Designing Privacy Affordances for Data Release when Searching

by

Anisha Fernando

Master of Business Information Systems
from the University of South Australia (2013)
BSc. (Hons) 1st Class in Information Systems with Business Management
from the University of Westminster (2011)
Bachelor of Information Technology (External)
from the University of Colombo (2010)

A thesis submitted for the degree of

Doctor of Philosophy in
Computer and Information Science

School of Information Technology and Mathematical Sciences
Division of Information Technology, Engineering and the Environment

University of
South Australia

March 2017

Contents

List of Figures vi
List of Tables viii
Glossary ix
Abstract xi
Declaration xiv
Acknowledgments xv
1 Introduction 1
1.1 Motivation 1
1.1.1 Data: Its Value, Purpose and Flow in a Hyperconnected World 1
1.1.2 Web Search and Personalisation 4
1.1.3 Privacy and Data Flows in the Context of Web Search 6
1.2 Research Objectives, Methodology and Contributions 8
1.2.1 Research Objective 1: Technical Dimension - To Investigate the Effects of Data Release on Personalised Search Results 9
1.2.2 Research Objective 2: Empirical Dimension - To Understand the Privacy Expectations Users have about Data Release 10
1.2.3 Research Objective 3: Conceptual Dimension - Towards Designing Privacy Affordances to Manage Data Release for Users when Searching 12
1.3 Thesis Outline 14
1.4 Contributions from this Doctoral Study 15
2 Literature Review 17
2.1 Introduction 17
2.2 Information Privacy 17
2.3 Affordances in Socio-Technical Interactions 22
2.3.1 Gap in Affording Protection via Technological Mechanisms 24
2.3.2 Gap in Affording Protection via Societal Mechanisms 27
2.3.3 Gap in Affording Effective Protections for Data Release 31
2.4 Conclusion 32
3 Methodology 33
3.1 Approach 33
3.1.1 Technical Investigations: Methodology 39
3.1.2 Empirical Investigations: Methodology 42
3.1.3 Conceptual Investigations: Methodology 49
3.2 Alternative Approaches 50
3.3. Conclusion 51

4 Technical Investigation: Identifying the Effects of Data Release on the Variance of Personalised Search Results 52
4.1 Introduction 53
4.2 Related Work: Data Release and Leakage in Web Search 56
4.2.1 Data Release in Web Search 56
4.2.2 Data Leakage Methods on the Web 57
4.2.3 Existing Mechanisms of Managing Personal Data in Search 57
4.3 Experiment Methodology 60
4.3.1 Experiment Design 61
4.3.2 Identifying Search Query Terms 65
4.3.3 Identifying Search Query Parameters 65
4.3.4 Identifying Search Scenarios 68
4.3.5 Identifying and Controlling Sources of Noise 68
4.3.6 Measuring the Variance of Personalised Search Results 72
4.3.7 Analysing the Variance of Personalised Search Results 72
4.4 Results, Analysis and Discussion 75
4.4.1 Identifying Query Terms that Influence Personalisation Variance 76
4.4.2 Search Scenarios 88
4.4.3 Summarising Significant Implicit Data Released in Personalised Search 88
4.5 Conclusion 90
4.5.1 Understanding the Effects of Implicit Data Release on Personalisation Variance of Search Results 90
4.5.2 Identifying Affordances for Privacy-Sensitivity in Personalised Search Results 91
4.5.3 Evidence of Information Asymmetries with Data Release in Personalised Search Data Flows 92
5 Empirical Investigations: Privacy Expectations of Users when Searching 99
5.1 Introduction 99
5.2 Search Behaviour 102
5.2.1 Demographic Perspectives 102
5.2.2 Search Behaviour Characteristics 104
5.2.3 Privacy Expectations around Data Flows: Quantitative Views 105
5.2.3.1 Awareness of Personal Data Release and Collection 105
5.2.3.2 Expectations around Data Flows: Behavioural Advertising, Tracking, Linkability and Aggregation of Data 106
5.2.3.3 Expectations around Privacy Risks and Impacts 109
5.2.3.4 Awareness of Mechanisms that Enable Privacy 110
5.2.4 Understanding the Relevance of Search Results to Users 117

5.2.5 Measuring Privacy Concern: Comparisons with Westin's Privacy Segmentation Index 123
5.3 Privacy Expectations around Data Release: Qualitative Thematic Analysis 126
5.3.1 Creating Data 130
5.3.2 Collecting Data 136
5.3.3 Processing Data 143
5.3.4 Disseminating Data 145
5.3.5 Deleting Data 159
5.4 Empirical Contributions: Explicating Data Flows and Privacy Expectations using the Contextual Integrity Framework 161
5.4.1 Practice 1: Collecting and Retaining Metadata 163
5.4.2 Practice 2: Collecting and Distributing Data for Commercial Purposes 166
5.4.3 Practice 3: Framing Choices through Ineffective Affordances 174
5.4.4 Summary: Analysis and Application of the Contextual Integrity Framework to Qualitative Thematically Analysed Privacy Expectations around Data Flows when Searching 177
5.5 Conclusion 178
6 Recommendations for Designing Data Release for Privacy-Sensitive Users when Searching 180
6.1 Introduction 180
6.2 A Concept for Understanding Privacy Expectations in Socio-Technical Interactions 181
6.2.1 Understanding Stakeholder Values in Socio-Technical Interactions 181
6.2.1.1 Societal Characteristics 186
6.2.1.2 Technical Characteristics 188
6.2.1.3 Stakeholder Values Analysis 191
6.2.2 Understanding Morals Mediating Interactions 193
6.2.2.1 Intent 196
6.2.2.2 Action 198
6.2.2.3 Effect 198
6.2.2.4 Understanding Mediations in Socio-Technical Interactions 199
6.3 Designing for Socio-Technical Interactions in Search: Contextually-Public Release of Data 205
6.3.1 Emergence of a Contextually-Public Sphere 205
6.3.2 Establishing Norms for Online Engagement 211
6.3.2.1 Norms of Appropriateness and Distribution 212
6.3.2.2 Norms of Motivation 212
6.3.2.3 Norms of Value 212

6.3.3 Designing for Proactive Privacy Protections 214
6.3.3.1 Necessity as a Transmission Principle in Relational Interactions 214
6.3.3.2 Detaching Self: Establishing Personal Spaces Online 215
6.3.3.3 Shedding Data: Designing for Relational Beings 216
6.3.4 Initiatives to Scaffold these Design Recommendations 217
6.3.4.1 User Empowerment by Perceiving Visibility 217
6.3.4.2 Stakeholder Empowerment by Aligning Incentives 217
6.4 Conclusion 219
7 Conclusion 221
7.1 Key Findings, Contributions and Applicability 221
7.1.1 Technical Dimension: Investigating the Effects of Data Release 222
7.1.2 Empirical Dimension: Investigating Privacy Expectations 223
7.1.3 Conceptual Dimension: Understanding Mechanisms for Data Release in Search that Afford Users Privacy 223
7.2 Future Work 226
7.2.1 Extending the Applicability of the Controlled Experiment Findings on Data Release in Personalised Search 226
7.2.2 Extending the Applicability of the Empirical Survey Study on Privacy Expectations associated with Data Release when Searching 227
7.2.3 Extending the Applicability of the Conceptual Findings and Design Recommendations 227
7.2.4 Validating Privacy Expectations and Online Social Norms 227
7.2.5 Determining Privacy-Sensitivity in Search 228
7.2.6 Communicating Online Privacy 228
7.3 Conclusion 229
References xvii
Appendix A4.1 – Selected Figures: Personalisation Variance Analysis xxxvi
Appendix A4.2 – Search Query Parameters List xlv
Appendix A4.3- Detailed Quantitative Data Analysis of the Controlled Experiment on the Effects of Data Release on the Variance of Personalised Search Results xlvi
Appendix A5.1 Survey Instrument on Web Search and User Privacy cxxii

List of Figures

Figure 1.1 Socio-Technical User Interactions....6
Figure 2.1 Typical Examples – Confusing Signifiers on Doors....23
Figure 2.2 Design of Privacy Settings of Mobile Applications (Author's own images) 23
Figure 2.3 Mozilla's Privacy Campaign – February 2016 (Author's own images)....24
Figure 2.4 Inequality in Socio-Technical Forces....32
Figure 3.1 Overall Methodology: Techniques and Contributions....38
Figure 3.1-3.3 Experiment Design Conceptualisation of Data Flows in Personalised Search from a Privacy-Sensitive User's Perspective (An expanded view is provided in section 4.3.1)....42
Figures 4.1- 4.3 Experiment Design Conceptualisation of Data Flows in Personalised Search from a Privacy-Sensitive User's Perspective....64
Figure 4.4a Wilcoxon Test of Jaccard Index: 2nd Run – Google Scenario 1....74
Figure 4.4b NMDS of Jaccard Index: 2nd Run – Google Scenario 1....74
Figure 4.5a GLM- Query Terms – Query Terms Across all Search Engines....79
Figure 4.5b GLM- Query Terms –Comparison of Search EngineIdentifying Search Query Parameters that Influence Personalisation Variance....79
Figure 4.6a GLM- Query Parameters – Location vs. User-Agent Across all Search Engines....86
Figure 4.6b GLM- Query Parameters – Query Terms Across all Search Engine....86
Figure 4.7a Current Data Flows in Search....93
Figure 4.7b Proposed Privacy-Sensitive Data Flows in Search....93
Figure 5.2 Age Range of Participants....102
Figure 5.3 Participant Gender....102
Figure 5.4 Participant Role....103
Figure 5.5 Area of Expertise of Participants....103
Figure 5.6 Distribution of Locations Associated with Participants (including cultures or places of origin or heritage)....103
Figure 5.7 Reliance on the Web to Search for Info....104
Figure 5.8 Familiarity with Web Search Use....104
Figure 5.9 Frequency of Browser Use....105
Figure 5.10 Search Engine Preference....105
Figure 5.11 Search Engine Use Frequency....105
Figure 5.12 Device Frequency....105
Figure 5.13 Search Scenarios....120
Figure 5.14 Search Engine Used....120
Figure 5.15 Browser Used....120
Figure 5.16 Use of Account Login....120
Figure 5.17 Sensitivity of Query Term: Blood Pressure....120
Figure 5.18 Sensitivity - Blood Pressure: Expected Use of Personal Information....120
Figure 5.19 Sensitivity of Query Term: Hepatitis B....121
Figure 5.20 Sensitivity: Hepatitis B: Expected Use of Personal Information....121
Figure 5.21 Search Result Comparison A: Blood Pressure....121
Figure 5.22 Search Result Comparison A: Hepatitis B....121
Figure 5.23 Reasoning: Comparison A....121
Figure 5.24 Sensitivity of Query Terms: Flight Tickets....122
Figure 5.25 Sensitivity: Flight Tickets: Expected Use of Personal Information....122
Figure 5.26 Sensitivities of transactional terms....122
Figure 5.27 Reasoning: Comparison B....122

Figure 5.23 Privacy Expectations around Data Flows 129
Figure 5.24 Privacy Sensitivities 130
Figure 5.25 Practice 1: Metadata Retention Practice 163
Figure 5.26 Practice 2a - Appropriateness of Data Flows 167
Figure 5.27 Practice 2b: Distribution Effects of Data Flows (Normal Scenarios) 170
Figure 5.28 Practice 2b: Distribution Effects of Data Flows (Exceptions) 170
Figure 5.29 Practice 3: Effects of Ineffective Affordances 175
Figure 6.1 Socio-Technical User Interactions 182
Figure 6.2 Socio-Technical User Interactions - Expanded View of Stakeholders 185
Figure 6.3 Socio-Technical Mediation Pattern 195
Figure 6.4 Nature of Data Flows 206
Figure 6.5 Visualising Contextually-Public Spheres: A Public Reading Place (Author's own image) 209
Figure 6.6 Progression of Identities in Socio-Technical Interactions 210
Figures A1-A56: Selected Figures: Personalisation Variance Analysis xxxvi
Table A4.2.6 Search Query Parameters and Impact on Personalisation Variance & Identifiability xlv
Figure A4.3.1- Wilcoxon Test of Jaccard Index: 2nd Run – Google Scenario 1 xlvii
Figure A4.3.2- NMDS of Jaccard Index: 2nd Run – Google Scenario 1 xlviii
Figures A4.1-A4.22 StartPage lxv
Figures 4.23 – A4.47 DuckDuckGo lxviii
Figures A4.82 – A4.206 Google lxxviii
Figures A4.207 – A4.261 Bing xcviii
Figures A4.262 – A4.325 Yahoo cvii
Figure A4.327-1st Run- Google & StartPage Comparison – Jaccard Index: Query Terms vs. Scenarios cxviii
Figure A4.327-1st Run- Google & StartPage Comparison – Edit Distance: Query Terms vs. Scenarios cxviii

List of Tables

Table 2.1 Overview of Technical Privacy Protections Afforded 27
Table 2.2 Overview of Societal Privacy Protections Afforded 28
Table 3.1 Overall Experiment Design 40
Table 3.2 Survey Study Design 45
Table 4.1 Overall Experiment Design 62
Table 4.2 Search Query Terms 66
Table 4.3 Search Scenarios and Search Query Parameters 67
Table 4.4 Controlling Release of Implicit Data in Search 71
Table 4.5 Generalised Linear Modelling for Query Terms 78
Table 4. 6 Generalised Linear Modelling for Query Parameters 85
Table 4.7 Experiment Findings – Summary of Analysis on Personalisation Variance ..89
Table 5.1 Survey Study Design 100
Table 5.2 Expectations about Awareness 107
Table 5.3 Online Privacy Expectations: Interactions 110
Table 5.4 Online Privacy Expectations: Preferences 112
Table 5.5 Expectations: Consent and Disclosure 114
Table 5.6 Awareness and Use of Privacy Info 115
Table 5.7 Privacy-enabling Mechanisms 116
Table 5.8 Characteristics of Consent and Disclosure 116
Table 5.9 User Evaluation of Obfuscated Personalised Search Results 118
Table 5.10 Reasoning of Search Results Selection 118
Table 5.11 Distribution of Privacy Concerns using Westin's Privacy Segmentation Index as a Baseline (as cited in (Kumaraguru & Cranor 2005)) * 125
Table 5.12 Categorised Data Flows and Privacy Risks 127
Table 6.1 Mechanisms of Implicating Human Values in Technology 182
Table 6.2 Characteristics of Socio-Technical User Interactions 183
Table 6.3 Summary – Stakeholder Value Analysis 192
Table 6.4 Types of Technological Mediations 195
Table 6.5 Mediation Effects 196
Table 6.5 Existing Affordances 202
Table A4.2.6 Search Query Parameters and Impact on Personalisation Variance & Identifiability xlv
Table A4.3.1- Appendix-Analysis: Experiment Findings Table xlix

Glossary

Affordances: Affordances in the context of this research expands beyond Gibson's concept of direct perception by using Verbeek's mediation analysis to offer new language or explanations in addressing privacy expectations of users around data release when searching (Gibson 1979, Verbeek 2011).

Contextual Integrity: Based on **Nissenbaum's theory of contextual integrity, t**he contextual integrity framework provides a definition for information privacy and a decision heuristic in identifying contexts in which this definition is not upheld, i.e. when privacy expectations are not met (Nissenbaum, 2010).

Data Flows: Information released when searching is characterised as data flows. The terms *flow of information, information flows, data release*, *information practice* or *practices* are used synonymously to refer to data flows in this thesis.

Data Leakage: Release of personal data without the explicit knowledge or consent of users is defined to be *data leakage* in the context of this thesis.

Data Release Types: Data release when searching in the context of the controlled experiments (Research Question 1 as detailed in Chapter 4) were defined as:

- *Minimal data release* is the lowest possible amount of data released to obtain search results. For most search engines, this was found to be the query term and IP address.
- *Default data release* is the default as-is query containing the data that is transmitted when individuals use any of the search engines investigated in the controlled experiment.
- *Obfuscated data release* is the release of minimal and changed data to obtain search results. For example, an obfuscated data release could involve the minimal data release (corresponding to the search engine investigated) and an obfuscated IP address.

Information Service Providers: Organisations or entities that provide users with services that involve data flows such as collecting, processing and disseminating data. These include search engine organisations and other commercial entities like third-party advertisers.

Information Privacy: The information privacy definition used in this thesis is based on the contextual integrity framework which states "we have a right to privacy, but it is neither a right to control personal information nor a right to have access to this information restricted, but it is a right to appropriate flow of personal information" (Nissenbaum 2010, p.127). This definition of information privacy is used because it is relevant to the nature of the problem as it allows the identification of privacy expectations and associated practices to be identified. These findings in turn determine the level of effectiveness of privacy protection mechanisms afforded to users when releasing data while searching.

Personal Data: Data that may contain characteristics which could be used to identify a person (European Commission 2015b).

Practices: "Practice is first and foremost a process by which we can experience the world and our encounters with it as meaningful." (Wenger 1998, p.51 as cited in Dourish & Anderson 2006). Information practice is about "the ways in which we collectively share, withhold, and manage information; how we interpret such acts of sharing, withholding, and managing; and how we strategically deploy them as part and parcel of every social interaction" (Dourish & Anderson 2006, p.335). Therefore, practices afforded to users can be viewed as mechanisms to practice privacy. This notion of privacy as a practice is used as a frame for analysis of privacy expectations around data release as expressed by the research participants. The affordances embedded in searching make these data flows meaningful in exercising the privacy expectations users hold.

Value Sensitive Design: The value sensitive design approach employs technical, empirical and conceptual investigations to study the effects of a phenomenon (i.e. privacy expectations and the effectiveness of privacy protection affordances) in a given application domain (i.e. data release in personalised search) (Friedman, Kahn & Borning 2009).

Abstract

Current information-seeking activities of users are driven by complex data processing, ease of access and ubiquitous services. When searching data-driven services online, a significant issue affecting users is privacy. The impacts of privacy loss are often instantaneous, reputational and may concern data about interconnected users. To address user privacy concerns, research has developed privacy-enhancing technologies, privacy-by-design concepts and policy mechanisms. Less adequate, however, is an understanding of online societal norms in relation to user privacy expectations around personal data flows. The nature of the Web adds to these complex interactions, making actors and purposes largely invisible.

This research study explored how and to what extent, technological affordances around managing personal data flows are effective in addressing people's privacy expectations in three phases. Using Nissenbaum's contextual integrity framework, a value sensitive design study was conducted to investigate the effectiveness of mechanisms enabling appropriate data flows. The effectiveness of privacy protections afforded to users were explored from technical, empirical and conceptual dimensions situated in the application domain of personalised search results.

In the technical phase, a controlled experiment was conducted on data release in search across seven search engines (both commercial and privacy-enhanced), four user scenarios, 29 query terms (informational, transactional, navigational and popular) across three distinct timeframes from 2014 to 2016. The analysis of personalisation variance using the Jaccard Index and Edit Distance showed that personalised search results are consistently and significantly affected by location, the user-agent parameter and query term popularity. These findings lend support to minimising data leakage through obfuscation without diminishing the variance of results. Where minimisation is not possible then data release should be made visible to users to enable them to make an informed choice.

The empirical phase surveyed a sample of 441 users aged between 18 and 70 years from a population of university students and staff. The aim was to identify online societal norms around privacy-sensitive data flows through quantitative and qualitative questions, and interactive search tasks. The findings revealed that users expect data release when searching online to be constrained to the primary purposes, actors and

conditions shaping the interaction. Users found value in releasing data, but for a specific purpose. Three practices identified were: collecting and retaining metadata, collecting and distributing data released for commercial purposes, and framing choices through ineffective affordances. Network effects were evident where a decision to opt-out of using a service meant to not engage, where the cost of engagement loss was a significant deterrent. When expectations were not met, users passively accepted the choices offered or found the interaction creepy, but were unable to take proactive action. Low levels of technical literacy inhibit the use of privacy-enhancing technologies. Users also had no significant preference between obfuscated results and non-obfuscated results except when location was a factor. These information and power asymmetries highlight the need for mechanisms that afford privacy protection by empowering the agency of the user.

In the conceptual phase, stakeholder and moral mediation analyses revealed the extent to which data is propagated, strengthening the information and power asymmetries. Data release was determined to be contextually-public because the properties of the Internet and its underlying infrastructure are public, although expectations of individuals are largely private or constrained to the purposes, actors and conditions. Norms governing data release include the appropriateness and distribution of data flows, motivations underpinning interactions and tensions between the social value of privacy and the market value of interactions. These findings give rise to the notion of relational interactions, where users interact in a capacity. Recommendations for design of relational interactions include adopting necessity as a transmission principle, detaching the whole self from interactions and shedding data based on temporal or local interests. Initiatives to scaffold these recommendations include empowering users to perceive the visibility of their data by creating a privacy mindset at a grassroots level in societal cultures, enabling accountable data protection and using social capital to incentivise organisations to whom data is released. The consequences of affording users with ineffective mechanisms for privacy protection include propagating the subliminality of data release, framing choices, conditioning beliefs and habituating norms. These effects if realised would risk the value and benefits of sharing data while compromising the protections needed to retain the integrity of socio-technical interactions that mediate human expression of thought and action.

Although the capabilities of technology are powerful and its impact and risks are far-reaching, the use of these capabilities should be aligned with moral values and not necessarily be exploited just because they exist. Being able to balance the integrity of social life with the immense benefits technology brings is important for technology-mediated interactions on the Web to be sustainable in the long term.

Keywords

Information Privacy, Contextual Integrity, Value Sensitive Design, Affordances, Personalised Search, Data Flows, Data Release, Data Leakage.

Declaration

This thesis presents work carried out by myself and does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; to the best of my knowledge it does not contain any materials previously published or written by another person except where due reference is made in the text; and all substantive contributions by others to the work presented, including jointly authored publications, are clearly acknowledged.

Anisha Thilini Jessica Fernando

25th March 2017

Acknowledgments

I would like to sincerely thank Helen Ashman for the opportunity to undertake this transformative journey, in exploring a socio-technical issue like information privacy and for your patience as I taught myself skills which enabled me to tell the data's story. They will take the form of papers soon! Also, I would like to thank Tina Du for her feedback and comments. I'd like to thank Kirsten Wahlstrom and Victor Prokhorenko for being good bookends of the socio-technical spectrum, for being willing and able to listen to my ideas, thoughts and views and debate these issues with me and in the process, facilitate learning.

I would like to sincerely thank the .auDA Foundation for funding this work, and for subsequently providing an opportunity to present it through the Australian Internet Governance Forum. In addition, I would like to thank Kathleen Lumley College, benefactors of the Harold Rodda, and Harry Medlin memorial scholarships, and the Cowan bursary for funding that supported my research endeavours.

I would also like to thank Prof. Vinny Wade for hosting me at Trinity College Dublin, for the research exposure and experience in the domain of personalisation, and Tim Brailsford for hosting discussions at the University of Nottingham in Malaysia which trigged my interest in the Internet's history and purpose as a social institution.

I would like to thank the staff at the school of ITMS, in particular Andy Koronios, Jiuyong Li, Amanda Johnston, Tricia Huffa, Kathleen Grosser and Elyse Perin for the work they do in supporting doctoral students.

My appreciation also goes out to Gaye Deegan for her persistent efforts in running the Masters in BIS program that led to a research industry placement with the City of Salisbury which opened doors of new possibilities and for her continuous support. To Kathy, Gaye, Kirsten and Georg for the opportunities to teach and translate knowledge into learning in action.

I also like to thank Monica Behrend, Judy Ford and others at the Research Education Support Activities group for the amazing work they do in running research workshops, which provide a much-needed vital infrastructure in our training as researchers, not just in terms of knowledge sharing, but what it means to be a part of research culture.

I also appreciate the feedback I've had from presenting my work whether it was through conference presentations, guest speaker talks, reviewer feedback on papers, all of which enabled continuous learning. I'd also like to thank the pilot and survey study participants for their time and effort. I deeply appreciate the teachers who've inspired me to keep preserving in pursuing my interest in computer science and my passion for learning.

To my friends both here in Adelaide and across the many oceans, for the gifts of friendship and shared experiences that have shaped and enriched our lives. To Aunty Maritana, Uncle Rohan, Sashika, Tanika and Aunty Ann, for sharing your kindness, laughter, love and company, for being a window through which I experience the rich diversity of Australian life.

To my brothers, Aiyya (Rukshan) and Malli (Shiran), for doing what brothers do and keeping this quintessential middle child toughened up, grounded and loved. To my sister-in-law, Rashmi Akki and baby nephew Ranik for being new sources of joy and love in our lives these past months and in the years to come.

To my parents, for their unconditional love and support as I adventured into unfamiliar terrains and took on new challenges and opportunities to grow. Thathi (Patrick), your work ethic, sacrifices and support gave me access to opportunities that lead to growth and learning. Ammi (Shanthi), as my first teacher for grounding in me the values and faith that underpin this thesis. For teaching me that what is right and good, may not be convenient or fast or easy, but that it simply matters in life. And for being my rock, patiently picking me up whenever I adventured into the unknown and stumbled and fell, only to try and rise again with determined belief and grace (as much as I could muster). For your lifelong countless efforts in speaking and standing up for an equal right to education in largely patriarchal cultures, so that your daughter could have access to the same opportunities in life as your sons. For these reasons and for so much more, this labour of love is dedicated to you, the strongest woman I have had the privilege to know.

1 Introduction

1.1 Motivation

Personal data release in the context of personalised search results is investigated in this research. It is motivated by the problems of pervasive data leakage, unclear privacy expectations and ineffective affordances when conducting Internet searches. These problems create and propagate information and power asymmetries between an individual's right to privacy and the capabilities offered to users in exercising this right. This research focuses on the notion of users managing their personal data flows from a user's perspective as a method to protect their identity online and, thereby, contribute towards managing their digital privacy. The research explores the effectiveness of privacy protection mechanisms afforded to users when releasing data while searching.

1.1.1 Data: Its Value, Purpose and Flow in a Hyperconnected World

We live in a hyperconnected world, where our interactions leave a sticky trace by default (Dewandre 2015b). Often, this sticky trace takes the form of a new type of oil, data (Humby 2006). Data is the single most valuable commodity in the information age (Spiekermann & Korunovska 2016). It fuels our everyday activities which are increasingly digital with technology-enabled experiences becoming the social norm. The nature of this hyperconnected medium is reactive and talkative, where the boundaries between our offline and online lives are fluid (Dewandre 2015b; Pasquale 2015). The data flowing in this hyperconnected medium may be of a personal nature.

Personal data is any piece of information that contains or makes it possible to infer identifiable characteristics about a person (European Commission 2015a). Personal data is constitutive, as it is about a person (Floridi 2015). Hence personal data is especially valuable to users, but also to information providers who monetise it. Arguments like "If you're not paying for it, you're not the customer. You're the product being sold" (Anon) have some merit as people and data about them become central to the business models on which many technology-centric organisations operate. A host of businesses act as online data vendors, information brokers or information services. These businesses are generally called omnibus information providers (Nissenbaum 2010).

Some widely-known examples of data brokers include Axicom Corporation, ChoicePoint Inc., and First Advantage Corporation.

Monetising data also includes mining user data and customer behaviour to derive insights for businesses, in addition to harnessing value from descriptive data collection. For example, Dunnhumby (a customer science company) specialises in customer loyalty programs. Hence, there is a clear commercial drive to monetise data collected about users and derive insights about customers to offer services of value to them whilst also generating business-related profits.

The value of data is not just accessible to corporations; it is accessible to everyday users as well. The ubiquitous nature of technology means that anyone can be easily connected to anyone else or interact with content and services in a seamless manner (Zimmer 2008). The accessibility to technologies with high-processing capabilities that are available anytime, anywhere and literally in the palm of a hand enables people to stay connected and have rich user interactions. This accessibility also means that people are more readily able to process and aggregate data of increasing complexity and large volumes. People can derive value through access to technology, processing capabilities and the pervasiveness of such technologies (Nissenbaum 2010). These data flows indicate information asymmetries where the value exchanged in an interaction is imbalanced (Nissenbaum 2010). The Internet is a general-public utility on which technologies act as a conduit for data flows (Naughton 2010). Increasingly, the corporations that create the technologies hold greater power in its design, use and dissemination, than the people from whom data is collected (Zuboff 2015). These corporations initiate new power asymmetries by influencing reputation (i.e. how people are rated and ranked), information-seeking needs (i.e. how ratings and rankings are used to classify search results of interest) and finance (where the value of data is central to the business models fuelling the commercial interests of information service providers) (Pasquale 2015). These power asymmetries are further aggravated with instances of misuse of information.

The implications of data misuse raise many privacy concerns. Data collection is extensive. Through personalised search, this collected data caters to the needs of advertisers (Feuz, Fuller & Stalder 2011). Data when aggregated and even if anonymised can be used to infer details about a person (Sweeney 2002). Seemingly

disparate search queries are often linkable and a person's identity may be inferred, as evident in the AOL search data leak where query logs of more than half a million AOL users between a three month period were released (Barbaro & Zeller 2006). Quite often the main consequence of data misuse is reputational damage which can be instantaneous, far-reaching and permanent. This is similar to social shaming cases on Twitter. Sometimes the implications of the misuse of data can have negative consequences particularly relating to a person's personal and immediate network (as evident in online bullying) and can extend into consequences in real life such as the case of Target predicting a teen's pregnancy before her father was aware (Ellenberg 2014).

Given the implications of data misuse, data collection also warrants protection (O'Dell 2015). The extent and scale of data collection increase the likelihood of companies being hacked. The exposure of personal data can have severe negative consequences on not just the person involved but their extended network (For example, as seen in the Ashley Madison hacking scandal). The widespread data collection enables surveillance. Schneier suggests surveillance as the business model of the Internet (Schneier 2015). This suggestion appears to be substantiated as user movements and preferences are routinely tracked, big-brother style with even personalised ads integrated to daily life (For example, Tesco's Amscreen face-recognition technology). Although users trust large corporations due to their ability to set up resources to safeguard our data, these organisations are also more attractive targets as it is more lucrative to attack an organisation with a sufficient critical mass of data.

In this project, the release of personal data without the explicit knowledge or consent of users is defined to be *data leakage*. Data collection presumes that personal data is released with user consent. However, data that could be personally-identifiable is also released during an interaction which may not be visible to users. The research explores the issue of data leakage as a problem from the user perspective.

Numerous methods of tracking user behaviour exist. For example, data leakage can occur through inadvertent sharing of personal information, Web tracking mechanisms and third-party agents and advertisers (Malandrino & Scarano 2013). Behavioural profiling, using data collected through persistent tracking mechanisms, is used to customise outcomes like Web search results (Toch, Wang & Cranor 2012). Targeted advertisements leak personal user information not only to advertisers, but to any party

that has access to those ads (Castelluccia, Kaafar & Tran 2012). On the Web, the pervasiveness of tracking extends not just to the Web page or the service the user currently visits, but when tracking mechanisms are not visible to users as user data maybe captured by third-party networks and advertisers. Online advertising behaviourally tracks Web users to capture aspects of their behaviour, such as interests and preferences. This relevance feedback is captured through mechanisms such as cookies and are used to build user models that generate personalised recommendations.

In summary, data is a valuable commodity and is monetised by service and information providers. When data is misused, privacy concerns arise. Data release without user awareness or knowledge, i.e. *data leakage* raises privacy concerns when searching. Unpacking the nature of these privacy expectations are the focus of this research.

1.1.2 Web Search and Personalisation

Searching is a fundamental component of online interactions. In today's information age, Web search engines are widely used to find specific information, explore content and obtain knowledge (Lewandowski 2012). Search engines play a significant role in facilitating information retrieval and knowledge transfer. Key market players include Google with a market share of over 80%, followed by other significant players like Baidu, Yahoo and Bing (Statista 2016).

A main source of income for search engine organisations is targeted advertising based on user profiles. Internet advertising revenues totalled US$32.74 billion for the first half of 2016, with desktop search accounting for $8.9 billion of this generated revenue (PriceWaterhouseCoopers 2016). So, whilst aiming to provide relevant results for users, it is beneficial for search engine organisations to also provide relevant audiences for advertisers.

Generally, the search process involves a series of three basic steps. These are: query submission by a user and retrieval of search results; the identification of relevant and non-relevant information by a user (captured through relevance feedback mechanisms) and the subsequent modified search results list (based on the relevance feedback received through techniques like query reformulation and re-ranking) (Steichen, Ashman & Wade 2012). Ranking algorithms are used to retrieve and rank search results.

In traditional search, retrieving results relevant to the user's query and needs is an existing problem as methods of expressing user needs are mostly linguistic (keywords). This approach has limits because keywords may not accurately capture all key aspects of user preferences in terms of relevance. Another problem is information overload or information oversupply, which exists where a vast amount of results are available for a specific query and users are not able to peruse them all to find the relevant results (Lincoln, A 2011). Users also expect instant gratification and hence search results need to be designed to meet user needs in an accurate and timely manner. Another issue is that information is constantly being changed, as the Web is a highly dynamic environment (Micarelli et al. 2007).

In order to address these problems of relevance and information overload, personalisation of Web search was introduced by Google in 2009, regardless of whether users were logged in, after trialling it since 2005 (Sullivan, D 2009a). Personalisation was a major shift in ranking methodology because it requires personal data in order to personalise content for the user (Berendt & Teltzrow 2003).

In general, search personalisation can largely be categorised into three types: *adaptive retrieval* where a query can be modified based on the user's intentions and context, *intelligent content reasoning* where similar ideas or concepts associated with the original query can be identified and *adaptive content composition* where the navigation, presentation and interaction elements of the resulting content page are enhanced (Steichen, Ashman & Wade 2012). Personalised search can be defined as session-based where an initial query is followed by subsequent queries from the user or multi-query based consisting of multiple query session. Personalised systems can also be categorised as adaptive to a user's need or adaptable by the user (Ashman et al. 2012). Personalisation in the context of this thesis is defined as the release of personal data with the intent of customising search results and providing relevant search results based on a user's needs. Personalised search in this context focuses on the initial query from a single session given that the ability to proactively mitigate the effects of personal data release and as mitigating privacy loss is most conducive with the initial release of data. The intensive use of personal data in personalisation raises many privacy implications.

1.1.3 Privacy and Data Flows in the Context of Web Search

"Privacy is a right to live in a world in which our expectations about the flow of personal information are, for the most part, met" (Nissenbaum 2010, p.231). Information released when searching is characterised as data flows. The terms *flow of information*, *information flows*, *data release*, *information practice* or *practices* are used synonymously to refer to data flows in this thesis. When data flows appropriately, privacy exists according to the contextual integrity framework (Nissenbaum 2010). Appropriateness of data flows depends on the social norms associated with it. Online social norms are formed through collective privacy expectations of people. Privacy expectations are shaped by a person's interactions with online information service providers.

Online interactions between users and information providers are mediated by socio-technical affordances embedded in the platforms that underpin these interactions. "An affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used." (Norman 2013, p.11). In this context, the object comprises of both the service provided, such as search, and the underlying mechanisms that enable the interactions, namely the Internet and its architecture. The agent in this case is the user interacting with the online service provider. During these socio-technical interactions involving the flow of personal data, privacy expectations are formed and shaped through affordances offered to users as illustrated in Figure 1.1[1].

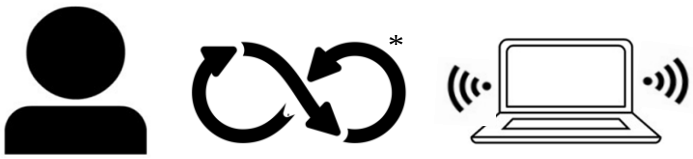


Figure 1.1 Socio-Technical User Interactions

Social interactions of people are immersed in technology. Searching is a type of socio-technology interaction between a user and a search engine. Searching online might raise privacy risks due to the release of personal information. These risks could take the form of banal spam emails to more intrusive ransomware scams, where the degree of

[1] Infinity by Olivier Guin from the Noun Project is licensed under CC BY 3.0 US

insecurity increases with the computational sophistication of the attacks. Often privacy is a secondary intangible goal of our interactions.

Consider a smartphone as an example, which if it is an iPhone with an iOS 7 or higher, it has by default an itemised list of frequent locations a person has been in based on their Wi-Fi activity (Apple Inc. 2016). The capability to harness this information exists beyond its intended purpose of delivering locally relevant information. Although this information was previously transmitted, it is only now that it is visible to iPhone users giving them some control in influencing how their personal location data may be used. The ability to commercialise personal data drives the development of technology that mediate user interactions.

When searching, the visibility of personally-identifiable data transferred is not typically apparent to users. Privacy expectations users hold around data flows and what constitutes appropriate data flows are unclear. It is important to determine the expectations, intentions and behaviour people associate with privacy in the context of searching to understand the implications from a user point of view.

Most users are unaware of the extent and scale of data collection. They also may not fully understand the implications of personally-identifiable data use. From a user's perspective, personal data is used to personalise search results and offer users the benefits of obtaining results relevant to their user's needs. However, this release and collection of personal data also provides significant commercial benefits as it provides mechanisms to target and track user preferences for behavioural advertising, which search engines commercialise as revenue streams. This purpose of personalisation based on targeted advertising from an organisation's perspective conflicts with the purpose of personalisation focusing on relevant results from a user's perspective, creating a personalisation-privacy trade-off. The effects of personal data in producing relevant personalised search results for users are unclear and is investigated in Chapter 4.

Providing relevant search results involves capturing user preferences and needs and storing them in user models. These user models should be scrutable where users can manage the data stored about them (Kay & Kummerfeld 2006). Context or contextual information is often used to determine relevance, but contextual information contains interrelated data that may be personally-identifiable (Pitkow et al. 2002). Hence, while context-awareness is an important factor in making results relevant to users (such as

search queries involving location-aware map directions), issues of relevance vs. privacy-sensitivity remain.

Commercial search engines collect longitudinal user data by monitoring and tracking user-browsing activities (Kosta et al. 2009; Tene 2008). In addition, third-party advertisers track users across different websites where not only is the primary user being tracked at risk, but information about other related users are captured as well (Krishnamurthy & Wills 2009). Third-party agents use stateful tracking methods via persistent supercookies and stateless tracking through active and passive fingerprinting techniques (Mayer & Mitchell 2012). Search engines or browser tool vendors may not necessarily make it easy for people to access these options to manage the user data being transferred or leaked.

Therefore, there exists a need to identify if existing mechanisms or affordances for managing personal data are satisfactory. From a human-computer interaction angle, it is important to empower users to manage personal data about them. As Web technologies mature, people become highly dependent on the Web for basic activities. To accommodate these changes, the provision of affordances needs to be integrated and transparent around personal data release.

In that regard, the application domain within the scope of this thesis is personalised search results from the perspective of users. Specifically, the study questions whether relevant personalised search results can be obtained with releasing minimal amounts of personally-identifiable user data or by obfuscating personally-identifiable information. This research also explores the mechanisms for proactive protection of personal data so that the choice to limit or transfer personal data occurs at the client-end with the explicit knowledge of the user.

1.2 Research Objectives, Methodology and Contributions

The overall question underpinning this research is: *To what extent are existing technological affordances around managing data flows effective in addressing people's privacy expectations when searching?*

This research explores data flows in search from a user's perspective and is motivated by the problems of data leakage, the lack of understanding of online privacy expectations and the effectiveness of affordances through a value sensitive design study.

As noted above, data leakage refers to the implicit release of personally-identifiable data whilst searching on the Web. Privacy expectations refer to beliefs people hold and their behaviour around searching. Affordances may enable people to manage their personal data when searching by exercising their privacy views.

"Value sensitive design is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman, Kahn & Borning 2009, p.69). It employs technical, empirical and conceptual investigations to study the effects of a phenomenon (such as the value of privacy) in a given application domain (i.e. data release in search). Value sensitive design is a suitable method for this research given the privacy concerns and implications of data release in search from a user's perspective. This study uses technical and empirical lenses to study the problem from a socio-technical angle to understand how effectively technological affordances can meet the privacy expectations of people and empower them to effectively manage data release when searching. To achieve this overall research aim, the following section outlines the research questions along with the method used to explore the question and resulting contributions.

1.2.1 Research Objective 1: Technical Dimension - To Investigate the Effects of Data Release on Personalised Search Results

Research Question 1: What is the effect of data release on the variance of personalised search results? To achieve this objective, variables which appear to influence the variance in personalised search results and those that contain personally-identifiable information were determined. Personalisation variance is an appropriate measure to investigate the effects of personal data on search results given the black-box nature of the investigation, where the focus is on controlling the release of data and identifying the subsequent effects on personalised search results. A set of controlled experiments were designed to investigate the effects of implicit data release on the variance of personalised search results across commercial and privacy-enhanced search engines. The controlled experiment was run across seven search engines (Google, Yahoo, Bing, DuckDuckGo, StartPage, Ixquick and Unbubble) and four user search scenarios (including both account and non-account login methods) using 29 query terms (sourced from popular search terms used) across three distinct timeframes (October 2014, May 2015 and January 2016).

A key contribution from the rigorous statistical analysis of this technical investigation is that overall, location, user-agent (the type of browser and operating system) and topicality of the query term influence the variance of personalised search results. Data release has an active (i.e. measurable) and passive (i.e. negligible) impact on the variance of search results. Three types of data release were defined as a result of this investigation:

- *Minimal data release* is the lowest possible amount of data released to obtain search results. For most search engines, this was found to be the query term and IP address.

- *Default data release* is the default as-is query containing the data that is transmitted when individuals use any of the search engines investigated in the controlled experiment.
- *Obfuscated data release* is the release of minimal and changed data to obtain search results. For example, an obfuscated data release could involve the minimal data release (corresponding to the search engine investigated) and an obfuscated IP address.

Hence, we propose that reducing data leakage in search queries can be achieved without diminishing the quality of search results. In the few exceptional circumstances where this data release is required, obfuscating such data (such as location or user-agent) or increasing the visibility of data flows to users are important steps in enabling users to manage data release in search. These contributions advance the state of the art as an understanding of the effects of personal data release on personalised search results are known and allows for the development of data release mechanisms that afford users privacy. These affordances can help establish trust in socio-technical interactions between users and information providers. Chapter 4 reports these findings in detail.

1.2.2 Research Objective 2: Empirical Dimension - To Understand the Privacy Expectations Users have about Data Release

Research Question 2: What privacy expectations do users have about the flow of personal data in search? In responding to this question, this research identified the privacy concerns of users and the implications of personal data use in personalisation. To address this objective, a survey was conducted between November 2015 and April

2016 to explore privacy expectations and personal data flows in search. The sample of 441 participants reflected a diverse mix of cultures commonly assoicated with Australian university populations including Australians (55.56%), Asian (18.08%), UK (9.5%), European (9.08%) with representation from South America (2.6%), North America (0.65%) and Africa (0.68%). The survey examined user intentions such as awareness about tracking, search history and use of data, and included an interactive component where respondents self-reported search behaviour across different data flows (such as accessing, tracking, selling personally-identifiable information). Respondents provided an evaluation of the affordances derived through Research Question 1 (specifically identifying if privacy-sensitive users preferred search results with minimal, obfuscated or default data release). A contextual understanding of the factors affecting privacy sensitivity in search was also developed through investigating search contexts and data flows.

Key contributions: The analysis of user privacy expectations identified three main practices through which users are afforded mechanisms to practice privacy when releasing data while searching.
Firstly, the practice of *collecting and retaining metadata* introduces risks of independently verifying surveillance oversight measures and the potential loss of personal autonomy. Secondly, the more common practice of *collecting and distributing data for commercial purposes* validates expectations around the appropriateness of the data flow and the distribution of data, in particular the relative visibility and access of these data flows to secondary information providers (who users may not have explicitly authorised visibility and access of these data flows). Lastly, the *practice of framing choices through ineffective affordances* raises significant privacy risks particularly around trust with actors when releasing data. These risks are propagated when releasing data through ineffective mechanisms such as existing consent measures, technology lock-in, and lack of viable alternate choices.

In addition, the findings highlight privacy expectations are met when the interaction is of value to users, where it achieves an information-seeking need. This value is supported by meaningful affordances that construct increased awareness and understanding of the release of personally-identifiable information in search. Often, the invisibility of these data flows acted as a deterrent to the user's awareness and understanding in how personal data is used in search. The sensitivity of contexts is

influenced through factors such as extent of use, unintended recipients, time dependency (habitual vs. short-term interests). Detailed findings are presented in Chapter 5. It is imperative that appropriate affordances are available to enable users to exercise their privacy choices and expectations, because privacy in a hyperconnected world is no longer about accessing and controlling information, but how users want to share data and obtain value through those interactions (such as consuming information services). These interactions are often contextually constrained and are transactional in nature. These contributions identify the state of the art by providing an understanding of online privacy expectations and online social norms governing data release contexts.

1.2.3 Research Objective 3: Conceptual Dimension - Towards Designing Privacy Affordances to Manage Data Release for Users when Searching

The conceptual dimension is investigated in two stages. Firstly, in order to investigate the effectiveness of affordances, related work on affordances in socio-technical interactions was reviewed. Secondly, informed by this review and findings from the previous technical and empirical stages, design recommendations on managing data release in search are presented.

Research Question 3: How effective are existing affordances in addressing the information privacy expectations of users? To explore the effectiveness of existing affordances, existing literature and related work were reviewed on affordances offered to users in socio-technical interactions such as searching. These are documented in Chapter 2. Affordances typically offered to users include privacy-enhancing technologies and policy mechanisms (such as the Do Not Track initiative and the Australian Privacy Principles, with a particular emphasis on personal data use in the Australian context). From this critique of existing affordances, privacy-enhancing technologies are found to be reactive and serve a singular functional purpose. For example, if users do not wish to have their activity tracked and targeted, they can install an ad-blocker. However, with such affordances, there is no holistic integration of capabilities to protecting user identity or anonymity.

Policy mechanisms have expanded protections in order to meet increasingly novel uses of personal data in technology. However, some limitations remain. For example, the EU cookie notification mechanism (which aims to provide users with increased awareness about cookie usage when interacting with websites) is not highly effective. Studies

show users display blindness to repetitive notifications, once its novelty is lost. However, the websites are still functional, offering no actionable privacy protection other than being notified to opt-out. While these online organisations achieve compliance with regulatory measures, they fall short of empowering users with informed privacy choices.

These policy-centric affordances require respect for human values and human dignity when implemented (Floridi 2016), which at times, conflicts with the commercial interests of information service providers. In other words, a tension or asymmetry exists between commercial interests and, human values and dignity. Further efforts to regulate data use by information providers reduces an individual's ability to determine the appropriate contexts for use. Current efforts on data collection should expand to data release as personal data released by the user is often combined with behavioural data released through the interaction medium.

Research Question 4: How can we proactively empower users to effectively manage data release in search? In order to address this question, a philosophically-informed conceptual analysis was conducted to understand the values at stake for users. This was further supplemented by a moral mediation analysis to understand how technologies mediate interactions. Incorporating the findings of the survey study, privacy expectations are met when asymmetries in data flows are minimised, i.e. when users regain the capacity to determine their identifiability online by minimising data released.

Most approaches have focused on empowering users through educative initiatives. However, the impact of this approach is limited given the subliminal nature of data release and the complexities in data collection and use. In addition, for most users this approach is limited by the lack of access to expert knowledge and the complexities around which technology operates. This technology-driven approach also assumes that the onus of protecting privacy in terms of personal data release should rest on the user.

This research indicates that to proactively and effectively empower users, affordances must be provided in the design of technology that appropriately meet the diverse privacy expectations of users. It is in the act of using or interacting with a technological tool that users obtain its value. Hence, privacy-centric values should be embedded in the design of technology, for users to proactively exercise their privacy choices.

Although different types of technological solutions (e.g. privacy-enhancing technologies) exist, users still need affordances embedded in the design to exercise their privacy expectations and these affordances need to have values that are adaptable to changing contextual data flows and privacy expectations. Further, although commercial incentives drive interactions on the internet, societal values around upholding human dignity are important to preserving notions of privacy expectations of users. These needs give rise to *a key contribution*, a concept for understanding online privacy expectations. *Data release in personalised search* is found to be *contextually-public* by nature. Users expect to constrain the access and visibility of data flows to primary information providers. This would allow users to separate their entire identity to more of a role-based identity and minimise the severity of privacy loss. *Role-based identity affordances* offer users more control over existing data minimisation techniques because it offers a new way of thinking about and conceptualising online identities. Users do not represent themselves as a whole person (one single identity) when interacting online. Instead these contributions provide evidence to support that privacy expectations of users are more effectively aligned to role-based identifies where the integrity of the interaction is constrained to the specific context.

1.3 Thesis Outline

This thesis explores the mechanisms afforded to manage privacy expectations around data flows in the personalised search context from a socio-technical standpoint. To achieve this aim, the value sensitive design method is used to articulate recommendations for concepts and practice for technologists designing these affordances in systems (RQ4). These recommendations stem from an analysis of technical considerations (RQ1), user perspectives (RQ2) and critique of existing affordances (RQ3).

A review of literature (Chapter 2 – RQ3) explores existing work and related studies around privacy, personal data use and affordances in addressing privacy expectations (including technology and policy mechanisms). The methodology chapter (Chapter 3) outlines a mixed-method approach in investigating personal data flows in search, with a controlled experiment to explore the technical considerations of data release and a survey study to understand privacy expectations around personal data release in search.

The following three chapters (Chapters 4, 5 and 6) function as the key data chapters of the thesis. They address the technical, empirical and second-half of the conceptual dimensions respectively. Chapter 4 reports on the detailed experiment design and findings on implicit data release in search. Chapter 5 presents a thematic analysis on the privacy expectations of participants when searching. Chapter 6 discusses an approach towards integrating the findings from the controlled experiment on implicit data release and the survey study on privacy expectations, and explores how privacy-sensitive affordances can be embedded in the design of technology. Chapter 7 concludes the thesis by advocating for designing affordances that enable users to practice privacy online and outlines areas for future work.

1.4 Contributions from this Doctoral Study

Contributions arising out of this doctoral study are listed below to identify the contributions arising out of the work of the thesis and its development. In all of these contributions, I am the first author as I conducted and wrote the research. Co-authors listed are my supervisors who reviewed these papers, provided critique and editorial feedback.

Published Publications:

1. [Forthcoming in Conference Proceedings; Presented at Conference] Fernando, ATJ, Wahlstrom, K, Du, T & Ashman, H 2017, 'Designing Privacy Affordances for Searching', CEPE (Computer Ethics Philosophical Enquiry) ETHICOMP 2017 Conference, Turin, Italy, pp.1-16.

2. A Doctoral Consortium paper on a synopsis of my PhD research titled, “Privacy-enhanced Personalisation of Web Search” was presented at the Doctoral Consortium of the UMAP 2015 Conference (*the 23rd Conference on User Modelling, Adaptation and Personalisation*) in Dublin (29 June- 3 July 2015) and was published as part of the UMAP 2015 Conference Proceedings (Accessible here).

3. Presented a Conference Workshop paper titled, “Personalisation of Web Search: Exploring Search Query Parameters and Information Privacy Implications” at the 37th Annual International ACM (Association of Computer Machinery) SIGIR (Special Interest Group in Information Retrieval) Conference – *PIR (Privacy-*

Preserving Information Retrieval) 2014 Workshop in Gold Coast, Australia (11 July 2014) (Accessible here).

Science Communication:

Invited Talks:

- Presented a talk titled, "Exploring Personal Data Use and Privacy Implications" based on my doctoral research as an invited speaker for the South Australian branch of the Australian Institute of Professional Intelligence Officers Inc. (AIPIO) in conjunction with the Australian Institute of International Affairs (AIIA) in Adelaide (26 November 2015)

- Presented a public forum talk titled, “Internet technology that is inclusive without being invasive” at the annual Australian Internet Governance Forum event in Melbourne (6-7 October 2015). This was part of the .auDA Foundation showcase.

Community Engagement:

- Conducted a privacy awareness workshop at Kathleen Lumley College in Adelaide (11 June 2015) to gauge and raise awareness around online privacy.

2 Literature Review[2]

2.1 Introduction

This chapter presents a review of related work on the mechanisms users have to protect their privacy online. Specifically, the review focuses on the effectiveness of existing affordances in addressing the information privacy expectations of users (Research Question 3). Related work is reviewed when the corresponding research question is investigated in each of the data chapters that follow the methodology chapter (Chapter 3). Hence, related work pertaining to those sections are considered out of scope in this chapter, as the main purpose of this chapter is to explicate and justify the research gaps to be investigated in each of those data chapters.

2.2 Information Privacy

Privacy is a defining issue of the information age (Acquisti, Brandimarte & Loewenstein 2015). Historically, scholars have defined privacy in a multitude of different, but largely complementary ways. Privacy is commonly viewed as a right to having "an inviolate personality" (Warren & Brandeis 1890, p.216). This right is essential to protecting human dignity because it concerns the ability to be free from authoritative control and intrusion, and is distinct from protection of property or free expression (Floridi 2016; Mill 1885). It is important to protect human dignity because humans are distinct in that we are "endowed with consciousness, intelligence, mental life and self-determination" (Floridi 2016). Privacy is tightly intertwined with personhood, where the right to privacy "protects the individual's interest in becoming, being and remaining a person" (Reiman 1976, p.44). This notion of personhood gives rise to the ability of a person to manage the details and the nature of information communicated, such as ensuring confidentiality and trust in disclosure of data.

Another aspect of personality is accessibility. Based on this perspective, privacy is "the claim of individuals, groups or institutions to determine for themselves when and how, and to what extent information about them is communicated to others" (Westin 1968,

[2] Parts of the text in this chapter has appeared in Fernando et al. 2017 (ETHICOMP Conference Paper)

p.7). This claim to privacy may involve "three dimensions: *physical privacy* characterised as seclusions and solitude, *informational privacy* characterised as confidentiality, secrecy, data protection over personal information and *proprietary privacy* characterised as control over names, likenesses and repositories of personal information" (Allen-Castellito 1999, p.723). In other words, privacy is about managing the access people may have to you in the form of physical proximity, information or attention (Gavison 1980).

In managing access, boundaries between what is private and what is public are created. These boundaries are stretched across a continuum of accessibility ranging from privacy (i.e. being inaccessible) to publicity (i.e. being openly accessible) through a "process of selective concealment and disclosure" (Nippert-Eng 2010, p.3). A public sphere is a space where ideas and thoughts are articulated freely by individual people gathered together under a common societal purpose. From this public sphere arise social institutions that direct and guide social norms, values and purposes. A public sphere is located within a private sphere. A private sphere is where an individual is able to freely express his or her ideas and away from the coercive influences of power structures like governments and social institutions. Boundaries between these spheres are permeable and constantly in flux (Habermas 1962 as cited in Thomassen 2010).

The private and public dichotomy of data is a well-studied characteristic of privacy. Boundary management has been essential to privacy from historic times. For example, in biblical times, people chose to place tent openings so that no two openings faced each other, affording nomad travellers privacy in the desert (Tene 2008). In recent times, passwords demarcate access to restricted online spaces by verifying authorised users. The tension between privacy and publicity has been extensively reported (Nippert-Eng 2010; Palen & Dourish 2003). This comparative accessibility creates tensions between a person's desire to relate to both an individual and a group identity (Nippert-Eng 2010). Traditionally, privacy is highly guarded in private spaces whilst individuals are open to some aspects of their lives being visible to the public (Westin 1967).

These concepts about accessibility have been formalised as theories. The *control theory* links privacy with freedom on the basis that people should be in control of their own information (Westin 1970). By contrast, the *restricted access theory* is concerned with access to personal data being limited because it is shared at a pace users cannot control

and manage (Moor 1990). The *integrative privacy theory* is a merger of both the control theory and the restricted access theory.

Contrary to Westin's ideas on access and control, privacy is not akin to a lack of social engagement (Steeves 2009). In dynamically negotiating boundaries between public and private spheres, interactions are situated in social contexts as opposed to Westin's sole focus on information flows. Privacy is not asocial where an individual does not construct social isolation, but discloses information based on their needs. Privacy is also not anti-social because disclosure does not solely depend on the individual's control, but also on the recipient's trustworthiness and sensitivity to the potential disclosure (Steeves 2009). Privacy is relational as disclosures of personal data elude to a certain degree of intimacy underpinned by trust (Rachels 1975; Steeves 2009).

These boundaries are dynamically negotiated in real-life as people determine what is public and private (Nippert-Eng 2010). Choice is central to these negotiations. Current negotiations (opt-in and opt-out) are not dynamic. These negotiations manifest as practices (Dourish & Anderson 2006). Therefore, privacy is a practice: "The right to privacy is the right to existence of a social practice which makes it possible for me to think of this existence as mine" (Reiman 1976, p.43).

This research is concerned with *information privacy*, particularly as it pertains to personal data released when searching. Information privacy is concerned with data that personally identifies a person. "Personal data is data containing any identifying characteristics about a person" (European Commission 2009, 2015a). Personal data is constitutive, where information belongs to a person and may not be easily detached from their identity (Floridi 2006, 2010). For example, ""my" in "my data" … is the same "my" as in "my hand"" (Floridi 2016). Personal data can also be descriptive, where it could be data about a person as opposed to solely a person's own data (Nissenbaum 2016). For example, the information could be about you, but it can also be about the people you interact with such as a spouse, neighbour etc. (Nissenbaum 2016).

The effects of personal data released when searching may have negative consequences on a person's right to inviolate personality. Behavioural profiling, targeted advertising, linkability of data even when aggregated, excessive data collection, poor user awareness and control over personal data are all significant concerns (Feuz, Fuller & Stalder 2011; Solove 2006, 2007; Sweeney 2002; Tavani 2008; Tene & Polonetsky 2013; Zimmer

2008). Often the consequences of personal data misuse are reputational, instantaneous, far-reaching and permanent. The permanent nature of data collection enables identification easier (Warren & Brandis 1890).

Trust is important. People typically disclose personal information to those they deem trustworthy (Nippert-Eng 2010). When fears are realised, privacy expectations are violated, resulting in a loss of control and choice (Nippert-Eng 2010). In such contexts, people engage in actions that pre-emptively provide protection for future actions or that enable regaining some sense of privacy retroactively (Nippert-Eng 2010). This creates a premise for mechanisms like privacy-enhancing technologies that afford users protection.

Personal data stores focus on the notion that privacy is obtained through rights concerned with ownership of property (Thomson 1975). However, the right to privacy does not stem from ownership or property, but a right to inviolate personality (Warren & Brandeis 1890). The theory of *privacy as a commodity* is based on the idea that if our data is valuable, then we can sell it (Schwartz 2004). However, it then raises the implication that only people with substantial means can afford not to sell their personal information, which goes against the fundamental human right to privacy. More recent trends into managing data as a commodity assumes that users have a basic understanding of the value of their data and the necessary knowledge to technically enforce access and sale of their data.

Often, privacy is conceptualised as a trade-off, with utility vs. privacy interests being a common perspective in literature(Awad & Krishnan 2006). Users have a tendency to risk privacy for short-term gains, and often lack information to make an informed decision (Acquisti & Grossklags 2005). The nature of the trade-off deals with uncertainty as it is difficult to know the detailed information required for decision-making and this is further complicated given the irrationality of decision-making (Acquisti, Brandimarte & Loewenstein 2015; Acquisti & Grossklags 2005).

Recent theories of privacy focus on information flows, because individuals find value in sharing personal data, which brings a new dimension to the question of accessibility. The *ontological theory* is based on the idea that privacy is friction provided by technologies in the infosphere (Floridi 2011). The *contextual integrity theory* is based on the underlying notion that privacy varies depending on the context (Nissenbaum

2010). According to the contextual integrity theory, “we have a right to privacy, but it is neither a right to control personal information nor a right to have access to this information restricted, but it is a right to appropriate flow of personal information” (Nissenbaum 2010, p.127). “Informational norms specific to contexts govern the flow of personal information from one entity to another” (Nissenbaum 2010, p.232).

Contextual integrity acknowledges the multiplicity of social contexts. This concept is deeply entrenched in social theory. Nissenbaum talks about the notion of privacy in public, where distinctions are made of data arising in public spaces such as within institutions, in the intimacy of a person’s home and sensitive information arising from any interaction (Nissenbaum 2010). This resembles a public vs private dichotomy, where public information should be accessible to all and private information is not. However, there is a concept of privacy in public where for example shoppers in a mall may consent to be seen by other shoppers, but not to be extensively video-recorded by a person (Nissenbaum 2010). Similarly, spheres of mobility pertain to interactions within a space where a person is able to freely determine his or her interactions (Zimmer 2007). Web search is one such sphere.

Often arguments around data flows are used to portray that privacy does not exist online. For example, “If you have nothing to hide, then you’ve got nothing to fear”, where privacy concerns are said to be overblown (Zimmer 2008). This argument relies on the fear instantiated by large-scale tracking of people’s activities. These activities may be private, but not illegal or immoral and may or may not embarrass or be harmful to the individual if exposed in a public domain. People knowingly share information, so they should expect no privacy, where people know that they share information on a public platform, but may have no knowledge or control over the continued sharing of their personal information through other people or services (Zimmer 2008). Young people do not care about privacy, because they actively embrace new digital technologies and seem to enjoy its wide, shared use (Nissenbaum 2010). This notion does not hold true, as numerous surveys exist of young people being more tech-savvy and privacy-aware in combating online intrusions into personal information.

The trade-off of privacy vs. security (Nissenbaum 2010), where transmission principles and roles of people are defined to ensure that even in cases where personal data is used for security purposes, it is only accessible to authorised people and transmitted using

secure methods. For example, NSA surveillance programs showed that masses of personal data collection occurred regardless of authorisation, and highlights the importance of having prescribed roles for access in place to ensure accountability. Although users release data in interactions, they rely on the protections afforded in the interaction to protect their right to an inviolate personality. The protections afforded to users when searching are explored in the following section.

2.3 Affordances in Socio-Technical Interactions

Affordances determine what actions are possible between people and the environment (Gibson 1967). This original definition from ecology by Gibson constrained affordances to only what could be perceived. However, affordances can include aspects that go beyond perception (Norman 2013). "The term *affordance* refers to the relationship between a physical object and a person… An affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used... The presence of an affordance is jointly determined by the qualities of the object and the abilities of the agent that is interacting. But affordance is not a property... An affordance is a relationship. Whether an affordance exists depends upon the properties of both the object and the agent" (Norman 2013 p.10). The qualities of the object include both its functional properties and the values embedded into it which are then afforded to people during the interaction. In the context of this study, mechanisms to manage data release when searching form socio-technical interactions.

Often affordances are associated with signifiers. "Affordances determine what actions are possible. Signifiers communicate where the actions should take place" (Norman 2013 p.14). A common example includes doors, that are often confusing in their ability to signify if people should push or pull to enter through them (Figure 2.1).







"The Door of Zero Affordances" Flickr photo by cogdogblog is licensed under CC BY 2.0 US

"push" by Rohan Gupta from the Noun Project is licensed under CC BY 3.0 US

"pull" by Rohan Gupta from the Noun Project is licensed under CC BY 3.0 US

Figure 2.1 Typical Examples – Confusing Signifiers on Doors

The purpose of affordances is three-fold: to be perceived directly by users, to perform purposive actions, and to enable meaningful interactions (Kaptelinin 2013). The design of privacy settings affords users protections (see Figure 2.2). Some apps include components that go beyond their primary functional purposes. For example, Viber a messaging app does not require location services to function, but has an option to enable it which is usually switched on by default. Such signifiers are user interaction mechanisms which enable a data trail to be left through the interaction.

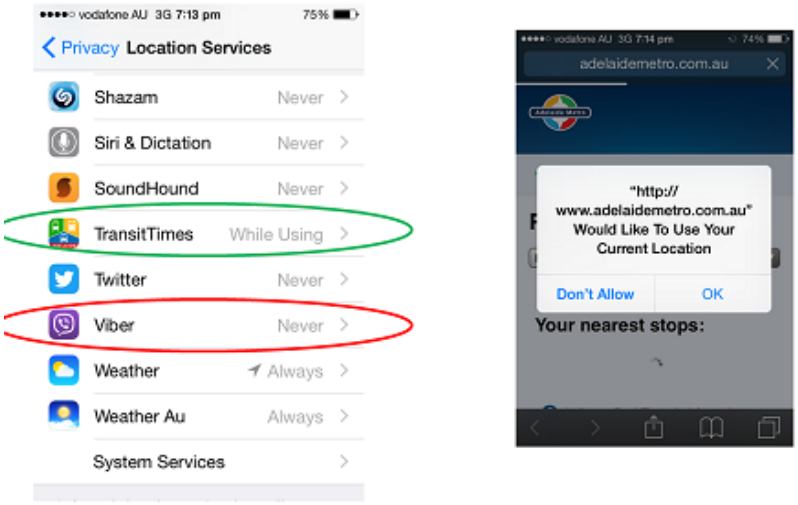


Figure 2.2 Design of Privacy Settings of Mobile Applications (Author's own images)

Affordances offered to users can include biases. For example, one privacy campaign, that aims to keep people informed about their data privacy, starts with its own data collection by asking to sign up to stay informed (see Figure 2.3). There are alternative ways to communicate information to users, but insisting that users sign up to access information is the norm. It is unclear how effective the number of signups are as a measure of disseminating information to a wider audience.

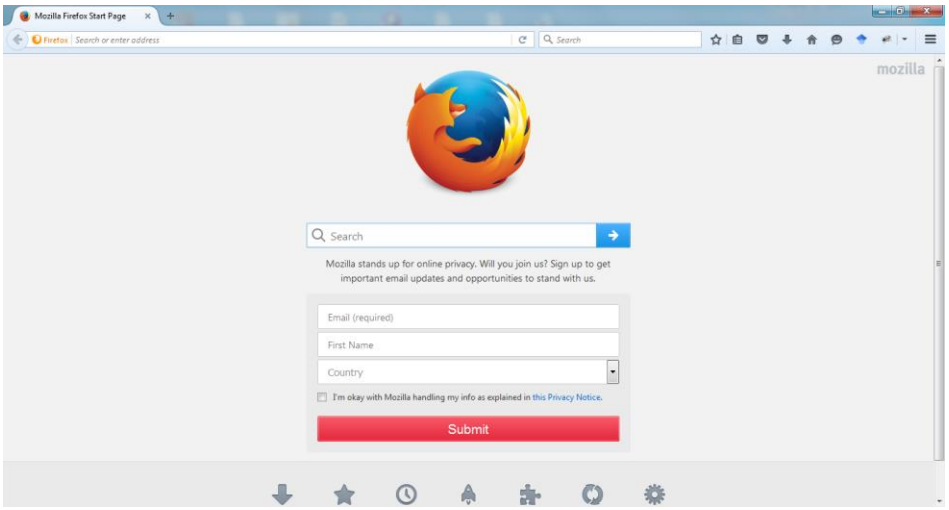


Figure 2.3 Mozilla's Privacy Campaign – February 2016 (Author's own images)

Affordances, however, are influenced by the mediums in which they are situated. Data flows through public and private spheres. A broader notion of affordances is required then to understand its uses and effects on people. Exploring mediated actions allows us to understand the opportunities for human action enabled through technology (Kaptelinin and Nardi 2012). Studies of mediated action primarily look at individual actions through an artefact as its focus. Collective mediated actions allow for understanding the socio-technical implications of technology use (Kaptelinin and Nardi 2012). These implications include the mechanisms through which people express privacy online. In this context, Norman's original definition of affordances is used to reflect the capabilities users are afforded through the socio-technical interaction. The effectiveness of privacy protections afforded to users when releasing data while searching is the focus of this review.

2.3.1 Gap in Affording Protection via Technological Mechanisms

Personalisation based on past searching history and habits may infringe on user privacy. Users may unknowingly be transmitting their personal data and thus no longer have control over their data. Factors affecting users' behaviour to disclose information include 'value of personalisation' where personalisation benefits must be what users want, 'knowledge and control over the user of personal information' where systems should elaborate on what details are stored and what options for users exist, 'trust in a website' which is based on positive past experiences, website design, reputation of website operator, presence of a privacy seal and privacy statement and benefits other than personalisation including 'financial rewards, social adjustment benefits' and

generally user disclosure behaviour is based on the trade-off between risks and benefits (Kobsa 2007).

Search engines collect large masses of user data. User data may be public personal information which is confidential and non-intimate in nature like place of work or it could be non-public personal information like medical histories (Tavani 2005). Google for example is said to collect user data automatically which are recorded in search logs when users type in search queries. These include information such as the search query, IP address, browser information, date and time of request, cookies that identify the browser, hyperlinks clicked, operating system, language, processor type, screen resolution and colour depth, active plug-ins etc. (Aljifri & Navarro 2004; Google 2013a). Cookies alone may not be effective in personalising as cookies are browser-dependent and a computer may be used by more than one user or a user may use multiple computers, thereby an inaccurate or incomplete user profile may be created (Aljifri & Navarro 2004). Alternatives such as 'flash cookies' (local stored objects) behave differently to normal cookies and can be managed independently of the browser through the Adobe flash security setting (Adobe 2013).

Privacy-enhancing technologies (PETs) afford users confidentiality or anonymity of personal data when searching (Gurses 2010). These technologies are used to provide protection for particular dimensions of privacy such as confidentiality and control (Danezis & Gurses 2010). For example, data confidentiality is provided through cryptographic techniques like k-anonymity and differential privacy. Other PETs provide anonymity when data is in transmission over communication networks like Tor or when stored and utilised in databases such as privacy-preserving data publishing techniques (Danezis & Gurses 2010).

Another category of PETs aim to provide users with the control and ability to determine disclosure of data. Most of these PETs although providing users with some level of control, are functional and not integrated to the user's experience of searching. The burden is placed on the user to understand the complexities of the information being released and take action. Further these PETs are removed from the social context in which the interaction is situated (Danezis & Gurses 2010).

A subset of privacy-enhancing technologies focus mainly on personalised search. PETs in the domain of personalised search may include use of pseudonymous user identifiers

and user models, client-side personalisation, techniques that employ collaborative filtering including distribution, encrypted aggregation, perturbation and obfuscation and privacy solutions specifically catering to individual user needs (Kobsa 2007). For example, disclosure is explored in context-aware recommender systems to uncover privacy decision-making insights (Knijnenburg & Kobsa 2013). However, these techniques do not explore the social context of the interaction, where privacy is viewed as a practice.

From this analysis of related work on embedding privacy-by-design principles in data protection regimes, a number of design considerations for investigation emerge. Currently, there is a lack of visibility of personal data from a user's point of view in particular pertaining to awareness and understanding of how personal data flows in online contexts. Existing approaches (like PETs) function in silos serving a particular functional or technical purpose (refer Table 2.1). This may be the reason that it is often easy to build technology or code software to cater to a concrete functional requirement as opposed to embedding privacy values which are often fluid in nature.

In such instances, a value sensitive design approach embedding privacy-by-design principles will enable the exploration of privacy expectations and subsequently privacy values in a particular context of personal data use. It is also important to note the privacy-by-design principle of being proactive and not reactive in terms of personal data release (Cavoukian 2009). The implications of personal data release may be instantaneous with reputational damage and social harm to a person's identity occurring without their awareness. Further, the ineffectiveness of privacy-by-design principles lies in the conflicting interests between data collection practices and privacy practices in businesses (Rubinstein & Good 2013). Inculcating an organisational culture which supports building privacy-by-design principles into the product is important (Koops & Leenes 2014).

In summary, technical mechanisms that address users' privacy expectations are reactive and serve a functional purpose. There is a disconnect between managing data flows and managing appropriate data flows. This leads to major gap: What are the effects of data release on personalised search results? This is investigated in Chapter 4.

Table 2.1 Overview of Technical Privacy Protections Afforded

Mechanism	Example	Purpose	Placement	Protection Afforded
Ad-Blockers **Script Blockers** **Cookie Management**	AdBlock NoScript BetterPrivacy	Prevent release of data	Browser-based but user has to search for and install to make it a part of their experience	Identity Management – Against Invasive Data Collection
Anti-Tracking	Ghostery PrivacyBadger			Identity; Expression of Privacy needs
Manage Data Leakage	NoTrace			Identity Management
Anonymiser	Tor	Anonymises or obfuscates data release		Anonymity
Obfuscates	TrackMeNot			Obfuscation of search queries
Anonymiser/ Obfuscates	Private Web Search, Milk			Anonymity with obfuscation of queries, cookies
Anonymising VPNs	Disconnect Search			Anonymity/ Identity
Decentralised architecture to manage digital identities	Databox Solid	To extract personal data stored in applications and provide users with ownership		Control over personal data by separating and storing personal data at the client-end with the aim of the user releasing it for authorised applications.
Allows tracking protection at client-end	Private Browsing in Mozilla Firefox	To prevent saved history, searches, cookies, and other temporary files		Anonymity among shared users of accessible devices
Visualises data leakage in terms of actors	Lightbeam	Allows users to visualise their interactions with actors online		Raises awareness of the extent to which data released is shared among actors, both first-party and third-party sites.
User Account-based control of data	Google Dashboard	Allows visibility of data collected while logged in.	Tied to User Account	Manage control over data and account while using services when logged in.

2.3.2 Gap in Affording Protection via Societal Mechanisms

Societal mechanisms that exist to afford users privacy protection are described in this section (refer Table 2.2). The contextual integrity theory argues that the current approach to online privacy is based on a ‘notice and consent’ or a ‘default opt-in’ approach which fails to recognise that users may not completely understand the facts in the instance of providing consent to use of their personal information. Users may opt in simply because opting out prevents them from using the online service and it is a trade-off made so that they are able to access the service. Such a trade-off does not provide the user with a genuine choice or with control over personal information.

A key concern is that users may not be aware that their personal user data is collected and how it is being used (Brandi & Olivier 2010). Users may have also little control over how their personal data is being used and how they may retrieve and eliminate personal information (Tavani 2005). This could be one of the reasons why there is a low uptake in adopting PETs relating to Web search. Users may not have an actual choice in using search engines, as opting not to use search engines results in the inability search and retrieve information (Nissenbaum 2010). Also, the privacy policies of search engines are usually concise to avoid lengthy details, but in doing so may be vague and cover a broad spectrum of areas. Hence, in such a context, people do not know what they are agreeing to and as such privacy policies fail in adequately addressing how user data is captured and used (Nissenbaum 2010). The language used in these privacy policies may also not accurately convey meaning with terms like user profiling not used (Aljifri & Navarro 2004). Therefore, this may create a false sense of security in enticing users to opt-in.

Table 2.2 Overview of Societal Privacy Protections Afforded

Mechanism	Example	Purpose	Placement	Protection Afforded
Policy Initiatives	DoNotTrack (DNT) Cookie Notice (EUROPA 2016)	To prevent collection of data used for targeting and tracking individuals when they express a desire to do so	Browser-based	User consent preferences are captured but are not technically enforced
Consent and Disclosure	Notice and Consent, Informed Consent, P3P	To inform users about data collection P3P (Platform for Privacy Preferences) provides information on data collection by websites.	In application	Implementation of a social contract in the online world.
Protection afforded by law	Australian Privacy Principles	Outlines best practices for collection of personal data that organisations have to comply with	Information service provider/ organisation	Australian privacy principles do not afford Australians privacy as a human right, but instead it focuses primarily on compliance
Design Principles	Privacy-by-Design Principles	To design mechanisms that embed privacy-centric values when data collection occurs	In application	Provides mechanisms to manage data release if practiced.

Most online services offer assurances in terms of anonymised data, but users can be recognized from usage trends alone, even without contextual information or content (Ars Technica 2013). The theory outlines that context-specific norms should be adopted, so that for example websites are limited with respect to what user information is collected, shared and the conditions of use. Information is more accessible, easily transferable and shared and can be easily aggregated and analysed than before because

of technologies (Nissenbaum 2010). Technology improves the quality of functions in daily life. It entices people to use it and privacy is usually traded off against the capability to use it (Zimmer 2008).

National legislation, such as the EU Data Protection Directives and the Australian Privacy Principles (APPs), mandates collection of personal data solely for the purpose of conducting a business-related transaction or a service. Limitations exist with respect to use, maintenance, transfer and quality of data. In personalised Web search, the basic presumption when users search online using search engines is that users agree to the terms of service this action entails. However, privacy policies are usually ambiguous and vague or users may not fully understand the conditions of use to which they readily agree to (Nissenbaum 2010). In most cases, if users wish to opt out, they encounter a denial of service rather than the option to leverage their consumer's choice. Public perception has heightened in recent years in the wake of surveillance scandals. Even so, people often trade security measures or privacy considerations for the ease of use of a service, the convenience it entails, the accessibility it provides and the benefits of sharing personal data with others in a community.

People should be able to actively control personal information and know how non-personal information about them is being used. Privacy requirements that should be upheld when data is transmitted include transparency, openness, notice and consent (Office of The Australian Information Commissioner 2016). The Australian Privacy Principles have significantly changed from its earlier form as National Privacy Principles to include specific privacy protections in collecting and managing personal data, and in maintaining its integrity and accuracy. However, the APPs do not afford Australians privacy as a human right, but instead it focuses primarily on compliance. While the privacy commissioner's mandate has been advanced to enable protection, the current Australian Privacy Principles are weak in relation to transfer of data across jurisdictions, credit reporting and does not adequately enforce accountability, or the choice to be anonymous (Waters and Greenleaf 2012).

The European Union's Data Privacy Directive has similar standards on how personal data must be collected and processed and transparent privacy policies where explicit user consent should be obtained (European Union 2011). The Data Privacy Directive (Directive 95/46/EC) outlines specific guidelines in relation to notice, purpose, consent,

security, disclosure, access and accountability of data collected and processed. The Directive on Privacy and Electronic Communications (Directive 2002/58) was amended by the EU Cookie Directive and The Data Retention Directive. The EU Cookie Directive (Directive 2009/136/EC) calls for websites to acquire informed consent from users before any information is stored on any type of electronic device. This directive aims to raise awareness of cookies used to track user activities on websites. The 'Track Me Not' initiative is as a result of this law, where organisations need to explain the cookies used on their site and obtain explicit user consent for the use of cookies (European Union 2009). The Data Retention Directive (Directive 2006/24/EC) states that telecommunications data such as an IP address can be held for 6 to 24 months where police personnel can access such data using court orders in the European Union (European Union 2006). Similarly, the Regulation of Investigatory Powers Act (RIP 2000) in the UK ensures that all ISPs need to collect and store all data going through them (United Kingdom 2000). Therefore, anonymity is not fully possible. In comparison to the Australian Privacy Principles, the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) provides protection against algorithmic decision-making by providing a case for a 'right to an explanation' by claimants (Goodman and Flaxman 2016).

The lack of awareness or ignorance by users of the value or importance of behavioural tracking through capture of personal data on a routine basis, may lead to personal or sensitive details gleaned from aggregated data. This loss of privacy through data leakage may potentially lead to misconceptions or inaccurate deductions about an individual if taken out context. As is the case often, perception about such an entity (be it an organisation or an individual) may change and the damage is often reputational and instantaneous, and difficult to repair. Alternatively, personalised search results may contribute towards a bias in perception of search results and may influence a person's selection of results. Users may receive results closely associated with their browsing history and captured preferences. These captured preferences may sometimes not reflect an individual's sole needs in the case of a shared device. Information needs and user preferences change over time. Both these aspects may distort filter bubbles present in personalised search (Pariser 2011).

In summary, the ineffectiveness of these societal protections lead to key questions: what are people's preferences when searching? Where do they find value? This study aims to

understand: What constitutes appropriate flows of information and the privacy expectations that underpin these data flows? A detailed investigation is presented in Chapter 5.

2.3.3 Gap in Affording Effective Protections for Data Release

Data leakage is a pervasive problem due to the growth in the number of user data sources, the integration of numerous data sources in large corporate contexts and the increasing capability of technologies to process this data in real time (Kobsa 2007). The accessibility of both user data and the ability to process it using sophisticated technologies without having specialised knowledge further challenges the boundaries of access to personal data (Nissenbaum 2010). Low levels of user awareness and poor control over collection of personal data lead to growing implicit data leakage.

The ability to make accurate inferences between two disparate groups of user data even if it is putatively anonymised is a cause for concern (Sweeney 2002). Linkability of queries in a user's search history can be used to identify users (Barbaro & Zeller 2006). Behavioural profiling, using data collected through persistent tracking mechanisms, is used to customise outcomes like Web search results (Toch, Wang & Cranor 2012). Targeted advertisements leak personal user information to not just the advertisers, but also to any party that has access to those ads (Castelluccia, Kaafar & Tran 2012). The large-scale aggregation of longitudinal user data enables behavioural advertising with the potential for misuse of data or possible misinterpretation of a person's profile as the data would contain only those aspects of a person that are revealed. These implications contribute to a potential loss in anonymity, integrity and control of a user's identity on the Web.

Loss of personal data has a substantial impact on user privacy. Routine leakage of data may be aggregated to reveal personally-identifiable information about a person, even though originally the individual bits of data may not be by themselves personal or sensitive (Teltzrow & Kobsa 2003). Privacy in the broader context of this research study is not just limited to individual privacy (which is ever-changing and fluctuates depending on the scenario), but is also an important aspect to consider from the perspectives of national legislation or corporate communication restraints.

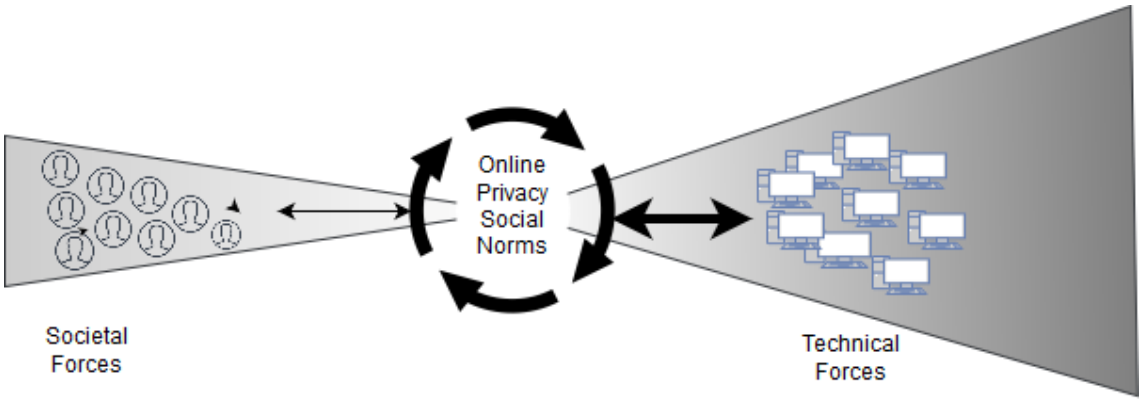


Figure 2.4 Inequality in Socio-Technical Forces

Overall, we need to understand how to effectively empower users to manage data release when searching. The existing mechanisms do not treat the problem, just the symptoms. Advancements in search enable access to information in a faster and efficient manner solving existing information and power asymmetries around access to information. However, because of the nature of data collection, and the targeting and tracking mechanisms used, the mechanisms and practices in place create and propagate information and power asymmetries. This disconnect creates an inequality as visualised in Figure 2.4. This reiterates the need to further understand the effect of data release on personalised search results. To what extent, are the mechanisms and the protections afforded to users effective? The study focuses on data release because it stands to provide more protection from a user's point of view. Chapter 6 explores this gap in further detail.

2.4 Conclusion

A review of the effectiveness of existing affordances users have to manage data release when searching is presented in this chapter. In summary, privacy is a practice where users exercise their right to an inviolate personality. In the context of searching, the protections afforded to users often have a functional purpose and place the burden on the user when attempting to manage release of data aligned with their privacy expectations. The study will investigate the effects of data release on personalised search results, the privacy expectations underpinning this interaction, and explore how the mechanisms afforded to users can be improved upon to bridge the inherent inequalities in data release mechanisms.

3 Methodology

"Value sensitive design is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman, Kahn & Borning 2009, p.69). The approach employs technical, empirical and conceptual investigations to study the effects of a phenomenon (i.e. privacy expectations and the effectiveness of privacy protection affordances) in a given application domain (i.e. data release in personalised search).

Privacy is a very diverse concept with different points of view relating to privacy that people associate with personalised search. In attempting to understand privacy expectations and the effectiveness of protections afforded to users with scholarly rigour, value sensitive design is appropriate as it allows investigation of the problem from technical, empirical and conceptual dimensions to uncover the values at stake for users and information service providers. A generic overview of the methodology of this research study is presented in this chapter. Detailed methodological considerations are described in the relevant data chapter for each phase.

3.1 Approach

The study explores the effectiveness of mechanisms affording users privacy protection around data release when searching in three phases. Mixed methods are used to apply the value sensitive design framework. Mixed methods use a combination of qualitative and quantitative methods to obtain varied data when investigating a problem gap, as diverse types of data will aid in understanding the research problem (Creswell & Plano Clark 2011). Mixed methods would also capture viewpoints from a range of different stakeholders, providing a multi-dimensional angle to understanding different phenomena in a context (Barbour 2008). For example, in this research, quantitative techniques (i.e. a controlled experiment and survey) were used to collect and analyse data, and was further supported by qualitative data analysis techniques (i.e. thematic analysis of privacy expectations gathered and conceptual analysis techniques via stakeholder values and moral mediations analyses) Refer Figure 3.1 for more details.

Traditionally value sensitive design has been applied commencing with a focus on conceptual or empirical investigations, followed by a technical investigation re-

designing a technical artefact. This thesis applies value sensitive design in a flipped-order because of the nature of the problem. There is a multitude of existing affordances, but an understanding of the extent to which these affordances effectively address privacy expectations associated with data release when searching is lacking. Hence, value sensitive design is used to guide this exploration of this overall research problem. Firstly, the effectiveness of existing affordances was reviewed in Chapter 2 (Literature Review) which represented the first part of the conceptual investigation. Secondly, the effects of data release in personalised search were explored via a controlled experiment under the technical investigation phase (Chapter 4). Thirdly, to build an understanding of privacy expectations users associated with data release is investigated via an empirical survey study (Chapter 5). Lastly, findings from technical and empirical investigations are synthesised through a conceptual investigation of stakeholder values and moral mediation analyses and recommendations on effectively designing affordances that address privacy expectations associated with data release when searching are presented.

In phase one – *technical investigations*, a controlled experiment design identified the effects of data release on the variance of personalised search results and to derive if these effects impacted use of any personally-identifiable information. "Technical investigations focus on the design and performance of the technology itself." (Friedman & Freier 2005, p.369). The goal of this phase was to explore whether relevant search results can be obtained by minimising the level of personal data released through search query requests. Quantitative methods refer to examining subjects through use of mathematical, statistical or computational methods. Experimental research, a type of quantitative method, aims to identify if specific conditions impacts an outcome (Creswell 2009). A controlled experiment is a type of experimental research design.

The controlled experiment design captured a sample of search query parameters for defined search scenarios using an HTTP intercepting-proxy tool. Search requests were then automated to compute the variance between two search results (minimum data released as a baseline vs. iterated combination of parameters based on the scenario). The variance was computed using two measures namely Edit Distance and Jaccard Index for each search result comparison. This automated computation of variance was run across commercial search engines like Google, Bing and Yahoo and privacy-enhanced search engines like DuckDuckGo, StartPage, Ixquick and Unbubble in three distinct

timeframes (October 2014, May 2015 and January 2016). Additional controls for variables affecting variance such as for the user-agent parameter, and controlling for sources of noise were also embedded into the design. The overall results were analysed statistically using Wilcoxon Signed Rank Test to demonstrate the significance of personalisation variance and Non-Metric Multidimensional Scaling to identify the similarity or dissimilarity of variables affecting personalisation variance. This analysis contributed the finding that minimised or obfuscated data release has a negligible effect on personalisation variance. In the few instances where minimisation or obfuscation is not possible, implicit data release should be made visible to users.

The second phase of the research – *empirical investigations,* sought to build an understanding of the privacy expectations people associate with searching. "Empirical investigations focus on the human response to the technical artefact, and on the larger social context in which the technology is situated" (Friedman & Freier 2005, p.369). This objective was achieved through a combination of a survey instrument that gauged expectations quantitatively and qualitatively and also incorporated user evaluations of interactive search tasks involving obfuscated search results. Privacy expectations expressed qualitatively were analysed using the Contextual Integrity framework. Practices involving release of data that impact privacy concerns were subsequently derived. The main contribution of this phase was to analyse user expectations towards privacy when searching and to validate if their behaviour matches their level of concern or intention. This validation was done using the quantitative and qualitative questions and the behaviour reported in the interactive search tasks. The survey obtained demographic information from the user sample, rich quality data on privacy concerns and personalisation preferences and a quantitative measurement of search behaviour and privacy expectations. Statistical analysis of data from questionnaire using 5-point Likert scales measured the level of expectation based on privacy constructs derived from literature.

Qualitative methods aim to understand user behaviour and reasoning behind interactions in specific contexts (Creswell 2009). Qualitative insights provide researchers with rich data for a robust comparison of factors (Barbour 2008). In this study, the qualitative aspect helps to understand privacy expectations, whilst the quantitative aspect adds a different dimension to the data by assessing the effects of implicit data release on search results and expectations around privacy concerns relating to data release such as data

misuse and disclosure. Both these methods are ideal in this instance as the research study is exploratory and it allows the researcher to understand how different users perceive personalised search results and the use of their personal data in this process.

The use of the Contextual Integrity framework as a guideline for evaluating privacy expectations will benefit from having foundations in scholarly rigour, will better suit peoples' needs and would help bridge the gap between privacy theories and technologies. Application of the framework would help reinforce the principles of the framework and would make it more robust. Privacy is a social construct with different people having different attitudes towards privacy. These differences could be identified through assessment of the varying levels of expectations associated with privacy through the Contextual Integrity framework. Application of the framework in this context is also suitable as search queries and personalisation build a context of user details over time and involve a number of actors playing different roles, information types, and information collected, shared and aggregated under different transmission principles.

In the third phase – *conceptual investigations*, data from the technical and empirical investigations were used as a basis to inform the design of effective privacy affordances for data release. "Conceptual investigations comprise philosophically informed analyses of the central constructs and issues under investigation" (Friedman & Freier 2005, p.369). The Contextual Integrity framework has a largely functional purpose in investigating the appropriateness of data flows. However, to further understand the context and the extent of information and power asymmetries that underpin the gap between user privacy expectations and the effectiveness of privacy protections afforded to users, stakeholder and moral mediation analyses are used.

Stakeholder analysis was used to consider the values of interest to key stakeholders. This conceptual analysis was further augmented by employing mediation analysis and a moral assessment of the interaction because stakeholder analysis does not incorporate how values are mediated by technology. Exploring technology mediation was important because expression of privacy expectations is shaped by the interaction medium. Properties of the interaction medium consisting of the Internet, the Web, the Web browser and the search engine were included as part of the assessment for the existing product-in-design that users rely on when searching.

From this analysis, a concept for privacy expectations in socio-technical interactions was identified, where data release is contextually public by nature. Online social norms are influenced by the appropriateness and distribution of data flows, and the motivation and values of information service providers. The nature of the socio-technical interaction is relational, where design practice should include release of data necessary to the purpose of the interaction where possible, separation of the core identity from the relational identity, and shedding data due to temporal or local changes in user interests. These design initiatives can be scaffolded by instilling a privacy mindset where data release is visibility perceived, accountable data protection through effective protection mechanisms and creating incentives for information service providers based on social capital to maintain the integrity of socio-technical interactions. An overview of the overall methodology including techniques and contributions is visualised in Figure 3.1.

