal data

This theme evolves mainly around big companies' monitoring abilities, private data collection and their unprecedent power. For example, the second most commented post in the study's time period (2019) was an article about Google attaining 50 million health records without the owners' consent. Further examples are articles and posts about corporations' use of personal data, regulation and the investigations of the "big four" – Google, Facebook, Amazon and Apple. The sentiment surrounding this issue is mostly negative and reflects fear and anger (Figure 18). This concern about the misuse of power by corporations and commercial companies is reported by numerous studies (Dror and Gershon, 2014; Raban and Soffer, 2014; Dialogue Organizational Consulting, Research and Training, 2019), with almost half of the Israeli population (49%) feeling that the use of personal data by private companies jeopardizes their privacy (Raban and Soffer, 2014). Apparently, this feeling is much more prevalent among older age groups (55+) than younger (12-17; 35-54) age groups (Dror and Gershon, 2014).

**Figure 18: Concerns regarding corporations' use of personal data**




## A triangulated approach for investigating online privacy

The availability of social media discourse data and digital trace data allows us to triangulate it with the survey data and to deepen our understating on online privacy.

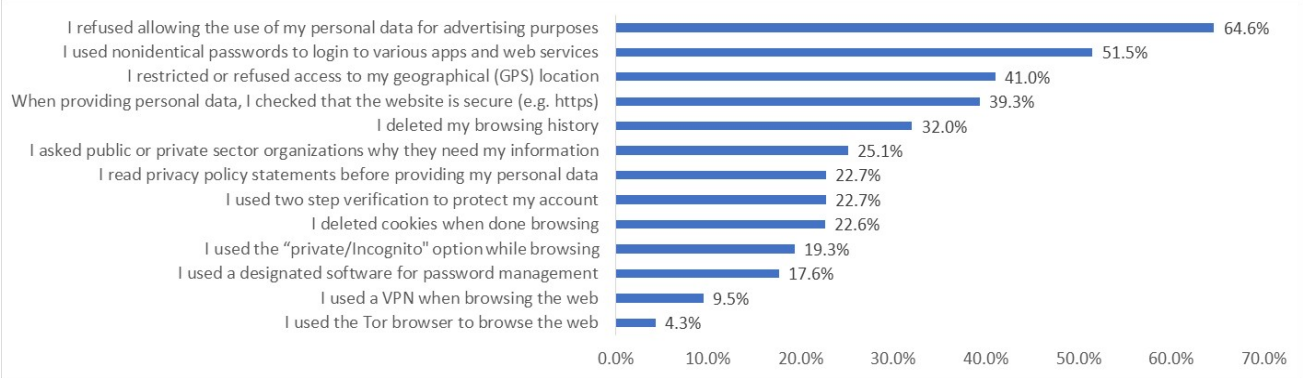
Table 7 and Figure 19 present descriptive statistics for the privacy and data security items for the Israeli population. As can be observed from the data, the most frequent precaution that users exercise in protecting or maintaining their privacy online is “refusing to allow the use of their personal data for advertising purposes” (65% of the respondents exercise it often or very often ; mean score of 3.8 on a 5 point scale), followed by "using nonidentical passwords to login to various apps and web services” (52%; mean: 3.5) and “restricting or refusing access to their geographical (GPS) location” (41%; mean: 3.5). The least frequent precaution in the protection of privacy or data security online is using a designated software for password management browser (18% use it often or very often;

mean: 2.0) and using online tools such as VPN (10%; mean: 1.8) and the Tor Browser (4%; mean: 1.3).

**Table 7: Privacy and data security items in the Binational Survey**

| Privacy and data security items | N | Std. Deviation | Mean |
|---|---|---|---|
| I refused allowing the use of my personal data for advertising purposes | 1191 | 1.2 | 3.8 |
| I used nonidentical passwords to login to various apps and web services | 1255 | 1.3 | 3.5 |
| I restricted or refused access to my geographical (GPS) location | 1208 | 1.2 | 3.2 |
| When providing personal data, I checked that the website is secure (e.g. https) | 1089 | 1.5 | 2.9 |
| I deleted my browsing history | 1206 | 1.3 | 2.9 |
| I read privacy policy statements before providing my personal data | 1212 | 1.3 | 2.5 |
| I deleted cookies when done browsing | 1089 | 1.3 | 2.5 |
| I asked public or private sector organizations why they need my information | 1155 | 1.5 | 2.5 |
| I used the “private/Incognito" option while browsing | 1129 | 1.2 | 2.4 |
| I used two step verification to protect my account | 1010 | 1.3 | 2.4 |
| I used a designated software for password management | 1124 | 1.4 | 2.0 |
| I used a VPN when browsing the web | 949 | 1.1 | 1.8 |
| I used the Tor Browser to browse the web | 876 | 0.8 | 1.3 |

**Figure 19: Distribution of online privacy and data security items - percent replaying "often or always"**




The analysis of social media discourse enables us to determine which type of audience takes part in online privacy discussions. In order to understand who participates in online privacy technical related discussions, we created several queries to help in identifying the discourse and where it takes place. We searched for discussions that include the terms: “Incognito/InPrivate Browsing”, “Tor Browser” and “Browsing History”. These terms appear in the survey questions and it is reasonable to assume that they will reflect interest in online privacy by people who use them. The term “incognito/InPrivate browsing” was

mentioned 499 times during 2019 (Figure 20). The analysis shows that 73% of the discourse was taking place in forums, where people can maintain anonymity (Figure 21). In addition, the data indicates that 48% of the discourse in the forum arena was taking place in forums of ***Jewish religious communities*** such as Prog and Netfree, and 46% of the discourse took place in ***teenagers forums*** like Stips and Fxp (Figure 22).

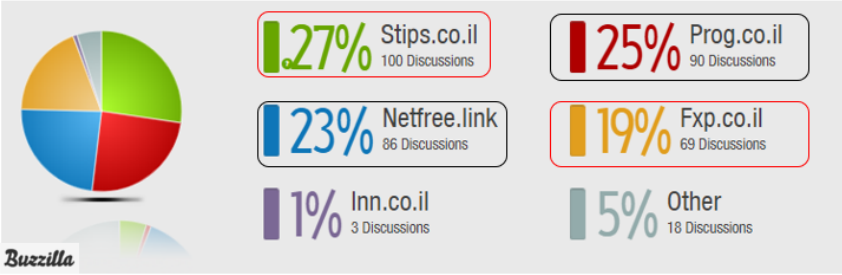
**Figure 20: "Incognito/InPrivate browsing" discourse volume**


Source: query and data are powered by Buzzilla

**Figure 21: "incognito/InPrivate browsing" discourse arena distribution**


Source: query and data are powered by Buzzilla

**Figure 22: "incognito/InPrivate browsing" discourse forum distribution**



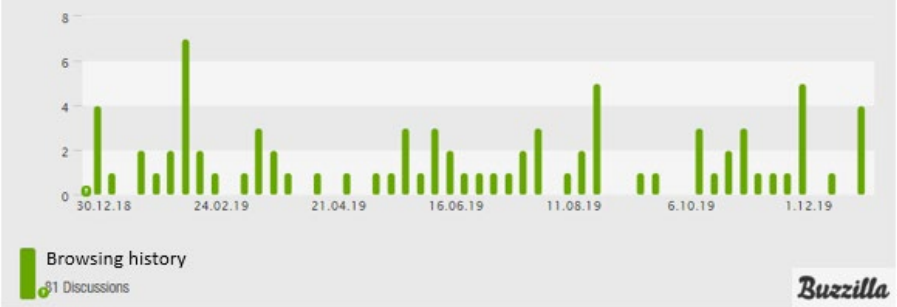

Source: query and data are powered by Buzzilla

In Prog and Netfree the term "incognito/InPrivate browsing" is mentioned frequently in shopping related issues, where users advise others to use incognito mode to get better

prices or to use discount coupons. In contrast, in the teenagers' forums Stips and Fxp the term was frequently mentioned by teenagers who wished to hide their browsing data (from their parents, school etc.). For example, one user asked: "can my school see my browsing history, even if I am in incognito mode"? Another one asked: "Is there a way to see what websites I go to while browsing in an incognito mode"?

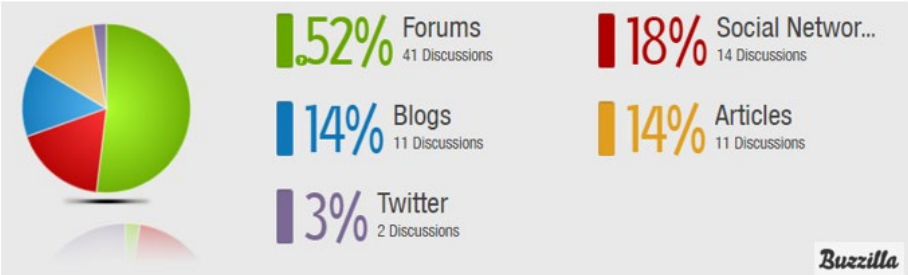
The term "Browsing history" was mentioned 81 times during 2019 (Figure 23). It is important to note that the discourse around "Browsing history" was mentioned in a variety of technical contexts, however the focus here is on privacy concerns ("how can I delete my browsing history"?). The data shows that 52% of the discourse took place in forums (Figure 24). The majority (64%) of the discourse in the forum arena was taking place in teenagers forums such as Stips and Fxp (Figure 25).

**Figure 23: "Browsing history" discourse distribution**



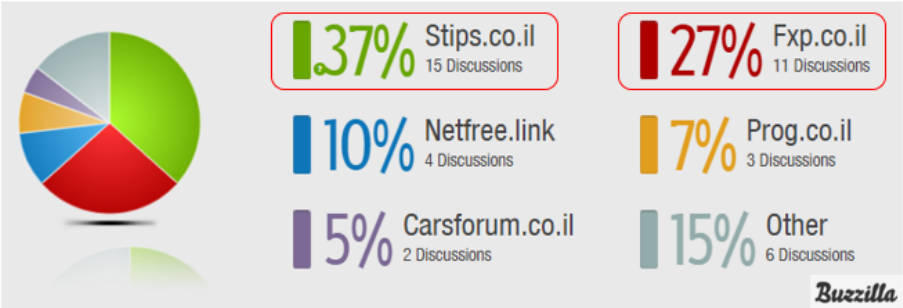

Source: query and data are powered by Buzzilla

**Figure 24: "Browsing history" discourse arena distribution**




Source: query and data are powered by Buzzilla

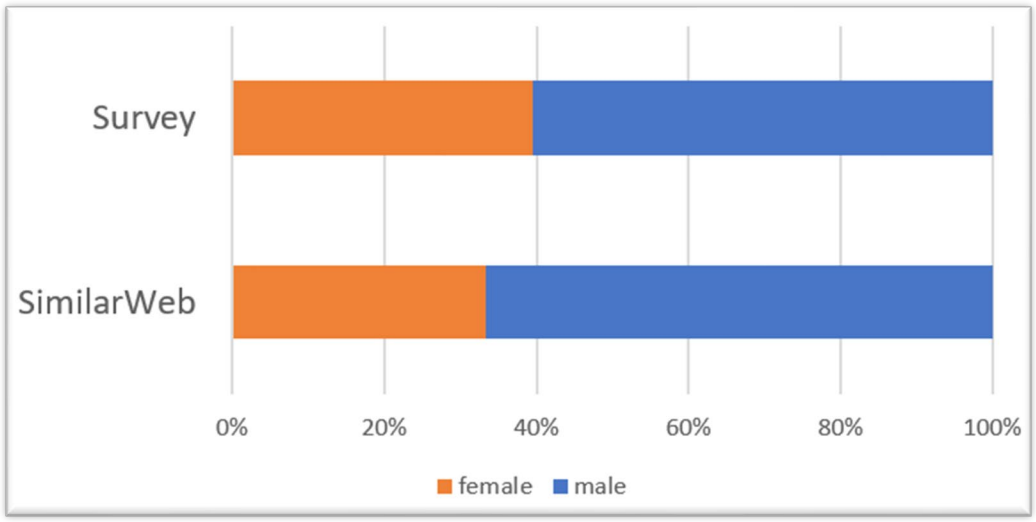
**Figure 25: "Browsing history" discourse forum distribution**




Source: query and data are powered by Buzzilla

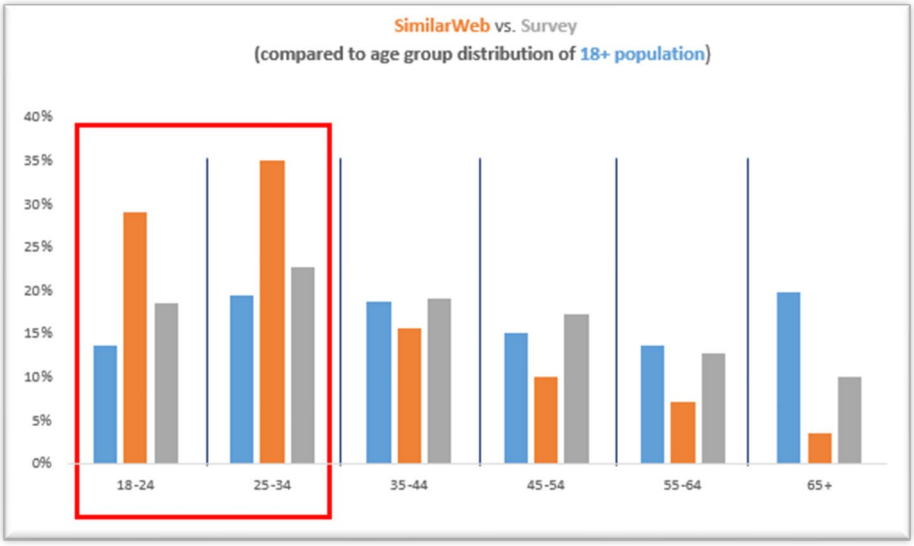
VPN (virtual private network) is a service or tool that enables users to maintain online privacy and anonymity by creating a private network from a public Internet connection. Figure 26 and Figure 27 describe gender and age differences in VPN use among Israeli 18+ population as exemplified by both self-report data (Binational Survey) and digital trace data (SimlarWeb) for the 2019 time period. As can be clearly seen from the figure, both data sources show substantially higher signals of VPN use among male users (Figure 26) and younger age cohorts (Figure 27).

**Figure 26: Distribution of VPN use by gender and data source**




Data Sources: Binational Survey and Similarweb – *18+ population (Israel)*

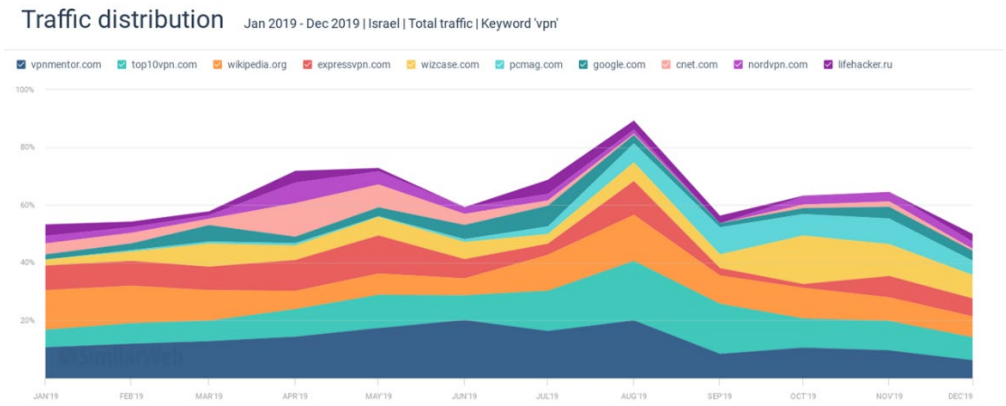
## Figure 27: Distribution of VPN use by age group and data source




Data Sources: Binational Survey and Similarweb – *18+ population (Israel)*

The SimilarWeb tool enables us to analyze website traffic share for a specific search term over time. As can be seen from Figure 28, the two main websites used by Israeli online users to receive information about VPN use are Vpnmentor.com (which aims "to offer users a fair, committed and efficient tool for VPN navigation and web browsing while maintaining privacy") and top10vpn.com. It is important to note however that it is not possible to determine from the data if the purpose of the user in using/receiving information about VPN is related to maintaining privacy or for other technical reasons such as bypassing geographical blocking (e.g. consuming TV series and movies from overseas).