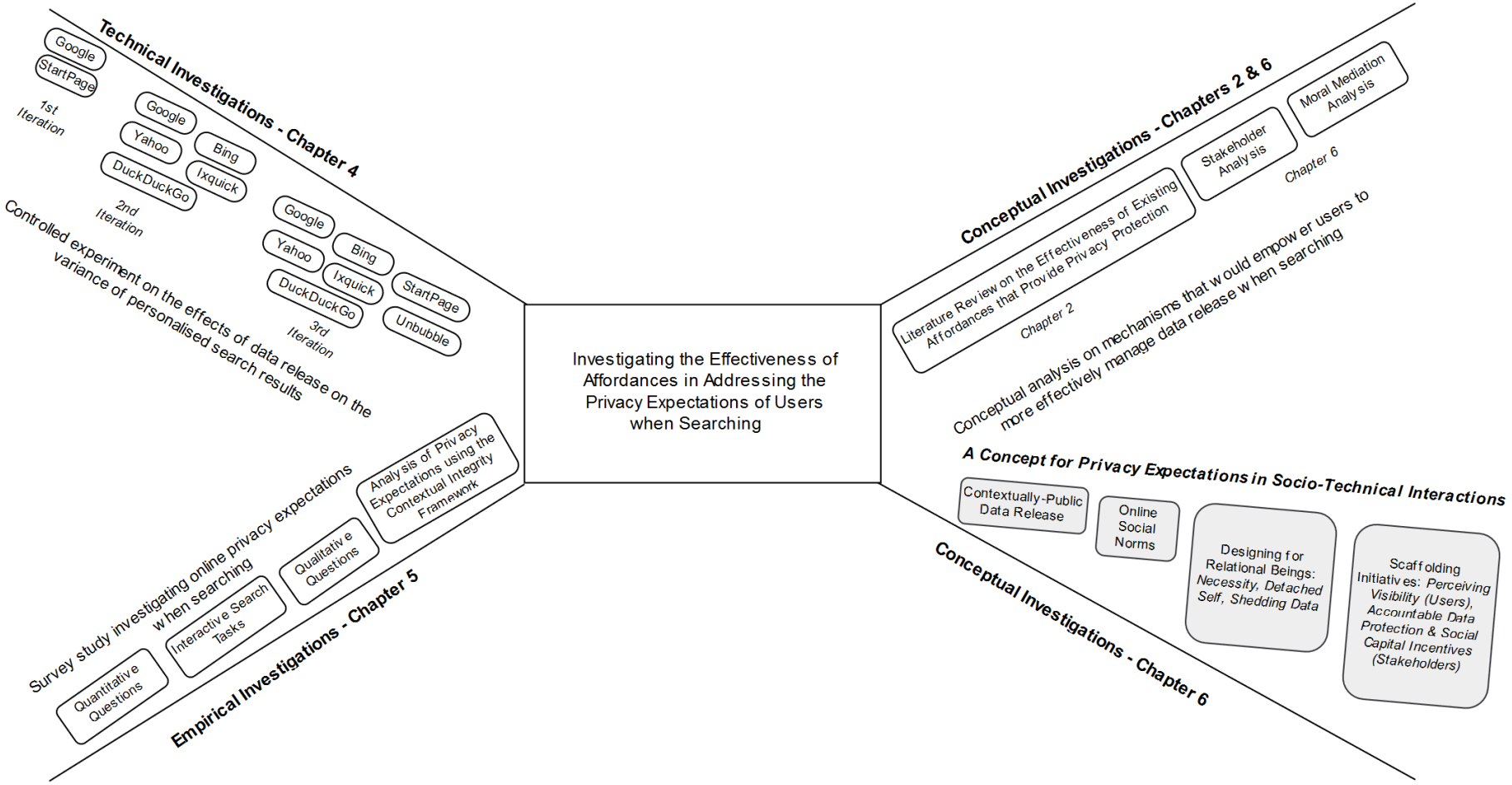


Figure 3.1 Overall Methodology: Techniques and Contributions

3.1.1 Technical Investigations: Methodology

The goal of the experiment is to identify the effects implicit data has on the variance of personalised search results. To achieve this, a controlled experiment was designed to explore the flow of implicit user data through the HTTP channel, consisting of three key steps (see Figures 3.1-3.3). Firstly, a sample of search query parameters was captured for each defined search scenario across each search engine and query term using an HTTP intercepting proxy tool (Fiddler). Secondly, search requests were automated (for each query term across each defined search scenario and search engine) to compute the variance between search result lists in different search scenarios using the Edit Distance and Jaccard Index measures (for each query result comparison retrieved). Thirdly, the personalisation variance comparison datasets were analysed to identify factors significantly affecting the variance of personalised search results using the statistical measures – Wilcoxon Signed Rank Test, Non-metric Multi-Dimensional Scaling and Generalised Linear Modelling. A summary of these controls and the overall experiment design is provided in Table 3.1.

This section describes the experiment design including details on how query terms and search query parameters were identified, and how search scenarios for each search engine and controls for sources of noise were defined. This section concludes with a description on the experiment methodology with a description of the personalisation variance measures (Jaccard Index and Edit Distance) and the statistical analysis measures (Wilcoxon Signed Rank Test, Non-metric Multi-Dimensional Scaling and Generalised Linear Modelling) used and its application in the context of the controlled experiment.

Experiment Design: Experimental research is a type of behavioural research that explores the causation between variables, typically through use of a controlled experiment method (Lazar, Feng & Hochheiser 2010; Rosenthal & Rosnow 2008). The experiment was designed to reflect the search behaviour of a typical user and explored how data is released in the information-seeking activity of obtaining personalised search results on the Web. These activities are explained in Figures 3.1a, 3.2a and 3a. Data is implicitly released through search query parameters including HTTP headers, GET/POST parameters and cookies, and explicitly through search query terms as shown in Figure 2. A summary of the experiment design is provided in Table 3.1.

Table 3.1 Overall Experiment Design

Category	Feature	Description	Example
Independent Variable	Mechanisms of Data Release: Query Terms, Experiment Conditions	Informational, Navigational and Transactional Query Terms	Refer Table 4.2
Dependent Variable	Effects of Search Query Parameters, Query Terms and Location on the Variance of Personalised Search Results	Randomly occurring based on account and non-account login search scenarios	Refer Table 4. 3 for examples for commercial and privacy-enhanced search engines
Experiment Conditions	Search Scenarios for Personal Data Release	Four common scenarios for searching defined based on account-login and non-account login methods	Refer Table 4.3
	Comparing personalisation variance across two search result lists.	Defined two search result lists: A *minimum* set of query parameters (control) vs. a *comparison set**	*A *comparison set* could be of three types: Each *Individual* Parameter, *Default* (set of parameters), *Location-Spoofed* (i.e. a default set of parameters with the location-spoofed)
	Search Engines	Commercial Search Engines and Privacy-enhanced Search Engines	*Commercial* -Google, Yahoo, Bing *Privacy-enhanced*- StartPage, Ixquick, DuckDuckGo, Unbubble
Experiment Run	Experiment data collected at three distinct time points with a gap of 7 months. This allowed a longitudinal view of the data.	The experiment design was optimised with each iteration	*1st Run - October 2014* – Google and StartPage *2nd Run - May 2015*- Google, Yahoo, Bing, DuckDuckGo and Ixquick *3rd Run – January 2016*- Google, Yahoo, Bing, DuckDuckGo, Ixquick, StartPage, Unbubble; Further experimental controls using Google as a lens
Randomisation of Experiment Run Order	Inclusion Effect	Systematic addition of search query parameters to search request	To identify if a particular combination of search query parameters has an effect on personalisation variance. This is the default type performed across all 3 runs.
	Exclusion Effect	Systematic removal of search query parameters to search request	To identify if the absence of a particular search query parameter influences personalisation variance. This control was performed in the 3rd run.

Category	Feature	Description	Example
User-Agent Control	To identify personalisation variance influenced by the User-Agent	Use of multiple values for the user-agent parameter to determine reliable significance effecting personalisation variance. This control was performed in the 3rd run, with the default being tested across all runs.	Default User-Agent Parameter was: *Mozilla Firefox* [Mozilla/5.0 (Windows NT 6.1; rv:32.0) Gecko/20100101 Firefox/32.0] Tested Controls were *Google Chrome* [Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2228.0 Safari/537.36] *Internet Explorer* [Mozilla/5.0 (Windows NT 6.1; Trident/7.0; MATP; rv:11.0) like Gecko]
Geolocation Control	Location-Spoofing the minimum query set comparison	To test the effect location has on personalisation variance of search queries	Use of an Amazon Web Services proxy server hosted in the US. Default location value for queries is Australia (dynamic IP addresses). This control was performed across all 3 runs.
Controls for Sources of Noise	Carry-over Effect	Each manual default search query was performed on a clean fresh instance of the browser with a gap of at least 11 minutes to minimise the carry-over effect	
	Stability of Search Results	Three datasets of personalisation variance across a sample of commercial and privacy-enhanced search engines were collected during a period of two years with a gap of 7 months in between each data collection point. This longitudinal view acted as a control for plausible variations in the stability of results.	
Privacy-Sensitive User Controls	To investigate only implicit data leakage through search queries	Use of independent search requests by removing the association among subsequent queries. Use of cURL and disabling JavaScript to limit dynamic tracking, collection of clickthrough data and Web bugs. This allowed personalisation variance to be attributed to implicit data release through search queries.	
Personalisation Variance Measures	Jaccard Index	Provides a measure of the level of reordering between two similar lists.	See section 4.3.6 on Measuring the Variance of Personalised Search Results
	Edit Distance	Provides a measure of the differences between two pairs of similar lists	See section 4.3.6 on Measuring the Variance of Personalised Search Results
Statistical Analysis Measures	Wilcoxon Signed Rank Test (non-parametric)	Demonstrates significance of personalisation variance through identification of statistically significant median increases	See Figure 4a and related description
	Non-metric Multidimensional Scaling (NMDS)	Identifies the similarity or dissimilarity of a personalisation variance measure	See Figure 4b and related description
	Generalised Linear Modelling (GLM)	To identify and attribute the variation in the response variable (personalisation variance for query terms, and personalisation variance for query parameters) to the explanatory variables (factors such as search engine and query type, or search engine and query parameter respectively). See section 4.3.6 on Analysing the Variance of Personalised Search Results	

For scope, the focus was on the data released through the process of an initial search request, i.e. the first interaction the user has when seeking information by querying the search engine. Hence, to identify the effects implicit data has on personalised search results, data flows were controlled by selectively releasing the query terms, the implicit data (search query parameters), and the scenarios through which users search in iteration.

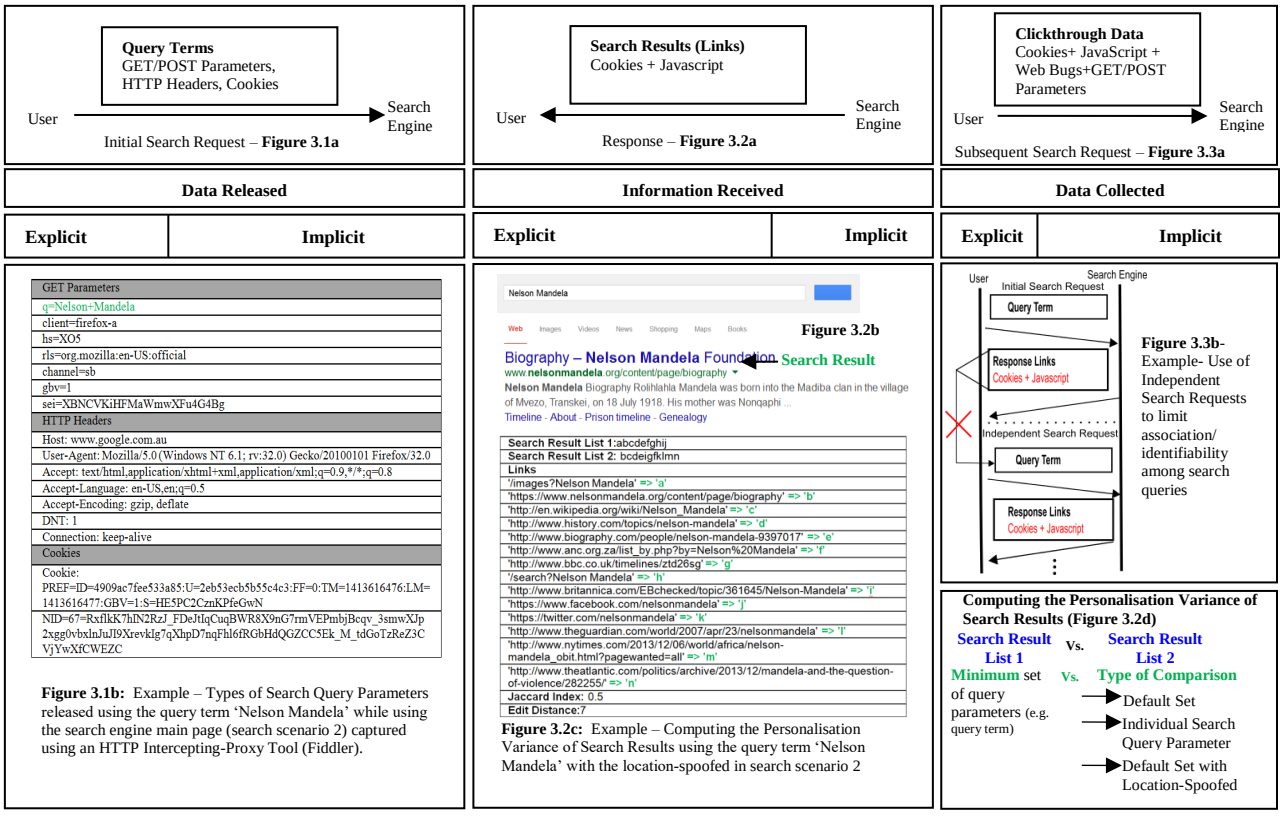


Figure 3.1-3.3 Experiment Design Conceptualisation of Data Flows in Personalised Search from a Privacy-Sensitive User's Perspective (An expanded view is provided in section 4.3.1)

3.1.2 Empirical Investigations: Methodology

"Empirical investigations focus on the human response to the technical artefact, and on the larger social context in which the technology is situated" (Friedman & Freier 2005, p.369). Privacy expectations of users around data release in search were explored in this phase.

Current available material around privacy concerns and risks of populations largely comprise of generic community attitudes such as the 2013 Office of the Information Commissioner's Privacy Attitudes Survey of 2013 (n=1000) or the 2015 Victorian survey

on attitudes about privacy and data protection by the Office of the Commissioner for Privacy and Data Protection and the survey on privacy concerns around Internet use (n= 965) (Arnott & Andrejevic 2012; Office of the Australian Information Commissioner 2013; Victorian Government 2015). However, these surveys do not explore the effects of privacy in personalised search and contributes to a survey of privacy expectations from a South Australian population.

The design of the survey was based on a review of literature and related work on information privacy concerns and risks (see Section 2.2). Quantitative questions were modelled on questionnaires from related work (Earp et al. 2005; Office of the Australian Information Commissioner 2013). The literature review revealed no pre-existing questionnaire to the best of the author's knowledge, that explored user privacy expectations around data release when searching. Therefore, a specific survey instrument was designed.

Care was taken to ensure reliability (i.e. the extent to which the findings can be successfully replicated) and validity (i.e. the extent to which the findings accurately reflect reality) of the survey instrument (Lazar, Feng & Hochheiser 2010). Interactive search scenarios were included given the effectiveness of the method in providing empirical perspectives (Preibusch, Soren 2013a, 2013b). The survey was pilot-tested in June 2015 with a sample of 21 participants who were representative of the targeted demographic population. To ensure reliability, items in a 5-point Likert scale were randomised to minimise random errors and care was taken to clarify wording of the items where ambiguity existed, along with ensuring question types and response formats appropriately tested the construct.

Internal consistency assesses the extent to which scale items measure the attribute being tested (Pallant 2007). Cronbach's alpha is one such measure, where good internal consistency is indicated with a value greater than 0.7. Cronbach alpha value coefficient for the Likert scales on expectations around data flows, and consent and disclosure mechanisms were 0.828 and 0.872 respectively indicating good internal consistency.

Construct validity measured the extent to which the survey instrument is useful in practice (Litwin 1995). Feedback on how participants perceived questions were used to refine them to minimise the effects of preconceived biases and the influence of sensitisation to the

survey instrument by participants. The survey design was also critically reviewed by two experts with experience in the design and administering of survey instruments to ensure accuracy and precision. The approved ethics protocol was amended with these changes to the survey study protocol and subsequent human research ethics committee approval was obtained for these amendments. These two measures provided internal validity through verification and refinement of the constructs tested.

External validity (i.e. generalisability of the findings) is constrained to population centres with similar metropolitan characteristics as outlined in section 5.2.1. The sample was an educated population from a university with 37,000 students and staff in Adelaide city. Random sampling was used for participant selection from this population, with the unit of sampling being participants that met the qualifying criteria. The key qualifying criteria were whether participants regularly searched online and if they are either a current student or staff member of the university. All participants undertook the same survey minimising potential participant biases that threaten internal validity. Survey findings were compared with findings from related work (such as findings from Australian privacy surveys) where applicable to provide a more representative view. Quantitative data was analysed using SPSS, a statistical analysis software. Data was cleaned for errors where they existed and scale items were coded during data analysis.

The survey comprised of a mix of close-ended questions (multiple choice Likert scales, short-form text responses and multiple selection options) and open-ended questions (long-form text responses). A random sample of 120 students and staff was sought. The survey constructs are present in Table 3.2 (see Appendix 5.1 for the detailed survey).

Table 3.2 Survey Study Design

Factor	*Rationale*	
Demographics [Q1]: Age, Gender, Role, Area of Expertise, Associated Culture, Heritage or Ethnicity	To provide a description of the survey sample.	
Quantitative Questions		
Search Behaviour [Q2]: Reliance on Web, Level of Expert, Browser Use, Search Engine Use, Preference & Frequency, Device Frequency	To understand characteristics around how users search.	
Awareness of Personal Data Release and Collection [Q3a-c]: Awareness of tracking, accessibility & use of personal data in searches	To identify level of awareness of how personal data flows in search.	
Expectations around Data flows [Q3d-i]: Public availability of search history, Release and use of personal data, Linkability of data, Capture of clickthrough data, Tracking search behaviour, Aggregation of data	To identify level of expectations around different types of data flows.	These expectations were identified based on a synthesis of literature review findings such as (Awad & Krishnan 2006; Teltzrow & Kobsa 2003; Toch, Wang & Cranor 2012).
Consent and Disclosure Mechanisms [Q3o-v, Q6]: Choice on Identifiability & Anonymity, Transparent notice and indicative level of transparency, Control around collected data, Notification of security of data, Notice and Consent Privacy Policies, Length and Complexity of Terms and Conditions	To identify expectations around existing consent and disclosure mechanisms.	General questions around the expectation of data flows and consent and disclosure mechanisms were used to ascertain if there was evidence to prove/ disprove the findings from chapter 4.
Privacy Risks and Impacts [Q3j-n]: Unauthorised access, Identifiability, Monetizing personal data, Accessibility of data to family/friends or co-workers/employer	To identify level of expectations around typical privacy risks associated with searching. These risks were identified based on (Solove 2009)	
Privacy-enabling Mechanisms [Q4]: Awareness and Use of Privacy Policy Mechanisms and Privacy-enhancing Technologies	To identify awareness and use of existing privacy-enabling mechanisms such as policy information communicated to users and privacy-enhancing tools. This question included images of mechanism and served to validate the levels of awareness and expectations addressed.	
Privacy Concerns [Q9]: Privacy Unconcerned, Fundamentalist and Pragmatist based on Westin's Privacy Segmentation Index	To evaluate if Westin's categorisation of privacy concerns is still relevant in the current online information privacy context. The questions were based on a review of Westin's privacy segmentation index (Kumaraguru & Cranor 2005).	
Interactive Search Tasks		
Validation of Search Behaviour [Q7]: Informational Scenario: Health Terms (Blood Pressure/ Hepatitis B), Transactional Scenario: Flight Tickets	Users were asked to perform these search tasks and report their answers. This question validated search behaviour and provided indications of sensitivity of query terms (as reported in Chapter 4).	
User Evaluation of Effectiveness of Obfuscated Search Results [Q8]: Minimum vs. Default, Minimum vs. User-Agent Obfuscated, Minimum vs. Location Obfuscated	Users were asked to identify which search result was preferred to identify effectiveness of obfuscated techniques based on findings from the technical controlled experiment reported in Chapter 4.	

Factor	***Rationale***
Qualitative Questions [Q2f, Q3w, Q5]:	
Open-ended question on search behaviour: Use of search engines and/or browsers (Optional) (*"Do you have any additional comments on how you use search engines or browsers?"),*	To provide users an opportunity to offer alternate explanations of their choices on search behaviour. 114 provided usable views. This qualitative question on search behaviour was used to support the quantitative close-ended questions.
Expectations around use of personal data online (Optional) *("Do you have any other expectations about the use of personal data online?)*	To provide users an opportunity to offer alternate explanations of their choices on privacy expectations. 93 provided usable views.
Narrative on privacy expectations (Mandatory) *(*"*Give an example of an instance where you felt your expectations about privacy were met or not met when you searched online and your reasons why you think this is so."*).	By providing an illustrative example of how people perceive privacy online, to further understand privacy expectations and the underlying social norms and values. This question was formed based on the key premise of the contextual integrity framework (Nissenbaum 2010).

The survey outreach plan included invitations to students via recruitment flyers posted on approved notice boards, an announcement on the student portal and email requests for a call for survey participants (staff and students) sent to the respective school administrator and approval sought prior to circulation in October 2015. The online survey was open for 6 months receiving 441 responses. A prize draw to win one of four $50 vouchers was offered in appreciation of participant's time and effort.

Data analysis of qualitative responses commenced with coding to uncover distinct themes (Bryman 2012). The coding process involves the development of a codebook which outlines "operational definitions for all variables" (Neuendorf 2002, p.111). For the questions on privacy expectations, the coding schedule included first identifying data flows and then uncovering potential instances of privacy risks. Data flows and privacy risks include those outlined in Table 5.1 after a review of related literature. These findings were used to guide the identification of privacy risks and expectations. These elements were coded by the researcher over three iterations for reliability and included cleaning data which were unusable or unclear.

Question 5 on understanding privacy expectations qualitatively was made mandatory because it is necessary to validate the expectations sought from a quantitative point of view and to place it in context. A wide scope of responses was provided ranging from expectations expressing concern to those not concerned. All responses were included in the thematic analysis if they were presented in a complete format. The alternative option to interview a sample of participants to gather a qualitative interpretation may present issues of potentially forced answers. The survey method allows the user to be anonymous and freely voice their opinion regardless of what it may be as evidence by the responses.

Thematic analysis identified important themes pertaining to privacy expectations around data flows in search. The thematic analysis was performed using a data-driven code approach consisting of the following steps adapted from Boyatzis (1998)'s data-driven approach to thematic analysis: designing the sample to be surveyed, collecting the data from the sample, identifying themes within the sample, categorising the themes, comparing the themes across different subcategories, creating a code and determining its reliability. Coding was performed manually over three iterations.

This thematic analysis was framed by the decision heuristic in the contextual integrity framework to uncover any instances of privacy violations (Nissenbaum 2010). Codes reflected instances of data flows and privacy expectations. (see Table 5.12 in Section 5.3 in Chapter 5 for more details). No conflicting narratives were discarded. The range of privacy expectations articulated was considered when identifying the themes. Only unusable (insufficient text to derive meaning) narratives were discarded. This categorisation was done iteratively on paper and on excel. Themes were identified after three iterations. Idea networks were used to visualise these themes. Idea networking is a technique used "to identify the conceptual frames from participant statements" which helps to "conceptualise problems" via "analysis and synthesis"(Metcalfe 2007, p.141). The creation of idea networks means ideas are "expressed as statements" which become "nodes on the network" (Metcalfe 2007, p.144). If two ideas are linked through synthesis, a connection is made. The strength of the connection increases the size of the node forming clusters in the network. These clusters with their connections visualise conceptual frames which can then be used to guide analysis and interpretation of findings. To provide some consistency and rigour, idea networking was used to provide a qualitative visualisation of these thematic concepts, including the degree of strength (i.e. connections) between key emerging concepts. These conceptual frames are visualised in section 5.3

Another way to ensure validity and reliability in qualitative research, is that researchers should assess the trustworthiness and authenticity of the data. Trustworthiness includes credibility, transferability, dependability and conformability whilst authenticity includes fairness, ontological authenticity, educative authenticity, catalytic authenticity (Guba & Lincoln 2008; Lincoln, YS & Guba 1985). This can be achieved through triangulation which ensures that several methods or sources of data are used (Bryman 2012).

In this study, qualitative privacy expectations were triangulated with quantitative and interactive responses, ensuring the ability to validate data from multiple perspectives and providing data credibility (i.e. internal validity). Transferability (i.e. internal validity) is provided through the value of in-depth narratives which can also be referenced against the demographics of the sample. Dependability (reliability) and confirmability (objectivity) is ensured through effective data collection and analysis practices, presentation and

refinement of ideas through scholarly debates, and overall research being underpinned by adhering to human research ethics protocols.

To ensure authenticity, fairness calls for data from diverse perspectives to be represented which is seen in the detailed accounts of where users find and do not find value of personal data release and collection. Ontological (i.e. better understanding), educative (better appreciation of perspectives), catalytic (i.e. a catalyst for change) and tactical authenticity (i.e. empowering action) are factors that address participant perspectives. Although these factors were not the key focus of the study, given the focus on privacy expectations, by partaking in this study participants gained some value through self-awareness of their own privacy expectations and understanding behavioural implications. Participant statements are represented as-is, with spelling errors corrected and the use of [] to denote missing words.

3.1.3 Conceptual Investigations: Methodology

"Conceptual investigations comprise philosophically informed analyses of the central constructs and issues under investigation" (Friedman & Freier 2005, p.369). Value sensitive design employs a philosophically-informed analysis of the values at stake for key actors, which in this case are users and information service providers. However, perception of values and its understanding changes through interactions because technology mediates user interactions (Verbeek 2016). Therefore, to strengthen the conceptual analysis, technology mediation theory as developed by Verbeek is used (Verbeek 2011). Traditionally value sensitive design has been used towards informing and refining specific technology designs, such as embedding informed consent in Web browsers (Friedman, Felten & Millett 2000). As this project aims to understand the effectiveness of privacy protections afforded to users, it was necessary to flip the order in which value sensitive design is more commonly-used, to understand the research problem from technological (technical) and societal (empirical) dimensions. The object of study is the socio-technical interaction. These findings can then inform a holistic approach to technology design which is more inclusive of the diverse privacy sensitivities users express in relation to appropriate flows of information.

The conceptual investigation reported in chapter 6 consists of two main components, namely a theoretical and a practical contribution. The theoretical contribution comprises of

an analysis of the values the key stakeholders hold during these interactions and an understanding of how technology mediates these interactions. A concept for privacy expectations in socio-technical interactions is presented. The practical contribution takes the form of design recommendations outlined to assist designers, empower users and equip policy-makers in the roles these key stakeholders play in addressing information privacy challenges.

3.2 Alternative Approaches

Alternative methodological approaches to value sensitive design were considered and are described in this section. Action research advocates the study of a problem in two phases namely a diagnostic stage, where analysis and understanding of the context of the problem is built through efforts by the researchers and the participant and a therapeutic phase, where hypotheses are identified, changes or interventions are introduced to the context and their effects studied (Baskerville & Wood-Harper 1996).

A similar approach is the interpretivist paradigm based on the notion that people create subjective and multiple meanings through interaction with the world. Interpretivist research aims to understand social phenomena through the meanings associated to them by people in a given context (Creswell 2009). Constructivism which is used in this research extends interpretivism by attempting to understand users' expectations of privacy and how they construct notions of privacy in contexts where data is released when searching.

Action research extends the interpretivist paradigm by allowing the researcher to study the problem through continuous staged iterations. Exploring this research problem through the lens of action research provides theoretical grounding and rigour to first build an understanding of the context before applying interventions. However, given the exploratory nature of the problem, the unit of sampling measures user expectations at a particular point in time and does not extend to measurement across iterative cycles. Therefore, any proposed interventions cannot be rigorously tested in iterative cyclic stages. If for example, the study aimed to develop and investigate suitable privacy-enhanced techniques around data release, action research is an appropriate methodology as it would allow development

of the proof-of-concept in cyclic iterations and enable measuring effectiveness of privacy expectations and protections afforded at various time points.

Alternative design methodologies to value sensitive design include reflective design, which uses critical reflection to uncover hidden biases and assumptions in socio-technical interactions, so that it is used to inform design proactively (Sengers et al. 2005). Participatory design on which the value sensitive design method is developed, explores incorporating democratic values in systems design (Schuler & Namioka 1993). While participatory design is a useful approach, value sensitive design was determined to be a more appropriate method to explore the underlying privacy problems associated with data release when searching from multiple perspectives (i.e. technical, empirical and conceptual). Another approach seeks to embed design values in practice by using the "Balls in Play" metaphor exploring technical, philosophical and empirical issues in three stages namely discovery, translation and verification (Flanagan, Howe & Nissenbaum 2008, p.328). However, while each of these methods deal with improving a specific functional technical artefact, whereas this research study is concerned with the effectiveness of privacy protections afforded to users during the interaction.

3.3. Conclusion

In conclusion, the overall approach guiding this research uses value sensitive design given the exploratory nature of the research problem in three dimensions to investigate the overarching research question i.e. the effectiveness of privacy protection mechanisms afforded to users when releasing data while searching. Value sensitive design was adopted because it allowed for the investigation of existing affordances, understanding online privacy expectations and exploring mechanisms that bridged the gap between these two important issues.

4 Technical Investigation: Identifying the Effects of Data Release on the Variance of Personalised Search Results

This chapter investigates the effectiveness of technical affordances involved in data release on a specific application domain – i.e. personalised search results. Specifically, this chapter explores if personalised search results can be obtained with limited release of implicit data which is typically released without a user's awareness or understanding. In this study, this exploration is done by firstly, discovering implicit data being transferred while searching, and secondly, investigating the effects of implicit data on the variance of personalised search results from a privacy-sensitive user perspective. It is important to determine the effects of implicit data on search results to preserve the notion of personal data being protected from an individual point of view in the context of search by providing users with affordances to exercise their privacy choices.

A controlled experiment study was conducted on seven different search engines (both commercial and privacy-enhanced) sourcing 29 popular query terms across 3 timeframes over 2 years (2014-2016). Overall, through rigorous statistical analysis, the variance of personalised search results are consistently affected mostly by location, the user-agent parameter, topicality of the query term (popular informational terms), and query terms where the user intent was transactional or navigational, had significant impact on personalisation variance across both commercial and privacy-enhanced search engines over these timeframes. For some commercial search engines, cookies and parameters carrying personal-identifiable data such as the referrer header had a minimal, but inconsistent impact. For privacy-sensitive users, these findings give rise to an opportunity to proactively control the distinguishability of their queries through privacy-sensitive socio-technical affordances that obfuscate the data released in queries. By minimising data release and increasing visibility of data flows in search, users may be empowered to make proactive and informed privacy choices.

4.1 Introduction

Data is the 'new oil'(Humby 2006). Data is a valuable business commodity and for users, personal data is the currency fuelling our interactions on the Internet. In Web search, personal data is used to produce relevant search results for users (Berendt & Teltzrow 2003). As commercial search engines are profit-oriented, personalisation also serves the purpose of creating targeted groups of users relevant for advertisers (Feuz, Fuller & Stalder 2011). Whilst data collected is used to improve product and service offerings, there is also a clear commercial need to monetise the personal data collected to derive insights about customers, increase service value and generate business-related profits.

From a user's point of view, personal data is released to obtain the benefits of consuming a data-driven service like Web search. *Data flow* in the context of this chapter is concerned with the transfer of personally-identifiable information when users search on the Web to obtain personalised search results (Kobsa 2007). When users release this personal data implicitly (i.e. without a user's explicit awareness or knowledge) it is *data leakage* (Micarelli et al. 2007). In Web search, implicit data leakage occurs when personal and contextual data is transmitted through search query parameters. Personalisation in search is said to represent "the context of you, the searcher, including your interests and your network of contacts" (Singhal 2011). Implicit data released may also benefit search engine optimisation and form signals that trigger personalisation as observed with algorithmic changes (Bing 2011; Google 2009a; Horling & Kulick 2009; Singhal 2011; Sullivan, D 2003, 2009b, 2009a, 2011b, 2011a, 2012, 2015; Yahoo 2015). Hence, it is believed that search engines use personal data to produce personalised search results (Feuz, Fuller & Stalder 2011).

However, contrary to this view, some commentators mention that data collected may not be completely used in search results to avoid instability of search results (Chi 2015). In addition to the explicitly specified query term, data such as HTTP headers, GET/POST parameters and cookies are implicitly transferred. Although individually these data points may not contain personally-identifiable information, when combined together, it can be used to uniquely identify a user, resulting in linkability between queries (Sullivan, D 2003; Sweeney 2002).

Although it is evident that such identifiable data is collected, it is interesting to know the effects this released implicit data has on the variance of personalised search results, particularly for those users concerned about the privacy of their online personal data, and its value and impact on their technology-mediated interactions. Of particular concern is understanding how release of implicit data impacts a user's privacy choices given that this release may occur without their awareness or knowledge.

Privacy is a social construct and is contextual in meaning as it means different things to different people, and at different times (Reiman 1976). Information privacy is concerned with data that personally identifies a person. "Personal data is data containing any identifying characteristics about a person" (European Commission 2009, 2015a). Privacy-sensitivity in this context deals with the visibility of data flows in search. Of importance is the visibility of how personal data flows pertaining to a user's awareness and understanding. If a user has expectations or concerns about how their personal data impacts their searches, they express some degree of privacy-sensitivity.

There's an observable increase in awareness and concern about privacy of digital personal data among people. This is reflected in consumer views regarding loss of control in managing personal data (91% of consumers in the US were of this view in 2014) and willingness to release their personal data for a benefit (51% of the same surveyed consumers) (Madden 2014). In a similar 2015 Pew survey, consumers expressed concerns about secondary or invasive purposes of personal data collected where their willingness to share personal data was highly reliant on its context of use (Rainie & Duggan 2015). These views reflect a growing trend in the lack of choice for consumers in making informed decisions when interacting with information service providers and a willingness to be able to control how their personal data is being used. Hence, they seek openness and transparency on how their personal data is used.

The implications of misuse of personal data release far exceeds the monetary value attach to it. Prior work has reported on a variety of data privacy issues including behavioural profiling, targeted advertising, linkability of data even when aggregated, excessive data collection, poor user awareness and control over personal data (Feuz, Fuller & Stalder 2011; Solove 2006, 2007; Sweeney 2002; Tavani 2008; Tene & Polonetsky 2013; Zimmer,

M. 2008). Often the consequences of personal data misuse is reputational, instantaneous, far-reaching and permanent as evident in Target's infamous targeted advertising teen pregnancy case (Ellenberg 2014).

In this chapter, the effects of removal of personally-identifiable data in personalised search results are investigated through a *controlled experiment* design which compares the variance of personalised search result lists by incrementally controlling the data flow released across both commercial and privacy-enhanced search engines. The experiment reported in this chapter forms part of a larger project to investigate the effectiveness of data release approaches in obtaining satisfactory personalised search results for users.

Establishing the effects of implicit data on personalised search results is of importance to the preservation of individual privacy. By identifying affordances on the transfer of personal data, individuals can proactively manage their identifiability online before data leakage occurs. From a privacy-sensitive user perspective, if implicit data released has negligible impact on the variance of personalised results, can such users still obtain the information they wish to consume through limited release of personal data? To achieve this, variables that appear to influence the variance in personalised search results are determined and which of those contain personally-identifiable data.

Determining these variables and their effect on personalised search results is the subject of this chapter while the inherent problem of data leakage underpins the overall motivation. The main focus of the controlled experiments is not to predict search engine personalisation algorithms. Rather, it is to investigate and determine the types of personal data that leak through personalised search and identify the effects implicit data release on the variance of personalised search results.

The chapter structure includes a brief background on data release in Web search discussing both related work and data leakage techniques. The experiment methodology section details the identification of different implicit data variables and measures the effect of this implicit data on the variance of personalised search results. The analysis of the controlled experiment results focuses on the effects implicit data has on personalised search. The

chapter concludes with a discussion on the role of implicit data in personalised search and identifies possible avenues for future research.

4.2 Related Work: Data Release and Leakage in Web Search

Literature on data release and leakage in personalised search, and existing mechanisms of managing data release is reviewed in this section.

4.2.1 Data Release in Web Search

When users search online, search query, browser environment and user details are sent both *explicitly* (i.e. the query term the user types) and *implicitly* (i.e. various search query parameters consisting of HTTP Headers, GET/POST parameters and Cookies). Search engines personalise search results through account sign-in and non-account sign-in methods. For example, commercial search engine account login personalisation, in Google's case, a user's search history coupled with cookies enables personalisation. In the non-account login scenario, anonymous persistent cookies linked to the user's Web browser, active for 180 days, enable personalised search results to be generated (Enge 2011; Google 2017; Horling & Kulick 2009).

Privacy-enhanced search engines claim not to personalise search results because they do not store user details such as a user's IP address as in StartPage's case (StartPage 2014a). Personalisation is also observed to be impacted by country, locality, personal history and social connections (Ellis 2013; Meyers 2015; Search Engine Land 2016). Eye-tracking experiments provide evidence that users tend to focus only on the top two search results in a given search engine results page suggesting an F reading pattern (Google 2009b).

Even with changes to search result ranking and reordering, users tend to click on an organic search result on the first result page (Pan et al. 2007). For example in a 2014 clickthrough rate study,71.33% of searches resulted in a click on an organic search result on the first page with the top 5 links accounting for 67.60% of clicks (Advanced Web Ranking 2014). Similarly the effectiveness of advertisements or personalised ads has decreased with banner blindness or right-sided blindness being a commonly studied phenomenon (Brajnik & Gabrielli 2010). Users essentially focus less on the right side of the page, by unconsciously

filtering out information with repeated familiarity. This places importance on the order and rank in terms of the variance of personalised search results.

4.2.2 Data Leakage Methods on the Web

Prior studies have investigated dynamic Web environments to identify sources of leakage and user tracking mechanisms on the Web or variances in personalisation (Hannak et al. 2013; Malandrino & Scarano 2013). Commercial search engines collect longitudinal user data by monitoring and tracking user-browsing activities (Kosta et al. 2009; Tene, O 2008; Zimmer, M. 2008). In addition, third-party advertisers track users across different websites and may cause indirect damage where not only is the primary user being tracked at risk, but information about other related users is captured as well (Krishnamurthy & Wills 2009). Third-party agents use stateful tracking methods via persistent supercookies and stateless tracking through fingerprinting via both active and passive techniques (Mayer & Mitchell 2012).

Passive fingerprinting concerns leakage of browser-related information including those explored in this experimental study, namely IP address, user-agent and HTTP headers. Recent studies explored the impact of new tracking mechanisms through session hijacking personalised recommendations, canvas fingerprinting, evercookies and cookie syncing and the subsequent privacy loss and risks posed by these techniques (Acar et al. 2014; Castelluccia, C., De Cristofaro & Perito 2010). The ability to make inferences about a person is further enhanced through these persistent tracking mechanisms and raises the issues of protecting the identity of users online and preserving the option of anonymity.

4.2.3 Existing Mechanisms of Managing Personal Data in Search

In the context of implicit data leakage in personalised Web search, the variance of personalised search results can be measured by the Edit Distance and Jaccard Index, assessing the level of reordering and amount of difference between two result pages. Implicit data leakage is limited to HTTP Headers, GET/POST parameters, cookies, query terms and disabling of Javascript, eliminating browser history and having clean scenarios (no potential to capture data through other scripts or methods). The types of data transferred implicitly through Web browsing are represented in Table 4.1.

The implicit data released is used to inform user profile models and recommend personalised suggestions to users. Personally-identifiable information such as device, log and location information collected are often shared among external third-parties in cases where consent is retrieved (such as the 'I Agree' click sign-in), external processing among trusted affiliates, and for legal reasons (Koene et al. 2015). A similar line of work found that the ability to accurately detect sensitive topic profiling increased when monitoring changes in advertisements and found evidence of search engines learning topic sensitivity without a user's Web history or through non account-login (anonymous) search scenarios (Mac Aonghusa & Leith 2016). Related work in this area has explored the use of filtering mechanisms to filter the search query request and response while obfuscating the location (Saint-Jean et al. 2007). Another study focused on learning which domains plant and retrieve cookies, so as to only allow first-party domain cookies (Walls, Clark & Levine 2011). This study extends this related work by analysing the variance of implicit data released in terms of the value it provides to personalised search results. By focusing on data flows before data acquisition, the aim is to investigate if the implicit data released has any effect on personalisation variance, with the possibility of investigating obfuscating of search query requests for privacy-sensitive users.

Existing mechanisms of expressing user consent and choice is sometimes inadequate. When this is coupled with poor security mechanisms, leakage of personal data can be detrimental as in the case of Baidu apps leaking personal data (BBC 2016). DoNotTrack (DNT) is an HTTP header that captures the user's choice over tracking preferences (Mayer & Mitchell 2012). However, the lack of consensus among technology organizations to respect, and action in technically implementing user preferences expressed through DNT is a major weakness and leaves the DNT as an ineffective mechanism for expressing user choice particularly through robust technical means.

The EU Cookie law is meant to raise awareness about cookies, tracking and behavioural profiling, and provide users with a choice over tracking on websites (EUROPA 2016). However, this mechanism is inadequate for a number of reasons: cookie notification messages are too generic, with repeated familiarity users are blind to perceiving the notice, it assumes that users already know what cookies are, and often the website is just ensuring compliance with law but leaves loopholes by not specify exact purposes (including

secondary use) of the cookies for data collection, and is a passive form of communicating privacy expectations as the website remains operational regardless of whether users see the cookie notice or not.

There is also a visible default laziness effect, where users hardly change the default settings in Web browsers due to convenience and lack of awareness of the importance of these settings in capturing privacy preferences. An emerging line of work calls for user-centric affordances in managing personal data. The databox is one such concept where a technical platform to securely manage personal data at the client-end is advocated for with due consideration given to key characteristics of agency, legibility and negotiability (Chaudhry et al. 2015; Mortier et al. 2014). An alternate example is IBM's Identity Mixer, which provides a platform for developers to verify users through public keys that correspond to identities with users having secure access to a single private key (IBM Corporation 2015).

There has been educative improvements in raising the awareness of user controls, such as controlling information collected when account logged-in through a central dashboard and the importance of educating consumers on aspects like private browsing, use of cookies and other means for tracking (Google 2016a, 2016b; Mozilla 2016). Google's dashboard for example, was created with the aim of providing transparency and control among users. The dashboard focused on addressing users concerns such as who had access to the kinds of data about them, and steps to remedy the indeterminacy reflected by users in exercising personal data protection strategies (Ortlieb 2014). However, these educative goals centre on control of personal data after data release, which is a passive or reactive way of providing user choice. These goals aim to educate users on data collection and choice, but mostly represent only a portion of the data collected (such as when actively logged-in to accounts or data stored at the client-side).

The effectiveness of existing mechanisms of data collection without a user's knowledge and consent has been debated much in the legal community, with rulings often advocating for the privacy rights of the user. For example in the Google Inc v Vidal-Hall & Ors case the claimants successfully argued that their personal information was tracked and collected when using the Apple Safari browser without their consent which was an invasion of their privacy (*Google inc v vidal-hall & ors* 2015). While the law allows citizens to exercise

their privacy rights, this approach is often limited in that it is retrospective and the nature of the loss of privacy is tightly intertwined with a person's identity.

The risks associated with the unauthorised access, collection, release and misuse of personal data may be irreparable and instantaneous damage, most often with a user's reputation or corporation's intellectual property at stake. Further, the use of pervasive methods of tracking user activities whilst searching makes it imperative to minimise the leakage of personal data through different media. It is therefore, important to consider methods of preventing leakage of personal data before the act occurs to minimise the loss associated with the unauthorised access, release and misuse of personal data and protect the identity or anonymity of users on the Web. This would provide better socio-technical affordances that are privacy-sensitive to the needs of users. Hence, this experimental study explores the role of data (i.e. the search query components) transferred implicitly whilst users search on the Web and its impact on personalisation.

In this section related work on data leakage is reviewed in Web search focusing particularly on implicit data release. Existing mechanisms of managing data release are often reactive with singular functional focus and occurs in isolation. This body of existing work is extended, by attempting to understand the value of implicit data through investigating the effect this implicit data release has on the variance of personalised search results from a privacy-sensitive user perspective.

4.3 Experiment Methodology

A controlled experiment was designed to identify the effects implicit data has on the variance of personalised search results.

This section describes the experiment design including details on how query terms and search query parameters were identified, and how search scenarios for each search engine and controls for sources of noise were defined. This section concludes with a description on the experiment methodology with a description of the personalisation variance measures (Jaccard Index and Edit Distance) and the statistical analysis measures (Wilcoxon Signed Rank Test, Non-metric Multi-Dimensional Scaling and Generalised Linear Modelling) used and its application in the context of the controlled experiment.

4.3.1 Experiment Design

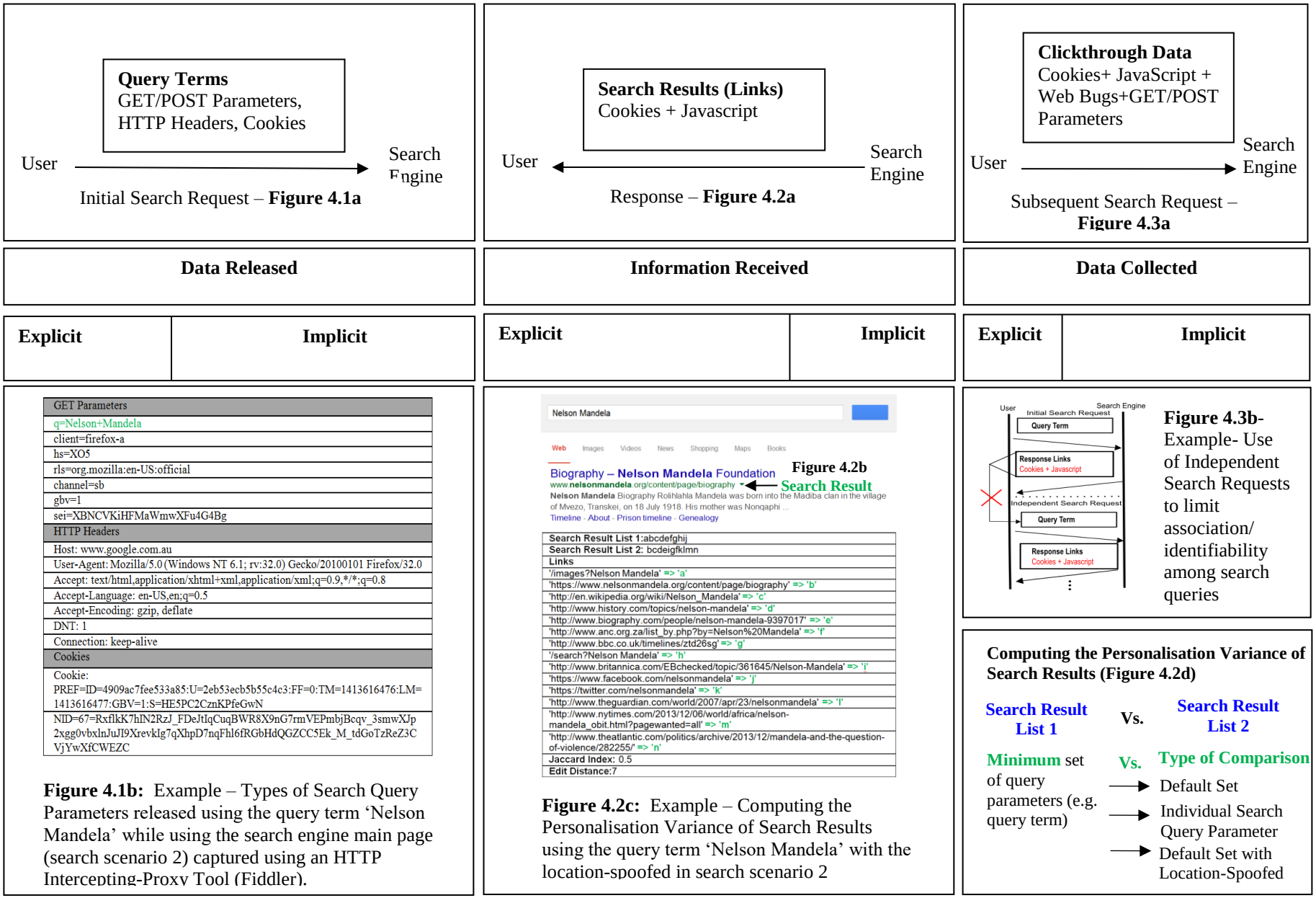
This controlled experiment investigated the flow of implicit user data through the HTTP channel, consisting of three key steps (see Figures 4.1-4.3). Firstly, a sample of search query parameters was captured for each defined search scenario across each search engine and query term using an HTTP intercepting proxy tool (Fiddler). Secondly, search requests were automated (for each query term across each defined search scenario and search engine) to compute the variance between search result lists in different search scenarios using the Edit Distance and Jaccard Index measures (for each query result comparison retrieved). Thirdly, the personalisation variance comparison datasets were analysed to identify factors significantly affecting the variance of personalised search results using the statistical measures – Wilcoxon Signed Rank Test, Non-metric Multi-Dimensional Scaling and Generalised Linear Modelling. A summary of these controls and the overall experiment design is provided in Table 4.1.

For scope, the focus was on the data released through the process of an initial search request, i.e. the first interaction the user has when seeking information by querying the search engine. Hence, to identify the effects implicit data has on personalised search results, data flows were controlled by selectively releasing the query terms, the implicit data (search query parameters), and the scenarios through which users search in iteration.

Table 4.1 Overall Experiment Design

Category	Feature	Description	Example
Independent Variable	Mechanisms of Data Release: Query Terms, Experiment Conditions	Informational, Navigational and Transactional Query Terms	Refer Table 4.2
Dependent Variable	Effects of Search Query Parameters, Query Terms and Location on the Variance of Personalised Search Results	Randomly occurring based on account and non-account login search scenarios	Refer Table 4. 3 for examples for commercial and privacy-enhanced search engines
Experiment Conditions	Search Scenarios for Personal Data Release	Four common scenarios for searching based on account and non-account login methods	Refer Table 4.3
	Comparing personalisation variance across two search result lists.	Defined two search result lists: A *minimum* set of query parameters (control) vs. a *comparison set**	*A *comparison set* could be of three types: Each *Individual* Parameter, *Default* (set of parameters), *Location-Spoofed* (i.e. a default set with the location-spoofed)
	Search Engines	Commercial Search Engines and Privacy-enhanced Search Engines	*Commercial* -Google, Yahoo, Bing *Privacy-enhanced*- StartPage, Ixquick, DuckDuckGo, Unbubble
Experiment Run	Experiment data collected at three distinct time points with a gap of 7 months. This allowed a longitudinal view of the data.	The experiment design was optimised with each iteration	*1st Run - October 2014* – Google and StartPage *2nd Run - May 2015*- Google, Yahoo, Bing, DuckDuckGo and Ixquick *3rd Run – January 2016*- Google, Yahoo, Bing, DuckDuckGo, Ixquick, StartPage, Unbubble; Further experimental controls using Google as a lens
Randomisation of Experiment Run Order	Inclusion Effect	Systematic addition of search query parameters to search request	To identify if a particular combination of search query parameters has an effect on personalisation variance. This is the default type performed across all 3 runs.
	Exclusion Effect	Systematic removal of search query parameters to search request	To identify if the absence of a particular search query parameter influences personalisation variance. This control was performed in the 3rd run.
User-Agent Control	To identify personalisation variance influenced by the User-Agent	Use of multiple values for the user-agent parameter to determine reliable significance effecting personalisation variance. This control was performed in the 3rd run, with the default being tested across all runs.	Default User-Agent Parameter was: *Mozilla Firefox* [Mozilla/5.0 (Windows NT 6.1; rv:32.0) Gecko/20100101 Firefox/32.0] Tested Controls were *Google Chrome* [Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2228.0 Safari/537.36] *Internet Explorer* [Mozilla/5.0 (Windows NT 6.1; Trident/7.0; MATP; rv:11.0) like Gecko]

Category	Feature	Description	Example
Geolocation Control	Location-Spoofing the minimum query set comparison	To test the effect location has on personalisation variance of search queries	Use of an Amazon Web Services proxy server hosted in the US. Default location value for queries is Australia (dynamic IP addresses). This control was performed across all 3 runs.
Controls for Sources of Noise	Carry-over Effect	Each manual default search query was performed on a clean fresh instance of the browser with a gap of at least 11 minutes to minimise the carry-over effect	
	Stability of Search Results	Three datasets of personalisation variance across a sample of commercial and privacy-enhanced search engines were collected during a period of two years with a gap of 7 months in between each data collection point. This longitudinal view acted as a control for plausible variations in the stability of results.	
Privacy-Sensitive User Controls	To investigate only implicit data leakage through search queries	Use of independent search requests by removing the association among subsequent queries. Use of cURL and disabling JavaScript to limit dynamic tracking, collection of clickthrough data and Web bugs. This allowed personalisation variance to be attributed to implicit data release through search queries.	
Personalisation Variance Measures	Jaccard Index	provides a measure of the level of reordering between two similar lists.	See section 4.3.6 on Measuring the Variance of Personalised Search Results
	Edit Distance	provides a measure of the differences between two pairs of similar lists	See section 4.3.6 on Measuring the Variance of Personalised Search Results
Statistical Analysis Measures	Wilcoxon Signed Rank Test (non-parametric)	Demonstrates significance of personalisation variance through identification of statistically significant median increases	See Figure 4a and related description
	Non-metric Multidimensional Scaling (NMDS)	identifies the similarity or dissimilarity of a personalisation variance measure	See Figure 4b and related description
	Generalised Linear Modelling (GLM)	To identify and attribute the variation in the response variable (personalisation variance for query terms, and personalisation variance for query parameters) to the explanatory variables (factors such as search engine and query type, or search engine and query parameter respectively). See section 4.3.6 on Analysing the Variance of Personalised Search Results for further information.	



GET Parameters
q=Nelson+Mandela
client=firefox-a
hs=XO5
rls=org.mozilla:en-US:official
channel=sb
gbv=1
sei=XBNCVKiHFMaWmwXFu4G4Bg
HTTP Headers
Host: www.google.com.au
User-Agent: Mozilla/5.0 (Windows NT 6.1; rv:32.0) Gecko/20100101 Firefox/32.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Cookies
Cookie: PREF=ID=4909ac7fee533a85:U=2eb53ecb5b55c4c3:FF=0:TM=1413616476:LM=1413616477:GBV=1:S=HE5PC2CznKPfeGwN
NID=67=RxflkK7hlN2RzJ_FDeJtlqCuqBWR8X9nG7rmVEPmbjBcqv_3smwXJp2xgg0vbxlnJuJI9XrevkIg7qXhpD7nqFhl6fRGbHdQGZCC5Ek_M_tdGoTzReZ3CVjYwXfCWEZC

Figure 4.1b: Example – Types of Search Query Parameters released using the query term 'Nelson Mandela' while using the search engine main page (search scenario 2) captured using an HTTP Intercepting-Proxy Tool (Fiddler).

Figure 4.2b
Search Result

Search Result List 1:abcdefghij
Search Result List 2: bcdeigfklmn
Links
'/images?Nelson Mandela' => 'a'
'https://www.nelsonmandela.org/content/page/biography' => 'b'
'http://en.wikipedia.org/wiki/Nelson_Mandela' => 'c'
'http://www.history.com/topics/nelson-mandela' => 'd'
'http://www.biography.com/people/nelson-mandela-9397017' => 'e'
'http://www.anc.org.za/list_by.php?by=Nelson%20Mandela' => 'f'
'http://www.bbc.co.uk/timelines/ztd26sg' => 'g'
'/search?Nelson Mandela' => 'h'
'http://www.britannica.com/EBchecked/topic/361645/Nelson-Mandela' => 'i'
'https://www.facebook.com/nelsonmandela' => 'j'
'https://twitter.com/nelsonmandela' => 'k'
'http://www.theguardian.com/world/2007/apr/23/nelsonmandela' => 'l'
'http://www.nytimes.com/2013/12/06/world/africa/nelson-mandela_obit.html?pagewanted=all' => 'm'
'http://www.theatlantic.com/politics/archive/2013/12/mandela-and-the-question-of-violence/282255/' => 'n'
Jaccard Index: 0.5
Edit Distance:7

Figure 4.2c: Example – Computing the Personalisation Variance of Search Results using the query term 'Nelson Mandela' with the location-spoofed in search scenario 2

Figure 4.3b- Example- Use of Independent Search Requests to limit association/ identifiability among search queries

Computing the Personalisation Variance of Search Results (Figure 4.2d)

Search Result List 1 **Vs.** **Search Result List 2**

Minimum set of query parameters (e.g. query term) **Vs.** **Type of Comparison**

→ Default Set
→ Individual Search Query Parameter
→ Default Set with Location-Spoofed

Figures 4.1- 4.3 Experiment Design Conceptualisation of Data Flows in Personalised Search from a Privacy-Sensitive User's Perspective

4.3.2 Identifying Search Query Terms

A sample of 29 search query terms was sourced by combining the most popular search terms in 2013 on Google, Yahoo, Bing, Amazon (for transactional terms) and WebMD (for health terms) respectively (Amazon 2013; Bing 2013; Google 2013; WebMD 2013; Yahoo 2013). Table 4.2 shows the search query types and terms selected. The search query terms were classified under the established query categories of *informational* – containing knowledge content; *navigational* – seeking a particular page or site; and *transactional* – conducting a Web-based activity, aligned with user intent (Broder 2002).

As information query terms contain a wide variety of topics, ten different types of informational search query terms were defined (e.g. news, political, health, entertainment) and selected along with navigational and transactional terms. Two query terms were sourced for each of these query category types (20 informational terms, 2 transactional terms and 2 navigational terms). To assess the impact of locally-relevant results based on a user's physical location (Ellis 2013; Sullivan, D 2012), local public transport terms were used for informational lifestyle and navigational search terms. As a control, a comparison set of five popular and recent search query terms for 2014 were selected using Google Trends (as the experiment was first designed and run in 2014, complete lists of popular search terms were not available at the time). This control was to ascertain if personalisation variance is more significant for recent search terms in use. Any of these terms may be transient over a specific period. To control for these fluctuations, the statistical analytical techniques used (GLM, NMDS) take into account its cumulative variance.

4.3.3 Identifying Search Query Parameters

A default set of search query parameters were obtained by performing a manual search using the Firefox browser for each of the defined search scenarios (see Table 4.3 and the explanation in the following paragraph) across each search engine and query term. Search query parameters in the search requests were captured using Fiddler, an HTTP-intercepting proxy (Telerik 2014). A search query parameter was selected if it naturally or randomly occurred in a given scenario. Each search was done on a fresh instance of the Firefox browser with cookies and browser data cleared after each instance. The implicit data released (i.e. the search query parameters) was determined based on a particular search

scenario and focused on first-party access. The types of implicit data were produced based on the search scenario. While not every search query parameter is documented, the effect on personalisation variance was deduced through this controlled experiment for each parameter. The documented definitions of search query parameters are listed in Appendix 4 - B.

Table 4.2 Search Query Terms

Category	Query Term
Informational- News	Nelson Mandela; Royal Baby Born
Informational- Political	North Korea; Syria Conflict
Informational- Shopping	Cards Against Humanity; Playstation 4
Informational- Sports	Boston Marathon; 2014 FIFA World Cup
Informational- Quirky	Harlem Shake; Grumpy Cat
Informational- Entertainment	Beyonce Knowles; Paul Walker
Informational- Health	Health Reform; Diet Drugs
Informational- Lifestyle	Adelaide to Mawson Lakes Directions; Hawaii
Informational- Tech	iPhone 5S; Samsung Galaxy s4
Informational- Science	Typhoon Haiyan; Tiger
Navigational	Adelaide Metro; Dominos
Transactional	Australia to US Flight Tickets; Lean In Sheryl Sandberg
Recent (Topical) 2014 Query Terms as a Control	Ukraine Crisis; ISIS; Ebola Virus; Malaysia Airlines; Game of Thrones

Search query parameters were classified to be of two types: *Minimum Set* (the minimum required set of parameters compulsory for search to function which acts as the control) and *Comparison Set* (any search query parameters transmitted which are not absolutely required for the search query to function, i.e. the query would allow a result to be retrieved if the parameter was removed from the search request). The comparison set of search query parameters could be of three types: the default set of parameters, the default set with the location-spoofed, and each individual search query parameter. A minimum set was identified to establish a standard baseline for computing the personalisation variance between two search result lists. It was important to define a minimum set to establish a baseline for minimal data release. For most commercial and privacy-enhanced search

engines, just supplying the query term ('q' HTTP GET parameter), for example *http://www.google.com/search?q=query+term* is sufficient for search to function. Hence, all other search query parameters were additional and formed part of the comparison set. These search query parameters may influence how search results are generated. Depending on the personalised search algorithm employed by the search engine, these search query parameters can either affect results independently or have a combined interdependent effect. The parameters outlined in Table 4.3 were identified based on its random occurrence as per the defined scenarios (See Table 4.3). Establishing minimum and comparison sets formed the basis for computing the personalisation variance between two search results lists (i.e. personalisation variance across the minimum set vs. comparison set) across search engines, query terms and search scenarios.

Table 4.3 Search Scenarios and Search Query Parameters

Search Scenarios	Commercial Search Engine Example- Google	Privacy-enhanced Search Engine Example- StartPage
Minimum Set	q=query+term	query=query+term
Scenario 1 - Using the browser search bar with no cookies set and no account sign-in. (Regular use)	GET parameters: ie, oe, aq, rls, client, channel, gfe_rd, ei, Headers: HOST, User-Agent, Accept, Accept-Language, DNT	POST parameters cat, language; Headers: Host, User-Agent, Accept, Accept-Language, DNT
Scenario 2 - Using the search engine page with cookies set and no account sign in (First time visiting the search engine page manually and performing a search)	Same GET params as in Scenario 1 plus btng, hs, gbv or btnk & sclient or hl, source, sei Same Headers as in Sce.1 plus Referer Cookies: PrefID, NID	Same POST params as in Scenario 1 plus cmd, engine, abp, x, y. Same Headers as in Scenario 1 GET parameters such as cmd, pid,
Scenario 3 - Using a search engine associated account having browsing history with the browser search bar (Regular use – single persona)	Same GET params, Headers in Scenario 1 plus Referer. Same Cookies in Scenario 2 plusGoogleAccountsLocale_session, SID, HSID, SSID, APISID & SAPISID	Not applicable as no account login
Scenario 4 - Using a newly created search engine account with the browser search bar (Control for account sign-in – single persona)	Same GET Params, Headers, Cookies as in Scenario 3 except GoogleAccountsLocale_session	Not applicable as no account login

4.3.4 Identifying Search Scenarios

The ways in which people interact with search engines through Web browsers was looked at and modelled. For commercial search engines, users are expected to obtain better personalised search results when logged into accounts associated with the search engine (Enge 2011; Horling & Kulick 2009; Pan et al. 2007; Singhal 2011). Taking this into consideration, four scenarios of searching were defined. These include using the *search bar*, the *search engine page* (both of which are non-account login methods), the *search engine page with account-login* through a single active user account with at least 5 years of associated account activity. Accounts with at least 5 years of ativity were considered as personalisation is said to effectively and prominent occur when users are logged into their accounts) and this period of time would provide sufficient data on which to make inferences can be made for search personalisation. The last scenario consisted of a s*earch engine page with a control account login* (an account with no associated Web search history and activity was created). Examples of these search scenarios are provided in Table 4.3. Note that search scenarios 3 and 4 (account login based methods) do not apply to privacy-enhanced search engines as there are no accounts associated with them.

4.3.5 Identifying and Controlling Sources of Noise

The controlled experiment was designed to simulate the environment of a user performing typical Web searches. The experiments were run on a Windows 7 OS using a Mozilla Firefox Web browser across four search scenarios, seven search engines (both privacy-enhanced and commercial) and 29 query terms at three distinct timeframes (see Table 4.1). Privacy-enhanced search engines that were selected include StartPage, DuckDuckGo, Ixquick and Unbubble on the basis of their commitment to protecting user privacy as articulated in their privacy policies and mission statements (DuckDuckGo 2015; Ixquick 2015; StartPage 2014a; Unbubble 2016b). Commercial search engines that were selected include Google, Yahoo and Bing, three of the most widely used commercial search engines (NetMarketShare 2016a).

Some privacy-enhancing search engines source search results from other commercial search engines. For example StartPage sources search results from Google while aiming to protect user privacy (StartPage 2014b). In these cases, similar levels of personalisation variance due to the original search engine source being the same were expected. For example, StartPage could act as a useful comparison to Google's personalised search results. Some search engines supply other search engines, for example Google supplies AOL Search (Watson 2011) and Ask.com (Bloomberg 2010), whereas Bing supplies Yahoo (Oreskovic & Rigby 2007; Yahoo 2015). Search engines like Google and Bing may have higher

stakes and influence in the search engine market. For the period January 2015 – March 2016, Google had 66.50% of the global search engine market share on desktop (with it increasing to 92.8% on mobile/tablet for the same period), while Baidu had 11.66% (0.31% on mobile); Bing 10.96% (1.72% on mobile); Yahoo 8.61% (4.76% on mobile); and a host of other smaller search engines having a combined stake of 1% (NetMarketShare 2016a, 2016b). These statistics reflect the popularity of commercial search engines among consumers.

Initially in October 2014 when the experiment design was implemented, personalisation variance comparisons for a single commercial search engine (Google) and a single privacy-enhanced search engine (StartPage) were run. For the second iteration in May 2015 the method was expanded to include other commercial search engines (Bing and Yahoo) and privacy-enhanced search engines (DuckDuckGo and Ixquick) along with Google (as a comparison measure for stability of search results).

In January 2016, in the third and final iteration of this experiment, comparisons were run across all selected commercial (Bing, Google, Yahoo) and privacy-enhanced search engines (DuckDuckGo, Ixquick and StartPage) with the addition of Unbubble (a privacy-enhanced search engine specifically catering to the legal privacy requirements of the European Union and supports net neutrality (Unbubble 2016b). Collecting data during this longitudinal timeframe with data points at 7 month gaps allowed us to also investigate the stability of personalised search results across both commercial and privacy-enhanced search engines whilst controlling implicit data release.

When manually identifying search query parameters using Fiddler, each default search query was performed with a gap of at least 11 minutes to minimise the carry-over effect between subsequent searches for both account login and non-account login scenarios for each search engine (Hannak et al. 2013). For scope, organic Web search results were retrieved to compute personalisation variance. All search requests were directed to the search engine's main search page and not to an API to better reflect search results that everyday users would receive. The experiments were conducted through a custom script written using PHP and libcurl to form search query requests and capture the responses for analysis of personalisation variances between each comparison search result list combination (cURL Haxx 2013). Each automated query is treated as an individual interaction. Within a given scenario (or session), when measuring the variance of search query parameters depending on the comparison, existing cookies were reused, although they are not updated with subsequent received cookies to minimise association between query requests.

Statistical analysis measures were used (Wilcoxon Signed Rank Test and NMDS in particular) to uncover any potential dependencies associated with the reuse of cookies if query sources are recognised. To test the variances in *geolocation* on search query terms, an external IP address was used through an Amazon Web Services proxy server hosted in the US based on the minimum set. The default location (via the IP address) for all queries without geolocation spoofing is Australia. On occasion when testing, invisible links were appended to the search results to detect automated search query requests and capture clickthrough data. In those cases, the experiment scenario was rerun.

Controlling the Release of Implicit Data: Having designed the experiment to control the release of data when searching, further controls were integrated to reflect search behaviour of a privacy-sensitive user. Though search engines can perform personalisation based on a multitude of factors (e.g. IP address, Web browser version), the most valuable factor is considered to be knowledge about previous Web searches (Castelluccia, C., De Cristofaro & Perito 2010; Sullivan, D 2012).

To obtain this knowledge on a per-user basis, search engines must be able to uniquely identify users. Such identification mainly happens through cookie usage. If the identification of users is prevented, the search engine is unable to link users back to their profile. In general, this prevention of identification happens by analysing and restricting implicit data leakage for users. This prevention mechanism was implemented by only allowing the query term to be transmitted during search (See Figure 4.3b). By removing the association between subsequent queries, each query was submitted as an independent, history-free search request and restricted implicit data leakage. This limited association between queries allowed us to assess only the effect that each type of data item released has on the personalisation variance comparison.

Although HTTP is a stateless protocol, initial investigations revealed the use of JavaScript and JSON to track queries within a Web session dynamically. Through the use of cURL and disabling JavaScript, noise and tracking from the dynamic nature of the Web was limited. Due to the nature of this restriction mechanism used, clickthrough data, Web bugs and other tracking mechanisms were essentially disabled, enabling clean queries to be released. Deliberate and active user data disclosure is out of scope for this experiment as the focus is on implicit data release. For privacy-sensitive users, who intentionally do not want to release personally-identifiable information as they wish to avoid recognition from search engines and negate the effects of personalisation, implementing this control of removing association between queries is an important step in managing the distinguishability of their search queries. Hence, the experiment method enabled accurate observation of the variance of only implicit data released (i.e.

search query parameters as shown in Figures 4.1-4.3) on personalised search results. Table 4.4 provides an overview of how the controlled experiment design addressed implicit data release with comparisons to related work.

Table 4.4 Controlling Release of Implicit Data in Search

Data Released	Experiment Context	Related Work
Location	IP Address, Identifying parts such as .com.au of a URL; or inferring the location of an organization based on the IP Address	(Krishnamurthy, Malandrino & Wills 2007; Krishnamurthy, Naryshkin & Wills 2011; Saint-Jean et al. 2007)
Referer header	Normalising HTTP Headers like the Referer header	(Krishnamurthy, Malandrino & Wills 2007; Malandrino & Scarano 2013; Saint-Jean et al. 2007)
Cookies	Cookies – first party cookies set on initial search request; Flash cookies, Web bugs and beacons not explored	(Castelluccia, C., De Cristofaro & Perito 2010; Krishnamurthy, Naryshkin & Wills 2011; Walls, Clark & Levine 2011)
Search strings	HTTP GET and POST parameters	(Krishnamurthy, Naryshkin & Wills 2011; Saint-Jean et al. 2007)
Account identifiers such as Email addresses, Passwords, Account Numbers	Presumably stored in cookies set; subsequent means of leakage eliminated	(Krishnamurthy, Naryshkin & Wills 2011)
Leakage via Script Executions	Javascript was disabled in the browser	(Krishnamurthy, Naryshkin & Wills 2011; Malandrino & Scarano 2013)
Leakage via browser fingerprinting; browser history and auto-complete	Browser history was minimized; auto-complete applied in Scenario 2; browser fingerprinting is limited to identifiers leaked via search query parameters User-Agent String (Header Parameter): Browser and Operating System Version captured	(Castelluccia, C., De Cristofaro & Perito 2010; Krishnamurthy, Naryshkin & Wills 2011)
Leakage via Globally Unique IDs	Not explored as out of scope for implicit data leakage in search queries	(Krishnamurthy, Naryshkin & Wills 2011)
Obfuscation of queries by adding noise	Sends multiple queries to disguise real query via the TrackMeNot tool. Obfuscation was improved not by adding noise but by improving the quality of the signal sent out (to limit identifiability).	(Toubiana, Vincent, Subramanian & Nissenbaum 2011)

As Google is the most widely used search engine (Pan et al. 2007), Google was used as a lens to explore further *controls for rigour* in our experimental design (in addition to the standard personalisation variance comparisons). These controls are randomisation of the experiment run order, use of the inclusion/exclusion effect to determine if addition or removal of search query parameters introduces no

other related effect and confirmation of the significance of the user-agent as an important parameter affecting personalisation variance.

Randomisation of Experiment Runs: As an experiment control, the order of the experiment runs ('default', 'individual search query parameter', 'location-spoofing') was randomised to identify if the experiment run order had any effect on the results and to avoid introducing new biases with the comparisons.

Inclusion and Exclusion Effect: For Google - Initially, to address the research question (identify the effect of removal of personally-identifiable information on personalisation variance of search results), the comparisons were run by investigating the effect of systematically adding search query parameters (starting from the minimum required – the baseline). This type of approach helped identify the usefulness of the search query parameter. The effect of eliminating the search query parameter from the comparisons was investigated [default vs. factor]

Longitudinal Time Perspective of Personalisation Variance: The experiment runs were conducted from October 2014 to January 2016 across 3 specific timeframes with at least a 7-month gap. This longitudinal perspective will control for potential noise associated with A/B testing, search index improvements, distributed infrastructure upgrades as identified in the search personalisation measurement method by Hannak et al. (2013). In the first iteration of this controlled experiment, comparisons of unique search query parameters captured for scenarios was run and analysed. In the subsequent iterations of the experiment (2^{nd} and 3^{rd} runs), the experiment design was expanded to include all combinations of the three main comparison types (minimum vs. default query set, minimum vs. search query parameter, minimum vs. location-spoofed query set).

4.3.6 Measuring the Variance of Personalised Search Results

An automated script was written to measure personalisation variance, using Edit Distance and Jaccard Index to compare personalisation variance between a pair of personalised and non-personalised search results lists. Edit Distance provides a measure of the reordering between two similar lists. The higher the Edit Distance, the greater the level of reordering between the two search results lists. Jaccard Index provides a measure of the differences between two pairs of similar lists (Hannak et al. 2013). If the Jaccard Index is low, the difference between the two search engine result pages is larger.

4.3.7 Analysing the Variance of Personalised Search Results

The datasets were analysed using three statistical analysis measures using R and RStudio (see Table 4.1) (Crawley 2013; RStudio 2015; The R Foundation 2015). The Wilcoxon signed rank test is a non-

parametric method that is used to analyse the difference in median values between variables in two data sets that are paired or have a dependency on each other (Laerd Statistics 2015). In this chapter, the Wilcoxon signed rank test was used to determine if there is a difference between personalisation variance results for search query terms and parameters as per the defined experiment conditions (Refer Table 4.1). For scope, Wilcoxon tests were run on parameters such as default vs. location comparisons that showed visible variation in the retrieved dataset. To illustrate our use of this measure, see Figures 4.4a and 4.4b as examples. The Wilcoxon signed rank measure identified the effects of controlled data release on personalisation variance comparisons (e.g. minimum vs. default, against minimum vs. location-spoofing). For example, the 29 search query terms run in Google Scenario 1, with the location spoofed, the Wilcoxon test elicited a statistically significant median increase in personalisation variance in the Jaccard Index comparison to the default query ($W_{(1,29)} = 434$, $P < 0.001$; Figure 4.4a).

Non-metric Multi-Dimensional Scaling (NMDS) visualises dimensions representing distances between data points and approximately ranks the order of distances between data points in a set by comparing search query parameters in different conditions (Myers, Well & Lorch 2010). In this context, NMDS was used to identify the similarity or dissimilarity of a personalisation variance measure (i.e. either Jaccard Index or Edit Distance) between factors (i.e. search query parameters or query terms) for a particular experiment scenario. For example, the significant search query parameters for Google Scenario 1 (2^{nd} Run) are location, default (which is nearly equal to User-Agent), and gfe_rd as these parameters are dissimilar (shown through distances among clusters) to all the other parameters which are grouped together (See Figure 4.4b). If any comparisons in the datasets had errors or missing values, the data was cleaned (by introducing zeros in most cases) to obtain the NMDS computation as the statistical measure does not deal with missing values.

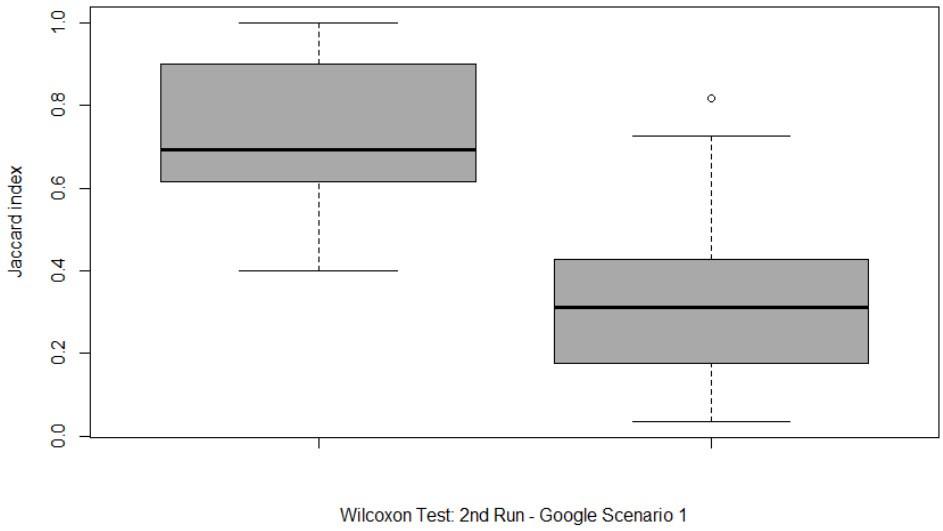


Figure 4.4a Wilcoxon Test of Jaccard Index: 2nd Run – Google Scenario 1

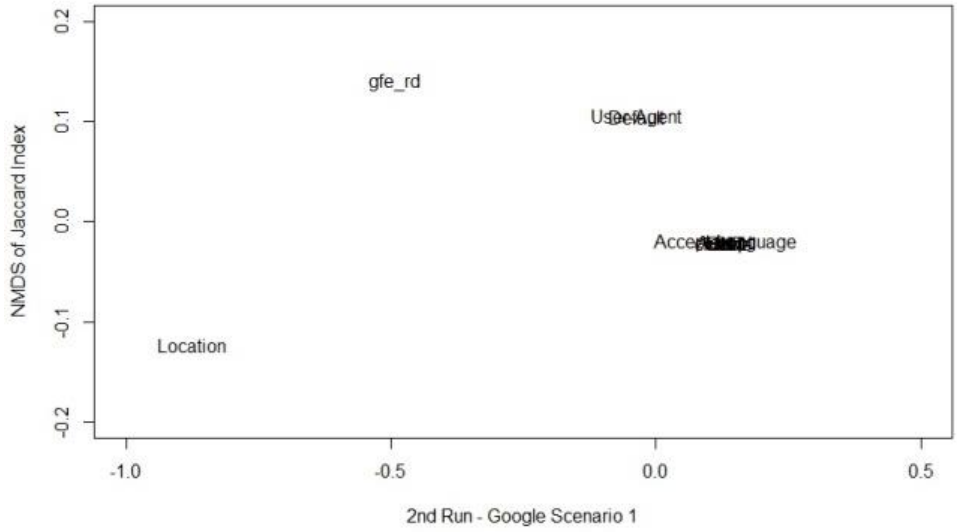


Figure 4.4b NMDS of Jaccard Index: 2nd Run – Google Scenario 1

Generalised Linear Modelling (GLM) is a method that identifies and attributes explanatory factors for particular response variables where the variation is not constant and the errors are not normally distributed (Burnham & Anderson 2002; Crawley 2013). In the context of this chapter binomial generalised linear modelling was used to identify and attribute the variation in the response variable (personalisation variance for query terms, and personalisation variance for query parameters) to the explanatory variables (factors such as search engine and query type, or search engine and query parameter respectively). When analysing personalisation variance computations based on the Wilcoxon

signed rank measure and GLM, particular attention was paid to significance and its associated confidence interval estimate to provide a balanced interpretation of significance (Sauro & Lewis 2012).

In summary, the focus of the controlled experiment was to investigate which types of data affected the variance of personalised search results and which of those contain personally-identifiable information. The effects on data flows were investigated by controlling the data released via an automated search script incorporating query terms, query parameters search scenarios. Personalisation variance was measured using Jaccard Index and Edit Distance (by comparing two search result sets – one was a baseline – the 'minimum' vs. a 'factor' which could be of three types: 'default', 'individual search query parameter', 'location-spoofed'. These personalisation variance comparisons were run across both commercial search engines (i.e. Google, Yahoo, Bing) and privacy-enhanced search engines (DuckDuckGo, Ixquick, StartPage, Unbubble) across three distinct timeframes. To protect the identity and anonymity of users and limit implicit information being transmitted the linkage associated with queries was removed. As evident in Table 4.1, the types of data leaking and the methods of leakage are established. The types of data leaked were identified along with its impact on the personalised search results and if personalisation can still occur at a minimum level of leakage, i.e. removing query link association as shown in Figure 4.3.

4.4 Results, Analysis and Discussion

This section presents the results of the controlled experiment investigating the effects of implicit data released on the variance of personalised search results. The controlled experiment measured the variance in personalised search through changes in re-ordering and difference between search result lists (i.e. Edit Distance and Jaccard Index). Subsequently, for each search scenario across each query term and search engine using the developed script, personalisation variance comparisons were run in the minimum set vs. comparison set format described in section 3.

In this section, the results were categorised thematically to reflect analytical perspectives on the effects of search query terms, search query parameters, search scenarios on personalisation variance and other insights of interest from the data. The analysis focused on whether the implicit data collected had any impact on the variance of personalised results and if this data was personally-identifiable. Rigorous analysis confirms the importance of location on personalisation variance across commercial and privacy-enhanced search engines. The user-agent parameter carrying browser identifiable details is also a significant factor affecting personalisation. Topical query terms were found which are popular informational terms or query terms where the user intent was transactional or navigational had

significant impact on personalisation variance. This section concludes with a discussion of the implications of these findings to a privacy-sensitive user concerned about implicit data release.

4.4.1 Identifying Query Terms that Influence Personalisation Variance

Firstly, the effects of query terms on personalisation variances across privacy-enhanced and commercial search engines over the three experiment datasets are described. An aggregate summary using GLM to reflect personalisation variances of significant query terms across search engines and query categories is provided. Refer to Table 4.2 for a full list of query terms used.

Privacy-enhanced Search Engines: For *StartPage* in general, query terms showed variation in clusters. In the third run for example informational- health, entertainment and navigational terms have significant variation when location-spoofed (see Figure A9 in Appendix A.4.1). Similar variations were seen in *DuckDuckGo* with query terms that showed variations include informational tech, shopping, lifestyle and transactional terms in the NMDS computation (see Figure A14). *Ixquick* had some distinct variation visible for certain query terms; particularly for informational (science, shopping, health, lifestyle, tech, sports, news term types), transactional, navigational terms and topical (popular in the news) queries of 2014 (see Figure A15). In the first experiment run, minimum result lists of privacy-enhanced search engine (StartPage) and commercial search engine (Google) were compared. Some query terms such as transactional terms, health and tech terms showed variations for Jaccard Index with variations seen for transactional, tech and topical 2014 terms for Edit Distance (see Figures A3-A4).

Commercial Search Engines: For *Bing*, clustered variation with some query terms varying distinctly (informational–news, shopping, lifestyle, science, transactional terms, navigational terms and popular 2014 query terms) (see Figure A19). In *Yahoo,* generally there was clustered variation with some query terms varying distinctly (informational–sports, tech, lifestyle, navigational, and transactional terms (see Figure A27). For *Google*, during the 1st experiment run, there was clustered variation seen with some query terms varying distinctly (such as informational–political, lifestyle, quirky, transactional terms, navigational terms and popular 2014 query terms for the NMDS computation of the Jaccard Index- see Figure A1).

Similar variation was seen for the Edit Distance computation. In the 2nd run, clustered variation with some terms varying distinctly (informational –tech, shopping, political, navigational terms and a popular 2014 query term) was seen (see Figure A39). A true anomaly showing a high degree of personalisation was found for the query term ‘Malaysia Airlines’ in Scenario 2 during the 2nd experiment run which had a Jaccard Index of 0.045 and an Edit Distance of 13 showing both vastly different sets of results and

levels of reordering. Similarly, the predicted set of five topical terms of 2014 showed high levels of personalisation when compared to the top 24 terms list of 2013, suggesting that the current terms could possibly have more varying search results given their topical nature. In the 3rd run, query terms are variant, though clustered with some types (informational – tech, navigational, transactional, a popular 2014 term) standing out (see Figure A40).

Overall, the personalisation variance observable across informational, navigational, transactional and popular query terms indicates topicality of query terms for informational and recent terms, and significant personalisation variance for query terms where the user intent was to navigate to a particular site or make a transaction. The variance in query terms are largely representative of results from previous experimental studies measuring personalisation where personalisation variance occurs based on the query terms and its search query category type (Hannak et al. 2013).

However, compared to existing related work, this study found that certain terms could be privacy-sensitive to a user based on what is sensitive and identifiable to a user's own context. For privacy-sensitive users, these findings indicate that with limited implicit data release, personalised search results can be obtained. Personalisation variances can be observed with limited implicit data release due to the effects of the query term being topical or popular, or where the results incorporated localisation (such as maps or local consumer stores where the user intent was navigational or transactional). In addition, the controlled experiment design extends the methods found in related work by focusing on data release when searching through identification of search scenarios and controls for different sources of noise.

For the privacy-sensitive user, having found that personalisation variance is impacted by the topicality of query terms, means that even if such users limit the flow of personally-identifiable information, the variance or the variety in personalised search results can still be obtained, provided the query terms themselves do not contain personally-identifiable information.

Statistical Analysis on Personalisation Variance of Query Terms: Using Generalised Linear Modelling (GLM) via the tool R, for query terms (based on the NMDS computation), navigational query terms are highly variant than informational, transactional or topical query terms (of 2014) when compared across commercial and privacy-enhanced search engines. From this statistical analysis, an association between personalisation variance and informational terms can be inferred. Informational terms also have a much narrower estimate interval than other terms indicating higher precision and less variability, and its confidence interval does not overlap with other estimates signalling a statistical significant difference between it and other estimates. Navigational, transactional and recent 'topical'

terms have a greater influence in commercial search engines as opposed to privacy-enhanced search engines. Conversely confidence interval estimates for navigational, recent 'topical' and transactional query types overlap while each of their estimate intervals is wider indicating higher variability and lesser precision. Overall, these observations allude to the notion of topicality of query terms influencing personalisation variance (when p values are significant). Topicality in this context refers to the popularity of the search term both from an information producer (a large amount of information is produced from diverse sources and search results are available for the query term) and information consumer (a large number of people search for the term) perspectives.

Table 4.5 Generalised Linear Modelling for Query Terms

Coefficients	Estimate	Standard Error	z value (Estimate/ Standard Error)	P value (significant when p<0.05)
Intercept (Control for Query Type is Navigational and for Search Engine is Google)	0.2389	0.2446	0.977	0.3287
Query Type - Informational	-1.5723	0.2468	-6.371	1.88e-10 ***
Query Type - Topical Queries of 2014	-0.3668	0.3246	-1.130	0.2584
Query Type - Transactional	-0.5785	0.3265	-1.772	0.0764.
Search Engine - Bing	0.2988	0.1978	1.510	0.1309
Search Engine- DuckDuckGo	-0.5707	0.3242	-1.760	0.0783.
Search Engine- Ixquick	-0.4804	0.2801	-1.715	0.0863.
Search Engine- StartPage	-0.6465	0.3588	-1.802	0.0716.
Search Engine- Yahoo	0.4983	0.2048	2.433	0.0150 *

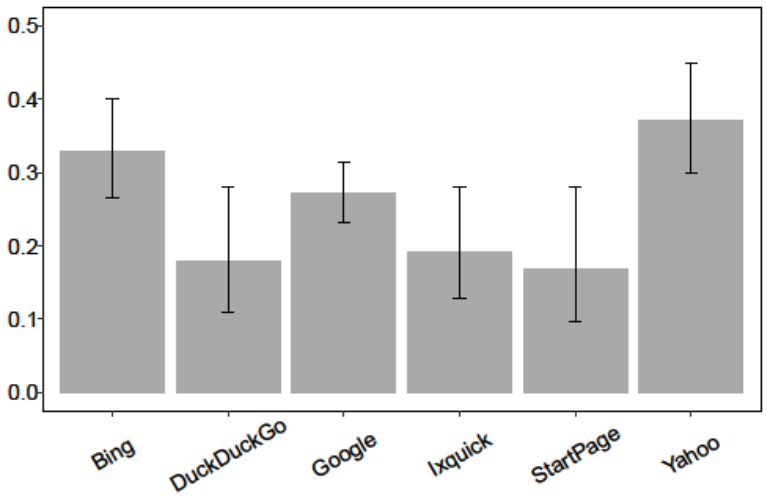


Figure 4.5a GLM- Query Terms – Query Terms Across all Search Engines

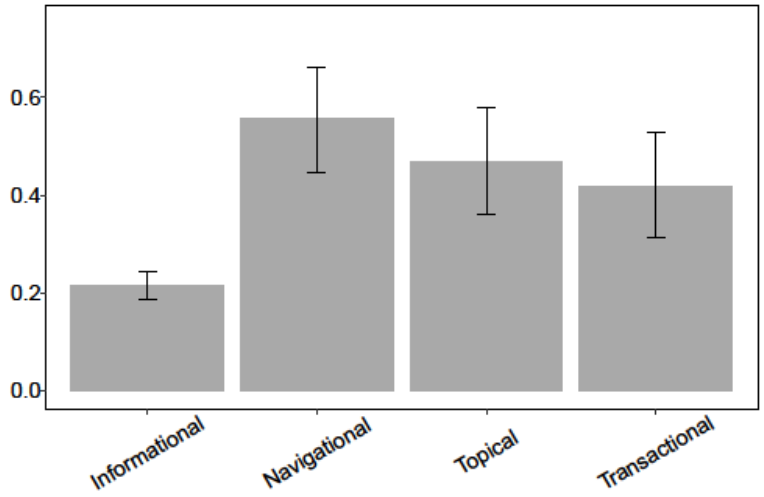


Figure 4.5b GLM- Query Terms –Comparison of Search EngineIdentifying Search Query Parameters that Influence Personalisation Variance

When analysing the association between personalisation variance of query terms across search engines, the variance across query terms there is a statistically significant difference with Yahoo. The variance of Google's search results is lower and more precise as indicated by a narrower estimate interval. The variance of search results from privacy-enhanced search engines seem to have some overlap with Google and Bing, perhaps given that they source search results from these commercial search providers (DuckDuckGo 2015; Ixquick 2015; StartPage 2014b). The variance across query terms is greater across commercial search engines. This variance was expected given that their primary business is in the ranking and rating of search results. See table 4.5 for a detailed view.

The effects of search query parameters released on personalisation variance across privacy-enhanced and commercial search engines over the three experiment datasets is reported in this section. This is followed by an aggregate summary using GLM to reflect personalisation variances of aggregate search query parameters across search engines and query categories. Refer to Table 4.3 and Appendix 4.B for a list of query parameters used.

Privacy-enhanced Search Engines: For *StartPage*, in the first run, significant variance was visible for location when compared across both scenarios for Jaccard Index (W = 841, p-value = 2.721e-12; Figures A5) and Edit Distance (W = 0, p-value = 2.59e-12; Figure A6). In the 3rd run, significant variance was visible for location (W = 840.5, p-value = 4.094e-12; Figure A7) and no variance for User-Agent (W = 421, p-value = 1; Figure A8) as both user-agent and default values are the same. Query parameters show variation in 2 clusters with location identified as significant as it is a sole separate cluster (see Figure A10). Location spoofing generated huge variances in personalisation on StartPage (See Figure A10). This is interesting given that StartPage claims that it does not *record* a user's IP address (StartPage 2014a). However, a user's IP address affect search results from Google as shown by the variance in different results for the same query term whilst having one query term with the location spoofed.