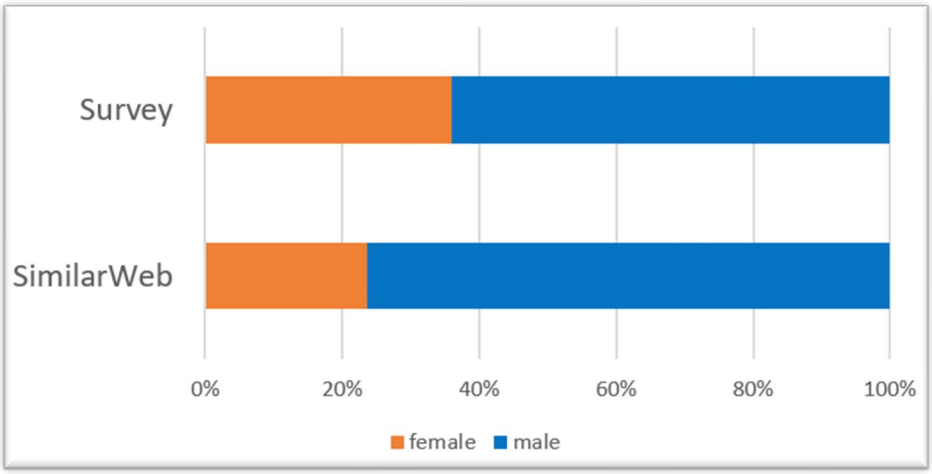
## Figure 28: Website traffic distribution for the search term "VPN"




Source: SimilarWeb

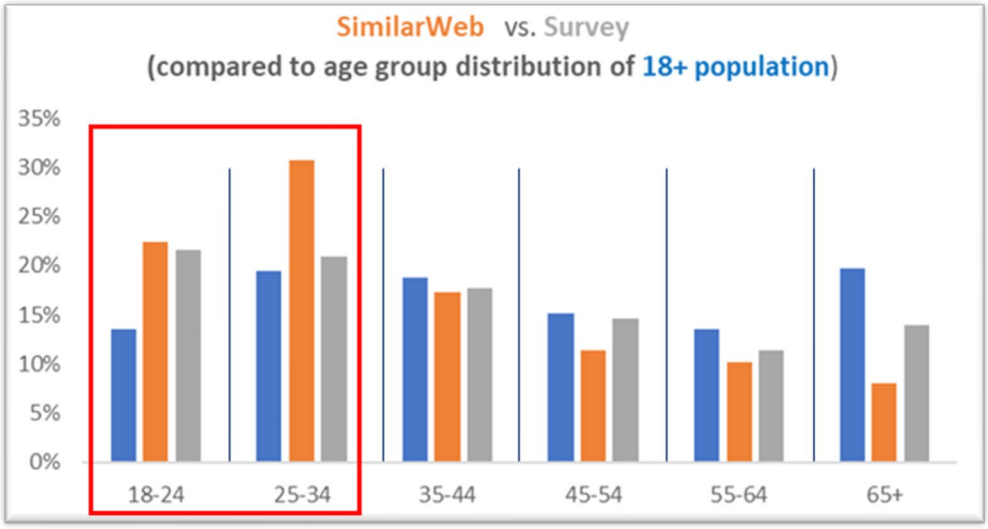
The Tor Browser is a free and open-source software for enabling anonymous communication online. Figure 29 and Figure 30 describe gender and age differences in TOR browser use among Israeli 18+ population as shown by the survey data and digital trace data (SimlarWeb) in 2019. In accordance with our findings on VPN use, both data sources show strong signals of TOR Browser use among male users (Figure 29) and young age cohorts (Figure 30).

**Figure 29: Distribution of TOR Browser use by gender and data source**




Data Sources: Binational Survey and Similarweb – *18+ population (Israel)*

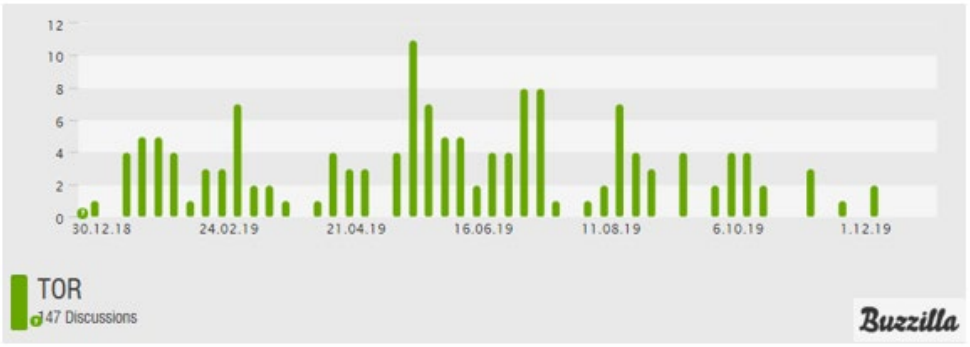
**Figure 30: Distribution of TOR Browser use by age group and data source**




Data Sources: Binational Survey and Similarweb – *18+ population (Israel)*

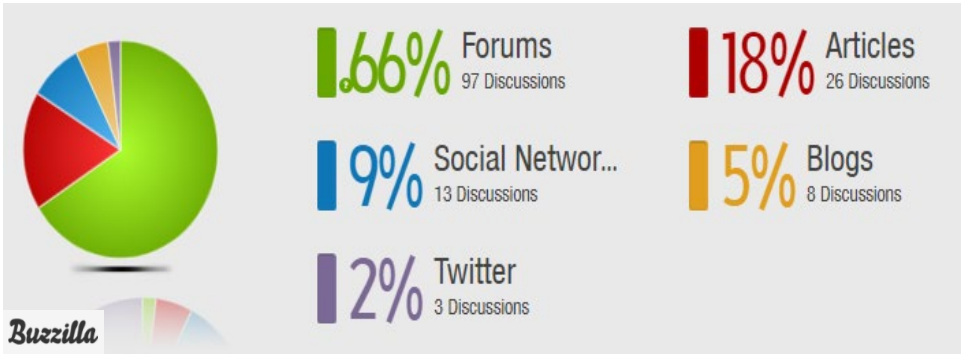
The analysis of public Israeli social media discourse shows that the term “Tor Browser” was mentioned 147 times during 2019 (Figure 31). About 66% of the discourse took place in forums (Figure 32) and 79% of the discourse in the forum arena was taking place in teenagers forums such as Stips and Fxp (Figure 33). The term is mentioned mainly in technical related issues, where users discuss using Tor Browser for anonymity and data protection. For example: “What is the best way to remain anonymous online? Can Tor Browser assure anonymity? What is the best way to do something without leaving traces?”

**Figure 31: “Tor Browser” discourse volume**



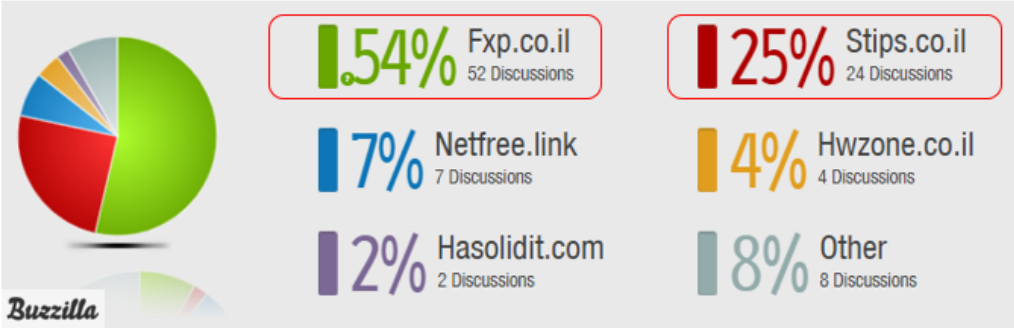

Source: query and data are powered by Buzzilla

**Figure 32: “Tor Browser” discourse arena distribution**




Source: query and data are powered by Buzzilla

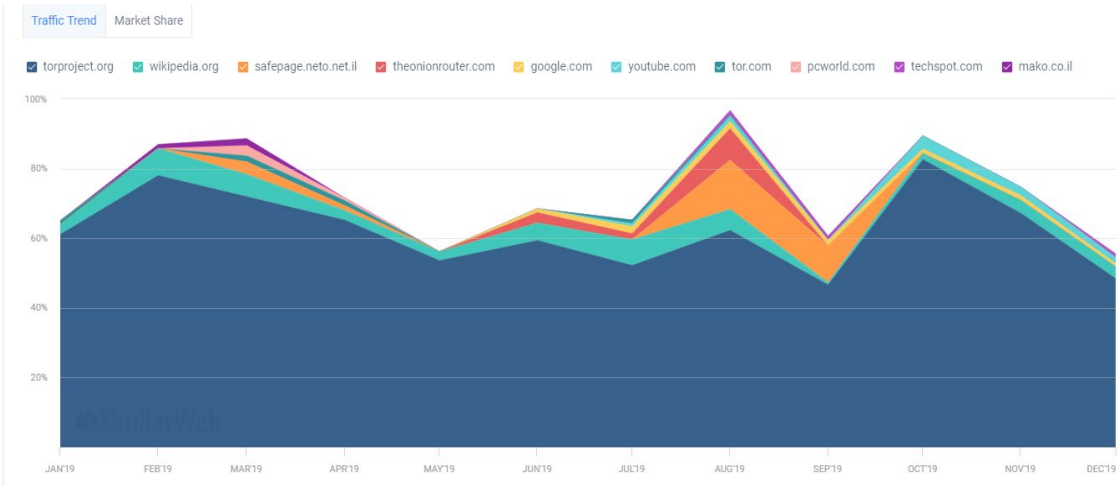
Figure 33: “Tor Browser” discourse forum distribution




Source: query and data are powered by Buzzilla

As can be seen from Figure 34, the main website used by Israeli online users to receive information about TOR use is **torproject.org** (aims at “defending against tracking and surveillance and circumventing censorship”), which also allow users to download the browser. Over 60% of the website traffic relating to “TOR” by Israeli online users in 2019 was conducted via this website.

Figure 34: Traffic distribution for the search term “TOR”




Source: SimilarWeb

## The privacy indices

As mentioned earlier in this chapter, the online survey included 13 different privacy questions. Factor Analysis was performed on the 13 various online privacy and data security items in order to reduce the number of variables into a set of aggregated

underlying variables. This statistical procedure identified three aggregated factors for privacy (Table 8) which explained 51% of the variance of the squared loadings.

**Table 8: Factor analysis results - rotated component matrix**

| | Component | | |
|---|---|---|---|
| **Privacy questions** | 1 | 2 | 3 |
| I refuse to allow personal information to be used for advertising purposes | .750 | | |
| I restrict or refuse access to my geographical location (GPS authorization) | .675 | | |
| I read the privacy regulations before I disclose personal information | .586 | | |
| When providing personal information, I check that the site uses secure protocol (using HTTPS) | .525 | | |
| I ask companies or private or public organizations why they need my personal information | .482 | | |
| I use passwords that are not identical | .436 | | |
| I use the Tor Browser | | .764 | |
| I use a VPN (Virtual Private Network) | | .695 | |
| I use dedicated password management software | | .619 | |
| I use 2-step verification | | .482 | |
| I delete cookies | | | .797 |
| I delete browsing history | | | .741 |
| I use the private / incognito option (In private / in Cognito mode) in your browser | | | .685 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

We labeled the first component or factor (shaded in yellow in Table 8) as ***"General Privacy"***, the second factor (shaded in orange) as ***"Hard Technical"*** and the third factor shaded in green) as ***"Soft Technical"***. The following bullets present a definition for each factor:

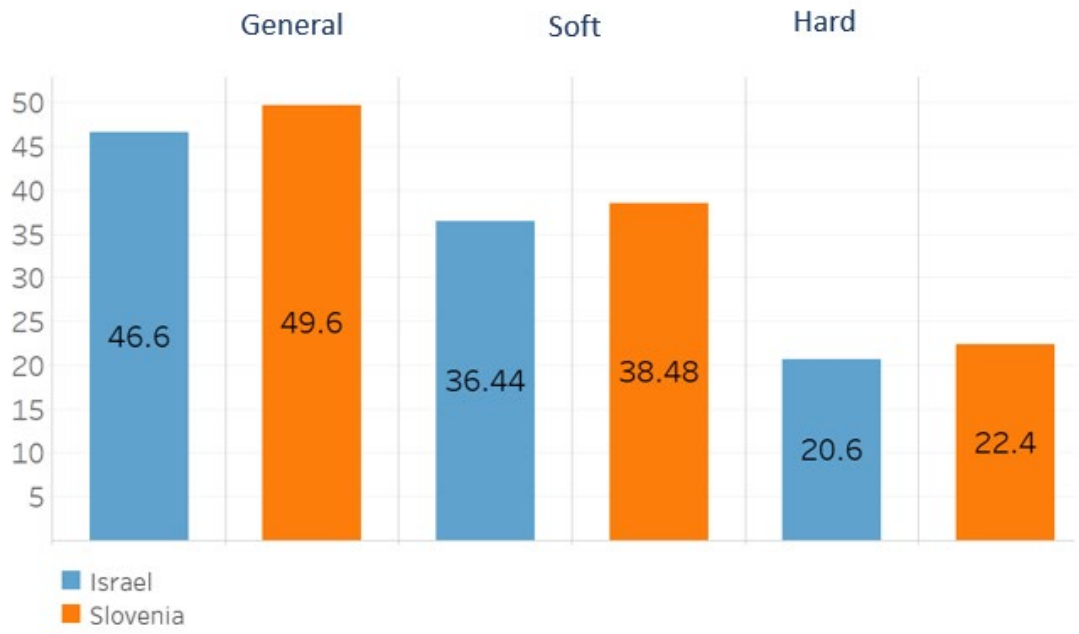
- **General Privacy (GP)** – Reading privacy statements and being aware of the use of personal information by third parties; restricting access to personal data.
- **Soft Technical (ST)** – Carrying out simple, routine measures to maintain/secure user anonymity & privacy online.
- **Hard Technical (HT)**– Using more complex and designated tools, technologies and software in order to protect privacy, data security and anonymity online.

A mathematical transformation was used to constrain the **sum score of each set of variables** or factor to be at the 0-100 range (this was done for each respondent/observation) in order to create a common metric (index) that enables to conduct comparison within and between the three factors:

1. *GP index for respondent i= round(100*((SUM $GP_i$- MIN $GP_i$)/(MAX $GP_i$-MIN $GP_i$))*
2. *ST index for respondent i= round(100*((SUM $ST_i$- MIN $ST_i$)/(MAX $ST_i$-MIN $ST_i$))*
3. *HT index for respondent i= round(100*((SUM $HT_i$- MIN $HT_i$)/(MAX $HT_i$-MIN $HT_i$))*

Figure 35 and Figure 36 present the mean scores of these three indices, parsed by population (Israeli, Slovenian) and gender. Figure 35 presents the mean scores of the privacy indices, by population breakdown. As can be seen from the graph, the score for general privacy is significantly higher than the soft privacy and the hard privacy scores (P<0.001). This means that relatively small share of online users takes active and serious technical measures to protect their privacy. The Slovenian privacy scores are slightly higher than the Israeli privacy scores and are statistically significant (P<0.001 for General privacy; P<0.05 for soft and hard technical).

**Figure 35: Population differences in privacy indices**




As for gender-based differences with regards to privacy (Figure 36), it is evident that male users display much higher soft technical and hard technical skills than female users. These differences are statistically significant (P<0.001). Similar trend can be observed from the analysis of digital trace data (Figure 37) which shows higher signals for hard technical skills among male users.

**Figure 36: Gender differences in privacy indices**

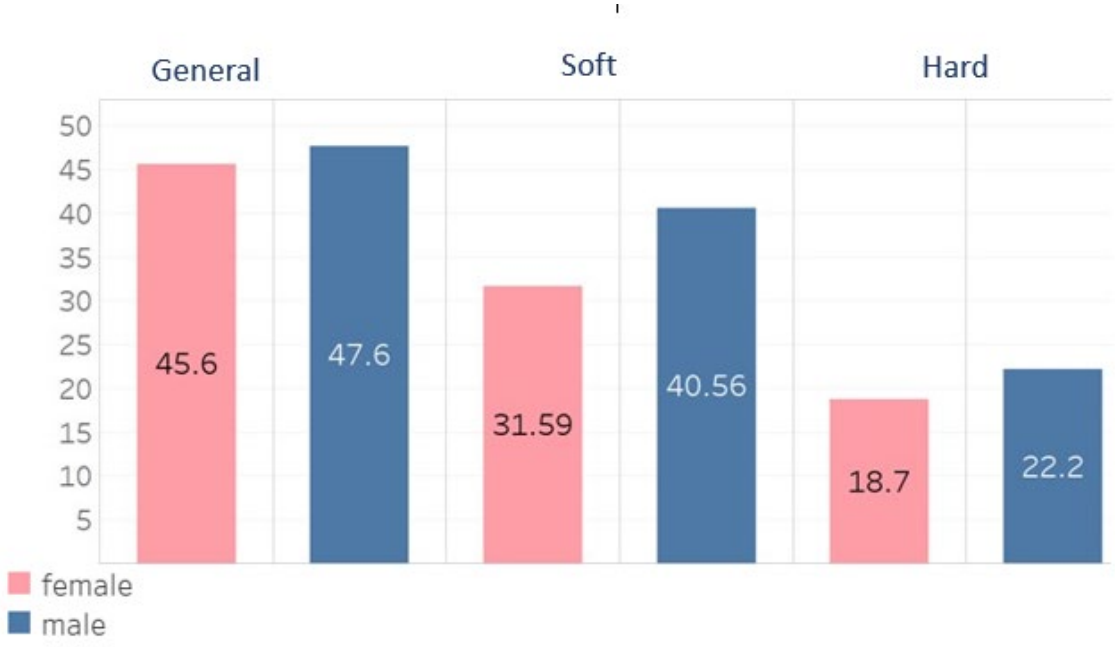
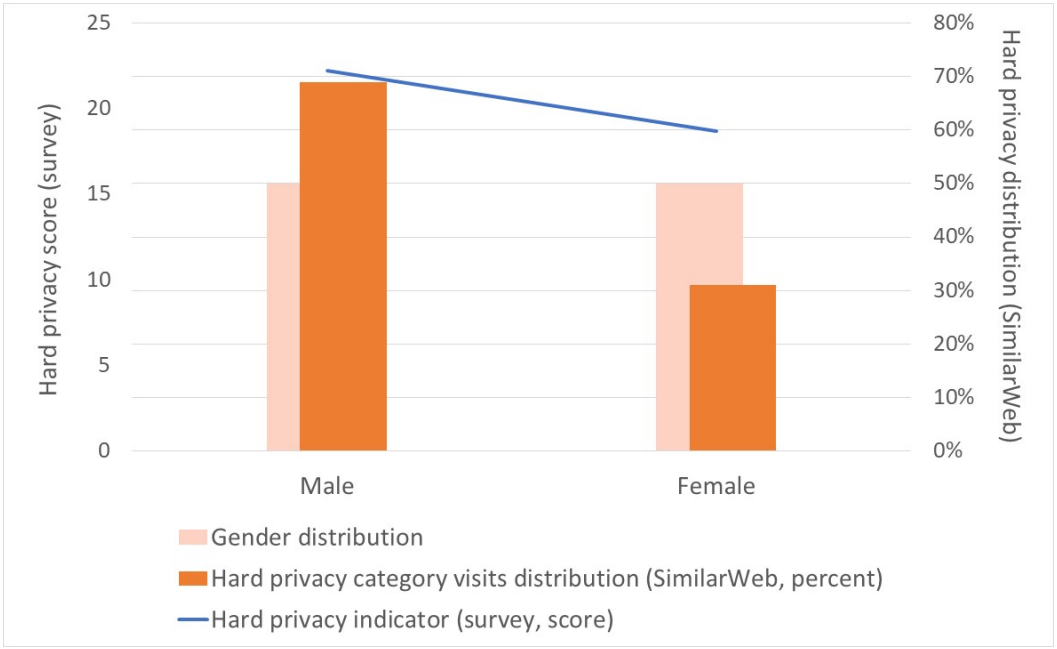



**Figure 37: Gender differences in hard privacy index – survey data versus digital trace data**

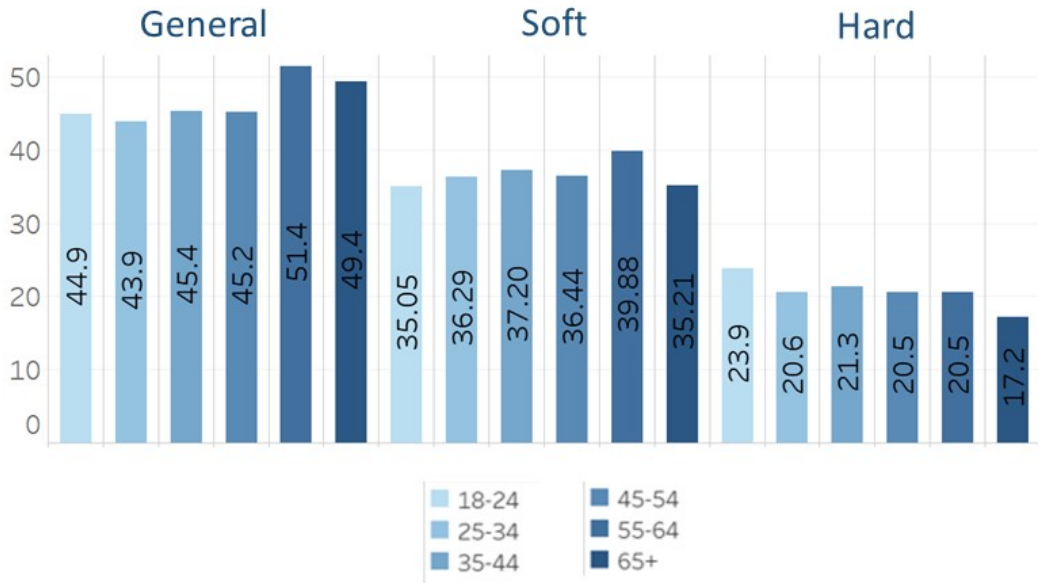



The hard-technical category is combined by torproject.org and vpnmentor.com

In accordance with our findings on soft and hard technical privacy skills, a research conducted in 2019 by the Israeli organizational consulting firm Dialogue for the Israel Protection Authority has found that female users were less likely than male users to use

firewalls, spam filters and antiviruses to protect their privacy online. The next three figures present the mean scores for the general privacy, soft technical and hard technical privacy indices by age and education breakdown. As can be noticed from Figure 38, the hard technical privacy scores are higher for the younger age groups (18-24) and they statistically differ from the older age cohorts (65+; 55-64; 65+) at the 0.01 level. The general privacy score (e.g. carefully reading privacy statements and being aware of the use of personal information by third parties) is higher for the older age cohorts (65+) and statistically differs from the younger age cohorts (18-24; 25-34) at the 0.01 level

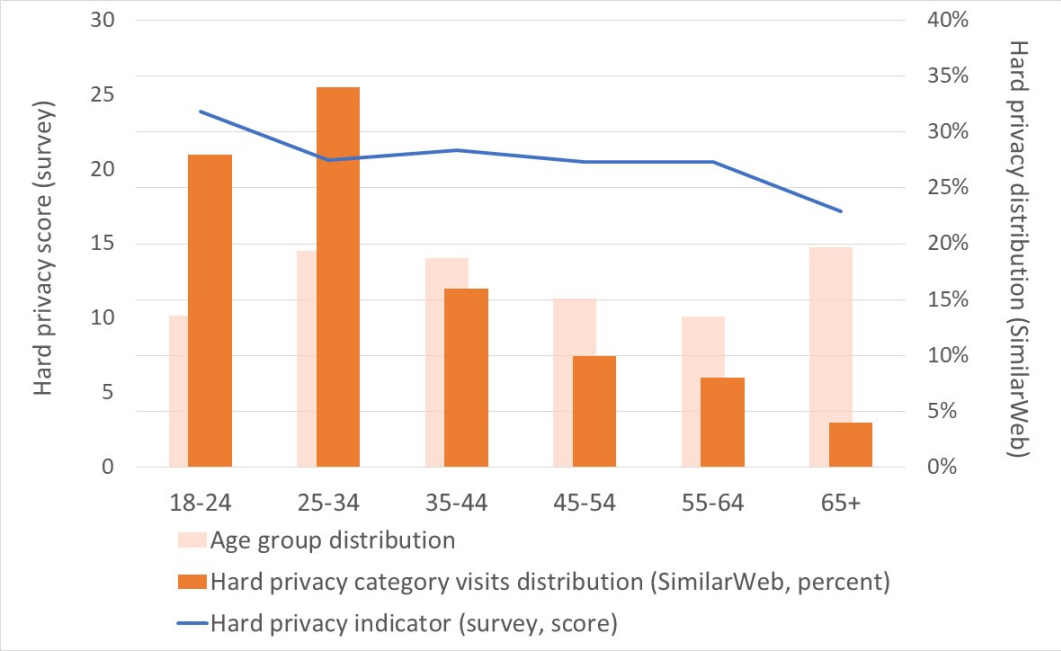
**Figure 38: Age differences in privacy indices**




Similar trend can be observed from the analysis of digital trace data (Figure 39) which shows higher signals for hard technical skills among younger online users. In accordance with our findings on general privacy, the Dialogue survey on online privacy reports that general privacy attributes such as carefully reading privacy statements and refusing to authorize app permissions are much more prevalent among older age cohorts than young age cohorts (Dialogue, 2019). In concordance with our findings on soft and hard technical skills, Dror and Gershon (2014) report that 73% of young online users in Israel declared that “they know what tools to use in order protect their privacy”. As age increases, the proportion of online users who concord with this statement at age 55 and over stands on 42%. According to the authors, these findings may teach one of two things: either young online users express youthful arrogance that is not necessarily based on reality or, alternatively, as digital natives and unlike older online users, they are better skilled in

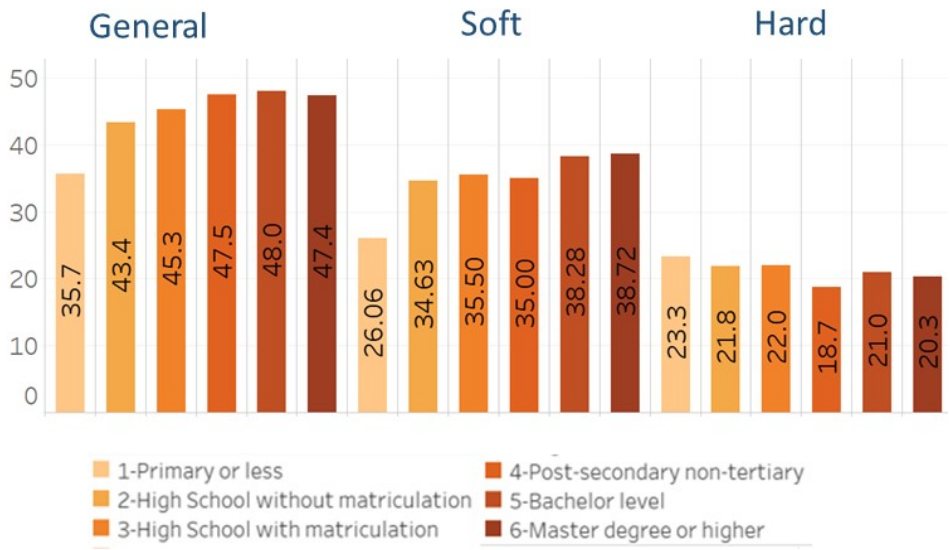
online activities and more acquainted with the tools designed to protect their privacy. Privacy scores rise with the education level, especially for the general and soft technical indices (Figure 40).

**Figure 39: Age differences in hard privacy index – survey data versus digital trace data**




The Hard-privacy category is combined by torproject.org and vpnmentor.com

**Figure 40: Education differences in privacy indices**



### Modeling the relationship between socio-demographic and behavioral attributes and online privacy

Three linear regression models were applied to test the relationship between various socio-demographic and behavioral attributes of online users and online privacy (Table 9). The independent socio-economic variables and behavioral attributes include: A dummy variable representing male users; age (continuous); education level (ordinal); two ordinal (big-five) variables denoting self-perception of order ("I get chores done right away", "I like order") and four dichotomous variables denoting behavior in social networks (use of real name, stating personal status, posting family photos, indicating geographical location). The dependent variables (one in each models) are the privacy indices scores for "general privacy", "soft technical" and "hard technical".