The variance in personalisation with geolocation followed a similar trend in comparison to personalisation variance from other search engine results, with exceptions relating to shopping, health, transactional and navigational terms. Possible reasons for these variations may be due to specific terms being more topical in the US and the presence of different online shopping retailers (See Figures A5 and A6). Therefore, location may be considered as personally-identifiable based on the context of its use when searched. Overall, search results from StartPage may not be overly personalised, at least not using users' personally-identifiable data or browser information as there was negligible difference between the minimum subset (i.e. the query term) and the default set of parameters. Also of interest is that no obvious cookies are being sent and received with requests and responses, limiting implicit data leakage.

Similar personalisation variances were observed with *DuckDuckGo* across the 2nd and 3rd data sets. Key representative examples are described here. Significant variance was visible for location (W = 529.5, p-value = 0.03903; Figure A11), and user-agent (W = 319, p-value = 0.004786; Figure A12). However, this is mostly due to the variation caused by the minimum set comparisons, rather than the user-agent parameter itself. Query parameters vary in clusters but some distinction is shown for the user-agent parameter (see Figure A13). In *Ixquick*, similar personalisation variances were seen across the 2nd and 3rd experiment runs. Key representative examples are described here. Query parameters significant

(p<0.5) for location is significantly variant (W = 29.5, p-value = 0.09824; Figure A16), Host (W = 429.5, p-value = 0.3638). User-Agent is significant though not consistently (W = 14.5, p-value = 0.3991). The parameter engine0 was significant and unique to Ixquick (W = 21, p-value = 0.02354; Figure A17). When the NMDS of query parameters was computed, overall there was variation but no distinct clusters were observable (see Figure A18).

Commercial Search Engines: When comparing the minimum result lists from *commercial* (Google) and *privacy-enhanced* (StartPage) search engines in the 1st experiment run, variances were significant when location-spoofed for both Jaccard Index and Edit Distance (Figures A3-A4). In *Bing*, there was significant variation (p<0.5) shown for location (W = 534, p-value = 0.0689; Figure A21) and user-agent (W = 218.5, p-value = 9.733e-05; Figure A20) though not strong variation. The cookie _EDGE_V (W = 469, p-value = 0.1871; Figure A22) varied significantly. Other cookies MUID (W = 401.5, p-value = 0.7697; Figure A23) and MUIDB (W = 416.5, p-value = 0.9558; Figure A24). In the NMDS computations, there was visible variation. In one example, query parameters such as location and the cookie _EDGE_V are variant and separate from the other cluster of parameters (see Figure A25). In another example, query parameters such as location, default, and the cookies _EDGE_V, MUID, MUIDB are variant and separate from the other cluster of parameters (see Figure A26). With *Unbubble*, perceivable differences in personalisation variance for search query parameters or query terms were not observed. However, the presence of a derefer header, which kills the referer heading from the originating search engine was detected, but Unbubble could plausibly capture clickthrough data (because it's redirected to their server). Unbubble are open about the use of this referer header and invite users to use this as a means to conceal or obfuscate query links (Unbubble 2016a).

In *Yahoo*, personalisation variance was dynamic and inconsistent. In some cases, the parameters hspart (W = 560, p-value = 0.0282; Figure A28) and hsimp (W = 580.5, p-value = 0.01231; Figure A29) and the header Referer (W = 564.5, p-value = 0.02444; Figure A30) were significant. Other parameters that showed significance include ei (W = 305, p-value = 0.07251; Figure A31), DSS (W = 295, p-value = 0.05066; Figure A32), ypcdb (W = 288, p-value = 0.03922; Figure A33), sSN (W = 263.5, p-value = 0.01408; Figure A34), Location (W = 534, p-value = 0.0689; Figure A35) and User-Agent (W = 218.5, p- value = 9.733e-05; Figure A36). When computing the NMDS, query parameters such as user-agent, location and default are clustered on one side and separate from the other cluster of parameters. In another example, query parameters such as hspart, hsimp, accceptlang, accept are variant and separate from the other cluster of parameters (see Figures A37 and A38).

During the 1st experiment run for *Google*, for the NMDS computation of Jaccard Index, query parameters such as location, gfe_rd and the cookie GoogleAccountsLocale are variant and separate from the other cluster of parameters (Figure A2). Edit Distance showed similar variation along with cookies NID, PREFID showing some variation. In the Wilcoxon computation of Edit Distance, significant variation ($p<0.5$) visible for User-agent ($W= 25.5$, p-value = 0.272), the parameter gfe_rd ($W = 3$, p-value = 3.319e-06) and location ($W = 14$, p-value = 1.616e-05). Variances were similar for the Jaccard Index measure. In the third iteration of the experiment run, significant variation ($p<0.5$) shown for location ($W = 471$, p-value = 0.2172; Figure A43) and user-agent ($W = 756$, p-value = 1.975e-08; Figure A44). When computing the NMDS of search query parameters, in the 2nd iteration, query parameters such as location, user-agent, gfe_rd and default are variant and separate from the other cluster of parameters whereas an example from the 3rd iteration, shows query parameters are in two clusters with user-agent and location in one (see Figures A51 and A52).

Cookies set on initial search: Contrary to expectations and findings from previous studies (Hannak et al. 2013; Kosta et al. 2009; Krishnamurthy, Malandrino & Wills 2007; Malandrino & Scarano 2013), cookies by themselves largely did not appear to have any significant impact on the variance of retrieved personalised search results. For example, performing a search in Google via the search engine page set two anonymous cookies PrefID and NID which tracks the user anonymously, whilst also transferring a Referer header in some cases. Both account login scenarios were the least anonymous (one with search history and the control account without search history) and set 6 additional cookies namely GoogleAccountsLocale, SID, HSID, SSID, APISID and SAPISID.

Given that cookies from the point of an initial search were utilised (cookies set when users first search either via the search engine page or through account login, and not through subsequent searches where cookie values may change based on clickthrough data), these cookies that are initially set did not impact personalisation variance other than exceptions seen with Google (PREF_ID, NID) and Bing (_EDGE_V, MUIDB). Cookies contain personally-identifiable data. This finding supports the premise that the data stored in the cookies is valuable from a search engine perspective in terms of preserving user interaction across multiple sessions. However, the impact cookies have on personalisation variance is limited when each cookie query result set is compared against the minimal result set. The effect of subsequent cookies on personalised search results is identified as an important avenue for future work on further analysis of the data flows in search. Cookies have a limited effect in this experiment because of the difference in scope and experiment method compared to related work. The method only focused on initial queries and

not subsequent queries. The findings are constrained to this scope. In other words, cookies do not have a significant impact on the variance of personalised search results for initial queries.

Search Query Parameter Values. Table A4.2.1 in Appendix 4.2 lists the different types of search query parameters encountered across the scenarios and query terms. The purposes of these parameters were derived from official documentation where available or from relevant exploratory sources (Google 2014b, 2014a, 2015; Google Operating System Blog 2006; Internet Engineering Task Force 2014; StackExchange 2014; Wailes 2008). While formal definitions for some of these search query parameters exist, the key purpose was to evaluate if the data impacted personalisation variance and/or contained personally-identifiable information. For example, for Google, certain search query parameters had constant values across all four scenarios (GET parameters *ie, oe, aq, rls, client and channel*; the HTTP headers *Host, User-Agent, Accept, Accept-Language* and *DNT*). Other search query parameters were present based on the scenario such as the GET parameters: *btng, hs, gbv* and the Header: *Referer* for Scenario 2. Cookies followed the same pattern with *PREFID* and *NID* being fixed cookies across Scenario 2, 3 and 4 whilst, *SID, HSID, SSID, APISID and SAPISID* were constant across Scenarios 3 and 4 (having account login) with the *GoogleAccountsLocale_session* cookie exclusively present in Scenario 3.

Geolocation and Personalisation. Location generated a huge variance in personalised search results across all query terms, particularly with commercial search engines. The effect of geolocation was attributed to the change in IP address. In particular, transactional queries such as those seeking a public transport site 'Adelaide Metro' and online pizza ordering via 'Dominos' had largely different result sets which for the location-specific public transport site is unusual, but expected for 'Dominos'. Similarly, predicted top 2014 terms like 'ISIS' and 'Malaysia Airlines' were highly varied in personalisation through location spoofing. Large differences in personalised result sets produced substantial variances in reordering as expected. This is similar to results from prior related work where location, user-agent attributes and the minimum vs. default comparison yielded significant reordering of results (See Figures A3 and A4) (Fernando, Du & Ashman 2014; Hannak et al. 2013; Krishnamurthy, Naryshkin & Wills 2011). Hence, even though data leakage was minimised (location was tightly controlled via cURL and an external proxy when location was spoofed), personalisation varied greatly just based on the query term and location.

Overall, across both commercial and privacy-enhanced search engines location was consistently observed as a key factor affecting personalisation variance. Location can be highly privacy-sensitive

depending on its context of use. User-agent (the type of browser and operating system – which is an HTTP header sent across with each search request) may not be personally-identifiable on its own, but coupled together with location and identifying data points such as time zones, language, can be used to browser fingerprint and uniquely identify browsers with high degrees of accuracy (Eckersley 2010). Currently search engines organisation and to some extent (if captured), any information service provider would have access to this data as it is freely transmitted in search queries given the nature of the protocol. Persistent identifiers such as location and browser fingerprinting data have recently been considered personally-identifiable by the Federal Trade Commission in the US because it can be used to distinguish users across multiple independent search sessions (Rich 2016). The high likelihood of browsers being uniquely identified raises further implications such as if information service providers like search engines are able to personalise more effectively based on fingerprint data acquired from their own browsers, but could be a scenario for future work that extends this controlled experiment design method.

User-agent had significant personalisation variance across commercial search engines and some privacy-enhanced search engines (DuckDuckGo, Ixquick with StartPage not showing consistent significance). The control for user-agent (changing it from default value of Mozilla Firefox to Google Chrome and Internet Explorer) also showed similar significant personalisation variance. The referer header showed some significant impact across both commercial and privacy-enhanced search engines similar to prior studies (Krishnamurthy, Malandrino & Wills 2007; Malandrino & Scarano 2013). It carries the clickthrough header info which is used to capture and analyse user feedback with search results.

Some cookies had a significant effect on personalised search results (EDGE_V (Wilcoxon and NMDS) and MUIDB specific to Bing for Wilcoxon and PREF_ID and NID (NMDS) specific to Google) although this was inconsistent across the longitudinal timeframe. The parameters hspart and acceptlanguage also showed some significance at times for personalisation variance in search results across Yahoo. Most GET/POST parameters also had insignificant impact on the level of personalisation in comparison to related work (Krishnamurthy, Malandrino & Wills 2007). Overall though, location and user-agent seemed to show the most observable significant personalisation variance across commercial and privacy-enhanced search engines.

Statistical Analysis on Personalisation Variance of Search Query Parameters. Using Generalised Linear Modelling (GLM) variance among search query parameters across both commercial and privacy-enhanced search engines was influenced significantly and consistently when the location was spoofed

and by the user-agent parameter (See Appendix 4.1 for selected figures or for detailed information refer table A4.3.1 in Appendix A4.3). Generalised linear modelling was used to determine the causal effect between two factors (Rosenthal & Rosnow 2008). Generalised linear modelling was applied to this case, using the key search query parameters affecting personalisation variance across search engines based on the Wilcoxon test measure (see Table 4.6). Location and user-agent (the type of browser and operating-system) are statistically significant parameters as evident in Figure 4.6a and Table 4.6. However, the mean and the confidence interval for location show more consistent and significant variation when compared with the user-agent suggesting that location is a significant factor in Figure 4.6a.

Commercial search engines have higher means indicating greater personalisation variance. In table 4.6, when compared the variance of personalised search results, Bing and Yahoo (commercial search engines) differ significantly from Google, but this is less evident for privacy-enhanced search engines. This indicates a greater reliance on location and user-agent as key search query parameters affecting personalisation variance across commercial search engines. Location is a more consistently significant factor affecting personalisation variance. In Figure 4.6b, the confidence interval for Google's variance of personalised search results is much narrower as opposed to the other search engines, owing to a greater degree of consistency in variance of personalised search results and precision across the estimate intervals (as opposed to the personalisation variance shown by privacy-enhanced search engines). Although search engines routinely optimise search results, and incorporate improvements, this does not seem to affect the stability of personalised search results over a longitudinal timeframe, which is in line with expert commentary on the topic (Chi 2015).

Table 4.6 Generalised Linear Modelling for Query Parameters

Coefficients	Estimate	Standard Error	z value (Estimate/ Standard Error)	P value (significant when p<0.05)
Intercept (Control for Query Parameter is Location and for Search Engine is Google)	1.59546	0.35370	4.511	6.46e-06 ***
Query Parameter-User Agent	-1.51555	0.36878	-4.110	3.96e-05 ***
Search Engine -Bing	-0.83769	0.51359	-1.631	0.103
Search Engine-DuckDuckGo	0.95959	0.84384	1.137	0.255
Search Engine-Ixquick	-0.83769	0.63305	-1.323	0.186
Search Engine-StartPage	-0.03684	0.77978	-0.047	0.962
Search Engine- Yahoo	-0.01643	0.50663	-0.032	0.974

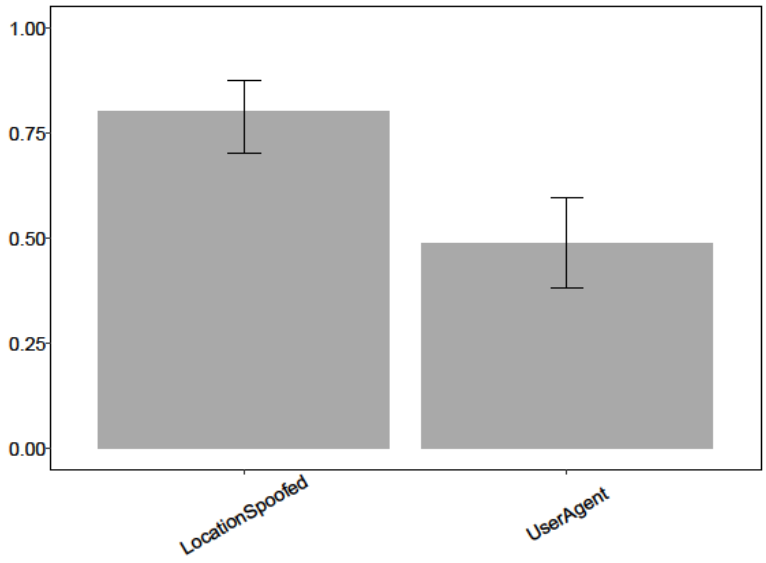


Figure 4.6a GLM- Query Parameters – Location vs. User-Agent Across all Search Engines

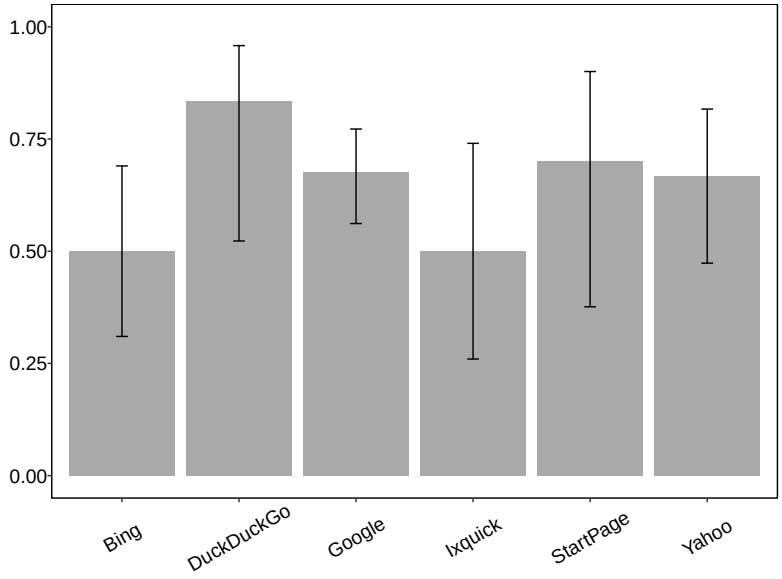


Figure 4.6b GLM- Query Parameters – Query Terms Across all Search Engine

Experiment Control: With the additional experiment controls, similar variances were observable. When data release was randomised similar variations were shown for Jaccard Index and Edit Distance using the Wilcoxon Test measure. Significant variation ($p<0.5$) shown for location ($W = 342$, p-value = 0.1298; Figure A45) and user-agent ($W = 82.5$, p-value = 3.256e-08; Figure A46). However, user-agent is not so significantly variant across all scenarios (as much as location-spoofing is). This was similar in the NMDS computation of personalisation variances. When randomised, location user-agent and default distinctly variant from other query parameters (see Figure A53). For the exclusion effect, query

parameters such as location, user-agent and default are variant and separate from the other cluster of parameters (see Figure A54). In some cases, there's further correlation supporting strong similarity between location, gfe_rd (redirect parameter –significant only in some cases) and user-agent (see Figures A55 and A56). When randomised, clustered variation with some query terms distinctly seen (informational – news, navigational, popular 2014 query term) was seen (see Figure A41). There is evidence to support the argument that the experiment order has no perceived effect on the personalisation variance comparison results.

For the *exclusion* effect, there is similar variation. User-agent is less consistently significantly variant than location with significant variation ($p<0.5$) shown for location ($W = 327$, p-value = 0.05362; Figure A47) and user-agent ($W = 391.5$, p-value = 0.4999; Figure A48), similar values for personalisation variance for default and user- agent resulting in a few points of difference. When the *user-agent was changed*; using Google Chrome there was significant variance for the user-agent ($W = 15$, p-value = 1.022e-07; Figure A49). With Internet Explorer, the p value was just slightly greater than 0.5 for the user-agent ($W = 449$, p-value = 0.6605; Figure A50). Hence, similar values for personalisation variance were observed for default and user-agent resulting in a few points of difference. For query terms with the exclusion effect, clustered variation was observed with some terms varying distinctly (informational–tech, shopping, navigational terms and a popular 2014 query term) (see Figure A42).

Longitudinal Time Perspective of Personalisation Variance. The overall experiment was run *longitudinally* over 2014-2016, which allowed data collection at three discrete time points across the experiment iterations (October 2014, May 2015 and January 2016 with a gap of at least 7 months). Potential instability of search results retrieved and are mainly due to search engine development testing, and changes in rank and reorder mechanisms or algorithms. When findings on personalisation variance were aggregated over time using statistical measures like NMDS and GLM, this aggregation controlled for potential instabilities that may be present in specific experiment runs. These results indicate there is evidence to support that search engine optimisation and other time related effects have little impact on personalisation variance of search results and subsequently, when exploring the removal of personally-identifiable data in search results.

Personalisation without query association. The experiment method verified that personalisation still occurs without explicit personal data (such as clickthrough data from JavaScript; stored cookies) being required. This is similar to related studies using quasi-identifiers to de-anonymise and sanitise search query logs (Toubiana, V & Nissenbaum 2011) . This was achieved by breaking the association between

linked or subsequent queries. Hence, the link between queries and search history can be removed but still maintain contextual or personalisation relevance. This effectively demonstrates that the leakage of data can be prevented or minimised depending on the user's choice.

4.4.2 Search Scenarios

The Jaccard Index difference in personalisation variance between the minimum set and the default set of parameters was significant in Scenario 1 when compared with the rest of the scenarios in Google. This finding was counter-intuitive given that the account login-based personalisation (Scenarios 3 and 4) might be expected to impact personalisation variance more than Scenario 1. The variance in the search histories of the accounts used and the variance in the browser fingerprint data offered to the search engine were controlled for as explained in Table 4.1. The variance in personalisation can be attributed generally to the user-agent (browser) parameter, showing a degree of correlation between the user-agent parameter and default set of parameters for a given scenario. This correlation is unusual as it suggests that personalisation is impacted by the type of browser and operating system employed, raising questions of relevance of results to users. However, as users can be uniquely identified based on their browser and operating system versions; this raises significant identifiability concerns, especially since this information is leaked implicitly with every query submitted. The impact of this finding is in line with growing interest in browser fingerprinting, a related area of work which aims to protect the identifiability of browsers as a means to identify users (Castelluccia, De Cristofaro & Perito 2010; Krishnamurthy, Naryshkin & Wills 2011). These findings highlight that account-login does not necessarily produce better variances in personalised search results, as you can obtain the same level of variant personalised search results without being logged into accounts associated with commercial search engines (where Web search history of user interactions are collected across its range of information services). This finding is contrary to popular thought and publicly available information on how commercial search engines provide personalised search results (Enge 2011; Hannak et al. 2013).

4.4.3 Summarising Significant Implicit Data Released in Personalised Search

A summary of significant search query parameters and query terms affecting personalisation variance across commercial and privacy-enhanced search engines is described in Table 4.7. Variance of personalised search results is mostly impacted by the topicality of the query term (which are popular informational terms or query terms where the user intent was transactional or navigational had significant impact on personalisation variance), user-agent and location for personalised search results.

Other personally-identifiable data such as cookies and certain headers have a negligible effect on the variance of personalised search results.

Table 4.7 Experiment Findings – Summary of Analysis on Personalisation Variance

Search Engine	Significant Search Query Parameters	Significant Query Terms	Longitudinal Consistency & Stability
Commercial Search Engines			
Google	Location, User-Agent, Minimum vs. Default query set comparison, Referer header with some inconsistent variation seen across the cookies PREF_ID, NID	Significant variance across topical informational, transactional, navigational and popular 2014 terms	Over 3 timeframes
Yahoo	Location and User-Agent with some inconsistent variation seen across the parameters hspart and acceptlanguage	Significant variance across topical informational, transactional and navigational	Over 2 timeframes
Bing	Location, User-Agent, Cookies (MUIDB, _EDGE_V)	Significant variance across topical informational, transactional, navigational and popular 2014 terms	Over 2 timeframes
Privacy-enhanced Search Engines			
StartPage	Location (consistently)	When location-spoofed, variation is greater across topical informational and navigational terms	Over 2 timeframes
Ixquick	Location (consistently). User-Agent, engine0 and referer header (inconsistently)	Significant variance across topical informational, transactional, navigational and popular 2014 terms	Over 2 timeframes
DuckDuckGo	Location and User-Agent (inconsistent though significant when variation is observable)	When location-spoofed, variation is greater across topical informational and transactional terms	Over 2 timeframes
Unbubble	No perceivable personalisation variance can be detected using this experiment method; use of referer headers capturing clickthrough data is visible		N/A; tested during the 3rd experiment run (2016)

4.5 Conclusion

This chapter reports on the effectiveness of technical mechanisms users are afforded in releasing data when searching. This chapter describes a controlled experiment investigating the variance of retrieved personalised search result lists with the aim of identifying the effects of implicit data released. Location, user-agent and topicality of the query term have a greater impact on the variance of personalised search results. The findings provide evidence that implicit data released through queries, can be selectively suppressed or released to obtain personalised search results. There is a need to qualitatively assess whether these data release or suppression techniques provide satisfactory personalised search results for users and if these privacy-sensitive data release techniques could provide the basis for improved affordances for expressing user privacy expectations.

4.5.1 Understanding the Effects of Implicit Data Release on Personalisation Variance of Search Results

Contributions: Personalisation variance was affected by factors such as location, type of browser used (user-agent) and topicality of the query term (which are popular informational terms or query terms where the user intent was transactional or navigational had significant impact on personalisation variance). Location in particular is likely to be personally-identifiable to an individual. Geolocation is used to provide search results based on a user's location. Whilst this may be helpful for most location-specific query terms, for generic query terms users may prefer results from a broader geographic area. Additionally, location could be aggregated with other data sets to infer details about a person's identity such as places they frequent often. Most GET parameters and HTTP headers implicitly transferred were not sources of personal data. The query term itself, depending on whether the keyword is topical, influences the personalised search result set. Interestingly, the browser header containing the type of browser and operating system impacts personalisation.

Certain user scenarios are more anonymous and provide fewer opportunities for implicit data leakage. For example in Scenario 1, where a user performs a search using the browser bar, was the most anonymous of the four, as no cookies were set and browser identifier data is at a minimum. However, account-login in commercial search engines does not necessarily provide improved variances in personalised search results. Cookies surprisingly had little impact on personalised search results across the 29 query terms and four typical user scenarios. Personalisation variation may not be influenced by cookies alone, at least not in the initial data flow of search queries, given the isolated experiment environment. Cookies could be used for ensuring Web session continuity and could also be a useful form of data collection. The captured

cookies leaked through the initial search showed no dependency on the search results and were fault-tolerant (this was assumed and later validated as cookie values could be submitted multiple times without adverse impacts on search results). However, this may change with time, subsequent searches and a more longitudinal browsing or search history data.

The major contribution of this experimental study is an analysis of how the different types of implicit data impact the variance in personalised search results in both modes of personalisation (non-account login and account login) search scenarios, which advances the current understanding around the effects of data release on the variance of personalised search results in literature. Personalised search results can be generated with limited implicit data leakage and reduced ability to distinguish queries. This ability to have personalisation with limited identifiability would be highly appropriate in some contexts. One example includes restrictions involving corporate communications where the sensitivity of searches conducted and the identifiability of users is minimised. From an individual privacy point of view, this would empower the user who chooses to limit the leakage of personal data when searching, especially pertaining to certain sensitive contexts. From a national legislation angle, this would enable personalisation to still occur without excessive collection of data or the potential for misuse or inaccurate inferences from this aggregated personal data.

Applicability: The scope of the controlled experiment was constrained to using a single user persona, partly to assess if implicit data had any effects on personalisation variance. Having found the effects of implicit data release on personalisation variance, the qualitative assessment of personalised search results will be scaled with a larger sample of 50 users. Also, only organic Web search result links were investigated for the scope of this controlled experiment. However, there is a possibility to expand the same method to explore other search result links like news, images, advertisements or sponsored search results. The first step in data release in search was explored as it provides a better way to effectively manage data release. Similarly, the experiment method can be expanded upon to investigate the effects of implicit data release with subsequent personalised search results.

4.5.2 Identifying Affordances for Privacy-Sensitivity in Personalised Search Results

Contributions: The significance of the study lies in the value derived from identifying the effects of implicit data in producing personalised results. In general, personalisation variance is impacted by the user-agent, location, query term topicality across commercial and privacy-enhanced search engines. Most variables impacting personalisation variance based on these results are variables that users cannot control, whilst some user scenarios provide a minimum level of data leakage.

Two distinct roles played by implicit data in personalised search can be inferred. These are: data that actively (measurable variance) and passively (negligible variance) impacts personalised search results, with some of this data being personally-identifiable. User-agent and location are instances of personally-identifiable data that actively impact personalisation variance with significant implications of use of such data. For example, the aggregation of a user's location data over a period of time coupled with other data sources that uniquely identify users such as through browser fingerprinting may make it possible to infer potentially personal or sensitive details. Persistent identifiers such as location and browser fingerprinting data have recently been considered personally-identifiable by the Federal Trade Commission in the US because it can be used to distinguish users across multiple independent search sessions (Rich 2016). Information service providers have a mandate to convey how their services use these persistent identifiers in enabling users to make informed decisions.

Privacy-sensitive users can obfuscate their location and user-agent and still obtain personalised search results by reducing the implicit data released as indicated by the findings. Obfuscating is an important goal and although simplistic, the strength of an obfuscation technique is in its purpose and goal (Brunton & Nissenbaum 2015). Meaning of data in a given context is important in deriving its value (Boyd 2014). In this context, by obfuscating the location and user-agent, the meaning associated with each individual search query or data release was disambiguated, so that users can still obtain personalised search results and technology affords users a proactive choice in terms of implicit data release. This finding implies that by obfuscating data released when searching, the variance of personalised search results does not largely diminish. This is an important finding that highlights the importance of technical mechanisms and the role they can play in affording privacy if appropriately implemented.

4.5.3 Evidence of Information Asymmetries with Data Release in Personalised Search Data Flows

Contributions: **Invisibility of Data Flows:** A majority of the data release in search flows implicitly (see Figure 4.7a). When these data flows occur without the user's awareness or understanding, they were categorised under implicit leakage. Users most often are consciously aware of just the query term they type in. Account logins associated with commercial search engines can be either explicit or implicit depending on if users are aware of being logged into accounts when searching. Two strategies for dealing with the invisibility of data flows from a privacy-sensitive user perspective are recommended (see Figure 4.7b).

Firstly, data can be *minimised* when transmitted with search query requests. A parallel line of thought calls for leaner HTTP requests in HTTP/2 through client-side cache digests purely from a performance improvement standpoint (Nottingham 2016). These findings contribute to the viability of a similar ideal of minimalism in search query requests and the potential it could serve the purposes of maintaining pseudonymity of users or reduced distinguishability of queries through obfuscation.
Secondly, if this ideal of minimalism cannot be achieved, or is impractical in certain scenarios (such as transactional or navigational queries where disclosing location is a conscious and active user choice) then there should be *increased visibility* of such data flows for users, to aid understanding and avoid misinformation. While modern Web browsers aim to warn users about possible location leakage, a website might be able to transmit location in a (possibly encrypted) cookie, to bypass browser protection. Similar design recommendations (privacy patterns) were suggested for Web-based interactions; namely informed consent, masked online traffic and minimal information asymmetry (Romanosky et al. 2006). The notions underpinning these design recommendations for privacy-sensitive affordances are extended by investigating its applicability in practice as a proactive measure to afford the user with managing data release before it actually occurs (as it is at the base of technology-mediated interaction for data flows in search).

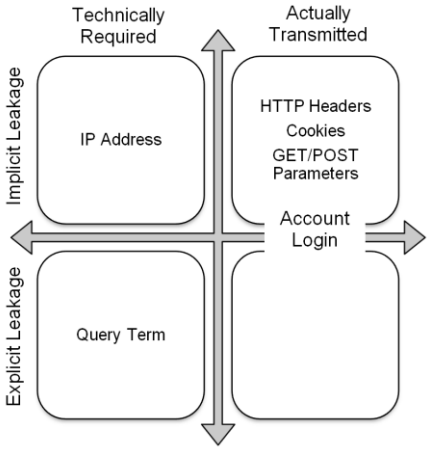


Figure 4.7a Current Data Flows in Search

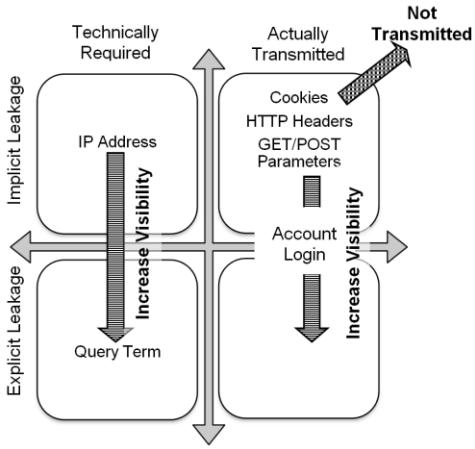


Figure 4.7b Proposed Privacy-Sensitive Data Flows in Search

Contributions: Implicit Data Leakage Inhibits User Choice: Generally, users trade their privacy (through implicit data loss) for the benefits of personalisation. This exploratory study provides value by keeping the nature of data loss to a minimum and reducing the distinguishability of each query. This enables protection of privacy principles as affirmed in national legislation as only minimal data required for the service (search) to function is released. Corporate communication restraints may also benefit

from this identification, as employees searching sensitive or personal data for business purposes may leak only required data with the ability to distinguish individuals kept to a minimum.

Currently, there is no real choice if users wish to consume information via search providers. The current weaknesses of existing consent mechanisms such as notice and consent hinder informed consent because it makes use of the vulnerabilities of users via a lack of knowledge or access to such knowledge to make rational decisions, and the large amounts of processing time required to make a decision with the way information is currently represented to users (Pasquale 2015; Solove 2007).

If users express concerns with these existing mechanisms, they are often faced with a choice that is either denial-of-service or to keep using these inadequate affordances, with convenience and ease of effort being key drivers in decision-making (Nissenbaum 2010). This is ineffective in building trust and transparency with personal data flows because existing consent methods are misinforming. Increasing the visibility of implicit data use builds trust and empowers users as the value proposition would be openly articulated. This visibility of data use is a key step in enabling users to understand the extent of their data use (including who has access to what kinds of data) and resolving the indeterminacy effect concerning users (Ortlieb 2014). This finding contributes towards the importance of effectively providing visibility of data flows which validates claims made in related work such as those of Nissenbaum (2010) and Ortlieb (2014).

Contributions: **Implications of Power Asymmetries:** The information asymmetries observed in data flows create power asymmetries where users lose agency in the release of data when consuming services. This power asymmetry and loss of agency creates negative user interactions and experiences as evidenced by the growing practice of providing false information to use a service to avoid privacy loss (resulting in inaccurate data being collected (Fuster 2010).

However, if organisations improve transparency around data collection and trust with users, the level of inaccurate information can be reduced. The lack of visibility of how data is used creates a socio-technical vulnerability that is often manipulated by technology organisations complying with existing data practices but circumventing the impact of this in practice (through prescribing how information is shared in data collection practices (Koene et al. 2015)). It is imperative to investigate proactive approaches to managing personal data release as most existing approaches are reactive and may be ineffective in remedying privacy loss. Location for example can be used to track and distinguish a person's activity and by extension, infer aspects of their identity (Clarke & Wigan 2011). This systemic

and pervasive tracking gives rise to the notion of dataveillance, or surveillance capitalism, a similar scholarly school of thought, where monopolistic markets such as search have a dominant influence in the consumption of information by users (Clarke & Wigan 2011; Luca et al. 2015; Zuboff 2015).

Online interactions should be trust-based value exchanges, but users are apprehensive in trusting organisational structures that resemble black-boxes where interactions are continuously tracked and plausibly permanent (Dewandre 2015b; Pasquale 2015; Zuboff 2015). In a hyperconnected world, human interactions leave a trace by default, calling for new affordances that can adapt to balancing competing interests of accountability of data-driven service providers and the freedoms exercised in privacy (Dewandre 2015b). There is growing interest in transparency and accountability of information governance to minimise these power asymmetries, but the onus is on the service provider whose interests are mainly commercial (Dewandre 2015b; Pasquale 2015). While ethical and legal frameworks need to expand to meet these changing needs (Mayer-Schönberger 2015), these findings contribute to the need to create socio-technical affordances which are embedded in the design of technology, and reflects the moral values and privacy expectations of users around personal data release. In this context, by minimising information asymmetries, users are afforded a more reasonable expectation of privacy.

In conclusion, prior work has studied methods of data leakage in relation to personal data in online interactions and its implications. This controlled experiment study on commercial and privacy-enhanced search engines contributes to the existing body of work, by identifying the effects implicit data released have on personalisation variance of search results. Location, the user-agent parameter and query term topicality (which are popular informational terms or query terms where the user intent was transactional or navigational had significant impact on personalisation variance), have a significant impact on personalisation variance. When location and user-agent are obfuscated, there is a negligible effect on the variance of personalised search results, giving rise to an opportunity to limit distinguishability of queries, particularly for privacy-sensitive users. It is recommended that by minimising data released and increasing the visibility of data flows in search, users are provided with proactive privacy choices. The effects of these privacy-sensitive affordances around data release from a user-centric perspective will be investigated in Chapter 5.

5 Empirical Investigations: Privacy Expectations of Users when Searching

5.1 Introduction

This chapter investigates the societal expectations that shape the privacy protections afforded to users. Specifically, this chapter reports an empirical investigation on online privacy expectations that users experience when they search online. The empirical gap this chapter addresses is to build an understanding of the online privacy expectations and social norms that govern data flows and influence the integrity of the contexts in which data flows (see Section 2.3). Central to maintaining the integrity of socio-technical interactions is an understanding of online social norms, which at present is lacking in clarity and depth. These socio-technical observations underpin the motivation for an empirical study on privacy expectations around the flow of data during search This exploration of privacy expectations is grounded by the contextual integrity framework (Nissenbaum 2010). Using techniques derived from the prior technical controlled experiment in Chapter 4, search behaviour was quantitatively analysed in terms of awareness of data release, effectiveness of existing privacy-enabling mechanisms, and usefulness of search results produced with minimal (query term and IP address) and obfuscated (location and user-agent) data release (Section 5.2)

Narratives that participants provided including how they experienced privacy online were also categorised and thematically analysed as in the form of data flows (Section 5.3). These expectations were further evaluated using the decision heuristic outlined in the contextual integrity framework (Nissenbaum 2010). Key findings include factors affecting privacy-sensitivity such as evidence of network effects, passive acceptance, the ‘creep’ factor, technical literacy and a contextual dependence in sharing data for primary purposes, actors and conditions. Practices that impact the appropriateness of data flows include collection and retention of metadata, collection and distribution of data for commercial purposes and framing choices through ineffective affordances.

Privacy expectations around data flows were explored using a survey comprising of a mix of close-ended questions (multiple choice Likert scales, short-form text responses and multiple selection options) and open-ended questions (long-form text responses). Responses from 441 students and staff were received. The survey constructs are presented in Table 5.1 (see section 3.1.2 for methodological details on the survey design and Appendix 5.1 for the detailed survey) and includes references to specific survey questions as denoted in [] under factors.

Table 5.1 Survey Study Design

Factor	*Rationale*	
Demographics [Q1]: Age, Gender, Role, Area of Expertise, Associated Culture, Heritage or Ethnicity	To provide a description of the survey sample.	
Quantitative Questions		
Search Behaviour [Q2]: Reliance on Web, Level of Expert, Browser Use, Search Engine Use, Preference & Frequency, Device Frequency	To understand characteristics around how users search.	
Awareness of Personal Data Release and Collection [Q3a-c]: Awareness of tracking, accessibility & use of personal data in searches	To identify level of awareness of how personal data flows in search.	
Expectations around Data flows [Q3d-i]: Public availability of search history, Release and use of personal data, Linkability of data, Capture of clickthrough data, Tracking search behaviour, Aggregation of data	To identify level of expectations around different types of data flows.	These expectations were identified based on a synthesis of literature review findings such as (Awad & Krishnan 2006; Teltzrow & Kobsa 2003; Toch, Wang & Cranor 2012).
Consent and Disclosure Mechanisms [Q3o-v, Q6]: Choice on Identifiability & Anonymity, Transparent notice and indicative level of transparency, Control around collected data, Notification of security of data, Notice and Consent Privacy Policies, Length and Complexity of Terms and Conditions	To identify expectations around existing consent and disclosure mechanisms.	General questions around the expectation of data flows and consent and disclosure mechanisms were used to ascertain if there was evidence to prove/ disprove the findings from chapter 4.
Privacy Risks and Impacts [Q3j-n]: Unauthorised access, Identifiability, Monetizing personal data, Accessibility of data to family/friends or co-workers/employer	To identify level of expectations around typical privacy risks associated with searching. These risks were identified based on (Solove 2009)	
Privacy-enabling Mechanisms [Q4]: Awareness and Use of Privacy Policy Mechanisms and Privacy-enhancing Technologies	To identify awareness and use of existing privacy-enabling mechanisms such as policy information communicated to users and privacy-enhancing tools. This question included images of mechanism and served to validate the levels of awareness and expectations addressed.	
Privacy Concerns [Q9]: Privacy Unconcerned, Fundamentalist and Pragmatist based on Westin's Privacy Segmentation Index	To evaluate if Westin's categorisation of privacy concerns is still relevant in the current online information privacy context. The questions were based on a review of Westin's privacy segmentation index (Kumaraguru & Cranor 2005).	
Interactive Search Tasks		
Validation of Search Behaviour [Q7]: Informational Scenario: Health Terms (Blood Pressure/ Hepatitis B), Transactional Scenario: Flight Tickets	Users were asked to perform these search tasks and report their answers. This question validated search behaviour and provided indications of sensitivity of query terms (as reported in Chapter 4).	
User Evaluation of Effectiveness of Obfuscated Search Results [Q8]: Minimum vs. Default, Minimum vs. User-Agent Obfuscated, Minimum vs. Location Obfuscated	Users were asked to identify which search result was preferred to identify effectiveness of obfuscated techniques based on findings from the technical controlled experiment reported in Chapter 4.	

Factor	***Rationale***
Qualitative Questions [Q2f, Q3w, Q5]:	
Open-ended question on search behaviour: Use of search engines and/or browsers (Optional) (*"Do you have any additional comments on how you use search engines or browsers?"),*	To provide users an opportunity to offer alternate explanations of their choices on search behaviour. 114 provided usable views. This qualitative question on search behaviour was used to support the quantitative close-ended questions.
Expectations around use of personal data online (Optional) *("Do you have any other expectations about the use of personal data online?)*	To provide users an opportunity to offer alternate explanations of their choices on privacy expectations. 93 provided usable views.
Narrative on privacy expectations (Mandatory) (*"Give an example of an instance where you felt your expectations about privacy were met or not met when you searched online and your reasons why you think this is so."*).	By providing an illustrative example of how people perceive privacy online, to further understand privacy expectations and the underlying social norms and values. This question was formed based on the key premise of the contextual integrity framework (Nissenbaum 2010).

5.2 Search Behaviour

This section presents the results from the survey outlining demographic details, search behavior and quantitative expectations around data release.

5.2.1 Demographic Perspectives

Key demographic factors of participants such as age ranges, participant gender, role, area of expertise and countries or heritage associated with their identities are described in this section (See Survey Question 1 in Appendix A5.1). The 18-3 age range account for 74% of respondents. Gender distribution was 63.72% female to 36.28% male (see Figures 5.1 & 5.2) The student participants consisted of undergraduate (72.79%), postgraduate by coursework (14.97%), postgraduate research (7.7.1%), and staff participants were academic (2.04%) and professional (1.59%) (see Figure 5.3). The role the participants play and the area of knowledge expertise informs their level of awareness and understanding about privacy expectations (see Figure 5.4).

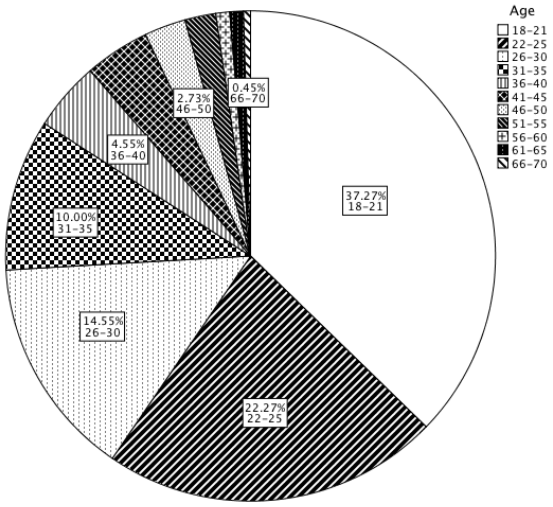


Figure 5.1 Age Range of Participants

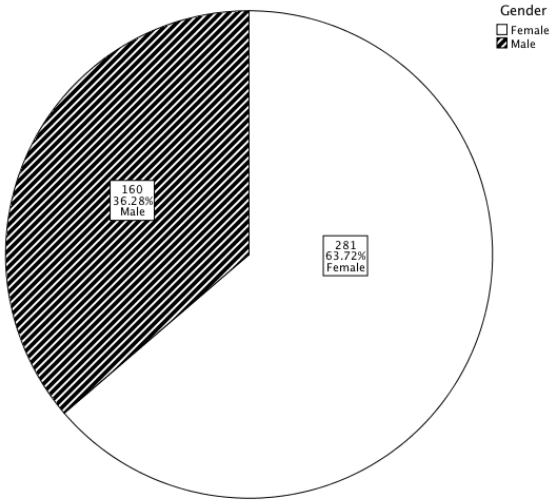


Figure 5.2 Participant Gender

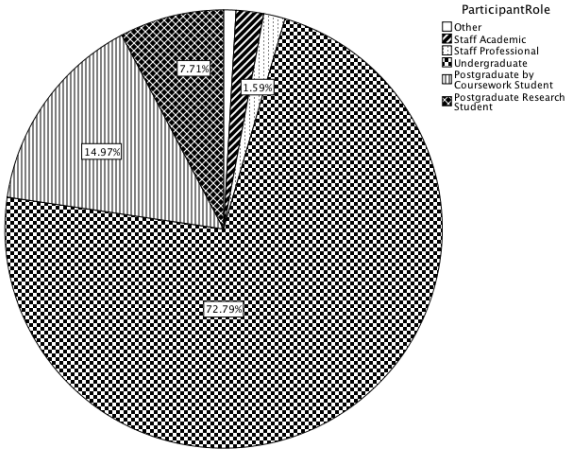


Figure 5.3 Participant Role

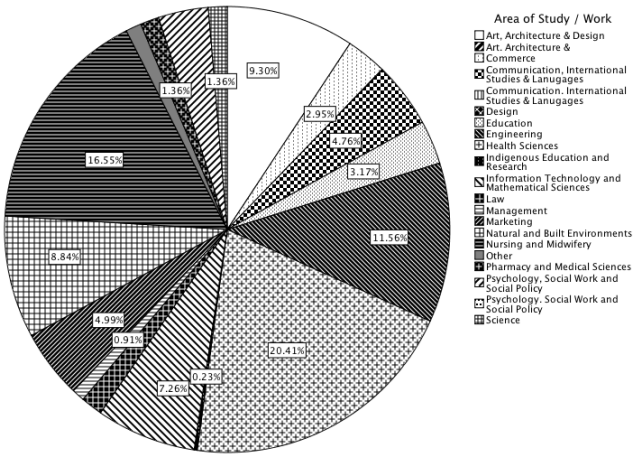


Figure 5.4 Area of Expertise of Participants

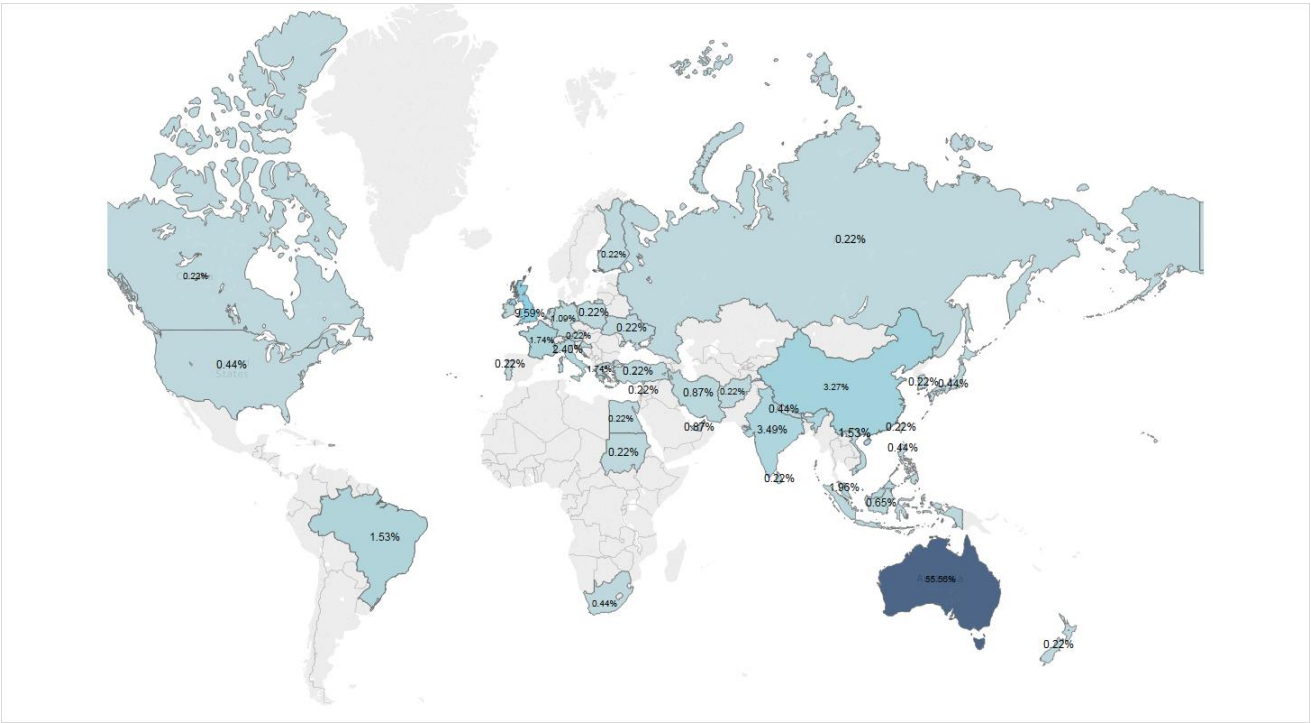


Figure 5.5 Distribution of Locations Associated with Participants (including cultures or places of origin or heritage)[3]

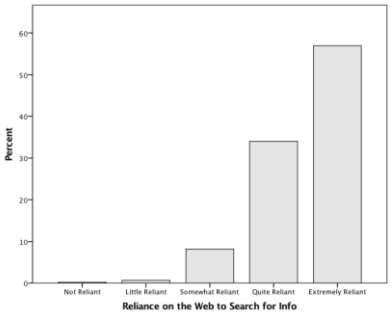
Participant awareness and understanding about privacy expectations can be influenced by their role, area of knowledge, expertise and culture. The participants were from 21 knowledge areas with the largest

[3] In instances where users reported geographic regions like Asia, Africa and Latin America, the most frequently reported location was chosen for the purposes of the visualisation, which required exact geographic coordinates.

representation 20.41% from Education, 16.55% from Nursing and Midwifery and 11.56% from Engineering. The student respondents reflected a diverse mix of cultures dominated by Australians (55.56%). Other dominant cultures were Asian (18.08%), UK (9.5%), European (9.08%) with reprsentation from South America (2.6%), North America (0.65%) and Africa (0.68%) (see Figure 5.5). This diversity closely resembles a typical multiculutral population commonly assoicated with Australian university study populations (Australian Bureau of Statistics 2011). Culture has been widely associated with informing an individual's views of the world and shaping expectations people have about privacy. Social norms about public and private spheres vary greatly across cultures (Acquisti, Brandimarte & Loewenstein 2015).

5.2.2 Search Behavior Characteristics

Survey Findings: Search behaviour characteristics corresponding to Survey Question 2 are reported in this section. (see Appendix A5.1). This sample of users is highly reliant on search engines to find information with 56.92% reporting a strong dependence on the Web as a primary measure. In terms of familiarity with Web search use, 43.08% of users have used the Web for 10 years or more (see Figures 5.6 & 5.7). **Comparisons with Related Work:** In a survey of similar demographics, 68% of respondents reported using the Internet daily, with 63% using it for search on a daily basis (Arnott & Andrejevic 2012).



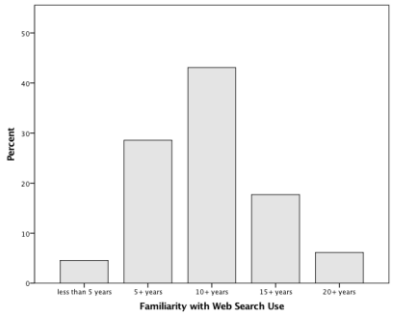


Figure 5.6 Reliance on the Web to Search for Info **Figure 5.7 Familiarity with Web Search Use**

Survey Findings: To assist information-seeking needs, Web browsers and search engines are used. Google Chrome (45.04%) was the popular browser, with Safari, Mozilla Firefox and Internet Explorer also used (18.51%, 19.10% and 16.76% respectively – see Figure 5.8). Google was the most popular search engine used (89.41 %). Yahoo and Bing were also used (less than 3.87% and 3.46%). DuckDuckGo, a privacy-enhanced search engine was used (2.04%) providing evidence of people being aware of privacy concerns associated with search engines (see Figure 5.9). Reliance on these search engines were also significant with 73.24% of participants accessing their prefered search engine multiple

times a day and more than 95.01% reporting access to this search engine across two or more devices (see Figures 5.10 and 5.11)

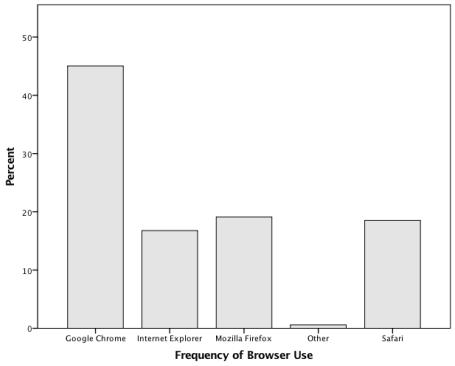


Figure 5.8 Frequency of Browser Use

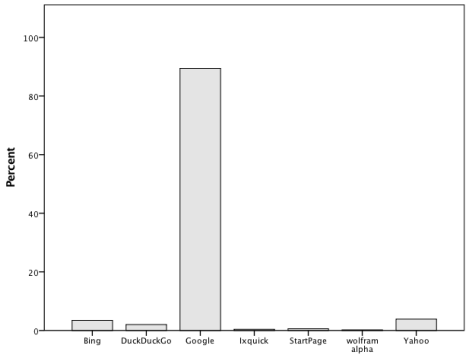


Figure 5.9 Search Engine Preference

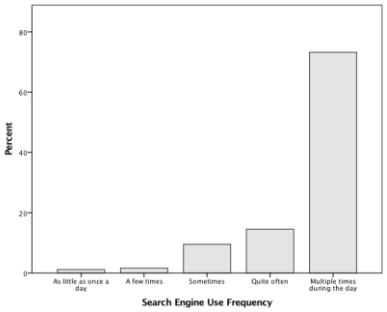


Figure 5.10 Search Engine Use Frequency

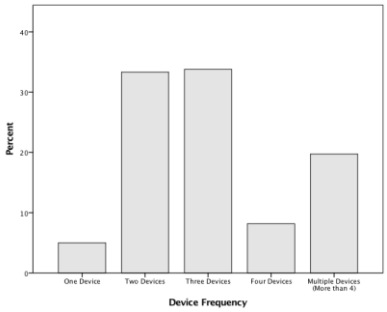


Figure 5.11 Device Frequency

5.2.3 Privacy Expectations around Data Flows: Quantitative Views

User expectations around privacy sensitivities are reported in this section. These include expectations around awareness of personal data release, privacy expectations of data flows, privacy risks, impacts and concerns, consent and disclosure mechanisms and privacy-enabling mechanisms.

5.2.3.1 Awareness of Personal Data Release and Collection

In this section, findings relating to privacy expectations around the awareness of personal data release and collection corresponding to Survey Questions 3a-c are reported (see Appendix A5.1). **Survey Findings:** Participants were asked about their expectations regarding tracking of their search behaviour. This included capture of search terms, search preferences, clickthrough data. Although 57.60% of participants were aware that tracking occurs, when asked about use of personal data shared or disclosed

in subsequent searches, 60.77% said they were not aware[4] (see Table 5.2). 92.06% of users also overwhelmingly do not expect search history to be made publicly available (refer Table 5.2).

Comparisons with Related Work: In other similar samples, this lack of understanding dips to just 8% with 31% of consumers in the UK not knowing how personal data is used, whereas 68% of marketing professionals were aware of personal data use (Chartered Institute of Marketing UK 2016). This inability to understand personal data use is universal with respondents in the US with 47% of respondents unsure about data collection practices (Rainie 2016).

Overall, these expectations indicate the users may be aware that tracking of personal data occurs, but are unsure as to the specific details on how this pervasive data collection occurs.

Comparisons with Related Work: In a survey of generic privacy community attitudes by the Office of the Australian Information Commissioner, 48% of survey respondents believed that the biggest privacy risks are caused by online services. Identity theft (23%), data security (16%), misuse of financial data (11%) were some of the key types of privacy risks mentioned with 63% of consumers concerned of being at risk and 33% experienced dissatisfaction with management of personal data (Office of the Australian Information Commissioner 2013). Key privacy risks around personal data as perceived by European consumers include fraud (50%) or misuse of online identity (40%) (European Commission 2015b). In a survey of 700 visitors to the Australian eBay website, identity theft (86%) and financial data (83%) are perceived as key privacy risks in a commercial context. These consumers also desired mandatory data breach notifications when privacy breaches occur (MacGibbon & Phair 2012).

Overall, these expectations indicate that trust underpins online interactions even when such interactions have a transactional foundation.

5.2.3.2 Expectations around Data Flows: Behavioural Advertising, Tracking, Linkability and Aggregation of Data

In this section, findings relating to privacy expectations around data flows specific to behavioural advertising, tracking, linkability and aggregation of personal data corresponding to Survey Questions 3d-i are reported (see Appendix A5.1). **Survey Findings:** Similar to awareness about personal data collection, when asked in this survey if expectations were met if personal data was made publicly

[4] A small proportion of users chose the option 'other' to reflect their view 'that they are aware and they don't like it'. It was not a standardised option, but people who had this view chose to express it in this manner.

accessible, 50.1% were unsatisfied, 24.7% somewhat unsatisfied, while a further 21.7% were indifferent (see Table 2).

Comparisons with Related Work: For example, privacy expectations were clearly not met in instances where personal data was exposed to another customer (97%) or if it was used for purposes beyond its original purpose underpinning data release (97%) in the 2013 OAIC survey (Office of the Australian Information Commissioner 2013). Exposure of data for unsolicited marketing is considered inappropriate (96%) with 56% of consumers feeling annoyed or 39% being concerned. Similarly, data being transferred to external parties offshore remained a key issue with 79% believing it to be inappropriate or 90% having concerns about it (Office of the Australian Information Commissioner 2013).

Overall, expectations that the publicity of search histories or personal data be limited are at odds with pervasiveness of tracking in reality. Data collection driven by the use of persistent cookies and other tracking mechanisms extends the visibility of each search interaction to hidden actors like behavioural advertisers and commercial information aggregators.

Table 5.2 Expectations about Awareness

Expectations about Awareness	Awareness about Tracking	Awareness about Use of Personal Data in Subsequent Searches	Expectation about User Search History being Publicly Available
Yes, I'm aware	57.60%	37.41%	6.58%
Depends on some contexts	2.49%	0.23%	0.68%
I'm aware of this, but I don't like it	1.13%	0.91%	0.23%
I don't know	0.45%	0.68%	0.45%
No, I'm not aware	38.32%	60.77%	92.06%

Survey Findings: Expectations around data release mechanisms revealed lower levels of awareness and understanding. For example, when asked about if expectations were met when results users clicked on a search page are recorded (i.e. capture of clickthrough data) 39.2% expressed indifferent views, whilst 23.8% and 24.9% said expectations were not met or somewhat not met respectively (see Table 5.3). Similarly, when a scenario question described linkability of search queries (i.e. when ads or search results are presented to users in subsequent linked search queries), users conveyed similar views with 36.5% being indifferent and 20.9% satisfied and 18.1% somewhat unsatisfied (see Table 5.3). These expectations confirm that users largely lack understanding about the extent to which personally-identifiable data is explicitly or implicitly released (through their search behavior).

Survey Findings: The value of personalised search results seemed to be largely ambivalent. Users were asked if expectations were met when personally identifiable information was used to tailor and present relevant search results that better aligned with their needs. 39.7% of respondents said that it does not really affect them, while more than half of the respondents were not satisfied (36.3% were not satisfied and 15% were somewhat unsatisfied respectively see Table 5.3).

Comparisons with Related Work: In related surveys, when asked about willingness to trade off information for improved services or discounts only 34% and 28% of participants respectively expressed satisfaction in the 2013 OAIC survey (Office of the Australian Information Commissioner 2013).

Overall, both these perspectives (this survey and comparisons with related work) show that use of data to personalise results or the extent to which this data capture may influence results, may be largely invisible to users.

Survey Findings: A majority of the expectations around the use of tracking mechanisms were expressions of dissatisfaction. Users were asked if having their search history, cookies and other technologies to track online behaviour was acceptable. 39.7% and 28.8% of respondents said their expectations were not met, and somewhat not met respectively while a further 21.1% were undecided (see Table 5.3). These responses contrast with expectations around recording of clickthrough data and the inherent commercial value regarding linkability of data. In prior scenarios (see Table 5.3) the level of expectations of specific tracking activities was gauged, whilst in this scenario (see Table 5.3) a direct generalised statement about tracking in commonly-used language such as search history and cookies is presented. Both these scenarios test the same data collection and tracking activities by looking at different effects (i.e. linkability and capture of clickthrough data vs. generalised tracking of search histories through cookies). Survey respondents may be more familiar with the latter scenario (due to more mainstream privacy efforts such as the EU Cookie Law) than the former, regardless of both scenarios occurring on a regular basis. From these responses, it can be inferred that the ability to comprehend technical terminology and its effects play a part in influencing expectations of users.

Comparisons with Related Work: In related studies, tracking online activities or potential commercialisation of their data for behavioural advertising purposes were considered to be instances in which privacy expectations were not met (78% and 69% respectively). 59% of users were aware of website tracking with 48% of users expecting that mobile apps tracked user behaviour as well. Expectations around targeted advertising using online behaviour were uncomfortable to 60% of people,

while 78% were uncomfortable with storage of this information in databases (Office of the Australian Information Commissioner 2013). Likewise, in another Australian study 36% of users approve of targeted advertising with 56% disapproving. Of those that approved of targeted advertising, 73% approved if this information was collected during the primary interactions with that website, while only 38% approve of aggregating data across websites or combining it with offline data (29%). Similarly, for targeted news, 31% approve with 64% disapproving. Of those that approve of targeted news, 79% approve from primary sources, 51% from secondary aggregated sources and 32% from offline sources (Arnott & Andrejevic 2012).

Overall, privacy expectations are not met when interactions are commercialised for behavioural advertising purposes, although these expectations are influenced by the level of understanding of terminology and the complexity underpinning the interactions.

Survey Findings: Views around aggregating data to create profiles largely indicated dissatisfaction with 50.1% of participants reporting dissatisfaction whilst 23.4% were somewhat unsatisfied with a further 18.6% ambivalent (see Table 5.3). When asked whether expectations were met if undisclosed parties had access to browsing behaviour and search preferences of users, 70.3% of respondents were not satisfied with 18.1% somewhat unsatisfied and about 7.9% reporting indifferent views (see Table 5.3).

Comparisons with Related Work: In a study of Australian consumers, 48% thought permission was required to aggregate data and link users across different sessions while 49% did not think so (Arnott & Andrejevic 2012). In a 2015 Pew Center survey, 51% of users were unwilling to accept disclosing personal details in exchange for advertisements on social media whilst only 33% found it acceptable (Rainie & Duggan 2015).

Overall, aggregating data across different interactions to unspecified parties presents a privacy concern. This indicates that users expect data released to be used for specific purposes, among authorised actors constrained to those contexts.

5.2.3.3 Expectations around Privacy Risks and Impacts

In this section, findings relating to privacy expectations around privacy risks and impacts corresponding to Survey Questions j-n are reported (see Appendix A5.1). **Survey Findings:** Similar views were expressed around inference and identification based on personally-identifiable information released when searching with 72.1% and 15.6% of respondents articulating strong or weak dissatisfaction

respectively (see Table 5.3). These views underscore the importance of combating browser-fingerprinting techniques that can add to implicit release of identifiable data. Particularly strong views were expressed around the sale of personally-identifiable information to undisclosed parties with 78.2% and 12.2% of participants not satisfied or somewhat not satisfied respectively (see Table 5.3). Contrary to these views about selling personal data, data brokers capture interactions, aggregate data and harness value from it and are seemingly invisible to a typical user.

Table 5. 3 Online Privacy Expectations: Interactions

Expectations	Expectations not satisfied	Expectations somewhat unsatisfied	Doesn't really affect me	Expectations somewhat satisfied	Expectations satisfied
Personal Data being Publicly Accessible	50.1%	24.7%	21.7%	3.2%	0.9%
Capture of Clickthrough Data	23.8%	24.9%	39.2%	8.2%	3.9%
Linkability of Search Queries	20.9%	18.1%	36.5%	16.6%	7.9%
Value of Personalised Search Results	36.3%	15%	39.7%	2.7%	6.3%
Use of Tracking Mechanisms	39.7%	28.8%	21.1%	6.8%	3.6%
Data Aggregation	50.1%	23.4%	18.6%	5.7%	2.3%
Access to Personally-Identifiable Information (PII)	70.3%	18.1%	7.9%	2.3%	1.4 %
Inference and Identification around PII	72.1%	15.6%	7.7%	2.7%	1.8%
Selling PII	78.2%	12.2%	6.1%	2.3%	1.1%
Access to PII by Family or Friends	56%	19.5%	19.5%	3.2%	1.8%
Access to PII by Co-workers or Employers	70.3%	14.3%	12.2%	1.8%	1.4%

Survey Findings: Access to personally-identifiable information remains an important issue with many respondents considering it to be personal. 56% and 19.5% expressed dissatisfaction whilst 19.5% were ambivalent when asked if personally-identifiable information could be accessed by friends or family

(see Table 5.3). Similarly, if access to PII was extended to co-workers or employers 70.3% reported being dissatisfied with 14.3% somewhat unsatisfied whilst a further 12.2% were ambivalent. A relatively low number of people were satisfied (1.8% and 1.4% respectively) with granting access (see Table 5.3). **Comparisons with Related Work:** Contrary to this, US users are particularly concerned when personal data is accessible to hackers or criminals (33%) or advertisers (28%) but have higher levels of acceptance among friends (19%), family (14%), employers or coworkers (11%) with people not actively trying to avoid government surveillance or law enforcement (5% and 4% respectively) (Rainie 2016). Overall, these findings indicate the contextual nature of privacy, where it matters which recipients have access to this data, particularly those that users may not have intended to have visibility when that data was originally released.

5.2.3.4 Awareness of Mechanisms that Enable Privacy

Survey Findings: This section explores awareness of mechanisms that enable privacy such as PETs, policy mechanisms and preferences around these mechanisms. These findings reported correspond to Survey Questions 3o-v and 6 (consent and disclosure mechanisms), and Survey Question 4 (privacy-enabling mechanisms) (see Appendix 5.1). Some users also reported on privacy habits such as *"I delete cookies often"* (P#104); *"I'm not comfortable with Google Chrome. Google already has my search history, I don't want them having my entire browsing history. It seems to be the fashionable browser, but I don't care"* (P#132); *"Typical use privacy option for transactions."* (P#322). Others reported on the use of privacy-enhancing technologies such as ad-blockers and plugins *"I install plug-ins to stop ads and customise the Browsers"* (P#10) or isolating potential incidents of risk *"I also use IE5.5 in a Windows Me virtual machine to visit links from spam emails and see how badly their terrible sites break on a terrible browser. Nothing has escaped the VM yet."* (P#51).

A highly-informed user reports on the weaknesses of using these existing privacy-enhancing tools.

> Usually use AdBlocker or script blockers to block annoying ads. Although, this is somewhat detrimental to privacy because the fact that you use a script blocker can be identified. The unique setup of your browser, including the plugins you use, produce a so-called fingerprint that can identify you publically. The premise is that every single person has their browser setup differently, providing useful data to Google Analytics. Even if a user refuses to use Google, websites they visit may use Google services, making the person identifiable through the Google APIs that are downloaded by scripts ran on the client's computer. (P#12).

Users also report trying to minimise data release with extensive effort undertaken to avoid dissemination of data and its linkability across different services. *"I used additional browsers to test learning objects which I am not logged into on my main browser. I also use alternative browsers to be logged into*

different Google IDs" (P#419); "I sandbox different tasks in browser sessions, e.g. webmail is always in its own sandbox and doesn't share cookies with a general purpose browser. I use extensions like NoScript, Ghostery, AdBlock Origin. I delete cookies and LSOs frequently. I use a VPN. I know this probably only strengthens browser fingerprinting. Damned if you, damned if you don't, right?" (P#361). This last sentiment underscores the difficulty users currently face as the more protections put in place themselves increase the identifiability of the browser. There was strong agreement (83.9%) for preferences to be anonymous or pseudonymous among participants. 12.93% were of not of the same view. Some people said that it depends on some contexts (2.27%), with some expressing that they were not concerned (0.68%) or not sure (0.23%) (see Table 5.4).

Table 5.4 Online Privacy Expectations: Preferences

Expectations	Yes	No	Other
Preferences about being Anonymous or Pseudonymous	83.9 %	12.93%	Depends on the context of use (2.27%) Not concerned (0.68%) Not sure (0.23%).
Preferences about Transparency regarding Collection, Use and Storage of Personal Data	85.71 %	11.79%	My personal data should not be collected (0.45%) Not concerned (1.36%) Not sure (0.68%).
Preferences regarding Access and Control to Sensitive Information	85.03 %	13.38%	My personal data should not be collected (0.23%) Not concerned (0.45%) Not sure (0.23%).
Preferences regarding Informed Consent about Collection, Use and Storage of Data	87.53 %	10.66%	Depends on the context of use (0.68%), My personal data should not be collected (0.45%) Not concerned (0.45%) Probably wouldn't understand (0.23%)
Preferences around using Anonymised IDs	68.03 %	26.53%	Depends on the context of use (3.4%) It shouldn't be identified at all (0.91%) My personal data should not be collected (0.23%) Not concerned (0.23%) Not sure (0.68%).

When asked if transparency regarding collection, use and storage of personal data by search engines was important, 85.71% said yes while 11.79% did not prefer transparency (see Table 5.4). Some people were of the view that personal data should not be collected (0.45%), not concerned (1.36%) or not sure (0.68%). The ability to have transparency and anonymity is an expectation of importance to users. 85.03% of participants preferred some degree access and control to sensitive information held by search engines whilst 13.38% were not of the same view (see Table 5.4). Some other views include that personal data should not be collected (0.23%), with some not concerned (0.45%) and some not sure

(0.23%). Similarly, 87.53% said they preferred to be informed regarding consent about collection, use and storage of data whilst 10.66% said they didn't prefer (See Table 5.4).

A further 0.68% of participants said it depends on some contexts. Other views expressed include 'my personal data should not be collected' (0.45%), not concerned (0.45%) and 'probably wouldn't understand' (0.23%). 68.03% of participants preferred the use of anonymous IDs as opposed to being identified by their name (see Table 5.4). Comparatively 26.53% didn't prefer this option, while 3.4% said it depends on the context of use. Other views include that 'they shouldn't be identified at all' (0.91%), 'my personal data should not be collected' (0.23%) and not concerned (0.23%) and not sure (0.68%).

Comparisons with Related Work: In a study of Australian consumers, 75% of users believed telecom providers collected location data, while 65% believed websites will track personal data. However, 50% believed that websites need to obtain permission to track personal data while 49% believed that they did not need permission (Arnott & Andrejevic 2012). Likewise 69% of European respondents expect websites to ask explicit permission before collection of data (European Commission 2015b). Overall, the consensus among users globally is to be asked for their agreement when releasing data. These findings indicate that existing affordances (such as informed consent via terms and conditions) do not effectively address privacy expectations around consent.

Survey Findings: Terms and conditions are the primary mechanism in which user agreement is sought in interactions. In line with prior work highlighting the ineffectiveness of the notice and consent mechanism, 36.7% said they do not read terms and conditions at all, while 24.5% said a few times and 28.8% on some occasions[5]. Relatively few said quite often (6.3%) and all the time (3.6%). Length and complexity of the terms and conditions expressed in its current form affect a user's ability to fully understand what it is they are agreeing to. 63.3% expressed that length of terms and conditions affected the effectiveness of notice and consent, whilst 54.4% were of the view that complexity impacted effectiveness. (See Table 5.5).

Articulating a clear privacy policy is a first step, but should not be the primary method in which consumers are reassured of data protection as only 30% of UK consumers feel reassured through this

[5] Note that these expectations were ranked on a Likert scale of 1 to 5 where 1 denoted *not at all*, 2 denoted *a few times*, 3 denoted *sometimes*, 4 denoted *quite often* and 5 denoted *all the time*.

mechanism. In contrast, 59% feel assured through the verification of encrypted communications (Chartered Institute of Marketing UK 2016). Similarly, 18% of respondents in an Australian study read privacy policy statements with 36% reporting they never read them (Arnott & Andrejevic 2012).

Putting into place legal mechanisms for data protection is a strongly desired expectation of consumers. Scenarios include when deleting data (96%), notification of collection (95%), Do Not Track (DNT) (92%) and the right to view collected data (89%) (Arnott & Andrejevic 2012). Other mechanisms include protections for children under the age of 13 (98% on personal data collection, and 91% on a targeted ad ban). This is particularly difficult to enforce as it means identifying that the user is in fact a child, unless generic usable settings are built into browsers (Arnott & Andrejevic 2012).

Table 5.5 Expectations: Consent and Disclosure

Expectation: Consent and Disclosure	Awareness about Notice and Consent	Effectiveness of Notice and Consent: Length	Effectiveness of Notice and Consent: Complexity
Not at all / No	36.7%	10.9%	12.2%
A few times / Probably Not	24.5%	2.9%	3.9%
Sometimes	28.8%.	10.7%	16.1%
Quite often / Maybe	6.3%	12.2%	13.4%
All the time / Yes	3.6%	63.3%	54.4%

Trust in relation to how personal data is managed is important to Australian users with 69% of respondents trusting public organisations more than private with exceptions accounted for financial (74%) or health (90%) institutions. Transparency is key to building this trust with users expecting to know how personal data is collected and stored (96% for public, 95% for private), or if it is lost (96% for both) (Office of the Australian Information Commissioner 2013). Similarly, low trust is seen among Australian consumers with 4% indicating that all businesses provide sufficient information regarding personal data use while 19% indicating none do. 59% of consumers believe too much personal data is collected (Arnott & Andrejevic 2012).

In a survey of UK marketing professionals like those involved in online behavioural advertising, 70% of marketers reported not placing trust as a key important issue over personal data use, whilst 57% of UK consumers reported not trusting brands with 70% not understanding the benefits of sharing data. A further 51% of consumers reported negative effects of unsolicited marketing while 50% of marketers

were not aware that opting out of direct marketing was a consumer right. 81% of UK marketing professionals reported sharing personal data without explicit permission from consumers aggravating trust issues (Chartered Institute of Marketing UK 2016). The Chartered Institute of Marketing recommends redefining the role of marketer to be more authentic about data collection practices to build trust among consumers and aligning strategic business practices around data analytics (Chartered Institute of Marketing UK 2016).

Similarly in the US, 91% of users say they do not have control over the release of their personal data with few consumers articulating they are adequately equipped to deal with dataveillance (Rainie 2016). In Europe, 67% of respondents are concerned about this ineffective level of control over data release (European Commission 2015b).

Overall, these tensions indicate the complex nature of personal data use and collection and why consumers struggle to act according to expressed concerns. The lack of knowledge, training and time to implement existing privacy-enhancing measures as practices are significant barriers people face in exercising their privacy.

In this survey study, users were also asked to interactively evaluate if they have seen or used privacy-enabling mechanisms such as policy mechanisms or privacy-enhancing technologies. Awareness of static privacy policies (15.1%) and its various information formats (19.2%, 10.2%) are quite low compared to more dynamic mechanisms like the Google Privacy CheckUp (42.5% - see Tables 5.6 and 5.7). This indicates that dynamic mechanisms that explain concepts simply to users through a short sequence of steps have better reach and effectiveness than simply putting up a static policy, which primarily exists for legal purposes. Designing mechanisms like the Privacy CheckUp highlights the role affordances play in communicating key privacy information to users.

Table 5.6 Awareness and Use of Privacy Info

Awareness and Use of Privacy Info	%
Google Privacy CheckUp Notice	42.5
Info from the Google Privacy Information Center (2 sections)	19.2 10.2
Google Privacy Policy	15.1
Not seen or used any of these options	12.7
Other services	0.1

Table 5.7 Privacy-enabling Mechanisms

Privacy-enabling Mechanisms	%
Anonymising VPNs	0.7
Anonymous browsers (Tor)	3.2
Anonymous browsers (EPIC)	1.6
Do Not Track	7.7
Not Seen or Used Any of These Options	28.7
Other	0.7
Privacy-enhanced Search Engines	4.3
Tools that block ads	19.6
Tools that let you control your identity	5.9
Tools that prevent 3rd Party tracking	6.6.
Tools that prevent tracking	5.2
Your browser privacy settings	15.9

Table 5.8 Characteristics of Consent and Disclosure

Characteristics of Consent and Disclosure	%
I should be notified when my information is disclosed, sold or shared with third-parties	18.6
I should be notified when privacy policies or practices change	15.4
Option to decide how my Personally-Identifiable Information is used	18.2
Other	0.4
Search engines should disclose how my data is to be used	15.9
Search engines should disclose security safeguards used to protect my PII	14.9
Search engines should inform me before using my data	16.6

Adoption of privacy-enhancing technologies among users is low with 28.7% of respondents in this survey indicating that they have not used any of the mainstream options (see Table 5.7). This trend is similar to prior related work (Wills & Zeljkovic 2011). 15.9% had seen or used privacy settings embedded in the browser, while 19.6% use AdBlock tools. Use of privacy-enhanced search engines, anonymising tools or tools that prevent tracking were low.

Comparisons with Related Work: A similar Australian survey awareness about privacy-enabling mechanisms indicated that most users have rudimentary knowledge of key concepts with 73% indicating they wished they had access to clearer information on how personal data is used by websites. 54% reported understanding what a cookie is with 85% of those respondents having cleared cookies. Only

17% had used anonymous browsing through incognito mode (Arnott & Andrejevic 2012). The measures that people take to protect release of personal data include refusing to release information in specific scenarios (90%), verifying security of websites prior to release (78%), deleting browser search history (72%), choosing not to interact on apps that may leak information (62%) , falsifying personal data such as name (30%) or details (32%) (Office of the Australian Information Commissioner 2013). Similarly, 78% of respondents reported not providing personal data with 69% choosing to not engage with websites that could potentially leak data (Arnott & Andrejevic 2012). In the US, awareness about privacy-enhancing mechanisms is more prevalent with 86% of users reporting use (Rainie 2016).

Overall, these findings highlight the need to develop meaningful mechanisms that enable appropriate data flows. Users characterise expectations of consent and disclosure by an intention to be informed when any change to the previously agreed-upon data flow happens. Even though privacy policies have been widely debated upon as ineffective in scholarly research, users do not seem to find the existing methods used by organisations effective. Perhaps the underlying social norm of asking someone's permission before using data about them, instead of a wide one-stop license to use information without checks and balances still persists in the online space. Meaningful consent respected contextually is a key user requirement that needs to be addressed.

5.2.4 Understanding the Relevance of Search Results to Users

This section reports on Survey Questions 7 and 8 which required users to perform a comparative evaluation of data released including minimised data release (query terms and IP address) and obfuscated data release (where search results were produced after obfuscating the user-agent and location). Incorporating findings from the technical controlled experiment in Phase 1 (see Chapter 4, RQ1), respondents were asked to evaluate search result comparisons. Search results from Google were used as a basis for the comparison as it is the leading search engine (as evident in the survey findings), using popularly-sourced query terms (refer Chapter 4). The evaluations consisted of three different experimental conditions: minimised, location obfuscated and user-agent obfuscated query release as these three factors had a significant impact on the variance of personalised search results in the technical controlled experiment (see Table 5.9).

Survey Findings: Users have a marginal preference for the minimalist query (i.e. the minimum amount (query term and location) of technically possible data released - 57.6%) over the default query (i.e. with no modification - 41.4%) (see Table 5.9). Similarly, users had a slight preference for the minimalist query (54.5%) compared to a query with an obfuscated user-agent parameter (i.e. describing the type of

browser and operating system - 41.6%) (see Table 5.9). In both these prior cases, users expressed that there was largely no distinct difference separating the two choices. Similar ambivalent views around the value of personalisation were reported earlier in the survey (see Table 5.3). The preferences were more distinct between location-obfuscated results (i.e. spoofing the location to appear from the US vs. the default Australian results) with users preferring Australian results (61.9%) over the location obfuscated result (34.6%) (see Table 5.9). This indicates that location has some value in terms of relevance of search results particularly for transactional queries (see Table 5.10).

Table 5.9 User Evaluation of Obfuscated Personalised Search Results

Minimised **Query Term: Informational** **(Harlem Shake)**	**Location Obfuscated** **Query Term: Transactional** **(Dominos)**	**User-Agent Obfuscated** **Query Term: Informational** **(Cards Against Humanity)**
Search Result A (57.6%)	Search Result E (61.9%)	Search Result C (54.5%)
Search Result B (41.4%)	Search Result F (34.6%)	Search Result D (41.6%)
Same (1.1%)	Same (1.6%)	Same (2.5%)
Neither (0.1%)	Query Term Reformulation (0.2%)	Neither (0.7%)
	Not Sure (0.7%)	Not Sure (0.7%)
	Neither (0.9%)	
	*Other options also include	

Table 5.10 Reasoning of Search Results Selection

Reasoning – Search Results Selection	**%**
Both search results do not meet my expectations about the information contained in the search results	3.4
I can find what I would look for in the selected search result	26.9
I don't mind if identifiable information such as location is used to generate the result as long as it meets my needs	20.5
I want to be able to better control how identifying information is reflected in search results	13.2
I would like to prevent any type of data leakage	12.5
Other	0.2
The selected search result somewhat matches the info I would look for	23.2

Comparisons with Related Work: In prior work, measuring the variance of personalised search results and evaluations of the privacy-personalisation trade-off and privacy-enhancing tools have been extensively conducted among users (Awad & Krishnan 2006; Hannak et al. 2013; Malandrino & Scarano 2013; Toubiana, Vincent, Subramanian & Nissenbaum 2011). To this end, these results in this chapter present a novel contribution on the user evaluation of obfuscated search results. Providing users

with affordances to manage data flows appropriately will enable users to take proactive measures for data protection if they choose to do so.

From related work, when client-side personalisation is coupled with privacy-sensitive behaviour such as periodic data backups and remote access to secure devices improvements were seen with perceive levels of privacy protection (Kobsa, Knijnenburg & Livshits 2014). Users tend to adopt privacy-enabling measures in evaluation of prototypes that embed these affordances (Preibusch 2013a). The strength of an obfuscation technique is in its purpose of use (Brunton & Nissenbaum 2015). Therefore, tailoring privacy-enabling measures such as obfuscation techniques to address expectations affords users the ability to be manage data release when searching.

Survey Findings: In the second set of interactive search tasks, users were to search on their devices for particular search term and report their findings. This was used to validate previously-reported search behaviour. Most users reported using the search engine page (63.19%), with 97.72% using Google (see Figures 5.7 and 5.8). 53.99% accessed Google using Google Chrome, with a further 18.91% via Mozilla Firefox, 15.26% via Safari, and 11.16% via Internet Explorer (see Figure 5.12-5.15). Compared to the previously reported findings earlier in the survey, actual use of search engines and browsers mirrors reported intent. 60.77% were logged into their accounts indicating that behaviour among most users is to use default mechanisms which enable collection of data linked to their identity (through account logins – see Figure 5.10). Effective account-based privacy-enabling mechanisms would be of value, if verification of functions (such as deleting search history) could be made visible to users.

To understand sensitivity of search terms informational health terms were used in the first scenario. A control was included where for 51 users the search scenario was on ‘Blood Pressure’ while for the remaining 390 the scenario was based on ‘Hepatitis B’. Sensitivity preferences were marginal between the terms with 82.35% considering ‘Blood Pressure’ as sensitive while 71.72% considered ‘Hepatitis B’ as sensitive (see Figures 5.16-5.19).

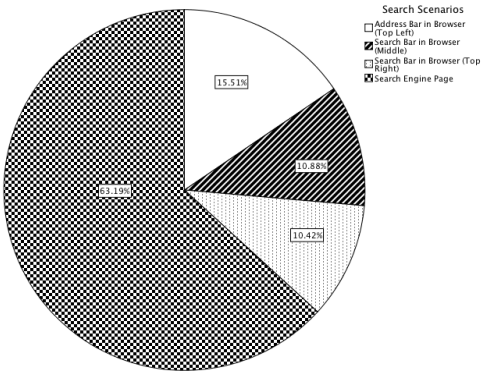


Figure 5.12 Search Scenarios

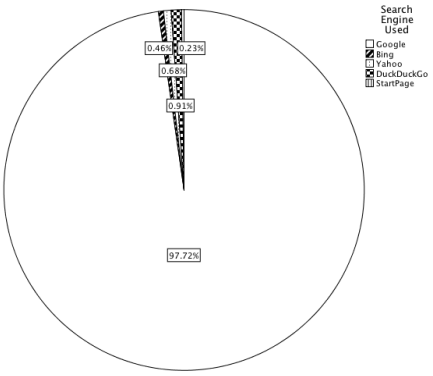


Figure 5.13 Search Engine Used

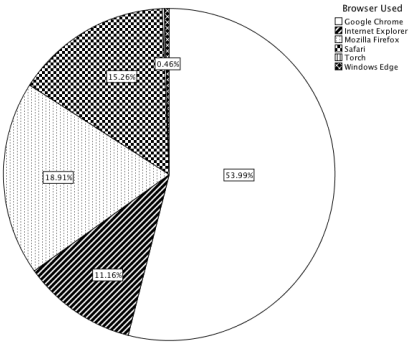


Figure 5.14 Browser Used

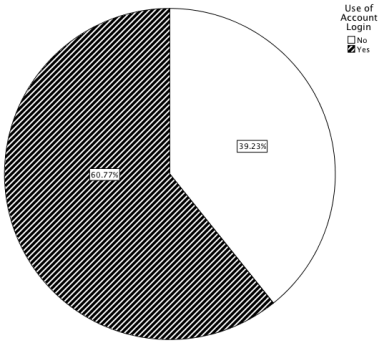


Figure 5.15 Use of Account Login

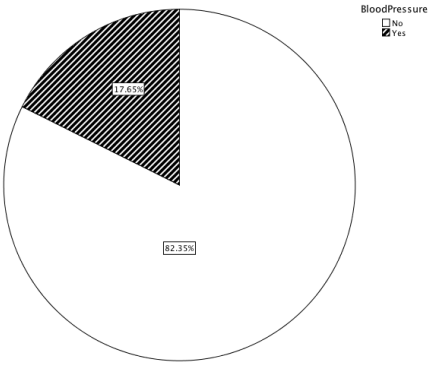


Figure 5.16 Sensitivity of Query Term: Blood Pressure

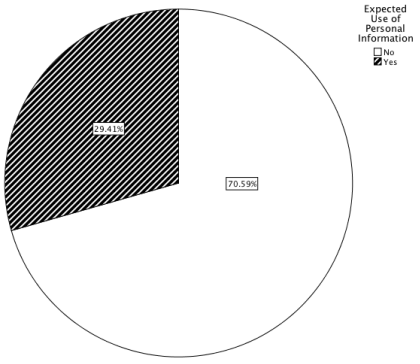


Figure 5.17 Sensitivity - Blood Pressure: Expected Use of Personal Information

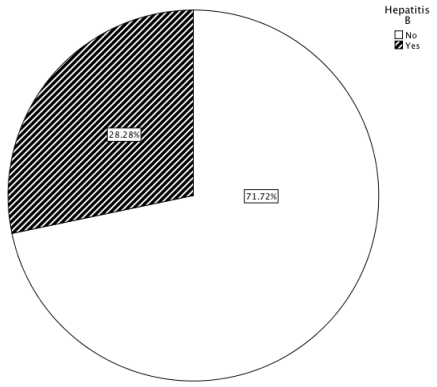


Figure 5.18 Sensitivity of Query Term: Hepatitis B

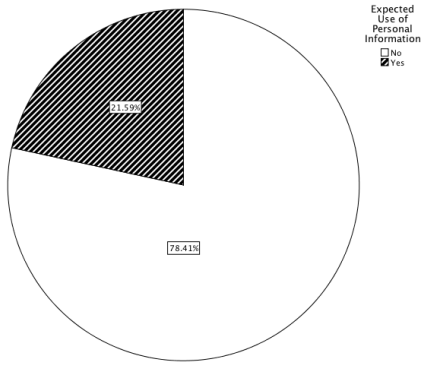


Figure 5.19 Sensitivity: Hepatitis B: Expected Use of Personal Information

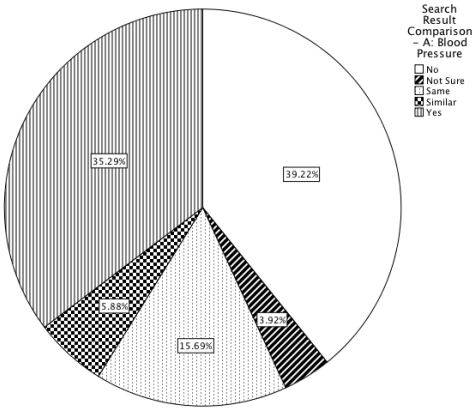


Figure 5.20 Search Result Comparison A: Blood Pressure

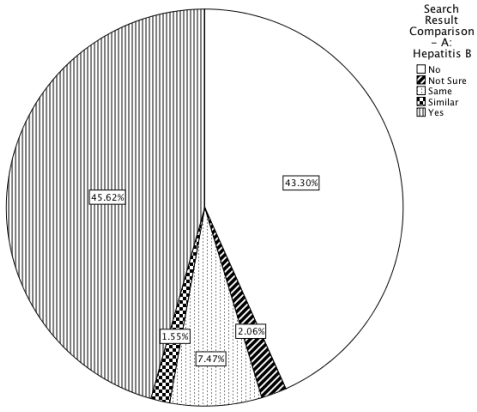


Figure 5.21 Search Result Comparison A: Hepatitis B

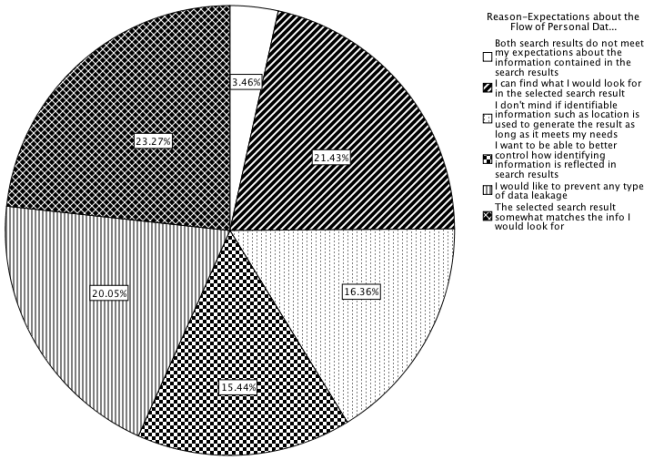


Figure 5.22 Reasoning: Comparison A

However, when asked if they expect personal data to be used to generate personalised search results, in both instances the response was overwhelmingly no (70.59% not expecting personal data to be used for 'Blood Pressure' while 78.41% for 'Hepatitis B' – see Figures 5.17 and 5.19). When asked to compare if they preferred their own search result or the provided search result, the differences between preferences were marginal (39.22% indicating no for 'Blood Pressure' while 45.33% indicating yes for 'Hepatitis B' – see Figures 5.20 and 5.21). Most users reported that the search results were similar with little to choose between either choice (see Figure 5.22). This indicates that although users performed a search task taking into account their own data release (personally-identifiable data and search behaviour data), when it comes to users selecting between choices, neither option holds a distinct majority

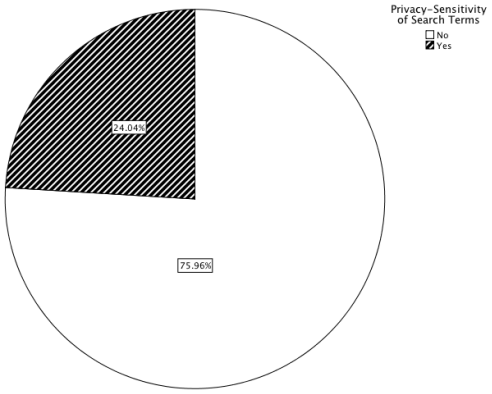


Figure 5.23 Sensitivity of Query Terms: Flight Tickets

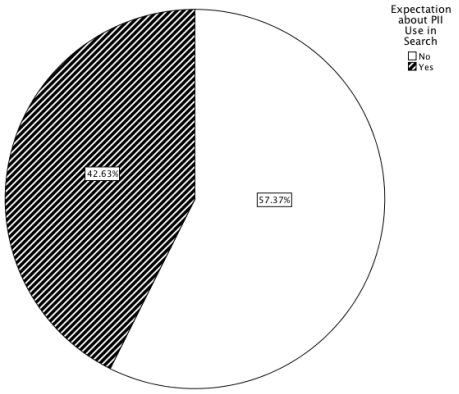


Figure 5.24 Sensitivity: Flight Tickets: Expected Use of Personal Information

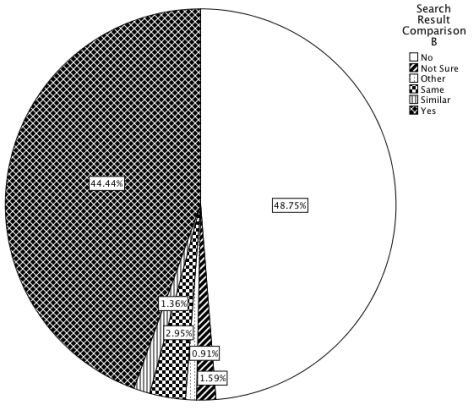


Figure 5.25 Search Result Comparison B

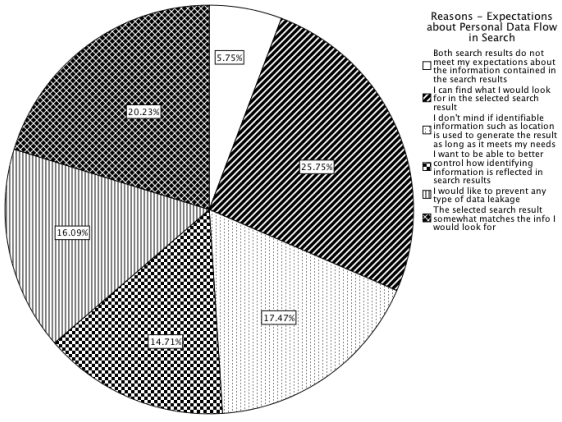


Figure 5.26 Reasoning: Comparison B

In the second scenario, users were provided with a transactional search scenario, where they were asked to search for flight tickets between two major Australian cities (Adelaide and Melbourne). Sensitivity of the transactional search term was similar in relation to the informational search scenario (75.96%

considering it sensitive – see Figure 5.23-5.24). However, expectations about personal data use increased to 42.63% compared to the information search scenario although 57.37% still preferred personal data to not be used (see Figure 5.24).