**Table 9: Factors explaining online privacy – results of OLS regression models**

| | General Privacy | | | Soft technical | | | Hard technical | | |
|---|---|---|---|---|---|---|---|---|---|
| | Beta | Std. Error | Sig. | Beta | Std. Error | Sig. | Beta | Std. Error | Sig. |
| (Constant) | 25.870 | 4.523 | 0.000 | 31.695 | 3.133 | 0.000 | 21.051 | 3.328 | 0.000 |
| Gender (male dummy) | 3.306 | 1.293 | 0.011 | 9.008 | 1.502 | 0.000 | 4.767 | 1.347 | 0.000 |
| Age | 0.137 | 0.038 | 0.000 | | | | -0.108 | 0.041 | 0.008 |
| Education level | 1.530 | 0.501 | 0.002 | 1.326 | 0.586 | 0.024 | | | |
| Big five - I get chores done right away | 1.322 | 0.604 | 0.029 | | | | | | |
| Big five - I like order | 1.603 | 0.706 | 0.023 | | | | | | |
| Using social networks | 1.377 | 0.544 | 0.012 | | | | 1.538 | 0.570 | 0.007 |
| Using my real name in social networks | -5.256 | 1.758 | 0.003 | -5.772 | 1.987 | 0.004 | -3.695 | 1.752 | 0.035 |
| Stating my personal/marital status in social networks | -3.996 | 1.358 | 0.003 | | | | | | |
| Posting family photos and clips in social networks | -2.473 | 1.525 | 0.105 | | | | | | |
| Indicating geographical location in social networks | -7.022 | 1.688 | 0.000 | -4.227 | 1.913 | 0.027 | -4.079 | 1.679 | 0.015 |
| R | 0.286 | | | 0.223 | | | 0.203 | | |
| R Square | 0.082 | | | 0.05 | | | 0.041 | | |
| N | 1109 | | | 1017 | | | 1017 | | |

As can be seen from the table, all variables (with the exception of "indicating geographical location in social networks- e.g. living address) are statistically significant at least at the 0.05 level. The dummy variable for male users was found to be positively and significantly correlated with all three privacy indices, implying higher perception of privacy and data security among the male population. Age was found to be positively and significantly correlated with general privacy and negatively and significantly correlated with hard privacy. This means that general privacy skills are high among older age cohorts, whereas younger age cohorts display high rates of hard technical skills. Education level was also found to be positively correlated both with general privacy and with soft technical skills. This means that there is linkage between higher education levels and enhanced

privacy/data security attributes and skills. Users of social networks (e.g. Facebook) display higher privacy attributes and skills than non-users. The model shows that the use of social networks is positively and significantly linked both with the general and the hard technical indices. Despite this, certain activities in social networks such as displaying the users' real name online and indicating their geographical location were found to be negatively and significantly associated with general privacy attributes and hard technical skills. This means that users that were engaged in these two activities in social media were less likely to perform specific actions which enhance their privacy and data security online. Additional social media activity - stating the user's personal status online was found to be negatively and significantly associated with the general privacy indicator. Finally, two interesting "Big Five" behavioral attributes pertaining to self-perception of order ("I get chores done right away", "I like order") were found to be positively and significantly correlated with general privacy attributes. These findings are in line with other studies (e.g. Osatuyi, 2015) who found that "Big Five" indicators relating to conscientiousness behaviour (the tendency to be orderly, logical, rational and competent and attentive to details) positively influence the concern for information privacy. A conscientious individual will sift through a variety of reputable information on privacy on social media sites before using one. He or she will be more informed and educated about risks associated with the use of personal information on online platforms (e.g. Osatuyi, 2015; Mccrae, and Costa, 1991).

# Chapter 5: Visualizing Survey Data– Lessons learned

A large variety of survey platforms support the development, design and technical creation of online questionnaires and the collection of complex data (e.g. 1KA system4). However, these tools are not as robust when it comes to visualizing the collected data. While some survey tools supply diverse visual solutions (e.g. SurveyMonkey and Qualtrics[5]), the best stories in the survey data remain hidden behind canned reports that are too difficult or expensive to customize. Presenting survey data in an appealing and efficient manner is challenging.

In this chapter, we discuss some lessons learned while developing a generic interactive visualization tool for our Binational Survey data in the context of online users' behavior using Tableau[6]. It is important to note however, that the lessons learned are not tool dependent, as we aim at shedding light on the visual design space of survey data and highlight some suggested design guidelines. **The tool (which is under construction) can be found here (Ctrl+Click):**

Following Munzner (2014), we first describe the data characteristics ("What"), then we indicate the relevant tasks ("Why") and finally we suggest some options for visual solutions ("How"), highlighting specific issues to be considered. Many ideas that are presented here are adopted from Steve Wexler's blogs.

## Survey data (what?)

Survey data in the Social Sciences usually include four different elements:

1. Respondents' socio-demographic data (e.g. gender, age, income level etc.).
2. Arrangement of responses in numeric and/or text format (survey observations).
3. Calculated variables derived from the observations (e.g. calculated indicators).
4. Meta data describing the survey data (e.g. questions types and wording, response status indicating if the survey was properly completed etc.).

### Question types

Surveys include the following types of questions (examples from the Binational Survey are shown in parentheses):

[4] https://www.1ka.si/d/en
[5] https://www.surveymonkey.com/ ; https://www.qualtrics.com/
[6] https://www.tableau.com/

1. Single-Punch questions (e.g. “yes” / “no” / “maybe” or “Specify the device from which you bought on the Internet – mostly from PC / mostly from Smartphone”),
2. Multi-Punch questions (e.g. “Which social networks do you use? Check all that apply – Facebook / Twitter / Instagram…”),
3. “Enter a value” (e.g. “What is your age?”; “What is your salary?”),
4. Likert-Scale questions (agreement, frequency, satisfaction, importance, e.g. “To what extent do you agree: User ratings and reviews of products and services can be trusted”).

Surveys often include ***open-ended questions*** which usually involve text visualization (e.g. sentiment analysis). In this chapter we do not discuss in detail this type of questions, however text visualization techniques can be found in the Text Visualization Browser.

Survey raw data is usually organized in a table or spreadsheet-like format (i.e. each respondent forms a single observation). The items of the table (i.e. respondents) are identified by a unique key (e.g. ID number or record number). The scalability of survey data largely varies. The number of respondents in typical surveys usually varies from hundreds to thousands, and the number of questions is likely to be several dozen.

### Data Preparation

The data preparation stage for survey visualization is a prerequisite task that requires smart and efficient organization of the data. Successful execution of this initial but crucial step will save time and effort and will allow the generation of meaningful outputs. Generally, we can speak of three levels of data preparation: meta data level, variable level, and table-level:

**Meta-data level**: Preparation of a number of well-organized tables: socio-demographic tables (codes and text), questions table (type, group, and wording/labeling) and answers table (values and labels) that could be interlinked by a common key.

**Variable-level**: Cleaning and harmonizing data variables (e.g. transforming blanks to zeros); Calculating derived attributes (e.g. aggregated indicators).

**Table-level**: Using the status attributes for removing invalid records (e.g. incomplete questions or questionnaires), joining numeric codes with textual labels using meta-data tables, and creating **pivot** tables based on the questions. The reshaped “long” format of the table (e.g. each row represents a single answer given by each respondent in contrast

to the “wide” format in which each row represents a single respondent) facilitates data coding into **visual** channels. Keeping the respondent’s characteristics in each row is advisable, as it simplifies the visualization process and contributes to better user orientation. Each row of the new “long” table includes two parts. The first element contains the respondent ID with the related independent demographic variables. The second element includes the dependent variables - the question’s identifier (including the question type, group and wording) and the related response both in numerical (value) and textual (label) formats.

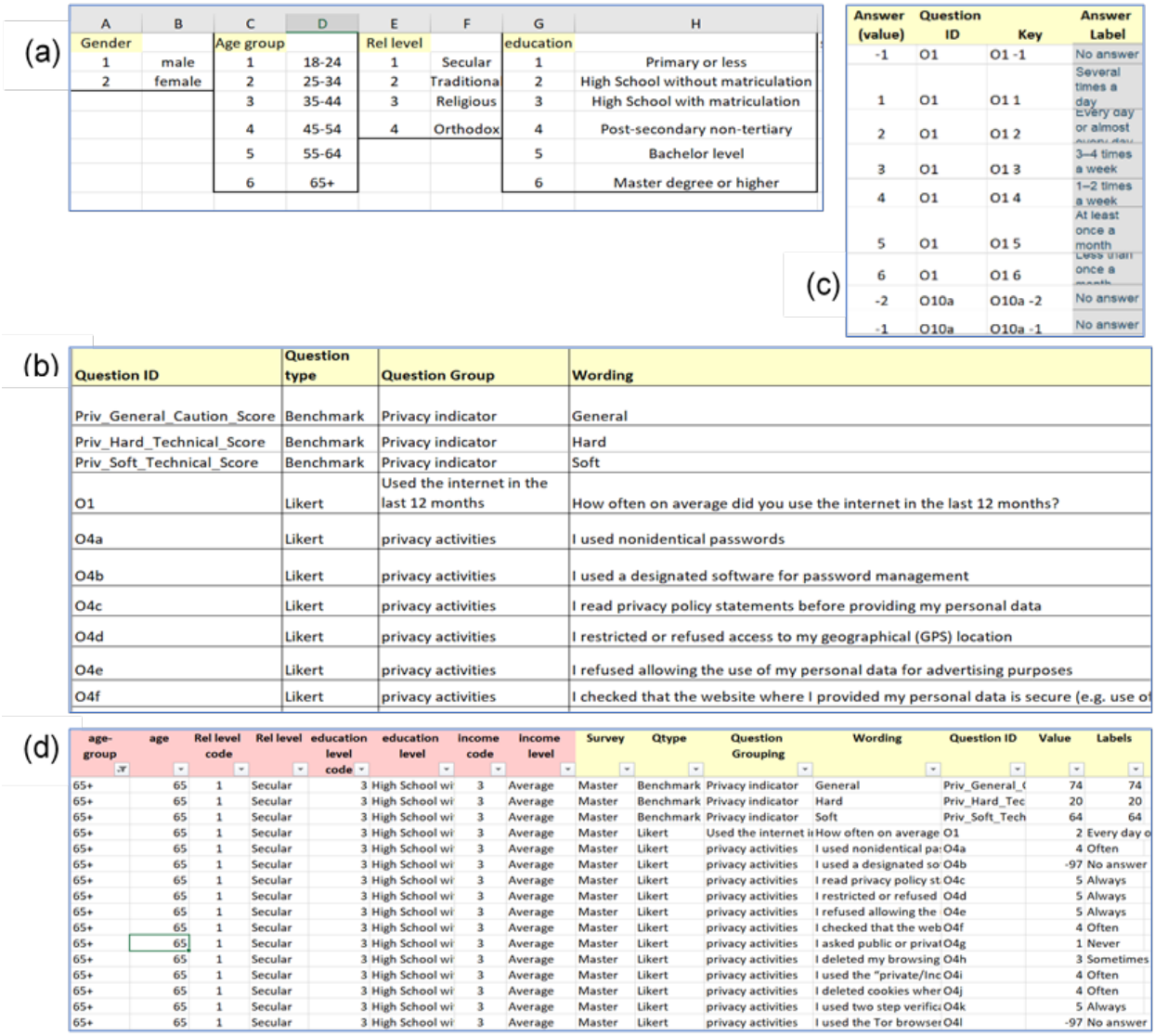
Figure 41 shows examples of demographic tables (a), question table (b), answers table (c), and most important – the reshaped survey data table (d).

### Survey tasks (why?)

The users’ tasks (“why” – the user’s justifications for using the visualization tool), are an equally important constraint for visualization designers, almost as the type of data that they possess (Munzner, 2014). The following bullets present some of the general tasks to consider for visual survey analysis:

- Who are the respondents (research population)? What is the focus or theme of the survey?
- How do the subjects respond to a single question? (Within-question comparison)
- How do the subjects respond to a group of questions? (Between-question comparison)
- Which socio-demographic factors best explain differences in the respondents’ attributes (e.g. online behavior)? This task can be interpreted by asking what are the differences between the different socio-demographic groups (e.g. male vs. female) or by asking how did a selected group responded compared to the unselected group (e.g. age group 18-24 vs. all others).
- Are there any significant relationships (e.g. correlations) between the variables?
- Is the *sample size* (n) big enough for statistical inference?

**Figure 41: Data preparation stage**



(a)

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| Gender | | Age group | | Rel level | | education | |
| 1 | male | 1 | 18-24 | 1 | Secular | 1 | Primary or less |
| 2 | female | 2 | 25-34 | 2 | Traditiona | 2 | High School without matriculation |
| | | 3 | 35-44 | 3 | Religious | 3 | High School with matriculation |
| | | 4 | 45-54 | 4 | Orthodox | 4 | Post-secondary non-tertiary |
| | | 5 | 55-64 | | | 5 | Bachelor level |
| | | 6 | 65+ | | | 6 | Master degree or higher |

(c)

| Answer (value) | Question ID | Key | Answer Label |
|---|---|---|---|
| -1 | O1 | O1 -1 | No answer |
| 1 | O1 | O1 1 | Several times a day |
| 2 | O1 | O1 2 | Every day or almost |
| 3 | O1 | O1 3 | 3–4 times a week |
| 4 | O1 | O1 4 | 1–2 times a week |
| 5 | O1 | O1 5 | At least once a month |
| 6 | O1 | O1 6 | Less than once a |
| -2 | O10a | O10a -2 | No answer |
| -1 | O10a | O10a -1 | No answer |

(b)

| Question ID | Question type | Question Group | Wording |
|---|---|---|---|
| Priv_General_Caution_Score | Benchmark | Privacy indicator | General |
| Priv_Hard_Technical_Score | Benchmark | Privacy indicator | Hard |
| Priv_Soft_Technical_Score | Benchmark | Privacy indicator | Soft |
| O1 | Likert | Used the internet in the last 12 months | How often on average did you use the internet in the last 12 months? |
| O4a | Likert | privacy activities | I used nonidentical passwords |
| O4b | Likert | privacy activities | I used a designated software for password management |
| O4c | Likert | privacy activities | I read privacy policy statements before providing my personal data |
| O4d | Likert | privacy activities | I restricted or refused access to my geographical (GPS) location |
| O4e | Likert | privacy activities | I refused allowing the use of my personal data for advertising purposes |
| O4f | Likert | privacy activities | I checked that the website where I provided my personal data is secure (e.g. use of |

(d)

| age-group | age | Rel level code | Rel level | education level code | education level | income code | income level | Survey | Qtype | Question Grouping | Wording | Question ID | Value | Labels |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Benchmark | Privacy indicator | General | Priv_General_( | 74 | 74 |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Benchmark | Privacy indicator | Hard | Priv_Hard_Tec | 20 | 20 |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Benchmark | Privacy indicator | Soft | Priv_Soft_Tech | 64 | 64 |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | Used the internet i | How often on average | O1 | 2 | Every day o |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I used nonidentical pa | O4a | 4 | Often |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I used a designated so | O4b | -97 | No answer |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I read privacy policy st | O4c | 5 | Always |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I restricted or refused | O4d | 5 | Always |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I refused allowing the | O4e | 5 | Always |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I checked that the web | O4f | 4 | Often |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I asked public or privat | O4g | 1 | Never |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I deleted my browsing | O4h | 3 | Sometimes |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I used the "private/Inc | O4i | 4 | Often |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I deleted cookies wher | O4j | 4 | Often |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I used two step verific | O4k | 5 | Always |
| 65+ | 65 | 1 | Secular | 3 | High School wi | 3 | Average | Master | Likert | privacy activities | I used the Tor browser | O4l | -97 | No answer |

Demographic tables (a), Question Table (b), Answers Table (c), Reshaped pivoted survey data (d).

Source: screen shots taken from the Binational Survey data excel files.

## Survey visualization (how?)

Following the introduction of the survey's data and tasks, we now suggest some relevant design guidelines for survey visualization. These guidelines are relevant for the display possibilities of the various chart types as well as for demonstrating interactions.

In this section we describe six suggested steps for developing survey data visualization, assuming that the data was prepared as suggested above. Screen shots from the Binational Survey visualization tool are used for illustrating and describing each step.

### Mapping the questions and responses (What is the survey about?)

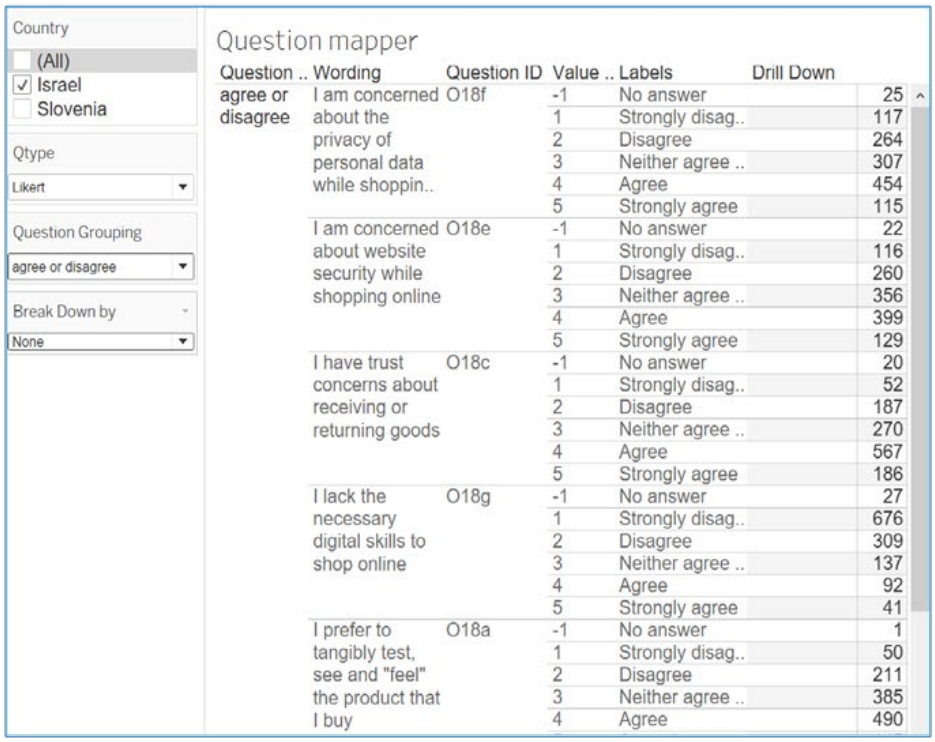
The formulation of a “single view table” that simultaneously maps and summarizes all question types, grouping of variables, IDs, wording and responses (values and text) could greatly contribute for better data management. The addition of interactive filters (e.g. by question type, demography, question grouping) to this “at-a-glance” table could also be beneficial for more efficient orientation as well as for tracking and fixing coding errors. Displaying the number of responses could also enrich this inventory. Figure 42 shows an example for a “Question Mapper” based on the Binational Survey.

### Visualizing a demographics dashboard (Who are the respondents?)

Figure 43 presents the demographics dashboard for the Binational Survey. The values are presented by the number of cases (n) and by the percentage of cases. The total n, which is dynamic, is displayed on the top-right corner. The maximum scale value for this particular example uses the un-fixed-scale option, but could be altered to represent a user-defined option (e.g. 0-100 scale). The tool can also facilitate a cross-variable breakdown (e.g. Figure 43-b the socio-demographic characteristics of Arab females), as each item can be used as a filter.

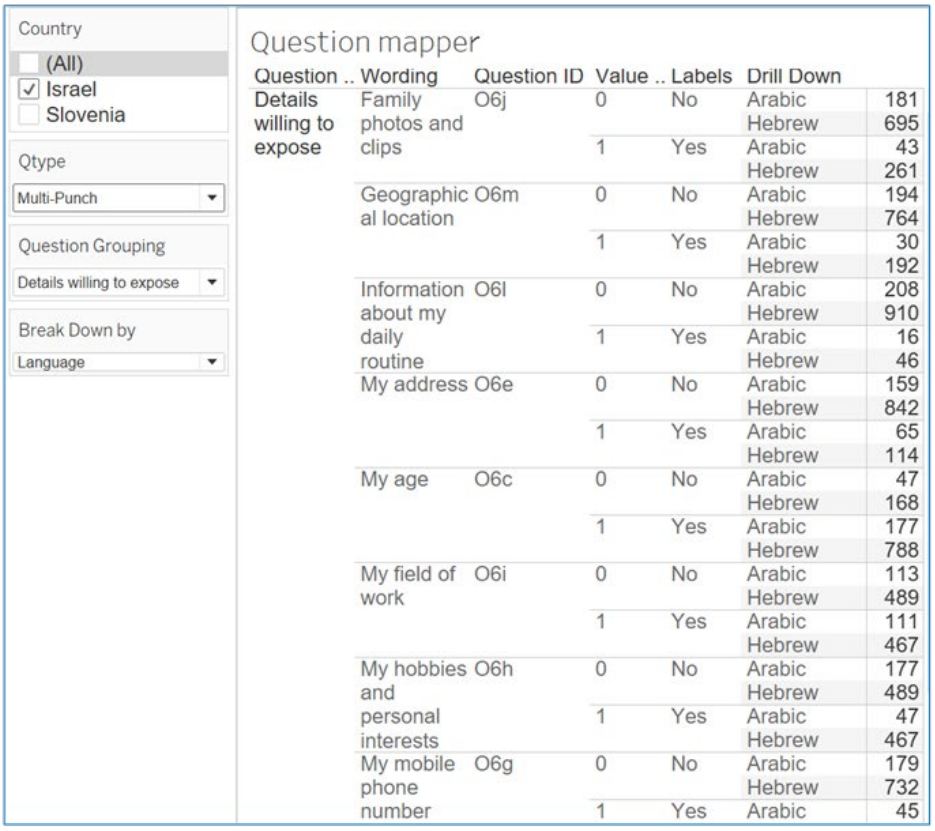
## Figure 42: Question mapper – questions and responses inventory

(a)




| Question .. | Wording | Question ID | Value .. | Labels | Drill Down | |
|---|---|---|---|---|---|---|
| agree or disagree | I am concerned about the privacy of personal data while shoppin.. | O18f | -1 | No answer | | 25 |
| | | | 1 | Strongly disag.. | | 117 |
| | | | 2 | Disagree | | 264 |
| | | | 3 | Neither agree .. | | 307 |
| | | | 4 | Agree | | 454 |
| | | | 5 | Strongly agree | | 115 |
| | I am concerned about website security while shopping online | O18e | -1 | No answer | | 22 |
| | | | 1 | Strongly disag.. | | 116 |
| | | | 2 | Disagree | | 260 |
| | | | 3 | Neither agree .. | | 356 |
| | | | 4 | Agree | | 399 |
| | | | 5 | Strongly agree | | 129 |
| | I have trust concerns about receiving or returning goods | O18c | -1 | No answer | | 20 |
| | | | 1 | Strongly disag.. | | 52 |
| | | | 2 | Disagree | | 187 |
| | | | 3 | Neither agree .. | | 270 |
| | | | 4 | Agree | | 567 |
| | | | 5 | Strongly agree | | 186 |
| | I lack the necessary digital skills to shop online | O18g | -1 | No answer | | 27 |
| | | | 1 | Strongly disag.. | | 676 |
| | | | 2 | Disagree | | 309 |
| | | | 3 | Neither agree .. | | 137 |
| | | | 4 | Agree | | 92 |
| | | | 5 | Strongly agree | | 41 |
| | I prefer to tangibly test, see and "feel" the product that I buy | O18a | -1 | No answer | | 1 |
| | | | 1 | Strongly disag.. | | 50 |
| | | | 2 | Disagree | | 211 |
| | | | 3 | Neither agree .. | | 385 |
| | | | 4 | Agree | | 490 |

(b)

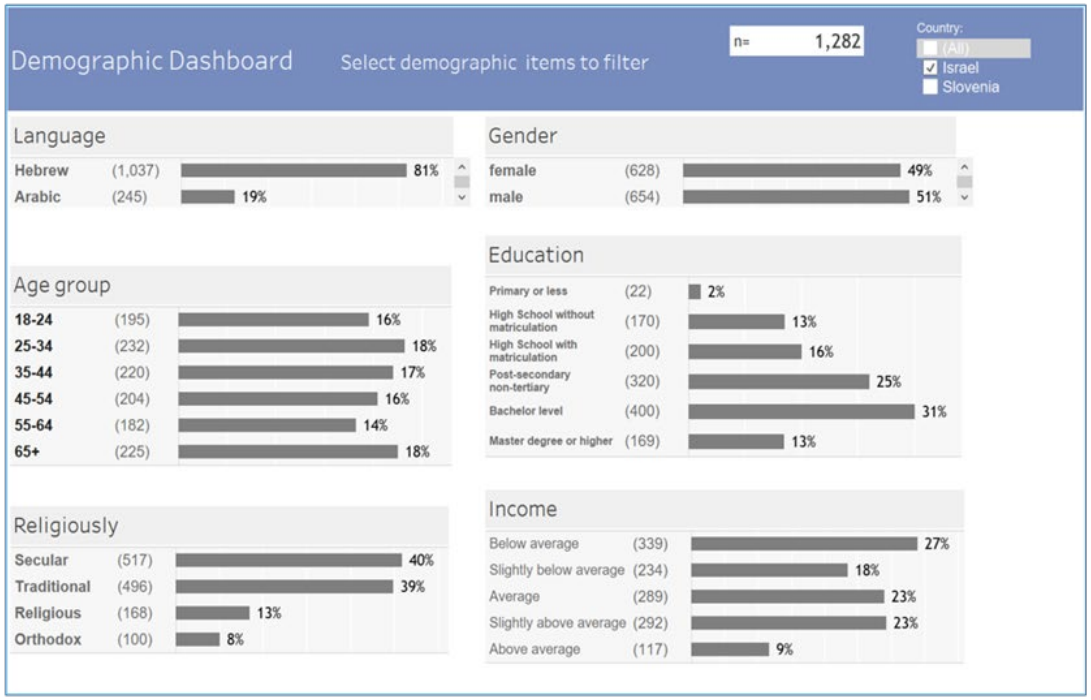



| Question .. | Wording | Question ID | Value .. | Labels | Drill Down | |
|---|---|---|---|---|---|---|
| Details willing to expose | Family photos and clips | O6j | 0 | No | Arabic | 181 |
| | | | | | Hebrew | 695 |
| | | | 1 | Yes | Arabic | 43 |
| | | | | | Hebrew | 261 |
| | Geographical location | O6m | 0 | No | Arabic | 194 |
| | | | | | Hebrew | 764 |
| | | | 1 | Yes | Arabic | 30 |
| | | | | | Hebrew | 192 |
| | Information about my daily routine | O6l | 0 | No | Arabic | 208 |
| | | | | | Hebrew | 910 |
| | | | 1 | Yes | Arabic | 16 |
| | | | | | Hebrew | 46 |
| | My address | O6e | 0 | No | Arabic | 159 |
| | | | | | Hebrew | 842 |
| | | | 1 | Yes | Arabic | 65 |
| | | | | | Hebrew | 114 |
| | My age | O6c | 0 | No | Arabic | 47 |
| | | | | | Hebrew | 168 |
| | | | 1 | Yes | Arabic | 177 |
| | | | | | Hebrew | 788 |
| | My field of work | O6i | 0 | No | Arabic | 113 |
| | | | | | Hebrew | 489 |
| | | | 1 | Yes | Arabic | 111 |
| | | | | | Hebrew | 467 |
| | My hobbies and personal interests | O6h | 0 | No | Arabic | 177 |
| | | | | | Hebrew | 489 |
| | | | 1 | Yes | Arabic | 47 |
| | | | | | Hebrew | 467 |
| | My mobile phone number | O6g | 0 | No | Arabic | 179 |
| | | | | | Hebrew | 732 |
| | | | 1 | Yes | Arabic | 45 |

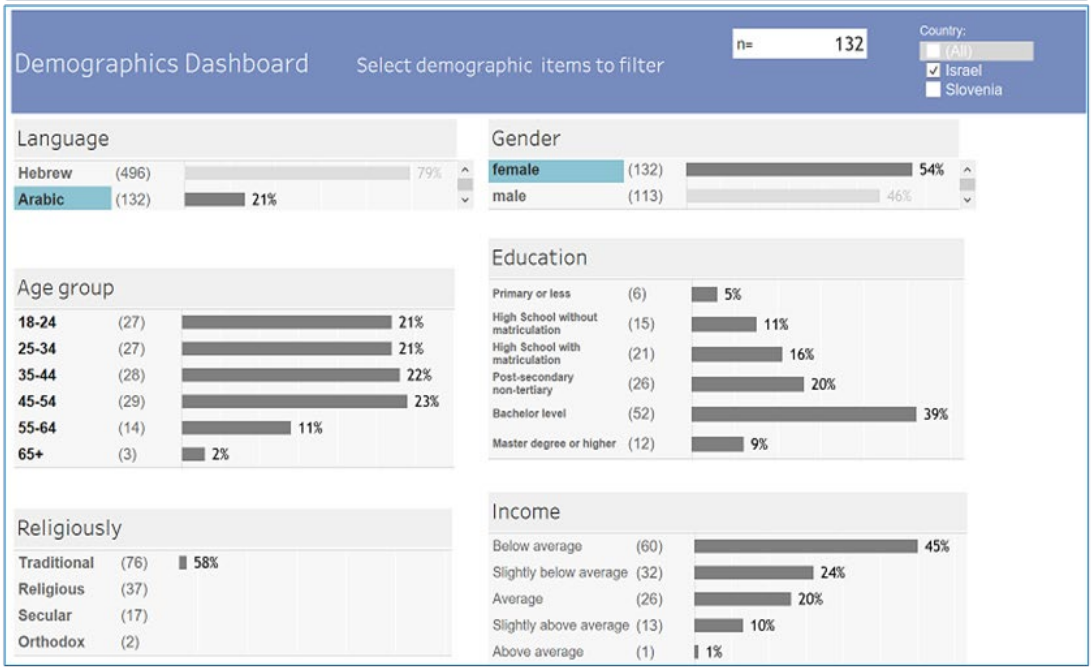
Source: screen shots taken from the Binational Survey visualization views.

**Figure 43: Demographics dashboard – who are the respondents?**

(a)




(b)




The view can be filtered by any item (single item (a) or multiple items (b))
Source: screen shots taken from the Binational Survey visualization views.

## Visualization of single-punch questions

Visualizing response distribution of a single punch question (e.g. “Yes” / “No” / “Maybe”) seems trivial, as horizontal bar charts are perfect for comparisons. However, when there

is a need to show a **group** of questions (i.e. a multi-item situation as shown in Figure 44), there are some issues to consider:

- Synchronization of horizontal axis of all items (surprisingly it is not always the situation).
- Exclusion of irrelevant labels (for example the “Maybe” response) could reduce clutter.
- Displaying both numbers and percentage of respondents could aid in reducing uncertainty and increasing validity.
- In a multi-item situation (e.g. a group of single-punch questions) the focus of the comparison can be alternated between the question item and the answer, meaning that the breakdown order can include the question item first and answer second, or vice versa. Figure 44 presents the response distribution of both options:

**Figure 44: Visualization of a multi-item single-punch questions**


Side-by-side bars (Question break (a) vs. Answer break (b)) or stacked bars (c).
Source: screen shots taken from the Binational Survey visualization views.

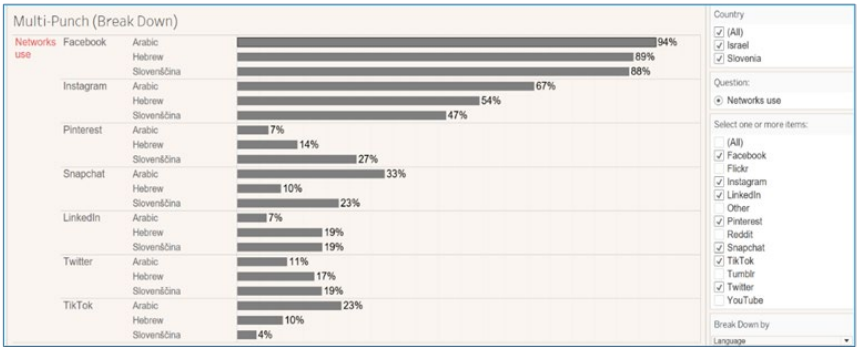
In the shown example (“Device used when placing the order – mostly for PC / mostly from smartphone”), the question-break option seems to be more relevant as it highlights the effect of the product price on the device used for its purchase. A toggle interaction between the two views facilitate user’s control. When the single-punch answers are “Yes” / “No” only, the illustration of the “No” segment might be redundant. However, when there are two (relevant) sides for the coin, a stacked-bar type chart could be useful for illustration purposes, as shown in Figure 44c. As can be seen from the figure, both views which in fact represent two sides of the same “coin” are well demonstrated: smartphone use diminishes as the product price increases (orange) and PC use increases as the product price rises (blue).