Similar preferences were seen with the transactional search scenario as with the informational search one. Differences between preferring their own search result (44.44%) vs. the search result provided (48.75%) were marginal (see Figures 5.25-5.26). This finding further supports the notion that personally-identifiable data (either account based personal details or behavioural data released through the interaction) impacts users' choices to a lesser degree than current assumptions made by information providers like search engines.

5.2.5 Measuring Privacy Concern: Comparisons with Westin's Privacy Segmentation Index

In this section, findings relating to a measurement of privacy concern (based on Westin's Privacy Segmentation Index) corresponding to Survey Question 9 are reported (see Appendix A5.1). Often relevance is framed as a trade-off between utility and privacy concerns. For example this trade-off is seen in studies that aim to increase levels of personalisation using significantly low amounts of personal data (Krause & Horvitz 2010). A commonly-used metric to measure privacy concerns is Westin's privacy segmentation index which he developed and iteratively refined using surveys since 1978 to 2004**.** The index asks respondents to rate on a scale of 1-4 (from strongly disagree to strongly agree), their responses to three questions:

"1. Consumers have lost all control over how personal information is collected and used by companies.
2. Most businesses handle the personal information they collect about consumers in a proper and confidential way.
3. Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today." as cited in (Kumaraguru & Cranor 2005, p.19)

Survey Findings: In this survey study, privacy concerns for this sample was tested using the privacy segmentation index method outlined in Westin's 2001 study A (as calculated and cited in (Kumaraguru & Cranor 2005)). According to Westin's privacy segmentation index, 65.99% of survey respondents identified as privacy fundamentalists, 20.86% as privacy pragmatists and 5.9% as privacy unconcerned (see Table 5.11). Privacy fundamentalists are deeply concerned about their privacy; do not particularly like excessive data collection and typically want strong data protection policies in place. Privacy unconcerned individuals are on the opposite end of this spectrum and are of the view that the benefits

offered by online information service providers are more than the cost of releasing personal data. Typically, they are also unconcerned about data protection regimes. Privacy pragmatists as the name suggests are practice-oriented evaluating potential benefits and losses before deciding whether to release their data. Generally, they consider a number of factors such as the trust they have with online information service providers and the level of protection afforded to them through privacy-enabling mechanisms. These findings as indicated by search behaviour highlights an interesting trend – a rise in the number of people who identify as privacy fundamentalists – those that are deeply concerned about privacy.

Comparisons with Related Work: This outlook is in line with reported trends showing larger numbers of privacy fundamentalists (Tsai et al. 2006; Woodruff et al. 2014). However, this trend somewhat contrasts Westin's own research in surveying people's privacy, which showed moderate levels in each category with the majority of people identified as privacy pragmatists.

Perhaps privacy issues being reported in the public media may engineer a shift in privacy attitudes and expectations in the collective consciousness of people. A recent analysis of privacy behaviour after the Snowden revelations demonstrated little impact in user behaviour change with low uptake in privacy-enhancing technologies beyond initial interest, although significant interest persisted in academic and journalistic domains (Preibusch 2015). These findings indicate that users perceive privacy issues involving governments and corporations differently (Preibusch 2015). However, one possible explanation is that this trend indicating a gap in expectations versus actual behaviour may be impacted by the fact that existing privacy-enhancing technologies are not accessible to average users, with factors such as technical literacy, awareness and understanding of behavioural implications constraining uptake as indicated by this survey's findings. Another plausible interpretation is that change in privacy behaviour as measured by uptake in privacy-enhancing technologies is low because there is little meaningful change in the capabilities offered by service providers, affording users more effective privacy protection given the reliance on embedded defaults. Participants recruited from a university environment may present a potential bias in having a stronger understanding of the complex offerings and privacy-enabling mechanisms than the general population. However, despite this potential bias from an educated sample of participants, respondents struggled to grasp the complex service offerings and terms and conditions. These perspectives need to be further explored in future research.

Table 5.11 Distribution of Privacy Concerns using Westin's Privacy Segmentation Index as a Baseline (as cited in (Kumaraguru & Cranor 2005)) *

Category	This study (2017)	Woodruff et al. 2014	Tsai et al. 2006	Westin 2003 study*	Westin 1996 study*	Westin's original study 1990*
Sample Size and Origin	n = 441 Australia	n = 844 US	n=276 US	n= 101 US	n=1005 US	n=1255 US
Privacy Fundamentalist	65.99%	49.1 %	70.3%	26%	25%	25%
Privacy Pragmatist	20.86%	39.8%	15.6%	64%	59%	57%
Privacy Unconcerned	5.90%	11.2%	14.13%	10%	16%	18%

However, as with most quantitative metrics indicating privacy sensitivities, there are weaknesses with this method. Quantitative evaluations are limited in their ability to elicit a richer and deeper understanding of the reasoning behind these privacy sensitivities. One of the weaknesses of the privacy segmentation index is that drawing correlations between classifications of privacy types and behavioural intent or actions is difficult although connects may exist between attitude vs. behaviour or attitude vs. correlation (Woodruff et al. 2014). Another plausible reason is for the high level of privacy fundamentalists in this survey is through as an impact of a learning effect as the users completed this privacy survey. However, the high levels of privacy concern are demonstrated in earlier questions on awareness and expectations of tracking and data collection.

Using only quantitative techniques is insufficient to explain the depth of expectations. This depth is represented by the contextual nature of privacy, where respect for purposes, boundaries and context of personal data use adds meaning (Nissenbaum 2015). Privacy of personal data involves not just data about you, but is constitutive and tightly linked to one's identification of self and notions of human dignity (Floridi 2016). Further recent empirical analysis puts forth a compelling argument that the gap between intent vs. action persists because behaviour as measured via quantitative survey instruments are ambiguous and fail to capture the contextual nature of privacy behaviour in reality (Martin & Nissenbaum Forthcoming). To further explore these nuances in privacy sensitivities, a qualitative analysis of privacy expectations is undertaken in the next section.

5.3 Privacy Expectations around Data Release: Qualitative Thematic Analysis

A qualitative thematic analysis of privacy expectations of users around data release is presented in this section. The main goal of the thematic analysis is to uncover what constitutes appropriate data flows. This analysis identified instances of appropriate data flows using the types of data flows and their associated privacy risks which extends the work reported in literature, primarily from a legal perspective (see Table 5.12 for an overview).

Data flows can also be called information practices in social interactions such as search because data is in constant motion when users consume online services. “Practice is first and foremost a process by which we can experience the world and our encounters with it as meaningful.” (Wenger 1998, p.51 as cited in Dourish & Anderson 2006). Information practice is about “the ways in which we collectively share, withhold, and manage information; how we interpret such acts of sharing, withholding, and managing; and how we strategically deploy them as part and parcel of every social interaction” (Dourish & Anderson 2006, p.335). Therefore, practices afforded to users can be viewed as mechanisms to practice privacy. This notion of privacy as a practice is used as a frame for analysis of the respondents’ privacy expectations around data release. The affordances embedded in searching make these data flows meaningful in exercising the privacy expectations users hold. These data flows include creation, collection, processing dissemination of data, and deletion are further explored in Section 5.3.

These data flows were analysed through the narrative thematic analysis to identify information practices. The links between the data flows identified from literature, potential privacy risks and its rationale, reference to the thematic analysis scenarios and the derived information practices were amended in Table 5.12 (Categorisation of Data Flows and Privacy Risks). These derived information practices are subsequently evaluated under the contextual integrity framework to determine if information privacy expectations are met or not. This evaluation provides an understanding of the current online privacy expectations in addressing Research Question 2.

As a foundation for this analysis of data flows Solove's taxonomy of privacy was used to extend the identification of privacy risks associated with the data flows (Solove 2009). Solove's taxonomy of privacy is based on problems associated with privacy. Although it is often through privacy loss that the importance of privacy is experienced, data flows that involve not just data loss, but data creation are covered to provide a more holistic view of the movement of data and associated privacy sensitivities.

Table 5.12 Categorised Data Flows and Privacy Risks

Data Flow	Privacy Risk	Rationale	References	Thematic Analysis Reference	Information Practices Derived
Creation Data created through the process of interacting with an online information provider.	Power and Information Asymmetries	Social Production – Data freely created by the user is exchanged for online services. However, the nature of information goods is that it appreciates with time and its value increases with aggregated use. The authorship of content is lost and users are not rewarded in a typical model.	(Benkler 2006) (Lanier 2014) (Ariely 2008)	Scenarios mentioned in 5.3.1	The following information practices were derived from the thematic analysis (including all data flows) and evaluated using the contextual integrity framework: 1. Collecting and Retaining Meta-data 2. Collecting and Distributing Data for Commercial Purposes 3. Framing Choices through Ineffective Affordances
Collection Data collected by providers from a user through an interaction	Surveillance	Surveillance is the observation and recording of a person's actions as captured through data flows.	(Solove 2009)	Scenarios mentioned in 5.3.2	
	Interrogation	Interrogation is the use of questions and other techniques to seek information.	(Nissenbaum 2011; 2015 Solove 2006)		
Processing	Aggregation	Aggregation of data creates links between different interactions.			
The transfer, communication and transmission of data between a user and a provider	Identification	Identification is associating different data flows with a person through various mechanisms such as linkability and personalisation	(Nissenbaum 2011; 2015) (Solove 2009) and Survey Findings	Scenarios mentioned in 5.3.3	
	Insecurity	Insecurity is the poor protection of data from significant risks such as unauthorised access. "Insecurity involves carelessness in protecting stored information from leaks and improper access" (Solove 2008 pp. 104).	(Solove 2009)		
	Secondary Use	Secondary use is the use of dat gathered data for purposes other than its original stated use.			
	Exclusion	Exclusion occurs when data is used for purposes other than originally prescribed and the user is not informed.			
	Inconvenience	The effect experienced by users when they have to put in significant time and effort into implementing privacy protection mechanisms because existing affordance are inadequate.	From Survey Findings		
Dissemination The distribution, sharing or disclosure of data to a provider during an interaction.	Breach of Confidentiality	Breach of confidentiality is the breech of an agreement to maintain the confidentiality of collected data.	(Nissenbaum 2010, 2011, 2015)	Scenarios mentioned in 5.3.4	
	Disclosure	Disclosure is the release of accurate data about an individual that may have an effect on the individual including his or her reputation.	Solove 2009		
	Exposure	Exposure is the disclosure of a person's nudity, sorrow or bodily processes.			

Data Flow	Privacy Risk	Rationale	References	Thematic Analysis Reference	Information Practices Derived
	Increased Accessibility	Increased accessibility is the ease in which data can be accessed and collected.			
	Blackmail	Blackmail is the intent to release personal data about a person usually not in their best interests.			
	Appropriation	Appropriation is the use of a person's identity and associated data to assist another entity's purposes.			
	Distortion	Distortion is spreading inaccurate data about a person.			
	Misinformation	The risk incurred when users take decisions based on incorrect information which they are unable to otherwise independently verify and make their own judgement on.	From Survey Findings		
	Invasion	Invasion concerns mostly with physical privacy, although the risks can be of an intrusive or an interference nature	(Solove 2009)		
	Intrusion	Intrusion is the disruption of a person's peace of mind.			
	Decisional Interference	Decisional interference is the ability to intrude into a person's personal life and influence decisions made.			
Deletion Users want to delete data that is no longer of use or relevance to them	Inhibition	Deleting data is not easily permitted by current design architecture. This inability is further supported technological capacities to store and the commercial incentives to use data analytics to drive business profits and goals. The risk is temporal because information taken out of context loses meaning, may be used for purposes beyond the original intent of releasing and may simply become redundant with time and evolving beliefs and interests. This interaction is also open to the risks under processing and dissemination. Currently only account signed-in deletion functions exist, although users have no method to independently verify if this has occurred.	From Survey Findings	Scenarios mentioned in 5.3.5	

Idea networks were used to visualise the thematic analysis, drawing connections between key conceptual frames. Figure 5.27 illustrates an idea network, exploring expectations associated with data flows. The size of the nodes represents the strength of the connection between these conceptual frames. Each expectation was categorised under a data flow (in colour) and linked with an action (in black). In some instances privacy risks (such as aggregation, identification) were also used to differentiate expressed data flows from similar expectations.

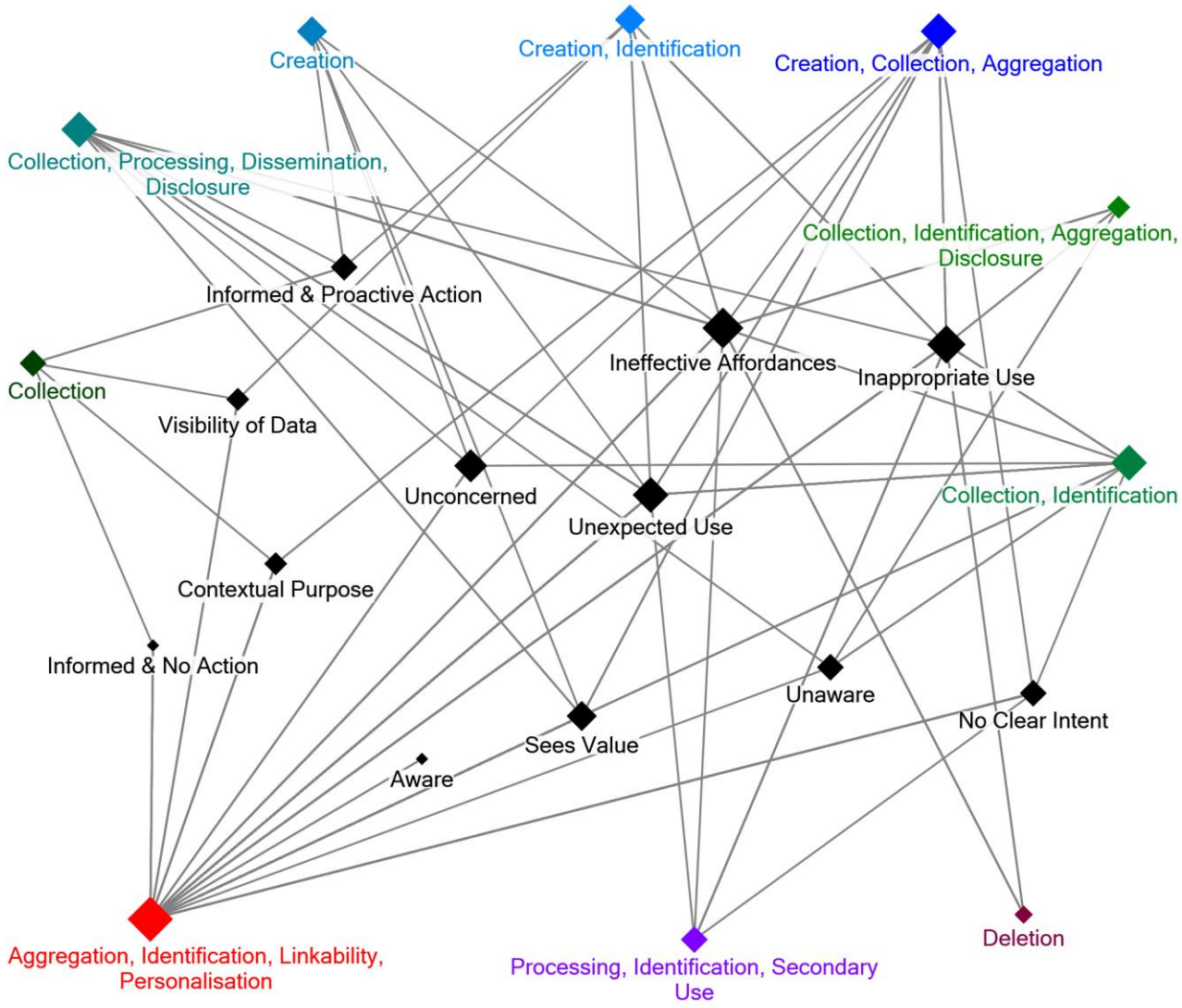


Figure 5.27 Privacy Expectations around Data Flows

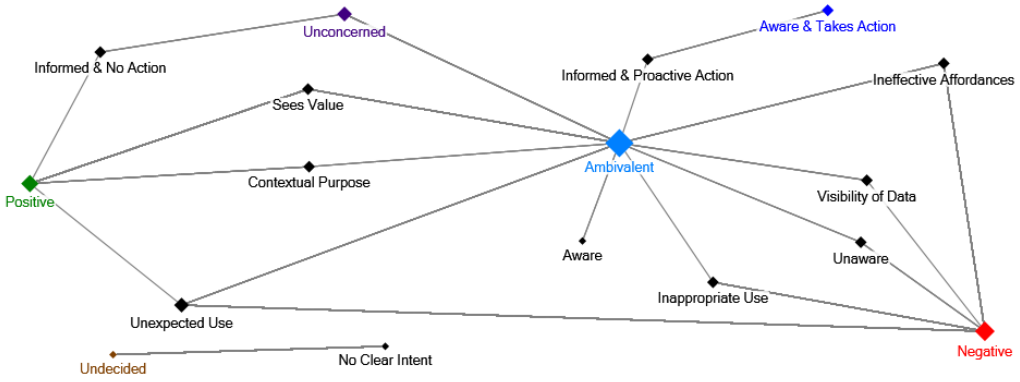


Figure 5.28 Privacy Sensitivities

Figure 5.28 sketches a second idea network, demonstrating the types of privacy sensitivities attached with data flows. There was a wide spectrum of sensitivities users associated with interactions that influenced their expectations. Sentiments expressed include positive, negative and ambivalent expectations, but also those such as unconcerned, undecided and informed views (illustrated in colour). Actions or behaviour relating to these sentiments are depicted in black. The detailed qualitative analysis of privacy expectations is presented in this section.

5.3.1 Creating Data

Creation of data involves both personal data released by the user and data about them, such as personally or uniquely identifiable data like location or browser-fingerprinting identifiers. Expectations under this data flow contain instances of creating data but also include cases involving privacy risks of identification and aggregation.

Scenario 1 – Awareness about Data

i. Awareness about Surveillance Risks with Metadata Retention

Awareness about how data is created and released is growing. “I hope my ISP will not profile me and sell that data but I wouldn't be surprised if they did. I expect government will collect everything they can but I wish they wouldn't” (P#361).Heightened awareness was seen in the attentiveness to the recently introduced metadata retention law in Australia, where ISPs need to collect and store metadata for 2 years (*Telecommunications (interception and access) amendment (data retention) act 2015 (cwlth)*)

> YES I expect to be tracked, but do I like it? Absolutely NOT. I am not 'satisfied', I do NOT WANT to be identified or have my information stored and used to formulate a profile of me… I DO care, I

> DON'T like it, I don't want it to occur. What I particularly hate is that service providers must store 2 years worth of meta data so ASIO can search through it at their leisure and only one search warrant is required to access a 'network' of computers, with no limit being placed on the number they can access. We are turning into a totally surveillance state, do I like it? NO (P#304)

On the flipside of these concerns underlying surveillance, users are sometimes willing accept the use of these practices on data retention and profiling for public good. These users place trust in well-established social institutions, such as governments, considering them to be trusted parties working for end-user interests and benefits such as safety of communities and nations. *"I'm happy to allow government to track my personal data use so that they can track/ prevent crime."* (P#404). Others express a willingness to accept practices of retention and profiling by established social institutions for public good such as a "*Government website [like] (centerlink)."* (P#307).

ii. Awareness about Commercial Risks with Data Release

Others display awareness of more prevalent commercially-driven privacy risks. These views are: *"Generally unhappy about profiles being built up"* (P#441);*"Automatically using cookies when you go onto the particular website"* (P#10) There is a need for more effective legislation regarding this commercial data collection. "*Shouldn't be tracked at all, government should introduce regulations to prevent unnecessary tracking by third parties"* (P#119).

Some users display a low level of awareness about how tracking occurs as seen in *"wasn't informed that my personal information would be stored"* (P#140) or may be unconcerned *"don't have any such experience" (*P#424). This lack of awareness highlights the importance of educative goals in providing visibility on how data flows through search. This gap in awareness affects how other information practices such as aggregation, collection and disclosure of data are viewed. *"I wasn't aware that any personal data was actually used when I searched."* (P#395) or *"I am not comfortable with any of my growing history being retained to refine search results."* (P#396) or *"I don't feel safe if people can keep track of me."* (P#423). These views express a lack of awareness of how personal data is collected and represented across different sites and increasing privacy risks in the online space.

The hidden and complex nature of the medium in which data flows makes it easy for users to be uninformed or allows room to be misinformed. *"I have only heard that our Web searches are tracked, I guess I really have no idea about what goes on. However, Google comes out every year with most*

searched items, and I feel like individual tracking must take place to find cyber criminals, inevitably resulting in law abiding citizens having their searches tracked also." (P#421).

The degree of visibility in how data is represented and linked across different information services is not easily apparent to users. *"I have not been directly involved in an instance where I felt my privacy 'invaded' whilst using a search engine. If I have, I was not aware of it…. It does not affect me for the most part, so I cannot say I mind or do not mind…"* (P#418).The privacy risk to the individual is in acting on these views which are not fully formed, as users blindly trust popularly disseminated views in the media. The media itself driven by advertising attempts to propagate stories that are likely to go viral or gain traction in the race to compete for diminishing consumer attention spans.

Other users are aware of the pervasive nature of data collection. *"Marketing utilise personal information to sell you products. The amount of cookies cached in Web browsers is crazy. My expectations are that embedded or popup ads should not be allowed on Web pages that interfere with the main intent of that Web page. E.g. Looking for recipes and getting sports bet pop ups."* (P#62).

However, some users are aware and take action in creating and releasing data. *"I expect that everything I enter online to be recorded therefore I don't enter anything that I don't want recorded."* (P#13); *"I'm very serious about privacy, my real name and any details about me are not online."* (P#42); *"I know that personal data is being retained and this makes me uncomfortable so I do all I can to minimise the risk to me e.g. using a VPN and adblock software, deleting history every day*" (P#430). These descriptions fit users who are informed, understand how data created through different transactions can be linked and take pre-emptive actions to not create (type in) data that he or she does not want captured. This type of user is probably aware of data minimisation strategies and treats the online world as a public space. For others expectations are met by controlling release and taking specific precautionary measures. *"I don't really have an issue because I don't tend to let out personal information on the Web or allow for passwords or IDs to be saved for future use"* (P#92).

Others are comfortable in releasing data, but constrain it depending on what is termed as sensitive to them. *"I don't mind if Google used my search history of places, shops, etc." (P#151); "I don't mind my personal information being used if it's just to determine what sort of things I shop for but not really anything more personal"* (P#138). It is therefore important to understand what kind of data is personal or sensitive to users.

Scenario 2 – Identifying Sensitivities

Some users are unconcerned with releasing personal information, given that it has become normal for users to sign up and authenticate their interactions with service providers. *"I don't mind birthdays and names being provided" (P#28).* Others tie sensitivity with money where vested interests lie. *"More secure option for online banking" (P#29).* Users also express awareness of potentially sensitive data sources and impacts "*My Facebook profile can be used*" (105); *"Facebook ads know my sexual preferences"* (P#215); "Ads for IVF on Facebook" *(P#206).*

Others consider location data to be sensitive. *"When I found out that the Google maps tracks its users."* (P#409); *"I don't like it when an IP address is tracked and searches from that IP are used to make [recommendations] for a user"* (P#302). IP address or location is now considered to be personally-identifying information in a recent statement by the Federal Trade Commission which regulates advertisers and information providers in the US (Rich 2016). Other common sources of personal data considered sensitive include address (7%), date of birth (6%), home phone number (4%) (Office of the Australian Information Commissioner 2013). 71% of UK consumers feel comfortable sharing location data (Chartered Institute of Marketing UK 2016).

Some users outline interactions that taken out of context could be sensitive and have real-life impacts. *"when I was searching information about the genetic illness trials in my family, some personal information had to be given"* (P#217); "*When I have been looking on health sites to compare my health to regulation standards"* (P#369); *"for example searching for pregnancy information and seeing targeted advertisements on social media and in-site advertisements"* (P#281).

Scenario 3 – Revealing Identities

A key concern among users is tying real-world identities to online interactions. *"when I searched and I'm asked to sign up an account where I have to reveal my identity*" (P#75). Excessive release of personal details is viewed as non-essential from a user point of view. *"Many websites require registration of personal details in order to use their service which is annoying and unnecessary"* (P#83).

These users express dissatisfaction with having to sign up and release personal details to access services. *"I don't really care as long as my real-life name is not identifiable or able to be inferred."* (P#132). *"Facebook consistently finds new ways to creep me out. Ignoring their real names policy and using a fake name helped, but deactivating my account seems to have worked even better."* (P#51)

Users expect the ability be effectively anonymous online *"Although it is inevitable that our online behaviour is tracked, I wish we could remain as anonymous as possible."* (P#290); *"I expect personal data used for improving the quality of search result, however those data could be under anonymous ID and should not directly linked to my personal detail"* (P#161); "*Should be treated the same as if I was browsing in a library i.e. anonymous."* (P#82). The ability to separate offline and online identities and affordances that allow users to implement these expectations are important to users.

Another a user acknowledges that accountability for interactions is important, but emphasises that associations made to link identities through aggregation, linkability and dissemination of that data is valuable to organisations. *"I expect that information remains 'private', but understand that there needs to be accountability for online use of the internet. I don't think anonymity is necessarily the answer, but would like not to be traceable by corporations who are interested in profiting from information they may have gained from me."* (P#390).

Scenario 4- Disconnected Interactions

The pervasiveness of the interaction medium in capturing data is a concern. *"I recently looked up places I had stayed last year in Thailand for a friend, who is thinking of travelling there. For the next week or so, I had adverts all over my Google searches and Facebook for places to stay in Thailand. I don't like the idea that my searches are recorded in this way and used to create a personal selection of advertisements."* (P#181). *"Sometimes it's brilliant at jogging a memory to do something and other times it just plain old creeps me out."* (P#274). In these cases, the user finds the ability to track search history creepy.

Sometimes users are misinformed about the extent of existing data protection mechanisms. *"I don't mind when my info is used online because if I need to use PrivateBrowsing then I will enable this feature but otherwise I'm not concerned if my private data is tracked."* (P#320). Although the user is aware about private modes in browsers where data is not stored at the client-side, this does not mean that data leakage does not occur with data captured by third-parties to drive advertising.

Other users report on the ineffectiveness of affordances in expressing privacy expectations. The ability to adapt privacy settings according to a user's needs is a requirement that design of technology products and services need to address to facilitate appropriate release of data. *"I stopped using Google Plus because I couldn't adapt my privacy settings as I required."* (P#280). In other cases, this requirement is

important to prevent exposure *"Settings meant to prevent explicit material coming up in searches but they still did*" (P#417).

Some users illustrate examples where existing measures fall short of expectations.

> I feel it should be easier (and people more aware) that we can reset the unique identification number, and even disable it in some applications/interfaces. I also feel that the general public are not aware of the data collection and its implications. After I deactivated my unique identification number I still find ads are targeted at me following searches. (P#35).

Scenario 5 – Actioning User Choice and Consent

It is clearly important to provide users with the freedom to choose. Autonomy is central to the privacy of an individual.

> How can somebody else deem what is relevant to me? I would prefer anonymous searching so that I can get all relevant results, not what someone else deems as relevant. (P#184)

Interestingly 98% of respondents in a separate Australian survey felt that Australian law should include a right to privacy which at present it does not (Arnott & Andrejevic 2012). Similarly 90% of European respondents perceive that individuals should have a right to privacy (European Commission 2015b). Respondents in Europe are of the view that responsibility for data protection lies with organizations (67%), people (66%) and public authorities (55%) (European Commission 2015b).

Providing users with choice is still important in establishing trust with online information service providers.

> I am concerned about the amount of tracking that happens that I am unaware of. Who is collecting this data and for what purpose? I do not like the idea of direct marketing, if I am interested in a product/service I will contact you. (P#184)

Others express desires for data to not be collected *"No personal data of the user should be recorded and stored as there is no real need of it*" (P#415). Others say collected data should be privately held *"I expect for it to be completely private, and when search history is deleted it should be cleared. However, I am under the impression that this is not the case."* (P#433). These diverse views further support the argument that personal data is constitutive, influencing and shaping a person's being and their interactions (Floridi 2016).

The mechanisms through which users provide consent need to be direct, visible and transparent. For users, it is still an important consideration to be made aware of and be able to freely provide permission for organisations to use their personal data in return for perceived value gained through the interaction. *Okay with using relevant terms for search results when the user has actually given their consent, rather than indirectly through terms and conditions" (*P#119). Consent is important when disclosing data to parties. The ineffectiveness of existing consent measures is evident: *"I expect it to be not used at all unless I give explicit permission"* (P#98). Convenience with consent mechanisms is useful, but when users are unaware of its presence or effects it leads to dissatisfaction. "*I think the main reason is because you are not generally informed (that I know of) and then suddenly similar items to your search start coming up and you feel as though you are being tracked. It's not necessarily negative, just a little unnerving."* (P#228).

Similarly, confidentiality is a key theme impacting unexpected data use. *"Don't share [data] with 3rd parties for beneficial purpose"* (P#25); *"Personal data should keep as private and confidential" (*P#29). Others would like to be periodically notified when their data is used.

> I'd like to receive reports periodically (monthly) informing how my data has been used. Even if I don't always read through the reports, I'd keep them as evidence of how my info has been used (P#104).

Existing informed consent mechanisms are ineffective in providing users with choices without the effect of technology lock-in (Naughton 2012). This user articulates dissatisfaction in not having any real choice when wanting to engage with online information service providers.

> Unfortunately I'm well aware that users have very little control over what we share online. Policy at the moment is, Accept the T's & C's or you can't use the service. Accept it or go away essentially. (P#302).

Another highlights the importance of storing the data at the client (user's) end. "I expect it to be stored on my device not on an external source" (P#99).

5.3.2 Collecting Data

This data flow involves the collection of data from a user through the process of interacting with an online information service provider. Privacy risks concerning data collection has been widely studied in related work (Nissenbaum 2010; Solove 2009).

Scenario 1- Perceived Uses and Value of Data Collection

One clear requirement for beneficial use is that users want convenience. *"When I'm typing in my address and details for a parcel, the browser will automatically fill it in. I like this, it makes it a lot easier and quicker"* (P#37). Some users find value when data is aggregated and updated real-time to notify of changes that could potentially impact them. *"When Google started to read Gmail and identified my flight details and notified my [me] of delays to my flight"* (P#419). This instance speaks of the visionary goal behind Google; which Eric Schmidt calls as its key purpose: "I actually think most people don't want Google to answer their questions, they want Google to tell them what they should be doing next" (Jenkins, HW 2010). The danger in adopting such a view is that the autonomy of an individual to freely decide and make associations online is threatened. Users interact with a medium where their affordances for expressing privacy are largely framed by the algorithmic decisions and a lack of meaningful choices.

Perhaps the usefulness of these recommendations are constrained to specific instances and not all facets of life, which is supported by user expectations. Users are happy to disclose data if the value exchange is beneficial to them. *"I don't mind if it's something academic related or part of sporting clubs where they post your achievements"* (P#319). Others consider this convenience negatively as they are unaware of how personal data is collected and represented across different sites. *"Sometimes the personal information might save as 'Autofill", and the information were appeared on some website without any permissions"* (P#71).

They question the trade-off between convenience and security of personal details against disclosure. *"when my credit card detail were saved and I didn't want it to be"* (P#358).*"E.g. sometimes user name will stay on the page for next person. This is not good for security although it makes people easier to sign in again."* (P#297). Others expect data to be tracked, but held securely, particularly sensitive personal data. *"Maybe user can't stop their information would be collected by the data service, but at least please keep them in a secure place and not sell user details to other companies for any reason."* (P#401); "*Parties you are unaware of should not be accessing it."* (P#337). Users would also like to be notified when unauthorised access occurs. *"To identify if other people / hackers had accessed my account because of different behaviour has been found*" (P#22).

When data is collected beyond its primary purpose, it does not sit well with users.

> I attempted to compare health insurance online and instead ended up being hassled by telemarketers for the next month trying to sell me health insurance (P#189).

Here instances of personal intrusion that have offline impact and real-world implications are seen, through actions that occurred in the online world. The risk here is that of disclosure of details to third-parties (actors not privy to the interaction)

Scenario 2 – "I want this, but it depends on …" Understanding Contextual Interactions

Some users express a positive requirement to gather data to provide recommendations, but stipulate a temporal constraint for these types of interactions. *"It would be relevant to save the data that I searched for first time and toss/forget it after a month. For example, if I am interested on buying a watch, I might search them and considering buying them. However, after a month, there could be chances that I had bought it or changed my mind. At this time, it would be useless for such things to be shown again."* (P#22). As such the user wants the measure of relevance of recommendations to appropriately fit changing user interests and not just be based on clickthrough data.

Another wants relevance of search results to include locality which is of value in particular contexts. Location is consented to be released when searching for locally relevant search results. *"While I don't have examples of my privacy being circumvented by online searches, I believe that privacy information is used. I try not to look at the inevitable adverts, so I'm not sure if my privacy information has been used to market things at me. I know that my location is used, but I'm not sure how the information was gathered - but I don't mind that as some things I search for are local."* (P#36). These views also underscore the relative difficulty for users in understanding the extent of the data they release in each interaction.

Some users prefer constraining interactions to specific actors and purposes choosing to limit the visibility of data exchanged to within the context.

> Personal data should only be used within the confines of what they want to be used for. If they want a more customized search it should be made clear that they are using data that they have collected from previous searches and Identity information. (P#10).

"I'm a private person so I don't like the idea of my information/ details being tracked and retained at all. This is probably particularly so with App's such as fb where information is shared for my family and friends to see. E.g. when I comment or like something. It limits my interactiveness with FB to a bare minimum, opting usually to pm or browsing but not actually commenting, sharing or liking." (P#106). In these instances, the disconnect between existing mechanisms and user expectations is visible. The

user has a specific purpose behind these interactions and does not believe in the semi-public or public modes for expression embedded by the application medium. The user alludes to the freedom to associate, considers these associations personal and implements different mechanisms to manage control of the flow of data and the actors with which s/he associates with.

Information should be constrained to the purposes of collection. It is increasingly becoming the norm to harness the capability of technology-rich information service providers such as Web analytics at the cost of information privacy. *"UniSA embedding YouTube and Google Analytics into their website to track when uni students log onto the site. This info is made publically available via Google."* (P#12). These specific capabilities are no longer viable to be grown in-house in organisations and as mainstream technology companies establish prominence in the marketplace, it is easy to essentially outsource the capability and manage it through vendor contracts.

While users may agree that the benefits supplied to the recipient organisation are immense, the lack of controls embedded in the platform to protect information privacy (e.g. the data being publicly available and visible) concern users. Organisations seeking to tap into these capabilities should make these embedded privacy controls part of the negotiating process and not simply or passively accept the license to use the platform as-is. Organisations may tend to do this once a critical point has been reached when there is visible public interest. Another mechanism that could facilitate this is when enforced through legislation. For example the metadata law requires strict controls for collection of data. This has seen some technology service providers move data centres within the jurisdiction of the country for ease in management and compliance with the law.

Linking data across different services is not desired by users although the capability exists as the underlying architecture is the Internet and applications run on the WorldWideWeb. They expect each interaction to be constrained to a primary actor and purpose. *"I expect when I am requested to give personal information that it remains private and secure between the two parties"* (P#264); *"When the ads on Facebook are targeted to the things I have searched on eBay. I do not think Google and eBay sites should be sharing information with Facebook about my buying habits."* (P#364). Others extend this notion that data should not be shared with third-parties unless consent is clearly provided "*I don't believe personal data should be given to a third party unless consent is provided."* (P#265) and identify actors who they would rather not have access to their personal data *"I would not expect family members or employers to be able to see my personal data or search history"* (P#240).

This contextual nature of interactions is also seen among US consumers who are reported to make risk-benefit decisions when choosing to share data. Criteria considered include trustworthiness of the organisation, assurance about data collection and protection practices, accessibility of data to third parties and limits on data retention practices (Rainie & Duggan 2015).

Scenario 3 - Available Affordances Ineffective in Managing Data Release

The ability to easily commodify personal data is at odds with information-seeking goals of users. *"I was buying flights online, and the prices kept increasing. Even though I was clearing my full search history/cookies/data, the flights were still at a high price. So, I felt that even though I made an effort to clear this data, it was not actually cleared. When I used another computer, the flight prices were significantly cheaper. However, if I searched on incognito mode, the flights would not load on Skyscanner, until I enabled cookies."*(P#86). In this instance, the user wishes to research different flights. However, the use of persistent cookies and ineffectiveness of mechanisms such as incognito mode and client-side options to delete search history hinders management of data release.

> Since upgrading to Windows 10, I've been a little concerned about how my browsing history seems to be used to serve ads within Skype. Even from things I have liked or shared on Facebook, despite my Facebook account not being linked with Skype. I've turned off as much of the tracking as I can in the operating system, but I'm still not happy with how much it still seems to track. (P#194).

This user talks about not being able to manage release of data, where data capture now occurs at the operating system level. This example indicates the increasing difficulty users have with to contend with if they want to keep using established software. Presumably users are locked-in to the platform either through accounts or familiarity via use. Significant user effort in relearning new software is needed to combat this difficulty and changeover to a new platform. This evidence points to the malleability of privacy expectations where implicit factors such as default settings can induce a decrease in privacy concerns and affect user behaviour (Acquisti, Brandimarte & Loewenstein 2015).

Some users report on the effectiveness of existing privacy-enhancing tools such as ad-blockers to achieve some level of privacy protection. *"As I have AdBlock enabled, obvious events of privacy exploitation are not immediately visible to me, as things such as personalising ads based on previous searches don't show up in my browser."* (P#116). It is hard to take effective action with the multiplicity of erroneous software that exists when users seek privacy controls. *Searched for an ad blocker or antivirus software and an advertisement was redirected to a fake website for malicious intent. After*

going to this website my information was used to display more ads about antivirus and the browser kept sending me to dodgy websites." (P#62). Here, a strong foundation in the fundamentals of privacy protection and genuine tools that may assist the user may shed some light in implementing effective privacy and works towards building user trust with technology to safeguard their privacy.

The pervasiveness of tracking is visible across different information services as collection is linked and data is processed and disseminated across these platforms. *"After sending emails regarding self-publishing for an assignment, I found many of the ads on sites I visited changed to promote publishing offers and deals. This was particularly concerning for me given I had only exchanged emails on the topic, and hadn't use search engines, making it obvious that the subject of my emails was being used to target ads."* (P#248).

Users express frustration regarding cookie notifications which function merely as compliance measures and not usable options. *"when all the sudden they list that cookies are being used and continuing means that you are agreed with the terms and conditions and it is on the bottom of the page which hardly can be noticed"* (P#344). For users, that don't agree with cookie usage, the only real option is to not engage with products or services. This view that respondents do not really have a viable choice to not engage is reflected by European respondents as well with 71% saying some personal data exchange is part of normal life online (European Commission 2015b).

Scenario 4 – The Pervasiveness of Data Collection: The Creep Factor is Real

Users are aware when data is collected through the effects on its subsequent interactions. *"When adds pop up that are based upon my previous searches, when my email is given to other parties and I receive emails from them"* (P#278). Awareness about potential data collection sources is seen when profiles are visibly created as with Facebook (*"My Facebook profile can be used."* (P#105)). Some users expect that data should not be collected at all. "*I don't want it to be used"* (P#74). Others voice a strong need for autonomy. *"I do not want my info used at all ever for anything. If I want to find something I will look for it end of story."* (P#253).

Given the complexity in how data is captured, there are also indications that users are unaware of how personal data is collected and represented across different sites.

> When I shop online, then go onto Facebook and find an ad for the object I was just looking at for that price. I just wonder how they are able to get that information and I get bit creeped out by that. (P#333).

Another user reports on value obtained when collection occurs within a specific context (i.e. a shopping site in this case)

> if my shopping choices within a specific site are recorded i.e. an online supermarket then I don't mind choices coming up in that instance if I have logged in to the site. Coles online will bring up gluten free items for me for example and that doesn't bother me but I don't like to feel like I am being watched! i.e. the way Facebook takes your Google searches and makes them targeted ads for you, creeeeeepy. (P#252).

Creepiness occurs when an activity is ambiguous enough to not be specifically threatening, is somewhat in accordance with data collection purposes and does not evade privacy settings (Tene, O & Polonetsky 2013). The risk to users is when data is disclosed to parties not privy to the originating interaction. “I am active when it comes to online campaigns to help animals or the environment. Unfortunately this always results in letters & pleas for donations from multitudes of companies & 'charities'.” (P#273). This disclosure to unauthorised parties adds to the feeling of creepiness expressed by these users.

The creepiness extends to security risks when users become targets for hackers as seen with instances like cyber-bullying, revenge porn cases, the celebrity nude photo scandals and the iCloud account hacks (Boyd 2014; Citron 2014). A key feature underpinning these attacks is the use of social engineering, which stems from a lack of awareness and understanding of technology use to safeguard privacy and security. This is aided by the practice of setting default settings to maximise exposure of data.

Increasingly collection of data happens pervasively with user consent and controls in managing it becoming less of the norm. *“How when for example you are logged into a Google account it automatically finds your location but doesn’t tell you that it is looking for it.”* (P#202). *“Lately Google has not been asking if it can use my GPS to track my location and has just been doing it anyway. There are certain people and organisations I'd rather not have know where I am 24/7.”*(P#213). Here the technology mediates release of location information for all interactions, whereas users differentiate between interactions that require location being released and those that do not. Through these examples it is clear that release of location information is context-dependent.

5.3.3 Processing Data

The data flow about processing involves the transfer, communication and transmission of data between a user and a provider. It may involve privacy risks experienced through effects of identification and aggregation.

Scenario 1 - Inappropriate Use

Some instances of processing data revolve around insecurity when personally-identifiable details are stolen and disclosed to unauthorised parties, who then may use it for malicious purposes. *"I had fraudulent activity on my credit card that was only used for online purchases."* (P#58). Although it is unclear if these cases were influenced by carelessness or ignorance by the user in interactions involving sensitive personal details, what is clear is that the risk is financial and clearly visible to the user, who expects this sensitive information to be securely accessed only by authorised parties. Others expect personal data to be securely held and not be sold. "*Personal information entered into "secure sites" (i.e. payment information) are not disclosed to third parties under any circumstances - nor browsing history etc."* (P#396); *"Sold to international bodies etc Indian call centres...drives me nuts. Personal info has been sold for years."* (P#169).

Inconvenience is another type of impact on the user when data is processed for unexpected purposes. Spam is a common outlet for inconveniencing people such as through *"Unsolicited emails*" (P#159). *"I received emails from people I did not know"* (P#67).

Scenario 2 - Constraining Use

The freedom to think and explore topics of interest without being surveilled and risking negative consequences is key to this user's trust and freedom of speech? "*When I used to go to school, in biology we had to look up the human anatomy. I was tracked because I was using penis, vagina, etc. and I felt like the school were not trusting us enough as year 12's."* (P#94). The nature of the medium easily allows the capability for information-seeking actions to be tracked and then evaluated according to institutional guidelines whereas previously users would not have been subject to this level of immediate surveillance. As this type of big brother watching becomes commonplace, users may continue to keep searching as-is or perhaps self-censor their actions depending on their level of concern.

Scenario 3- Managing Impacts of Use

To aid their information-seeking goals users are provided with conveniences such as continuous account login across browser search sessions. However, this user reports a case where convenience was at the cost of security.

> I searched online when I was logged in on google chrome and when I signed out of my profile, someone else used the computer and saw what I had been browsing, even though I logged out. (P#149).

The saved search history could be potentially accessible across different search sessions on the browser, assuming the same login to the computer was used. Here, the user wishes to protect visibility of information considered personal across different interactions on the same device. Incognito browsing may have helped the user in this instance assuming the user is aware of this affordance.

Scenario 4- Non-Targeted Uses

Non-targeted use is when users interact with a technology in ways that its designers did not intend (Friedman, B & Hendry 2012). Some value is seen in using tracking to monitor online activities of children. *"It's helpful having previous searches shown up especially when using it to monitor kids' activities"* (P#282). Perhaps the convenience of tracking, inadequate knowledge or insufficient effort in using online child safety tools influences this non-targeted use of tracking.

Another non-targeted use is deliberate use of search to find out exposure or publicity in relation to a person's identity and activities. "*When I googled my full name and it came up with my Facebook profile page as one of the first options."* (P#432). These instances also show the ease and the extent to which data capture and dissemination is prevalent among different information service platforms, so as to provide a connected view of individuals. One user reports on an instance which stretches the curiosity boundary from plain inquisitiveness to creepy intrusiveness, where the user presumably wishes not to be tracked when searching for data on other people. *"When searching for data on other people"* (P#41). Perhaps it is becoming commonplace to google people particularly before you actually meet them as echoed by this user *"I'm afraid that companies can use people's personal data as a tool to select them for jobs".* (P#439), a practice driven by job recruitment sites such as LinkedIn and presence of public-facing employee profiles.

5.3.4 Disseminating Data

A majority of privacy expectations (216 of 441 reported expectations) dealt with aspects of data aggregation, identification, linkability and personalisation. These sentiments were then categorised into scenarios on value, convenience, relevance of recommendations, contextual value, linkability and value exchanges, perceptions around data dissemination and tensions in information asymmetries.

Scenario 1 – Aggregating Data for Positive User Value

It is important to understand if users value aggregated data and uncover the underlying reasons, as the use of aggregated data is a key benefit often cited by information service providers as essential to their service operations. Some value is seen with realising information-seeking goals of a transactional nature that usually takes the form of online shopping. *"I don't mind search engines tracking what I'm shopping for"* (P#44). Users also find value particularly if the ads match their current interests and intents during that specific interaction. *"When holiday/clothing/ any item for purchase is being advertised in my relevant results. it makes it convenient, as it is what I'm looking to purchase online at the time."*(P#363). In addition, users like the customisation offered where preferences for regularly bought items are remembered. *"When shopping searches are used to find better deals" (P#384).* Often users see the impact of data released via a primary search query in subsequent interactions on the Web. *"I do not have an example however I have noticed that if I have been on clothes websites that ads regarding the website have been on the side of Facebook or other sites. And this has been useful as I have picked up an item on sale" (P#118)*

These recommendations are visible across interactions on different service platforms given the ease of data capture allowed by the openness of the interaction medium, the Internet.

> I don't mind my information being used when websites such as eBay offer similar items to what i have previously purchased as I find this occasionally helpful. (P#365).

Others extend this value found in transactional information-seeking goals so that recommendations align with their interests. "*When online shopping, data I enter about myself or what i am looking for. I don't mind this being used to help me*." (P#120); *"I purchase a number of books online and don't mind suggestions based on previous purchase history. I often find new and interesting authors this way"* (P#211).

Value is also seen with information-seeking goals that address purposes beyond commercial motivations. *"I don't mind if the data isn't personal- for example uni research"* (P#353). This user reports on the perceived benefits of mining aggregated search data to uncover scholarly recommendations and gain new insights. *Contributing to non-commercial purposes "I do not mind when I am searching things for assignments, because then the data can see what is helpful and being used regularly on the internet and what isn't."* (P#435).

Scenario 2- Convenience is Expected

Positive uses for personalised recommendations are seen among users, as they value the convenience it brings provided it's not overly intrusive. *"I don't mind targeted advertising based on my purchase history as most of the time this is convenience, appreciated to a degree as long as it's not too intrusive/personal"* (P#11). These users report on monetary benefits received as it is convenient to users when data is profiled and represented as recommendations.

User intent is key to these recommendations being perceived positively as most users had intended to browse options that could later inform their purchase or to actually purchase an item. *"When I went online to find a hotel room to book for my birthday, i was really happy the next year when i found a really good deal on the ads on Facebook for a hotel room in the area i wanted to stay the year after."* (P#236); *"When searching for a hotel or cheaper prices on sportswear, I was able to end up saving over $100 on items due to customised ads to my preferences being used on websites."* (P#436). These users report on relying on the convenience data collection and aggregation brings to help them stay informed of potential bargains of interest.

The conveniences embedded in defaults that guide search browsing behaviour are useful for some users. *"When previously searched pages are highlighted it can be easier to find them later"* (P#315). As users become habituated to these embedded defaults, they tend to rely on them when navigating search results. *"I don't mind it when my searches of clothes or flights are kept and used to tailor the ads that are seen on the side bars of internet pages. This allows me to kind of keep on top of specials etc. without flooding my email account"* (P#359 *"I don't mind when the search uses my search history to predict what I am typing in"* (P#179). Hence, convenience is a value embedded into the design of these interactions, both in outcomes (personalised ads or search results) and in interaction mechanisms. Convenience has become a key incentive that's tangible to users and is often synonymous with technology use given its computational efficiencies.

Scenario 3 - Understanding the Relevance of Recommendations to Users

Relevance of recommendations has been extensively studied and reported on using largely quantitative measures that measure the variability between a pair of lists. Examples of such measures used in the technical controlled experiment in Phase 1 include Jaccard Index and Edit Distance (Hannak et al. 2013). However, it is important to uncover measures of relevance to a user. A sample of these qualitative measures of relevance are reported in this section.

Positive Value

The relevance of personalised recommendations be it ads or search results are important to users. This measure may be subjective as alluded to by this user *"It depends on how relevant it is and what pages are"* (P#286); *"Used for ethical reasons is ok"* (P#215). To some users, it is clear they see the impact of personal data being aggregated to provide ads. *"I had searched on magnetic rowing machines a month or so ago, the next time I looked at news.com.au, the side-bar advertising had changed to BeFit Hire and showed a rowing machine even though it was days apart and I only clicked on the page once."* (P#274).

Some users find personalisation valuable when the data signals captured provide recommendations that align with their interests such as better search results.*"I'm fine with it tracking my use to help with future searches, but not happy with it sharing this information with other people"* (P#293);*"I didn't mind when my search preferences were being used when I'm looking through videos in YouTube, since it comes up with similar type videos of what I want to watch as recommendations"* (P#38).

Indifferent Perceptions

> I don't mind if ads are targeted at me based on my searches because ads don't influence what I spend my money on (P#354).

"I don't care if the internet uses my stuff for statistics etc. I don't shop online so the ads etc. won't entice me to buy stuff anyway" (P#177). To some users, the use of personal data for behavioural advertising has no perceivable impact as it does not affect their decision-making.

However, given the increasing advances of innovations in payment, the ability to link transaction information in the near future is a possibility particularly if it is the same parent company offering multiple services. The ability for users to then make independent purchasing decisions becomes slightly

ambiguous as capturing of data is embedded into the platform used for purchasing. This trend indicates the inherent monetary value in linking different transactions on the internet. Others constrain impact to their information-seeking goal, where financial information is considered sensitive. *"general searching doesn't bother me, it's when its personal such as banking"* (P#96).

Negative Perceptions

Although there are benefits in users receiving recommendations as a result of data collection, there are indications of discomfort to users. They see the pervasiveness of tracking as increasing the publicity of interactions which may have been originally perceived to be between the sender (user) and the intended recipient (search engine). *"I was looking into a new laptop and afterwards ads for it would pop up on other unrelated sites like Facebook. I didn't mind too much because it brought up a few sites I haven't seen but it was a bit uncomfortable to think that everything I was searching was being tracked."* (P#340). Sometimes the inability to gain bargains adds to the negative user perceptions around data collection. *"In past when searching airline tickets, accommodation, I've noticed that sometimes cheaper prices seem to disappear the more I search a specific term. Remembers me as being interested in those search terms"* (P#222).

Often this discomfort is termed by instances of use deemed inappropriate to users, frequently expressed as an annoyance. "*searching for trampolines online and then being inundated by trampoline adds"* (P#72); *"Ads relating to my searches. Too intrusive."* (P#313); *"Annoying when I search an online store, and then ads for the same product/brand come up in other pages"* (P#383).

Others don't like the pervasiveness of data collection and dissemination that happens in the background when searching. *"Expectations were not met when search results are used to create ads"* (P#205).

> When you search for something say furniture and then when you are in eBay or Facebook and all these ads keep popping up relating to searches I have performed. Annoying and I'd prefer them not to come up. (P#157)

Some users also do not like the impact of extensive profiling across different interactions. They specifically commented on ads relating to previous searches popping up for example from searching for flights or Gumtree. *"I don't like when I search for something and the ads then appear that are similar to my previous history."* (P#385). The people expected to only view the information or data supplied by the primary information providers, such as Google. They are not expecting the interaction with a different

party to interfere with this primary interaction. They wanted a clear demarcation between the separate interactions.

This discomfort with irrelevant recommendations once normalised with time turns to the inattentiveness that is associated with spam. Users articulate the ineffectiveness of ads by considering it as spam, where personal details are disclosed to third parties. *"I have Google searched the price of an item and later on Facebook that information was being used to spam me with unwanted adds."* (P#275).

Irrelevant recommendations and associated user discomfort is also perceived as creepiness when profiling people's personal interests. Dating ads are sometimes deemed intrusive when inferring sensitive information considered private.

> There's always that 'Date James' advertisement which annoys me, you know? "James is 32, smart and cute". (P#251);
>
> When ads display items that I have searched for previously. Or, these days often times there are ads to 'meet Asian women' when I am browsing the internet. Considering I am Asian, this seems too much of a coincidence and I don't like seeing these ads. (P#290).

This user reported a similar instance encountered with job ads. "I was searching for potential job openings and later whilst scrolling through my Facebook, I found ads for job openings that had my name and other details about me. It took me by surprise that they would use my name and details in an ad." (P#109)

Sometimes recommendations can introduce real-life risks particularly if this type of data leakage is combined with spyware. *"I was looking up flight details for an upcoming holiday. All the advertisements that I saw after this were related to different tourist destinations and flight centre based around my searches. I would personally like to know that if I do go away, not a great deal of people know how long I'll be gone, just due to safety concerns for my house when I leave."* (P#374).

Questioning Relevance

Questions on who it is relevant for remain. Some users are unconvinced that these personalised recommendations accurately target their interests. *"Was on a site once and they recommended I watch certain videos and visit certain sites because they believed they would "interest me"."* (P#337); *"…Google was trying to be 'helpful' by saving me time."* (P#433).

Users also question the relevance of sources used in producing these recommendations.

> While searching for online resources for a report I was writing, the results I was shown were being influenced by past searches done on similar topics. This became frustrating as the relevance of some of these resources to my search was questionable. (P#23).

Users prefer unbiased search results that are not influenced by profiled search history and driven by a filter bubble style selection of interests and recommendations. *"…I prefer searches focus on showing results for what I ask rather than what they think I want based on my history."* (P#50);

> It would be nice to see default pages that have not been affected by search history. It would also be nice to see advertising and content suggestions that are completely different from my usual browsing choices. Seeing the same things all the time is boring, it is good to see what other content is available and might trigger creative ideas. (P#83).

Popularity may be a deciding factor in choosing sources as evidenced by this user's experience in dealing with a Taylor Swift YouTube recommendation, which clearly is not within his or her's musical tastes. *"In subsequent visits to YouTube I have found recommended channels on the home page, which remain in every visit. However, my issue is that channels that I have rarely used, or cannot recall using at all (in this case it was TaylorSwiftVEVO) appear more than channels I use regularly. This stirs some concern in how my browsing data is being collected."* (P#254).

Others are unconcerned with release of data, but generally form their own judgements with received search results and do not overly rely on the relevance of results as stipulated by the information service provider. *"Personal searches like flights, eBay or sales purchase suggestions, etc. are fine for some personal info used to narrow search results. However, I generally ignore suggestions, searching fresh each time I search something new, personal or impersonal to me."* (P#226).

Recommendations are sometimes not considered as personalisation. Instead to this user, this is viewed as observation of search interactions online and this pervasiveness motivates the user's choice to not engage in social media. *"Often when I go online shopping there will ad banners from the websites I have visited of the items I have browsed. Personally, I don't consider this a personalisation of my browsing habit if anything I think it's invasive makes me wary of who's observing my online presence. For this reason, I do not have social media profiles."* (P#322).

Sometimes excessive profiling is at odds with the nature of how users interact and seek information. *"I don't like when ads are tailored to my previous search histories. my boyfriend searched for shower heads and then I received ads for shower heads throughout the week- I did not need this"* (P#218). Similarly, users also reported a few instances where the pervasiveness of data collection let planned surprises go astray. "*Searching for birthday presents for wife on Google on my phone - turned up as suggested searches when she used my phone"* (P#108).

Sometimes information users search are not about them or their personal interests, but an exercise in expanding their knowledge. *"After doing some general health research for my Uni studies, Ads related to these searches started appearing in my browser, especially on Facebook. These searches were to help my knowledge about health conditions and were not personal at all, which made me uncomfortable about how much search information is passed on to third parties."* (P#389). The presumption that all data released is essential to providing users with recommendations and extensively profiling user interests based on these data signals perhaps may not be the most effective strategy in terms of generating relevant recommendations for users. These expectations indicate there is scope to differentiate the nuances behind user intentions around data release.

Scenario 4 – Contextual Value: Location and Time

Relevance of personalised results (search results or advertisements) to users also depends on the context. Certain value is seen in providing locally relevant search results. "*I don't usually mind my info being used for travel information e.g. fastest way to get to X store bringing up an embedded Google maps instance"* (P#87).

As opposed to earlier instances where users found location information sensitive (see section 5.3.1), users find value in releasing their location when a specific tangible benefit is seen such as local travel map directions. *"I appreciate when my location information is used to provide relevant search results."* (P#406); *'Location services when searching for restaurants and addresses don't bother me." (P#390).*

Contextual value may also be temporally constrained. For this user, recommendations based on mining captured search history over a longitudinal timeframe are irrelevant. *"On one or two occasions ads have appeared on my computer for products that I had searched for previously, where the advertised products were more suitable than the ones found in my initial search. This happy result is rare as I normally purchase during my initial search session, so ads that appear later are just a nuisance."*

(P#391). Additionally, the primary search request may satisfy the user's browsing needs with subsequent recommendations deemed irrelevant. *"If I search for an item, ads for the same item appear on majority of pages. This does not meet my expectations as I have probably already done the research I required on the item, it also gives insight into my private life to others who see the ads or collect the data."* (P#49).

Often recommendations target items already bought, indicating ineffective personalisation. There is a disconnect in translating offline norms around consumer purchasing to the online world. In real-life once a consumer has bought a product, it is unlikely that the same purchase will occur again immediately.

> I had purchased a watch for a sibling's birthday and the next day advertisements for watches began appearing on my search home screen. This frustrated me as my privacy felt invaded. I did not wish to be bombarded with advertisements 'screaming' for my attention. I had clearly already purchased a product, I did not desire to purchase another in the immediate future (P#408).

These expectations indicate temporal constraints in determining relevance of search results to users.

Scenario 5 - Understanding the Value Exchange in Linking Data Sources

Some benefits in linking different data sources either across the same platform or different accounts. *"I found it helpful for use of my Pinterest account, my Facebook account (even though it is private)."* (P#150). Another user talks about the assurance provided in containing user interests within the primary platform s/he has engaged in. "*I have seen Google ads on pages that aren't Google about test kits used in biomedical science. I didn't mind this because I knew the data was maintained within Google, and hadn't been sent to a third party."* (P#305)

Users are sometimes aware of the impacts of linking aggregated data between consolidated services. *Searched for whiskey in Google and my Gmail account had related whiskey advertising."* (P#183). The impact of linking data across devices with being able to profile and identify a user across different machines can be seen. This goes beyond uniquely identifying each machine. *"When search suggestions in Google are the same between my different home computers"* (P#299); *"Strange that my search and my son's search for the same acronym had different results"* (P#440).

In this case, an unexpected use which combines different data sources is reported. *"One time when I search for the directions to Adelaide airport, my future flight details were marked on the map. I haven't*

published these anywhere, expect in my email attachments. They must have read my Gmail attachments about flight details and place in the map, which is clearly not expected." (P#1). Here, the expectation was that different interactions are kept separate. However, this instance shows the fluidity of data flowing across use of different products of the same parent company. It is a phenomenon known as a privacy lurch, when organisational data practices are modified without prior warning, creating potentially new privacy risks (Grimmelmann 2009; Ohm 2013). Technology is innovative, constantly evolving to bring in new functionality. Although it is indeed convenient, expectations are not met when users are not conscious of being informed of this new data use and exposure of data thought to be private. The norm here is for technology to innovate and push functionality out to the market, users then encounter the new use, and then become habituated to it. It is unclear if the lack of alternate choices shapes this technology push to form new habits.

Some users question why parties other than primary actors (for the purpose of the interaction) should have access to the data. .*"Facebook advertisements tailored to my search history - should not be linked in my opinion."* (P#3); "*When an entirely unrelated site (e.g. social media) presents advertisements strikingly similar to searches I have conducted within another application"* (P#103).

These sentiments echo the need to separate linking different interactions as users perceive each interaction to be visible to the information service they are interacting with. *"*

> Facebook and Google constantly show links to items I've viewed on fashion websites. I'm browsing a separate website, why should G&FB get access to my browsing history like this? (P#81);

"… This was disturbing to realise that what I'm searching for has been linked to something which I use for recreational use." (P#100); *"I have searched on Google while logged off from Facebook. Then I logged in to FB there were ads related to the Google search which means I've been tracked even when not logged in to other sites"* (P#104). The risk to users here is the increased disclosure of information deemed only visible to parties that are privy to the original interaction and the secondary use of data for purposes unintended by the user when releasing the data. As a result of these risks, information asymmetries persist in the value exchange.

Scenario 6 – Perceiving the Impact of Aggregated Data: Ambivalent, Misinformed and Unconcerned Views

An array of different perceptions regarding the impact of aggregated data was seen. Most users are aware of the effects of aggregated data through personalised recommendations (ads or search results).*"when stores have my recent history and try to personalise items around me."* (P#243).

Other users express ambivalent views about data collection, stating they are aware and unconcerned about its impact. *"Getting personalised ads based on search history, however this doesn't really bother me"* (P#20); *"I suppose I see ads on social media of things I've previously searched for. I don't mind that."* (P#195).

A few express ignoring ads as a strategy in dealing with the pervasiveness of data collection. This user thinks that there is no privacy online, treating it as a public medium.

> I feel it's ridiculous to assume privacy. I don't care if Google tries to aim ads at me, I am more than happy to ignore them. My favourite ads lately have been advertising for the hotel I already have booked. (P#19)

> Not really. I've just come to expect that personal data once online can't really be taken off (P#222).

Similarly, another user expresses being unconcerned and does not value putting in the effort to prevent tracking. "I don't mind sites using my cache and cookies to get information about me, if I wanted to hide it, I would..." (P#70); "I expect that I am in control of the data. E.g. I can wipe it if I chose to" (P#5). However, the increasing difficulty in preventing release of data perhaps may indicate naivety or being misinformed on part of the user.

Others are more visibly misinformed underscoring the need for effective user awareness initiatives on data release. *"I assume it's all anonymous and not tracked and stored. I have noticed ads pop up though after I search for specific items. This isn't horrible but still quite creepy."* (P#124);

> as long as it doesn't reveal my identity, I am okay with others tracking it. I feel as though, I'm 1 in millions of internet users, it doesn't concern me if they have my information (P#291).

This last sentiment echoes a common misconception where the user is mistakenly reassured given the large amount of people who use the internet, it is relatively safe to do so and be less identifiable in that

large pool of users. However, the ease in which you can be identified and have personalised ads or search results targeting your interests underscores the difficulty that users have in understanding the computational capabilities and the extent to which data aggregation, linkability and personalisation can occur.

These perceived views highlight a need to increase awareness and understanding around tracking and the importance of educative goals in empowering privacy-sensitive behaviour online, so that users are able to take informed decisions.

Scenario 7 – Tensions in Information Asymmetries: Passive Acceptance of Affordances

Information asymmetries occur when the exchange between the user and the information provider is imbalanced, due to constraints posed on the abilities of an agent (user or technology) to fully access or understand the effects of information ((Nissenbaum 2010). In online information societies, asymmetries often take the form of information and power where users are disadvantaged against information service providers.

Tensions in passively accepting the release, collection and dissemination of data start with a lack of awareness. Being aware of how data is collected and linked across multiple devices is a first step in reducing information and power asymmetries that persist. *"Until very recently I didn't realise how much of my info and search history was collected online..."* (P#214); *"I'm really unsure how companies use the data and how much companies or tracking devices understand who I am and what I use, somewhat scary."* (P#436).

Similarly, some users articulated a lack of understanding in comprehending how data is collected, processed and disseminated. *"The only evidence I have seen is advertisements related to previous searches"* (P#338); *"I often find advertisements for products pop up when I browse public sites e.g. weather sites. Unless my information is being sold I don't understand the Web well enough to guess why this happens"* (P#294); *"I was surprised that websites "know" my location. For example, when I access a store website, the website is able to tell which of their stores is closest to where I am."* (P#163).

This lack of awareness then propagates into discomfort as users realise the far-reaching impacts of data collection once made aware. *"I'm naively surprised when ads pop up [on] omg my Facebook pages that relate to businesses I may have searched through Google e.g. online shops I've visited. I find it very*

annoying - when I think about it." (P#233); *"There is obvious tracking when you log in... Very underhanded!"* (P#169);

> at first, when searching, and the internet gave preferences I might like I realised I was being tracked and didn't like that but didn't know there was anything I could do about it. I don't want to give free info to a data collection agency that they will then benefit from. (P#101).

In this last data flow example the user implies the ease of access and visibility of data to omnibus information providers not authorised by the user.

Other manifestations of these tensions take the form of passive acceptance, where users are aware and understand the impacts of data collection, but are helpless in doing anything constructive about it according to their expectations. Some of the views expressed indicate making do with what is available. *"I accept gathering information is their currency..."* (P#264); *"…since there's going to be an ad there anyway I would rather it interest me"* (P#112); *"My expectations of online privacy are very low. Sometime YouTube will suggest something interesting based on what it thinks I like."* (P#361); "*I expect them to record all of my data but I am not happy about it."* (P#13); *"Even though it is all recorded; I sincerely wish it wasn't."* (P#16)*; "I would prefer that personal information is not used, however I am very aware that it is being used. I suppose we get used to it, it's how we live now."* (P#187). As users continue to passively accept the choices offered to them, they become habituated over time to this lack of real choice*. "I expect it to continue to creep me out, but I also expect to become accustomed to it over time."* (P#280).

Others express this as a trade-off that they do not exactly like, but to access a particular information service, they are aware they need to play by the rules the organisation sets.

> I expect and somewhat accept that my personal data will be stored online, in this day and age. However, as a normal person I feel that there is not much I can do about it. At the moment, it doesn't affect my daily life. I don't like the thought of cookies being stored and as a consequence, flight / purchase prices increasing because I am being 'watched' online. Having to bear the cost, whether financial or not, is my major concern." (P#86)

Others express value for the conveniences embedded (which can be implemented securely at the client-end), but also articulate the excessive collection and monetisation of data. *"Seeing advertisements of things I have previously googled like books or electronics makes me feel uncomfortable; I don't like it*

that my search history has been saved and used in order to make money or 'draw me in' to click on an advertisement. I would prefer if my search history was completely eradicated. Things such as auto-fill forms like address or postcode are acceptable, and actually quite helpful to avoid re-writing." (P#431).

Sometimes the popular 'I've got nothing to hide' argument is used in public discourse to justify that there is no perceived harm in data collection. For example, attorney-general George Brandis could not clearly articulate the nuances of data collection and its effects with respect to the Australian metadata retention laws (Dingle 2014). This notion mistakenly underscores the range and extent of privacy protection (Solove 2007). It oversimplifies the risk by assuming that surveillance is harmless when no illegal actions have taken place.

Given the longevity and permanence of data, it is plausible that collection and aggregation of data may have unforeseen future impacts and risks. These users report instances acknowledging that data collection has become normal and expresses disinterest or helplessness in data collection as it does not. *"My use of the internet is very generalised, study, shopping, research for health info, news, or items of interest. I don't mind anyone looking at it really. Facebook already grabs as much data about your interests and put up advertising to match. I don't think it's something you can stop. And I am not engaging in anything illegal on line - so I probably don't care."* (P#238)

> I do expect that my usage patterns are being tracked and recorded. I know search engines are using it to make my experience better. I am not happy that undisclosed parties may be tracking what I search and read but I have nothing to be ashamed of really (P#44).

However, this conflation between illegal and privacy underscores the tensions in the existing information asymmetries as 'illegal' is not the same thing as private or personal data.

Others take a stand against passively accepting available options. This user reports on proactive actions taken to misinform profiling.

> I don't mind if ads are targeted. But I have a habit about 'misinforming' Google. Thus I receive a healthy mix of 'you should take these pills if you want a baby' and 'hey look at these power tools, they are rad' advertising. Both neither worries me nor causes me issues. What I hate is the location service which now puts me into my city. I don't see how my location is relevant if I want to know the nutritional value of a certain type of food. I can't get my phone to not include this now in the bottom of the Web browser. (P#133).

This effect is common among pragmatic users to actively misinform providers (Fuster 2010). Obfuscation is a further step that extends this line of thinking, a form of protest perhaps against information and power asymmetries (Brunton & Nissenbaum 2015). Hence, there is a clear need to better manage release of data such as location considered personal and its relevance tied to the purpose of search request.

It is also hard for users to keep up with continuous technology changes in privacy affordances such as default settings that are set to maximise exposure or complicated privacy settings in profiles. "*I believe if used transparently and made understandable to all parties, then the use of personal data should be allowed. This is only if the party providing the data can determine what information is shared.*" (P#116). Often the information service providers themselves evolve their privacy policies and settings to address changing user beliefs. As privacy-by-design is not practised holistically in the development of these products, these *ad hoc* pushed-out changes inevitably result in a burden on the user to first understand the changes and then figure out how to action these changes in-line with their expectations.

These weaknesses in addressing privacy expectations result in low levels of trust of information service providers among users "*No respect for privacy by internet providers*" (P#160); "*I don't want my personal information to be on-sold to other companies and I'd rather not have my name spread around as this I feel is a major breach of my privacy and if they are not upfront about it, why aren't they being upfront about it?"* (P#213). This last thought underscores an important point in building trust by information service providers where being honest about the extent to which data collection occurs could be positively perceived by users, as opposed to current practices such as default settings that are set to expose and other limiting affordances that users encounter.

Other users perform interactions fully aware of the information and power asymmetries at play that contribute to a lack of trust.

> I expect Google and other advertisers will build a comprehensive profile on me. I expect them to sell that data. I expect most advertisers to have terrible data security and they will hide data breaches. I think there need to be strong privacy laws put in place both by our government and the USA. (P#361)

> I have selected the option for Google not to collect any data about me (including search history), but behind the scenes I know they probably still collect data, just without telling me. So transparency about their data collection probably won't change anything, as they can easily lie about it. (P#158).

This last sentiment highlights the importance in being able to independently verify initiatives and actions that aim to build transparency, because as this user points out, commercial incentives cause a conflict of interest in companies being able to do this independently.

Further, evidence points to existing mechanisms for expression of privacy expectations as being ineffective or inflexible. This user reports on not liking the expected norm of having to register to use services**.** *"I think there should be options as to whether your searches should be recorded and used for ads…"* (P#268); *"I would prefer if they didn't ask as many personal questions, I don't like putting out my name and my address together etc."* (P#127). Although accountability is important in online interactions, perhaps there is a need to rethink how user identities are represented and managed online.

5.3.5 Deleting Data

There were 5 scenarios that described privacy expectations in the act of deleting data. All 5 of them had a negative connation around either inappropriate use of data or ineffective affordances in expressing privacy needs.

Scenario 1 – Erasing Data is Difficult

The ineffectiveness of options to manage or control saved client-side data is a persistent issue for users. Of particular concern is the difficulty in verifying if data is actually deleted once actioned through client-side options. *"Regardless of deleting browser history/cookies, I realised Google had a saved search feature and webpage that was difficult to navigate. To remove searches, the user had to clear each search one by one."* (P#14); *"you should have the option to erase as and when you'd like to"* (P#66). Here the user intent was to manage and control saved search history that was visible to the user. The user displays awareness of tracking not just within browsers or sessions, but of the linkability of data through shared online services (like Google) and seeks to remove data stored through account log-in.

In some instances, depending on the purpose of the interaction, users expect a clean and fresh session that is not linked to prior sessions. *"I was buying flights online, and the prices kept increasing. Even though I was clearing my full search history/cookies/data, the flights were still at a high price. So, I felt that even though I made an effort to clear this data, it was not actually cleared. When I used another computer, the flight prices were significantly cheaper. However, if I searched on incognito mode, the flights would not load on Skyscanner, until I enabled cookies."* (P#86). The user reports trying existing

affordances such as private browsing (where no data is stored client-side), breaking functionality when cookies are blocked and removing session data to avoid being linked to earlier transactions. From a user's point of view, transactional queries often are preceded by queries that seek out and uncover different options. The information service provider's main goal is to maximise profit for each sale. The expectations of the two parties are contradicting, resulting in the user having to improvise around available options to address the user's information-seeking goal.

Collection and dissemination of search session data once made visible (through advertisements), this user reports on action taken to delete search history stored at the client-end and change advertising preferences. *"Having searched for a financial product I was disturbed to find advertising for these services on most website I looked at. This prompted me to immediately change my advertising preferences with Google, as well as delete all histories."* (P#5). However, this marginally solves the problem of managing the dissemination of information to third-parties as this capture happens through the interaction, but may not be visible. The current affordances (deleting account logged-in search history) provides the user with some control over managing future search results, but is reactive and does not affect the instantaneous tracking that has already occurred.

Scenario 2 – To Use or Not To Use: Framing Choices through Technology Lock-in

Another commonly encountered scenario is technology lock-in through an integrated platform (Naughton 2012). The cost of not participating in an online service may be more than the cost of ineffective affordances to manage the release of data (Nissenbaum 2011). *"After quitting Facebook, my personal email was utterly inundated with spam. One could assume that they sold my details after I closed my account."* (P#16).

The ineffectiveness of these affordances leads to users having to choose between engaging and not engaging. *"If you don't want to share it, I guess advice is not to use the website. Expectations using websites that it is transparent for all to know. However it ends up just being data and nothing personal."*(P#122). If they choose to engage then, the expectation here is that the data is visible to all. However, data taken out of context can loose meaning and certain data particularly when linked can be personally-identifiable.

In this case, the user was not expecting to contend with spam after choosing to opt-out, which is a negative consequence through potentially secondary use of identifiable data. The privacy risk to the user is inconvenience which adds to the difficulty in choosing to opt +out of a service. Another question that

arises is what happens to the data when the user chooses to opt-out as users often are not given the chance to verify if data has been deleted.

In another instance, a user wanted to delete an account that had no purpose. *"I created a Skype account because it was recommend[ed], then when I didn't want it any more I discovered that I could not delete the account."* (P#39). The user reports not being able to delete the account. It is plausible that user knowledge may be a factor at play here regarding awareness and understanding of appropriate mechanisms to close an account. In applications such as Skype, data on the actual content of the calls are not recorded. Information privacy in this case deals with being able to delete data about interactions such as saved conversation histories. Users may also choose to de-identify themselves by creating accounts with fake names.

5.4 Empirical Contributions: Explicating Data Flows and Privacy Expectations using the Contextual Integrity Framework

The major contributions of this empirical work are described in this section. These empirical contributions consist of an analysis explicating data flows by applying the contextual integrity decision heuristic to investigate if appropriate flows are upheld. The key arguments stemming from this analysis is summarised to articulate why addressing online privacy expectations is a design challenge and sets the scene for the conceptual phase of the study presented in Chapter 6.

To identify if these privacy expectations are violations and derive the underlying norms guide this deletion data flow, the contextual integrity decision heuristic is applied. The overarching principle of the contextual integrity framework is that privacy is a "right to *appropriate* flow of personal information" (Nissenbaum 2010 pp. 127). Central to this framework is the notion of norms which "prescribe the flow of personal information in a given context, are a function of the types of information in question; the respective roles of the subject, the sender (who may be the subject), and the recipient of this information, and the principles under which the information is sent or transmitted from the sender to the recipient" (Nissenbaum 2010 p. 127).

Contexts are defined as "structured social settings characterised by canonical activities, roles, relationships, power structures, norms (or rules), and internal values (goals, ends, purposes) (Nissenbaum 2010 pp. 132). Contexts are constrained to "framing a response to socio-technical systems and practices that have radically altered information flows" (Nissenbaum 2010 pp.132). Constructs are

"intended as abstract representations of social structures experienced in daily life" (Nissenbaum 2010 pp. 134). Roles are "typical or paradigmatic capacities in which people act in contexts (Nissenbaum 2010 pp.133). Activities are "canonical activities and practices in which people, in roles, engage" (Nissenbaum 2010 pp. 133). Norms are "behaviour-guiding norms prescribe and proscribe acceptable actions and practices" (Nissenbaum 2010 pp. 133). Typically norms of appropriateness and distribution exist. Norms of appropriateness describe the kinds of data about a person that is appropriate to be released in a given context, while norms of distribution mandate the data flows between two actors (Sar and Al-Saggaf 2014). Values are "goals, purposes, or ends; that is, the objectives around which a context is oriented" (Nissenbaum 2010 pp.134).

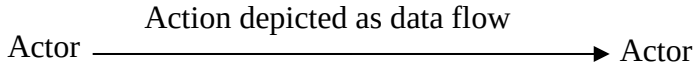
The decision heuristic is used to evaluate if the practices derived from the privacy expectations expressed address appropriate flows of information and uphold contextual integrity. Three such practices were identified from this in-depth qualitative analysis, namely *collecting and retaining metadata*, *collecting and distributing data for commercial purposes* and *framing choices through ineffective affordances*. These practices which are viewed as a frame for analysis of privacy expectations as they afford users privacy, are then analysed in relation to the contextual integrity framework's decision heuristic to ascertain if data flows are appropriate. This application of the contextual integrity framework's decision heuristic consists of 9 steps which are followed for each practiced identified. This application consists of 9 steps as defined in the contextual integrity framework:

1. Identify new practices in terms of information flows
2. Describe the Prevailing Contexts and Impacts for these data flows
3. Describe the Information subjects, senders and recipients for these data flows
4. Describe the Transmission Principles under which these data flows occur
5. Locate any entrenched norms or identify significant points of departure from existing practice
6. Prima facie assessment: what are the ways a system or practice defies entrenched norms?
7. Evaluation 1: Consider moral and political factors affected by the practice in question
8. Evaluation 2: Ask how the system or practices directly impinge on values, goals, and ends of the context. Consider the meaning or significance of moral and political factors
9. In conclusion, recommend in favour or against systems or practices on the basis of findings

Note that the recommendation provided in step 9 is an evaluation if the information practice described upholds stated privacy expectations. These recommendations are not overarching recommendations for the data flows/ information practices discussed (see Chapter 6 for high-level design recommendations

around data release in search), but recommendations presented in light of its evaluation when the contextual integrity framework is applied.

There is no standard notation used to illustrate data flows under the contextual integrity framework. Instead, for the purposes of illustrating these data flows in this thesis, actions taken by actors are denoted as data flows. For example:



5.4.1 Practice 1: Collecting and Retaining Metadata

Before applying the contextual integrity decision heuristic to determine if privacy expectations are violated or not addressed, the key practices described in the above scenarios are two-fold:

i. Metadata retention practice by Internet service providers introduced through law by government
ii. Data release by a user when interacting with an information service provider

A majority of the expectations address the second practice, which occurs in a commercial context as opposed to the metadata retention practice, which is considered as surveillance. The multiple risks arising from the expectations of users around data flows in a commercial context include lack of awareness which could lead to misinformed or inauthentic actions with the user not being aware or having the capability to mitigate the risks addressed.

In addition to these risks, some users accept that data release involves capture of data by primary providers, though mention of omnibus information providers (or third-party data aggregators) is lacking. They then choose to accept the practice of data capture and are unconcerned by the continuation of these practices.

1. *New practices in terms of information flows:*

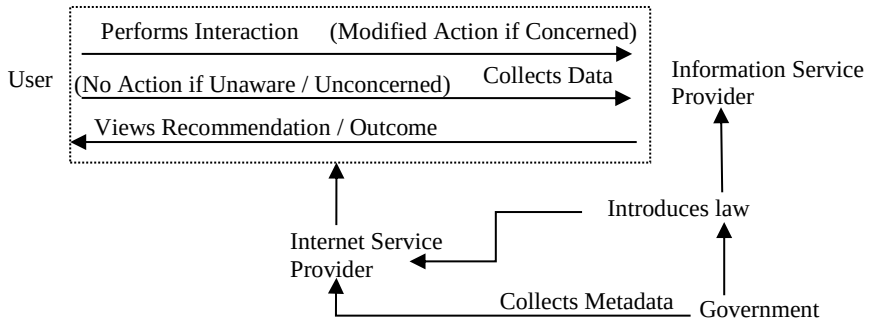


Figure 5.29 Practice 1: Metadata Retention Practice

Awareness about surveillance risks is seen with an introduction of a law by the Australian government that mandates collection of meta-data of user interactions from internet service providers.

2. *Prevailing Contexts and Impacts:*

Data collection occurs already on a commercial basis when users interact with online service providers. To this prevailing context, the practice of retention of metadata by internet service providers is introduced. The main goal of the law is to help law enforcement uphold the safety of communities by being able to access telecommunications data sources (either phone or internet-based communications) as evidence and to identify suspects in crimes (planned or committed).

3. *Information subjects, senders and recipients:*

Online interactions between a user (subject, sender) and an information service provider (recipient) are now accessible to government agencies through internet service providers (secondary recipient as their owners of the infrastructure through which data flows) in addition to existing third parties such as omnibus information providers that harness the data for behavioural advertising purposes in the commercial context. These third parties are largely invisible to most users.

4. *Transmission Principles:*

From a user's point of view these online interactions take place under assurances that data is only visible between the sender and recipient whilst acknowledging that data flows through the medium (Internet), where access to it is owned by internet service providers.

5. *Locate entrenched norms, identify significant points of departure:*

Metadata is data about an online interaction such as identifiers about the subscriber using the service (information subject) including names, addresses and unique identification numbers attached to devices, source, destination, date, time and duration of a communication, type of communication and service and location of equipment ((Attorney-General's Department of the Australian Government 2015) and does not contain information about the substance of the communication like a person's browsing history, visited URLs and IP addresses). The established norm here is that interactions are visible only to parties privy to the interaction (sender, recipient, secondary recipient). The significant point of departure is that the new practice (data retention law) mandates the collection of information about the interaction which may retrospectively be used in evaluating the sender's actions in terms of crime and safety.

6. *Prima facie assessment: what are the ways a system or practice defies entrenched norms?:*

Because the law mandates retention of metadata for 2 years in its aim to reduce crime, users are aware that their actions are being monitored. This might cause self-censoring of actions deemed suspicious as it pertains to the law. Although only metadata is collected and stored, it is possible to infer details about activities the information subject (user) may have partaken in. It is relatively easy to aggregate and link different data sets to uncover patterns and make inferences about individuals, given the vast quantities of data sources and advances in computational abilities. Of particular concern to users is the implementation of oversight and accountability mechanisms that guide government agencies' ability to exercise these capabilities. For example in 2015 the Office of the Australian Information Commissioner was allocated reduced funding and limited staff by the then Australian Government, which limited its ability to achieve its mandates as an independent watchdog (Farrell 2015; Office of the Australian Information Commissioner - Australian Government 2016).

Further, the approving authority for most warrants is internal to the government agency, which, while increasing the ease in which these warrants can be obtained, also introduces questions of the independent verification or judicial oversight of these warrants prior to approval ((Australian Internet Governance Forum 2015). Also data collection for 2 years is longer time period than other metadata retention regimes in European countries (Electronic Frontier Foundation 2016; Lawrence & O'Donnell 2015). Any substantial data collection such as this may be an attractive target for attackers and security measures must be implemented by telecommunications companies to safeguard the data collected.

7. *Evaluation 1: Consider moral and political factors affected by the practice in question:*

Users may have to adjust to the notion that the internet is a public medium and any loss of personal freedom is offset by the public good in reducing crime and increasing safety of communities. Increasingly, government authorities are relying on metadata retention regimes to combat cybercrime and crime in general.

8. *Evaluation 2: Ask how the system or practices directly impinge on values, goals, and ends of the context. Consider the meaning or significance of moral and political factors:*

There is a loss of autonomy and personal freedom experienced by the user because information about their interactions is now recorded. It increases the reach and capabilities of power structures like governments to monitor activities of residents in their jurisdiction. The risks to the user are the lack of

independent verification of these practices and being obliged to passively accept the collection and profiling of potentially identifiable information.

9. Recommend in favour or against systems or practices on the basis of findings:

In its current state, the practice of collecting and retaining meta data violates contextual integrity and is identified as an information practice where privacy expectations are not met. Given the pervasive nature of online surveillance and the ease in which computational technologies facilitate this collection, data retention regimes will continue to be used by law enforcement agencies to combat crime and mitigate national safety risks. Users that expressed willingness to release data for surveillance or other government-related activities implied a level of trust with the government. However, if this practice was conducted in a transparent mater it would reassure information subjects that due diligence and accountability measures underpin these activities such as when releasing data when searching. Given, the lack of visibility of data flows from the findings of the technical controlled experiments (in Chapter 4) and the complexity in understanding the nature of these data flows from the findings of the empirical survey study (Chapter 5), providing these oversight measures will help strengthen trust in established power structures.

5.4.2 Practice 2: Collecting and Distributing Data for Commercial Purposes

Given the wide range of privacy expectations received and scenarios derived, this information practice will be evaluated in two stages, namely an evaluation of the collection, followed by an evaluation of the distribution of data.

Practice 2a: Collecting Data for Commercial Purposes

1. New practices in terms of information flows:

Awareness about commercial risks is more prevalent among users compared to awareness and understanding of surveillance risks. These practices are now considered normal, but when technology companies change or tweak these practices, users may encounter unfamiliar situations.

This existing practice explores data collection, processing, dissemination combined with data release. Although this is an existing practice, it is enhanced by considering not just personal data, but data released about the user. This includes behavioural data such as data the medium (i.e. the browser, Internet) implicitly releases in addition to what users explicitly release through an interaction such as search.

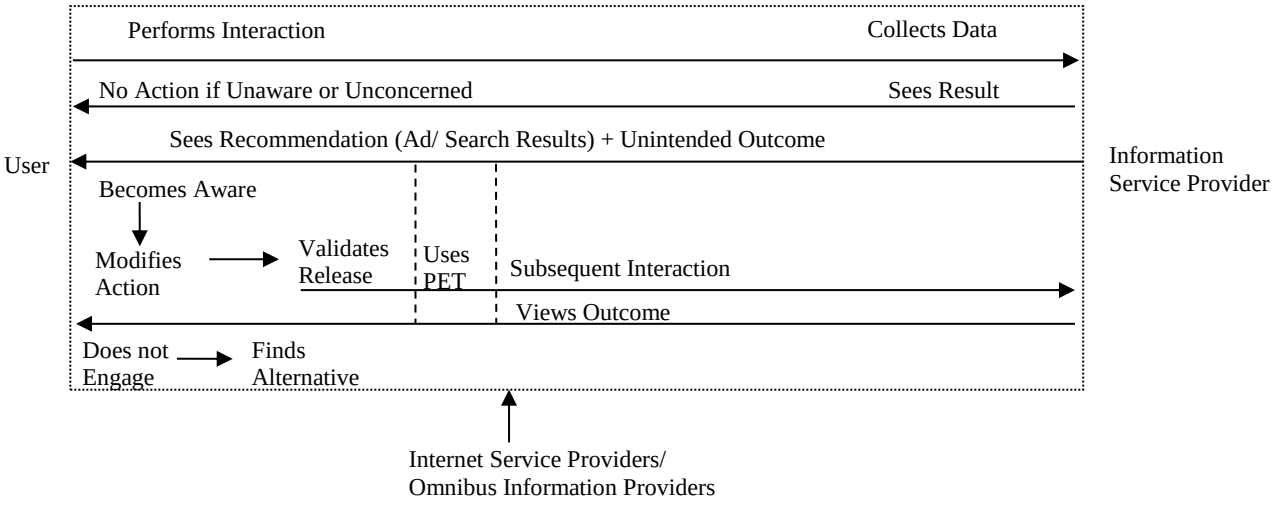


Figure 5.30 Practice 2a - Appropriateness of Data Flows

2. Prevailing Contexts and Impacts:

Even in a largely commercial context, user interactions may take place with users working the interaction with their own presumptions of how data flows as it may be unfamiliar with them. Those users who are informed, may take a conscious decision to accept the collection that occurs because the benefit perceived is of use to them. Other informed users may choose to take action to minimise release of data or choose not to engage in some extreme cases.

3. Information subjects, senders and recipients:

Online interactions between a user (subject, sender) and an information service provider (recipient) are accessible to existing third parties such as omnibus information providers that harness the data for behavioural advertising purposes in the commercial context. The effects of these third-party actions are largely invisible to users because although 57.6% of users said they were aware about tracking, in subsequent questions that tested the validation of this intent via specific examples about tracking through behavioural advertising effects, most users expressed dissatisfaction or indifference to these activities. For example, 60.77% of users where unaware of personal data use in subsequent searches and a combined 48.7% and 39% were unsatisfied about capture of clickthrough data and linkability of search queries (with a further 39.2% and 36.5% expressing indifference) (see Tables 5.2 and 5.3) .

4. Transmission Principles:

From a user's point of view these online interactions take place under assurances that data is only visible between the sender and recipient whilst acknowledging that data flows through the medium (Internet), where access to it is owned by internet service providers.