### Visualization of multi-punch questions

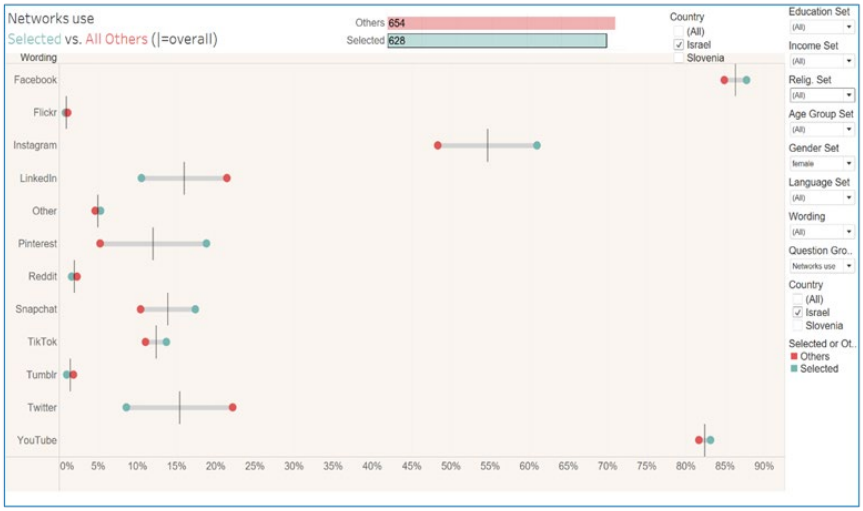
Figure 45a presents an example of side-by-side bars illustrating a multi-punch question, parsed by a single demographic variable.

**Figure 45: Visualization of multi-punch questions**

(a)




(b)




Source: screen shots taken from the Binational Survey visualization views.

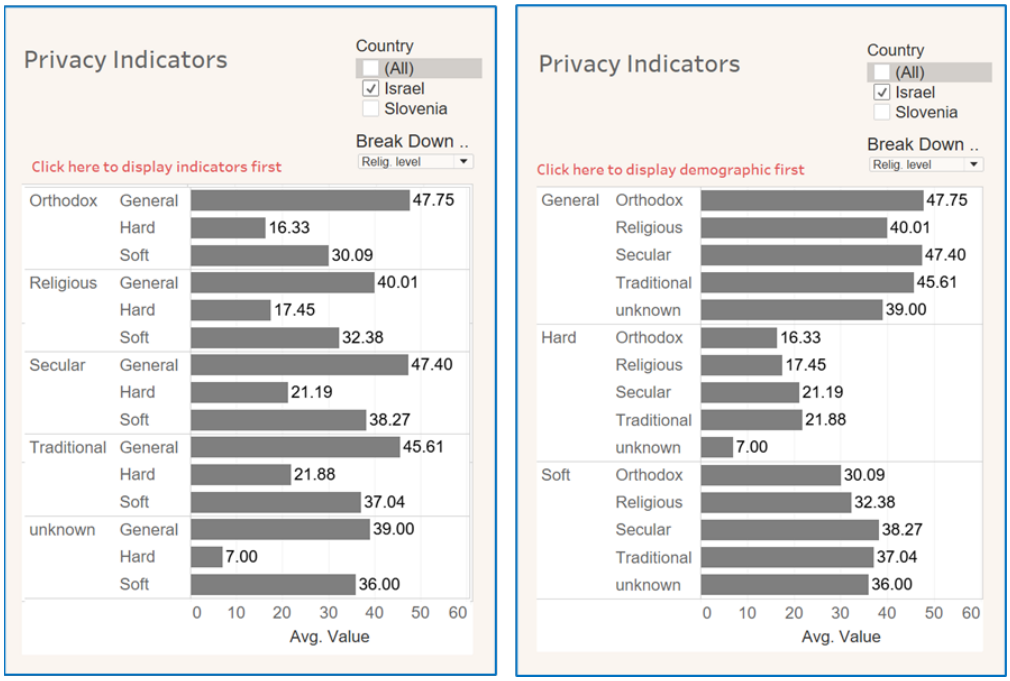
side-by-side bars (a) and gap chart (b)

This visualization view includes an item filter that enables the breakdown of the multi-punch questions by selected socio-demographic variables. Figure 45b shows an example of a gap chart (also called a dumbbell chart or a connected dot plot) illustrating a comparison of social networks use among female users (selected-blue dot) as compared to male users (others-red dot) and all users (overall-black line). This technique can be applied to single-punch questions visualization as well. This particular visualization is used when there is a need to show how a selected group responded to a question as compared to the group which was not selected and to the general population. Displaying a bar chart that illustrates the number of selected respondents versus the number of unselected respondents (as can be seen above the gap chart in Figure 45b) can illustrate the size of the various groups or sub-samples.

### Visualization of quantitative variables

In the visualization of quantitative variables, the variable values can be either flatly displayed using the original distribution, or they can be aggregated to form new values or indices. Figure 46 shows an example of side-by-side bars illustrating three calculated privacy indicators (general, hard, and soft indicators). The user can toggle between the views by clicking on the “Click here to display indicators / demographic first.”

**Figure 46: Visualization of quantitative variables**



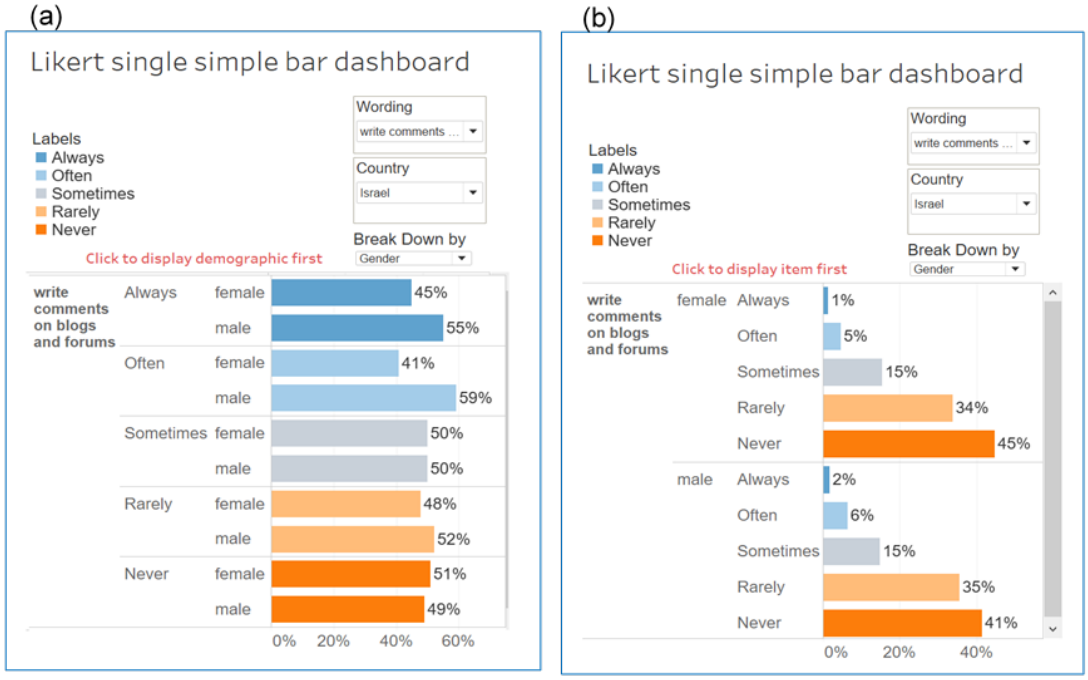

Side-by-side bars with an interaction enabling controlling the break down order

Source: screen shots taken from the Binational Survey visualization views.

### Visualization of Likert-scale questions

Originally introduced by psychologist Rensis Likert in 1932 (Likert, 1932), the Likert scale has become the most widely used psychometric response scale. The scale consists of several statements, or Likert items, in which respondents specify their level of agreement to a particular statement comprised of several ordered response alternatives. The trivial chart type option for Likert scale type questions is the bar chart (Figure 47), which places the count or percentage of cases in one axis (usually the vertical) and the ordered response categories on the other axis (usually the horizontal). The figure below illustrates user-defined visualization, enabling to control various interactions such as the breakdown of Likert scale questions by socio-demographic variables, toggling between items' views and the order of the various categories (e.g. frequency of writing comments on blogs and forums by gender vs gender by writing comments on blogs and forums by gender).

**Figure 47: Likert Scale simple bar chart**

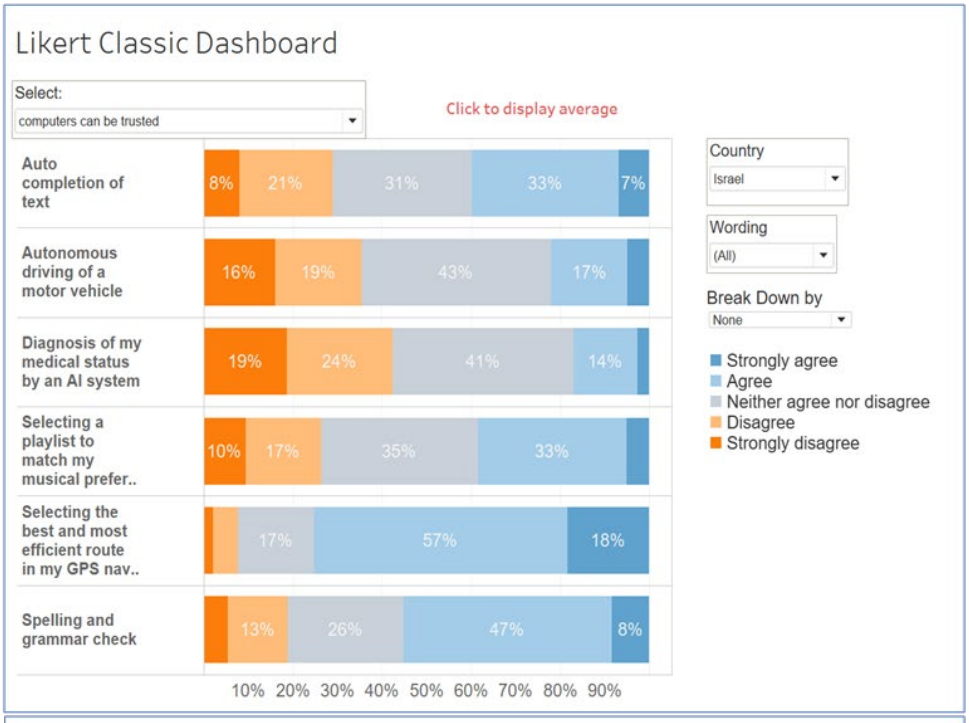



Toggle views interaction (in red) enables display of items first (a) or with demographic first (b)
Source: screen shots taken from the Binational Survey visualization views.

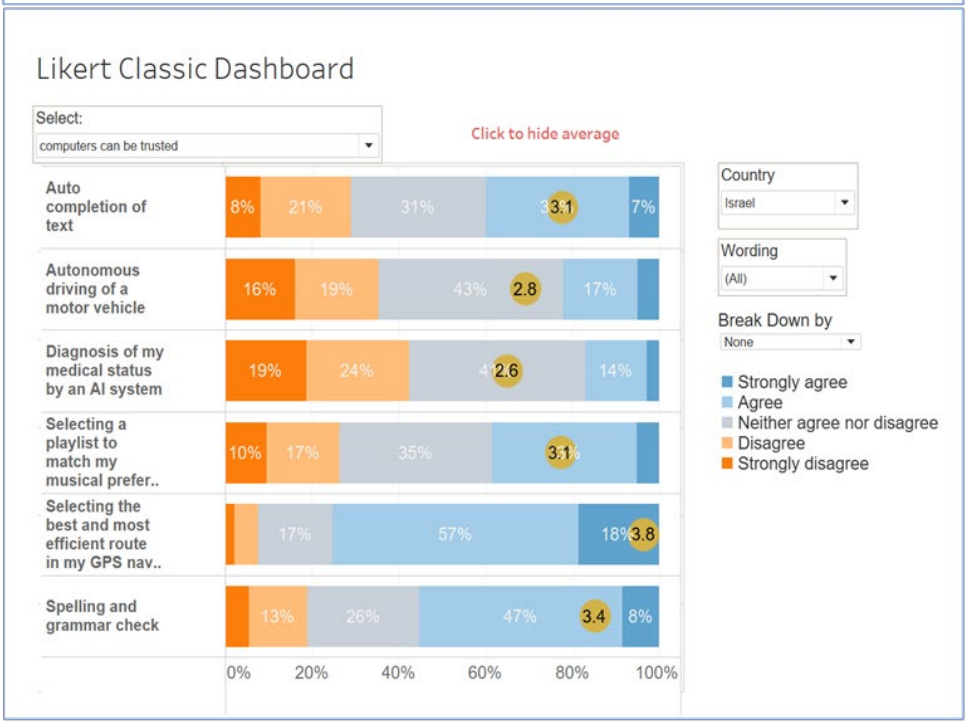
Stacked bar charts are commonly used for displaying multiple statements simultaneously (Figure 48a). These charts show the response distribution by subdividing a single bar into several response categories (represented by different colors). Displaying the mean response values can be used for highlighting responses differences (Figure 48b).

**Figure 48: Likert stacked bar chart**

(a)




(b)



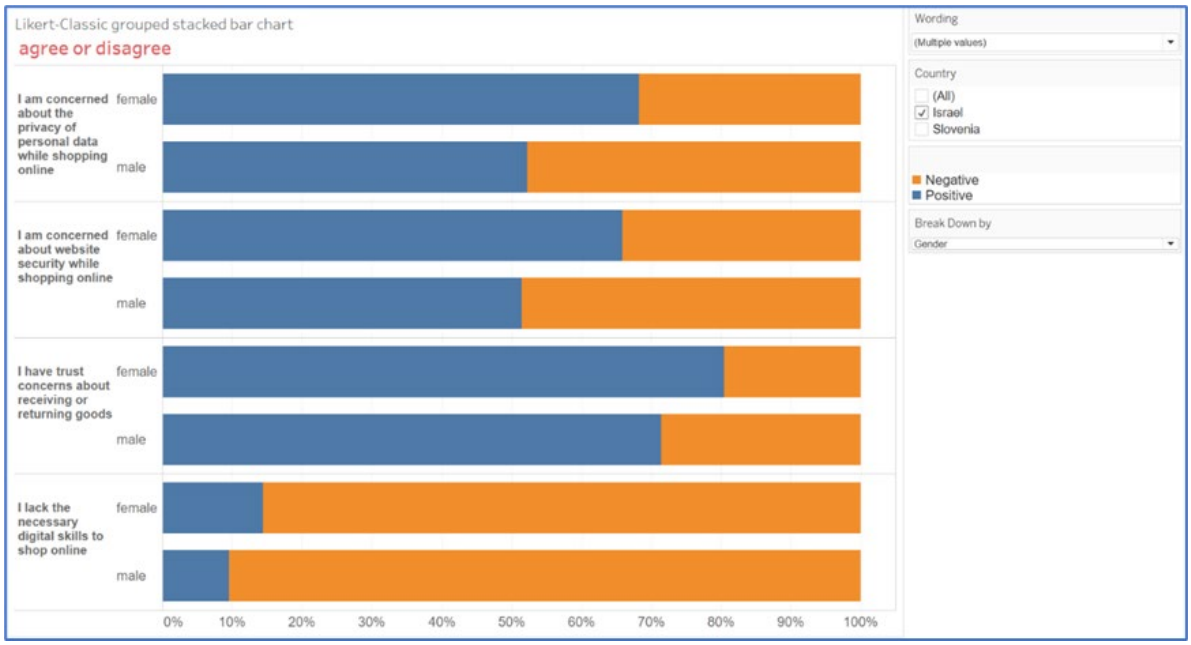

Source: Screen shots taken from the Binational Survey visualization views.