5. *Locate entrenched norms, identify significant points of departure:*

In a commercial interaction context, users are surprised when personal data is used for purposes other than they intended, for example, aggregation visible through mechanisms such as personalised ads or search results. To address their expectations, they modify their actions by seeking alternate mechanisms of managing data release either through use of PETs or by minimising data release.

Misinformation among users about data release and leakage to third-parties is a cause for real concern. For example, PrivateBrowsing hides data from being stored and linked at the user (client) end. This is insufficient as data released to the primary party the user is interacting with also enables visibility and accessibility of this said data to third parties that capture it for behavioural advertising purposes.

6. Prima facie assessment: what are the ways a system or practice defies entrenched norms?:

In a commercial interaction context, the practice of enabling data capture by omnibus information providers negatively impacts user experience for individuals. For those that want to be able to manage their data, the existing practice does not allow this and they are coerced to either comply by the practices and rules of the information service provider or seek alternative mechanisms to manage data release if they so choose to.

7. Evaluation 1: Consider moral and political factors affected by the practice in question:

Morally, the pervasiveness of data collection plays on users not being aware or able to understand the extent to which this data capture occurs. This difficulty is compounded by the fact that users need to be technically literate and to know of practical ways to mitigate this risk online. Often this technical literacy takes years of specialist learning and practice.

8. *Evaluation 2: Ask how the system or practices directly impinge on values, goals, and ends of the context. Consider the meaning or significance of moral and political factors:*

In commercial contexts where users are unaware, their trust is not reciprocated because the monetary value of the personal data produced through the interaction often outweighs a company's desires to proactively inform users about the extent to which data collection takes place.

9. *Recommend in favour or against systems or practices on the basis of findings:*

In its current state, the practice of collecting data for commercial purposes violates contextual integrity and is identified as an information practice where privacy expectations are not met. To address the awareness issues in online interactions, these practices need to build trust by being open and upfront about the extent to which data collection takes place, and by providing users with a mechanism to effectively manage data release, as choosing to simply opt-out is insufficient for meeting consumer expectations. Given, the lack of visibility of data flows from the findings of the technical controlled experiments (in Chapter 4) and the complexity in understanding the nature of these data flows from the findings of the empirical survey study (Chapter 5), it is imperative to communicate and raise awareness of the depth and breadth of commercial data collection in a manner that everyday users understand when they consent to releasing data when searching.

Practice 2b: Distributing Data Released for Commercial Purposes

1. *New practices in terms of information flows:*

This practice involves the distribution effects of data. It takes into consideration expectations listed under the processing data flow including those involving unexpected and non-targeted processing and dissemination of data. These data flows consist of risks arising from unexpected uses in processing data, the capability to process data and evaluate interactions according to institutional guidelines and the ability to process data beyond the context of the original interaction and perceived user intent. These information flows are represented below.

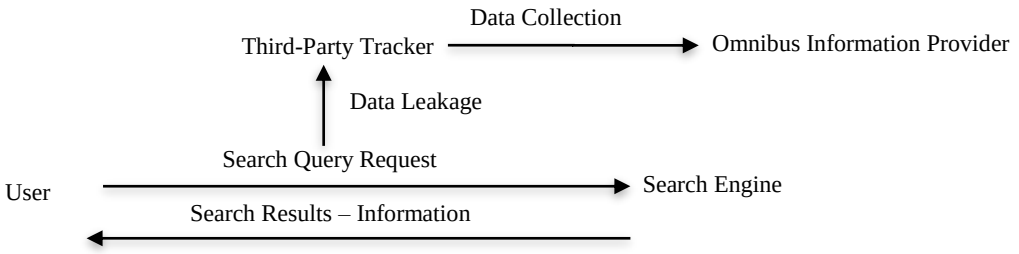


Figure 5.31 Practice 2b: Distribution Effects of Data Flows (Normal Scenarios)

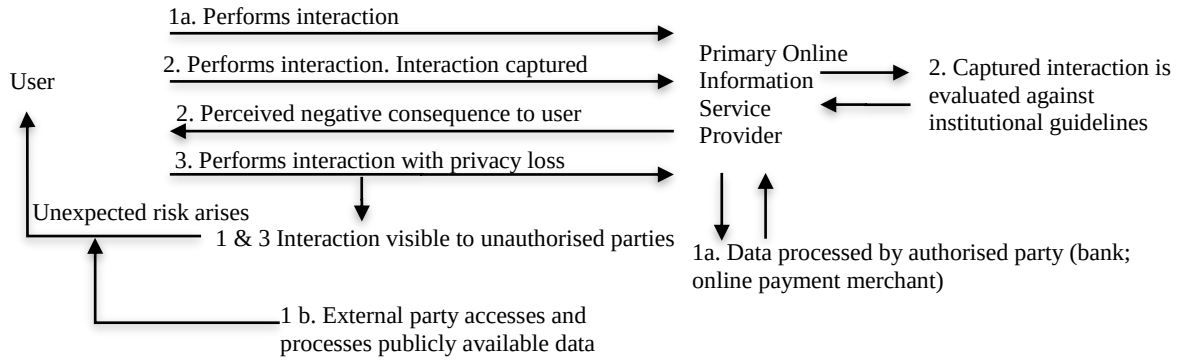


Figure 5.32 Practice 2b: Distribution Effects of Data Flows (Exceptions)

2. Prevailing Contexts and Impacts:

The first diagram describes an example of a typical practice of data release in an information-seeking context, where the user (subject, sender) intent is to search available information resources on the Web via a search engine (recipient), so that user can make an informed decision on which health insurance to purchase.

The described exceptions to this general practice involve information-seeking contexts with a primary information service provider. The user intent is transactional in scenario 1a with the purchase of a product or service and confirmation of the purchase being the main end goals whereas the user intent is informational in scenarios 2 and 3 to seek relevant information. Scenario 1b describes processing of spam, where external parties access and process publicly-available data.

3. Information subjects, senders and recipients:

The user expected to interact with only the search engine. The user didn't expect to be hassled by external parties such as telemarketers or unwanted ads. The user may have been unaware of how the

information is captured by or accessible to third parties. Third parties may not have been visible to the user.

Scenario 1a illustrates a transaction between the user and a retailer of online products or services with data processing occurring among authorised parties such as banks or online payment merchants. Scenario 1b is a hidden interaction between a user and an invisible actor whose intent is to process publicly available data (email addresses) usually with malicious intent (perhaps social engineer users into leaking credentials and other personal information). Scenario 2 depicts the use of a search engine by the user. In all cases, the information subject is the user.

4. Transmission Principles:

It is a moral violation in terms of the user's expectation and the subsequent inconvenience to the user because of this interaction. For third parties, it has become expected practice to be able to collect search data. For search engines, it is accepted practice to collect data to optimise the effectiveness of ads relevant to users and to advertisers. Privacy expectations of users are not violated when ads are perceived as having some value to the user. This indicates the contextual nature of user acceptance of data leakage, where the boundaries between acceptable and unacceptable practices are influenced by user intent.

In scenario 1, the transmission principle underpinning both types of data flows includes security (authorised access) and confidentiality (visible to only authorised parties). In scenarios 2 and 3, the user expects the data flows to be confidential i.e. constrained to information-seeking purposes and visible only to primary actor (search engine) in this information-seeking context.

5. Locate entrenched norms, identify significant points of departure:

Regulatory factors affect online behavioural advertising and actors (search engines, advertisers, third-party trackers, omnibus information providers) have to comply with this. However, data protection lacks provider user protection as it largely compliance-driven. Technological affordances that help the user include ad-blockers, spoofing data whilst searching, but this puts the onus on the user in terms of effort (installing the ad-blocker) and knowledge (understanding technical how their data flows and make an informed decision).

In scenario 1, the entrenched norm is that transactions should be secure regardless of whether users initiated the interaction. When unauthorised parties gain access to data or misuse publicly-available data

for undesirable purposes, the security of the interaction is weakened, placing the data at risk. In scenarios 2 and 3 the norm for the user is to seek information freely without interactions being excessively captured and viewed out of context.

6. Prima facie assessment: what are the ways a system or practice defies entrenched norms:

Scenario 1a is a clear prima facie violation as the security of transactions is paramount not just to users, but to providers of these online services as it is key to building trust and keeping businesses operational. It involves instances of privacy expectations not being met when personal details were visibly seen to be sold to or accessible by unauthorised third parties. Scenario 1b is also a clear violation as publicly-available data is used for purposes beyond the original intention (of providing a mechanism for communication) with the risk of inconveniencing the user.

In scenarios 2 and 3, a disconnect between norms perceived by the user (interactions constrained to primary actor and purpose) and norms practiced by the recipient exists. In scenario 2 an enforcement of institutional guidelines by a school to moderate Web interactions taking placing within their jurisdiction is seen. In scenario 3 the practice of commercial actors providing information-seeking services is to provide users with convenience whilst capturing data with the presumption of providing users with relevant search results gleaned through processing of this captured data.

7. Evaluation 1: Consider moral and political factors affected by the practice in question:

In scenario 1a, the risk caused to the user is financial and results in a weakening of trust with primary and secondary authorised providers. In scenario 1b, the risk to the user is inconvenience at the very least and anxiety about personal data if sophisticated social engineering attacks are employed. In both cases organisations invest significant resources in providing safe and secure contexts for interactions.

In scenario 2, the user's autonomy and freedom is impinged upon by the organisational practice to moderate Web interactions. The user expects to be able to freely seek knowledge without dire consequences. In scenario 3 users expect convenience, but the embedded defaults are found to be limiting in this case, as search history was visible to an unexpected user.

8. *Evaluation 2: Ask how the system or practices directly impinge on values, goals, and ends of the context. Consider the meaning or significance of moral and political factors:*

In scenario 1, organisations usually communicate best practices for users to abide when conducting these transactions and have contingencies in place to minimise the risk to the user. Both the organisation and the user work towards the goal of securing transactions as it is a common interest. In scenario 2, the nature of the medium given its capabilities to immediate capture and aggregate data, makes information about the interaction mobile. Then tension between the user's perceived beliefs of freedom and autonomy clash with institutional guidelines on best practices for Web interactions.

A caveat to this claim may be that these tensions may depend on the exact nature of these institutional guidelines. However, increasingly institutions (both commercial and non-commercial) rely on core services and associated products and services for assurances in terms of business continuity and security. For example, in an educational institution, students are expected to search for information as part of their learning activities. Most institutional Web environments are monitored for the purposes of online safety. Further teaching and learning is enhanced by improvements in state-of-the-art technology. In cases of monopolistic power such as Google (the leading search provider), tensions may arise when institutions that already monitor Web environments for security purposes, further collaborate to obtain business benefits such as via Google Analytics and Google classroom. In such an instance, the educational environment which traditionally is constrained by institutional guidelines, is liable to being influenced by guidelines and practices associated with its collaborators. These tensions risk the nature of learning environments and its associated privacy expectations where increasingly learning goals (such as quality of learning) may be influenced by pseudo learning goals (such as optimising performance and learning outcomes), where the latter may be influenced by commercial profit-making motives.

In scenario 3, the goal of the commercial search engine provider is to ensure continuity between interactions and hence users have come to appreciate the convenience. However, this capability in linking disparate interactions also makes it hard for users to keep interactions private if different interactions are shared across the same device. Existing affordances such as use of incognito browsing and not saving session data are only available to the user if they are aware of them.

9. *Recommend in favour or against systems or practices on the basis of findings:*

In its current state, the practice of distributing data for commercial purposes violates contextual integrity and is identified as an information practice where privacy expectations are not met.

Scenario 1 is a known problem and the primary actors involved are driving efforts to inform users of best practices and create stronger platforms for secure and safe transactions.

Scenario 2 is an increasingly common problem as users like digital natives expect to use their preferred medium (the Web) for information-seeking goals without fearing being observed or judged. However, the ease in which information about interactions can be recorded and evaluated limits the space in which this unobserved exploration can freely take place. These tensions perhaps will be normalised as the Web, a relatively young social institution, matures.

Scenario 3 is a classic case in the ineffectiveness of affordances to manage data release. Existing capabilities are not reachable or usable to users because they may lack awareness and understanding of the ease in which data is linked across interactions. This requires more user focused education or training initiatives on the extent and scope of data release and measures to control data flows particularly across different interactions.

Given, the lack of visibility of data flows from the findings of the technical controlled experiments (in Chapter 4) and the complexity in understanding the nature of these data flows from the findings of the empirical survey study (Chapter 5), the extent and scale of the distribution effects of data should be conveyed to users before data is released when searching, so that these dataflows occur as as interactions based on trust instead of based on lack of understanding or awareness, misinformation or disinformation.

5.4.3 Practice 3: Framing Choices through Ineffective Affordances

1. *New practices in terms of information flows:*

The practice of practice of framing choices through ineffective affordances is analysed through a scenario that looks at deleting data released through searching. In this scenario data has already been released through previous interactions and has been collected, processed and stored. These information flows are represented below.

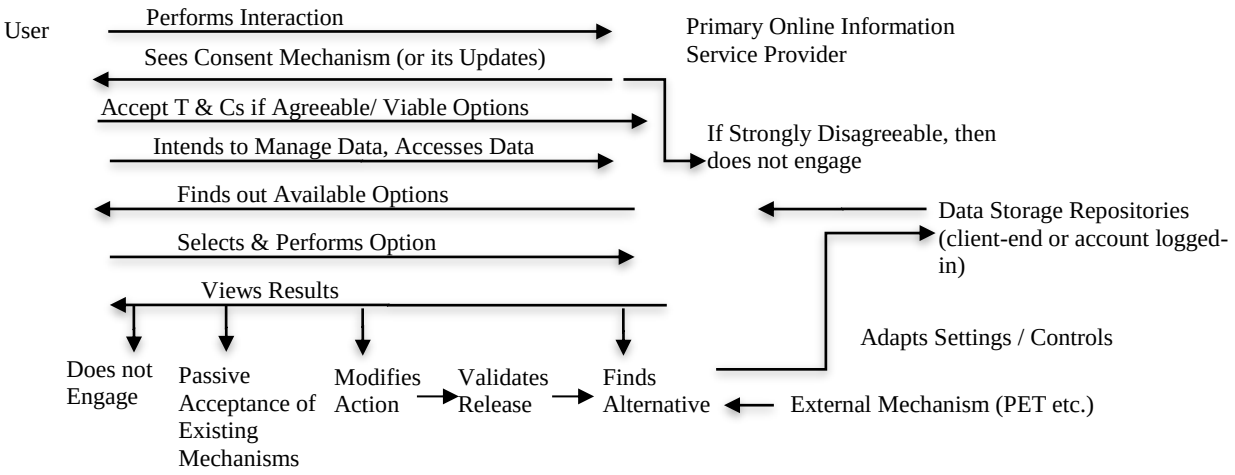


Figure 5.33 Practice 3: Effects of Ineffective Affordances

2. Prevailing Contexts and Impacts:

The prevailing context is interaction with an online information service provider such as a search engine, online website offering flight tickets, social network service such as Facebook or communication platform such as Skype (based on the examples participants discussed). Nested within this context are interactions with primary information providers and interactions with omnibus information providers that are not visible to users. Potential impacts include dissemination to these omnibus information providers.

3. Information subjects, senders and recipients:

The user is the information subject and sender (in most cases). The primary online information service provider is the recipient/sender, based on the information flow.

4. Transmission Principles:

The practice of deleting data involves the transmission principles of user choice in deciding to delete data that they deem no longer relevant or purposeful to their interactions.

5. Locate entrenched norms, identify significant points of departure:

The entrenched norms underpinning this practice are two-fold. When data is released, it is collected and stored in multiple methods, given the capabilities in storage and ease of recording and duplicating data. This norm is driven by the technological ability to capture and store data. Assuming the user has not actively attempted to prevent data leakage, data may also be captured by omnibus information providers.

This second norm is societal and revolves around user effort in setting up data release controls that are aligned with their privacy sensitivities.

6. *Prima facie assessment: what are the ways a system or practice defies entrenched norms:*

Technology created by the primary information service provider does not afford the user a range of mechanisms to manage deletion of data according to their privacy sensitivities. Deletion of data is only possible regarding data stored at the client end and account logged-in data. There is a certain degree of technology lock-in to platforms, making it harder to leave integrated platforms or opting out is inconvenient to users as choosing to opt out might mean choosing to not engage with that specific type of information service.

7. Evaluation 1: Consider moral and political factors affected by the practice in question:

The risks to users include inconvenience and a dependency or over-reliance on information providers in creating the controls to manage user information. Users are conditioned to passively accept the inability to delete or completely delete information as a new normal. Evidence is also seen of a perspective that data released online is public (as evident through the technical controlled experiment findings) even though it may be constrained to a contextual purpose (as evident through user expectations about the value of personal data released to obtain benefits). This undermines the autonomy of an individual in being able to manage information about their self.

8. Evaluation 2: Ask how the system or practices directly impinge on values, goals, and ends of the context. Consider the meaning or significance of moral and political factors:

Users desire to be able to delete data that is deemed no longer relevant to them. The current technological affordances are ineffective in addressing this user expectation. This disconnect creates information asymmetries between the user and the information service provider, as users realise they may not realistically have any control over deletion of data. The information asymmetry involves distinguishing between data that is released, data that is captured, and data that is captured and visible to users through being located at the client end and account-login. A lot of the data released and captured may not be visible to users and adds to this asymmetry as it may be used for purposes beyond the user's intention to release it.

9. Recommend in favour or against systems or practices on the basis of findings:

In its current state, the practice of framing choices through ineffective affordances violates contextual integrity and is identified as an information practice where privacy expectations are not met.

There is a need to design more effective affordances that allow ease in deleting data users deem no longer relevant. Currently options exist to remove data captured through account logins or visible at the client-side. However, verification that this removal of data has occurred is also important in sustaining user trust in Web interactions in the long term. Gaps exist in managing deletion of data released through the medium of interaction (automatic capture of browser fingerprints etc.).

Given, the lack of visibility of data flows from the findings of the technical controlled experiments (in Chapter 4) and the complexity in understanding the nature of these data flows from the findings of the empirical survey study (Chapter 5), it is important to consider designing mechanisms that afford users a reasonable expectation of privacy when releasing data while searching.

5.4.4 Summary: Analysis and Application of the Contextual Integrity Framework to Qualitative Thematically Analysed Privacy Expectations around Data Flows when Searching

The analysis of privacy expectations using the Contextual Integrity Framework involved the identification of three main practices. These practices are viewed as mechanisms afforded to users through which users practice privacy. Firstly, *the practice of collecting and retaining meta data:* the newly-introduced metadata regime introduces risks of independently verifying surveillance oversight measures and a potential loss of personal autonomy as individuals contend with their actions online being monitored.

Secondly, there is significant evidence to support that the known practice of *collecting and distributing data for commercial purposes* impacts user privacy. More common were expectations around data flows involving commercial data creation, collection, processing, dissemination and deletion. Evidence for two sub-practices were subsequently identified, including appropriateness of data flows (i.e. what constitutes appropriate flows of information) and the distributions effects of data (i.e. actors who have access to and visibility of these data flows) (Nissenbaum 2010). This existing practice introduces risks such as lack of awareness about data released both implicitly by the technology medium and explicitly through data users type in. Users with this lack of awareness and understanding are not in the best position to make informed decisions in managing their personal data. Users who are aware and wish to

manage their personal data modify their actions by self-censoring, minimising or obfuscating or using alternative tools. The risk here is that the burden is on the user to put in the effort to learn and understand the extent to which they can manage their data flows. Another risk is the exposure of details deemed personal or sensitive and the need to separate personal identity characteristics from their online interactions.

Lastly, the practice of *framing choices through ineffective affordances* does not adequately address user privacy expectations. Some users are aware and unconcerned with the release of data, as is their right to choose freely keeping in mind the survey respondents were largely an educated population of university students and staff. Further work is needed in generalising the extent of these effects to larger mainstream populations. However, the freedom to choose is being impinged on by the practice of designing ineffective affordances, which are not aligned with addressing user expectations. This practice is strengthened by the lack of choice offered to users in adapting their privacy expectations consistent with online social norms. Often consent mechanisms are invisible or impractical for users to fully understand. Although confidentiality is expected, this is hard for users to verify and have trust in as when data is used for purposes contrary to their intentions. Often the risk of technology lock-in means that if the user wants to engage with the service they have to abide by the rules offered by the provider, leaving with little to no ability to negotiate their online interactions to align with their privacy expectations.

5.5 Conclusion

This empirical survey study investigated the nature of societal privacy affordances around data release in search by identifying and examining privacy expectations users have about the flow of personal data in search. The survey comprised of quantitative, qualitative and interactive questions about search behaviour and user privacy expectations around data release in search. These responses came from a sample of 441 university students and staff. Quantitative and interactive search behaviour responses were analysed with respect to user awareness about tracking, expectations about data flows, personalisation, consent and disclosure mechanisms, privacy-enabling mechanisms. Findings from the empirical and technical investigations have confirmed that people are unaware or lack understanding of how personalised ads in search results use their personal data. This lack of awareness and understanding perpetuates a systemic pattern of power asymmetry in decision-making between consumers and commercial organisations.

Qualitative responses were analysed, coded and categorised into data flows, affordances, privacy values and its effectiveness representing a snapshot of the collective information practices from this sample population. Users display an "it depends" frame of mind underscoring the need to understand contextual purposes. These constraints also expand across time, location, purpose and visibility dimensions. Further expression of privacy expectations is constrained by the affordances offered by the information service provider. Designing appropriate affordances that address user privacy expectations is crucial in enabling users to effectively manage data release.

The outcomes of the survey study contribute an understanding of the privacy expectations people associate with personal data release when searching. Data flows were categorised under creation, collection, processing, dissemination and deletion of data. These data flows contained scenarios that were thematically analysed and categorised with respect to the nature of the privacy expectations expressed. The contextually integrity framework was applied to these data flows to identify if the derived information practices upheld privacy expectations. The derived information practices of collecting and retaining metadata, collecting and distributing personal data for commercial purposes and framing choices through ineffective affordances largely do not uphold the privacy expectations surfaced in the given contexts associated with personalised search. These findings contribute towards an understanding of privacy expectations associated with data flows in personalised search, which previously had not been investigated in depth.

To explore the reasons for why these privacy expectations are not upheld, a stakeholder analysis will be conducted in Chapter 6 to elicit the values at stake and a moral mediation analysis will be performed to understand the forces at play that influence socio-technical interactions like search. Based on these analyses, a concept for understanding privacy expectations: *contextually-public data release* is outlined, followed by a presentation of design recommendations that aim to appropriately afford users privacy.

6 Recommendations for Designing Data Release for Privacy-Sensitive Users when Searching

6.1 Introduction

Building on the understanding gained from investigations on the effectiveness of privacy affordances around managing data release when searching from a technical perspective (Chapter 4 – technical controlled experiment) and a societal perspective (Chapter 5 – empirical survey study), this chapter explores how these findings can lead to designing mechanisms that enable users to more effectively manage data release when searching in alignment with their privacy expectations. This exploration commences with conceptual analyses of stakeholder values and moral mediations followed by design recommendations to effectively manage data release in online user interactions such as search.

The nature of user interactions is socio-technical. "A technology has its properties embedded in a social environment" (Nissenbaum 2016, p.6). Design of socio-technical systems is concerned with dynamic interactions of users in its design process (Scacchi 2004). When users interact with an online product or service, they draw upon the values embedded in the underlying technology. For example, users release data implicitly and explicitly while searching. The technical and empirical investigations in earlier chapters revealed a significant gap between people's expectations of privacy online and the capabilities afforded to users when exercising their privacy through search. This gap is a design challenge because designers create the tools and mechanisms for online interactions and most users have limited opportunity or flexibility to change the mechanisms underpinning these interactions.

This conceptual gap addresses Research Question 4 (RQ4), *'How can we proactively empower users to effectively manage data release in search?'.* In exploring this question, socio-technical interactions that take place when searching is the key focus. The purpose of this chapter is two-fold.

Firstly, a conceptual analysis illustrates the key values at stake when users negotiate privacy while searching. This analysis integrates findings from technical (Chapter 4) and empirical (Chapter 5) dimensions using the stakeholder analysis method in value sensitive design. Drawing on this analysis, technology mediation theory is employed to explain the different effects at play on these values. Mediation analysis is used to elicit these effects and uncover the ethical assumptions embedded in these

interactions. A concept for privacy expectations in socio-technical interactions namely contextually-public spheres such as data release when searching is presented.

Secondly, to augment this conceptual analysis in practice, a set of design recommendations is presented. These recommendations centre on an approach transforming the relationship users have with technology in scenarios where privacy expectations are negotiated. The guiding principle underpinning these recommendations is to afford users with the right to freely associate online. Mechanisms need to be constructed to bring the divergent interests and values of two key stakeholders (users and information service providers) into alignment with online social norms and morals. The recommendations are categorised under two main themes, namely *primary* recommendations tailored to designers of technology products and services, and *supporting* recommendations that scaffold the designer-based recommendations by focusing on user empowerment initiatives and policy mechanisms.

The scope of this chapter is not to develop a fully-fledged analysis of technological approaches to address the conceptual gaps. Instead the scope is limited to a discussion of the necessary conceptual building blocks emerging from the findings of the technical and empirical investigations. This chapter contributes design recommendations that maintain the integrity of data release in socio-technical interactions.

6.2 A Concept for Understanding Privacy Expectations in Socio-Technical Interactions

A philosophically-informed analysis of the values at stake for users and information service providers and an understanding of how technology mediates user interactions when searching is presented in this section.

6.2.1 Understanding Stakeholder Values in Socio-Technical Interactions

Technologies embed human values (Friedman, B & Kahn 2003). A summary of common mechanisms through which technologies foreground human values is presented in Table 6.1.

Value sensitive design embodies an interactional approach by reinforcing it with socio-technical analysis and is informed by key morals at stake. Values in the context of this study are defined as the “goals, purposes, or ends; that is, the objectives around which a context is oriented” (Nissenbaum 2010, p.134).

Table 6.1 Mechanisms of Implicating Human Values in Technology

Method	Description
Embodied Position	Technology shapes people's behaviour because designers of those technologies embed their beliefs and values based on their worldview into it. This is also known as *technological determinism.* An example is the Berliner lock and key that ensures people remember to lock their doors by embodying a rotating mechanism that secures the door after entry (Latour 1992 as cited in Friedman and Kahn 2003).
Exogenous Position	Societal forces have a major influence in how technology will be used by people. The evolution of the Internet from a military experiment, to an academic initiative to a commercial entity attests to this (Friedman and Kahn 2003; Naughton 2015).
Interactional Position	The functionality afforded to users enables or inhibits certain values more than others, with technology use determined by user intent. For example, a screwdriver can be used for multiple purposes (Friedman and Kahn 2003).

Socio-technical interactions involve data flows where privacy expectations are formed and shaped through technological affordances offered to users. This research's empirical findings include the identification of three practices (i.e. collecting and retaining meta-data, collecting and distributing data for commercial purposes and framing choices through ineffective affordances) that impact how people perceive privacy as a practice in online spaces. This is complementary to the notion that privacy is a right to an inviolate personality and needs to be practiced (Rachels 1975; Warren & Brandeis 1890). If values are not afforded in user interactions, users may not be able to exercise this right. A typical socio-technical interaction in the context of this research is illustrated in Figure 6.1.

This section explores societal and technical values underpinning user interactions. These values goals, purposes or objectives have been drawn out through a review of literature and informed by the empirical findings (see Table 6.2).

[6]

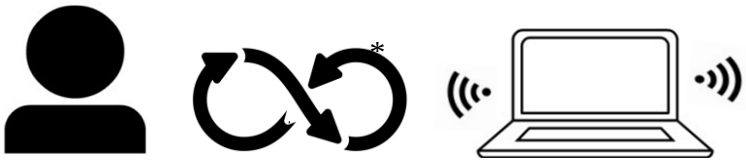


Figure 6.1 Socio-Technical User Interactions

[6] *The depiction of cyclic shaping uses "Infinity" by Olivier Guin from the Noun Project and is licensed under CC BY 3.0 US

Societal[7]	Socio-Technical Interactions	Technical
Trust Autonomy Choice Perception (Visibility) Knowledge (Awareness and Understanding) Transactional Benefits Identity Anonymity Transparency Accountability	Privacy as a Practice Interactions as a way of being (Winograd and Flores 1986 as cited by Friedman 2011) Meaningful Interactions Constitutive, Co-shaped (Verbeek 2011; Floridi 2016) Contextually Public Interactions Contextual Value: Local, Temporal, Purposeful (Intent) Ability to freely determine associations Information and Power Asymmetries (Nissenbaum 2010; Zuboff 2015) Hyperconnected Medium (Pasquale 2015; Dewandre 2015) Interactions are sticky by default (Dewandre 2015)	*Internet Infrastructure:* • Decentralised Control • End-to-End Principle (Naughton 2012) *Internet Affordances:* • "zero (or near-zero) marginal costs • powerful network effects; • dominance of power law statistical distributions • technological lock-in • low-entry thresholds (permissionless innovation) • combinatorial innovation • material – information" (Naughton 2016) Information and Power Asymmetries (Nissenbaum 2010); Siren Servers (Lanier 2013); Surveillance Capitalism (Zuboff 2015) Social Production (Benkler 2016) Remix Culture (Lessig 2008) Generativity (Zittrain 2008) Convenience, Efficiency and Speed Ubiquitous access to technology products and services Talkative and Reactive medium (Dewandre 2015)

Table 6.2 Characteristics of Socio-Technical User Interactions

According to social theory, particularly the notion of technology determinism, the basic premise is that people shape technology and over time technology shapes people, leading to social norms in motion (Feenberg 2009). The inherent complexity of the medium hides the visibility of information flows from everyday users (Naughton 2012). Given the nature of the medium, viewing it as dynamic ecosystem with forces at play is more useful as opposed to solely viewing it as a trade-off underpinned by economic incentives (Naughton 2012). Adopting this ecological metaphor, a representation of the forces at play in the data flow ecosystem is illustrated in Figure 6.2

This socio-technical interaction is mapped as a layered funnel, where with each wave or layer, the propagation of data released may be visible and data accessible to many indirect stakeholders. The

[7] Derived from the analysis of empirical findings as explained in section 6.3.1.1. Societal Characteristics

interaction between the user and the service is expanded to visualise the interests and impacts of different stakeholders, who are often invisible to everyday users. Some of these stakeholders aim to target, while others aim to track users across multiple browsing sessions (Nissenbaum 2016).

Currently the protection mechanisms afforded to users involve taking the initiative to use a privacy-enhancing technology to obfuscate data release, trusting external parties like a certificate authority to independently verify the integrity of transactions or relying on organisations that collect their data to uphold data protection laws as relevant to their jurisdiction. The most effective protection mechanism lies with the user being able to make informed decisions on data release before engaging with medium. Any layer that follows the user layer is susceptible to propagating data, given the generative property of the medium, network effects and the monetary value in information appreciating over time and use. Avenues for commercial data collection or surveillance of data to protect national security interests within regulations are also plausible given this propagation and openness of the medium. However, awareness and understanding of the importance of data release is at present comparatively low among general users. Interactions shape both online social norms and individual information practices. These interactions must be meaningful for users to draw on and evolve with the changing networked information economy.

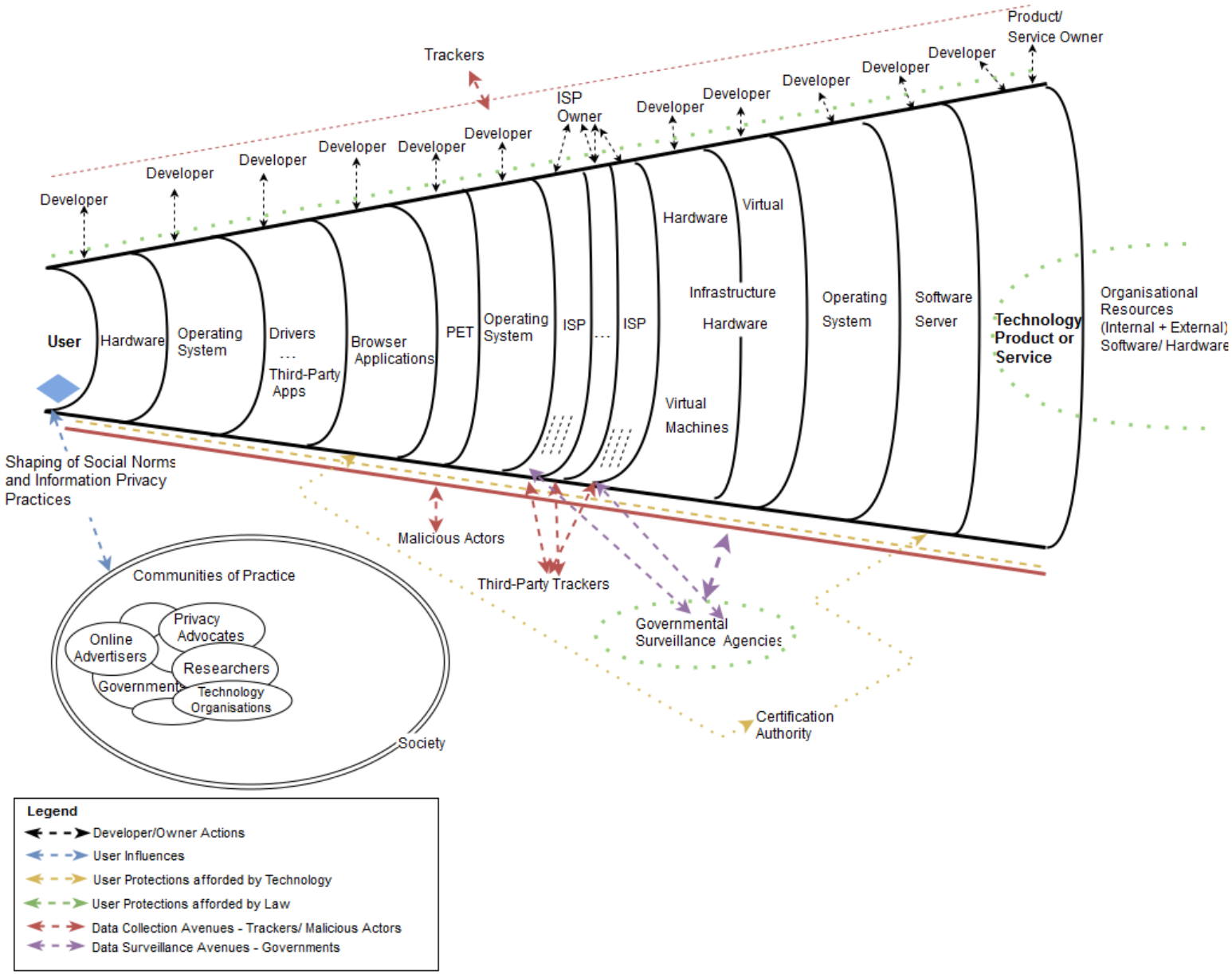


Figure 6.2 Socio-Technical User Interactions - Expanded View of Stakeholders

6.2.1.1 Societal Characteristics

Societal characteristics explore distinct features of interactions that arise from socially structured settings. Technology development and use shape societal norms. Societal characteristics relating to Web search as emerging from the empirical survey study are described in this section.

In socio-technical interactions, users typically assume their privacy expectations to be upheld. Values like trust, autonomy and transparency are commonly associated with these beliefs. "Trust is based on an assured belief that another person displays integrity, veracity, justice, and other aspects of ethical behaviour, and will continue to do so." (Goldberg, Hill & Shostack 2001, p.101).
These include data released being used for primary purposes ["*Personal data should only be used within the confines of what they want to be used for. If they want a more customized search it should be made clear that they are using data that they have collected from previous searches and Identity information."* (P#10)], being accessible to authorised stakeholders and situated in the context in which it was originally released [*"I expect when I am requested to give personal information that it remains private and secure between the two parties"* (P#264)"]. However, loss of trust is typically associated with privacy violations. Users currently are heavily dependent on technology services or products to honour their trust. There is a need to reciprocate trust online between users and information service providers.

Autonomy is closely associated with privacy. Autonomy is concerned with "control over the intimacies of personal identity" in relation to privacy (Gerety 1977, p.236).
Often autonomy in online interactions is manifested through an array of choices afforded to users. These choices assert that users have the ability to select a particular functional option. User choice is mostly predetermined ["*I expect them to record all of my data but I am not happy about it."* (P#13)"] with most users selecting the default option or passively accepting an option from the range of pre-determined options ["*"I expect it to continue to creep me out, but I also expect to become accustomed to it over time."* (P#280)"]. Given the complexities involved, choices in managing data release manifest as pre-determined options.

When choosing, user perception is affected by their level of knowledge, awareness and understanding as indicated by. Awareness and understanding that users have of how personal data is released and used online influences their expectations [*"I wasn't aware that any personal data was actually used when I searched."* (P#395) or *"I am not comfortable with any of my growing history being retained to refine search results."* (P#396) or *"I don't feel safe if people can keep track of me."* (P#423)]. Visibility of

data flows and its reach is also sub-liminal. Users tend to perceive the effects of privacy reactively. Data flows often reach audiences that are invisible to users ["*at first, when searching, and the internet gave preferences I might like I realised I was being tracked and didn't like that but didn't know there was anything I could do about it. I don't want to give free info to a data collection agency that they will then benefit from."* (P#101)].

When privacy choices are made online, users often deal with incomplete information making decision-making a complex task (Acquisti & Grossklags 2005). Users are largely unaware of the numerous mechanisms available to monetise and distribute personal data (Tene, Omer & Polonetsky 2012). Transparency with choices is commonly agreed to be important (Whitten et al. 2011). However, implementations of transparency mechanisms designed to reassure users about data collection are weak [*"I have selected the option for Google not to collect any data about me (including search history), but behind the scenes I know they probably still collect data, just without telling me. So transparency about their data collection probably won't change anything, as they can easily lie about it."* (P#158)]. Users are unable to form a trusting relationship with the service provider (Pasquale 2015). Although the complexity of the interaction is hidden from the user and is necessary to make the interaction functional, the lack of reciprocal transparency and accountability ensures that the interaction remains a 'black-box' to users.

In the digital world, a representation of a person's personality (as modelled by data released through transactions), functions as his or her identity (Clarke 1999). The value perceived by users in releasing their data is transactional [*"I didn't mind my information being used when clothes shopping because websites then promote relevant items and stores I might be interested in."* (P#115)]. This finding lends support to the notion of separating the transactional value users associate with data release from their digital identity. Transaction identity is used to authenticate users in transactions as opposed to tying each interaction to a user's digital identity (Sullivan, C 2013). A user's own digital identity is often invisible to the user and is fragmented across different services.

Anonymity allows users to freely express their intentions and interact online (Nissenbaum 199). In the offline world, being nameless in a social setting provides anonymity. The goal of anonymity online is to be unreachable. Anonymity remains a key concern for users [*"Although it is inevitable that our online behaviour is tracked, I wish we could remain as anonymous as possible."* (P#290)].

The boundary between publicity and privacy was also evident. The empirical findings indicate that people do want to share their data and find value in doing so, but not as publicly or as openly driven by technology. Instead, users perceive online privacy as a practice. In interacting with a service provider, users play a specific role that allow for managing this permeable boundary between privacy and publicity of personal data.

6.2.1.2 Technical Characteristics

Technology is also shaped by social norms. We cannot build technology in a vacuum or in isolation. For example, the Internet has matured to be a general-purpose technology on which new industries and business models are built (Naughton 2016b). The technical properties that underpin interactions are described in this section.

Certain key affordances enabled the development of the Internet from a military experiment to a civilian utility to the present day commercial entity, there are the Internet offers. These are "*zero (or near-zero) marginal costs*; v*ery powerful network effects*; the *dominance of power law statistical distributions* (which tend towards winner-takes-all outcomes); *technological lock-in* (where a proprietary technology becomes the de-facto technical standard for an entire industry); intrinsic facilitation of exceedingly fine-grained surveillance; *low entry thresholds* (which facilitates what some scholars call "permissionless innovation"); a development process characterised by *'combinatorial innovation'* which can lead to sudden and unexpected new capabilities, and an *acceleration in the pace of change and development*; and the fact that *the 'material'* that is processed by the technology is *information* — which is, among other things, the lifeblood of social and cultural life, not to mention of democracy itself." (Naughton 2016a)[8]. For example, technologies like homomorphic encryption enable utility of service and confidentiality as its design is malleable. However, many information service providers prefer to lock-in data within their application environment although the opportunity to ensure confidentiality exists and will be functional when resources become scalable.

It is important to take these characteristics into account when designing technologies that harvest personal data. This leads to power asymmetries between the service provider offering the service and the

[8] Emphasis in italics is by the referenced author.

user consuming it (Nissenbaum 2010). These asymmetries create new power structures called siren servers ((Lanier 2014) and enable surveillance capitalism (Zuboff 2015).

The internet which acts as the enabling infrastructure on which applications run was built on two fundamental principles. These are *decentralised control*, which allows no single entity to control how the internet operates and the *end-to-end principle* which embeds simplicity and flexibility in catering to different applications that may innovate as they see fit (Naughton 2012). These principles embody the *openness* of the internet that gives rise to much of the '*permissionless innovation'* of products and services that users consume.

The Internet created a backbone for future services to run on. One of these that revolutionised the use of the Internet is the World Wide Web (Berners-Lee 1989), which effectively applied existing technology (hypertext and a standard generalised markup language for communicating structured documents) to a new domain context. However, the architecture of underlying protocols such as HTTP only provide one-way linking. Ted Nelson's vision of how users could interact with the Internet was achieved although limited in comparison to his original vision (Ted Nelson's Project Xanadu (Nelson 1960)). For example Nelson advocated a two-way link architecture that would have preserved context when releasing data in search (Lanier 2014), as opposed to the one-directional HTTP transfers we have in existence now. The attribution of content authorship is lost because information about originating sources of content are not captured and the link relations not maintained (Lanier 2013).

This characteristic fuels *social production*, which is the creation of content by users during the process of interacting with an application for no financial payment (Benkler 2006). The conventional argument here is that users obtain benefits such as *efficiency* and *speed* through technology-mediated interactions that are not possible through similar engagement with traditional brick-and-mortar based businesses. These benefits of efficiency and speed could continue to be benefits even with two-way link architecture. Ideas have also been proposed on opening up the value of these interactions to people, enabling authorship of original material released through interactions by way of micropayments using a two-way link architecture (Lanier 2013). Lanier advocates that this would significantly empower creators of content by rewarding them for their original input (Lanier 2013). It would also enable accountability of content posted and act as a social reward in sustaining online interactions. Alternative approaches to embed context in data release include Microsoft's metadata-based architecture with the primary aim of aiding privacy policy initiatives around data protection (Nguyen et al. 2013).

The services running on this Web infrastructure create a networked information ecosystem that *disrupts* traditional modes of operation (Naughton 2012). The networked information economy is driven by two features, an economy based on information and cultural creation and with socially significant symbolism, and a reliance on a communications platform that utilizes the benefits of cheaper and efficient computational, communication and storage resources that are tightly integrated in a ubiquitous network (Benkler 2006). A key characteristic of this networked information economy is the importance of "decentralized individual action,−specifically, new and important cooperative and coordinate action carried out through radically distributed, nonmarket mechanisms that do not depend on proprietary strategies." (Benkler 2006 pp. 3). This networked information economy is generative. Generativity is determined by "(1) how extensively a system or technology leverages a set of possible tasks; (2) how well it can be adapted to a range of tasks; (3) how easily new contributors can master it, (4) how accessible it is to those ready and able to build on it and (5) how transferable any changes are to others – including (and perhaps especially) nonexperts." (Zittrain 2008, p.71). These characteristics fuel innovation through the openness of the underlying architecture (i.e. the Internet) and the programmable capability of the services offered on it (Naughton 2012). A similar concept is that of *logical malleability* where computers can be enabled to perform tasks that can be processed logically with inputs and outputs (Moor 1985). These features give rise to a '*remix culture'* where low-cost capabilities exist to slice content and reshape it into a new form (Lessig 2008). Users also harness the benefits of this automated processing particularly in terms of *efficiency* and *speed*.

Expanding on the material produced and exchanged through interactions, information *appreciates* in value compared to traditional market goods that depreciate over time (Ariely 2008). This is central to the extensive data collection that drives big data analytics to create business intelligence. Often, all plausible types of data are collected in the hope that future purposes may be served. However, not all patterns derived from such analysis create meaningful insights (at least, initially), and questions arise regarding the validity, currency and integrity of the data being input. Algorithmic biases may also influence the outcome of interactions. This can be through the biases held by programmers which are translated into code and the limits of data capture, where we collect what can be captured based on the capabilities of the technology and also what people choose to reveal consciously or unconsciously.

Of particular relevance to privacy, there are specific characteristics that underpin data flows. Linkability of interactions (i.e. search queries) enables disparate sets of personal data to be identified and inferences made from these established links (Sweeney 2002). De-anonymised data can be relatively easy to be re-

identified and linked (Narayanan & Shmatikov 2010). The ubiquitous use of digital devices and access to relatively free and low-effort technological capabilities to leverage value proliferate the impact of these concerns (Zimmer 2008).

Related work in privacy-by-design has called for the embedding of privacy into design (Cavoukian 2009). The reactive nature of embedding privacy-by-design features into functional products means often privacy loss occurs through data exposure. Interactions like search are also visible to third-party data aggregators, that then build repository and profiles often unknown to everyday Internet users. Omnibus information providers like Axicom operate on this business model (Nissenbaum 2010). The nature of the medium presents unique challenges for expression of privacy expectations and the designing of privacy-sensitive affordances that enable appropriate flows of information.

6.2.1.3 Stakeholder Values Analysis

These practices as derived from the empirical investigations are explicated in Table 6.3. An analysis of the values held by primary stakeholders is also mentioned along with its mediation effect. Although prior research has investigated embedding the value of privacy in technology design, tensions still persist for users in how they manage data release and subsequent privacy effects. This is evident through the survey finding results such as passive acceptance, the creep factor induced by the pervasiveness of data collection and tensions in interactions that have information asymmetries.

From the review on existing privacy-enhancing technologies, we argue that the visibility of data flows in search is not apparent to users. Existing affordances are function-specific, siloed and are often reactive. Through the controlled experiment on data release, when potentially personally-identifiable data were obfuscated, significant variance in personalised search results could still be obtained. Privacy-sensitive affordances were identified through obfuscation of location and user-agent by providing minimum data required for an online transaction and through a level of pseudonymity (detectable but not personally-identifiable search queries) for data release in personalised search.

Table 6.3 Summary – Stakeholder Value Analysis

Information Practice	Derived Privacy Expectation	Stakeholder (User) Value	Stakeholder (Information Service Provider) Value	Outcome
Practice 1: Collection and Retention of Metadata	Surveillance Risks	Convenience National Security Secure Data flows	Government – National Security and Law Enforcement Information Service Provider - Compliance	Mandates recording of interactions for surveillance purposes Normalises surveillance
Practice 2: Collection and Distribution of Data for Commercial Purposes	Commercial Risks Positive Uses of Data Collection and Aggregation Relevance of Recommendations Constraining Use and Managing its Impact Non-Targeted Uses Perceiving Impact of Aggregated Data (Ambivalent, Misinformed and Unconcerned Views) I want this, but it depends on (contextual interactions) Contextual value (location and time) Understanding the value exchange in linking data sources Pervasiveness of Data Collection (Creep Factor) Inappropriate Use	Interaction driven– wants to obtain benefit Convenience Ability to freely determine appropriateness and distribution effects of data flows	Targeted Advertising - Revenue based driven Accumulates wealth of data harnessed for computational intelligence	Passive Acceptance in cases where there is narrow framing of choice Transactional Interaction Viewed increasingly as 'public interactions' Visibility of indirect stakeholders reduced Increases accessibility of data and tools to process the data Enhances and improves productivity and efficiency of computational tasks Normalises interactions Habituates expectations
Practice 3: Framing Choices through Ineffective Affordances	Identifying Sensitivities, Revealing Identities Actioning User Choice and Consent Framing of Choice (To Use or Not to Use) Disconnected Interactions Ineffectiveness of Available Affordances Expectation of Convenience Tensions in Information Asymmetries Abilities in Deleting Data	Interaction driven– wants to obtain benefit and privacy is often a secondary goal during the interaction. Convenience Ability to freely determine associations	Defaults usually set to maximum exposure to aid data collection Basic privacy-enhancing steps such as encryption not readily adopted.	Frames selection of choices, guides expression of privacy expectations Nuances of expression not readily translatable as context specific Alternative measures are not reachable to everyday users. Conditions beliefs

From the empirical survey study, instances of appropriate flows and derived collective information practices around data flows in search were identified. Collectively users are aware of tracking, but less aware of the impacts of tracking such as linkability of search queries, behavioural profiling, data aggregation and secondary use of data.

From related work, societal norms for the privacy protection online are inferred to not be fully understood, a gap which this research aims to address. Societal norms are heavily entrenched or influenced in the medium of interaction, which in this case is the use of the Web on which information searching occurs. Affordances are key to enabling effective mechanisms for expression of societal privacy norms. However, there is an inconsistency between the values held by the user and the service provider. In order to the understand how these divergent views can be bridged, the socio-technical interactions are explored further in the following section.

6.2.2 Understanding Morals Mediating Interactions

Ethics of computers is the "analysis of the nature and societal impact of computer technology and the corresponding formulation and justification of policies for the ethical use of technology" (Moor 1985 pp. 266). Morals are central to this analysis of societal impact of technology use. Ethics uses moral theory to identify appropriate online social norms and the ways in which technology innovates and extends moral boundaries (Friedman and Kahn 2003). Traditional moral theory takes the form of consequentialist, deontological and virtue-based perspectives. A consequentialist position advocates that an entity should work towards generating optimum outcomes for a given situation. For example, utilitarianism is where the entity should produce the highest possible amount of utility for the most people (Friedman and Kahn 2003). This idea of utility in technology is a widely-employed concept. Trade-offs such as personalisation vs. privacy and the utility of PETs are often based on this premise. Deontological theories suggest that certain actions are morally unacceptable or inappropriate regardless of the outcome (Dworkin 1978; Gewirth 1978; Kant, 1785/1964; Rawls 1971 as cited in Friedman and Kahn 2003). For example, data collection by an entity without informed consent from a user falls in to this category. Virtue-based ethics imply that morally-good actions are at the discretion of the entity and if performed, establishes and reinforces the moral values at stake (Friedman and Kahn 2003). For example, if DNT (Do Not Track) technically honoured the moral values of anonymity (through not tracking), this would indicate that technology organisations reciprocate the morals that users hold.

The value sensitive design method was used to elicit key morals at stake for primary stakeholders in section 6.2.1. However, this method does not take into account how values are mediated by technology, given that values may change over the course of an interaction due to the different forces at play and the intentions and actions of different actors (as illustrated in Figure 6.2). Values may change over the course of the interaction because they may be influenced by the values imposed by different stakeholders and mediated by the affordances provided to users. The nature of values embedded in interactions is ephemeral because the values embedded in the interaction are not solely the user's, but influenced by the values held by different stakeholders in the affordances offered to users which mediate interactions. Asymmetries arise with such interactions afford mechanisms that are in conflict with the user expectations. For example, the current mechanism to obtain consent is a static generic one-time process, but consent is implied for all scenarios and subject to change, whereas data release is contextually-public (meaning consent should be obtained based on the nature of interactions particularly when contexts of use are subject to change). Regardless of related work that shows that notice and consent, and informed consent are not effective affordances, these mechanisms are still the preferred choice of use. In this section, we employ mediation analysis and moral assessment of socio-technical interactions that occur when searching to uncover the values at stake. We take into account properties of the Internet, the Web, the Web browser and the search engine as part of the existing product-in-design that people use to seek information.

"When technologies are used, they always help to shape the context in which they fulfil their function. They help to shape human actions and perceptions, and create new practices and ways of living. This phenomenon has been analysed as 'technological mediation': technologies mediate the experiences and practices of their users" (Latour 1992; Idhe 1990; Verbeek 2005 as cited in Verbeek 2008 p. 92). The notion that artefacts can have morality has been widely debated in philosophy. For example a speed bump warns drivers to slow down before reaching it to avoid damaging the car and ensure road safety (Verbeek 2008, p.Latour 1992 as cited in Verbeek). This debate is situated in phenomenology, "the analysis of the structure of the relations between human beings and their lifeworld." (Verbeek 2011, p.7). To have morality, artefacts should have some degree of intentionality and freedom (Verbeek 2008). Verbeek identifies two types of mediation that explain how artefacts have morality (Verbeek 2011). Mediations of experience (or hermeneutic) focus on the lifeworld and how reality is perceived by people as directed by artefacts, while mediations of praxis explore how human actions are shaped by artefacts (Verbeek 2011). Types of mediations from related work are summarised in Table 6.4. Both

mediations of praxis and experience delegate by deliberately inscribing scripts and are multistable by adopting different technological identities or intentionalities depending on the context (Verbeek 2011). Multistable means that use of technology is not solely dependent on the designer's original intention, but based on how users interpret its use in different contexts (Wendt 2015).

Table 6.4 Types of Technological Mediations

Method	Description	Reference
Mediations of Experience		
Embodiment Relation **Readiness-to-hand**	"Technologies are incorporated by their users, establishing a relationship between humans and their world through the technological artifact." Example a pair of glasses.	Idhe 1990, Heidegger 1927 as cited in Verbeek 2011 p.8
Hermeneutic Relation	"Provide a representation of reality which requires interpretation" Example a thermometer, infrared cameras Characteristic: has technological intentionality as it directs and transforms perception by amplifying or reducing dimensions of reality.	Idhe 1990 as cited in Verbeek 2011 p.8
Mediations of Praxis		
Script	"Actions are the results not only of individual intentions and the social structures in which human beings find themselves (the classic agency-structure dichotomy) but also of people's material environment." A script describes "the influence of artefacts on human actions" e.g. speed bumps, disposable coffee cups Characteristic: translates actions by inviting or inhibiting actions based on the context. A telephone enables social relationships to be sustained outside one's immediate local environment.	Latour and Akrich 1992 as cited in Verbeek 2011 p.10

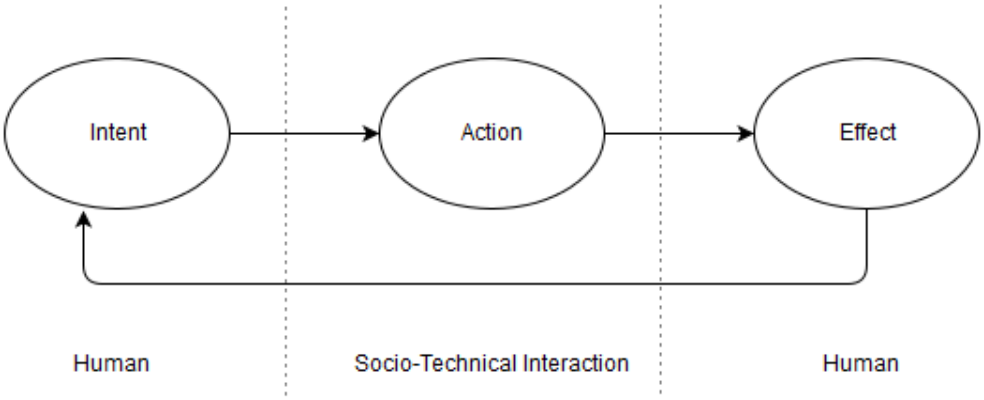


Figure 6.3 Socio-Technical Mediation Pattern

An analysis of socio-technical interactions in search is presented in this section. This analysis draws on findings of the technical controlled experiment in Chapter 4 and the empirical survey study of privacy expectations in Chapter 5. The overall pattern that these socio-technical interactions take became evident

through this analysis and is illustrated in Figure 6.3. Mediation effects have influence on user intent, action and its effect which over time reinforces intent. A summary of technological mediations identified through this research study is presented in Table 6.5 and described in the following section.

Table 6.5 Mediation Effects

Category	Mediation Effect	Description	Reference
Intent	Induces Liminality	The privacy protections afforded to users like PETs are not perceivable by users and their process of being awareness and taking action is in transit.	Study Findings
	Frames Choices	Options offered to users are framed from the perspective of the service provider	
	Influences Expectations	Reinforced commonly accepted views of online privacy as mediated through interactions and framed by choices influences expectations.	
	Contextualises Intent	Intent is contextualised based on time, location, purpose of interaction and recipient actor.	
Action	Observes Actions	An interaction is immediately observable given the nature of the medium.	Study Findings
	Repurposes Actions	Actions are automatically recordable and can be repurposed beyond a user's original intent.	
	Regulates Actions	This observability of actions also enables it to be regulated.	
	Guides/ Persuades Interactions	Users are also guided and persuaded by the capabilities offered to users within the interaction.	(Dorrestijn, van Der Voort & Verbeek 2014; Verbeek 2011) Study Findings
Effect	Adds Value	Users perceive value through interactions.	Study Findings
	Slicing Identities	Interactions slice identities as users play different roles.	
	Conditions Beliefs	Beliefs are conditioned over time through the capabilities and values embedded in the technology.	
	Habituates Norms	These conditioned beliefs over time habituate user expectations of online social norms.	

6.2.2.1 Intent

Heidegger reduces the world into a set of raw materials which we interact with; meaning is formed through these interactions and the meaning we hold in these raw materials (Verbeek 2011). If technologists design ways of being (Winograd and Flores 1986 as cited by Friedman, B 2011) and tools are extensions of self (Gibson 1979; Heidegger 1977), PETs are not reachable to individuals because they are not integrated to their beings (i.e. the socio-technical interactions we experience). PETs are not natural extensions of being (of the experience of searching). This is due to their invisibility which is affected by factors such as user awareness, understanding and effort (Moor 1985). This effect can be described as *inducing liminality*, where the usefulness of PETs cannot be clearly perceived by users and the process of raising awareness and taking action undergoes change at an individual level for each user.

The effects of objects being subliminal can be concerning as people may be unable to perceive the object without being made aware, but it affects their psyche. For example, legislations regulate subliminal advertising messaging in movies with companies paying instead for product placements as embedded advertising to make these objects noticeable. Subliminal affect is also described as mediation in instances where people are stimulated by barely conscious sensations such as use of fresh bread and coffee aromas to entice buying habits (Dorrestijn, van Der Voort & Verbeek 2014). Similarly, an interesting parallel is seen with online advertising. Pervasive data tracking that occurs is subliminal to most users without the necessary pre-requisite knowledge that take years of understanding and learning. Users cannot perceive data collection let alone fully understand the effects of its propagation. Consumers deal with ambiguity when attempting to distinguish organic and sponsored search results (Scardamaglia & Daly 2016).

PETs are imbued with purpose by users. However, PETs are also imbued with purpose by their designers, purpose which is often functional. Privacy is a social practice though. For artefacts like PETs, the purpose imbued by users is limited. This is because although intention exists with users (i.e. users may intend to have privacy online), the affordances that embedded into PETs are conditioned by the beliefs or purposes held by providers or prevailing online social norms. These mediations lead to a *framing of choice.* This framing may result in instances where users passively accept the options offered to them to engage with a technology product or service, because the cost of not engaging is no longer a viable option. The socio-technical interactions of users inform their information practices. Over time on an aggregate scale these practices form online social norms, reinforcing commonly accepted views of online privacy and *influencing expectations*.

Meaningful interactions are realised when *intent is contextual*. The empirical findings revealed users find value in data release when the intent is to find local search results such as location directions. Value is also realised when recommendations for ads that are personalised based on profiled data are temporally constrained based on intent. For example, consumers do not usually intend to buy the same product in succession if they are consumer durables (goods like bread and milk that are bought in predictive patterns). Interactions are also contextualised based on intended purposes and actors that underpin data released during an interaction. This mediation of intent is based on metrics that can be measured by the technology, given the limited nature of the interaction.

6.2.2.2 Action

In understanding how technologies mediate actions, we draw on Heidegger's notion of being, where tools are considered as extensions of self (Gibson 1979; Heidegger 1977). It is in the act of interacting with the tool, through which we draw on the value the tool affords us. In other words, if the tool does not afford symmetries in information flows addressing privacy expectations, the tool may not afford us a reasonable expectation of privacy. Heidegger advocates distilling objects down to their raw materials. By embedding these primary ingredients in interactions, a foundation is formed. However, privacy is already atomic, it cannot be broken down further into raw materials as embodied in the functional and siloed protections afforded to users. Instead the form privacy takes in socio-technical interactions is that of a practice. Form shaped by the technology offers privacy protections as functional components. This tension in protections afforded creates asymmetries because privacy as collectively perceived by society and protection afforded by technology are different.

When a user interacts online, each action by nature of the technology medium, observed by a trail of data leakage. In the offline world, actions by their very nature are not immediately observable, but certain factors allow for this, most of which are within the control of a person. For example, photos of daily life are not automatically recorded unless someone acts to capture it. Online actions are observed through tracking and targeting mechanisms which also allow for it to be propagated beyond a user's original intent. This repurposing of actions is often hidden from the user and visible to actors not privy to the original interaction. This observability also enables the possibility of regulating actions. More commonly, the affordances available to users guide and persuade them in shaping their interaction (Dorrestijn, van Der Voort & Verbeek 2014; Verbeek 2011). These mediations imply different consequences to agreements and attachments to actions.

6.2.2.3 Effect

User purposes or intents for privacy cannot be adequately expressed in its current form, which is akin to ticking boxes on a checklist. Instead design should inculcate habit-forming behaviour that enhances privacy practice. It should be incorporated to the whole being (part of the whole experience). Functionality is amenable to interactions based on raw material components. In turn, when users interact, they interact with these ways of being designed for them, and over time are *conditioned to these beliefs*. This conditioning of individual beliefs *habituates online social norms*.

When users are made aware of these limitations in their ways of being (interactions), information asymmetries become clearly visible and the disconnect in privacy expectations arise. From the survey results, a higher proportion of users reported being aware of Google's Privacy Checkup function. This is an indicator that when information is raised as part of a user's experience, it creates awareness. The next steps should be to inculcate habit-forming behaviour using feedforward (instead of feedback) loops that are designed to target user intention instead of reactive, action-based prompts. Signifiers during data release are missing so the focus is on proactive and preventative action rather than reactive. This would mean that when data release occurs, users are made aware and have options available to them to exercise their privacy expectation if they so choose to do so. To sum up, over time and through interactions, user expectations are habituated (Pasquale 2015): this creates a conditioning effect; and central to this habituation are the affordances offered to users.

6.2.2.4 Understanding Mediations in Socio-Technical Interactions

Affordances

From this analysis, the key organising principle that enables users to freely determine appropriateness and distribution of data released through interactions like search are the affordances afforded to users. Often this data leakage is not visible or is subliminal to the user. Affordances in the context of this research expands beyond Gibson's concept of direct perception by using Verbeek's mediation analysis to offer new language or explanations around addressing the privacy issues users face.

Affordances can also be conceptualised as a relationship between the properties of a technical artefact and the capabilities of a person in using that artefact (Norman 2013). The relationship between users and technology as perceived by the surveyed users is transactional. This is because the interaction is commercially driven, where the functionality offered to users is determined by profit-oriented incentives. For example, the hook model refers to habit-forming behaviour mechanisms embedded in interactions. A hook follows the pattern of internal or external triggers such as rewarding incentive embedded in pushing email onto your phone. When people try embedding privacy into this kind of transactional relationships, practicing privacy is not sustainable. People only want data release in contextual interactions. However, this is heavily mediated by technology often framing the transaction as intent vs. benefit to users.

The nature of privacy is such that it is a state that is experienced, like security, it is intangible, users often feel it when it is lost, which is irreversible as the fact of knowing can amount to that damage. By

expanding the design ethos to make tools reachable, they can then function more effectively as privacy protections. By embedding the right to associate freely in socio-technical interactions, design becomes robust in catering to privacy expectations. In a hyperconnected world, interactions only capture versions of ourselves. Hence, we should design for versions of ourselves because users perceive to be acting in a particular role when interacting with a service provider. The importance of this is that liminality, if not surfaced, will continue to condition beliefs and shape interactions. This degrades societal values and devalues the practice of privacy. The metrics that measure the usefulness of interactions are transaction-based. For example, these are 'clickthrough data' and other types of data leakage. This creates the need for reciprocal interactions among relational selves, where respect for the agency of the user is upheld, and the choice to determine release of data and aspects of identity rests with the individual (Arendt 1958; Dewandre 2015a).

Instead the need to shift these metrics to meaningful interactions where design is based on relational-selves of individuals based on their roles and sliced identity. While the value of privacy has been explored extensively, meaning plays a role too (particularly with respect to how people understand information privacy). "Raising the standards means altering the definition of the object, not paying a price for an alternate good or ideological value as the trade-off model holds." (Feenburg 2002, p.173)). We raise standards through embedding these human values in design through affordances (which people can then use to exercise their privacy). This is why it is a design challenge towards resolving the ineffectiveness of affordances in addressing privacy expectations.

Designing Appropriate Interactions

Appropriateness is defined as a flow of data that addresses user's privacy expectations. Current trends go beyond the development of apps, building on the links between interactions. The boundaries between subliminal tracking and conscious data release are blurred driven by developments in behavioural advertising and use of chatbots that rely on natural language processing in conversational commerce. How can we distinguish between content and sponsored content without form? Does this impact meaning?

The aim of the leading search provider is to be able to predict your query and present results relevant to you based on what they know of you. [Eric Schmidt, Google's CEO: "I actually think most people don't want Google to answer their questions. They want Google to tell them what they should do" (Jenkins, HW 2010). While this will be of immense value in terms of efficacy and speed of service in seeking

information, two key issues arise. One is that of relevance, which is currently determined by click-through data, user search history and behavioural data for example. This data collection is incentivised by search result algorithms and rankings which determine how your search results are tailored to you and the degree of its relevance. Commercial values influence this process.

Second, although search now provides much needed accessibility to a wider breadth of resources in an efficient and speedy manner, it does not freely allow users to determine what is relevant to them. The interaction with users is not reciprocal, where most data determining user input (such as click-through data) is passive and does not allow users a mechanism in which to determine the degree of relevance of search results. This gap in reciprocal interactions is evident when critiquing the differing purposes behind affordances (See Table 6.6).

Although old asymmetries of access and consuming information are broken, newer information and power asymmetries are created. Users will benefit from more accurately predicted recommendations based on larger sources of data, as sensors become more embedded in daily life recording their behaviour. Targeting and tracking will become more sophisticated. Experiences will be shaped more effectively to influence what to buy, so it will seem convenient to the user, but will also be convenient to the service provider as this process will still be driven by commercial purposes.

This process will shift from monetizing individuals' data to monetizing their experiences (Zuboff 2016). As individuals become highly networked beings, they need ways of interacting that allow them to distinguish between what is presented to them and what they would like to consume. This ability to identify the subliminal boundary between the two is important as autonomy and the ability to determine what is relevant for users is central to their core being and to their expectations of privacy in online spaces. Affordances currently offered to users to manage personal data released through search are examined in Table 6.6.

The nature of the medium is an important influence in shaping societal norms. Technological determinism is a school of thought that assumes that technology influences cultural and social norms (Feenberg 2009). This view is often reflected in the worldviews held by some technology entrepreneurs like Mark Zuckerberg ("a squirrel dying in your front yard may be more relevant to your interests right now than people dying in Africa") and Eric Schmidt ("We know where you are. We know where you've been. We can more or less know what you're thinking about"; "If you have something that you don't

want anyone to know maybe you shouldn't be doing it in the first place") (Jenkins, H 2010; Pariser 2011).

Table 6.6 Existing Affordances

Affordance-Action	Declared Purpose (Primary Providers)	Alternate Uses (Primary and Secondary Providers)	Ideal Use (User)
Single account sign-in	Enables users to stay engaged, convenient, don't have to remember usernames and passwords	Link different accounts and use of different services	Constrain each login by default only to the context of interaction
Personalised search results	Needs more data from users to make it more relevant	Is it actually better for users? Reasonable variation seen only with location, user-agent (browser) and search query term	Have options that focus on location (making it more visible) and query term (people type in)
Profiling	Need to know as much data to deliver better benefits (search results, recommendations)	Commercialise the data through monetising ad networks	Temporal interests means profiles need to be refreshed. What makes them relevant is not a well-defined metric
Ineffectiveness of Consent	Give users notice (to make an informed decision) and the option to consent	Relies on the inability for users to understand and to make use of options within the given time	Make it Privacy-by-design by default (indications point to Privacy Checkup working) Use of visual, clear, actionable (symbols etc.) Establish simpler rules for engagement
Cookie Notice	Make the user aware of cookies being used	Compliance; functionally the service works. No real option in refusing it	Make it visible, understandable and actionable
Privacy Settings – Account based	Make the user aware of information captured within account-based login interactions	Monetising data	Need transparency and visibility of data
Privacy Settings – Browser based	Make the user aware of information captured within a browser / session.	Monetising data	Need transparency and visibility of data
Policy Notice	Communicate personal data collected	Is freely changeable, hard to read and complex	Should communicate in a understandable format what personal data is necessary to be released for an interaction to take place.

However, many scholars offer dissenting views on the core principles of technological determinism. There are two key flaws with technological determinism, namely *unilinear progress,* i.e. technological progress does not follow a linear progressive path but instead a host of similar technologies emerge from which some are selected based on social needs and norms, and *determinisation by the base* where social reforms including technology-based initiatives are mandated by a governments or a trusted public body

in shaping society's acceptance as opposed to the society simply adapting to the technology (Feenberg 2009).

The forms through which technology mediates user interactions are forms of connectedness. Users find meaning in socio-technical interactions through the value these connections enable. When asymmetries occur such as subliminal tracking as opposed to feedback, privacy expectations are not met.

Explicating Privacy Expectations

From the survey findings, privacy expectations of users are impacted by factors such as network effects, passive acceptance, technical literacy, the 'creep factor' and contextual dependence regarding data release. Evidence of *network effects* was seen with users preferring the leading search engine provider and related products. It is most profitable for users to invest engagement in services that provide them with convenient, efficient and reliable search results. These user incentives create more user interactions and drive advertising revenue; this in turn allows the dominant player to establish monopolies in markets, giving them the opportunity to invest in resources (both human and infrastructure) and maintain the cost of the service (i.e. for search engines, the cost of 'free' search). It also allows these power players to have dominance over the market where they are able to influence market trends and user behaviour, which, when habituated over time, heavily influence social norms. The impacts of data release in search are far-reaching and driven by ease of access to data across linked services.

The power wielded by these general public utilities is substantial (Naughton 2016). From the survey findings, it is evidently becoming increasingly difficult for users to freely associate online. Often an algorithm predetermines their interactions. Although users appreciate the value in interacting with a service that has immense computational efficiency, there are increasingly fewer ways to choose freely, creating an illusion of choice. A decision to opt out means to not engage, and often the cost of not engaging is significant enough to deter users from leaving a platform. This tension was clearly visible in the survey findings, a phenomenon termed as *passive acceptance.* The argument that the lack of affordance for such control is a breach of contextual integrity still stands because these affordances are heavily influenced by technology forces, creating power and information asymmetries that perpetuate cycles of passive acceptance. Another deterrent is the high degree of specialist knowledge required to implement privacy-enabling measures. People are aware that tracking in general occurs online but may not be aware of the specific mechanisms in which data release and collection occurs or understand how to implement these privacy-enhancing measures. Ambivalent views occurred when users were aware

(informed) and no action (happy to share) taken or misinformed or they were aware and were able to take proactive action. Generally, when unconcerned, users express no perceived impact to them.

The privacy decisions users make are termed as trade-offs with numerous studies exploring the nature and the extent to which users are willing to release personal data for benefits (financial or otherwise). However, humans are cognitively limited in their ability to make complex decisions such as the decisions required in online privacy (Ariely 2008). Low levels of *technical literacy* and aliteracy are seen (Naughton 2012)as key inhibiting factors to adoption of privacy-enabling technologies. Technical literacy concerns the levels of awareness and understanding around privacy risks associated with data flows. Aliteracy occurs when users have the ability to understand but are not interested or find it difficult to convey their expectations due to the complexity in implementing and using privacy-enabling mechanisms. Both these factors play a role in restricting the abilities of users in making these complex privacy decisions. For example, a survey of Australian consumers reported that users found it difficult to distinguish the provenance and operation of different elements of a search engine results page, particularly with integrated vertical search services and less so with distinguishing ads from search results (Scardamaglia & Daly 2016).

Most users experienced ineffective affordances when trying to exercise privacy. These cases of inaction were due to limitations in the affordances experienced by users. *The 'Creep' factor* is in play driven by ineffective affordances. Further, when communicating transparency through data used in interactions, it is not enough to simply disclose details about the data to users. Techniques used to disclose privacy information in recommender systems have been shown to decrease user trust and satisfaction (Knijnenburg & Kobsa 2013). To constructively impact consumer behaviour around online self-disclosure of personally-identifiable information, suggestions call for changes to enabling mechanisms underpinning corporate policy and regulation of technology organisations and incorporating privacy protections in the roles of key decision-makers in data ecosystems (Chartered Institute of Marketing UK 2016; Pomfret & Previte 2014). It is also important for users to be able to independently verify if these actions have been performed. As most users lack the technical literacy to verify these actions themselves and are limited by the design of the system, a role exists for an independent agent to perform these trust-based verification measures, similar to the concept of a certification authority. Embedding these verification measures will enable users to form trust-based interactions.

There is evidence of the *contextual dependence of privacy expectations* reported in related work (Acquisti, Brandimarte & Loewenstein 2015; Martin & Nissenbaum Forthcoming). Privacy-sensitivity around data flows is influenced by the nature of the information being collected, shared and stored; with strong indications for its use constrained to the primary actors and purposes users socially contract with. Further contexts are constrained temporally (short-term interests) or locally (obtaining locally relevant information for directions on a map). The appropriateness of data flows are constrained to the privacy-sensitive norms governing these contexts. The value exchange is constrained to these contexts and is predominantly transactional. The current blanket approach of harvesting data signals for recommendations is not preferred by users. Fine-grained controls are needed to control visibility of information to honour these constraints users expect when their data is shared.

6.3 Designing for Socio-Technical Interactions in Search: Contextually-Public Release of Data

Interactions are situated in social contexts (Nissenbaum 2016). When users interact with technology, their interactions are situated in social contexts underpinned by a degree of connectedness. This section explores these social contexts further. Efforts to afford users privacy protection can take many different approaches given the diversity in how people perceive privacy.

One plausible approach is presented in the form of design recommendations. This approach is preventative and pre-empts action. To ensure contextual integrity, socio-technical interactions should first enable appropriate flows of data by providing affordances that people can use to exercise privacy. Affordances can take the form of technical capabilities designed into technology and the ability of users to perform actions that shape their privacy practices.

However, a weakness of private and public spheres is that it does not fully explain the nature of socio-technical interactions, where data release is both actively determined by the user and passively released through interacting with the medium. The protection mechanisms afforded to users are largely functional as opposed to being a practice and may not be reachable to users due to pre-requisite knowledge, expertise and effort constraints.

6.3.1 Emergence of a Contextually-Public Sphere

From the technical and empirical investigations, there is evidence to support the emergence of a contextually public sphere. This builds on the contextual integrity theory by explicating the nature of

privacy-sensitivity in this sphere, where evidence supports practices such as passive acceptance, or contextual value of transactions where users want to share data but have in place protection mechanisms. This socio-technical medium has the ability to track, record, and release data without inducing user awareness and understanding, as it simply happens through searching. In this shaping movement users have become passive (Shirky 2010). The social capital in the online medium is a lot less. It serves mostly materialistic needs and needs to be aligned with humanistic needs. The ability to exchange information about others or stay in social contact is a primal need. Social media for example facilitates that, but its current manifestation has weaknesses. From real-life neighbours, social contact has evolved to online 'friends', people can lurk looking at data leakage and not be seen, given the incentives to monetise click-through data.

Due to its decentralised nature and use of social production, the networked information economy shifts reliance from a mass-mediated public sphere to a networked public sphere (Benkler 2006). These innovations enable greater freedoms for individual actions. Social media taps into this primal need to stay in touch across physical boundaries in this globalised world. Data flows through a contextually public space as its meaning lies in contexts. Data is private and when released it is considered public. Efforts should focus on release of data because it provides the minimum source of leakage and the most effective control in creating that freedom to associate between separate interactions. Personal data is constitutive, being framed as being part of one's self, whereas it could also be data about a person (Floridi 2016). From the empirical findings data can be determined to be both constitutive and be descriptive, given that data is released through the course of an interaction. A visualisation of these data flows by the author is presented in Figure 6.4.

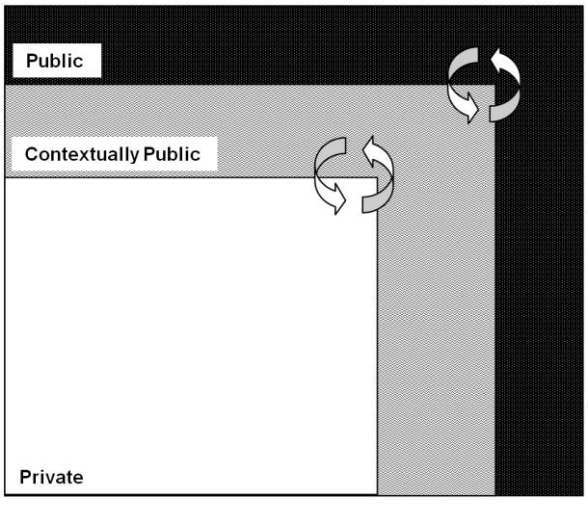


Figure 6.4 Nature of Data Flows

In this emerging contextually public sphere, one vital aspect that governs appropriate flows is the right for users to freely manage associations online. “Our [human] dignity rests in being able to be the masters of our own journeys, and keep our identities and our choices open.” (Floridi 2016). It is not about equating privacy with the notion that there should be nothing to hide, or on access and control to information, but about building interactions based on mutual respect for purposes, boundaries and context of personal data use that users may have and that service providers may need (Nissenbaum 2015). These contextual dependencies add value and meaning to otherwise seemingly transactional interactions. These interactions take place in an open decentralised network architecture that advocates for permissionless innovation (Naughton 2012) . However, openness does not mean entitlement to collection, processing and dissemination of data without limits. Further, the range of choices offered is framed by the provider as seen when embedded in defaults and existing alternate privacy-enabling mechanisms largely unreachable to users. It is clear from these privacy expectations that there is inherent value in online interactions, but this value is constrained by boundaries, purposes and contexts.

The nature of power and information asymmetries underpins online user interactions. Together with network effects, these black-box induced interactions create power asymmetries also known as surveillance capitalism and produce new power structures called siren servers (Lanier 2014; Naughton 2012; Pasquale 2015; Zuboff 2015). However, “the creepiness is not in the tech, but in the power we grant to siren servers” where “a siren server…is an elite computer or coordinated collection of computers, on a network.” (Lanier 2014, p.49;279). The agency of the user is severely limited in expressions of privacy in existing interactions as a result.

From the survey findings, it is clear that tensions in information asymmetries stem largely from ineffective affordances. In other words, if privacy-enabling properties are not designed into a product, it is very difficult for users to exercise privacy because the embedded capabilities are inadequate. Privacy values and design principles are enabling factors that guide expression of online societal norms. Mediating factors need to be taken into consideration when designing technology that proactively empower users to effectively manage personal data flows in search. Given the complexity of the medium, the basis for interaction must be visible and constrained to its contextual purpose.

When designing technologies in addition to the key functional criteria such as requirements, defaults and usability (for user convenience), designers should adapt appropriate contextually-relevant strategies that address the privacy sensitivities of users and which are aligned with online societal norms around privacy risks. For example, designers of recommender systems need to refresh profiles to reflect the

temporal interests of users. Personalised search or ad recommendations fall short of meeting user expectations when user profiles are not intuitive enough to reflect a user's changing interests. A user's worldview is changed by their experiences, which is a fluid exercise in discovering new interests, while some existing interests may lose its relevance to users. Profiling inhibits the ability to reflect this societal characteristic as the primary aim of profiling is to capture and collect as much information as possible to increase profiling accuracy. When this occurs, privacy expectations may not be met as users could have long moved on from the interests captured in the user profile. Because expectations are conditioned by the affordances offered to users, this disconnect in expectations builds a case for designing affordances that address the privacy sensitivities of users, as it is in upholding these societal norms that users find value in their interactions. Design should allow for negotiating the spectrum of privacy expectations that are deeply integrated into social life and not solely be driven by commercial motives.

"It is important to keep in mind that privacy norms do not merely protect individuals; they play a crucial role in sustaining social institutions." (Nissenbaum 2011, p.44). The institution in this case, is the Internet and the technology services and capabilities it enables. User interactions should not be actions that occur in commercial marketplaces, but as a value exchange on the Internet. For the institution to be sustainable, it should meaningfully cater to social norms. These empirical insights contribute important perspectives in understanding the intricate socio-technical relationship between privacy and technology (Palen & Dourish 2003).

These observations give rise to a contextually public sphere where users actively choose to share information for a perceived benefit for primary purpose of data collection and among actors. This idea extends the existing notion of private and public spheres for personal data flows, because evidence supports that users find immense value with releasing or sharing data, but is constrained to a context-dependent interaction. This data release may also differ conditionally within the same context due to changes in user expectations and behaviour.

In this image of a public reading place in Figure 6.5, passers-by can be seen to be reading in public, but details of the activity like the content of the book are not immediately visible. This type of interaction has its own properties. It is similar to socio-technical interactions where user actions determine properties, but the interaction is also shaped by the properties imbued by the medium. This is defined as a contextually-public sphere where interactions have some properties which are public by nature, with other properties that can be determined by the user. Web searching is an example of such an interaction as a lot of its properties are determined by the technology.



Figure 6.5 Visualising Contextually-Public Spheres: A Public Reading Place (Author's own image)

A key matter at stake for users in socio-technical interactions is the right to associate freely, and being able to manage those associations is important to users as reported from the survey data. The right to freely manage associations online is analogous to the freedom of privacy online. Users make implicit agreements pertaining to its appropriateness that underpin data flows. Framing of privacy choices and decision points at which users engage give us ideas of likely possibilities or scenarios in which privacy preferences may be less appealing in order to receive the benefits of interacting.

The inferences deduced from these interactions link back to a profile. However, the profile is not a complete representation of a person's identity. It is not holistic, but a set of scattered or routine iterations that inform a particular view or preference. Users on the other hand, act in a particular capacity during each interaction. This distinction is akin to Hannah Arendt's notion of relational selves, which is concerned with an individual's "humanness in a hyperconnected era, i.e. that they are truly interdependent from the mutual recognition that they grant to each other and that their humanity is precisely grounded in that mutual recognition, much more than in any "objective" difference or criteria that would allow an expert system to sort out human from non-human entities." (Dewandre 2015 p.8).

Although the interaction medium is contextually-public, private spaces are still required to sustain human dignity (Floridi 2016). Privacy in such spaces is not asocial or anti-social because it is central to

the creation and development of a person's being by affording people the ability to freely exercise their choices without every interaction being monitored and analysed (Steeves 2009).

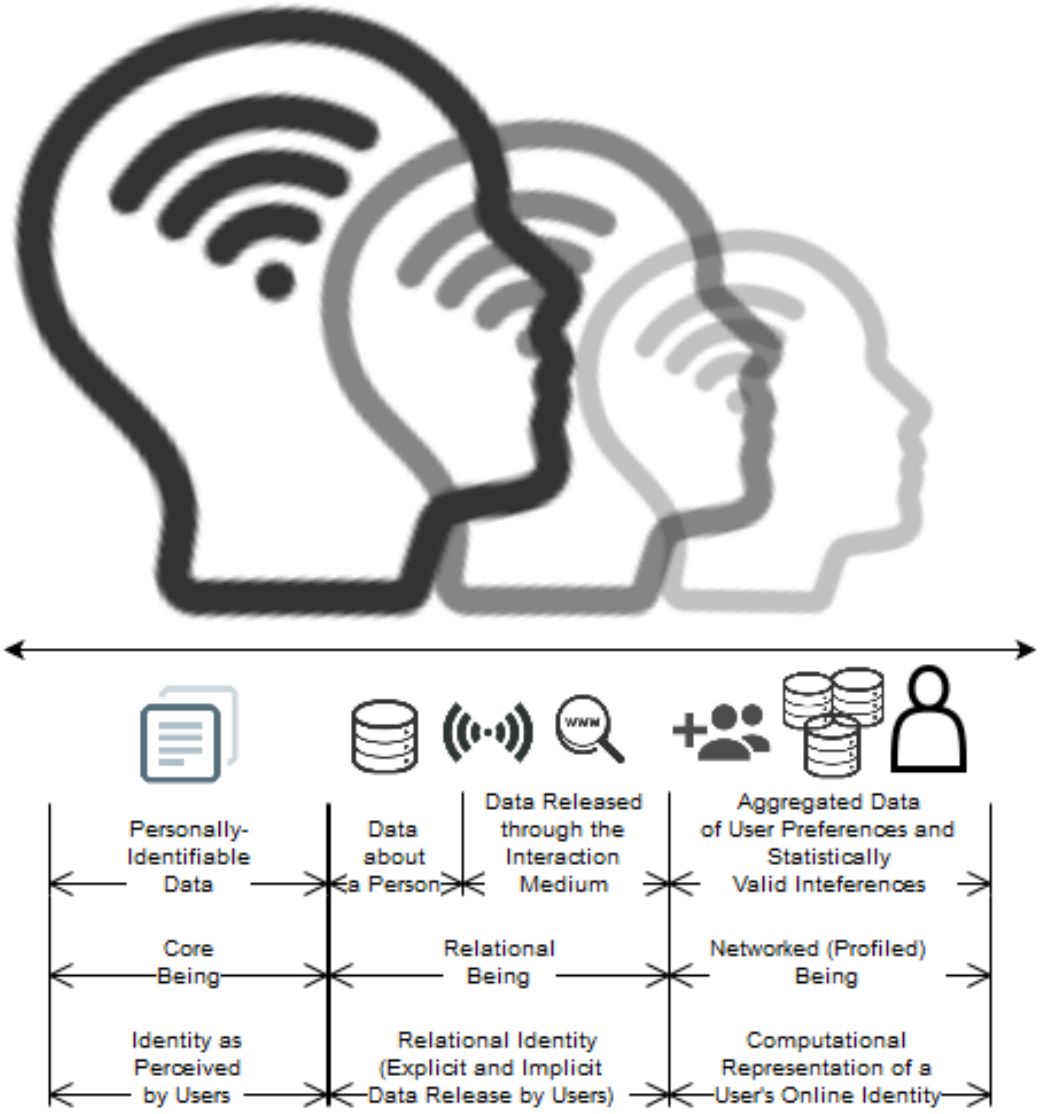


Figure 6.6 Progression of Identities in Socio-Technical Interactions

Private spaces online would facilitate the development of online identities that are not profiled and tracked. Humans are becoming networked beings where online profiles depict us as hyperconnected humans as visualised by the author in Figure 6.6. The dangers of hyperconnected selves include over-reliance to capture every interaction, in some cases at the risk of experiencing it first-hand. This over-reliance is habituated by users given the commercial incentives to monetise interactions by information service providers. Individuals become their recorded selves. Their experiences become less meaningful because they do not freely determine them. Regulation of reputation, as determined by networked

selves, is possible. If interactions allow for associations to be made, then interactions should also allow for associations to be managed. The boundaries between different ways of being are blurred.

In contextually public spheres, interactions are situated in socio-technical context through roles, instead of identities. Users primarily view this socio-technical interaction as role-based. They interact in a capacity, not with their whole self. They view release of data to be constrained contextually to that role's purpose. When data is explicitly released by the user, the medium adds to it by implicitly release data associated with each interaction. The permanent nature of this data release calls for the interaction to be contextually constrained. A blanket terms of service model while permissible legally does not adequately fit user expectations (they are only aware of release during the interaction) and neither does it fit the service provider's need to reuse.

6.3.2 Establishing Norms for Online Engagement

Social norms describe interactions that are appropriate among groups of people (Bicchieri & Muldoon 2014). Social norms can relate to commonly- held beliefs and expectations or identifiable patterns of behaviour (Bicchieri & Muldoon 2014). Social norms act as underlying implicit rules governing how societies function, simplifying the vast diversity of potential human activities through formation of routines and creation of habits (Bruce 1999).

Social frameworks like legal regulations are formalised from this shared identity and intention towards achieving common goals. Societal structures allow for human beings to play different roles as they participate in social life (Bruce 1999; Miller 2014). Norms are prescribed to these roles. Values are "what a person or group of people consider important in life" (Simpson and Weiner 1989 as cited in Friedman, B, Kahn & Borning 2009, p.70). These mechanisms of norms, roles and values over time create societal culture (Bruce 1999).

Social institutions are "a complex of positions, roles, norms and values lodged in particular types of social structures and organising relatively stable patterns of human activity with respect to fundamental problems in producing life-sustaining resources, in reproducing individuals, and in sustaining viable societal structures within a given environment" (Turner 1997, p.6). The Internet, given its importance as a general public utility, is a social institution that acts as a primary mechanism for practices like information-seeking. Norms underpin practices like data flows in search. Informational norms according to the contextual integrity framework comprise of contexts, actors, attributes and transmission principles (Nissenbaum 2010). A detailed analysis of social norms derived from privacy expectations expressed by

the survey study participants was conducted in Chapter 5. These derived social norms are discussed below.

In deriving social norms, information practices were analysed to identify commonly-held privacy expectations and how these beliefs formed appropriate socio-technical interactions.

6.3.2.1 Norms of Appropriateness and Distribution

Norms according to the contextual integrity theory are prescriptive and aim to identify privacy beliefs that are socially acceptable as opposed to descriptive norms that identify normal practices (Nissenbaum 2010). Norms of appropriateness and distribution associated with information privacy pertain to what is appropriate and are determined on what is relevant and to whom should it be distributed and when users interact with a service, they interact with a purpose (Nissenbaum 2010; Sar & Al-Saggaf 2014; Zimmer 2008)

When searching online, only the primary provider (consistent with their key purpose) is visible, while profile data may be shared to primary provider's business partners. From the empirical survey data, the level of appropriateness and distribution considered normal is influenced by the choices afforded to users. In these instances, passive acceptance is generated. This leads to normalising expectations by algorithmic governance, where passive acceptance becomes a habituated norm (Dourish & Anderson 2006; Naughton 2016b; Tufekci 2014).

6.3.2.2 Norms of Motivation

Users are habituated to norms that motivate key selling points like convenience, efficiency and speed. Convenience has become a necessity. Users associate efficiency and speed with online interactions. Design should be about giving users meaningful choices. Instead user behaviour is driven by business models that optimise the collection of click-through data. For example, the hook model uses four mechanisms (trigger, action, variable reward and investment) to capture and monetise user interactions (Eyal 2014). Measures of relevance as through click-through data lose meaning outside of monetary motivations.

6.3.2.3 Norms of Value

The disconnect between the social value of privacy and the market value around technology innovation triggers a third norm, where the market value take precedence over the social value of privacy. Traditionally, research has tailored privacy alternatives on top of market needs. In this instance, the social value of privacy will not hold in a market exchange because it is not premised on social values

(Ariely 2009). For example, standardised testing introduced mechanisms of measuring student performance in education. However, it has become the de-facto measure of student learning, where the incentive is to perform better on tests as opposed to creating a stable culture focused on learning where performances achieving good learning outcomes will naturally emerge (Ariely 2009). Both students and schools are measured against this indicator, which influences their roles in a learning context.