Toggle views interaction (in red) enables display without values average (a) or with values average (b)

Figure 49 shows another variation of the Likert Classic divergent stacked bar chart, where labels are grouped into three groups: “positive”, “negative” and “neutral” (which were

excluded in the example shown). This option is relevant when there is a need to highlight the contrast between two options or choices, whereas the precise level of agreement or disagreement is less important.

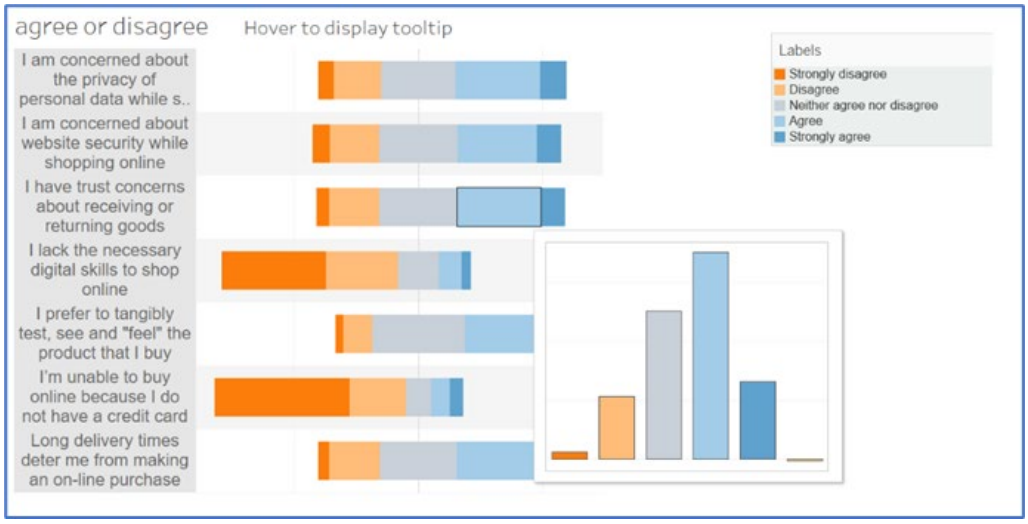
**Figure 49: A Likert grouped stacked bar chart**




Source: screen shots taken from the Binational Survey visualization views.

Figure 50 presents a variation for a stacked bar chart. This figure, which is known as **centered stacked bar chart** works in a similar way to a typical stacked bar, with the exception of using a central line distinguishing positive from negative responses, allowing for the skew between them to be seen more easily (Petrillo et al., 2011). The neutral distribution or values are shown in the center but can be also excluded. By hovering over an item, a tooltip is enabled, creating a simple bar chart. This chart displays the count or percentage of cases in one axis (usually the vertical) and the ordered response categories on the other axis (the horizontal). This presentation view can be especially beneficial for within-item comparison tasks.

**Figure 50: A Likert centered divergent stacked bar chart**




A central line is dividing positive from negative responses. A tooltip showing a simple common-scale bar chart can facilitate the comparison task

Source: screen shots taken from the Binational Survey visualization views.

To summarize, the following steps or guidelines should be considered for survey data visualization:

- Data preparation: create and arrange meta-data tables. Correct, harmonize and clean the data. Reshape the response data table to a “long” structure format including both values and text formats.
- Formulate a question mapper.
- Create an interactive demographics dashboard (showing both the number and percentage of cases).
- Create an interactive dashboard for each question type (i.e. single / multi / Likert / continuous values):
  - Choose the suitable chart types (e.g. bar chart, stacked bar chart, gap chart etc.) for each question type.
  - Enable relevant interactions to empower the user. The following interactions should to be considered: selection of demographic variables, selection of breakdown order, sort order, toggling between displays (e.g. show/hide mean values for Likert scale questions; group/ungroup Likert-scale values; show/hide neutrals etc.)
  - Include both the number and percentage of cases to reduce uncertainty and to increase validity.

In the future we plan to develop visual solutions for other relevant survey issues as cross-question analysis (e.g. finding correlations between question) and illustrating uncertainty issues.

# Chapter 6: Summary, Conclusions and Recommendations for Policy Makers

In this research, an innovative approach for profiling, studying and analyzing the socio-economic and personal trait characteristics of online behavior was applied using unobtrusive (digital trace analysis and social media analysis) and obtrusive methods (online surveys). The research employed a wide range of qualitative and quantitative research methods and tools including descriptive statistics and inferential statistic, in order to describe, characterize, explain and predict online user behavior. An important output of the research was the development of a generic interactive visualization tool for displaying and analyzing survey data.

A triangulation-based approach was used to evaluate and analyze differences in online user-behavior relating to various activities such as e-shopping, e-travel, e-finance, the use of social networks, search activity and the perception of privacy and personal data security. The fusion of survey data, digital trace data and social media data has enabled to deepen the understanding of investigated phenomenon and to construct more robust measurements. The triangulation methodology was further demonstrated by a case-study that investigated and analyzed data security and privacy aspects of online users using survey data, digital trace data and social media discourse data.

The findings of the research pointed out to differences in online behavior, as well as digital gaps, with respect to various online activities and the content consumed.

**Online shopping:**

- Higher rate of online shopping was found to be corelated with gender (male), higher education and higher income levels. Consumer related factors such the cost of the product (low) was found to exert significant and positive impact on frequent online shopping, whereas the “need to physically feel or test the product” was found to exert significant and negative impact.
- Personal or behavioral attributes of the online user were found to exert a significant impact on frequent online shopping. Impulsive (making unnecessary purchases frequently), active (submitting reviews for products frequently) and passive/lurking behavior (reading reviews for products frequently without participation) were found to be significantly and positively correlated with frequent online shopping.

- Other behavioral attributes such as the “lack of digital skills” and “having privacy concerns with regards to the leak of personal data when browsing” were found to be negatively associated with frequent shopping. Individuals who conveyed strong concerns for their privacy and fear for the leak of their personal information were 34% less likely to be frequent online shoppers than individuals who had no privacy or data security concerns.
- Both survey data and digital trace data revealed that special shopping days such “Black Friday” exert a strong influence on the propensity of users to conduct shopping online.
- Strong correlation was found between the type of device used in online purchases and the price of the good or service. Smartphone share use significantly diminishes as the cost of the ordered good or service rises. Similarly, there was a much higher propensity to use PCs over smartphones when making either high risk, rare or expensive orders.

**Online travel:**

- Online travel bookings (e.g. airline tickets, hotels) are much more prevalent among younger age groups, among individuals holding higher education degrees and among the secular population. A large gap in booking preferences was observed with respect to ethnic background, showing much more frequent use of online platforms among the Jewish population as compared to the Arab population.
- The ability to conduct a comprehensive search is the leading factor in the decision to book online, followed by the ability to compare costs, the ability to tailor a flexible flight that suits the traveler’s needs and the ability to receive more information about the flight.
- The leading factor for choosing a travel agent for booking travel accommodations (over online reservations) is the need to interact with a person who will answer questions and solve problems, followed by online privacy and data security concerns in online bookings and low digital skills of the user.
- User reviews and user rating on websites such as booking.com, trivago, Airbnb, TripAdvisor were found to exert strong influence, especially on younger age cohorts, on the decision to book travel accommodations.

**Online banking and e-finance**

- The share of online financial activities conducted by male users is higher than its comparable share among female users in almost all types of financial transactions (e.g. payments of bills, viewing details of provident funds and pensions, ordering credit cards, buying and selling stocks and bonds etc.).
- There is a clear linkage between the education level of the user and the scope of online financial transactions. This share of online use increases as the education level of the online user rises.

**E-health**

- The most frequent digital health activities conducted by Israeli online users are making appointments to a family doctor, followed by viewing laboratory tests and searching for doctors.
- Women were found to exercise higher online presence in all of the surveyed digital health activities (e.g. making appointments, viewing online medical records, requesting laboratory tests, asking for the renewal of prescription drugs etc.). Similar trend with respect to gender was observed from digital trace data, where women account for the majority of the traffic in the various sick-fund (Kupot-Holim) websites. Substantial gender gaps were also observed between online female users and online male users with respect to the search of health-related information, with female users exercising higher search activity.
- The research findings reveal stark and consistent gaps in the use of online health services and in the search behavior of health-related information between Jewish and Arab online users, with Arab users displaying much lower use of online digital platforms.

**Online privacy and data security**

- The most frequent precaution that users exercise in protecting and maintaining their privacy online is “refusing to allow the use of their personal data for advertising purposes”, followed by "using nonidentical passwords to login to various apps and web services” and “restricting or refusing access to their geographical (GPS) location”.
- The least frequent precaution in the protection of privacy or data security is the use of a designated software for password management and the utilization of online tools such as VPN and the TOR Browser.

- Both survey data and digital trace data show substantially higher signals of VPN use and TOR Browser use among male users and younger age cohorts.
- Factor analysis was employed on a set of 13 privacy questions from the online survey. This exercise has resulted in the identification of three factors or underlying variables for online privacy and data security which were labeled as ***"General Privacy"*** (reading privacy statements and being aware of the use of personal information by third parties; restricting access to personal data), "**Soft Technical Privacy"** (carrying out simple, routine measures to maintain/secure user anonymity & privacy online, e.g. deleting cookies and browsing history) and "**Hard Technical Privacy"** (using complex and designated tools, technologies and software in order to protect privacy, data security and anonymity online, e.g. VPN, TOR).
    - Gender was found to be positively and significantly correlated with all three privacy indices, implying higher perception of privacy and data security among the male population. Similar trend can be observed from the analysis of digital trace data which showed higher signals for hard technical skills among male users.
    - General privacy skills were found to be high among older age cohorts, whereas younger age cohorts display high rates of hard technical skills. Similar trend was observed from the analysis of digital trace data which showed higher signals for hard technical skills among younger online users.
    - Education level was found to be positively correlated both with general privacy and with soft technical skills.
    - The use of social networks was found to be positively and significantly correlated with the general and hard technical indices.
    - Two "Big Five" behavioral attributes pertaining to self-perception of order were found to be positively and significantly correlated with general privacy attributes.
- The discourse surrounding the concept of "online privacy" was focused on three sub-categories: Teenagers' (lack of) awareness to online privacy, Voyeurism & disrespect for privacy, and corporations' use of personal data. The discourse was mostly negative in its nature and included expressions of concerns about privacy and moral judgement of those who are blamed for breaching it.
- The discourse around hard technical aspects of online privacy (discussions which were related to the terms "Incognito browsing" and "Tor Browser") was most

prominent among teenagers' forums and religious Jewish communities forums, and its purpose was to provide users with tools to protect their data and receive better "deals" for flights and shopping.

- The content analysis of public social media shows that while the discourse surrounding the terms "online privacy" focuses on societal concerns and moral judgement, the discourse surrounding the terms "browsing history", "Tor Browser" and "Incognito browsing" ("hard privacy") is of technical/instrumental nature.
- The triangulation method facilitated the understanding and the perception of online privacy from different perspectives, which complement each other. While survey data allowed quantitative analysis of online privacy which could be parsed by socio-demographic and behavioral factors, the social data analysis enabled to investigate the context in which online privacy terms are used in public discourse, as well as the audiences who are involved in the discussions.

**Visualization of survey data**

The development of generic interactive visualization tool for displaying and analyzing survey data has highlighted the importance of following sequential steps or guidelines in facilitating the understanding of data stories and allowing to create an efficient framework for comparing and benchmarking survey data with other types of data (e.g. digital traces). These sequential steps include: data preparation, the formulation of question mapper, creation of interactive demographics dashboard, proper chart selection and enabling relevant interactions to empower the user.

## Recommendation for policy makers

The findings of this study can provide Israeli government ministries, business entities and the research community with insights that may contribute to the formulation of public policy in the fields of digital divide, online privacy, improvement of Internet user interfaces, as well as with methodological and procedural lessons that can be utilized for advanced research in the field of data integration.

We recommend stake holders and decision makers from the **public sector** to be active in the formulation of protocols aimed at defining and regulating the use of digital trace data**.** Such protocols should set clear guidelines for data collection and data mining from online sources; The anonymization of personal information on behalf of the data owner; Accepted practice and procedures for data processing, cross referencing and consolidation of digital

trace data and survey data from multiple sources; Guidance regarding the presentation of the data (on behalf of the researcher); The construction and maintenance of digital trace repositories (with or through entities such as the National Library or the Israel State Archives-ISA); Third party use; and the penalties that might be imposed on the researcher in case of breaching the contract terms.

**Our recommendations to government and public policy makers are:**

• Raise awareness about the consequences of **impulsive and addictive shopping behavior**.

• Raise awareness and enhance education, especially among women, of the importance of acquiring knowledge in the field of **e-banking and online financial transactions**.

• Raise awareness, especially among men and the Arab population regarding the **benefits and importance of online health services.**

• Raise awareness, especially among teenagers, regarding the issue of **online privacy** (e.g. reading privacy statements and being aware of the use of personal information by third parties; restricting access to personal data). In addition, raise awareness, especially among women as to the importance and advantages of using designated tools, technologies and software in order to protect privacy, data security and anonymity online.

**Our recommendations to the business sector are:**

• Raise awareness, especially among the religious and ultra-Orthodox populations, adults and Arab speakers regarding the benefits of using **online travel and tourism services**.

• Improve the friendliness of websites and applications especially in purchasing transactions interfaces on all types of devices (mobile phones, tablets and desktops of all types).

**Our recommendations to the research community are:**

• Promote and develop data triangulation methodologies and tools for the purpose of enhancing data reliability and understanding online behavior.

• Develop and improve existing methodologies for consolidating online surveys with digital traces for the purpose of deepening understanding of hidden and visible online behavior of users. This could be achieved through be the development of **visual components** as an integral and built-in part of survey platforms.

# List of References


Ajzen, I. (1991). The theory of planned behavior. Organizational behavior and human decision processes, 50(2), 179-211.

Alvarez-Galvez, J., Salinas-Perez, J. A., Montagni, I., & Salvador-Carulla, L. (2020). The persistence of digital divides in the use of health information: a comparative study in 28 European countries. *International Journal of Public Health*, 65(3), 325-333.

Amaro, S., & Duarte, P. (2013). Online travel purchasing: A literature review. *Journal of Travel & Tourism Marketing*, 30(8), 755-785.

Belk, R., & Kozinetz, R. (2017). Videography and netnography. In *Formative Research in Social Marketing* (pp. 265-279). Springer, Singapore.

Benckendorff, P. J., Xiang, Z., & Sheldon, P. J. (2019). *Tourism information technology*. Cabi.

Beyer, M. A., & Laney, D. (2012). The importance of "big data": A definition. Gartner. *G00235055*.

Bonfadelli, H. (2002). The Internet and knowledge gaps: A theoretical and empirical investigation. *European Journal of communication*, *17*(1), 65-84.