Privacy is social construct (Steeves 2009). The effects of privacy, whether positive or negative, impact human beings, not technology. Social cohesiveness is lost in these market exchanges. Ideally, interactions would progress from being transactional to transformative where users obtain its full value. For example, social media resembles transactional media. Masquerading friendship by the amount of likes a photo gets or the traction a post gets only shows a quantitative dimension of popularity, while the qualitative sentiments are lost. It should not be the only measure of relevance, to avoid risking or losing that richness or depth in interactions, as it would become artificial. When users provide inaccurate information to avoid privacy loss, this falsification of data is a dis-benefit to organisations because data quality diminishes too.

The market value stems from commercial motivations. The cost of interactions is considered free by users, largely influenced by the ethos stemming from academic activities that kick-started the development of the internet. Being on the internet was viewed as a social exchange. Established information providers use advertising to drive network effects, reduce operating costs and enable greater efficiency. The market wants to commodify privacy; whereas users see services as being free on the Internet.

Technology is constantly under innovation which brings in new functionality. Although it is indeed convenient, expectations are not met when users are not conscious of being informed of new data use and the exposure of data thought to be private. The norm here is for technology to innovate and push functionality out to the market, users then encounter the new use, and then become habituated to it. It is unclear if the lack of alternative choices shape the forming of new habits when using new technologies.

The value of data lies in aggregation from both points of view. Statistically-significant inferences are consulted to guide interactions. It is not profitable to solely target an individual person's preferences and habits with higher degrees of precision. Interactions need to distinguish between use of personal and aggregated data. A person's identity is intertwined with their experiences. Therefore, when using role-based identities, users should be provided protection in interactions.

Is there a place for privacy in a medium driven by market norms? For example, the CFS (Country Fire Service[9]) organize themselves around a community purpose for safety and well-being. Closer to home, in open source software development many amateur and expert programming enthusiasts spent countless hours problem-solving and helping each other out. These communities are widely popular and sustain themselves without any significant market incentive. In contrast, the products and services offered to consumers on the Internet, by its very design is incomplete. An open source business opportunity currently exists to disrupt the market norm for privacy protections afforded to users. For example, this need is viable in standalone compact products or services, where design principles can be validated. To obtain objective evaluations, these products or services must not be in a market where network effects can be felt because these effects create strong attachments that users have which may influence those online activities.

6.3.3 Designing for Proactive Privacy Protections

The following recommendations for designers are based on preventative actions to take when releasing data. Enabling users to freely interact and determine their associations online in relational interactions is the key organising principle. By embedding privacy protections in interactions, users are then afforded the ability to exercise their right to privacy and maintain an inviolate personality. By designing for proactive and meaningful choice when releasing data, users are afforded the ability to practice privacy online.

6.3.3.1 Necessity as a Transmission Principle in Relational Interactions

Like confidentiality, necessity is an important transmission principle that should underpin interactions. A lot of data flows leak data that are not essential to the service being provided. When users release data, they seem themselves interacting with a primary service provider. Designing user interfaces that incorporate only what is necessary (which may not be the same as minimalistic) is key to creating interactions based on trust and authenticity. This overcomes the limitations of the data minimisation principle, because it includes data that is necessary for the functioning of a transaction and data that could also be needed to validate a transaction. Although the user performs the interaction with the intent of receiving a service, feedback often is separate from the original interaction with continuous efforts to commercialise interactions.

[9] Acknowledgements to Kirsten Wahlstrom for the suggestion.

6.3.3.2 Detaching Self: Establishing Personal Spaces Online

From the empirical survey findings, one important aspect of privacy is the freedom for individuals to associate online, to the extent to which this is possible and the capability to manage associations. A device such as a phone is valuable to a user is because they think it is personal to them and may represent an extension of their self. The right to an inviolate personality relates to the ability to be anonymous. To be anonymous online is to be unreachable (Nissenbaum 1999). Given the permanence of data collection, the capabilities in identifying users for targeting and tracking, the ability to be anonymous is threatened.

One approach is to decouple constitutive and persistence data from data needed for role-based interactions. For example, decoupling a person's name from their online identities. Often a person's name forms an integral part of their identity, which coupled with a few other personal characteristics make them uniquely or personally-identifiable individuals. By using a number instead of a name to authenticate and validate transactions, interactions can still occur while ensuring the trustworthiness of both parties. This use of a number creates a personal space in a contextually public sphere where a user wants to release some data to obtain an information service. Although transactions can occur as is, meaning is still protected as an individual's name is detached from these online interactions.

This principle breaks the immediate linkability between a person and their associations, where data released is constrained to the context of the service engaged and enables users to detach their whole selves from their interactions. The conventional norm has been to profile excessively. A profile appears to be more constitutive than a name because a name is a label, but a profile contains details about you. However, by breaking the linkability, if the number assigned to each new session is random and does not persist, then profiling itself will become less detailed, role-based and in alignment with user expectations. This would change personal data that act as assigned identifiers (i.e. data that is reliably and persistently attached to one's identity) to opaque identifiers (i.e. data that reliably identifies a person, but is not constitutive) (Nissenbaum 1999). Using opaque identifiers would allow interactions to be authenticated where necessary and anonymous if desired.

While some data is required to authenticate users, there should be modes for 'cash-based' interactions online, where users are authenticated without risking their whole identity. The advantage in designing interactions is that it places the responsibility on the user, who is able to make that determination at an individual level. Typically, information service providers have been unwilling to take responsibility for

user actions as they have argued they are merely general-purpose technologies that enable transactions. For example, this is often provided as a defence in legal cases that challenge the claims of auto-correct search results. However, users currently have no real mechanisms to make this determination as an individual, because although users inherit responsibilities, rights are not reciprocated through data release mechanisms. For example, the EU Cookie Notice and DNT both which are not effectively enforced technically to uphold its protection aims.

Given the inherent value in linking user data financially, disassociating identities will probably not appeal to commercial information service providers as it is insufficient to disrupt existing business models of advertising. However, a growing body of related work explores this dissociation such as work on centralizing personal data and releasing it for authorization at the user end (examples include the concept of a databox and Solid in a Web 3.0 architecture). These concepts reflect role-based interactions where there is separation of their core being from the relational or networked being. Similar examples include biometric data that are not stored as-is, but only via a code containing the unique attributes. The personal attributes are not stored, deliberately creating a buffer zone to protect data that is constitutive (Floridi 2016).

Advances in machine learning can pick up and discern noise. However, although discernible patterns can be inferred, plausible deniability that can be built as a pattern may not be strongly tied to something uniquely and personally identifiable as a person's name. A person's name although it can be changed, is not easy to keep changing and the complexities of real-life associations add to this effect. A number however does not have these strongly personal properties and can be interchanged or refreshed updated. Anonymity is being unreachable (Nissenbaum 1999). In this age of connectedness, assigned and opaque identifiers can help with being unreachable.

6.3.3.3 Shedding Data: Designing for Relational Beings

Users perceive the benefits of data release such as recommendations, contextually. This context is constrained across temporal or location specific interests. User profiling inhibits the creation of these contexts. User profiling is based on the collection of extensive data about an individual in order to extract useful trends and insights, which are often useful to the collectors rather than the user. However, recommender systems need to include a measure of relevance from a user's perspective, which does not necessarily call for excessive data collection. From a user's perspective, these profiles need to be refreshed when interests are no longer relevant and its usefulness in terms of value expires.

6.3.4 Initiatives to Scaffold these Design Recommendations

When designing for role-based interactions to enable data release to be contextually-public, the above recommendations need to be scaffolded by practices that sustain them. Central to this is empowering users to perceive the visibility of subliminal targeting and tracking of data, so that the responsibility and the choice to determine release of personal data rests with the user.

6.3.4.1 User Empowerment by Perceiving Visibility

To scaffold these design recommendations, lasting change needs to be accompanied with empowerment at the grassroots level. Like teaching children about safety, privacy basics should be part of a school curriculum and conversations at home. Like occupational health and safety, employees should be educated on the privacy risks of data release in an institutional environment. Existing resources such as data protection information available with privacy commissioners such as the Office of the Australian Privacy Information Commissioner should be leveraged and communicated more effectively to the general public. Programs like Web literacy initiatives such as the open Web initiative by Mozilla and initiatives by privacy advocates create catalysts for change at the grassroots level. The aim of this is to make each individual responsible for their privacy in the same way that they take responsibility for their safety.

Most program initiatives focus on specific applications. However, a holistic focus is needed exploring the implication of data release through the data lifecycle stages of creation, collection, processing and dissemination. Given the subliminality of data tracking and the pervasiveness of profiling users, people should be made aware to recognise when interactions are influenced to freely determine their choices. This information on privacy protection needs to be communicated.

6.3.4.2 Stakeholder Empowerment by Aligning Incentives

Initiatives on aligning stakeholder incentives to afford users privacy protection are explored in this section.

Accountable Data Protection

The current focus on procedural rights are limited in terms of data protection as these rights primarily deal with decision-making processes (Kyriakakis 2016). From empirical studies on the privacy vs. utility trade-off, users cannot realistically make complex decisions when releasing their data and the hidden nature of the medium, i.e. the Internet compounds this complexity (Acquisiti et al. 2015; Naughton 2012). Data protection laws (including the APPs) are procedural and informed by Westin's findings on

access and control (Westin 1968). However, the nature of privacy is such that knowing a fact about another party is sufficient for that party's privacy rights to be at risk.

Nissenbaum calls for a substantive approach to privacy protection to address this inherent weakness in procedural-based data protection regimes (2016). Further, existing decision-making processes of users are largely influenced by the types of affordances offered to users, which in turn are currently determined by information providers. Given the findings that identified privacy as a practice, substantive data protection regulation should focus not just on data collection, but on release of data. Policy-makers can re-align existing data protection policies to explore the mechanisms in which data is released and the underlying privacy affordances embedded in the design of the socio-technical infrastructure. Regulating data released and used in interactions is akin to regulating the design of vehicles. This focus would shift data protection initiatives from a reactive regime looking at data leakage (for example, the collection and use of cookies) to the design of data release in interactions. Given that technology mediates this user interaction, regulating the design of data release mechanisms would ensure accountability is afforded to the user via socio-technical capabilities. Substantive data protection would look at data release in its original context, in particular the purpose, actors, location and time based on their findings. It is also important to regulate data release, given that most information providers offer a service utilising the internet as a general-public utility, where interactions are mediated by this socio-technical medium and the actions of individuals are impacted (Naughton 2016).

Given that most information providers deal with reputation of users and ranking of data, calls for fair data protection practices focus on the depth, scope and timing in the disclosure of data (Pasquale 2015). However, transparency and control are not sufficient in terms of regulation (Dewandre 2015). Accountability is key. The need to verify data released arises because normal expressions of trust are not reciprocated in socio-technical interactions. Users cannot build trust with an information provider who is invisible and complex (Pasquale 2015). Currently data protection efforts focus on collection, processing and dissemination of data. A weakness of these data protection mechanisms is that it is difficult to enforce (e.g. EU cookie law and DNT are not technically enforceable). Mechanisms should not just be limited to data collection or use. Instead by having mechanisms that verify data being released, the responsibility to manage data release rests with user who is then equipped to make those decisions.

Using Social Capital as an Incentive

Information-based service providers are known as general public utilities as they provide a vital part of the infrastructure underpinning the information economy (Naughton 2016). However, they also act as power structures in a monopolistic data market. These are private commercial companies driven mostly by the need to generate profits (as is normal with commercial companies). However, unlike other power structures that are also monopolies such as governments, these commercial companies are not directly or strongly influenced by the general public. The general public shapes the development of technology products and services primarily as consumers. These organisations are not elected by the public and, being commercial, serving the public good is not often a key part of their actioned agenda. In the current situation, these factors reduce the value exchange flowing on to users. Users are often passive consumers whose interaction with technology services and products create information asymmetries.

"The current bias in conceiving of the Net as a predominantly commercial enterprise seriously limits the privacy agenda" (Nissenbaum 2011 p. 32). It is a power structure that controls how technology is developed and the types of capabilities offered to users in exercising their privacy expectations. To sustain this social institution (the internet which is a young institution and its applications) in the future, social capital needs to come into play; aligning corporate commercial interests with user's social interests. Perhaps enough of a critical mass to affect social capital that in turn sees an impact on the public branding of the organization would be effective. Democratic freedoms of individuals are impinged upon by normalizing the array of choices afforded to them. That is dangerous because individuals will lose the ability to freely distinguish between what is morally acceptable and not. "In a democratic society, it is essential we continue to be able to watch the watchers. We must also maintain control of the controllers. Otherwise, a future of data exploitation will chill autonomy and diminish dignity." (Privacy International 2015).

6.4 Conclusion

In this chapter, findings from a conceptual investigation on the nature of socio-technical privacy protection mechanisms afforded to users is presented. This conceptual investigation consisted of a stakeholder values analysis and a moral mediation analysis. Building on these conceptual findings, a set of design recommendations to enable users to effectively manage data release when searching is presented. An analysis of the relationship between people and technology is presented. Users perceive privacy as a practice. Given that the nature of interactions is socio-technical, affordances should allow

for users to exercise their right to privacy. The right to determine and freely choose becomes redundant if the mechanisms are ineffective in addressing people's privacy expectations.

The freedom to choose is influenced by the choices offered to the user. There is inherent value in the right to associate freely online. Aggregation and linkability enable these inferences to be determined based on computations, given the market incentives. People perceive that access to information is provided to the information service provider they are interacting with. Clearly the model of notice and consent or informed consent through privacy policies has not translated well into the online world. Passive acceptance indicates that people are getting used to the idea that data they put out on the Internet is public and will plausibly be on the Internet forever. The norm for online interaction is shifting from something that is inherently private (i.e. being seen in your room) to something that is contextually public (e.g. the notion of being seen in the street) – people can see you and you can see them, so there is visibility and accountability for actions taking place in the context, but you still hold some sense of private space as you navigate through a busy space.

The motivations behind pervasive data collection are to build computational intelligence to better serve consumers. But they do not scale in the current system. Computational intelligence is fast and efficient, allowing us to grasp and access a large breadth of previously inaccessible resources. However, people create meaning through their experiences which form part of their human intelligence. Humans are not solely machines that feel or think, but are primarily feeling machines that think (Damasio 2006). Designers should realign their incentives and the values afforded to individuals in technology use. Instead of competing with the technology, use technology to liberate us. There is value in co-existing, in harnessing the intrinsic value of our human beings, of our artefacts and our interactions. What is currently at stake for users is the inability to effectively control release of data.

In the future, what is at stake will be their experiences, particularly when individuals do not have the ability to freely create them of their own free will. This will create new information and power asymmetries without the appropriate mechanisms to mitigate the ensuing privacy risks. Foucault's view of technological mediation explored that "humans are not only objects of power, but also subjects that create their own existence against the background of and in construction with these structures" (Verbeek 2011 p.18). In this hyperconnected age, his views can then be reinterpreted as 'what kind of humans do we want to be?'.

7 Conclusion

Data plays a central role in socio-technical interactions online. This research investigated the extent to which existing technological affordances for managing data release are effective in addressing people's privacy expectations when searching. The value sensitive design approach was used as the overall methodology to understand the nature of the problem from technical, empirical and conceptual dimensions.

Within these dimensions, the research questions of this study were to identify:

- **RQ1:** What is the effect of data release on the variance of personalised search results? *(Technical Dimension)*
- **RQ2:** How effective are existing privacy affordances in addressing the privacy expectations of users when searching? *(Conceptual Dimension)*
- **RQ3:** What privacy expectations do users have about the flow of data in search? *(Empirical Dimension)*
- **RQ4:** How can we proactively empower users to effectively manage data release in search? *(Conceptual Dimension)*

7.1 Key Findings, Contributions and Applicability

Key findings discovered from each investigation have been reported along with the primary contributions. The flipped-order approach in applying value sensitive design as a method is a methodological contribution as explained in Section 3.1 - Approach. Usually the order in which value sensitive design is applied consists of either empirical or conceptual investigations leading to re-designed artefacts as a result of technical investigations. Value sensitive design is applied in a flipped-order in this thesis because of the nature of the problem. There is a lack of understanding regarding the extent to which existing affordances effectively address privacy expectations associated with data release when searching. The application of the value sensitive method in this flipped-order is a methodological because application of this approach enabled data release to be studied from a socio-technical perspective and to present recommendations informed by analysis of data from technical and empirical investigations.

7.1.1 Technical Dimension: Investigating the Effects of Data Release

Personal data released when searching is used to produce personalised search results for users. Search engine providers argue that personal data produces more relevant results. However, releasing personal data does not necessarily lead to more relevant results from a user's point of view. Survey respondents reported expectations were largely not met when data was used for behavioural advertising purposes and if a perceived benefit was not visible. For example, when asked if expectations were met when personally identifiable information was used to tailor and present relevant search results that better aligned with their needs. 39.7% of respondents said that it does not really affect them, while more than half of the respondents were not satisfied (36.3% were not satisfied and 15% were somewhat unsatisfied respectively see Table 5.3). A set of controlled experiments was designed to determine if personalised search results can be obtained with limited release of data. Of particular concern was the effect of implicit data, which is typically released without a user's awareness or understanding and potentially may impact their privacy concerns. The controlled experiment study was conducted on seven different search engines (both commercial and privacy-enhanced), sourcing 29 popular query terms across three timeframes over two years (2014 – 2016). The experiment involved, firstly, discovering data transferred in typical search scenarios by using an HTTP intercepting proxy, and then secondly, investigating its effects on the variance of personalised search results using Edit Distance and the Jaccard Index as measures.

Overall, through rigorous statistical analysis, the variance of personalised search results was consistently affected mostly by location, the user-agent parameter, topicality of the query term (popular informational terms), and query terms where the user intent was transactional or navigational. These had significant impact on personalisation variance across both commercial and privacy-enhanced search engines over these timeframes. For some commercial search engines, cookies and parameters carrying personally-identifiable data such as the referrer header had a minimal, but inconsistent impact. For privacy-sensitive users, these findings contribute towards an opportunity to proactively control the distinguishability of their queries by obfuscating the data released in queries. This proactive measure allows the privacy risks to be contained. Minimising excessive data leakage does not adversely impact the variance of personalised search results. In the few instances where minimised obfuscation is not possible, this data release should be made visible to users, so that they are informed in determining their privacy choices.

7.1.2 Empirical Dimension: Investigating Privacy Expectations

After analysing the synthesised privacy expectations empirical data using Nissenbaum's contextual integrity framework, three information privacy practices stemming from privacy expectations associated with data release when searching were identified. The practice of newly-introduced metadata regime practice in Australia enables data collection of internet search histories of users (see section 5.4.1 Practice 1: Collecting and Retaining Metadata). This practice introduces a privacy risk of less autonomy as it is traded off for societal security. It could be improved however, by establishing a basis for independently verifying surveillance oversight measures, which reinforces integrity and trust in these practices.

The second practice identified explored collecting and retaining personal data for commercial purposes. This practice was concerned with the appropriateness of data flows (i.e. what constitutes appropriate flows of information) and the distributions effects of data released (i.e. actors who have access to and visibility of these data flows).

Both of these practices introduce privacy risks arising from the lack of awareness and understanding by users, of the extent and appropriateness of access and use of data released. Privacy-sensitive users who are aware of this data collection currently self-censor, use technologies to minimise or obfuscate the data released or choose to passively accept data collection as the new normal. Users who are not aware of this data collection, report evidence of the 'creep' factor at play, once the effects of targeting and tracking are made visible to them. The presumption here is that users should be able to take steps to protect their information should they choose to do so. However, this assumes that users should put in the effort to learn and understand the complexities of data collection and use.

The third practice identified explored the framing of choices offered through ineffective affordances, creating significant privacy risks., These risks can be increased through network effects of being locked in to commonly-used technology and the lack of choice offered to users in adapting their privacy expectations consistent with online social norms.

7.1.3 Conceptual Dimension: Understanding Mechanisms for Data Release in Search that Afford Users Privacy

The review on data release mechanisms identified gaps users have in exercising their privacy. Current technical affordances like privacy-enhancing technologies are siloed, as opposed to being holistically integrated within searching. These mechanisms assume that users are aware of and have an adequate

knowledge of the privacy risks concerned and often cater to a functional purpose. Given the commercial incentives involved in collecting and monetising data, these technologies are often subject to operating by the terms of its particular distribution platform and minor changes by search engine providers can often break or render its purposes less meaningful.

Social affordances often are not technically enforceable like DNT and the cookie notice, which although they capture and express important privacy information, do not make such expectations actionable. Typical consent mechanisms are ineffective in enabling users to make informed decisions. For example, end-user licence agreements and privacy policies are often ambiguously worded and do not reflect the socio-technical nature of the interaction, as it is modelled on existing social contracts. Privacy rights afforded by law concentrate primarily on access to data, which is at times at odds given that users want to share their data to obtain specific benefits.

Innovations in technology give rise to new capabilities that did not previously exist. The old information asymmetries that constrained access to and consumption of information have been broken. However, new information and power asymmetries are created in the process. The mechanisms currently afforded to users in protecting their privacy although embodying moral values are in most cases fragmented and not formally enforceable.

After a philosophically-informed analysis of the key values embedded in socio-technical interactions and a mediation analysis of its effects, the nature of data release is determined to be contextually-public. Users expect data to be constrained to primary actors and purposes that underpin these interactions. They perceive value in releasing data and find interactions meaningful, but its value is contextually constrained. Given the interaction medium, i.e. the Internet, Web browsers and other technologies that govern the flow of data, the social contract of engaging with an information service provider is public by nature of the technological properties at play. However, release and propagation of data is subliminal, given the relative invisibility of the extent of targeting and tracking. Technological affordances are imbued with purpose from users, but also by designers, which frames the range of privacy capabilities and choices offered to users. Over time, this framing of choices and capabilities leads to conditioning of privacy beliefs and habituating of online social norms. Often, the commercial market incentives of monetising data and interactions drive the development of technology. However, privacy is centred on social values. In an environment driven by market norms and values, social values cannot be sustained as the monetary gains outweigh and influence the range of protections offered to users. Hence, if

interactions are ways of being, then the current privacy-sensitive affordances offered to users are not natural extensions of ourselves.

This key contribution of a concept for privacy expectations: data release occurring in a *contextually-public sphere* has important implications for design of technology. Users expect their data to be contextually-public, i.e. where data release is constrained to primary purposes and actors concerned. Designers should design with necessity as a key transmission principle, where only necessary data is collected. Interactions are relational where it involves decoupling a real-world identity from the profiled identity. constrained to the roles at play. This relational view of interactions provide agency to users as it enables a hyperconnected user to detach their role-based interactions from their core identity. Also the value to users of releasing data arises from interests that may change with time or locality. Excessive targeting and tracking curtails this value for users. Therefore, profiles may need to shed data as opposed to extensive collection of data. For these design recommendations to be sustained, users need to be empowered to action change. Awareness and knowledge of information privacy should be communicated and be made part of the shared understanding that underpins civic values and social cultures. The key aim in this regard is to enable users to take responsibility for their privacy, as they do for their own personal safety.

Data protection needs to be substantive by regulating data release and the design of the mechanisms that enable data release in its interaction context. Accountability needs to be ensured by enabling verification of data release and protection. This reciprocates trust between the user and information service provider in socio-technical interactions. Similar data protection initiatives such as privacy by design, data minimisation, data portability, the right to access data, the right to be forgotten are mandated by the newly introduced General Data Protection Regulation (GDPR) by the European Union (European Union 2016). This work which predates the introduction of the GDPR validates the need and significance of implementing such measures, bearing in mind the European Union is leading the way in terms of data protection and viewing privacy as a human right when compared to initiatives of other governments. However, this research's findings contribute an understanding of privacy expectations associated with data when searching. In particular a case is made for contextually-public interactions as opposed to viewing search as a public or private interaction. Further, role-based identities are advocated for based on evidence, because in interactions people engage not with their entire identity or self, but within a particular role. These contributions extend and if applied can add value to existing legislative initiatives.

7.2 Future Work

Potential directions for future research into extending the work on the effectiveness of privacy protections afforded to users when searching are described in this section.

7.2.1 Extending the Applicability of the Controlled Experiment Findings on Data Release in Personalised Search

To further extend the applicability of the technical investigation findings, the controlled experiment on the effects of data release on personalisation variance in search results among users can be scaled to overcome current limitations. Most implicit data released (likely to be personally-identifiable) has a negligible passive impact on personalisation variance. If this implicit data is to be removed or obfuscated, personalised search results can be still be retrieved. Future work needs to examine the extent of usability of personalised search results with suppressed or obfuscated data release, and the level of trade-off for value and choice among user needs and privacy expectations around data release.

This could be achieved by a controlled experiment among across-section of heterogeneous expert users to obtain user-based evaluations on the effects of implicit data release on personalisation variance. One experimental condition would be to investigate if users perceive any difference with personalised search results through a longer timeframe with the effects of obfuscating search queries for location and type of user-agent could be determined where the controls would include minimal and default query release data being investigated.

Another approach would be to examine the effectiveness of four privacy intervention methods with varying degrees of data release in meeting satisfaction of users pertaining to the personalised search results, user experience and privacy-sensitivity. Hence, different privacy intervention scenarios can be set up, where users can be asked to carry out the same information search task across the four specific privacy intervention scenarios of clean query (fully parameter free), a default (control group) query option, two forms of user-controlled data release pertaining to location-spoofing and browser-spoofing. Think-aloud protocols can be used to analyse user satisfaction concerning privacy and personalisation during the search task in each scenario. Post-task interview questions will also be used to assess user satisfaction of the user experience, affordances and privacy sensitivity. Rich data insight into understanding user intentions and behaviour around privacy-sensitivity, user experience and affordances can be gathered. Content analysis can be performed to substantiate understanding of derived insights and patterns. Such a study would investigate if users would be happy to trade-off leaking these kinds of

personally-identifiable data to obtain relevant personalised search results or if they would be happy with personalised search results with known personally-identifiable data removed, to uncover the nuances that determine release such as query intent and other conditional factors.

7.2.2 Extending the Applicability of the Empirical Survey Study on Privacy Expectations associated with Data Release when Searching

By administering the survey in different environments such as general public populations or different national cultures, the applicability of the survey findings can be extended. Further, temporal stability can be built in using measures such as test-retest reliability, indicating the reliability of the scale and construct tested over time. Further research could extend the current limitations of the survey instrument by also expanding the range of privacy constructs tested. The survey instrument could be applied to understand and uncover privacy expectations in different domains such as social media use or the effects of algorithmic decision-making on individuals.

7.2.3 Extending the Applicability of the Conceptual Findings and Design Recommendations

The applicability of the conceptual findings can be extended by further understanding the effects of privacy-sensitive affordances in search. Affordances are a vehicle to embody privacy values. In information asymmetries, a user's need for information should be balanced with a competing need for privacy of personal data. The differences users perceive with privacy-sensitive affordances when searching can be examined, to determine the effects of contextually-public data release. By building a proof-of-concept embedding these privacy-sensitive affordances, further understanding on how to effectively align and embed privacy values in socio-technical affordances and provide users with proactive choices around personal data release can be uncovered.

7.2.4 Validating Privacy Expectations and Online Social Norms

It is important to qualitatively validate the privacy expectations and online social norms gathered in this research. Interviews with a sample of 25 users from this study are planned as a future project extending this work. Using grounded theory, the aim is to further explore the nature of privacy expectations and data flows to understand the nuances of the values at play. The effects of subliminal data release could be explored. A plausible outcome is to uncover a design pattern for practising privacy in the context of data release. A design thinking approach may be employed to rapidly prototype and investigate its effects. This could lead to future work in building this design pattern into technology such as a mobile

app with a tightly defined function. The effects of feedforward, perceived affordances and feedback in establishing reciprocal socio-technical interactions can also be investigated. Also, online social norms for searching could also be verified through crowdsourced survey techniques or via a sociological study using ethnographic techniques. These applications would enable generalisability to wider social cultures and demographics.

7.2.5 Determining Privacy-Sensitivity in Search

Building on the findings of the technical controlled experiment on identifying the effects of data release in personalised search, efforts could be made to determine the privacy sensitivity of search queries based on a user's context and search behaviour. The key aim is to derive a set of rules or generate a model that outlines the dominant factors that affect the privacy-sensitivity of users in typical search scenarios. The collected survey data (Chapter 5) contains data on specific search scenarios (tested and validated in Chapter 4) and a decision on whether this search scenario is sensitive or not to them. By building a feature pool of candidate search terms across the search scenario, features could be analysed to obtain dominant features. Using a supervised learning technique, a set of rules and models could generate from the data set. This extension would help determine the sensitivity of search terms in scenarios by harnessing the strength of machine learning techniques whilst incorporating user behaviour and contextual search data from the survey.

These efforts could then lead to a client-side recommendation tool that would nudge users when privacy-sensitive data is to be released and reinforces meaningful learning and accountability with users. Using empirical methods, a diverse range of opinions or the 'what' and uncover the 'why' behind effectiveness of privacy protection mechanisms. This in turn, will enable the development of effective data release or suppression techniques that enable people to manage their identifiability in relation to search results. A plausible outcome could be to create client-side mechanisms to manage contextually-public interactions by being able to detach a user's core identity from their relational or networked identity.

7.2.6 Communicating Online Privacy

There is a significant need to communicate the complexity of online privacy to people. Most participants reported being aware that tracking occurs, but not being proactively empowered to manage data release. Tools like Lightbeam help visualise the extent of data release when searching. However, these tools need to be adopted in an environment which encourages and sustains participatory public discussion and

where individuals have the ability to critically reason the information they consume and determine choices in line with their expectations.

Legislative privacy efforts like GDPR tackle the breadth of the problem, whereas grassroots efforts can affect the depth of the problem by shaping behavioural practices. Large regional initiatives like the GDPR help advance privacy efforts across the breadth of services offered to users by information service providers and are focused more inducing behaviour change of both providers and users from a legislative compliance angle. At the grassroots level to bridge the gap between understanding and action, the work of local government councils can be extended and informed by scholarly academic research through initiatives that communicate the importance of online privacy to the general public, similar to cyber-bullying and road safety initiative. Likewise, the work of start-up privacy groups like Hack for Privacy or CryptoParty Sydney, or privacy advocacy groups like the Australian Privacy Foundation could be extended to reach and inform mainstream audiences. These grassroots efforts can induce change across the depth of experiences and practices (such as privacy expectations associated with data flows when searching) over time as they are efforts that can proactively affect and shape social norms from a behavioural angle. These grassroots efforts help bring the voices of users and their privacy expectations to the table instead of it occurring reactively through instances of privacy loss. Privacy needs a healthy participative democracy for privacy rights mandated in laws such as the Australian Privacy Principles and the GDPR to continue to exist.

7.3 Conclusion

In conclusion, this research investigated the effects of data release on personalised search results, the privacy expectations people have during the search and the effectiveness of technological affordances in managing data release. The effects of data leakage when searching can be reduced by minimising the data released or through obfuscation. In terms of privacy expectations, users report evidence of passive acceptance of choices, the ‘creep’ factor and dislike of excessive targeting and tracking of users.

A concept for privacy expectations in online socio-technical interactions is derived from the empirical and technical findings and presented. Data release is found to be contextually-public in nature, where design needs to cater to relational selves in reciprocal interactions. Privacy is perceived as a practice by users. Users find value in data release. They also expect to be able to freely associate online. However, the current affordances do not offer users the ability to exercise their privacy views. To reduce the effects of these information and power asymmetries, technology should be designed incorporating the

requirements for data release in a contextually-public sphere, where both users and information service providers benefit from the value of releasing data.

> "An Orwellian world is much easier to recognize, and to oppose than a Huxleyan… Everything in our background has prepared us to know and resist a prison when the gates begin to close around us…[but] who is prepared to take arms against a sea of amusements?" (Postman, A 2017; Postman, N 1986).

These amusements have a tendency to make individuals too comfortable, where norms shaped by technology heavily influence behaviour, conditioning beliefs like the 'creep' factor and habituating norms like passive acceptance. The consequence of not providing users with adequate privacy protection is to lessen their ability to distinguish between a worldview that arises out of their own inviolate personality and one that is increasingly predetermined and created for them. To meaningfully afford users privacy protection is to provide them with the ability to change these norms and their accustomed behaviour. The relationship users have with technology could be transformed from being transactional by affording users privacy protection in the design of technology and ensuring that users have the ability to freely associate online.

References

Acar, G, Eubank, C, Englehardt, S, Juarez, M, Narayanan, A & Diaz, C 2014, 'The web never forgets: Persistent tracking mechanisms in the wild', *21st ACM Conference on Computer and Communications Security*, ACM, pp. 1–16. <https://securehomes.esat.kuleuven.be/~gacar/persistent/the_web_never_forgets.pdf >.

Acquisti, A & Grossklags, J 2005, 'Privacy and rationality in individual decision making', *IEEE Security & Privacy*, vol. 3, no. 1, pp. 26-33.

Acquisti, A, Brandimarte, L & Loewenstein, G 2015, 'Privacy and human behavior in the age of information', *Science*, vol. 347, no. 6221, pp. 509-514.

Adobe 2013, *Analytics and on-site personalization services - Adobe - Privacy Policy*, viewed 7 September 2013, <http://www.adobe.com/privacy/analytics.html>.

Advanced Web Ranking 2014, *Google organic CTR study*, Moz.com Blog, viewed 19 April 2016, <https://moz.com/blog/google-organic-click-through-rates-in-2014>.

Allen-Castellito, A 1999, 'Coercing privacy', *William and Mary Law Review*, vol. 40, no. 1, pp. 723-757.

Aljifri, H & Navarro, D 2004, 'Search engines and privacy', *Computers and Security*, vol. 23, no. 5, pp. 379-388.

Amazon 2013, *Amazon best sellers of 2013*, Amazon, viewed 19 October 2013, <http://www.amazon.com/gp/bestsellers/2013/books >.

Apple Inc. 2016, *About privacy and location services*, Apple Inc., viewed 1 September 2016, <https://support.apple.com/HT203033>.

Arendt, H 1958, *The human condition*, University of Chicago Press, Chicago.

Ariely, D 2008, *Predictably irrational: The hidden forces that shape our decisions*, Kindle edn, HarperCollins Publishers Ltd., New York, NY.

Arnott, C & Andrejevic, M 2012, *Internet privacy research*, Centre for Critical and Cultural Studies - University of Queensland

Ars-Technica 2013, *Yes, Internet: Your citizens still want their anonymity* viewed 7 September 2013, <http://arstechnica.com/business/2013/09/yes-internet-your-citizens-still-want-their-anonymity/>.

Ashman, H, Dagger, D, Brailsford, T, Goulding, J, O'Sullivan, D, Schmakeit, J-F & Wade, V 2012, 'HCI and the web', in A Sears & JA Jacko (eds), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* 3rd edn, CRC Press USA, pp. 565-588.

Attorney-General's Department of the Australian Government 2015, *Dataset*, Attorney-General's Department of the Australian Government, viewed 27 August 2016, <https://www.ag.gov.au/NationalSecurity/DataRetention/Documents/Dataset.pdf>.

Australian Bureau of Statistics 2011, *2011 census quickstats - people — cultural & language diversity: South australia* Australian Bureau of Statistics. <http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/quickstat/4?opendocument&navpos=220>.

Australian Internet Governance Forum 2015, *2015 Australian internet governance forum - plenary 2: Metadata retention in australia*, Australian Internet Governance Forum, viewed 6 October 2016, <https://www.youtube.com/watch?list=PLngt3M1CK_DyKFY0dfcQIaYRtDDHdDQ7R&v=ZvuakxsNuRM>.

Awad, NF & Krishnan, MS 2006, 'The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization', *MIS Quarterly*, vol. 30, no. 1, pp. 13-28.

Barbaro, M & Zeller, T 2006, 'A face is exposed for aol searcher no. 4417749', *New York Times*. viewed 4/08/2015, <http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=1&_r=0&ei=5070&en=6c5dfa2a9c1be4ec&ex=1155787200&emc=eta1 >.

Barbour, RS 2008, *Introducing qualitative research: A student's guide to the craft of doing qualitative research*, 1st edn, Sage Publications Ltd, Thousand Oaks, California.

BBC 2016, *Baidu apps are 'leaking' personal data*, BBC, viewed 27 February 2016, <http://www.bbc.com/news/technology-35669817>.

Baskerville RL & Wood-Harper AT 1996, 'A critical perspective on action research as a method for information system research', *Journal of Information Technology*, vol. 11, no. 3, pp. 235–246.

Benkler, Y 2006, *The wealth of networks: How social production transforms markets and freedom*, Yale University Press, New Haven, Connecticut.

Berendt, B & Teltzrow, M 2003, 'Addressing users' privacy concerns for improving personalization quality: Towards an integration of user studies and algorithm evaluation', in B Mobasher & SS Anand (eds), *IJCAI 2003 Workshop*, Springer-Verlag Berlin Heidelberg, pp. 69-88.

Berners-Lee, T 1989, *Tim Berners-Lee: Invention of the world wide web*, World Wide Web Consortium, viewed 1 March 2016, <https://www.w3.org/People/Berners-Lee/>.

Bicchieri, C & Muldoon, R 2014, *Social norms*.

Bing 2011, *Bing blogs: Adapting search to you*, Bing, viewed 12 April 2015, <http://blogs.bing.com/search/2011/09/14/adapting-search-to-you/>.

Bing 2013, *Bing blogs - top Bing searches for 2013*, Bing, viewed 19 October 2013, <http://blogs.bing.com/search/2013/12/01/top-bing-searches-for-2013/>.

Bloomberg 2010, *IAC's Barry Diller surrenders to Google, ends Ask.com's search effort*, Bloomberg, viewed 7 September 2013, <http://www.bloomberg.com/news/2010-11-09/iac-s-diller-surrenders-to-google-juggernaut-ends-ask-com-search-effort.html>.

Boyatzis, RE, 1998, *Transforming qualitative information: Thematic analysis and code development*, Sage Publications Ltd, Thousand Oaks, California.

Boyd, D 2014, *It's complicated: The social lives of networked teens*, Yale University Press New Haven, Connecticut.

Brajnik, G & Gabrielli, S 2010, 'A review of online advertising effects on the user experience', *International Journal of Human-Computer Interaction,* vol. 26, no. 10, pp. 971-997.

Broder, A 2002, 'A taxonomy of web search', *ACM SIGIR Forum,* vol. 36, no. 2, pp. 3-10.

Bruce, S 1999, *Sociology: A very short introduction* 1st edn, Oxford University Press Oxford, Great Britain.

Brunton, F & Nissenbaum, H 2015, *Obfuscation: A user's guide for privacy and protest*, The MIT Press, Cambridge, Massachusetts.

Bryman, A 2012, *Social research methods*, 4th edn, Oxford University Press, Oxford, New York.

Burnham, KP & Anderson, DR 2002, *Model selection and multimodel inference: A practical information-theoretic approach*, Springer-Verlag, New York.

Castelluccia, C, De Cristofaro, E & Perito, D 2010, 'Private information disclosure from web searches', *10th International Symposium on PETS 2010*, PETS 2010, pp. 38–55.

Castelluccia, C, Kaafar, M & Tran, M 2012, 'Betrayed by your ads ! Reconstructing user profiles from targeted ads', *12th International Symposium on PETS 2012*, PETS 2012, pp. 1-17.

Cavoukian, A 2009, *Privacy by design: The 7 foundational principles*, Information and Privacy Commissioner of Ontario, Canada, Information and Privacy Commissioner of Ontario, Canada.

Chartered Institute of Marketing UK 2016, *Whose data is it anyway?*, CIM, UK. <https://exchange.cim.co.uk/thought-leadership/whose-data-is-it-anyway/>.

Chaudhry, A, Crowcroft, J, Haddadi, H, Howard, H, Madhavapeddy, A & Mortier, R 2015, 'Personal data: Thinking inside the box', *Critical Alternatives*, pp. 29-32.

Chi, E 2015, *UMAP 2015 keynote speech: User modeling and the blurring of the boundary between interactive search and recommendation*, UMAP 2015 Conference, Dublin, Ireland, viewed July 3, 2015, <http://umap2015.com/speakers.html>.

Citron, DK 2014, *Hate crimes in cyberspace*, Harvard University Press, Cambridge, Massachusetts.

Clarke, R 1999, *Introduction to dataveillance and information privacy, definition of terms*, viewed 15 August 2014, <http://www.rogerclarke.com/DV/Intro.html>.

Clarke, R & Wigan, M 2011, 'You are where you've been: The privacy implications of location and tracking technologies ', *Journal of Location Based Services*, vol. 5, no. 3-4 (December 2011), pp. 138-155

Crawley, MJ 2013, *The R book*, John Wiley & Sons Ltd., Hoboken, New Jersey.

Creswell, JW 2009, *Research design: qualitative, quantitative, and mixed Methods approaches*, 3rd edn, Sage Publications Ltd, Thousand Oaks, California.

Creswell, JW & Plano Clark, VL 2011, *Designing and conducting mixed methods research*, 2nd edn, SAGE Publications, Thousand Oaks, California.

cURL Haxx 2013, *Curl and libcurl*, cURL Haxx, viewed 9 October 2013, <https://curl.haxx.se/>.

Damasio, A 2006, *Descartes' error: Emotion, reason and the human brain*, Kindle ebook edn, Vintage London, UK.

Danezis, G & Gurses, S 2010, 'A critical review of 10 years of privacy technology', *Surveillance Cultures: A Global Surveillance Society*, pp. 1-16.

Dewandre, N 2015a, 'Rethinking the human condition in a hyperconnected era: Why freedom is not about sovereignty but about beginnings', in L Floridi (ed), *The Onlife Manifesto: Being Human in a Hyperconnected Era*, Springer-Verlag Berlin Heidelberg Berlin, Germany, pp. 195-215. <http://dx.doi.org/10.1007/978-3-319-04093-6_20>

Dewandre, N 2015b, 'The human condition and the black box society: A review of frank pasquale, the black box society: The secret algorithms that control money and information ', *In boundary 2 : an international journal of literature and culture*, viewed 20 April 2016,<https://boundary2.org/2015/12/16/dewandre-on-pascal/>.

Dingle, S 2014, *Attorney-General George Brandis struggles to explain government's metadata proposal*, Australian Broadcasting Corporation, viewed 21 Nov 2016, <http://www.abc.net.au/news/2014-08-07/brandis-explanation-adds-confusion-to-metadata-proposal/5654186>.

Dorrestijn, S, van Der Voort, M & Verbeek, P-P 2014, 'Future user-product arrangements: Combining product impact and scenarios in design for multi age success', *Technological Forecasting & Social Change*, vol. 89, no. 2014, pp. 284-292.

Dourish, P & Anderson, K 2006, 'Collective information practice: Emploring privacy and security as social and cultural phenomena', *Human Computer Interaction*, vol. 21, no. 3, pp. 319-342.

DuckDuckGo 2015, *Duckduckgo privacy policy - we don't collect or share personal information* DuckDuckGo, viewed 24 April 2015, <https://duckduckgo.com/privacy>.

Earp, JB, Anton, AI, Aiman-Smith, L & Stufflebeam, WH 2005, 'Examining internet privacy policies within the context of user privacy values', *IEEE Transactions on Engineering Management*, vol. 52, no. 2, pp. 227-237.

Eckersley, P 2010, 'How unique Is your web browser?', in MJ Atallah & NJ Hopper (eds), *Privacy Enhancing Technologies: 10th International Symposium, PETS 2010*, Springer-Verlag Berlin Heidelberg, pp. 1-18. <https://doi.org/10.1007/978-3-642-14527-8_1>.

Electronic Frontier Foundation 2016, *Mandatory data retention*, Electronic Frontier Foundation, viewed 1 July 2016, <https://www.eff.org/issues/mandatory-data-retention>.

Ellenberg, J 2014, *What's even creepier than Target guessing that you're pregnant?*, Slate, viewed 8 June 2015, <http://www.slate.com/blogs/how_not_to_be_wrong/2014/06/09/big_data_what_s_even_creepier_than_target_guessing_that_you_re_pregnant.html>.

Ellis, M 2013, *Top 20 local search ranking factors: An illustrated guide*, Moz.com Blog, viewed 14 March 2014, <https://moz.com/blog/top-20-local-search-ranking-factors-an-illustrated-guide>.

Enge, E 2011, *How Google does personalization with jack menzel*, Stone Temple Consulting, viewed 19 October 2013, <http://www.stonetemple.com/how-google-does-personalization-with-jack-menzel >.

EUROPA 2016, *Cookie notice*, European Union's Web Portal, viewed 9 March 2016, <http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm>.

Directive 2009/136/ec of the European Parliament and of the Council.

Directive 95/46/ec *of the European Parliament and of the Council.*

Directive 2002/58 *of the European Parliament and of the Council.*

Directive 2006/24/ec *of the European Parliament and of the Council*.

European Commission 2015a, *Personal data definition*, European Commission, viewed 7 August 2015, <http://ec.europa.eu/justice/data-protection/individuals/index_en.htm >.

European Commission 2015b, *Special eurobarometer 431: Data protection* European Commission.

Eyal N 2014, *Hooked: How to build habit-forming products*, Penguin, London, England.

Farrell, P 2015, *'Governments do not like freedom of information': The war on australia's privacy and information watchdog* The Guardian, viewed 1 August 2016, <https://www.theguardian.com/australia-news/2015/oct/01/governments-do-not-like-freedom-of-information-the-war-on-australias-privacy-and-information-watchdog>.

Feenberg, A 2009, 'Democratic rationalization: Technology, power and freedom', in DM Kaplan (ed), *Readings in the Philosophy of Technology*, 2nd edn, Rowman and Littlefield Publishers, Lanham, Maryland, pp. 1-19.

Feenburg, A 2002, *Critical theory of technology*, Second Edition edn, Oxford University Press, Oxford, Great Britain.

Fernando, A, Du, T & Ashman, H 2014, 'Personalisation of web search: Exploring search query parameters and user information privacy implications – the case of Google', in S Luo & GH Yang (eds), *1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security*, ACM SIGIR, pp. 31-36.

Feuz, M, Fuller, M & Stalder, F 2011, 'Personal web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalisation', *First Monday*, vol. 16, no. 2, pp. 1-8.

Flanagan, M, Howe, D & Nissenbaum, H 2008, 'Embodying values in design: Theory and practice', in MJ van den Joven & J Weckert (eds), *Information Technology and Moral Philosophy*, Cambridge University Press, Cambridge, England, pp. 322-353.

Floridi, L 2006, 'Four challenges for a theory of informational privacy', *Ethics and Information Technology*, vol. 8, no. 3, pp. 109-119.

Floridi, L 2010, *Information a very short introduction*, Oxford University Press, Oxford.

Floridi, L 2016, *On human dignity as a foundation for the right to privacy*, Chefuturo, viewed 17 May 2016, <http://www.chefuturo.it/2016/05/dignity-foundation-right-to-privacy/>.

Friedman, Felten, E & Millett, LI 2000, *Informed consent online: A conceptual model and design principles*, CSE Technical Report.

Friedman, B & Kahn, PH 2003, 'Human values, ethics and design ', in A Sears & JA Jacko (eds), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* 2nd edn, Lawrence Erlbaum Associates, Inc., Publishers New Jersey pp. 1177-1201.

Friedman, B & Freier, N 2005, 'Value sensitive design', in KE Fisher, S Erdelez & LEF McKechnie (eds), *Theories of Information Behavior*, Information Today, Medford, New Jersey.

Friedman, B, Kahn, P & Borning, A 2009, 'Value sensitive design and information systems', in KE Himma & HT Tavani (eds), *The Handbook of Information and Computer Ethics*, John Wiley & Sons, Inc., Hoboken, New Jersey, pp. 69-101.

Friedman, B 2011, *Public lecture - The shape of being: Technology design, human values and the future* The University of Washington, viewed 14 August 2015, <https://www.youtube.com/watch?v=lnlRveMxEtQ>.

Friedman, B & Hendry, DG 2012, 'The envisoning cards: A toolkit for catalyzing humanistic and technical imaginations', in JA Konstan, EH Chi & K Hook (eds), *CHI 2012: The ACM SIGCHI Conference on Human Factors in Computing Systems* ACM CHI 2012, pp. 1145-1148.

Fuster, GG 2010, 'Inaccuracy as a privacy-enhancing tool', *Ethics and Information Technology*, vol. 12, no. 1, pp. 87-95.

Gavison, R 1980, 'Privacy and the limits of law', *The Yale Law Journal*, vol. 89, no. 3, pp. 421-471.

The General Data Protection Regulation (GDPR) (2016/679) (European Union)

Gerety, T 1977, 'Redefining privacy', *Harvard Civil Rights-Civil Liberties Law Review*, vol. 12, no. 2, pp. 233-296.

Gibson, J 1979, *The ecological approach to visual perception: Classic edition*, Psychology Press, Boston, Massachusetts.

Goldberg, I, Hill, A & Shostack, A 2001, 'Trust, ethics and privacy ', *Boston University Law Review*, vol. 81, no. 2, pp. 101-116.

Goodman, B & Flaxman, S 2016, 'European Union regulations on algorithmic decision-making and a "right to explanation"', 2016 *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, ICML Workshop on Human Interpretability in Machine Learning, pp. 1-9.

Google 2009a, *Google: Personalized search*, Google, viewed 17 June 2013, <https://www.youtube.com/watch?v=EKuG2M6R4VM>.

Google 2009b, *Eye-tracking studies: More than meets the eye* Google Official Blog, viewed 18 March 2016, <https://googleblog.blogspot.com.au/2009/02/eye-tracking-studies-more-than-meets.html>.

Google 2013, *Google trends - what did the world search for in 2013?* , Google, viewed 19 October 2013, <http://www.google.com/trends/topcharts#geo&date=2013>.

Google 2014a, *Search appliance documentation*, Google, viewed 9 March 2014, <http://www.google.com/support/enterprise/static/gsa/docs/admin/72/gsa_doc_set/xml_reference/request_format.html>.

Google 2014b, *Googel developers - XML API reference*, Google, viewed 9 October 2014, <https://developers.google.com/custom-search/docs/xml_results>.

Google 2015, *Search analytics help*, Google, viewed 6 February 2015, <https://support.google.com/analytics/answer/1012264?hl=en>.

Google 2016a, *Google dashboard: Personal info & privacy*, Google, viewed 18 February 2016, <https://myaccount.google.com/intro/privacy>.

Google 2016b, *View your data and account activity in the Google dashboard*, Google, viewed 9 March 2016, <https://support.google.com/accounts/answer/162744?source=gsearch&hl=en>.

Google 2017, *Google - privacy & terms: Privacy policy*, Google, viewed 9 March 2017, <http://www.google.com.au/policies/privacy/>.

Google inc v vidal-hall & ors (2015)

Google Operating System Blog 2006, *Meaning of parameters in a Google query*, Google Operating System Blog, viewed 9 March 2014, <http://googlesystem.blogspot.com.au/2006/07/meaning-of-parameters-in-google-query.html>.

Grimmelmann, J 2009, 'Saving Facebook (privacy on social network sites)', *Iowa Law Review*, vol. 94, no. 4, p. 1137.

Guba, EG & Lincoln, YS 2008, 'Competing paradigms in qualitative research ', in NK Denzin & YS Lincoln (eds), *The Landscape of Qualitative Research*, 3rd edn, Sage Publications, Los Angeles.

Gurses, S 2010, 'PETs and their users: a critical review of the potentials and limitations of the privacy as confidentiality paradigm', *Identity in the Information Society*, vol. 3, no. 3, pp. 539-563.

Hannak, A, Sapiezynski, P, Kakhki, AM, Krishnamurthy, B, Lazer, D, Mislove, A & Wilson, C 2013, 'Measuring personalisation of web search', in V Almeida, D Schwabe & H Glaser (eds), *World Wide Web 2013 Conference* IW3C2, pp. 527-538.

Heidegger, M 1977, *The question concerning technology and other essays*, Harper & Row, New York.

Horling, B & Kulick, M 2009, *Personalized search for everyone* Google, viewed 17 June 2013, <https://googleblog.blogspot.com.au/2009/12/personalized-search-for-everyone.html>.

Humby, D 2006, *Data defintion*, Dunn Humby, viewed 18 September 2015, <https://www.dunnhumby.com/ >.

IBM Corporation 2015, *Take control of your personal data with IBM identity mixer* IBM Corporation, viewed 3 March 2016, <https://idemixdemo.mybluemix.net/>.

Internet Engineering Task Force 2014, *Hypertext Transfer Protocol (HTTP/1.1): Semantics and content*. <http://tools.ietf.org/html/rfc7231>.

Ixquick 2015, *Startpage by Ixquick: Startpage protects your privacy*, Ixquick, viewed 24 April 2015, <https://ixquick.com/eng/protect-privacy.html>.

Jenkins, H 2010, *The Wall Street Journal: Google and the search for the future*, The Wall Street Journal, viewed 3 March 2016, <http://www.wsj.com/articles/SB10001424052748704901104575423294099527212>.

Jenkins, HW 2010, 'Google and the search for the future: The web icon's ceo on the mobile computing revolution, the future of newspapers, and privacy in the digital age', *The Wall Street Journal*. viewed 1 June 2014, <http://www.wsj.com/articles/SB10001424052748704901104575423294099527212>.

Kaptelinin, V 2013, *Affordances and design*, The Interaction Design Foundation, <https://www.interaction-design.org/>

Kaptelinin, V & Nardi, B 2012, 'Affordances in HCI: Toward a mediated action perspective', Proceedings *of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 967-976.

Kay, J & Kummerfeld, B 2006, 'Scrutability, user control and privacy for distributed personalization', in A Kobsa, RK Chellappa & S Spiekermann (eds), *CHI 2006 Workshop on Privacy-Enhanced Personalization*, CHI 2006, pp. 1-3.

Knijnenburg, B & Kobsa, A 2013, 'Making decisions about privacy: Information disclosure in context-aware recommender systems', *ACM Transactions on Interactive Intelligent Systems*, vol. 3, no. 3, pp. 1-33.

Kobsa, A 2007, 'Privacy-enhanced web personalization', in P Brusilovsky, A Kobsa & W Nejdl (eds), *The Adaptive Web*, Springer-Verlag Berlin Heidelberg, Berlin, Germany, pp. 628 – 670.

Kobsa, A, Knijnenburg, B & Livshits, B 2014, 'Let's do it at my place instead?: Attitudinal and behavioral study of privacy in client-side personalization', in M Jones et al (eds), *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 81-90.

Koene, A, Perez, E, Carter, CJ, Statache, R, Adolphs, S, O'Malley, C, Rodden, T & McAuley, D 2015, 'Privacy concerns arising from internet service personalization filters', *SIGCAS Computers & Society*, vol. 45, no. 3, Sept 2015, pp. 167-171.

Koops, B-J & Leenes, R 2014, 'Privacy regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data-protection law', *International Review of Law, Computers & Technology*, vol. 28, no. 2, pp. 159-171.

Kosta, E, Kalloniatis, C, Mitrou, L & Kavakli, E 2009, 'Search engines : Gateway to a new panopticon', in S Fischer-Hübner, C Lambrinoudakis & G Pernul (eds), *Trust, Privacy and Security in Digital Business.*, Springer-Verlag Berlin Heidelberg Berlin, Germany, pp. 11–21.

Krause, A & Horvitz, E 2010, 'A utility-theoretic approach to privacy in online services', *Journal of Artificial Intelligence Research*, vol. 39, no. 2010, pp. 633-662.

Krishnamurthy, B, Malandrino, D & Wills, C 2007, 'Measuring privacy loss and the impact of privacy protection in web browsing', *3rd SOUPS: Symposium on Usable Privacy and Security*, pp. 52-63.

Krishnamurthy, B & Wills, CE 2009, 'Privacy diffusion on the web : A longitudinal perspective', in WWW (ed), *18th International Conference on the World Wide Web*, ACM, pp. 541-550.

Krishnamurthy, B, Naryshkin, K & Wills, C 2011, 'Privacy leakage vs. Protection measures: The growing disconnect', *The Web 2.0 Security and Privacy Workshop*, pp. 1-10.

Kumaraguru, P & Cranor, LF 2005, *Privacy indexes: A survey of Westin's studies*, Carnegie Mellon University - Research Showcase at CMU. <http://repository.cmu.edu/isr/856/>.

Kyriakakis, J 2016, *Substantive data protection* Email correspondence - clarification on substantive data protection.

Laerd Statistics 2015, *Wilcoxon Mann-Whitney U test using SPSS statistics: Statistical tutorials and software guides*, Laerd Statistics, <https://statistics.laerd.com/>.

Lanier, J 2014, *Who owns the future?*, 1st edn, Penguin Books Ltd., London, England.

Lawrence, J & O'Donnell, L 2015, *European data retention laws: An update*, Electronic Frontiers Australia, viewed 1 August 2016, <https://www.efa.org.au/2015/07/29/european-data-retention-laws-update/>.

Lazar, J, Feng, J & Hochheiser, H 2010, *Research methods in human-computer interaction*, John Wiley & Sons., Hoboken, New Jersey.

Lessig, L 2008, *Remix: Making art and commerce thrive in the hybrid economy*, Penguin Press London, England.

Lewandowski, D 2012, 'New perspectives on web search engine research', in D Lewandowski & A Spink (eds), *Library and Information Science: Web Search Engine Research* Emerald Group Publishing Limited, Bradford, pp. 1-16.

Lincoln, A 2011, 'FYI TMI: Toward a holistic social theory of information overload', *First Monday,* vol. 16, no. 3, pp. 1-15.

Lincoln, YS & Guba, EG 1985, *Naturalistic inquiry*, Sage Publications, Beverly Hills, California.

Litwin, MS 1995, *How to measure survey reliability and validity*, Sage Publications, Thousand Oaks, California.

Luca, M, Wu, T, Couvidat, S, Frank, D & Seltzer, W 2015, *Does Google content degrade Google search? Experimental evidence*, Harvard Business School Harvard Business School Working Paper 16-035. <http://people.hbs.edu/mluca/SearchDegradation.pdf>.

Mac Aonghusa, P & Leith, DJ 2016, 'Don't let Google know I'm lonely!', *ACM Transactions on Privacy and Security*, vol. 19, no. 1(3), pp. 1-25.

MacGibbon, A & Phair, N 2012, *Privacy and the Internet: Australian attitudes towards privacy in the online environment*, Center for Internet Safety, Canberra.

Madden, M 2014, *Public perceptions of privacy and security in the post-Snowden era*, Pew Research Center. <http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/>.

Malandrino, D & Scarano, V 2013, 'Privacy leakage on the web: Diffusion and countermeasures', *Computer Networks*, vol. 57, no. 14, pp. 2833–2855.

Martin, KE & Nissenbaum, H Forthcoming, 'Measuring privacy: An empirical test using context to expose confounding variables ', *Columbia Science and Technology Law Review, Forthcoming*, pp. 1-40.

Mayer-Schönberger, V 2015, 'Connecting the dots', *Science*, vol. 347, no. 6221, p. 481.

Mayer, JR & Mitchell, JC 2012, 'Third-party web tracking: Policy and technology', *2012 IEEE Symposium on Security and Privacy*, IEEE, pp. 413–427.

Metcalfe, M 2007, 'Problem conceptualisation using idea networks', *Systemic Practice and Action Research*, vol. 20, no. 2, pp. 141-150.

Meyers, P 2015, *Google glossary: Revenge of Mega-SERP*, Moz.com Blog, viewed 15 March 2016, <https://moz.com/blog/google-glossary>.

Micarelli, A, Gasparetti, F, Sciarrone, F & Gauch, S 2007, 'Personalized search on the world wide web', in P Brusilovsky, A Kobsa & W Nejdl (eds), *The Adaptive Web*, Springer-Verlag Berlin Heidelberg Berlin, Germany, pp. 195-230.

Mill, JS 1885, *Principles of political economy abridged with critical, bibliographical, and explanatory notes, and a sketch of the history of political economy*, Project Gutenberg 2009 Ebook 30107 edn, D Appleton and Company, New York. <https://archive.org/details/principlesofpoli30107gut>

Miller, S 2014, *Social institutions*, Winter 2014 Edition edn, The Stanford Encyclopedia of Philosophy, viewed 1 February 2017, <https://plato.stanford.edu/archives/win2014/entries/social-institutions/>.

Moor, JH 1985, 'What is computer ethics?', *Metaphilosophy*, vol. 16, no. 4, pp. 266-275.

Moor, J 1990, 'Ethics of privacy protection ', *Library Trends*, vol. 39 no. 1 & 2, pp. 69-82.

Mortier, R, Haddadi, H, Henderson, T, McAuley, D & Crowcroft, J 2014, 'Human-data interaction: The human face of the data-driven society', CoRR, October 1, 2014.

Mozilla 2016, *Get smart on the web*, Mozilla, viewed 18 February 2016, <https://www.mozilla.org/en-US/teach/smarton/tracking/>.

Myers, J, Well, A & Lorch, R 2010, *Research design and statistical analysis*, 3rd edn, Routledge New York.

Narayanan, A & Shmatikov, V 2010, 'Privacy and security: Myths and fallacies of "personally identifiable information"', *Communications of the ACM*, vol. 53, no. 6, p. 24.

Naughton, J 2012, *What you really need to know about the Internet: From Gutenberg to Zuckerberg*, Quercus Great Britain.

Naughton, J 2016a, *Why the arrival, not the journey, matters*, Memex.Naugtons.org, viewed 3 June 2016, <http://memex.naughtons.org/archives/2016/05/15/23034>.

Naughton, J 2016b, 'The evolution of the internet: From military experiment to general purpose technology', *Journal of Cyber Policy*, vol. 1, no. 1, pp. 5-28.

Nelson, T 1960, *Project Xanadu*, Xanadu, viewed 1 September 2015, <http://xanadu.com/>.

NetMarketShare 2016a, *Desktop search engine market share*, NetMarketShare.com, viewed 25 April 2016, <https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0&qpsp=2015&qpnp=2&qptimeframe=Y>.

NetMarketShare 2016b, *Mobile/tablet search engine market share*, NetMarketShare.com, viewed 25 April 2016, <https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=1&qpsp=2015&qpnp=2&qptimeframe=Y>.

Neuendorf, KA 2002, *The content analysis guidebook*, Sage Publications, Thousand Oaks, California.

Nguyen, M-HC, Haynes, P, Maquire, S & Friedberg, J 2013, 'A user-centered approach to the data dilema: Context, architecture, policy', in M Hildebrandt, K Hara & M Waidner (eds), *Stand Alone: Digital Enlightenment Yearbook 2013 - the Value of Personal Data*, IOS Press, Burke.

Nippert-Eng, CE 2010, *Islands of privacy*, University of Chicago Press, Chicago.

Nissenbaum, H 1999, 'The meaning of anonymity in an information age', *The Information Society*, vol. 15, no. 2, pp. 141-144.

Nissenbaum, H 2010, *Privacy in context*, 1st edn, Stanford University Press, Stanford, California.

Nissenbaum, H 2011, 'A contextual approach to privacy online', *Daedalus*, vol. 140, no. 4, pp. 32-48.

Nissenbaum, H 2015, 'Respecting context to protect privacy: Why meaning matters', *Science and Engineering Ethics*, vol. 2015, no. July, pp. 1-22.

Nissenbaum, H 2016, *Public talk: Privacy at the intersection of ethics and technology* MCIR München, viewed 20 July 2016, <https://www.youtube.com/watch?v=Y2MdpfnZXt0>.

Norman, D 2013, *The design of everyday things*, Kindle: Revised and Expanded Edition edn, Basic Books, New York.

Nottingham, M 2016, *Ideal HTTP performance*, IETF HTTP Working Group, viewed 27 April 2016, <https://www.mnot.net/blog/2016/04/22/ideal-http>.

O'Dell, E 2015, *UMAP 2015 keynote speech: Privacy paradigm – towards a creative commons for privacy* UMAP 2015 Conference, Dublin, Ireland, viewed 2/07/2015, <http://umap2015.com/speakers.html >.

Office of the Australian Information Commissioner - Australian Government 2016, *OAIC-Office of the Australian Information Commissioner* Office of the Australian Information Commissioner - Australian Government, viewed 1 August 2016, <https://www.oaic.gov.au/>.

Office of the Australian Information Commissioner 2013, *Community attitudes to privacy survey*, Office of the Australian Information Commissioner - Australian Government, Canberra, ACT.

Ohm, P 2013, 'Branding privacy (companies' data handling policies)', *Minnesota Law Review*, vol. 97, no. 3, pp. 907-989.

Oreskovic, A & Rigby, B 2007, *Microsoft, Yahoo in 10-year web search partnership*, Reuters, viewed 7 September 2013, <http://www.reuters.com/article/2009/07/29/microsoft-yahoo-idUSN2921665320090729>.

Ortlieb, M 2014, 'The anthropologist's view on privacy', *IEEE Security & Privacy*, vol. 2014, no. May/June, pp. 85-87.

Palen, L & Dourish, P 2003, 'Unpacking 'privacy' for a networked world', *SIGCHI Conference on Human Factors in Computing Systems 2003*, CHI 2003, pp. 129-136.

Pallant, JF 2007, *SPSS survival manual : A step by step guide to data analysis using SPSS for Windows (version 15)*, 3rd edn, Allen & Unwin, Crows Nest, New South Wales.

Pan, B, Hembrooke, H, Joachims, T, Lorigo, L, Gay, G & Granka, L 2007, 'In Google we trust: Users' decisions on rank, position, and relevance', *Journal of Computer-Mediated Communication*, vol. 12, no. 3, pp. 801–823

Pariser, E 2011, *The opinion pages - when the internet thinks it knows you*, The New York Times, viewed 3 March 2016, <http://www.nytimes.com/2011/05/23/opinion/23pariser.html?_r=0>.

Pasquale, F 2015, *The black box society: The secret algorithms that control money and information* Harvard University Press, Cambridge, Massachusetts.

Pomfret, L & Previte, J 2014, 'Privacy behaviours: Protecting consumers from their friends and themselves', in S Rundle-Thiele, K Kubacki & D Arli (eds), *Australia and New Zealand Marketing Academy Conference 2014*, ANZMAC 2014, pp. 580-587.

Postman, A 2017, 'My dad predicted trump in 1985 – it's not orwell, he warned, it's brave new world ', *The Guardian* <https://www.theguardian.com/media/2017/feb/02/amusing-ourselves-to-death-neil-postman-trump-orwell-huxley>.

Postman, N 1986, *Amusing ourselves to death: Public discourse in the age of show business*, Heinemann, London.

Preibusch, S 2013a, 'The value of privacy in web search ', *12th Workshop on the Economics of Information Security (Weis 2013)*, pp. 1-14.

Preibusch, S 2013b, 'Guide to measuring privacy concern: Review of survey and observational instruments', *International Journal of Human-Computer Studies*, vol. 71, no. 12, pp. 1133-1143.

Preibusch, S 2015, 'Privacy behaviors after snowden', *Communications of the ACM*, vol. 58, no. 5, pp. 48-55.

PriceWaterhouseCoopers 2016, *IAB Internet advertising revenue report: 2016 first six months results - November 2016*, Internet Advertising Bureau, viewed 3 March 2016, <http://www.iab.com/wp-content/uploads/2016/04/IAB_Internet_Advertising_Revenue_Report_HY_2016___.pdf>.

Privacy International 2015, *Challenging data exploitation* Privacy International, viewed 3 March 2016, <https://www.privacyinternational.org/node/9>.

Rachels, J 1975, 'Why privacy is important', *Philosophy & Public Affairs*, vol. 4, no. 4, pp. 323-333.

Rainie, L & Duggan, M 2015, *Privacy and information sharing*, Pew Research Center. <http://www.pewinternet.org/2016/01/14/2016/Privacy-and-Information-Sharing/>.

Rainie, L 2016, *The state of privacy in America: What we learned*, PeW Research Center, US. <http://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/>.

Regulation of Investigatory Powers Act 2000 (United Kingdom)

Reiman, JH 1976, 'Privacy, intimacy, and personhood', *Philosophy & Public Affairs*, vol. 6, no. 1, pp. 26-44.

Rich, J 2016, *Federal Trade Commission: Keeping up with the online advertising industry*, Federal Trade Commission, viewed 3 May 2016, <https://www.ftc.gov/news-events/blogs/business-blog/2016/04/keeping-online-advertising-industry>.

Romanosky, S, Acquisti, A, Hong, J, Cranor, LF & Friedman, B 2006, 'Privacy patterns for online interactions', *2006 Conference on Pattern Languages of Programs - Plop '06*, ACM, New York, pp. 1-9.

Rosenthal, R & Rosnow, R 2008, *Essentials of behavioral research: Methods and data analysis*, McGraw-Hill, Boston, Massachusetts.

RStudio 2015, *RStudio*, RStudio.com, viewed 3 September 2015, <https://www.rstudio.com/>.

Rubinstein, I & Good, N 2013, 'Privacy by design: A counterfactual analysis of Google and Facebook privacy incidents', *Berkeley Technology Law Journal*, vol. 28, no. 2, p. 1333.

Saint-Jean, F, Johnson, A, Boneh, D & Feigenbaum, J 2007, 'Private web search', in P Ning & T Yu (eds), *2007 ACM Workshop on Privacy in Electronic Society*, ACM, pp. 84-90.

Sar, R & Al-Saggaf, Y 2014, 'Contextual integrity's decision heuristic and the tracking by social network sites', *Ethics and Information Technology*, vol. 16, no. 1, pp. 15-26.

Sauro, J & Lewis, JR 2012, *Quantifying the user experience: Practical statistics for user research*, Elsevier Inc., New York.

Scacchi, W 2004, 'Socio-technical design', *The encyclopedia of human-computer interaction*, vol. 1, pp. 656-659.

Scardamaglia, A & Daly, A 2016, 'Google, online search and consumer confusion in australia', *International Journal of Law and Information Technology*, vol. 24, no. 3, pp. 203-228.

Schneier, B 2015, 'Fear and convenience', in M Rotenberg, J Scott & J Horwitz (eds), *Privacy in the Modern Age: The Search for Solutions.*, Kindle edn, The New Press, New York.

Schuler, D. & Namioka, A 1993, *Participatory design: Principles & practices*, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Schwartz, PM 2004, 'Property, privacy, and personal data', *Harvard Law Review*, vol. 117, no. 7, pp. 2056-2128.

Search Engine Land 2016, *Chapter 7: Personalization & search engine rankings*, Search Engine Land, viewed 15 March 2016, <http://searchengineland.com/guide/seo/personalization-search-engine-rankings>.

Sengers, P, Boehner, K, David, S & Kaye, J 2005, *Reflective design*, 4th Decennial Conference on Critical Computing: Between Sense and Sensibility, Aarhus, Denmark, pp. 49-58.

Shirky, C 2010, *Cognitive surplus*, The Penguin Press, London, England.

Singhal, A 2011, *Google inside search: Some thoughts on personalization* Google viewed 17 June 2013, <https://search.googleblog.com/2011/11/some-thoughts-on-personalization.html>.

Solove, D 2006, 'A taxonomy of privacy', *University of Pennsylvania Law Review*, vol. 2006, pp. 477-564.

Solove, D 2007, "I've got nothing to hide' and other misunderstandings of privacy', *San Diego Law Review*, vol. 44, no. 1, pp. 745-772.

Solove, D 2009, *Understanding privacy*, Harvard University Press, Cambridge, Massachusetts.

Spiekermann, S & Korunovska, J 2016, 'Towards a value theory for personal data', *Journal of Information Technology*, vol. 32, no. 1, pp. 62-84.

StackExchange 2014, *Using quick-search with Startpage.com*, StackExchange, viewed 9 March 2014, <http://webapps.stackexchange.com/questions/63219/using-quick-search-with-startpage-com>.

StartPage 2014a, *Startpage protects your privacy*, StartPage, viewed 9 March 2014, <https://startpage.com/uk/protect-privacy.html>.

StartPage 2014b, *Startpage- about us*, StartPage, viewed 9 March 2014, <https://startpage.com/uk/what-makes-startpage-special.html?>.

Statista 2016, *Worldwide desktop market share of leading search engines from January 2010 to January 2016*, Statista, viewed 18 May 2016, <http://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/>.

Steeves, VM 2009, 'Reclaiming the social value of privacy ', in I Kerr, C Lucock & VM Steeves (eds), *Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society*, Oxford University Press, Cary, US.

Steichen, B, Ashman, H & Wade, V 2012, 'A comparative survey of personalised information retrieval and adaptive hypermedia techniques', *Information Processing and Management*, vol. 48, no. 2012, pp. 698–724.

Sullivan, C 2013, 'Digital identity, privacy and the right to identity in the united states of america', *Computer Law and Security Review: The International Journal of Technology and Practice*, vol. 29, no. 4, pp. 348-358.

Sullivan, D 2003, *Search privacy at Google & other search engines*, Search Engine Watch viewed 18 May 2016, <https://searchenginewatch.com/sew/study/2064327/search-privacy-at-google-other-search-engines>.

Sullivan, D 2009a, *Google now personalizes everyone's search results*, Search Engine Land, viewed 18 May 2013, <http://searchengineland.com/google-now-personalizes-everyones-search-results-31195>.

Sullivan, D 2009b, *Google's personalized results: The "new normal" that deserves extraordinary attention*, Search Engine Land, viewed 17 April 2013, <http://searchengineland.com/googles-personalized-results-the-new-normal-31290>.

Sullivan, D 2011a, *Bing results get localized & personalized*, Search Engine Land, viewed 12 April 2015, <http://searchengineland.com/bing-results-get-localized-personalized-64284>.

Sullivan, D 2011b, *Introducing: The periodic table of SEO ranking factors*, Search Engine Land, viewed 18 May 2013, <http://searchengineland.com/introducing-the-periodic-table-of-seo-ranking-factors-77181>.

Sullivan, D 2012, *Of "magic keywords" & flavors of personalized search at Google*, Search Engine Land, viewed 18 May 2013, <http://searchengineland.com/flavors-of-google-personalized-search-139286>.

Sullivan, D 2015, *FAQ: All about the new Google rankbrain algorithm*, Search Engine Land, viewed 12 March 2016, <http://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440>.

Sweeney, L 2002, 'K-anonymity: A model for protecting privacy', *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557-570.

Tavani, H 2005, 'Search engines, personal information and the problem of privacy in public'', *International Review of Information Ethics*, vol. 3, no. 1, pp. 39-45.

Tavani, H 2008, 'Informational privacy: Concepts, theories and controversies', in HTT Kenneth Einar Himma (ed), *The Handbook of Information and Computer Ethics*, John Wiley & Sons Inc., Hoboken, New Jersey, pp. 131-164.

Telecommunications (interception and access) amendment (data retention) act 2015 (cwlth)

Telerik 2014, *Fiddler : The free web debugging proxy for any browser, system or platform*, Telerik.com, viewed 18 May 2014, <http://www.telerik.com/fiddler>.

Teltzrow, M & Kobsa, A 2003, 'Impacts of user privacy preferences on personalized systems - a comparitive study', in C Karat, J Blom & J Karat (eds), *CHI 2003 Workshop - Designing Personalized User Experiences for Ecommerce: Theory, Methods and Research*, Kluwer Academic Publishers, Dordrecht, Netherlands pp. 1-11.

Tene, O 2008, 'What Google knows: Privacy and internet search engines', *Utah Law Review*, vol. 2008, no. 4, pp. 1433-1492

Tene, O & Polonetsky, J 2012, 'To track or "do not track": Advancing transparency and individual control in online behavioral advertising', *Minnesota Journal of Law, Science & Technology*, vol. 13, no. 1, pp. 281-357.

Tene, O & Polonetsky, J 2013, 'A theory of creepy: Technology, privacy and shifting social norms ', *Yale Journal of Law and Technology*, vol. 2013, no. 59, pp. 59-102.

The R Foundation 2015, *What is R?*, The R Foundation, viewed 3 September 2015, <https://www.r-project.org/about.html>.

Thomson, JJ 1975, 'The right to privacy', *Philosophy & Public Affairs*, vol. 4, no. 4, pp. 295-314.

Toch, E, Wang, Y & Cranor, LF 2012, 'Personalization and privacy: A survey of privacy risks and remedies in personalization based systems', *User Modelling and User-Adapted Interaction*, vol. 22, pp. 203-220.

Toubiana, V & Nissenbaum, H 2011, 'Analysis of Google logs retention policies', *Journal of Privacy and Confidentiality*, vol. 3, no. 1, pp. 3-26.

Toubiana, V, Subramanian, L & Nissenbaum, H 2011, 'Trackmenot: Enhancing the privacy of web search', CoRR,

Tsai, J, Cranor, L, Acquisti, A & Fong, C 2006, *What's it to you? A survey of online privacy concerns and risks*, Carnegie Mellon University US.

Tufekci, Z 2014, 'Engineering the public: Big data, surveillance and computational politics', *First Monday*, vol. 19, no. 7.

Turner, J 1997, *The institutional order*, Longman, New York.

Unbubble 2016a, *Derefer - for better privacy*, Unbubble, viewed 1 April 2016, <https://www.unbubble.eu/info/derefer>.

Unbubble 2016b, *Unbubble: Privacy & neutrality*, Unbubble, viewed 25 January 2016, <https://www.unbubble.eu/info/terms>.

Verbeek, P-P 2008, 'Morality in design: Design ethics and the morality of technological artifacts', in P Kroes et al (eds), *Philosophy and Design from Engineering to Architecture*, Springer Netherlands, Dordrecht.

Verbeek, P-P 2011, *Moralizing technology: Understanding and designing the morality of things*, University of Chicago Press, Chicago.

Verbeek, P-P 2016, *Public seminar: Ethics by design? On value sensitive design*, ADAPT Centre, Online viewed 30 August 2016, <https://www.youtube.com/watch?v=Eql0Qnrdte0>.

Victorian Government 2015, *Victorian community attitudes to privacy and data protection*, Victorian Government, viewed 1 July 2016, <http://www.vic.gov.au/news/victorian-community-attitudes-to-privacy-and-data-protection.html>.

Wailes, P 2008, *The ultimate guide to Google search parameters*, The You Moz Blog, viewed 1 July 2014, <http://moz.com/ugc/the-ultimate-guide-to-the-google-search-parameters>.

Walls, RJ, Clark, SS & Levine, BN 2011, 'Functional privacy or why cookies are better with milk', *7th USENIX Conference on Hot Topics in Security*, HotSec 2012, pp. 1-6.

Waters, N & Greenleaf, G, *A critique of Australia's proposed privacy amendment (enhancing privacy protection) Bill 2012 (August 15, 2012)*, UNSW Law Research Paper No. 2012-35, 19 October 2015, <https://ssrn.com/abstract=2134838>

Warren, SD & Brandeis, LD 1890, *The right to privacy* 2012 edn, University of Washington Press, Seattle, US. viewed 1 July 2015, <http://site.ebrary.com/lib/unisaau/docDetail.action?docID=10690526>

Watson, F 2011, *AOL serves Google search results. But is Microsoft doing their SEO?*, Search Engine Watch, viewed 2 September 2013, <http://searchenginewatch.com/article/2050043/AOL-Serves-Google-Search-Results.-But-Is-Microsoft-Doing-Their-SEO>.

WebMD 2013, *WebMD - 2012 year in health*, WebMD, viewed 19 October 2013, <http://www.webmd.com/news/year-in-health/2012>.

Wendt, T 2015, *Design for dasein: Understanding the design of experiences*, Kindle Edition edn, CreateSpace Independent Publishing Platform: On-Demand Publishing LLC, Scotts Valley, California.

Westin, AF 1968, 'Privacy and freedom', *Washington and Lee Law Review*, vol. 25, no. 166.

Whitten, A, Harvey, S, Fettec, I, Masielloc, B, Eisingerd, J & Horvathe, J 2011, *Transparency and choice: Protecting consumer privacy in an online world*, Google Research Publications, Google Inc. . <https://research.google.com/pubs/pub37350.html>.

Wills, C & Zeljkovic, M 2011, 'A personalized approach to web privacy: Awareness, attitudes and actions', *Information Management & Computer Security*, vol. 19, no. 1, pp. 53-73.

Woodruff, A, Pihur, V, Consolvo, S, Schmidt, L, Brandimarte, L & Acquisti, A 2014, 'Would a privacy fundamentalist sell their DNA for $ 1000 ... If nothing bad happened as a result ? The westin categories , behavioral intentions , and consequences', *Symposium on Usable Privacy and Security (SOUPS) 2014*, pp. 1-18.

Yahoo 2013, *Yahoo news - 2013 year in review*, Yahoo, viewed 19 October 2013, <http://news.yahoo.com/year-in-review/>.

Yahoo 2015, *Yahoo search: Information collection and use practices*, Yahoo, viewed 15 April 2015, <https://policies.yahoo.com/ie/en/yahoo/privacy/products/search/index.htm>.

Zimmer, M 2008, 'The gaze of the perfect search engine: Google as an infrastructure of dataveillance', in A Spink & M Zimmer (eds), *Web Search*, Springer-Verlag Berlin Heidelberg Berlin, Germany, pp. 77-99.

Zimmer, M 2008, 'The externalities of search 2.0: The emerging privacy threats when the drive for the perfect search engine meets web 2.0', *First Monday*, vol. 13, no. 3.

Zittrain, J 2008, *The future of the internet and how to stop it*, Penguin Books Ltd., London, England.

Zuboff, S 2015, 'Big other: Surveillance capitalism and the other prospects of an information civilization ', *Journal of Information Technology*, vol. 2015, no. 30, pp. 75-89.

Appendix A4.1 – Selected Figures: Personalisation Variance Analysis

Figures A1-A56: Selected Figures: Personalisation Variance Analysis

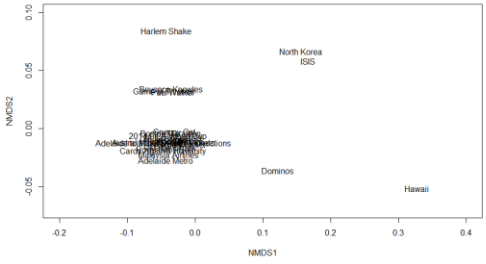


Figure A1– 1st Run-Google-Jaccard Index-Query Terms

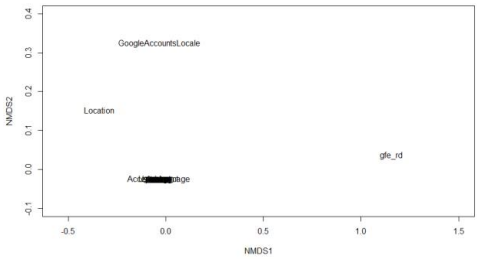


Figure A2– 1st Run-Google-Jaccard Index-Query Parameters

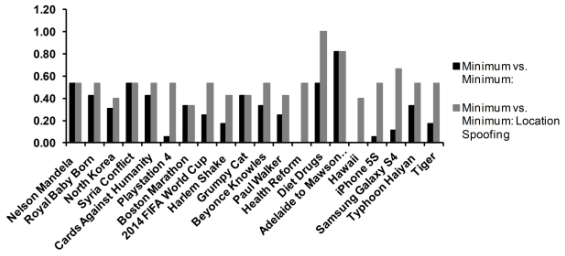


Figure A3-1st Run- Google & StartPage Comparison – Jaccard Index: Query Terms vs. Scenarios

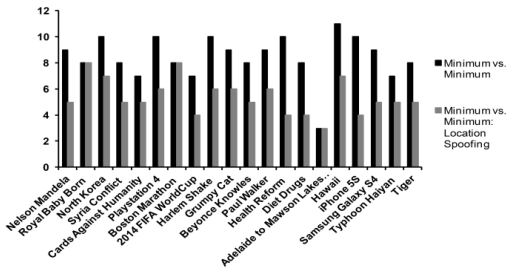


Figure A4-1st Run- Google & StartPage Comparison – Edit Distance: Query Terms vs. Scenarios

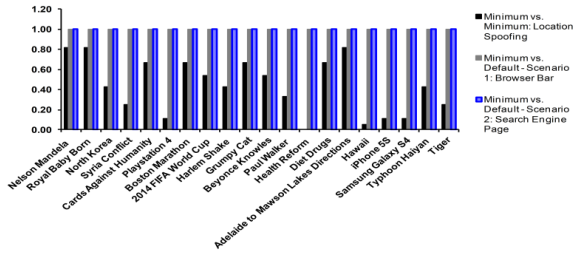


Figure A5– 1st Run - StartPage- Jaccard Index-Scenarios vs. Query Terms Variance

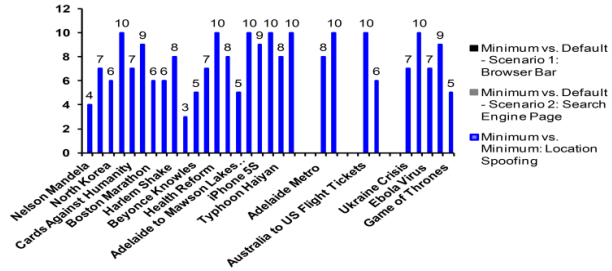


Figure A6– 1st Run - StartPage- Edit Distance-Scenarios vs. Query Terms Variance

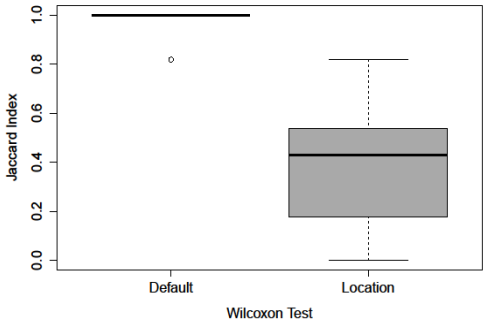


Figure A7 – 3rd Run- StartPage-Scenario 1-Jaccard Index-Wilcoxon Test- Default vs. Location

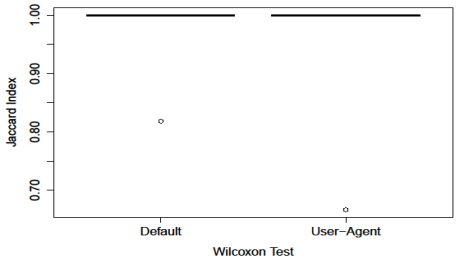


Figure A8 – 3rd Run- StartPage-Scenario 1-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

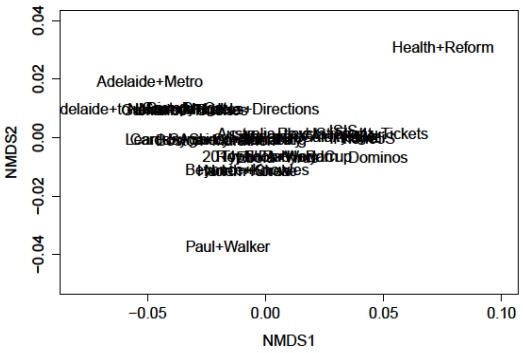


Figure A9 – 3rd Run- StartPage-Scenario 1-Jaccard Index-NMDS-Query Terms

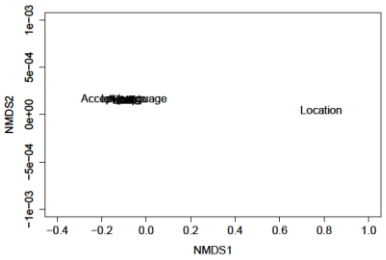


Figure A10 – 3rd Run- StartPage-Scenario 1-Jaccard Index-NMDS-Query Parameters

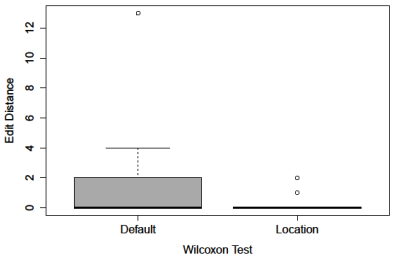


Figure A11 – 3rd Run-DuckDuckGo-Scenario 1-Edit Distance-Wilcoxon Test-Default vs. Location

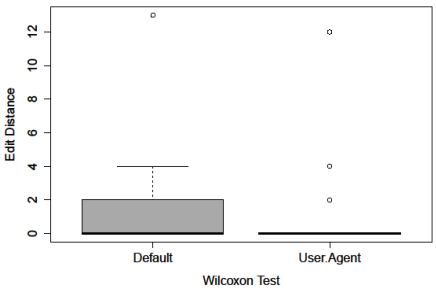


Figure A12– 3rd Run-DuckDuckGo-Scenario 1-Edit Distance-Wilcoxon Test-Default vs. User-Agent

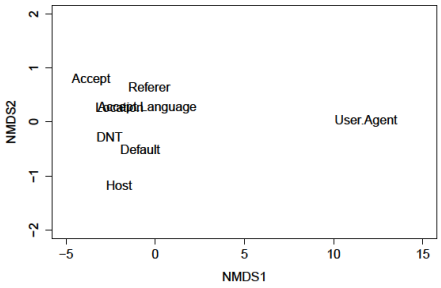


Figure A13 – 3rd Run-DuckDuckGo-Scenario 2-Edit Distance-NMDS-Query Parameters

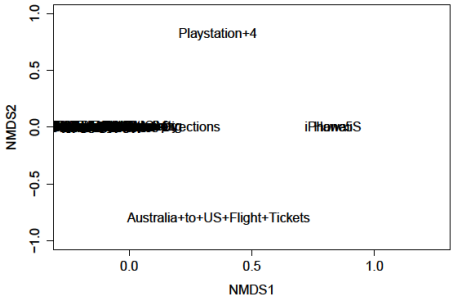


Figure A14 – 2nd Run-DuckDuckGo-Scenario 1-Jaccard Index-NMDS-Query Terms

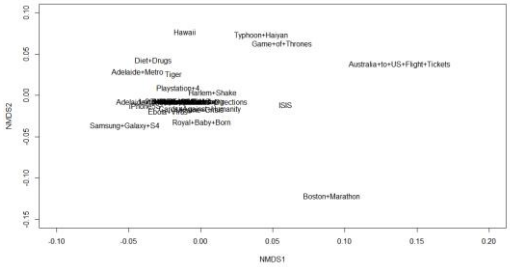


Figure A15– 3rd Run-Ixquick-Scenario 1-Jaccard Index- NMDS- Query Terms

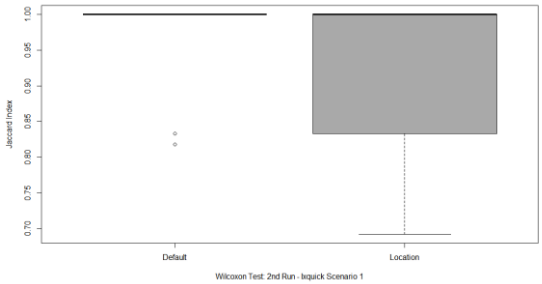


Figure A16– 3rd Run-Ixquick-Scenario 1-Jaccard Index- Wilcoxon Test- Default vs. Location

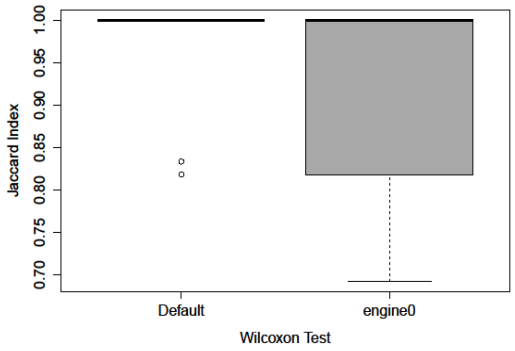


Figure A17– 3rd Run-Ixquick-Scenario 2-Jaccard Index- Wilcoxon-Default vs. engine0

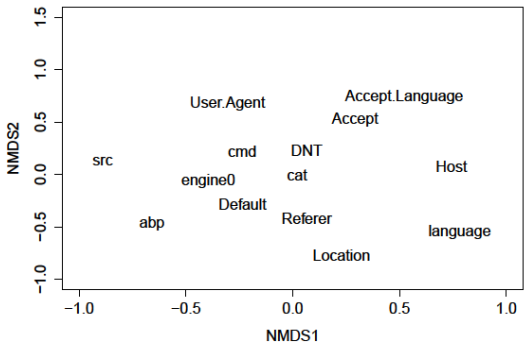


Figure A18– 2nd Run-Ixquick-Scenario 2-Edit Distance- NMDS- Query Parameters

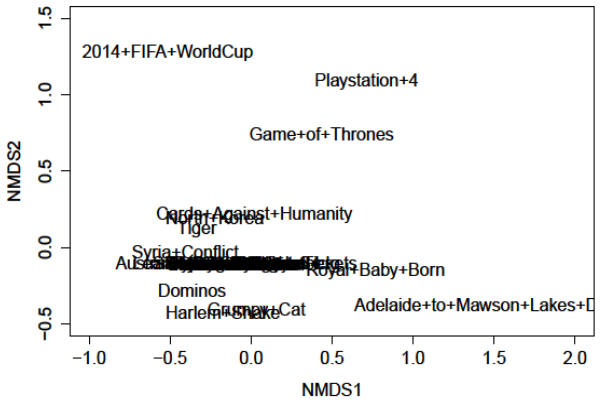


Figure A19– 2nd Run-Bing-Scenario3—Edit Distance-NMDS-Query Terms

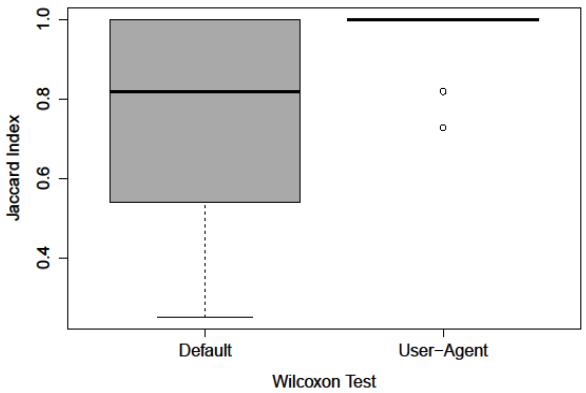


Figure A20– 3rd Run-Bing-Scenario1—Jaccard Index-Wilcoxon Test-Default vs. User-Agent

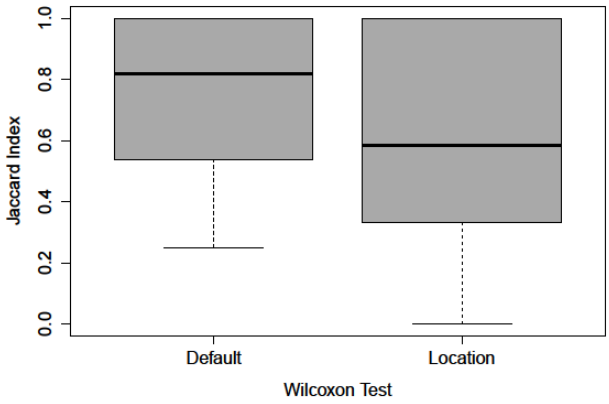


Figure A21– 3rd Run-Bing-Scenario1—Jaccard Index-Wilcoxon Test-Default vs. Location

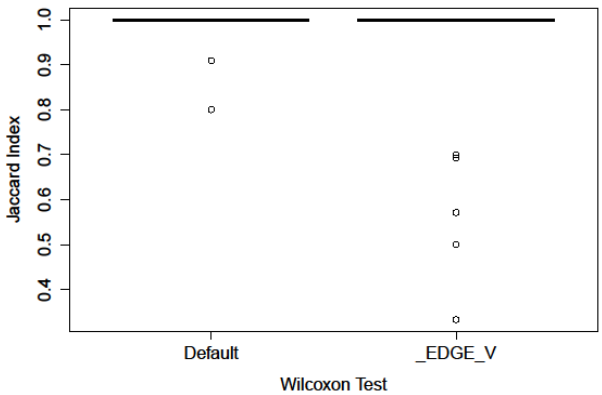


Figure A22– 2nd Run-Bing-Scenario2-Jaccard Index-Wilcoxon Test- Default vs. _EDGE_V

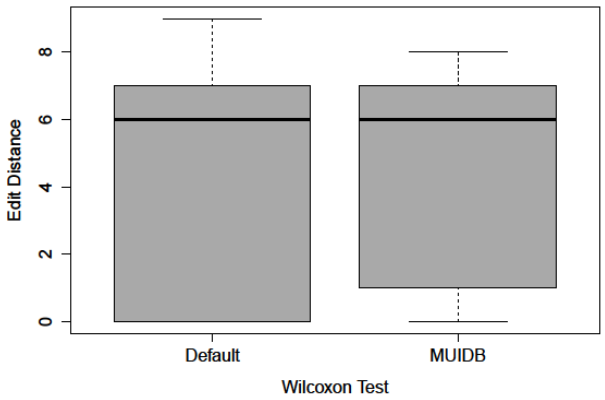


Figure A23– 3rd Run-Bing-Scenario4- Edit Distance –Wilcoxon Test-Default vs. MUIDB

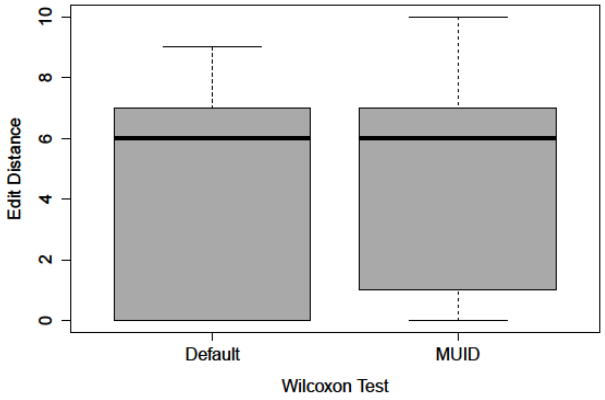


Figure A24– 3rd Run-Bing-Scenario4- Edit Distance –Wilcoxon Test-Default vs. MUID

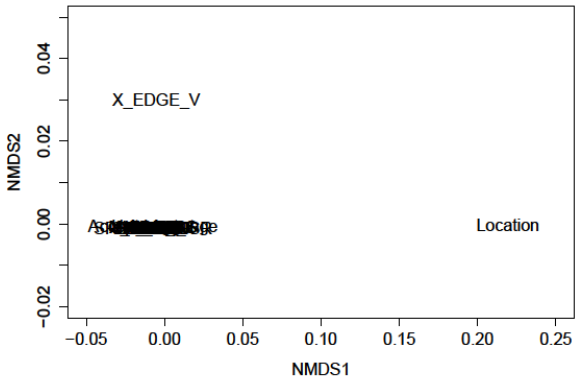


Figure A25– 2nd Run-Bing-Scenario2-Jaccard Index-NMDS-Query Parameters

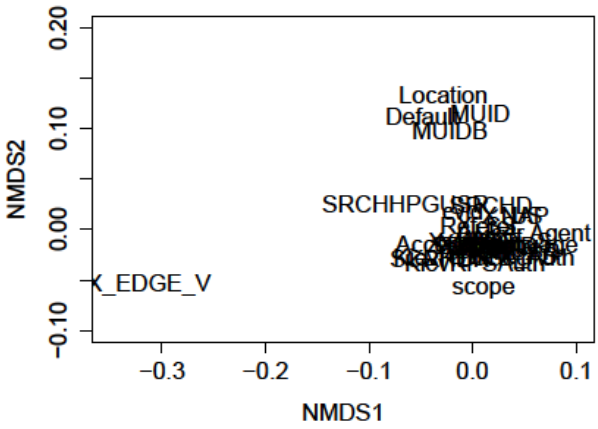


Figure A26– 3rd Run-Bing-Scenario3—Jaccard Index-NMDS-Query Parameters

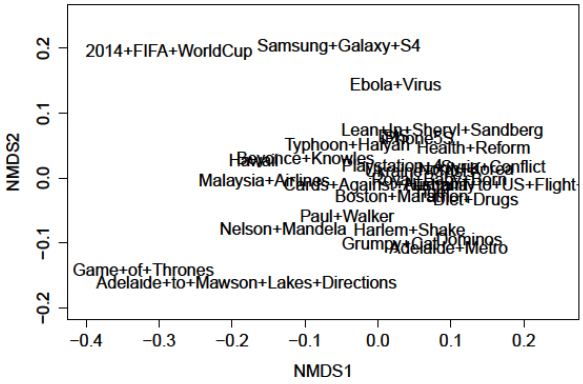


Figure A.27– 2nd Run-Yahoo-Scenario2-Jaccard Index-NMDS-Query Terms

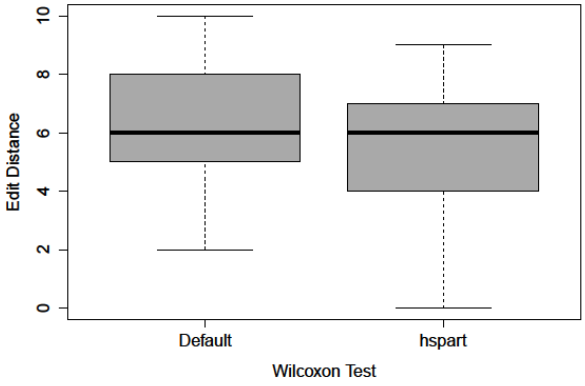


Figure A28– 2nd Run-Yahoo-Scenario1-Edit Distance-Wilcoxon Test- Default vs. hspart

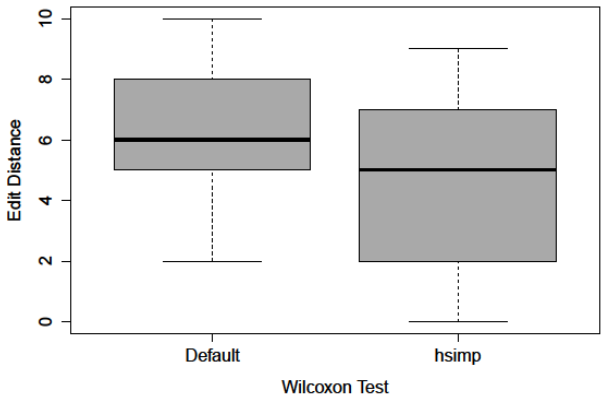


Figure A29– 2nd Run-Yahoo-Scenario1-Edit Distance-Wilcoxon Test- Default vs. hsimp

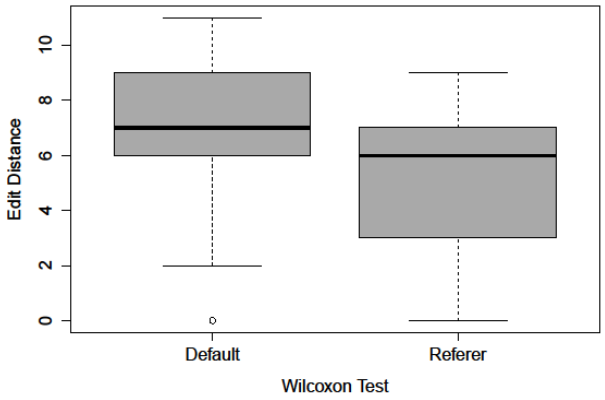


Figure A30– 2nd Run-Yahoo-Scenario2-Edit Distance-Wilcoxon Test-Default vs. Referer

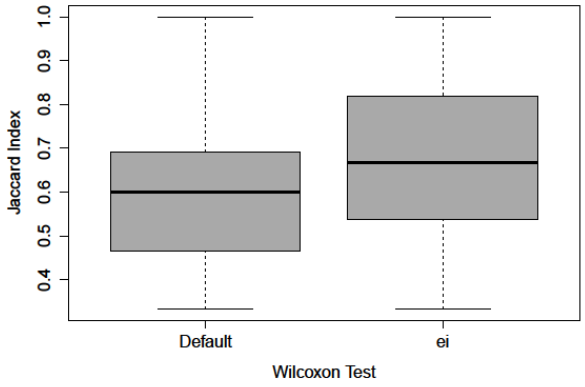


Figure A31– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. ei

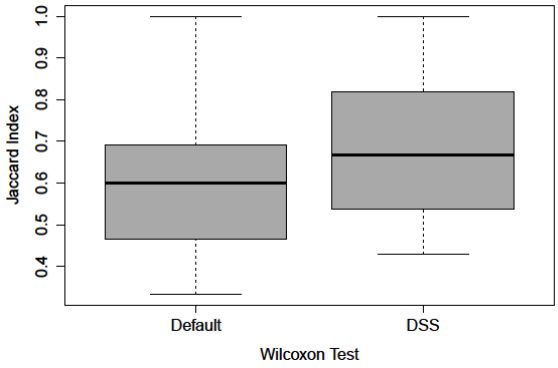


Figure A32– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. DSS

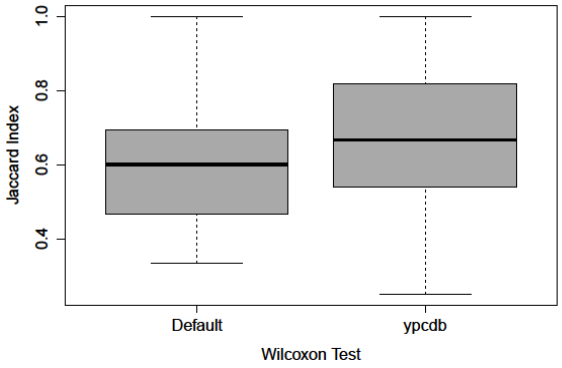


Figure A33– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. ypcdb

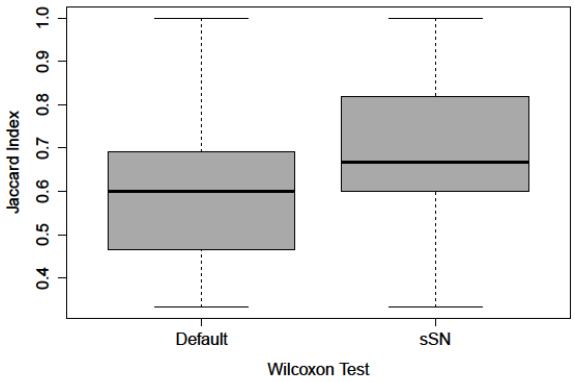


Figure A34– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. sSN

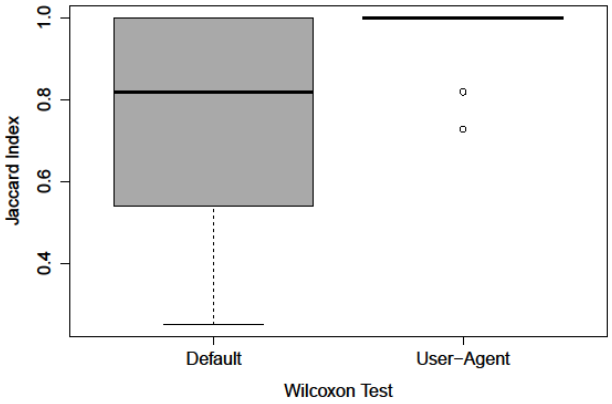


Figure A35– 3rd Run-Yahoo-Scenario 3-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

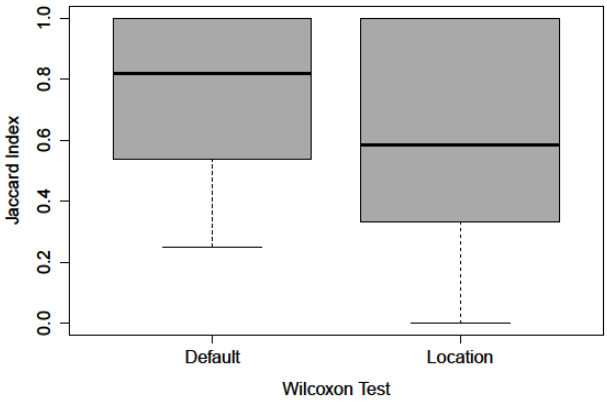


Figure A36– 3rd Run-Yahoo-Scenario 3-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

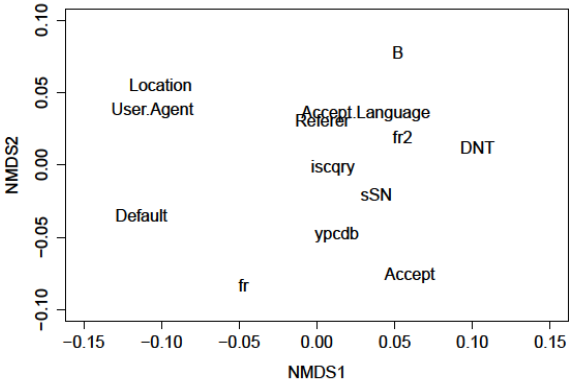


Figure A37– 2nd Run-Yahoo-Scenario2-Jaccard Index-NMDS-Query Parameters

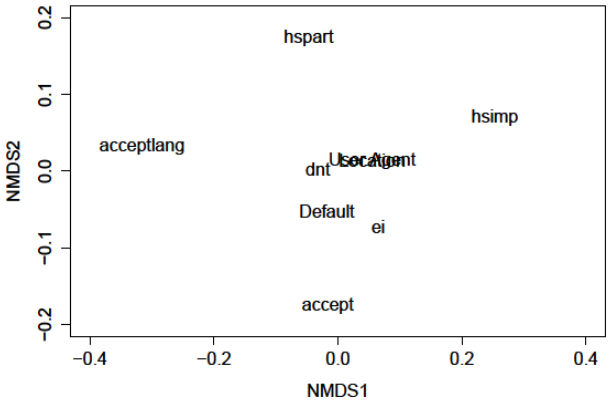


Figure A38– 2nd Run-Yahoo-Scenario1-Edit Distance-NMDS- Query Parameters

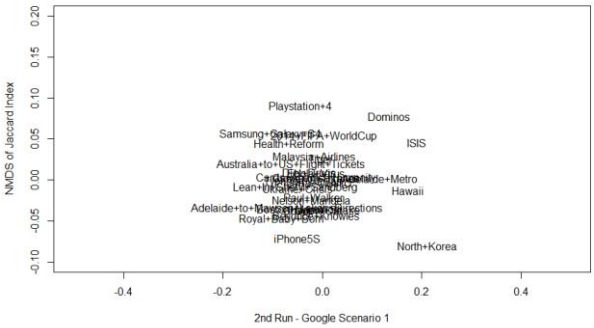


Figure A39– 2nd Run-Google- Scenario 1-Jaccard Index- NMDS- Query Terms

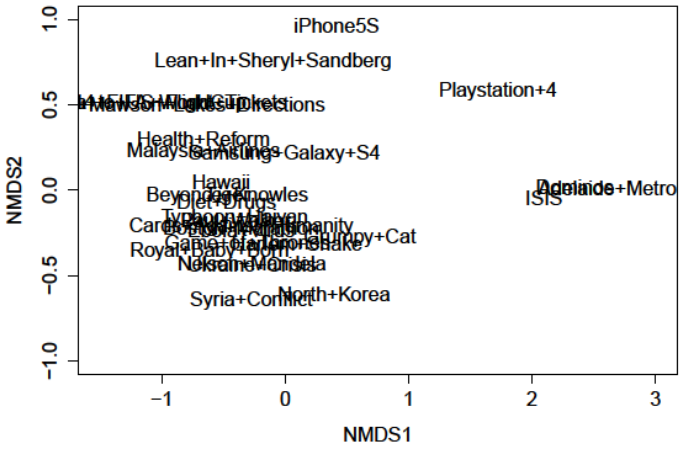


Figure A40– 3rd Run-Google- Scenario 4-Edit Distance-NMDS-Query Terms

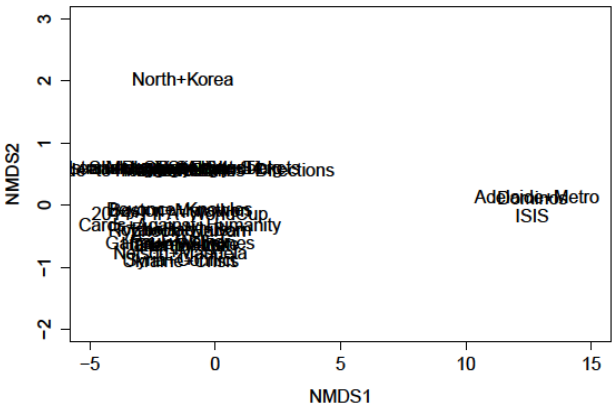


Figure A41– 3rd Run-Google- Randomised- Scenario 1-Edit Distance-NMDS-Query Terms

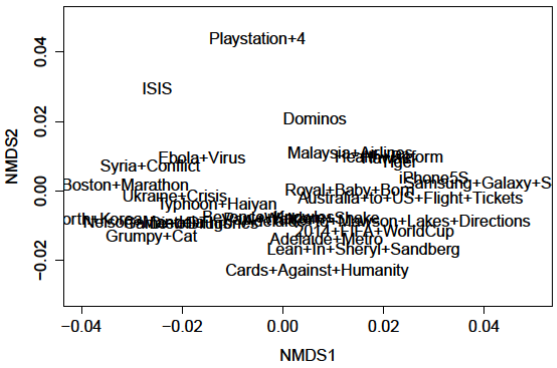


Figure A42– 3rd Run-Google-Exclusion- Scenario 3- Jaccard Index- NMDS- Query Terms

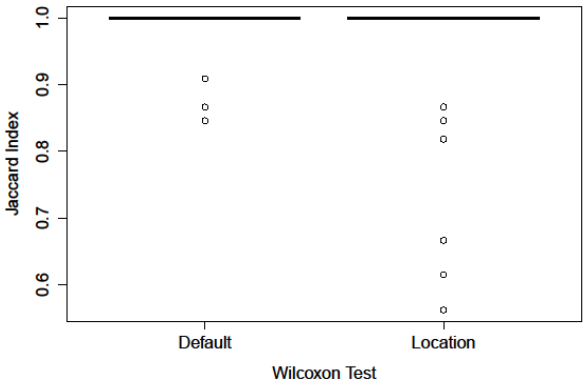


Figure A43– 3rd Run-Google- Scenario 2- Jaccard Index-Wilcoxon Test- Default vs. Location

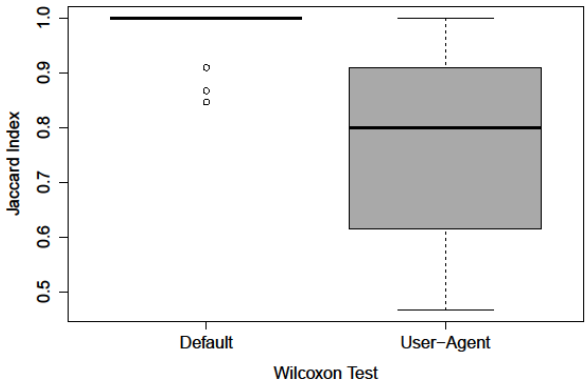


Figure A44– 3rd Run-Google- Scenario 2- Jaccard Index-Wilcoxon Test- Default vs. User-Agent

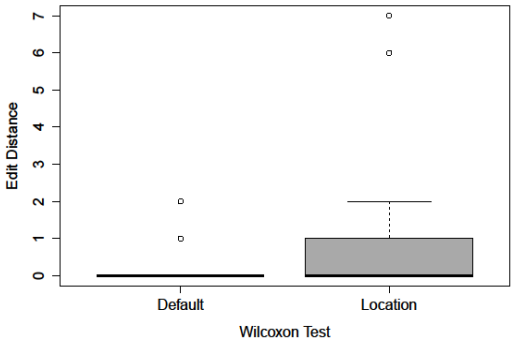


Figure A45– 3rd Run-Google- Randomised- Scenario 2-Edit Distance-Wilcoxon Test- Default vs. Location

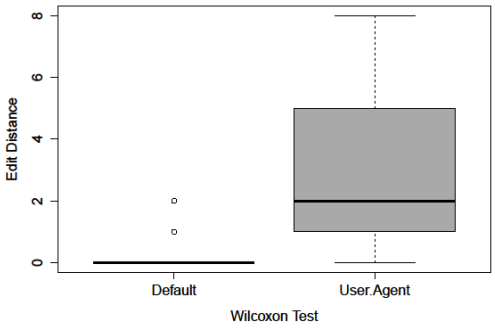


Figure A46– 3rd Run-Google- Randomised- Scenario 2-Edit Distance-Wilcoxon Test- Default vs. User-Agent

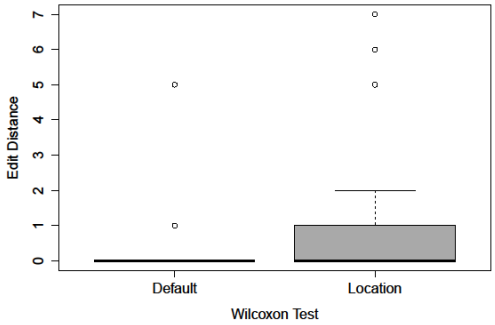


Figure A47– 3rd Run-Google-Exclusion-Scenario 2-Edit Distance- Wilcoxon test- Default vs. Location

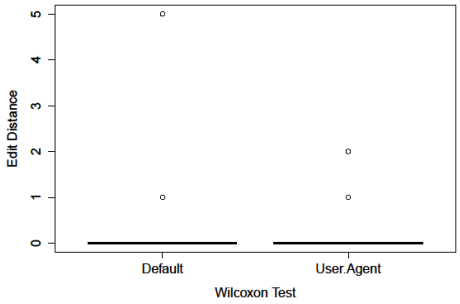


Figure A48– 3rd Run-Google-Exclusion-Scenario 2-Edit Distance- Wilcoxon test- Default vs. User-Agent

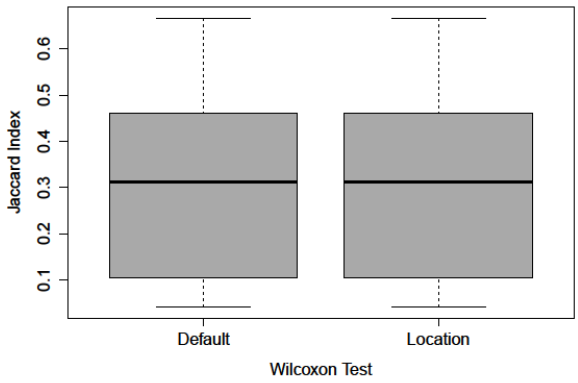


Figure A49– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Jaccard Index-Wilcoxon Test- Default vs. Location

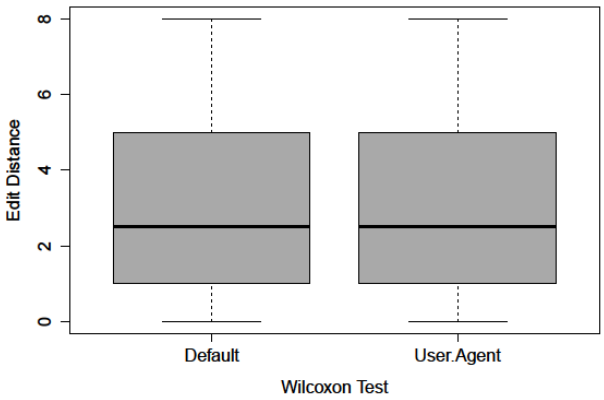


Figure A50– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3- Edit Distance-Wilcoxon Test- Default vs. User-Agent

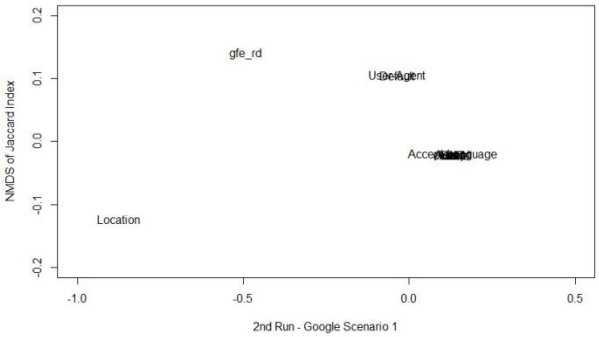


Figure A51– 2nd Run-Google- Scenario 1-Jaccard Index- NMDS- Query Parameters

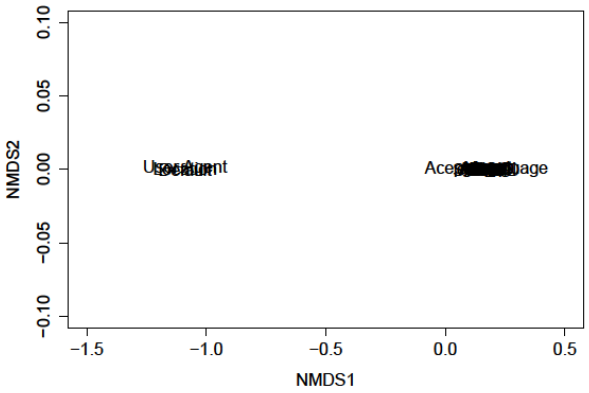


Figure A52– 3rd Run-Google- Scenario 4-Edit Distance-NMDS-Query Parameters

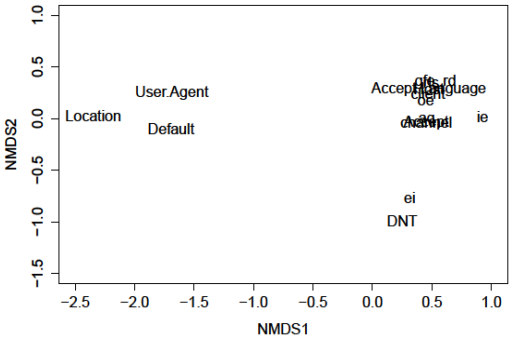


Figure A53– 3rd Run-Google- Randomised-Scenario 1-Edit Distance-NMDS-Query Parameters

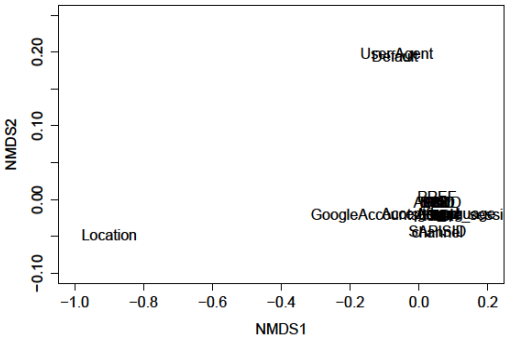


Figure A54– 3rd Run-Google-Exclusion-Scenario 3- Jaccard Index- NMDS- Query Parameters

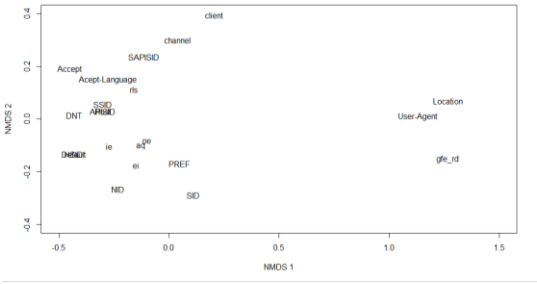


Figure A55- NMDS of Edit Distance: 2nd Run – Google Scenario 4

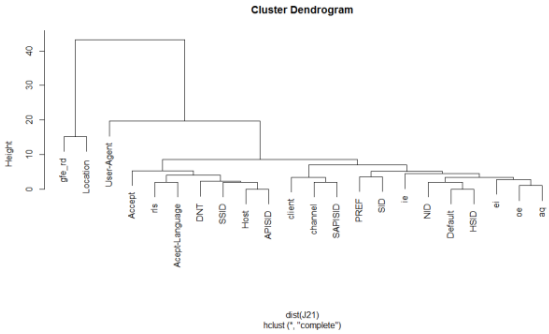


Figure A56- Cluster Analysis of NMDS of Edit Distance: 2nd Run – Google Scenario 4

Appendix A4.2 – Search Query Parameters List

Table A4.2.4 Search Query Parameters and Impact on Personalisation Variance & Identifiability[10]

Search Query Parameters	***Purpose***	***Impact on Personalisation Variance***	**Impact on Identifiability**
Query Terms			
q=query+term or *query=query+term*	Contains the query term	Significant- Jaccard Index/Edit Distance varied	High/Low: Variable based on sensitivity of terms
GET Parameters			
ie	Input encoding Output	Low	Low
oe	encoding	Low	Low
aq	Auto-complete	Low	Low
rls	Client version	Low	Low
client	Source type	Low	Low
channel	Tracks user	Low	High
gfe_rd	Redirects from country specific page	Moderate	Moderate
ei	UNIX timestamp	Low	Low
btng	Search button	Low	Low
hs	Undocumented	Low	Low
gbv	Undocumented	Low	Low
btnk	Search button	Low	Low
sclient	Undocumented	Low	Low
hl	Interface language	Low	Low
source	Type of the source	Low	Low
sei	Undocumented	Low	Low
*cmd**	Search button	Low	Low
*pid**	Undocumented	Low	Low
POST Parameters*[11]			
*cat**	Type of search	Low	Low
*language**	Language	Low	Low
*cmd**	Search button	Low	Low
*engine**	Undocumented	Low	Low
*abp**	Undocumented	Low	Low
*x**	Undocumented	Low	Low
*y**	Undocumented	Low	Low
HTTP Headers			
*HOST**	Host name	Low	Low
*User-Agent**	Browser and Operating System Version	Significant	High
*Accept**	Content Type	Low	Low
*Accept-Language**	Language Type	Low	Low
*DNT**	DoNotTrack		
Referer	Previous webpage link clicked	Low Moderate	Moderate Moderate
Cookies			
PrefID	Captures preferences	Low	High
NID	Tracks anonymous users	Low	High
GoogleAccountsLocale_session	Account sign-in authentication token	Low	High
SID	Authenticate services	Low	High
HSID	Authenticate services	Low	High
SSID	securely (encrypted)	Low	High
APISID	Undocumented Undocumented	Low	High
SAPISID		Low	High

[10] (using Google as a reference for commercial search engines and StartPage as a reference for privacy-enhanced search engines)

[11] *indicates search query parameter specific to StartPage

Appendix A4.3- Detailed Quantitative Data Analysis of the Controlled Experiment on the Effects of Data Release on the Variance of Personalised Search Results

The effects of personally-identifiable data in personalised search results were investigated through a controlled experiment design that compares the variance of personalised search result lists across both commercial and privacy-enhanced search engines by controlling the data flow released.

This appendix to Chapter 4 contains the detailed statistical analysis of the personalisation variance comparisons. This statistical analysis was done to identify significant factors impacting personalisation variance of search results. A brief introduction to the personalisation measures used (Jaccard Index and Edit Distance) and the statistical measures used (Wilcoxon Non-Parametric Test and NMDS - Non-metric Multidimensional Scaling) are provided. A summary table describes the experiment results and is used as a reference to illustrated figures.

Personalise variance across two search result lists was measured through Jaccard Index and Edit Distance. The Jaccard Index provides a measure of the differences between two pairs of similar lists. If the Jaccard Index is low, the difference between the two search engine result pages is larger. The Edit Distance provides a measure of the reordering between two pairs of similar lists. The higher the Edit Distance, the greater the level of reordering between the two search results lists.

The controlled experiment was conducted in three phases (3 data sets – October 2014, May 2015 and January 2016) across 6 search engines and 29 different query terms. To identify the effect personally-identifiable data has on personalised search results, the flow of data released to search engines was controlled by the query terms, the implicit data (search query parameters), and the methods in which we search. Personalisation variance using the Jaccard Index and Edit Distance was computed for each pair of search results (which compared the minimum set of parameters required for data release to obtain a satisfactory search query vs. a factor which could be either an individual parameter, the default set of search query

parameters or the default set with the location-spoofed). Detailed analysis of the experiment results are documented in Table A4.1 with references to the corresponding figure.

Statistics analysis measures were used to make meaning of the data (of the personalisation variance measured) namely the Wilcoxon Non-Parametric Test and NMDS - Non-metric Multidimensional Scaling.

The Wilcoxon signed rank test is a non-parametric method that is used to analyse the difference in median values between variables in two data sets that are paired or have a dependency on each other (reference). The Wilcoxon signed rank test measure was used to determine if there is a difference between personalisation variance results in the defined experiment conditions (Refer Table A4.3.1). For example, the Wilcoxon signed rank measure was used to identify the effects a type of controlled data release has on personalisation variance (e.g. minimum vs. default against minimum vs. location-spoofing). Of the 29 search query terms run in Google Scenario 1, with the location spoofed, the Wilcoxon test elicited a statistically significant median increase in personalisation variance in the Jaccard Index comparison to the default query ($W(1,29) = 434$, $P < 0.001$; Fig. A4.3.1) by comparing search query parameters in different conditions.

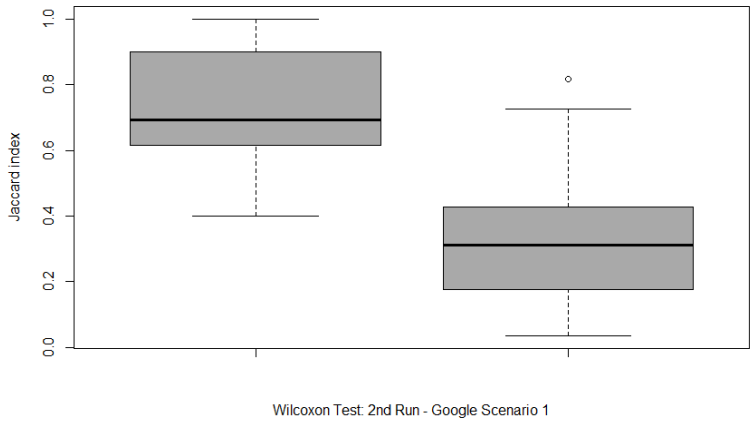


Figure A4.3.1- Wilcoxon Test of Jaccard Index: 2nd Run – Google Scenario 1

Non-metric multi-dimensional scaling (NMDS) visualises dimensions representing distances between data points and approximately ranks the order of distances between data points in a set. In this context, NMDS was used to identify the similarity or dissimilarity of a personalisation variance measure (i.e. either Jaccard Index or Edit Distance) between factors (i.e. search query parameters or query terms) for a particular experiment scenario. For example, the significant search query parameters for Google Scenario 1 (2nd Run) are

Location, Default which is nearly equal to User-Agent, gfe_rd as these parameters are dissimilar to all the other parameters which are grouped together (See Figure A4.3.2).

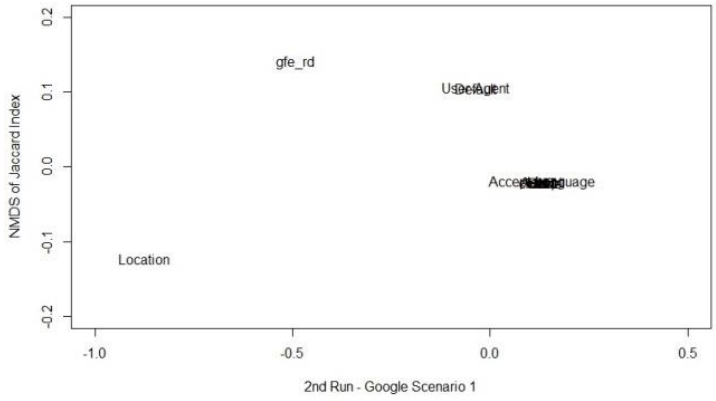


Figure A4.3.2- NMDS of Jaccard Index: 2nd Run – Google Scenario 1

Data Analysis of the Controlled Experiment on Data Leakage

This section presents a table summarising the detailed analysis.

Table A4.3.1- Appendix-Analysis: Experiment Findings Table

Figure Reference	Experiment Run	Search Engine	Scenario	Personalisation Variance Measure	Statistical Analysis Measure	Analysis Result
Privacy-enhanced Search Engines						
A4.1, A4.2	January 2016 3rd Run	StartPage	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variance for Location (W = 840.5, p-value = 4.094e-12) No variance for User-Agent (W = 421, p-value = 1) and both user-agent and default values are the same.
A4.3, A4.4					NMDS- Query Terms Query Parameters	Query Terms show variation in clusters. Health Reform (Informational- Health Term) has significant variation when location-spoofed. Query Parameters show variation in 2 clusters with location identified as significant as it is a sole separate cluster.
A4.5, A4.6				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variance for Location (W = 0, p-value = 3.659e-12) No variance for User-Agent (W = 420, p-value = 1) and both user-agent and default values are the same.
A4.7, A4.15					NMDS- Query Terms Location Parameter Variance	Query Terms are largely variant due to significant variances when location –spoofed (see A4.15 - hence, can't compute NMDS of Query Parameters). Variances seen for informational (health, entertainment, lifestyle terms) and some navigational and transactional terms.
A4.8, A4.9			2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variance for Location (W = 841, p-value = 2.676e-12) Although p <0.5, variance for User-Agent is dependent only on some query terms (not consistently variant across all query terms (W = 435, p-value = 0.3343).
A4.10, A4.11					NMDS- Query Terms Query Parameters	Query Terms show variation in clusters but not clearly distinct, although certain popular informational terms (entertainment, quirky and news term types) stand out. This variation is mostly due to location-spoofing. Query Parameters show variation in 3 clusters with location identified as significant as the parameter cmd is separate due to an experiment run-time error (query is broken when cmd is individually compared).
A4.12, A4.13				Edit Distance	Wilcoxon Test –	Significant variance for Location (W = 0.5, p-value = 3.663e-12)

					Default vs. Location, Default vs. User-Agent	No variance for User-Agent (W = 421, p-value = 1)
A4.14,					NMDS- Query Terms	Query Terms are largely variant due to significant variances when location –spoofed (Hence, can't compute NMDS of Query Parameters). Variances seen for informational (quirky and entertainment terms).
A4.16, A4.17, A4.18, A4.19, A4.20, A4.21, A4.22	October 2014 1st Run		Scenario 1 and 2	Jaccard Index Edit Distance	Only whole scenario comparisons with location-spoofing (Wilcoxon Test and NMDS)	Significant variance for Location when compared across both scenarios for Jaccard Index (both Default Scenario1 and 2- W = 841, p-value = 2.721e-12) and Edit Distance (both Default Scenario 1and 2: W = 0, p-value = 2.59e-12) NMDS shows location and default scenarios as separate entities.
A4.23, A4.24	January 2016 3rd Run	DuckDuckGo	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Variance seen for location-spoofing (W = 446, p-value = 0.6276), particularly for user-agent (W = 304.5, p-value = 0.002707).
A4.25, A4.26					NMDS- Query Terms Query Parameters	Query Terms show variance in clusters with informational-tech terms and informational-news terms being distinct. Query Parameters show variance for default and location-spoofing. Other terms are largely non-variant (with the exception of DNT which has a few deviations from 1 for some comparisons).
A4.27, A4.28, A4.29				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. t	Significant variance visible for Location (W = 529.5, p-value = 0.03903), User-Agent (W = 319, p-value = 0.004786) The parameter 't' has p<0.5 (W = 257.5, p-value = 0.6614)
A4.30					NMDS- Query Parameters	Variances seen between parameters, but more consistently for the user-agent parameter and when location-spoofed. Data not variant enough to compute NMDS of Query Terms.
A4.31, A4.32, A4.33			2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. Referer	Variance for location comparison tied to default (W = 450.5, p-value = 1), referrer comparison at Referer W = 253, p-value = 0.5728 and with p <0.5 for user-agent comparison (W = 435, p-value = 0.3337).
A4.34, A4.35					NMDS- Query Terms Query Parameters	Query terms vary in clusters, but more distinctly for informational-tech, informational-news and transactional terms. Query parameters vary in 2 cluster sets with location and referrer in the second cluster distinct from the rest of the parameter set.
A4.36, A4.37, A4.38				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. Referer	Significant variation (p<0.5) for location and user-agent (W = 288.5, p-value = 0.1407, W = 319, p-value = 0.004786 respectively). The referer parameter also varied with W = 257.5, p-value = 0.6614.
A4.39					NMDS- Query Parameters	Variances seen between parameters, but more consistently for the user-agent parameter. Data not variant enough to compute NMDS of Query Terms.

A4.40, A4.41	May 2015 2nd Run	DuckDuckGo	1	Jaccard Index	NMDS- Query Parameters Query Terms	Query parameters vary in clusters. Query terms also vary with distinct variation seen for informational term types like shopping, tech, lifestyle, and transactional terms.
A4.42, A4.43				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Not significant variation for user-agent (W = 645, p-value = 0.971) with moderate variation for location-spoofing (W = 675, p-value = 0.6945).
A4.44, A4.45			2	Jaccard Index	NMDS- Query Parameters Query Terms	Query parameters vary in clusters. Query terms also vary with distinct variation seen for informational term types like shopping, tech and transactional terms.
A4.46, A4.47				Edit Distance	Wilcoxon Test – Default vs. User-Agent, Default vs. Location	Significant variation seen for user-agent (W = 348, p-value = 0.02068) and location-spoofing (W = 260.5, p-value = 0.4766).
A4.48, A4.49, A4.50, A4.51	May 2015 2nd Run	Ixquick	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. cat Default vs. Host	Parameters significant (p<0.5) for Location (W = 364, p-value = 0.401), Host (W = 429.5, p-value = 0.3638). Although not particularly for cat (W = 391, p-value = 0.9893) and User-Agent (W = 378.5, p-value = 0.7107).
A4.52, A4.53					NMDS- Query Terms Query Parameters	Some information terms (sports, health, news, shopping term types) are more distinctly variant than others. The Host parameter varies significantly away from the other cluster.
A4.54, A4.55, A4.56, A4.57				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. cat Default vs. Host	For Edit Distance, the Host parameter has significant variation (W = 374.5, p-value = 0.3787) in comparison to location (W = 415, p-value = 0.9194), user-agent (W = 450.5, p-value = 0.5169 and the cat parameter (W = 408, p-value = 0.8082).
A4.58					NMDS- Query Parameters	Variances seen between parameters, but not distinct for individual parameters. Data not variant enough to compute NMDS of Query Terms.
A4.59, A4.60, A4.61, A4.62, A4.63, A4.64	May 2015 2nd Run	Ixquick	2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. cat Default vs. cmd Default vs. engine0 Default vs. Referer	User-Agent(W = 315, p-value = 0.08953), engine0 ((W = 318, p-value = 0.103) parameters and the Referer Header (W = 325, p-value = 0.1397) are significant when compared to others such as Location (W = 368, p-value = 0.6279), cat (W = 416, p-value = 0.6513) and cmd (W = 369.5, p-value = 0.6567).
A4.65, A4.66					NMDS- Query Terms Query Parameters	Parameters vary in groups, but not distinctly. Information tech terms distinctly vary in relation to other query term clusters.
A4.67, A4.68, A4.69, A4.70, A4.71,				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. cat Default vs. cmd	Parameters significantly vary (p<0.5) for: Location (W = 486, p-value = 0.2322), User-agent (W = 492.5, p-value = 0.1881), cat (W = 362.5, p-value = 0.3349), cmd (W = 362.5, p-value = 0.3349) and

A4.72					Default vs. engine0 Default vs. Referer	Referer (W = 487, p-value = 0.2177), but not for engine0 (W = 446, p-value = 0.655).
A4.73					NMDS-Query Parameters	Parameters vary, but not distinctly.
A4.74	January 2016 3rd Run	Ixquick	1	Jaccard Index	Wilcoxon Test – Default vs. Location	Location is significantly variant (V = 29.5, p-value = 0.09824) along with the language parameter (V = 30, p-value = 0.3657). User-agent is not variant enough to compute the Wilcoxon measure.
A4.75, A4.76					NMDS- Query Terms Query Parameters	Query terms and parameters both vary, with some distinct variation seen in query terms; particularly for informational (science, shopping, health, lifestyle, tech, sports, news term types), transactional, navigational terms and topical (popular in the news) queries of 2014.
A4.77				Edit Distance	Wilcoxon Test – Default vs. Location	Location (V = 7.5, p-value = 0.07274) and language (V = 5, p-value = 0.03697) are significantly variant.
A4.78					NMDS- Query Parameters	Not enough variation to compute NMDS of query terms. Variation in query parameters is not distinctive.
A4.79a A4.79b A4.80			2	Jaccard Index	Wilcoxon Tests NMDS- Query Terms Query Parameters	Wilcoxon test measures for location (variant V = 5, p-value = 1) are tied, but user-agent (V = 14.5, p-value = 0.3991) and engine0 parameters (V = 21, p-value = 0.02354) are significantly variant. Query terms vary in 2 clusters, with information tech term 'iPhone5S' forming its own sole cluster. Query parameters vary but not distinctly.
A4.81				Edit Distance	Wilcoxon Tests NMDS- Query Parameters	Location (V = 9, p-value = 0.6803) and user-agent (V = 15, p-value = 0.6729) are not significant, but engine0 is (V = 9, p-value = 0.2059). Query parameters vary but not distinctly. Not sufficient variation to compute NMDS of query terms
Commercial Search Engines						
A4.82, A4.83	January 2016 3rd Run	Google	1	Jaccard Index	Wilcoxon Tests – NMDS- Query Terms Query Parameters	Location (V = 434, p-value = 2.839e-06), which is significant variation. The variances in user-agent and default are tied. NMDS of query terms shows distinct variation for location. NMDS of query parameters show some variation in clusters with some informational (political), navigational and transactional terms along with a popular query term (ISIS- topical for 2014) showing variation.
A4.84, A4.85				Edit Distance	Wilcoxon Tests – NMDS- Query Terms Query Parameters	The variances in user-agent and default are tied. Location has significant variation with p <0.5 (V = 0, p-value = 2.493e-06). NMDS of query parameters show location, user-agent and default to be variant; whilst NMDS of query terms show variation in two distinct clusters (with some navigational and topical queries term being distinctly variant).
A4.86, A4.87			2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 471, p-value = 0.2172) User-Agent (W = 756, p-value = 1.975e-08)
A4.88, A4.89					NMDS- Query Terms Query Parameters	Some navigational and popular terms for 2014 are distinctly variant from clusters. Query parameters like location and user-agent are distinctly variant.

A4.90, A4.91				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 315, p-value = 0.04954) User-Agent (W = 91, p-value = 7.444e-08)
A4.92					NMDS- Query Parameters	Query terms can't be computed because of insufficient variation. User-Agent and Location are distinct from other parameters in the NMDS query parameter variation.
A4.93, A4.94	January 2016 3rd Run	Google	3	Jaccard Index	Wilcoxon Tests – NMDS- Query Terms Query Parameters	Significant variation (p<0.5) shown for: Location (V = 435, p-value = 2.561e-06) User-Agent (V = 3, p-value = 0.1797) Query term types like informational (science), navigational and popular 2014 terms are distinctly variant from other terms NMDS of query parameters show variation for location although stress is low.
A4.95, A4.96				Edit Distance	Wilcoxon Tests – NMDS- Query Terms Query Parameters	Significant variation (p<0.5) shown for: Location (V = 0, p-value = 5.417e-06) although User-Agent is at p=0.5 (V = 2, p-value = 0.5637) Some navigational and popular terms for 2014 are distinctly variant from clusters. Query parameters aren't distinctive.
A4.97, A4.98			4	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 821.5, p-value = 4.531e-10), but not for User-Agent (W = 402, p-value = 0.7781)
A4.99, A4.100					NMDS- Query Parameters Query Terms	Some informational–science, navigational and popular terms for 2014 are distinctly variant from clusters. Location, default, user-agent and the parameter oe are distinctly separate from the main cluster of query parameters.
A4.101, A4.102				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 26, p-value = 7.653e-10), but not for User-Agent (W = 434.5, p-value = 0.8321)
A4.103, A4.104					NMDS- Query Parameters Query Terms	Query terms are variant, though clustered with some types (informational – tech, navigational, transactional, a popular 2014 term) standing out. Query parameters are in two clusters with user-agent and location in one.
A4.105, A4.106	January 2016 3rd Run	Google – Randomised	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 821.5, p-value = 3.87e-10) while variance is tied between default and User-Agent (W = 420.5, p-value = 1)
A4.107, A4.108					NMDS- Query Parameters Query Terms	Some informational (news), navigational and popular query terms are important. Parameters show variation in 3 clusters - location by itself, user-agent and default, and the other parameters combined.
A4.109, A4.110				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 27, p-value = 1.173e-09) but not for User-agent (W = 407, p-value = 0.9934)
A4.111, A4.112					NMDS- Query Parameters Query Terms	Clustered variation with some query terms distinctly seen (informational – news, navigational, popular 2014 query term). Location, user-agent and default distinctly variant from other query parameters.
A4.113,	January 2016		2	Jaccard Index	Wilcoxon Test –	Significant variation (p<0.5) shown for:

A4.114	3rd Run	Google – Randomised			Default vs. Location, Default vs. User-Agent	User-Agent (W = 758, p-value = 1.628e-08) but not for Location (W = 439.5, p-value = 0.6111).
A4.115, A4.116					NMDS- Query Parameters Query Terms	Location, user-agent and the cookie PREF distinctly seen as variant. A few query term types (informational – news, navigational, popular 2014 query term) are distinctly variant apart from the cluster.
A4.117, A4.118				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 342, p-value = 0.1298) and User-Agent (W = 82.5, p-value = 3.256e-08)
A4.119					NMDS- Query Parameters	Clustered variation seen with user-agent and location set apart from other query parameters. Can't compute NMDS of query terms because of low variance.
A4.120, A4.121	January 2016 3rd Run	Google – Randomised	3	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 833, p-value = 1.439e-10), but not for User-Agent (W = 424, p-value = 0.9627)
A4.122, A4.123					NMDS- Query Parameters Query Terms	Query parameters vary with some navigational and a popular 2014 query term variant. Query terms such as location, user-agent and default are distinct from the rest of the cluster.
A4.124, A4.125				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 11.5, p-value = 1.834e-10), but not for User-Agent (W = 425, p-value = 0.9496)
A4.126, A4.127					NMDS- Query Parameters Query Terms	Clustered variation seen for query terms (such as (informational – news, tech, lifestyle plus navigational terms, a transactional and a popular 2014 query term). Query parameters such as location, user-agent and default are variant.
A4.128, A4.129	January 2016 3rd Run	Google – Random-ised	4	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 823, p-value = 3.975e-10), but not for User-Agent (W = 408, p-value = 0.8514)
A4.130, A4.131					NMDS- Query Parameters Query Terms	Clustered variation seen with a particular popular 2014 term (ISIS) slightly distinctly variant. Query parameters such as location, user-agent and default are variant from the cluster.
A4.132, A4.133				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 24, p-value = 6.21e-10), but not for User-Agent (W = 425.5, p-value = 0.9433)
A4.134, A4.135					NMDS- Query Parameters Query Terms	Clustered variation with some terms varying distinctly (informational –tech, lifestyle and a popular 2014 query term). Query parameters such as location, user-agent and default are variant and separate from the other cluster of parameters.
A4.136a A4.136b A4.137a A4.137b	January 2016 3rd Run	Google-Exclusion Effect	1	Jaccard Index	Wilcoxon Tests NMDS- Query Parameters Query Terms	Significant variation (p<0.5) shown for: Location (W = 798.5, p-value = 3.653e-09) but not for User-Agent (W = 430, p-value = 0.8845) Query terms vary significantly (informational – political, health, sports, lifestyle plus navigational terms, a transactional and a popular 2014 query term). Query parameters such as location, user-agent and default are variant.

A4.138, A4.139				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 36, p-value = 1.817e-09) but not for User-Agent (W = 412.5, p-value = 0.9031)
A4.140					NMDS- Query Terms	Clustered variation with some terms varying distinctly (informational –sports, news, quirky, lifestyle and a popular 2014 query term). NMDS of query parameters can't be computed because of low variance.
A4.142, A4.143, A4.144, A4.145	January 2016 3rd Run	Google-Exclusion Effect	2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. gbv Default vs. btnG	Significant variation (p<0.5) shown for: Location (W = 454.5, p-value = 0.4286) gbv (W = 722, p-value = 3.636e-07) but not for User-agent (W = 435.5, p-value = 0.7204) btnG (W = 425, p-value = 0.9174)
A4.146, A4.147					NMDS- Query Parameters Query Terms	Clustered variation with some terms varying distinctly (informational–science, lifestyle, a navigational term, transactional terms, and a popular 2014 query term). The parameter gbv is separate from the other cluster of query parameters.
A4.148, A4.149, A4.150, A4.151				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. gbv, Default vs. btnG	Significant variation (p<0.5) shown for: Location (W = 327, p-value = 0.05362), User-Agent (W = 391.5, p-value = 0.4999), gbv (W = 74, p-value = 1.011e-08) but not for btnG (W = 400.5, p-value = 0.6302)
A4.152					NMDS- Query Parameters	Query parameters are variant, but not distinctly. NMDS of query terms can't be computed because of low variance.
A4.153, A4.154	January 2016 3rd Run	Google-Exclusion Effect	3	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 794, p-value = 6.354e-09) but values are tied for the Default and User-Agent test (user-agent W = 420.5, p-value = 1)
A4.155, A4.156					NMDS- Query Parameters Query Terms	Clustered variation with some terms varying distinctly (informational –tech, shopping, navigational terms and a popular 2014 query term). Query parameters such as location, user-agent and default are variant and separate from the other cluster of parameters.
A4.157, A4.158				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 62, p-value = 2.315e-08) but not for User-Agent (W = 427, p-value = 0.9249)
A4.159					NMDS- Query Terms	Clustered variation with some terms varying distinctly (informational –tech, shopping, political, health and a popular 2014 query term). NMDS of query parameters can't be computed because of low variance.
A4.160, A4.161	January 2016 3rd Run	Google-Exclusion Effect	4	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 768.5, p-value = 6.236e-08) but not for User-Agent (W = 402, p-value = 0.7781)
A4.162, A4.163					NMDS- Query Parameters Query Terms	Clustered variation with some terms varying distinctly (informational –health, science and a popular 2014 query term). Query parameters such as location, user-agent and default are variant and separate from the other cluster of parameters.

A4.164, A4.165				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 78.5, p-value = 9.844e-08) but not for User-Agent (W = 449, p-value = 0.6605)
A4.166					NMDS- Query Terms	Clustered variation with some terms varying distinctly (informational –health, science, shopping). NMDS of query parameters can't be computed because of low variance.
A4.167	May 2015 2nd Run	Google	1	Jaccard Index	Wilcoxon Test – Default vs. Location	Significant variation (p<0.5) shown for: Location (V = 435, p-value = 2.544e-06) and gfe_rd(V = 367, p-value = 0.001225) but not for User-Agent (V = 4, p-value = 0.593)
A4.168, A4.169					NMDS- Query Parameters Query Terms	Clustered variation with some terms varying distinctly (informational –tech, shopping, political, navigational terms and a popular 2014 query term). Query parameters such as location, user-agent, gfe_rd and default are variant and separate from the other cluster of parameters.
A4.170, A4.171, A4.172				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. gfe_rd Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (V = 0, p-value = 3.542e-06), gfe_rd (V = 36, p-value = 0.0001371) but no values tied for the default and User-Agent comparison (V = 1.5, p-value = 1)
A4.173 A4.174					NMDS- Query Parameters Query Terms	Clustered variation with some terms varying distinctly (informational–tech, shopping, entertainment, quirky, political, science, transactional terms, navigational terms and popular 2014 query terms). Query parameters such as location, user-agent, gfe_rd and default are variant and separate from the other cluster of parameters.
A4.175a A4.175b	May 2015 2nd Run	Google	2	Jaccard Index	NMDS- Query Parameters Query Terms	Significant variation (p<0.5) shown for: User-Agent (V = 276, p-value = 2.527e-05) Location (V = 16, p-value = 0.2476) Clustered variation with some terms varying distinctly (navigational terms and popular 2014 query terms). User-agent is variant and separate from the other cluster of parameters.
A4.176a A4.176b A4.177				Edit Distance	Wilcoxon Test-s Default vs. Location Default vs. User-Agent NMDS- Query Parameters	Significant variation (p<0.5) shown for: User-Agent (W = 112, p-value = 5.241e-07) Location (W = 367.5, p-value = 0.3247) NMDS of query terms can't be computed because of low variance. Query parameters such as location, user-agent, the cookie NID and default are variant and separate from the other cluster of parameters.
A4.178, A4.179	May 2015 2nd Run	Google	3	Jaccard Index	Wilcoxon Tests NMDS- Query Parameters Query Terms	Significant variation (p<0.5) shown for: User-agent (V = 276, p-value = 2.527e-05) Location (V = 16, p-value = 0.2476) Clustered variation with some terms varying distinctly (informational–political, health, lifestyle, navigational terms and a popular 2014 query term). Query parameters such as location and user-agent, are variant and separate from the other cluster of parameters

A4.180a A4.180b A4.181a A4.181b				Edit Distance	Wilcoxon Tests NMDS- Query Parameters Query Terms	Significant variation (p<0.5) shown for: User-Agent (W = 119.5, p-value = 8.607e-07) Location (W = 8, p-value = 4.026e-11) Clustered variation with some terms varying distinctly (informational–tech, sports, shopping, entertainment, political, transactional terms, navigational terms and popular 2014 query terms). Query parameters such as location, aq, host are slightly variant from the other cluster of parameters.
A4.182a A4.182b A4.183a A4.183b	May 2015 2nd Run	Google	4	Jaccard Index	Wilcoxon Tests NMDS- Query Parameters Query Terms	Significant variation (p<0.5) shown for: User-Agent (W = 604, p-value = 2.48e-07) Location (W = 676, p-value = 1.601e-10) Clustered variation in 2 clusters including a smaller cluster with transactional terms, navigational terms and popular 2014 query terms. Query parameters such as location, user-agent, gfe_rd and default are variant and separate from the other cluster of parameters
A4.184, A4.185 A4.186				Edit Distance	Wilcoxon Tests NMDS- Query Parameters	Significant variation (p<0.5) shown for: User-Agent (V = 0, p-value = 1.081e-05) gfe_rd (V = 0, p-value = 7.516e-06) Location (V = 0, p-value = 2.288e-06) Query parameters such as location, user-agent, gfe_rd and default are variant and separate from the other cluster of parameters. This is further shown in a cluster dendrogram diagram showing how closely related the distances between parameters are.
A4.187, A4.188	October 2014 1st Run	Google	Parameters across all 4 scenarios	Jaccard Index	Wilcoxon Tests NMDS- Query Parameters Query Terms	Significant variation (p<0.5) shown for: User-Agent(V = 363, p-value = 0.001649) gfe_rd (V = 325, p-value = 1.173e-05) but not for Location (V = 15.5, p-value = 0.7995) Query parameters such as location, gfe_rd and the cookie GoogleAccountsLocale are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–political, lifestyle, quirky, transactional terms, navigational terms and popular 2014 query terms).
A4.189, A4.190				Edit Distance	Wilcoxon Tests NMDS-	Significant variation (p<0.5) shown for: User-agent (V = 25.5, p-value = 0.272) gfe_rd (V = 3, p-value = 3.319e-06) location (V = 14, p-value = 1.616e-05)

					Query Parameters	Query parameters such as location, gfe_rd, user-agent, default along with the cookies NID and PREFID are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–tech, shopping, political, lifestyle, quirky, transactional terms, navigational terms and popular 2014 query terms).
A4.191, A4.192	January 2016 3rd Run	Google-User Agent Experi-ment	Google Chrome – Scenario 1 with minimum identifiers to limit tracking	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: user-agent (W = 15, p-value = 1.022e-07), but not for location (W = 244.5, p-value = 0.9625)
A4.193, A4.194					NMDS- Query Parameters Query Terms	Query parameters such as location, gfe_rd, user-agent, default are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–tech, shopping, political, navigational terms and popular 2014 query terms).
A4.195, A4.196				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: user-agent (W = 482, p-value = 1.681e-08) but not for location (W = 234.5, p-value = 0.8676)
A4.197, A4.198					NMDS- Query Parameters Query Terms	Query parameters such as location, DNT, user-agent, default are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–tech, shopping, transactional terms, navigational terms and popular 2014 query terms).
A4.199, A4.200	January 2016 3rd Run	Google-User Agent Experi-ment	Internet Explorer-Scenario 3	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: location (W = 882, p-value = 1.652e-10) but not for user-agent (W = 230, p-value = 0.7848)
A4.201, A4.202					NMDS- Query Parameters Query Terms	Query parameters such as location, user-agent, default are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (navigational terms and popular 2014 query terms).
A4.203, A4.204				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: location (W = 5.5, p-value = 2.686e-08) but not for user-agent (W = 449, p-value = 0.6605)
A4.205, A4.206					NMDS- Query Parameters Query Terms	Query parameters such as location, user-agent, default are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–tech, shopping, political, transactional terms, navigational terms and popular 2014 query terms).
A4.207,	May 2015	Bing	1	Jaccard Index	Wilcoxon Test –	Significant variation (p<0.5) shown for:

A4.208	2nd Run				Default vs. Location, Default vs. User-Agent	Location (W = 841, p-value = 5.587e-12) but not for User-Agent (W = 406.5, p-value = 0.5843)
A4.209, A4.210					NMDS- Query Parameters Query Terms	Location is variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–sports, entertainment, transactional terms, navigational terms and popular 2014 query terms).
A4.211, A4.212				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 0, p-value = 3.968e-12) but not for User-Agent (W = 435, p-value = 0.5704)
A4.213					NMDS- Query Terms	Clustered variation with some query terms varying distinctly (informational–sport, transactional terms). NMDS of query parameters can't be computed due to 1 dimensional variance of distance.
A4.214, A4.215, A4.216	May 2015 2nd Run	Bing	2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. _EDGE_V	Significant variation (p<0.5) shown for: Location (W = 704.5, p-value = 5.796e-07), _EDGE_V (W = 469, p-value = 0.1871), but not for User-Agent (W = 406.5, p-value = 0.5843)
A4.217, A4.218					NMDS- Query Parameters Query Terms	Query parameters such as location and the cookie _EDGE_V are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–sports, entertainment, transactional terms, navigational terms).
A4.219, A4.220, A4.221				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. _EDGE_V	Significant variation (p<0.5) shown for: Location (W = 128, p-value = 2.597e-07) _EDGE_V (W = 325.5, p-value = 0.02551) but not for User-Agent (W = 435, p-value = 0.5704)
					NMDS	Can't compute NMDS of Query Terms or Parameters because variance is only 1 dimensional in distance measures used for NMDS.
A4.222, A4.223, A4.224	May 2015 2nd Run	Bing	3	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. _EDGE_V	Significant variation (p<0.5) shown for: Location (W = 840, p-value = 4.524e-12) but not for User-Agent (W = 435.5, p-value = 0.5568) _EDGE_V (W = 434.5, p-value = 0.5843)
A4.225, A4.226					NMDS- Query Parameters Query Terms	Location is variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–news, political, sports, entertainment, lifestyle, science, transactional terms, navigational terms and popular 2014 query terms).
A4.227,				Edit Distance	Wilcoxon Test –	Significant variation (p<0.5) shown for:

A4.228 A4.229					Default vs. Location, Default vs. User-Agent, Default vs. _EDGE_V	Location (W = 1, p-value = 3.434e-12) but not for User-Agent (W = 405.5, p-value = 0.5568) _EDGE_V (W = 405.5, p-value = 0.5568)
A4.230					NMDS- Query Terms	Clustered variation with some query terms varying distinctly (informational–news, shopping, lifestyle, science, transactional terms, navigational terms and popular 2014 query terms). NMDS of Query Parameters can't be computed because of low variance.
-	May 2015 2nd Run	Bing	4	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (V = 406, p-value = 3.932e-06), but values are tied for User-Agent and Default
A4.231, A4.232					NMDS- Query Parameters Query Terms	Location is variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–news, political, shopping, lifestyle, quirky, entertainment, transactional terms, navigational terms and popular 2014 query terms).
A4.233, A4.234				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 5.5, p-value = 7.388e-12) and moderately for User-Agent (W = 435.5, p-value = 0.5568)
A4.235					NMDS- Query Terms	Query Parameters can't be computed because of low variance. Clustered variation with some query terms varying distinctly (informational–news, political, shopping, quirky, navigational terms and popular 2014 query terms).
A4.236, A4.237	January 2016 3rd Run	Bing	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 534, p-value = 0.0689) User-Agent (W = 218.5, p-value = 9.733e-05)
A4.238, A4.239					NMDS- Query Parameters Query Terms	Location is variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–news, political, shopping, lifestyle, quirky, entertainment, transactional terms, navigational terms and popular 2014 query terms).
A4.240, A4.241				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 0, p-value = 6.113e-12) User-Agent (W = 376, p-value = 0.2773)
A4.242, A4.243					NMDS- Query Parameters Query Terms	Query parameters vary but not distinctly. Clustered variation with some query terms varying distinctly (informational–entertainment, transactional terms, navigational terms and popular 2014 query terms).

A4.244	January 2016 3rd Run	Bing	3 * Scenario 2 wasn't run as repetitive	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 534, p-value = 0.0689) User-Agent (W = 293, p-value = 0.01595) MUIDB (V = 4, p-value = 0.3452) _EDGE_V (V = 151, p-value = 0.2164) SRCHHPGUSR (W = 349, p-value = 0.211) but not for MUID (V = 4, p-value = 0.715)
A4.245a A4.245b					NMDS- Query Parameters Query Terms	Query parameters such as location, default, and the cookies _EDGE_V, MUID, MUIDB are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–lifestyle, tech, science, transactional terms, navigational terms).
A4.246				Edit Distance	Wilcoxon Test – Default vs. Location	Significant variation (p<0.5) shown for: Location (W = 420, p-value = 1) User-Agent (W = 606.5, p-value = 0.001068) SRCHHPGUSR (W = 526.5, p-value = 0.07406) but not for _EDGE_V (W = 380.5, p-value = 0.5285) MUIDB (W = 380.5, p-value = 0.5285) MUID (W = 449.5, p-value = 0.6455)
A4.247					NMDS- Query Terms	Query Parameters can't be computed because of low variance. Clustered variation with some query terms varying distinctly (informational-sports).
A4.248, A4.249, A4.250, A4.251, A4.252	January 2016 3rd Run	Bing	4	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. MUID, Default vs. MUIDB, Default vs. _EDGE_V	Significant variation (p<0.5) shown for: Location (W = 534, p-value = 0.0689) User-Agent (W = 218.5, p-value = 9.733e-05) but not for_EDGE_V (W = 408, p-value = 0.8427) MUID (W = 423.5, p-value = 0.9674) MUIDB (W = 423, p-value = 0.9739)
A4.253, A4.254					NMDS- Query Parameters Query Terms	Query parameters such as location and the cookie _EDGE_V are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–lifestyle, science, quirky, transactional terms).
A4.255, A4.256, A4.257, A4.258, A4.259				Edit Distance	Wilcoxon Test – Default vs. Location Default vs. User-Agent, Default vs. MUID, Default vs. MUIDB, Default vs. _EDGE_V	Significant variation (p<0.5) shown for: Location (W = 322, p-value = 0.1211) User-Agent (W = 496.5, p-value = 0.2264) but not for MUID (W = 401.5, p-value = 0.7697) MUIDB (W = 416.5, p-value = 0.9558)

A4.260, A4.261					NMDS- Query Terms Query Parameters	Query parameters such as Referer and _FS are variant and separate from the other cluster of parameters Clustered variation with some query terms varying distinctly (informational–sports, science).
---	May 2015 2nd Run	Yahoo	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (V = 149, p-value = 0.465) but not for User-Agent (V = 146, p-value = 0.6571)
A4.262, A4.263					NMDS- Query Parameters Query Terms	Query parameter variation is clustered but not distinctly variant. Similarly for query terms (no distinct variation)
A4.264, A4.265, A4.266, A4.267				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent Default vs. hsimp Default vs. hspart	Significant variation (p<0.5) shown for: hspart (W = 560, p-value = 0.0282) hsimp (W = 580.5, p-value = 0.01231) but not for Location (W = 403, p-value = 0.789) User-Agent (W = 387.5, p-value = 0.6096)
A4.268, A4.269					NMDS- Query Parameters Query Terms	Query parameters such as hspart, hsimp, accceptlang, accept are variant and separate from the other cluster of parameters. Clustered variation with some query terms varying distinctly (informational–lifestyle, science, entertainment, navigational terms, transactional terms).
A4.270, A4.271, A4.272, A4.273, A4.274, A4.275	May 2015 2nd Run	Yahoo	2	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. fr Default vs. sSN Default vs. ypcdb Default vs. Referer	Significant variation (p<0.5) shown for: Location (W = 346.5, p-value = 0.2523) ypcdb (W = 257, p-value = 0.01083) Referer (W = 248.5, p-value = 0.007171) sSN (W = 261, p-value = 0.01299) fr (W = 312.5, p-value = 0.09303) but not for User-Agent (W = 411, p-value = 0.8885)
A4.276, A4.277					NMDS- Query Parameters Query Terms	Query parameters such as user-agent, location and default are clustered on one side and separate from the other cluster of parameters. Clustered variation with some query terms varying distinctly (informational–sports, tech, lifestyle, navigational terms, transactional terms).
A4.278, A4.279, A4.280				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. Referer	Significant variation (p<0.5) shown for: Location (W = 464.5, p-value = 0.4959) Referer (W = 564.5, p-value = 0.02444) but not for User-Agent (W = 428.5, p-value = 0.9065)
A4.281, A4.282					NMDS- Query Parameters Query Terms	Query parameters such as fr, B, DNT, accept are variant and separate from the other cluster of parameters. Clustered variation with some query terms varying distinctly (informational– tech, political, lifestyle, science, navigational terms, transactional terms).
A4.283, A4.284,	May 2015 2nd Run	Yahoo	3	Jaccard Index	Wilcoxon Test –	Significant variation (p<0.5) shown for: hsimp (W = 339, p-value = 0.2062)

A4.285, A4.286, A4.287, A4.288, A4.289					Default vs. Location, Default vs. User-Agent, Default vs. DSS Default vs. hsimp Default vs. ei Default vs. ypcdb Default vs. sSN	ei (W = 305, p-value = 0.07251) DSS (W = 295, p-value = 0.05066) ypcdb (W = 288, p-value = 0.03922) sSN (W = 263.5, p-value = 0.01408) but not for Location (W = 421.5, p-value = 0.9938) User-Agent (W = 381, p-value = 0.5437)
A4.290, A4.291					NMDS- Query Parameters Query Terms	Query parameters such as location, Default, hsimp are slightly variant from the other cluster of parameters. Clustered variation with some query terms varying distinctly (informational– news, tech, political, lifestyle, science, navigational terms, transactional terms).
A4.292, A4.293, A4.294, A4.295, A4.296, A4.297				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. hsimp Default vs. ei Default vs. sSN Default vs. DSS Default vs. ypcdb	Significant variation (p<0.5) shown for: hsimp (W = 568, p-value = 0.0211) ei (W = 595, p-value = 0.006393) DSS (W = 572.5, p-value = 0.01739) ypcdb (W = 609, p-value = 0.00316) sSN (W = 582, p-value = 0.01142) but not for Location (W = 446, p-value = 0.6954) User-Agent (W = 407, p-value = 0.838)
A4.298, A4.299					NMDS- Query Parameters Query Terms	Query parameters such as Accept, DNT, AO are slightly variant from the other cluster of parameters. Clustered variation with some query terms varying distinctly (informational– news, tech, political, quirky, entertainment, lifestyle, navigational terms, transactional terms).
-	May 2015 2nd Run	Yahoo	4	Jaccard Index	Wilcoxon Test –	Significant variation (p<0.5) shown for: Location (V = 96, p-value = 0.2068) User-Agent (V = 68, p-value = 0.2858)
A4.300, A4.301, A4.302					NMDS- Query Parameters Cluster Dendrogram – Query Parameters Query Terms	Query parameters such as T, D, user-agent, location are slightly variant from the other cluster of parameters. Clustered variation with some query terms varying distinctly (informational– news, health, quirky, entertainment, lifestyle, science, navigational terms, transactional terms).
A4.303, A4.304, A4.305, A4.306, A4.307, A4.308, A4.309				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent, Default vs. hsimp Default vs. hspart Default vs. sSN Default vs. DSS Default vs. ypcdb	Significant variation (p<0.5) shown for: hsimp (W = 665, p-value = 0.0001282) hspart (W = 683.5, p-value = 3.815e-05) ypcdb (W = 665.5, p-value = 0.0001243) sSN (W = 654, p-value = 0.0002538) DSS (W = 680.5, p-value = 4.654e-05) but not for Location (W = 435, p-value = 0.8262) User-Agent (W = 382, p-value = 0.5504)
A4.310, A4.311					NMDS- Query Parameters	Query parameters such as hspart, acceptlang, F, ei, are slightly variant from the other cluster of parameters.

					Query Terms	Clustered variation with some query terms varying distinctly (informational– news, health, tech, sports, quirky, lifestyle, navigational terms, transactional terms).
A4.312, A4.313	January 2016 3rd Run	Yahoo	1	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: User-Agent (V = 0, p-value = 3.858e-05) but not for Location (V = 180, p-value = 0.909)
A4.314, A4.315					NMDS- Query Parameters Query Terms	Query parameters such as user-agent, default and in particular - location are clustered separately from the other cluster of parameters. Clustered variation with some query terms occurs in two clusters but not distinctly.
A4.316, A4.317				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 534, p-value = 0.0689) User-Agent (W = 218.5, p-value = 9.733e-05)
A4.318					NMDS- Query Terms	Query Parameters can't be computed because of low variance. Clustered variation with some query terms varying distinctly (informational–sports, political, entertainment, lifestyle, navigational terms, transactional terms).
A4.319, A4.320	January 2016 3rd Run	Yahoo	3	Jaccard Index	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 534, p-value = 0.0689) User-Agent (W = 218.5, p-value = 9.733e-05)
A4.321, A4.322					NMDS- Query Parameters Query Terms	Query parameters such as user-agent, SMW, default and in particular - location are clustered separately from the other cluster of parameters. Clustered variation occurs in two clusters but not distinctly. One cluster contains the informational-news term 'Nelson Mandela'.
A4.323, A4.324				Edit Distance	Wilcoxon Test – Default vs. Location, Default vs. User-Agent	Significant variation (p<0.5) shown for: Location (W = 534, p-value = 0.0689) User-Agent (W = 218.5, p-value = 9.733e-05)
A4.325					NMDS- Query Terms	NMDS of query parameters can't be computed due to low variance. Clustered variation occurs in four clusters but not distinctly. One cluster contains the informational-news term 'Nelson Mandela', another the informational-political term 'North Korea' and a third contains navigational and transactional terms.
A4.326	October 2014 1st Run	Google and StartPage	1	Comparing commercial and privacy-enhanced search engines	Jaccard Index – Minimum vs. Location-Spoofing	Variances were significant when location-spoofed. Some query terms such as transactional terms, health and tech terms also showed variations.
A4.327					Edit Distance – Minimum vs. Location-Spoofing	Similarly edit distance variations were significant when location-spoofed with some query term variations seen for transactional, tech and topical 2014 terms.