Bosnjak, M., Tuten, T. L., & Wittmann, W. W. (2005). Unit (non) response in web-based access panel surveys: An extended planned-behavior approach. *Psychology & Marketing*, *22*(6), 489-505.

Bronstein, J., Gazit, T., Perez, O., Bar-Ilan, J., Aharony, N., & Amichai-Hamburger, Y. (2016). An examination of the factors contributing to participation in online social *pla*tforms. *Aslib Journal of Information Management*.

Buchanan, T., Paine, C., Joinson, A. N., & Reips, U. D. (2007). Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American society for information science and technology*, 58(2), 157-165.

Buntain, C., McGrath, E., Golbeck, J., & LaFree, G. (2016, April). Comparing Social Media and Traditional Surveys around the Boston Marathon Bombing. In # *Microposts* (pp. 34-41).

Burgoon, J. K., Parrott, R., Le Poire, B. A., Kelley, D. L., Walther, J. B., & Perry, D. (1989). Maintaining and restoring privacy through communication in different types of relationships. *Journal of Social and Personal Relationships*, 6(2), 131-158.

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Sage.

Callegaro, M., & Yang, Y. (2018). The role of surveys in the era of “Big Data”. In *the Palgrave handbook of survey research* (pp. 175-192). Palgrave Macmillan, Cham.

Chan, M. S., Morales, A., Farhadloo, M., Palmer, R. P., & Albarracín, D. (2018). Harvesting and harnessing social media data for psychological research. *Social Psychological Research Methods: Social Psychological Measurement*.

Cherchye, L., Moesen, W., Rogge, N., & Van Puyenbroeck, T. (2007). An introduction to ‘benefit of the doubt’ composite indicators. *Social Indicators Research*, 82(1), 111–145.

Cho, Y. I., Johnson, T. P., & VanGeest, J. B. (2013). Enhancing surveys of health care professionals: a meta-analysis of techniques to improve response. *Evaluation & the health professions*, *36*(3), 382-407.

Christofides, E., Muise, A., & Desmarais, S. (2012). Hey mom, what’s on your Facebook? Comparing Facebook disclosure and privacy in adolescents and adults. *Social Psychological and Personality Science*, 3(1), 48-54.

Claessens, S., Glaessner, T. C., & Klingebiel, D. (2002). *Electronic Finance: a new approach to financial sector development?* (Vol. 431). World Bank Publications.

Cruz-Jesus, F., Oliveira, T., & Bacao, F. (2012). Digital divide across the European Union. *Information & Management*, 49(6), 278-291.

Dandapani, K. (2017). Electronic finance–recent developments. Managerial Finance.

DeCew, J. W. (1997). *In pursuit of privacy: Law, ethics, and the rise of technology*. Cornell University Press.

DiMaggio, P., Hargittai, E., Celeste, C., & Shafer, S. (2004). From unequal access to differentiated use: A literature review and agenda for research on digital inequality. *Social inequality*, *1*, 355-400.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of business and Psychology*, *17*(2), 245-260.

Dror, Y. (2014). Privacy on the Israeli Internet. Survey report on privacy on the web and in apps (Hebrew). Retrieved from: http://din-online.info/pdf/dig.pdf

Dutton, W., & Blank, G. (2011). The Internet in Britain: Oxford Internet Survey 2011 Report.

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet research*.

Finn, R. L., Wright, D., & Friedewald, M. (2013). Seven types of privacy. In *European data protection: coming of age* (pp. 3-32). Springer, Dordrecht.

Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research.

Fox, S., & Madden, M. (2006). Generations online (demographic report). *Pew Internet\& American Life Project.*

Gamliel, G. (2017). National Initiative Israel Digital–The National Digital Program of the Government of Israel. *The Office for Social Equality*.

Gan, C., Clemes, M., Limsombunchai, V., & Weng, A. (2006). A logit analysis of electronic banking in New Zealand. *International Journal of Bank Marketing*.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, *66*, 53-61.

Giannakoudi, S. (1999). Internet banking: the digital voyage of banking and money in cyberspace. *Information and Communications Technology Law*, *8*(3), 205-243.

Giles, D., Stommel, W., Paulus, T., Lester, J., & Reed, D. (2015). Microanalysis of online data: The methodological development of “digital CA”. *Discourse, Context & Media*, *7*, 45-51.

Graham, J. W., Collins, N. L., Donaldson, S. I., & Hansen, W. B. (1993). Understanding and controlling for response bias: Confirmatory factor analysis of multitrait-multimethod data. *Psychometric methodology*, 585-590.

Hampton, K. N. (2017). Studying the digital: Directions and challenges for digital methods. *Annual Review of Sociology*, *43*, 167-188.

Helsper, E. J., & Galácz, A. (2009). Understanding the links between social and digital exclusion in Europe. *Worldwide Internet: Changing societies, economies and cultures*, 146.

Hoskin, R. (2012). The dangers of self-report. Science for all brainwaves.

Howard, P. E., Raine, L., & Jones, S. (2001). Access, Civic Involvement, and Social Interaction. *American Behavioral Scientist*, *45*(3), 382-404.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 2.

Jackson, L. A., Ervin, K. S., Gardner, P. D., & Schmitt, N. (2001). Gender and the Internet: Women communicating and men searching. *Sex roles, 44*(5-6), 363-379.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., ... & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839-880.

Jones, Q., & Rafaeli, S. (2000, January). What do virtual " ells" tell? Placing cybersociety research into a hierarchy of social explanation. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences* (pp. 10-pp). IEEE.

Jones, Q., Ravid, G., & Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research*, *15*(2), 194-210.

Jones, S., & Fox, S. (2009). Pew Internet project data memo. *Pew Internet & American life project*.

Jones, A. S., Horsburgh, J. S., Jackson-Smith, D., Ramírez, M., Flint, C. G., & Caraballo, J. (2016). A web-based, interactive visualization tool for social environmental survey data. *Environmental Modelling & Software, 84*, 412-426.

Jungherr, A. (2018). Normalizing digital trace data. *Digital discussions. How big data informs political communication. Oxon: Routledge*, 19-45.

Kellehear, A. The unobtrusive researcher: A guide to methods. 1993.

Kim, Y. C., Jung, J. Y., & Ball-Rokeach, S. J. (2007). Ethnicity, place, and communication technology: Effects of ethnicity on multi-dimensional Internet connectedness. *Information Technology &* People, 20(3), 282-303.

Kim, L. H., Qu, H., & Kim, D. J. (2009). A study of perceived risk and risk reduction of purchasing air-tickets online. *Journal of Travel & Tourism Marketing*, *26*(3), 203-224.

Law, R., & Leung, R. (2000). A study of airlines' online reservation services on the *Internet. Journal of Travel Research, 39(2), 202-211.*

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of broadcasting & electronic media*, *57*(1), 34-52.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lim, Y. J., Osman, A., Salahuddin, S. N., Romle, A. R., & Abdullah, S. (2016). Factors influencing online shopping behavior: the mediating role of purchase intention. *Procedia economics and finance*, 35(5), 401-410.

Limayem, M., Khalifa, M., & Frini, A. (2000). What makes consumers buy from Internet? A longitudinal study of online shopping. *IEEE Transactions on systems, man, and Cybernetics-Part A:* Systems and Humans, 30(4), 421-432.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International journal of market research*, *50*(1), 79-104.

Mastrandrea, R., Fournet, J., & Barrat, A. (2015). Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one*, *10*(9), e0136497.

McCrae, R. R., & Costa Jr, P. T. (1991). Adding Liebe und Arbeit: The full five-factor model and well-being. *Personality and social psychology bulletin*, *17*(2), 227-232.

Mizrachi, Y., Shahrabani, S., Nachmani, M., & Hornik, A. (2020). Obstacles to using online health services among adults age 50 and up and the role of family support in overcoming them. *Israel Journal of Health Policy Research*, *9*(1), 1-10.

Müller, H., & Sedley, A. (2014, December). HaTS: large-scale in-product measurement of user attitudes & experiences with happiness tracking surveys. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design* (pp. 308-315).

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). Handbook on Constructing Composite Indicators: ECD Statistics Working Paper 2005/3.

Neirotti, P., Raguseo, E., & Paolucci, E. (2016). Are customers' reviews creating value in the hospitality industry? Exploring the moderating effects of market positioning. *International Journal of Information Management, 36*(6), 1133-1143.

Norman, C. D., & Skinner, H. A. (2006). eHealth literacy: essential skills for consumer health in a networked world. *Journal of medical Internet research*, *8*(2), e9.

O'Brien, M. (2010). Unobtrusive research Methods: An interpretative essay. *Practicing Media Research*, 1-15.

OECD/DSTI. 2001. Understanding the digital divide. OECD papers.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

Osatuyi, B. (2015). Personality traits and information privacy concern on social media platforms. *Journal of Computer Information Systems,* 55(4), 11-19.

Park, S., Kim, E. M., & Na, E. Y. (2007). Defining user groups on Internet usage pattern of adolescents and its relation to relationships with peers. *Journal of Cybercommunication Academic Society*, *22*(2), 39-82.

Park, S., & Tussyadiah, I. P. (2017). Multidimensional facets of perceived risk in mobile travel booking. *Journal of Travel Research*, 56(7), 854-867.

Peter, J., & Valkenburg, P. M. (2006). Adolescents' Internet use: Testing the "disappearing digital divide" versus the "emerging digital differentiation" approach. *Poetics*, *34*(4-5), 293-305.

Petrillo, F., Spritzer, A. S., Freitas, C. M. D. S., & Pimenta, M. S. (2011, October). Interactive analysis of Likert scale data using a multichart visualization tool. In *IHC+ CLIHC* (pp. 358-365).

Raban, Y., & Sofer, T. (2014). Perception of privacy in the face of accelerated use of new ICT technologies - Threats, challenges and opportunities (Hebrew). Retrieved from https://www.isoc.org.il/wp-content/uploads/2012/08/ISOC-Il_Grants_2012_Final_research_report_Yoel_Raban.pdf

Rafaeli, S., Ravid, G., & Soroka, V. (2004, January). De-lurking in virtual communities: A social communication network approach to measuring the effects of social and cultural capital. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (pp. 10-pp). IEEE.

Rafaeli, S., Albo, Y. and Shiti, I. (2013) Israel National ICT Index - Research progress report on the creation of ICT Index to promote international technology and Internet use in Israel. Haifa: The Center for Internet Research. (in Hebrew).

Rafaeli, S., Leck, E., Albo, Y., Oppenheim, Y., & Getz, D. (2018). An Innovative Approach for Measuring the Digital Divide in Israel: Digital Trace Data as Means for Formulating Policy Guidelines. *Samuel Neaman Institute for National Policy Research*.

Rudder, C. (2014). *Dataclysm: Who we are (when we think no one's looking)*. Random House Canada.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public opinion quarterly*, *80*(1), 180-211.

Schumacher, P., & Morahan-Martin, J. (2001). Gender, Internet and computer attitudes and experiences. *Computers in human behavior*, *17*(1), 95-110.

Shahrabani, S., & Mizrachi, Y. (2016). Factors affecting compliance with use of online healthcare services among adults in Israel. *Israel journal of health policy research*, *5*(1), 15.

Sheldon, P. J. (1997). *Tourism information technology*. Cab International.

Shih, T. H., and Fan, X. T. (2008). Comparing response rates from Web and mail surveys: A meta-analysis. *Field Methods, 20*, 249–271.

Singh, P. K., & Dutta, A. (2020). Socio-metrics of digital payments in demographic dividend: Descriptive analysis of dichotomous preferences. *Applied Innovative Research (AIR)*, *1*(3-4), 171-177.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39, 156-168.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: key issues in developing an emerging field.

Subrahmanyam, K. (2001). 2001 New Forms of Electronic Media: The impact of interactive games and the Internet on cognition, socialization, and behavior. *Handbook of children and the media*.

Triandis, H. C. (1979). Values, attitudes, and interpersonal behavior. In *Nebraska symposium on motivation*. University of Nebraska Press.

Tsao, W. C., & Chang, H. R. (2010). Exploring the impact of personality traits on online shopping behavior. *African Journal of Business Management*, 4(9), 1800-1812.

Van Deursen, A. J., & Van Dijk, J. A. (2014). The digital divide shifts to differences in usage. *New media & society*, *16*(3), 507-526.

Van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. *Poetics*, *34*(4-5), 221-235.

Vehovar, V., Sicherl, P., Hüsing, T., & Dolnicar, V. (2006). Methodological challenges of digital divide measurements. *The information society*, 22(5), 279-290.

Vehovar, V. and Lozar Manfreda, K. (2008). Overview: online surveys. In: Fielding, N. (ur.), Lee, R. M. and Blank, G. (eds). *The Sage handbook of online research methods.* Los Angeles: Sage. pp. 177-194.

Vehovar, V., Petrovčič, A. and Slavec, A. (2015). E-social science perspective on survey process: towards an integrated web questionnaire development platform. In: Engel, U.

(ed). Improving survey methods: lessons from recent research, (European Association of Methodology book series). New York; London: Routledge. 2015, pp. 170-183.

Wang, D., Park, S., & Fesenmaier, D. R. (2012). The role of smartphones in mediating the touristic experience. *Journal of Travel Research*, 51(4), 371-387.

Webb, Eu. (2000). Unobtrusive Measures. Rev. ed, Sage classics series. Thousand Oaks, Calif.: Sage Publications.

Wells, C., & Thorson, K. (2017). Combining big data and survey techniques to model effects of political content flows in Facebook. *Social Science Computer Review*, 35(1), 33-52.

Werthner, H., & Klein, S. (1999). *Information technology and tourism: a challenging ralationship*. Springer-Verlag Wien.

Wexler, S. (2016). Visualizing survey data. Tableau. Retrieved from https://www. tableau. com/sites/default/files/media/whitepaper_surveydata_v4. pdf.

Xiang, Z., Wöber, K., & Fesenmaier, D. R. (2008). Representation of the online tourism domain in search engines. *Journal of Travel Research*, *47*(2), 137-150.

Xiang, Z., Magnini, V. P., & Fesenmaier, D. R. (2015). Information technology and consumer behavior in travel and tourism: Insights from travel planning using the Internet. *Journal of retailing and consumer services*, 22, 244-249.

Yano, K., Akitomi, T., Ara, K., Watanabe, J., Tsuji, S., Sato, N., ... & Moriwaki, N. (2015). Measuring happiness using wearable technology. *Hitachi Review*, *64*(8), 517.

Yarger, J. B., James, T. A., Ashikaga, T., Hayanga, A. J., Takyi, V., Lum, Y., ... & Mammen, J. (2013). Characteristics in response rates for surveys administered to surgery residents. *Surgery*, *154*(1), 38-45.

Zillien, N., & Hargittai, E. (2009). Digital distinction: Status-specific types of Internet usage. *Social Science Quarterly*, *90*(2), 274-291.

# Annex 1: Cronbach's Alpha tests for realibility

**National Survey**

| | N valid | Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|---|---|
| Privacy and data security factors | 688 | 0.821 | 0.823 | 13 |
| Factors (reasons) for booking flights online | 678 | 0.858 | 0.860 | 5 |
| Factors (reasons) for not booking flights online | 310 | 0.728 | 0.740 | 3 |

**Binational Survey**

| | N valid | Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|---|---|
| Privacy and data security factors | 688 | 0.821 | 0.823 | 13 |
| Frequency of visiting shopping websites | 1026 | 0.814 | 0.852 | 15 |
| Reasons (factors) for online shopping | 1053 | 0.852 | 0.855 | 10 |
| Reasons (factors) for refraining from online shopping | 1250 | 0.749 | 0.748 | 5 |