List of Figures

Figures A4.1-A4.22 StartPage

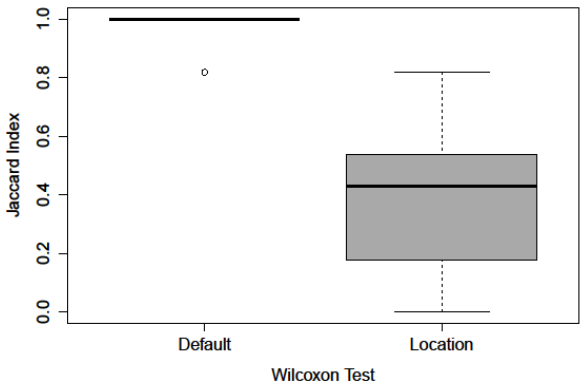


Figure A4.1 – 3rd Run- StartPage- Scenario 1- Jaccard Index-Wilcoxon Test- Default vs. Location

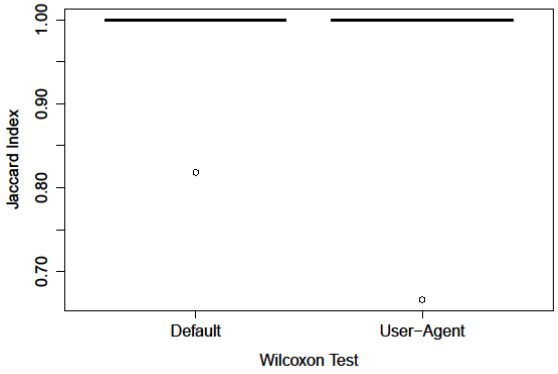


Figure A4.2 – 3rd Run- StartPage- Scenario 1- Jaccard Index-Wilcoxon Test- Default vs. User-Agent

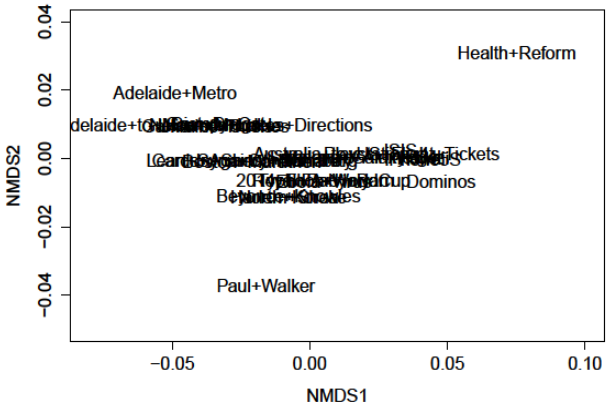


Figure A4.3 – 3rd Run- StartPage- Scenario 1- Jaccard Index-NMDS-Query Terms

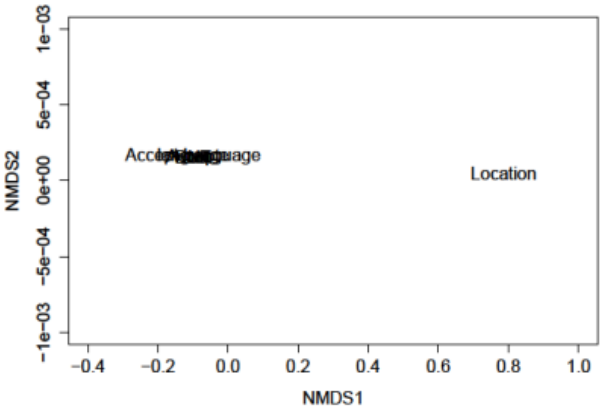


Figure A4.4 – 3rd Run- StartPage- Scenario 1- Jaccard Index-NMDS-Query Parameters

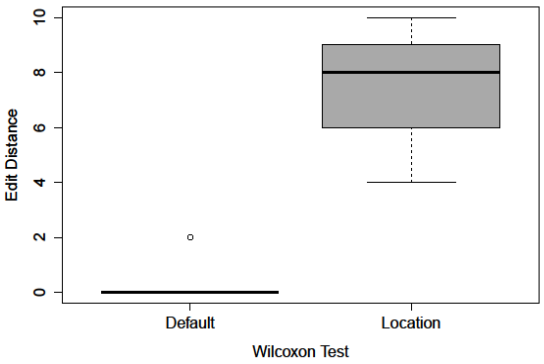


Figure A4.5 – 3rd Run- StartPage- Scenario 1- Edit Distance -Wilcoxon Test- Default vs. Location

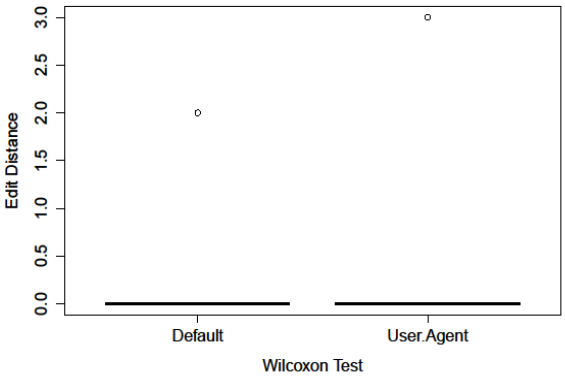


Figure A4.6 – 3rd Run- StartPage- Scenario 1- Jaccard Index-Wilcoxon Test- Default vs. User-Agent

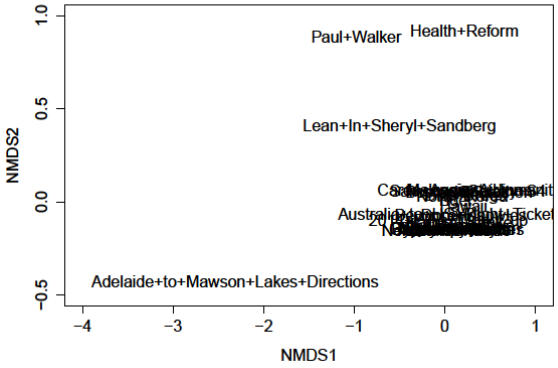


Figure A4.7 – 3rd Run- StartPage-Scenario 1- Edit Distance-NMDS-Query Terms

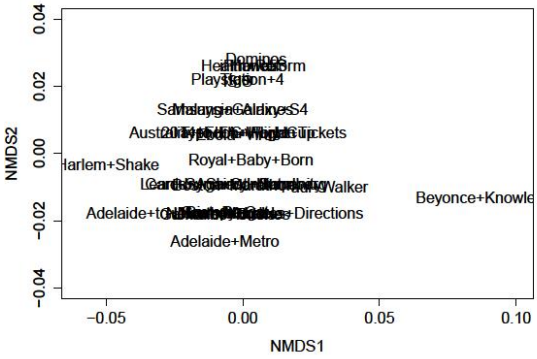


Figure A4.10 – 3rd Run- StartPage-Scenario 2- Jaccard Index-NMDS-Query Terms

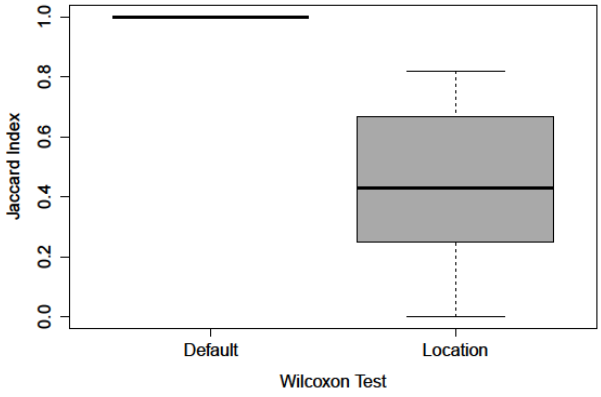


Figure A4.8 – 3rd Run- StartPage-Scenario 2- Jaccard Index-Wilcoxon Test-Default vs. Location

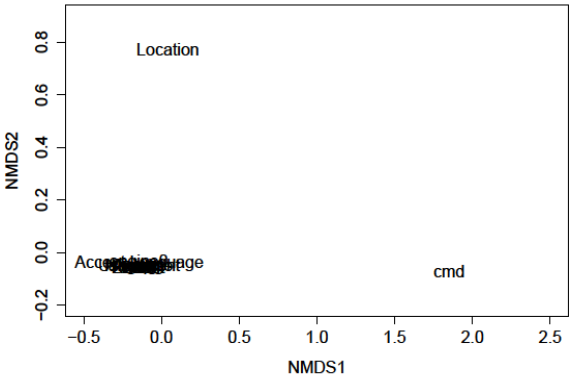


Figure A4.11 – 3rd Run- StartPage-Scenario 2- Jaccard Index-NMDS-Query Parameters

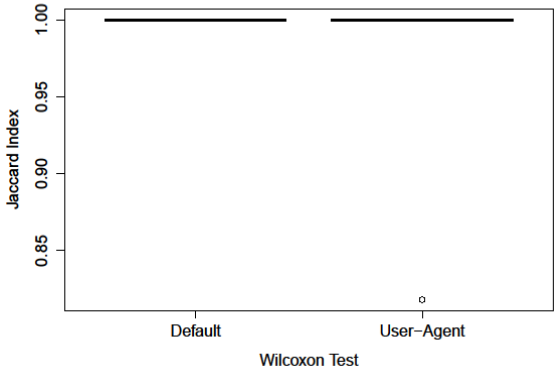


Figure A4.9 – 3rd Run- StartPage-Scenario 2- Jaccard Index-Wilcoxon Test-Default vs. User-Agent

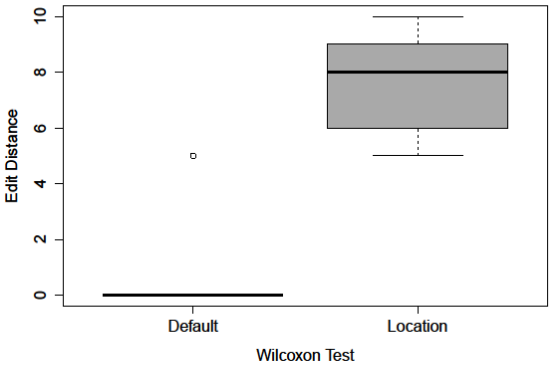


Figure A4.12– 3rd Run- StartPage-Scenario 2- Edit Distance-Wilcoxon Test-Default vs. Location

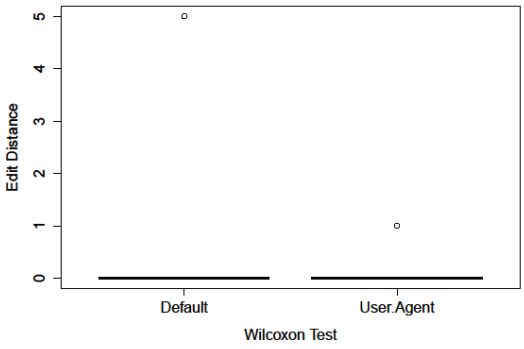


Figure A4.13 – 3rd Run- StartPage-Scenario 2- Edit Distance- Wilcoxon Test- Default vs. User-Agent

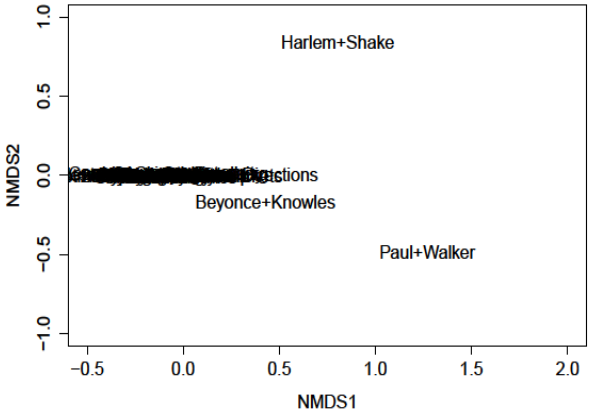


Figure A4.14 – 3rd Run- StartPage-Scenario 2- Edit Distance-NMDS-Query Terms

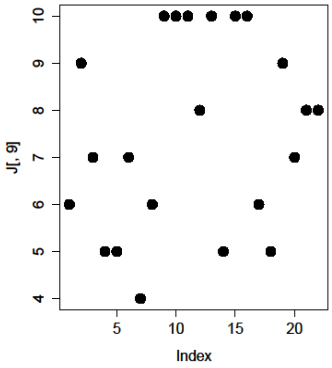


Figure A4.15– 3rd Run- StartPage-Scenario 1- Edit Distance-Location Parameter Variance

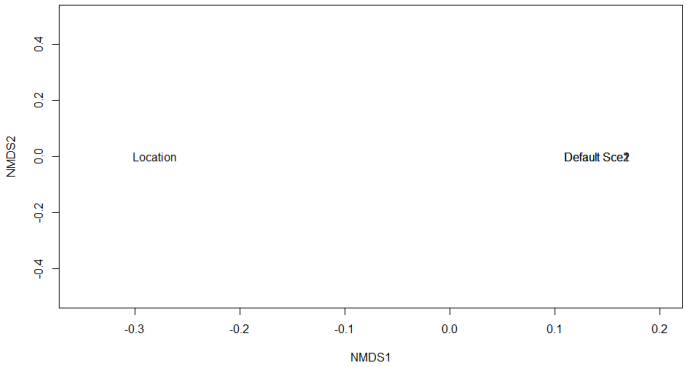


Figure A4.16– 1st Run - StartPage-Scenario 2- Jaccard Index- Default Variance vs Location-Spoofed Variance

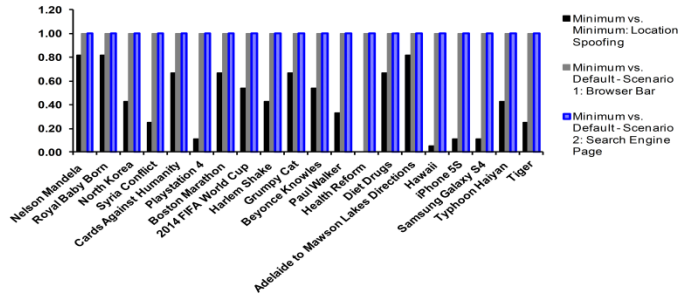


Figure A4.17– 1st Run - StartPage- Jaccard Index- Scenarios vs. Query Terms Variance

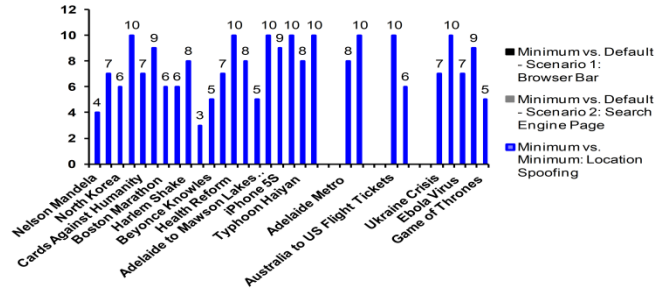


Figure A4.18– 1st Run - StartPage- Edit Distance- Scenarios vs. Query Terms Variance

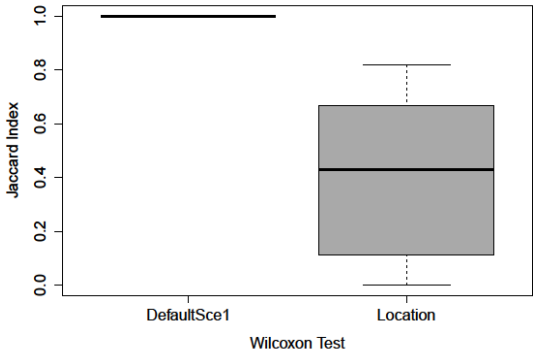


Figure A4.19 – 1st Run- StartPage- Jaccard Index- Wilcoxon Test- Default Scenario 1 vs. Location

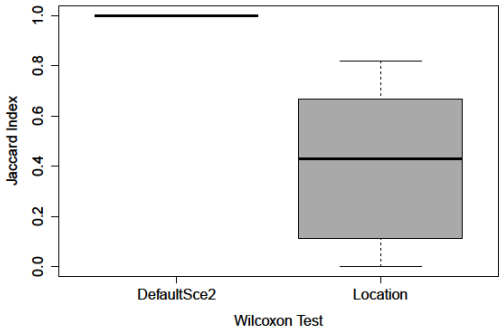


Figure A4.20 – 1st Run- StartPage- Jaccard Index- Wilcoxon Test- Default Scenario 2 vs. Location

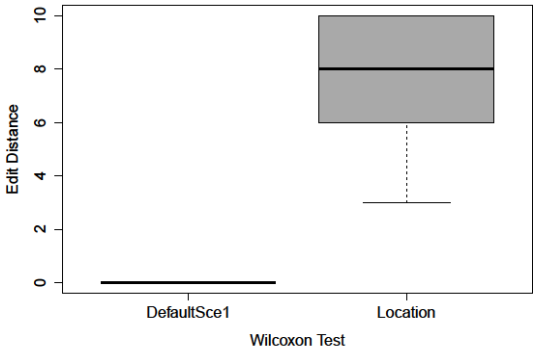


Figure A4.21 – 1st Run- StartPage- Edit Distance- Wilcoxon Test- Default Scenario 1 vs. Location

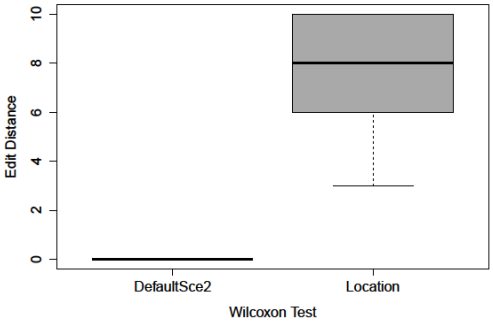


Figure A4.22 – 1st Run- StartPage- Jaccard Index- Wilcoxon Test- Default Scenario 2 vs. Location

Figures 4.23 – A4.47 DuckDuckGo

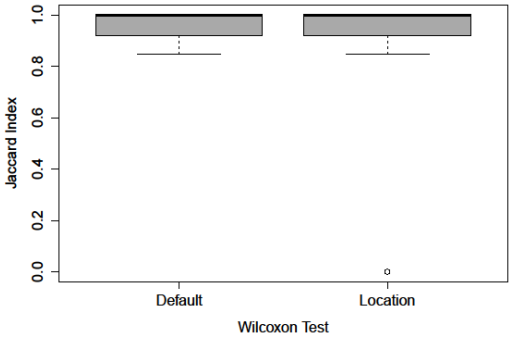


Figure A4.23 – 3rd Run-DuckDuckGo- Scenario 1-Jaccard Index-Wilcoxon Test- Default vs. Location

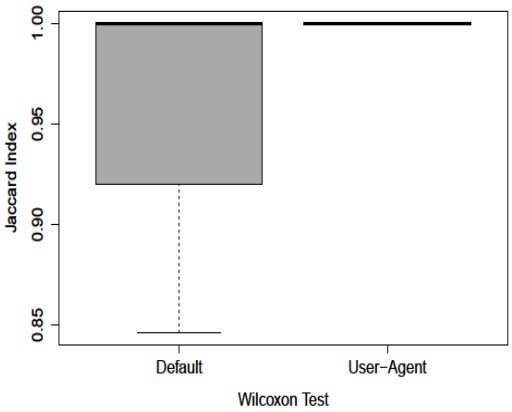


Figure A4.24 – 3rd Run-DuckDuckGo- Scenario 1-Jaccard Index-Wilcoxon Test- Default vs. Location

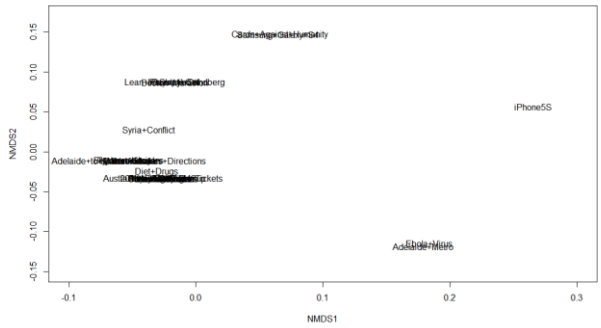


Figure A4.25 – 3rd Run-DuckDuckGo-Scenario 1-Jaccard Index-NMDS-Query Terms

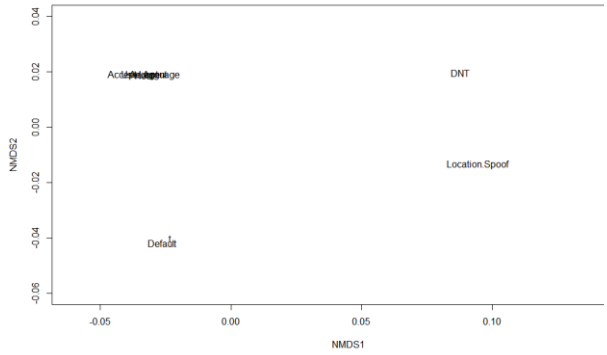


Figure A4.26 – 3rd Run-DuckDuckGo-Scenario 1-Jaccard Index-NMDS-Query Parameters

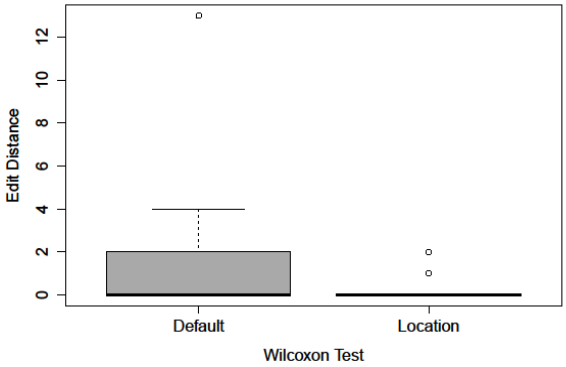


Figure A4.27 – 3rd Run-DuckDuckGo-Scenario 1-Edit Distance-Wilcoxon Test-Default vs. Location

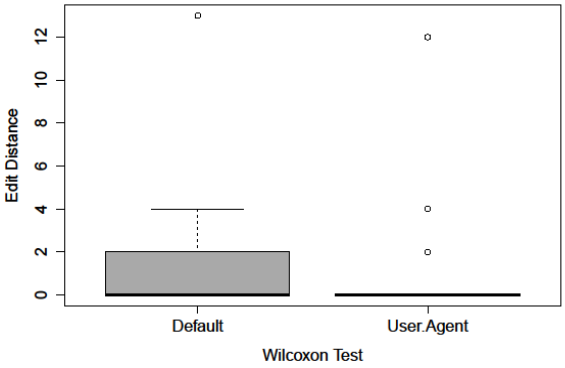


Figure A4.28 – 3rd Run-DuckDuckGo-Scenario 1-Edit Distance-Wilcoxon Test-Default vs. User-Agent

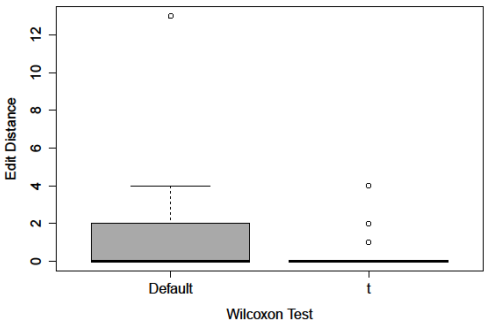


Figure A4.29 – 3rd Run-DuckDuckGo-Scenario 1-Edit Distance-Wilcoxon Test-Default vs. t

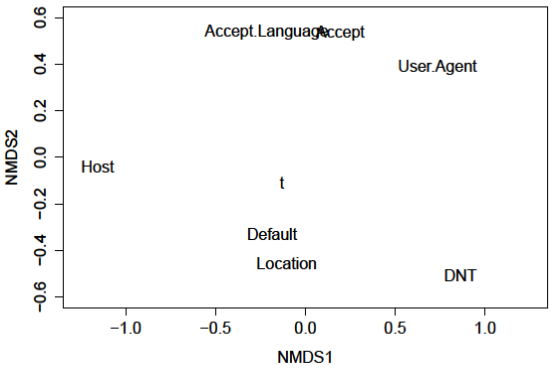


Figure A4.30 – 3rd Run-DuckDuckGo-Scenario 1-Edit Distance-NMDS-Query Parameters

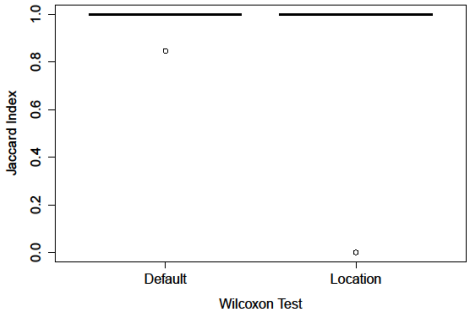


Figure A4.31 – 3rd Run-DuckDuckGo-Scenario 2-Jaccard Index-Wilcoxon Test-Default vs. Location

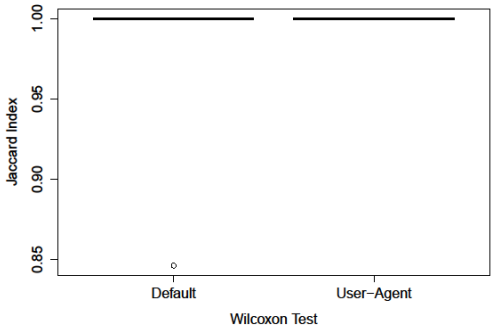


Figure A4.32 – 3rd Run-DuckDuckGo-Scenario 2-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

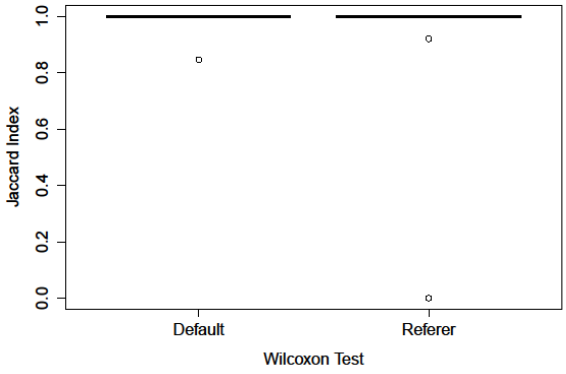


Figure A4.33 – 3rd Run-DuckDuckGo-Scenario 2-Jaccard Index-Wilcoxon Test-Default vs. Referer

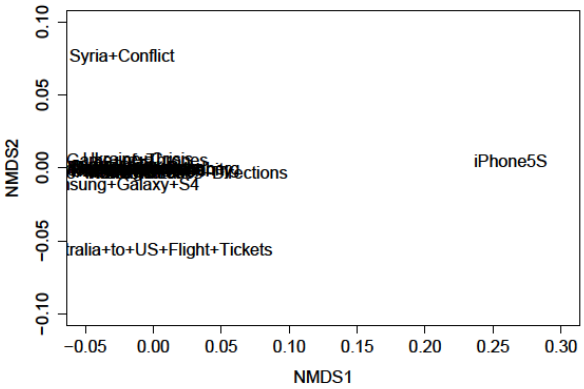


Figure A4.34 – 3rd Run-DuckDuckGo-Scenario 2-Jaccard Index-NMDS-Query Terms

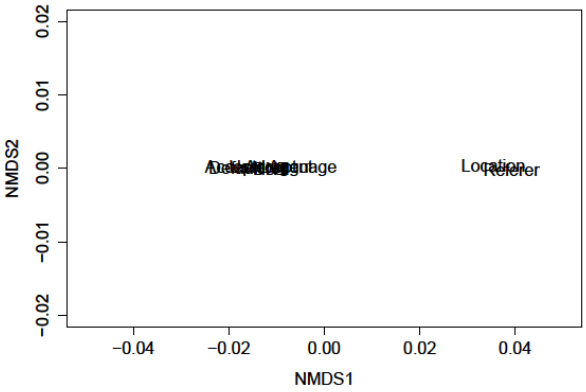


Figure A4.35 – 3rd Run-DuckDuckGo-Scenario 2-Jaccard Index-NMDS-Query Parameters



Figure A4.36 – 3rd Run-DuckDuckGo-Scenario 2- Edit Distance-Wilcoxon Test-Default vs. Location

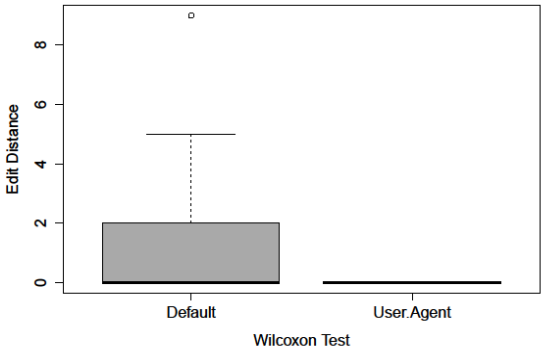


Figure A4.37 – 3rd Run-DuckDuckGo-Scenario 2- Edit Distance-Wilcoxon Test-Default vs. User-Agent

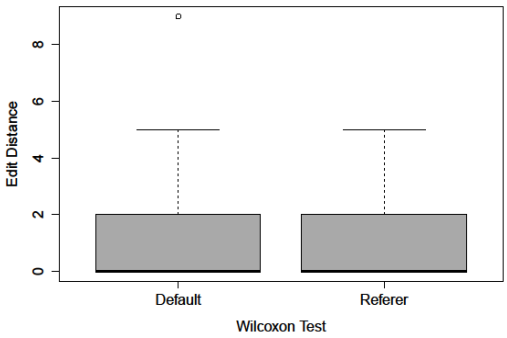


Figure A4.38 – 3rd Run-DuckDuckGo-Scenario 2- Edit Distance-Wilcoxon Test-Default vs. Referer

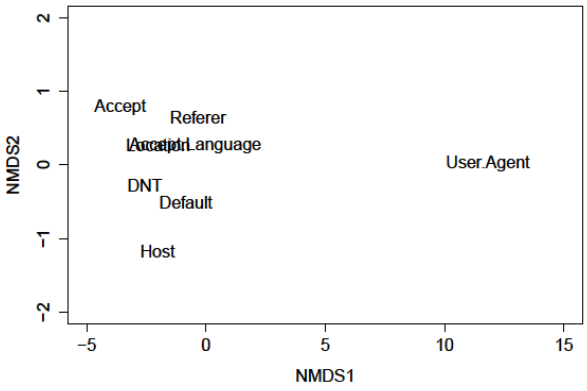


Figure A4.39 – 3rd Run-DuckDuckGo-Scenario 2- Edit Distance-NMDS-Query Parameters

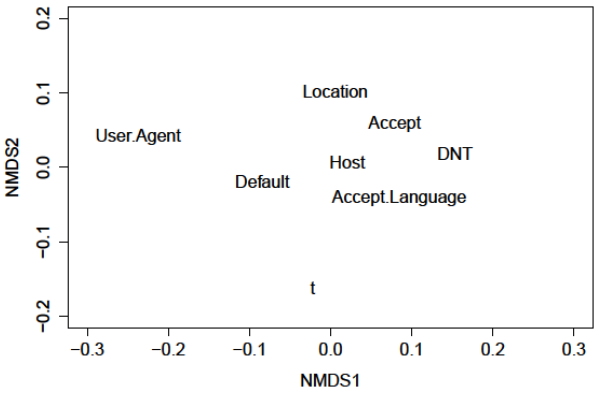


Figure A4.40 – 2nd Run-DuckDuckGo-Scenario 1-Jaccard Index-NMDS-Query Parameters

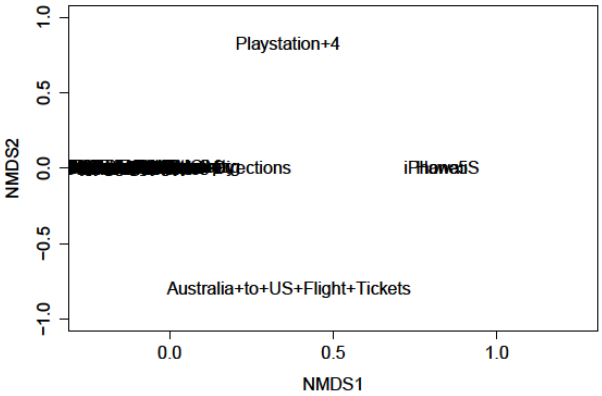


Figure A4.41 – 2nd Run-DuckDuckGo-Scenario 1-Jaccard Index-NMDS-Query Terms

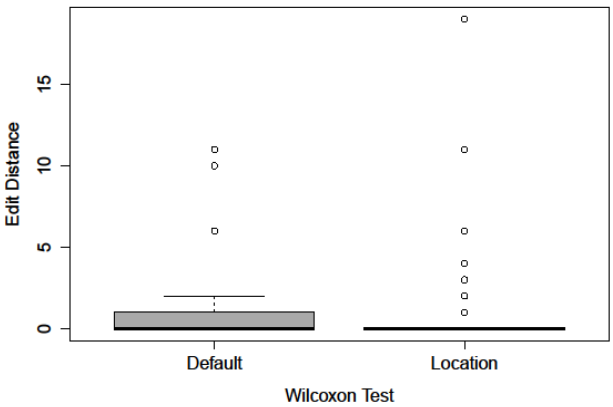


Figure A4.42 – 2nd Run-DuckDuckGo-Scenario 1- Edit Distance-Wilcoxon Test-Default vs. Location



Figure A4.43 – 2nd Run-DuckDuckGo-Scenario 1- Edit Distance-Wilcoxon Test-Default vs. User-Agent

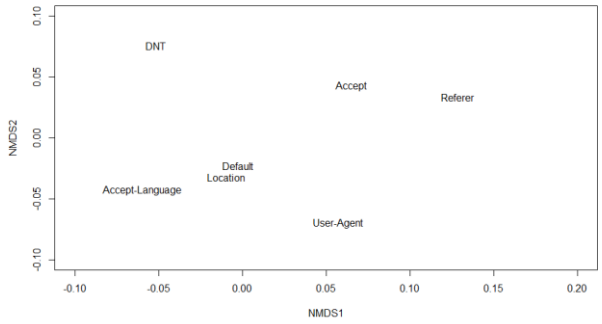


Figure A4.44 – 2nd Run-DuckDuckGo-Scenario 2-Jaccard Index-NMDS-Query Parameters

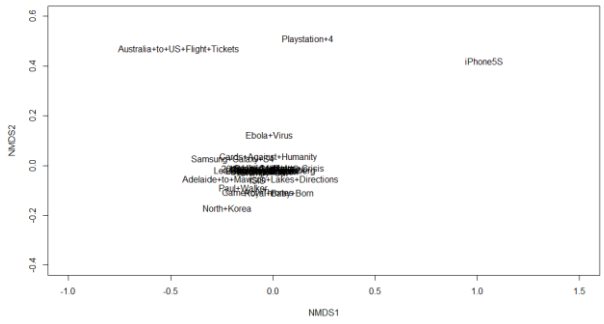


Figure A4.45 – 2nd Run-DuckDuckGo-Scenario 2-Jaccard Index-NMDS-Query Terms

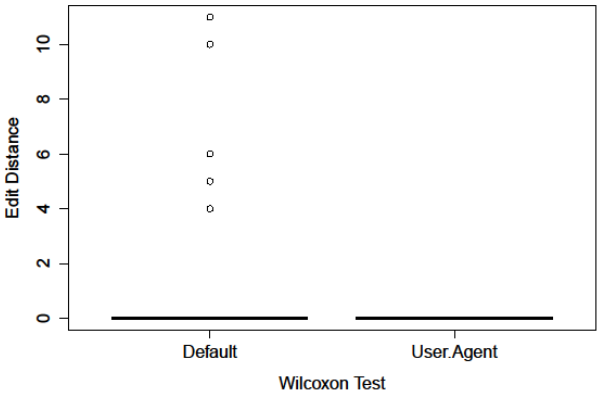


Figure A4.46 – 2nd Run-DuckDuckGo-Scenario 2- Edit Distance-Wilcoxon Test-Default vs. User-Agent

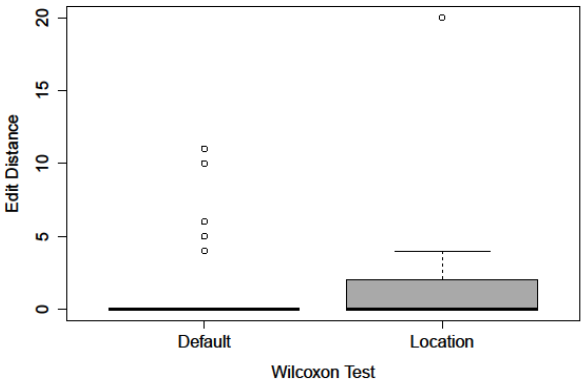


Figure A4.47 – 2nd Run-DuckDuckGo-Scenario 2- Edit Distance-Wilcoxon Test-Default vs. User-Agent

Figures A4.48-A4.81 Ixquick

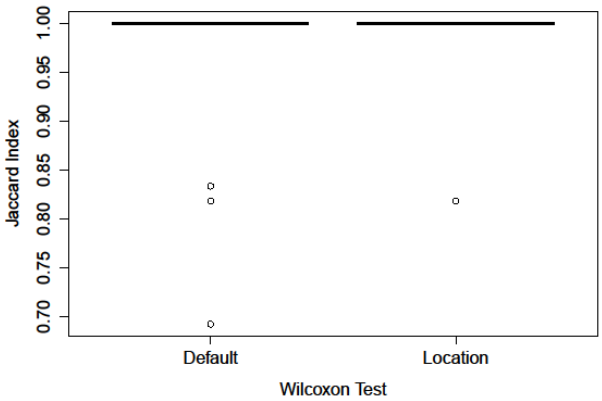


Figure A4.48 – 2nd Run-Ixquick-Scenario 1-Jaccard Index-Wilcoxon Test-Default vs. Location

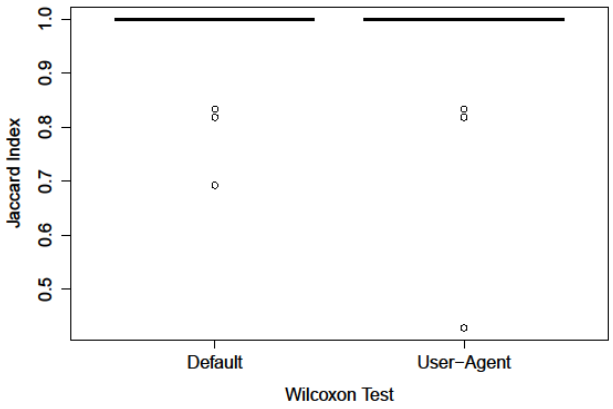


Figure A4.49 – 2nd Run-Ixquick-Scenario 1-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

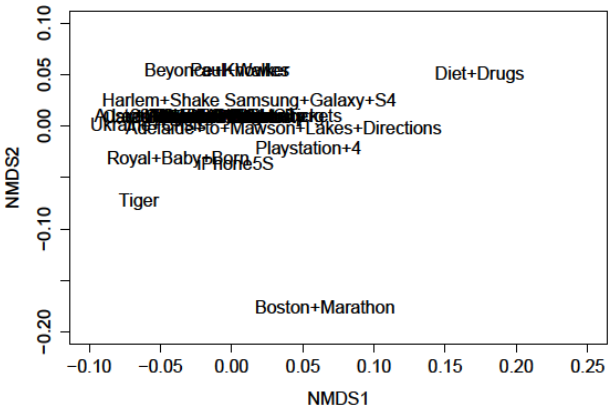


Figure A4.52 – 2nd Run-Ixquick-Scenario 1-Jaccard Index-NMDS- Query Terms

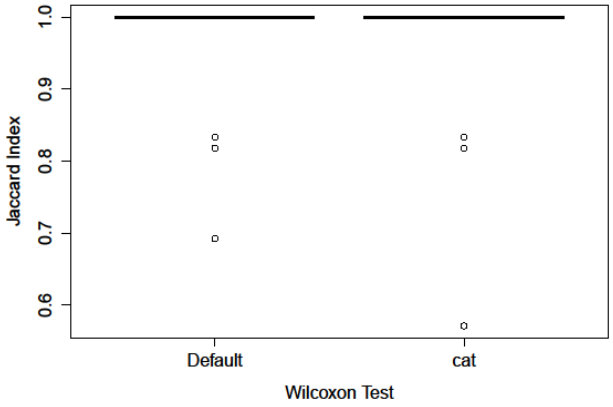


Figure A4.50 – 2nd Run-Ixquick-Scenario 1-Jaccard Index-Wilcoxon Test-Default vs. cat

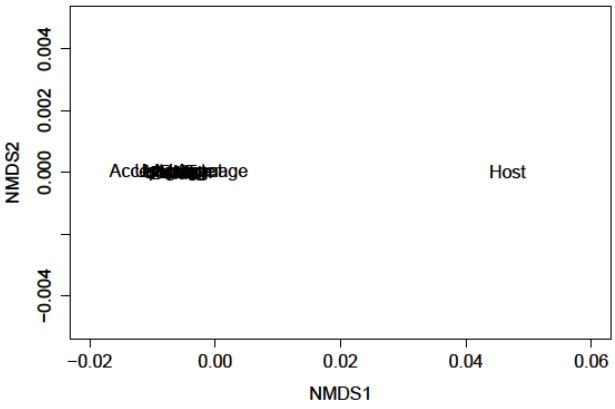


Figure A4.53 – 2nd Run-Ixquick-Scenario 1-Jaccard Index-NMDS- Query Parameters

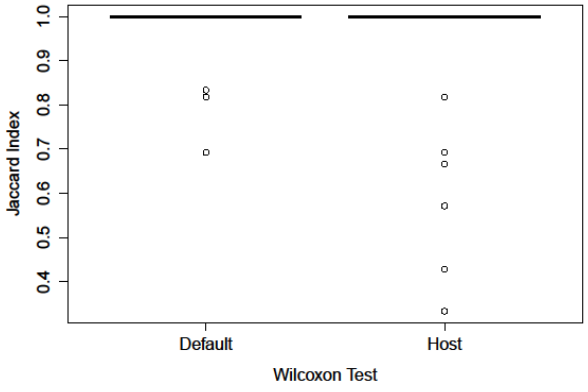


Figure A4.51 – 2nd Run-Ixquick-Scenario 1-Jaccard Index-Wilcoxon Test-Default vs. Host

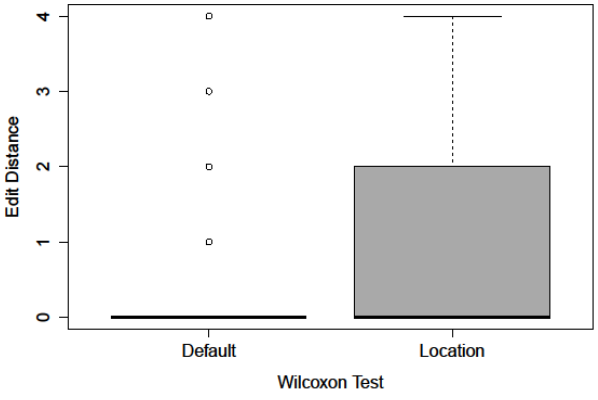


Figure A4.54– 2nd Run-Ixquick-Scenario 1-Edit Distance-Wilcoxon Test-Default vs. Location

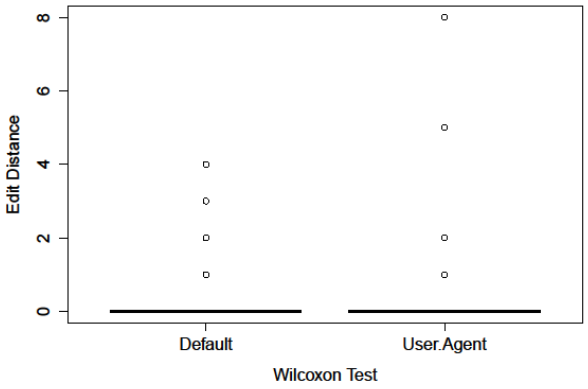


Figure A4.55– 2nd Run-Ixquick-Scenario 1- Edit Distance -Wilcoxon Test-Default vs. User-Agent

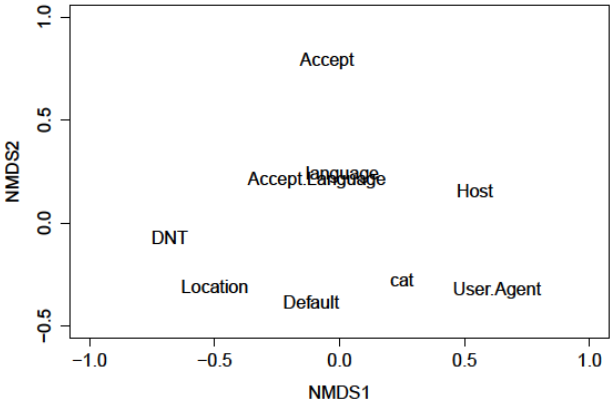


Figure A4.58– 2nd Run-Ixquick-Scenario 1- Edit Distance- NMDS- Query Parameters

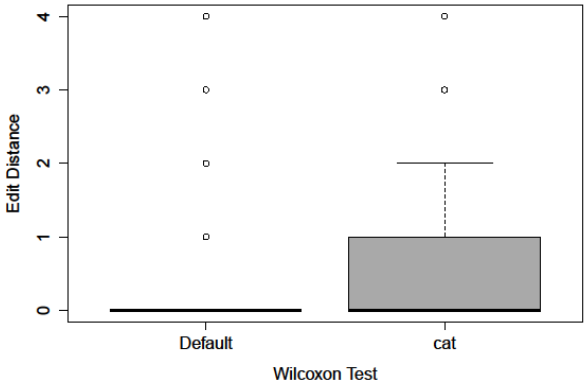


Figure A4.56– 2nd Run-Ixquick-Scenario 1- Edit Distance -Wilcoxon Test-Default vs. cat

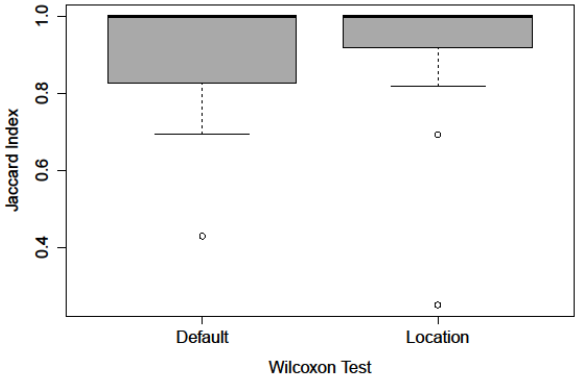


Figure A4.59– 2nd Run-Ixquick-Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. Location

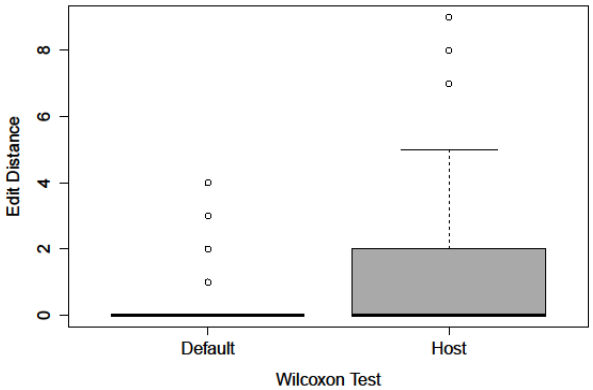


Figure A4.57– 2nd Run-Ixquick-Scenario 1- Edit Distance -Wilcoxon Test-Default vs. Host

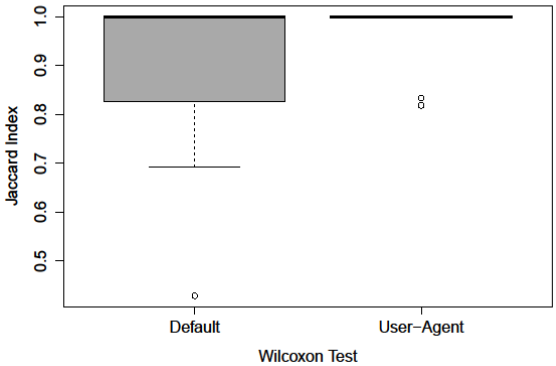


Figure A4.60– 2nd Run-Ixquick-Scenario 2-Jaccard Index-Wilcoxon Test- Default vs User-Agent

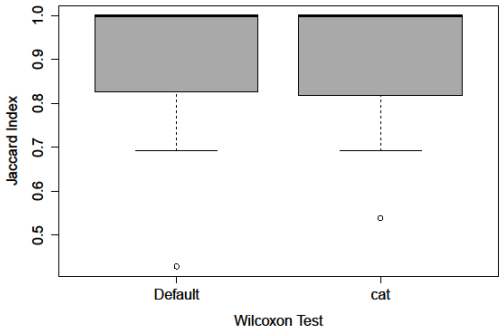


Figure A4.61– 2nd Run-Ixquick-Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. cat

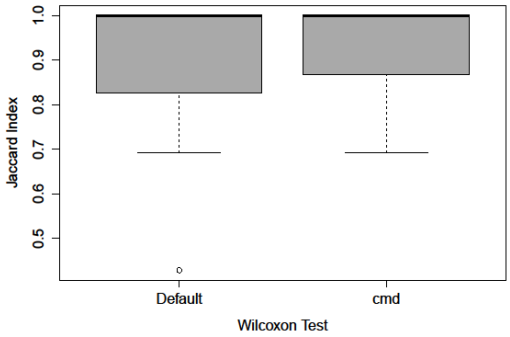


Figure A4.62.– 2nd Run-Ixquick-Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. cmd

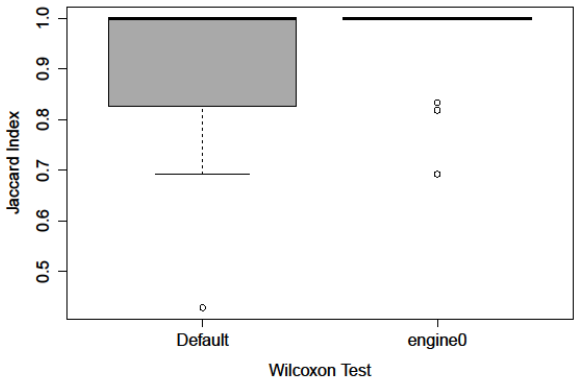


Figure A4.63– 2nd Run-Ixquick-Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. engine0

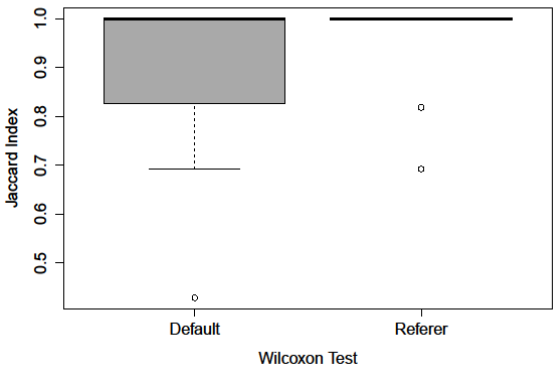


Figure A4.64– 2nd Run-Ixquick-Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. Referer

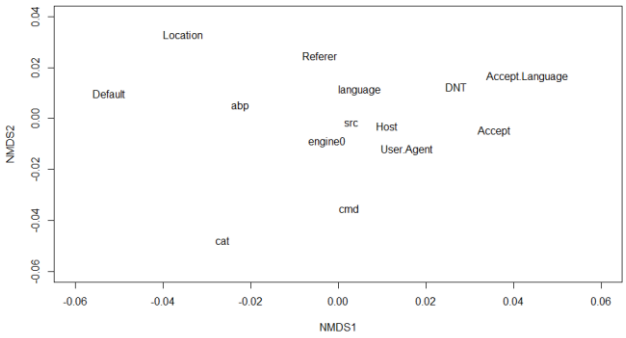


Figure A4.65– 2nd Run-Ixquick-Scenario 2-Jaccard Index-NMDS-Query Parameters

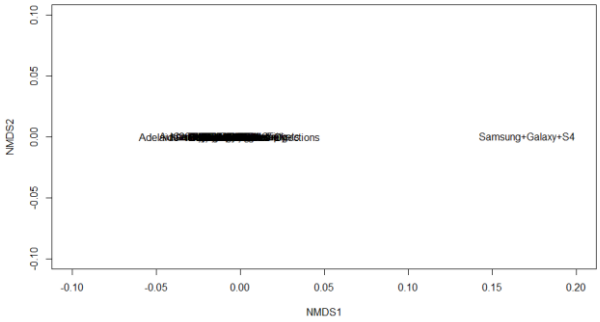


Figure A4.66– 2nd Run-Ixquick-Scenario 2-Jaccard Index-NMDS-Query Terms

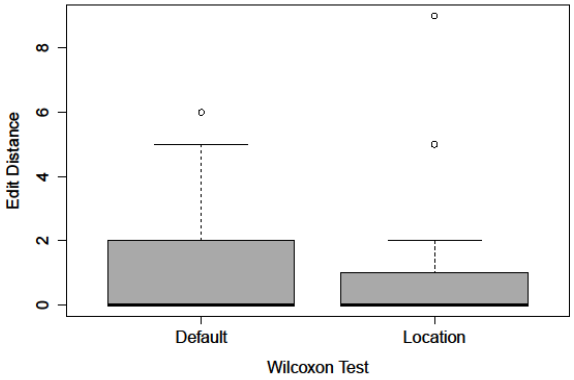


Figure A4.67– 2nd Run-Ixquick-Scenario 2-Edit Distance- Wilcoxon Test- Default vs. Location

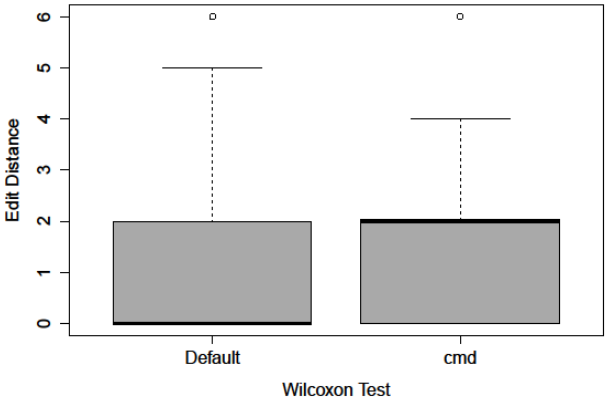


Figure A4.70– 2nd Run-Ixquick-Scenario 2-Edit Distance- Wilcoxon Test- Default vs. cmd

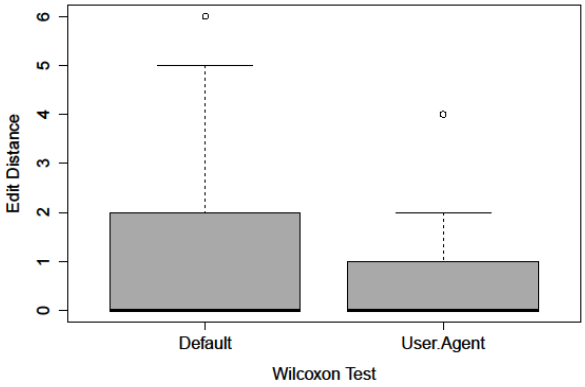


Figure A4.68– 2nd Run-Ixquick-Scenario 2-Edit Distance- Wilcoxon Test- Default vs. User-Agent

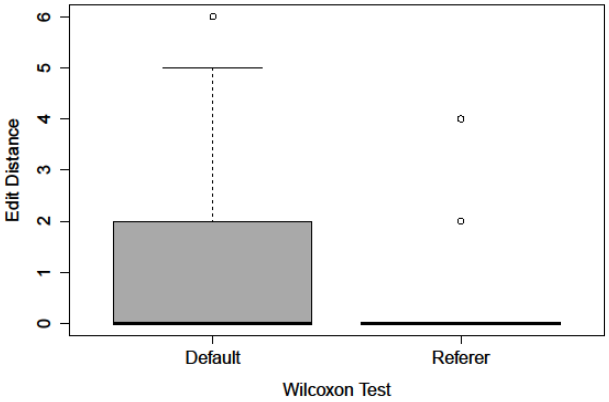


Figure A4.71– 2nd Run-Ixquick-Scenario 2-Edit Distance- Wilcoxon Test- Default vs. Referer

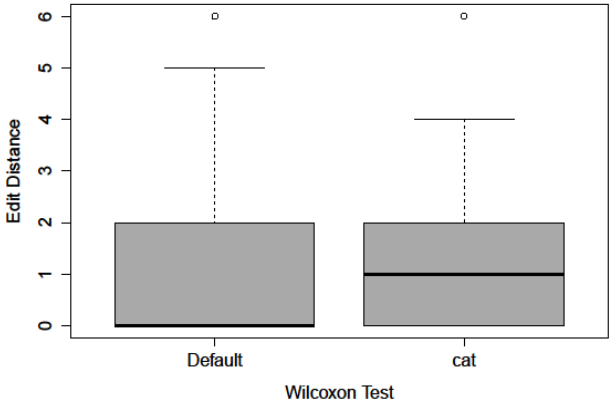


Figure A4.69– 2nd Run-Ixquick-Scenario 2-Edit Distance- Wilcoxon Test- Default vs. cat

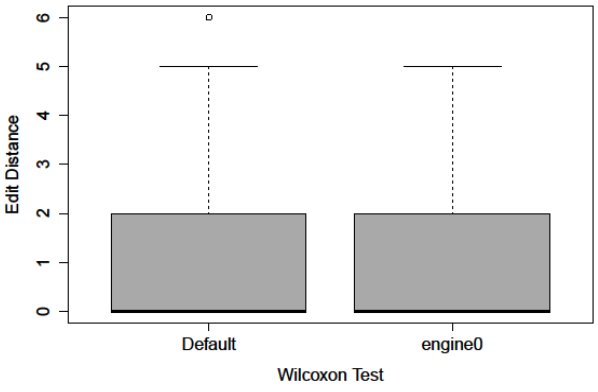


Figure A4.72– 2nd Run-Ixquick-Scenario 2-Edit Distance- Wilcoxon Test- Default vs. engine0

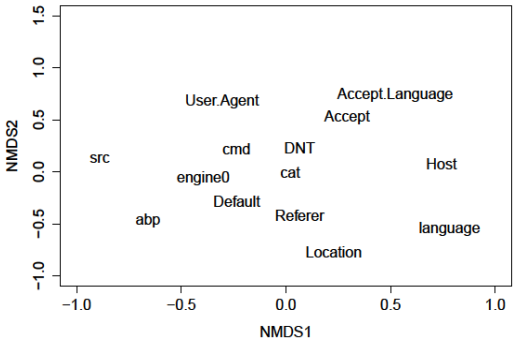


Figure A4.73– 2nd Run-Ixquick-Scenario 2-Edit Distance- NMDS- Query Parameters

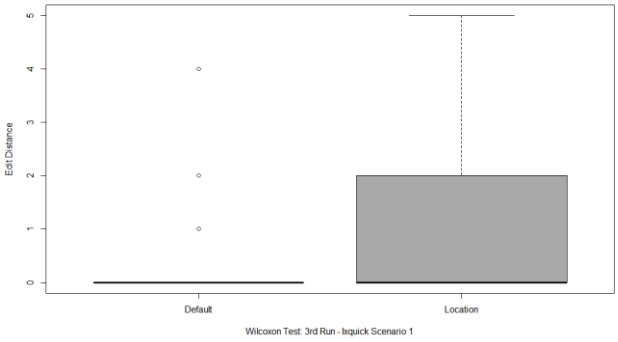


Figure A4.77– 3rd Run-Ixquick-Scenario 1-Edit Distance- Wilcoxon Test- Default vs. Location

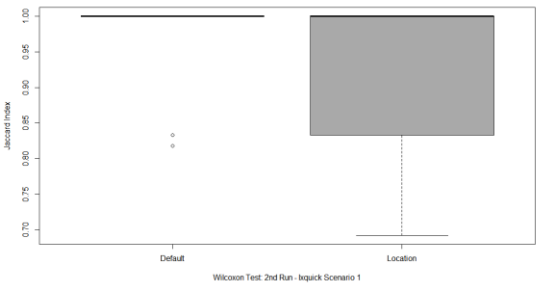


Figure A4.74– 3rd Run-Ixquick-Scenario 1-Jaccard Index- Wilcoxon Test- Default vs. Location

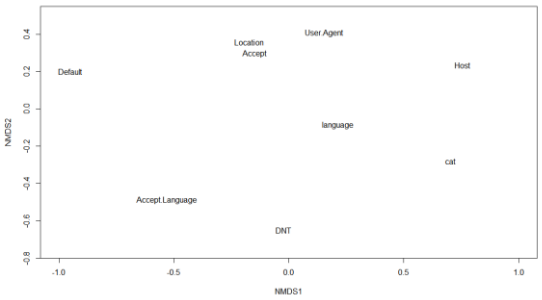


Figure A4.78– 3rd Run-Ixquick-Scenario 1-Edit Distance- NDMS- Query Parameters

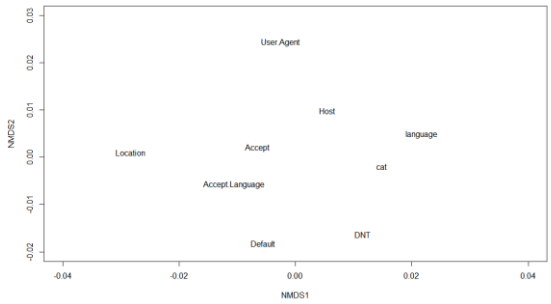


Figure A4.75– 3rd Run-Ixquick-Scenario 1-Jaccard Index- NMDS- Query Parameters

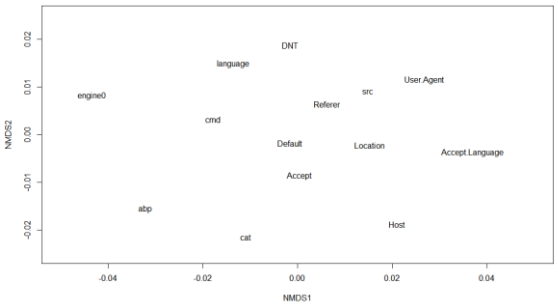


Figure A4.79a– 3rd Run-Ixquick-Scenario 2-Jaccard Index- NDMS- Query Parameters

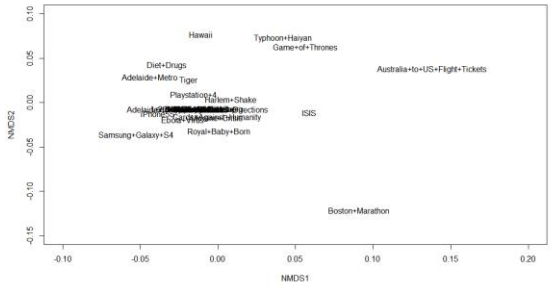


Figure A4.76– 3rd Run-Ixquick-Scenario 1-Jaccard Index- NMDS- Query Terms

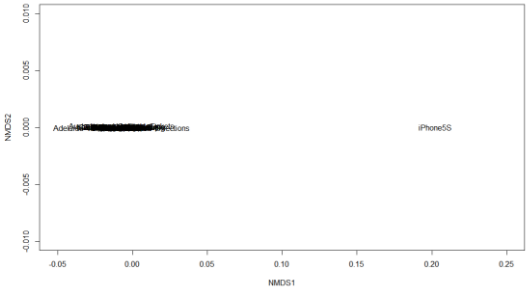


Figure A4.79b– 3rd Run-Ixquick-Scenario 2-Jaccard Index- NDMS- Query Terms

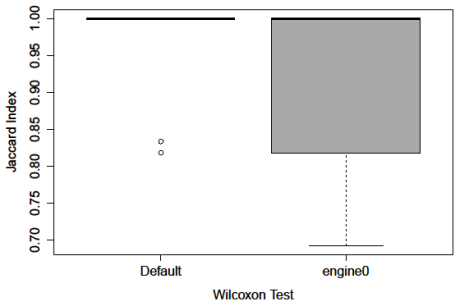


Figure A4.80– 3rd Run-Ixquick-Scenario 2-Jaccard Index- Wilcoxon-Default vs. engine0

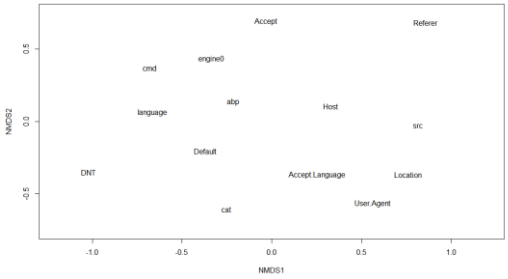


Figure A4.81– 3rd Run-Ixquick-Scenario 2-Edit Distance- NDMS- Query Parameters

Figures A4.82 – A4.206 Google

3rdRun – January 2016

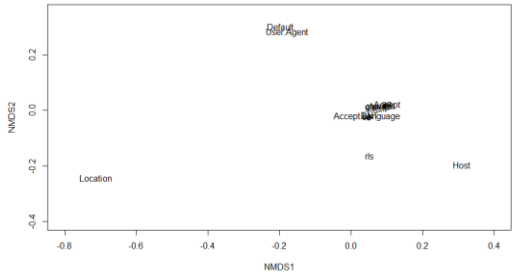


Figure A4.82– 3rd Run-Google- Scenario 1-Jaccard Index- NMDS- Query Parameters

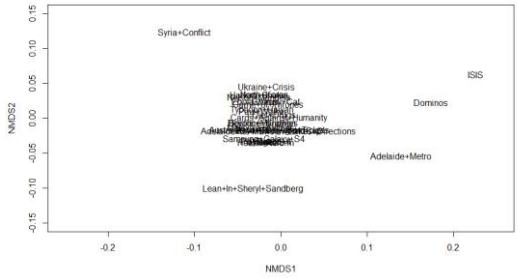


Figure A4.83– 3rd Run-Google- Scenario 1-Jaccard Index- NMDS- Query Terms

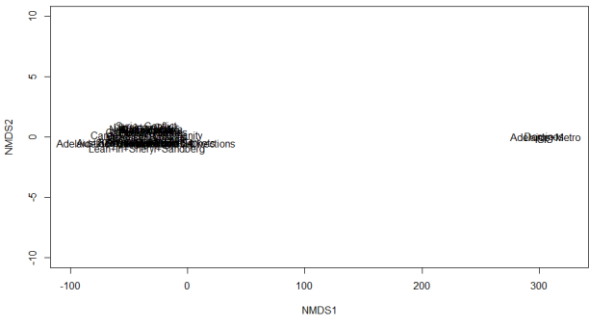


Figure A4.84– 3rd Run-Google- Scenario 1-Edit Distance- NMDS- Query Terms

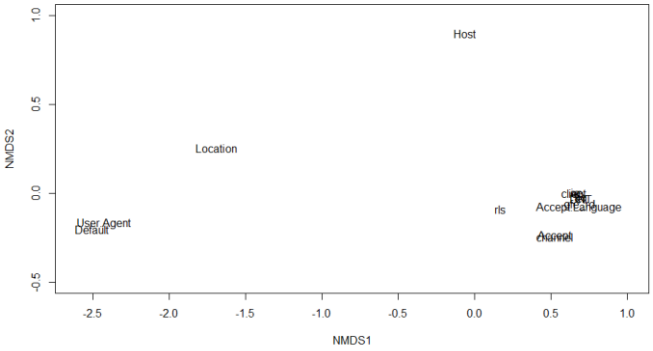


Figure A4.85– 3rd Run-Google- Scenario 1-Edit Distance- NMDS- Query Parameters

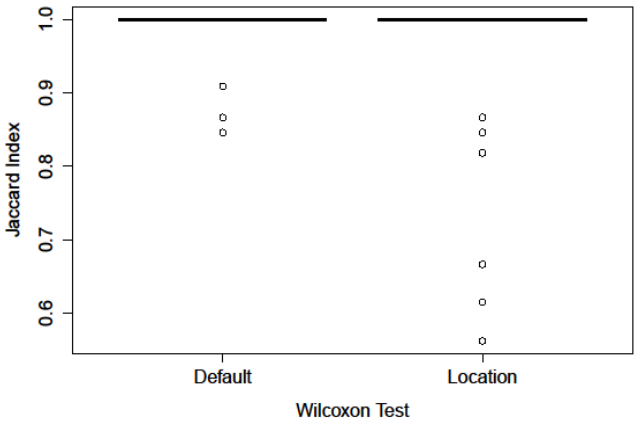


Figure A4.86– 3rd Run-Google- Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. Location

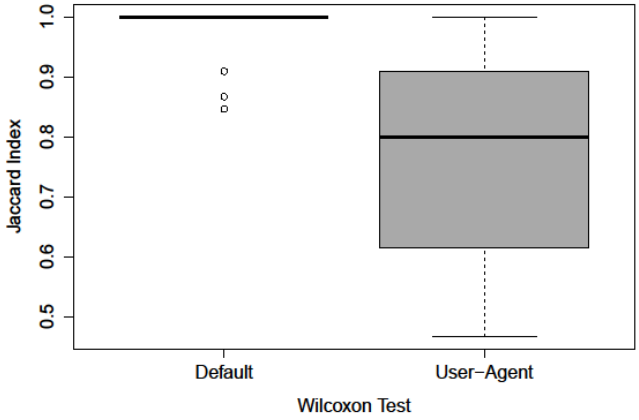


Figure A4.87– 3rd Run-Google- Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

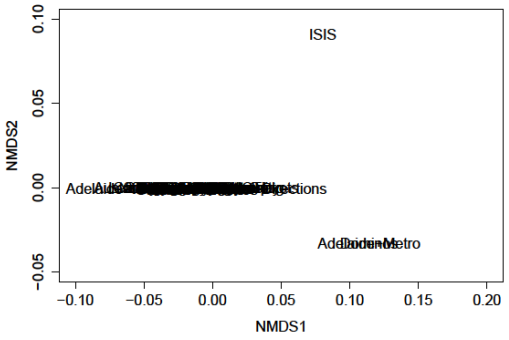


Figure A4.88– 3rd Run-Google- Scenario 2-Jaccard Index- NMDS- Query Terms

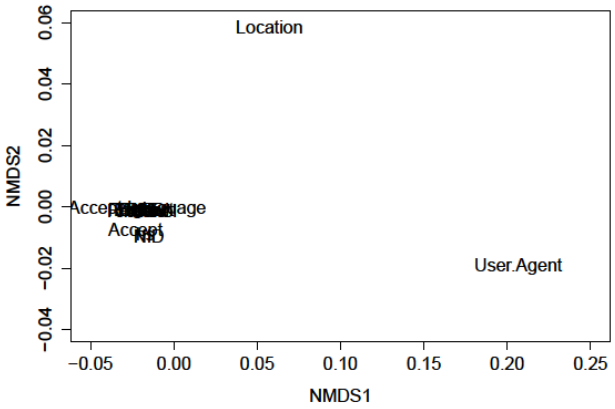


Figure A4.89– 3rd Run-Google- Scenario 2-Jaccard Index- NMDS- Query Parameters

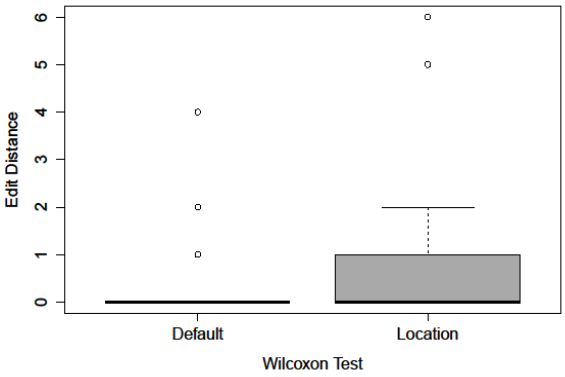


Figure A4.90– 3rd Run-Google- Scenario 2-Edit Distance-Wilcoxon Test- Default vs. Location

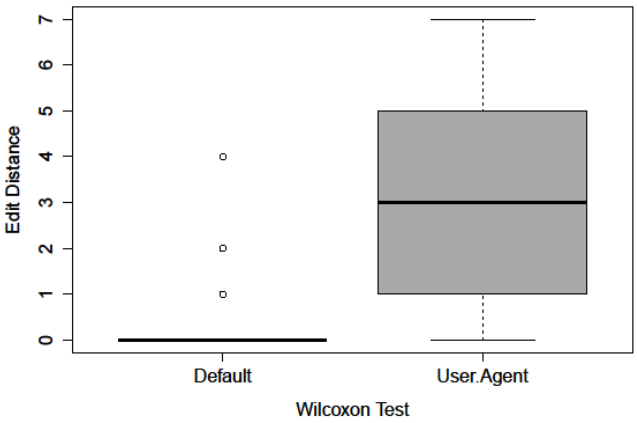


Figure A4.91– 3rd Run-Google- Scenario 2-Edit Distance-Wilcoxon Test- Default vs. User-Agent

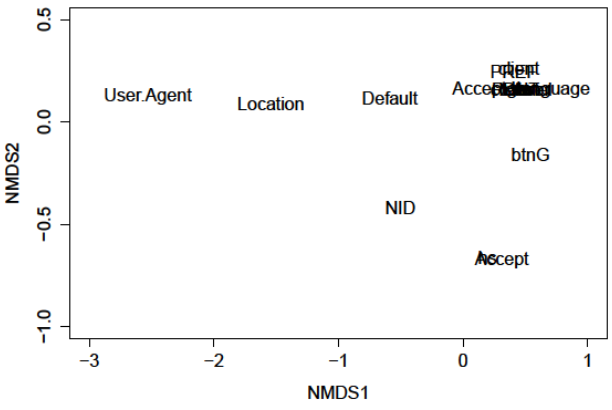


Figure A4.92– 3rd Run-Google- Scenario 2-Edit Distance-NMDS-Query Parameters

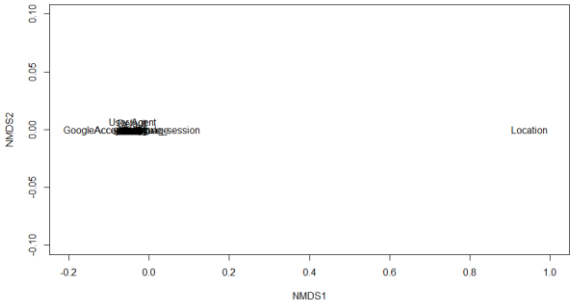


Figure A4.93– 3rd Run-Google- Scenario 3-Jaccard Index- NMDS- Query Parameters

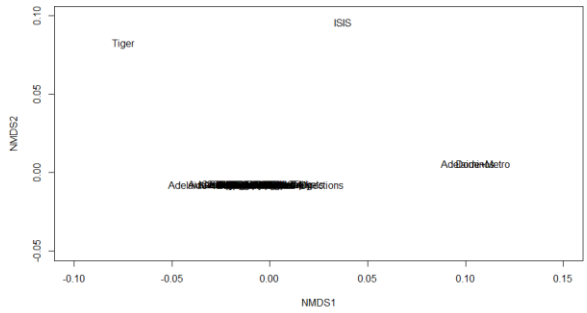


Figure A4.94– 3rd Run-Google- Scenario 3-Jaccard Index- NMDS- Query Terms

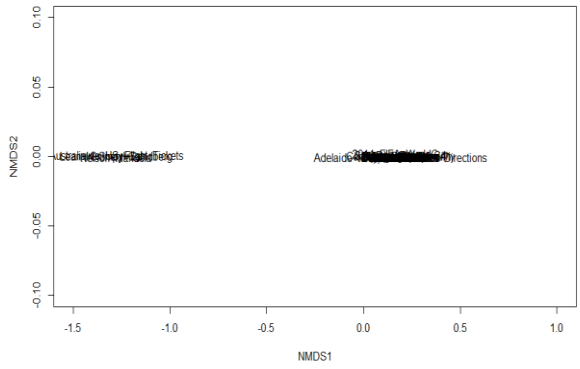


Figure A4.95– 3rd Run-Google- Scenario 3-Edit Distance-NMDS-Query Parameters

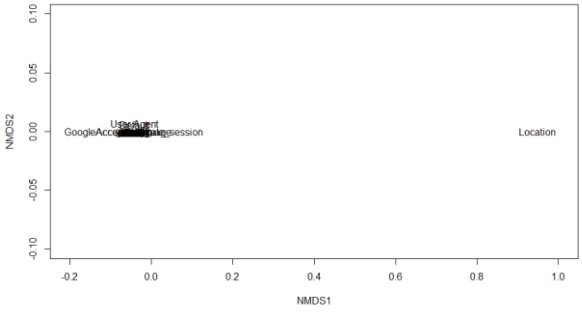


Figure A4.96– 3rd Run-Google- Scenario 3-Edit Distance-NMDS-Query Terms

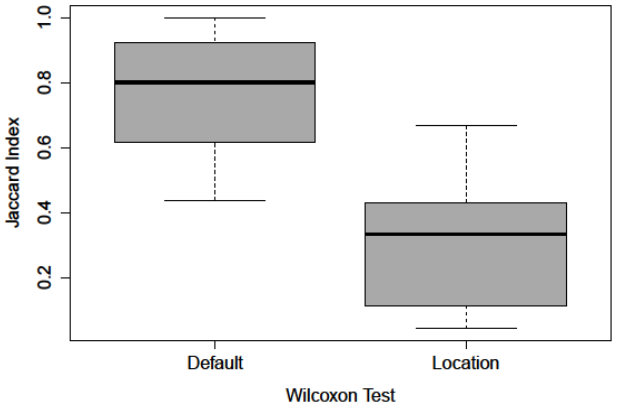


Figure A4.97– 3rd Run-Google- Scenario 4-Jaccard Index-Wilcoxon Test- Default vs. Location

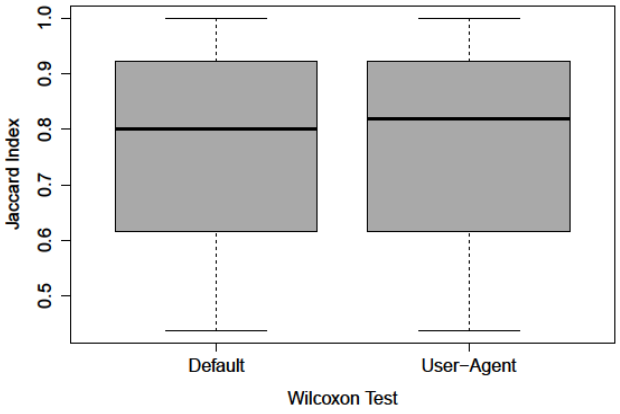


Figure A4.98– 3rd Run-Google- Scenario 4-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

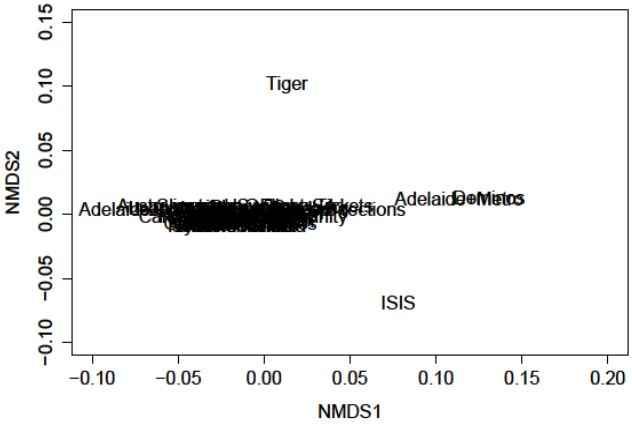


Figure A4.99– 3rd Run-Google- Scenario 4-Jaccard Index-NMDS- Query Terms

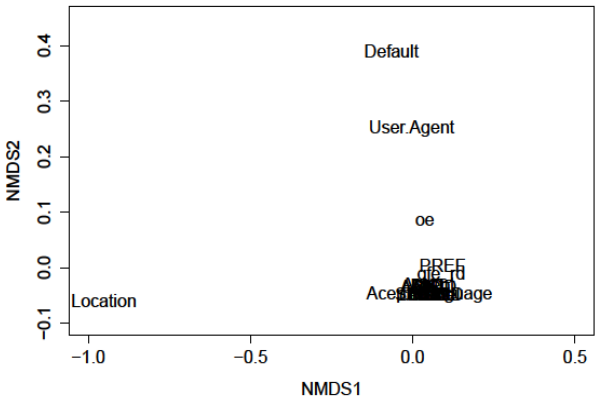


Figure A4.100– 3rd Run-Google- Scenario 4-Jaccard Index-NMDS- Query Parameters

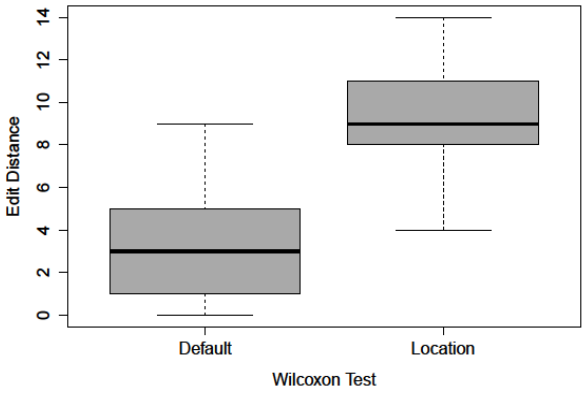


Figure A4.101– 3rd Run-Google- Scenario 4-Edit Distance-Wilcoxon Test- Default vs. Location

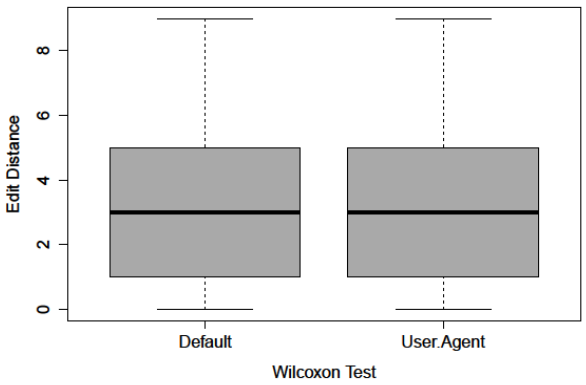


Figure A4.102– 3rd Run-Google- Scenario 4-Edit Distance-Wilcoxon Test- Default vs. User-Agent

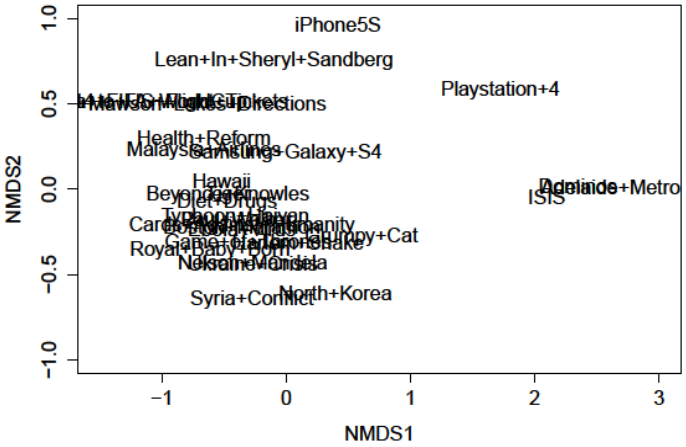


Figure A4.103– 3rd Run-Google- Scenario 4-Edit Distance-NMDS-Query Terms

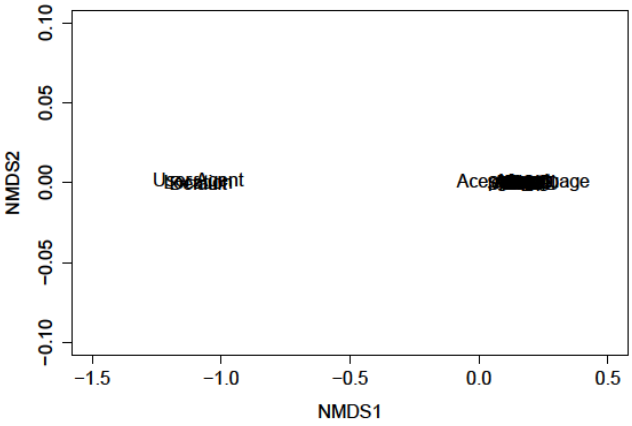


Figure A4.104– 3rd Run-Google- Scenario 4-Edit Distance-NMDS-Query Parameters

Randomised Effect – 3rd Run

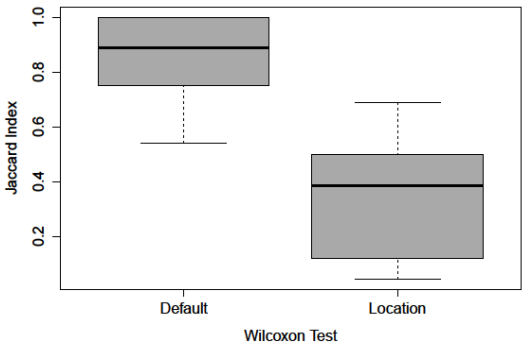


Figure A4.105– 3rd Run-Google- Randomised- Scenario 1-Jaccard Index- Wilcoxon Test- Default vs. Location

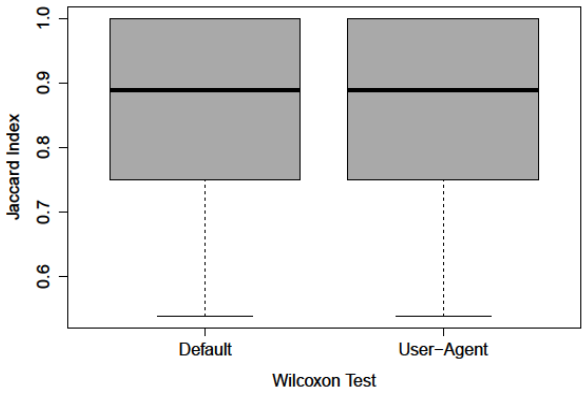


Figure A4.106– 3rd Run-Google-Randomised- Scenario 1-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

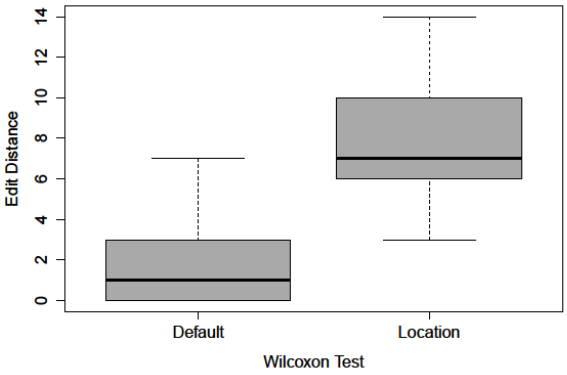


Figure A4.109– 3rd Run-Google-Randomised- Scenario 1-Edit Distance-Wilcoxon Test- Default vs. Location

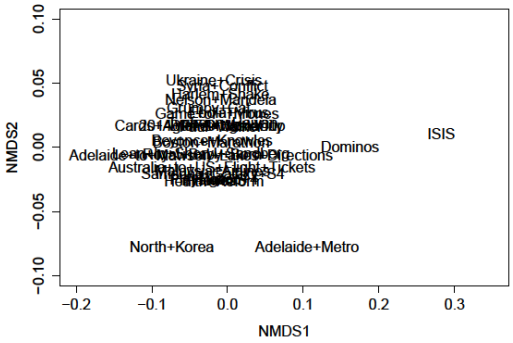


Figure A4.107– 3rd Run-Google-Randomised- Scenario 1-Jaccard Index-NMDS-Query Terms

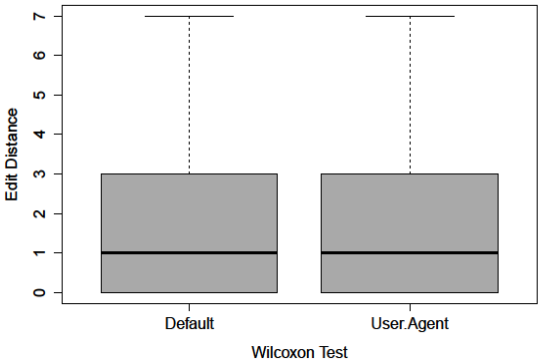


Figure A4.110– 3rd Run-Google-Randomised- Scenario 1-Edit Distance-Wilcoxon Test- Default vs. User-Agent

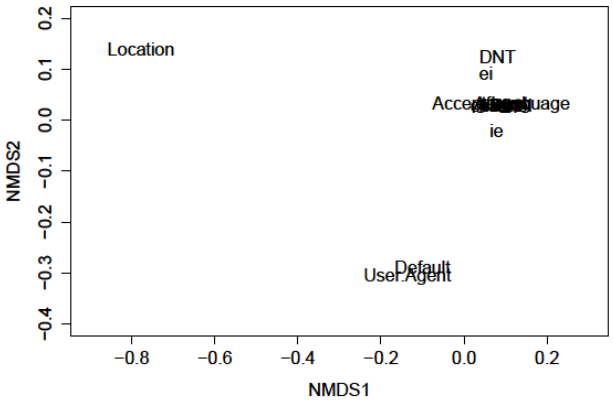


Figure A4.108– 3rd Run-Google-Randomised- Scenario 1-Jaccard Index-NMDS-Query Parameters

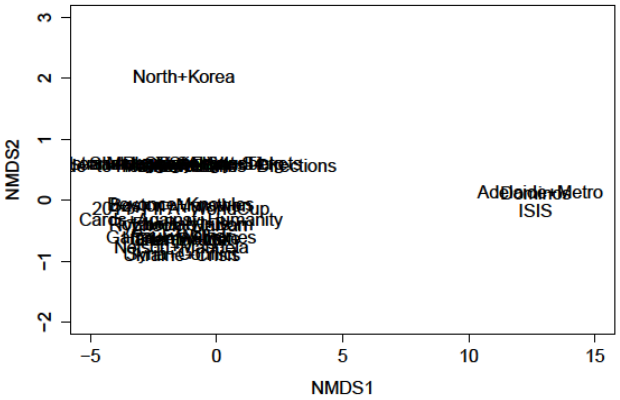


Figure A4.111– 3rd Run-Google-Randomised- Scenario 1-Edit Distance-NMDS-Query Terms

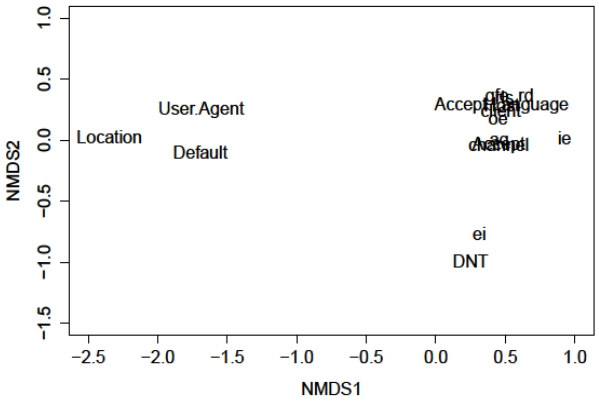


Figure A4.112– 3rd Run-Google-Randomised- Scenario 1-Edit Distance-NMDS-Query Parameters

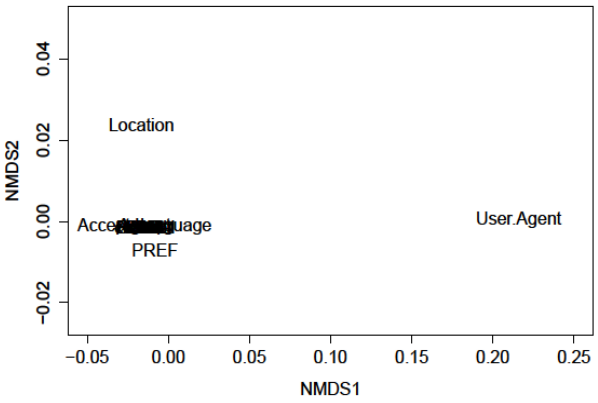


Figure A4.115– 3rd Run-Google-Randomised- Scenario 2-Jaccard Index-NMDS-Query Parameters

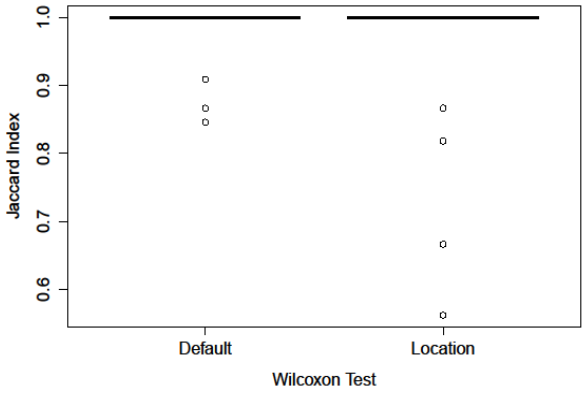


Figure A4.113– 3rd Run-Google-Randomised- Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. Location

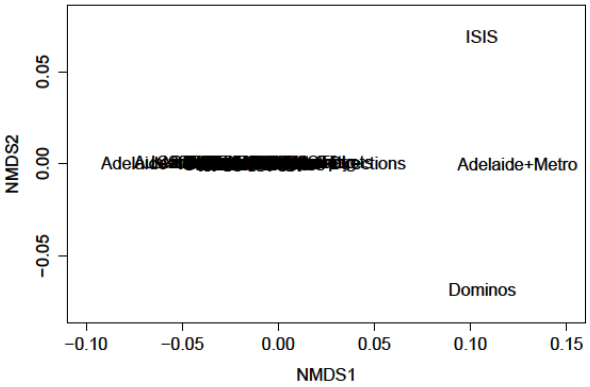


Figure A4.116– 3rd Run-Google-Randomised- Scenario 2-Jaccard Index-NMDS-Query Terms

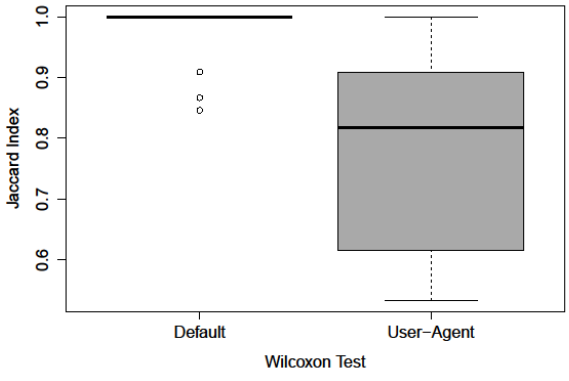


Figure A4.114– 3rd Run-Google-Randomised- Scenario 2-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

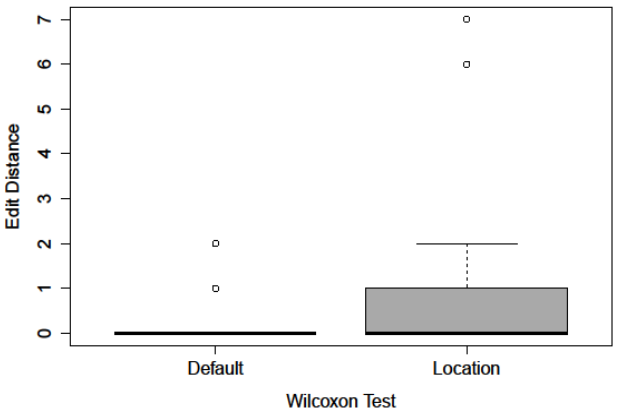


Figure A4.117– 3rd Run-Google-Randomised- Scenario 2-Edit Distance-Wilcoxon Test- Default vs. Location

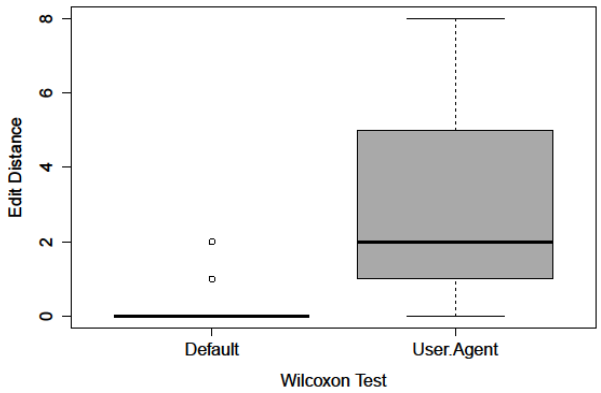


Figure A4.118– 3rd Run-Google-Randomised- Scenario 2-Edit Distance-Wilcoxon Test- Default vs. User-Agent

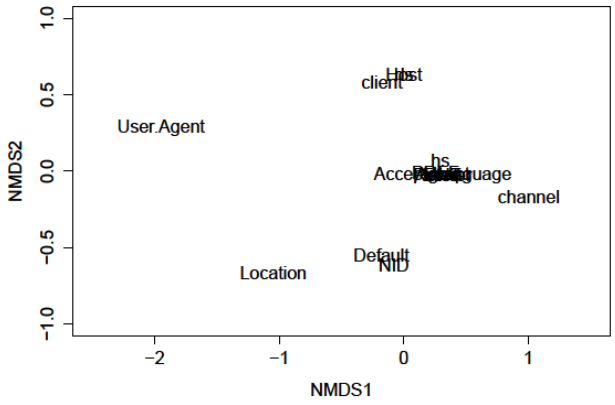


Figure A4.119– 3rd Run-Google-Randomised- Scenario 2-Edit Distance-NMDS-Query Parameters

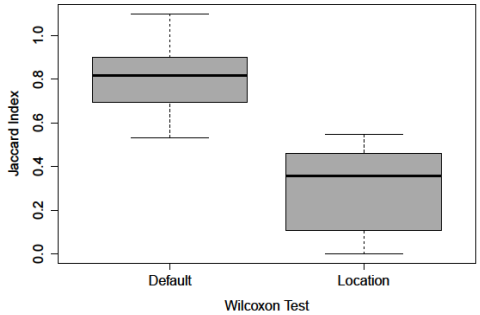


Figure A4.120– 3rd Run-Google-Randomised- Scenario 3-Jaccard Index-Wilcoxon Test- Default vs. Location

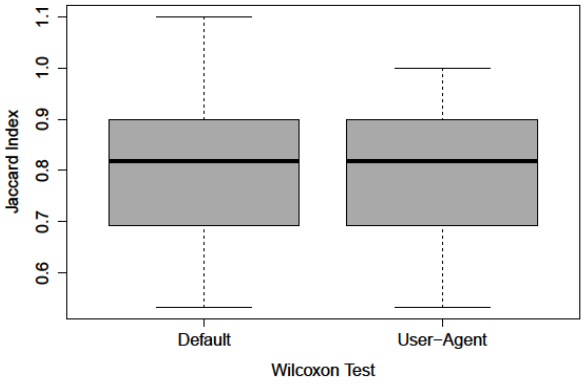


Figure A4.121– 3rd Run-Google-Randomised- Scenario 3-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

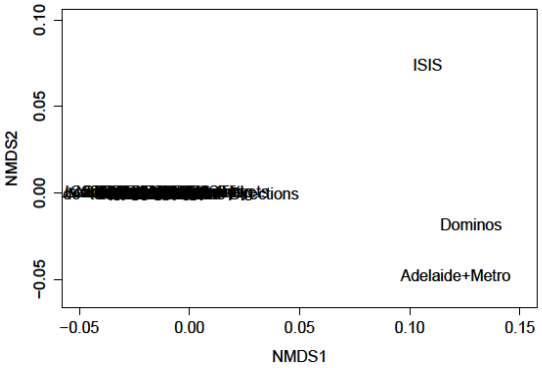


Figure A4.122– 3rd Run-Google-Randomised- Scenario 3-Jaccard Index-NMDS-Query Terms

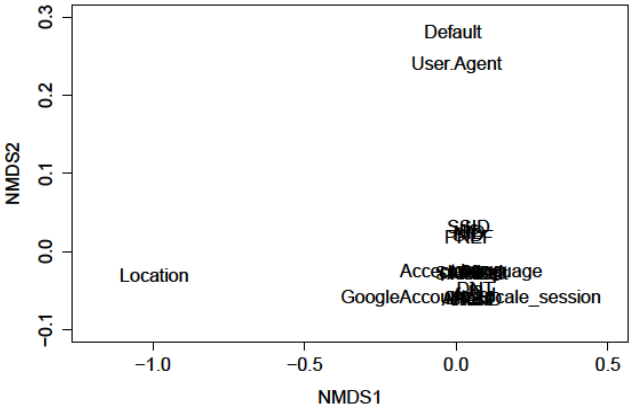


Figure A4.123– 3rd Run-Google-Randomised- Scenario 3-Jaccard Index-NMDS-Query Parameters

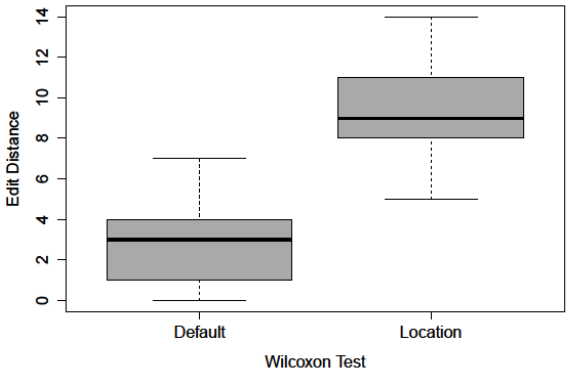


Figure A4.124– 3rd Run-Google-Randomised- Scenario 3-Edit Distance-Wilcoxon Test-Default vs. Location

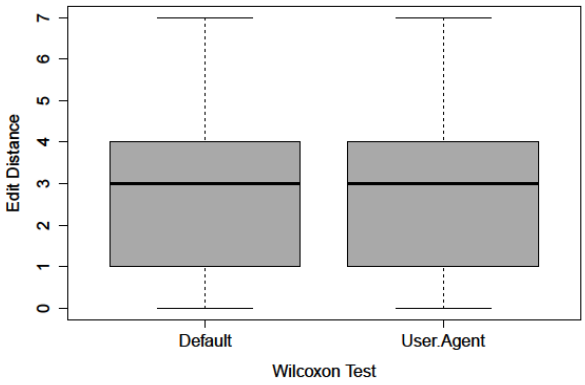


Figure A4.125– 3rd Run-Google-Randomised- Scenario 3-Edit Distance-Wilcoxon Test-Default vs. User-Agent

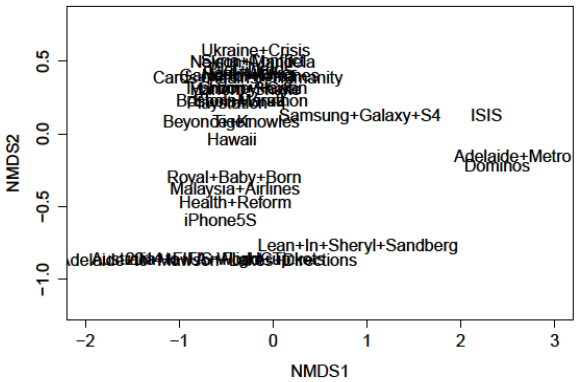


Figure A4.126– 3rd Run-Google-Randomised- Scenario 3-Edit Distance-NMDS-Query Terms

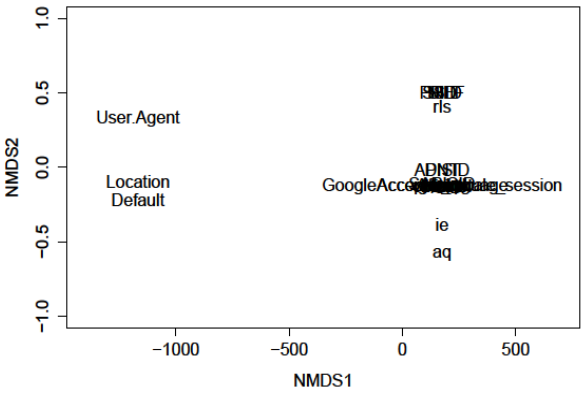


Figure A4.127– 3rd Run-Google-Randomised- Scenario 3-Edit Distance-NMDS-Query Parameters

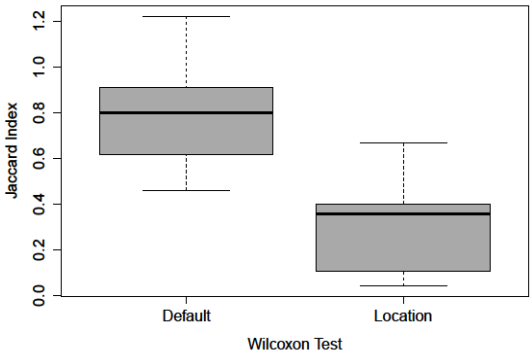


Figure A4.128– 3rd Run-Google-Randomised- Scenario 4-Jaccard Index-Wilcoxon Test- Default vs. Location

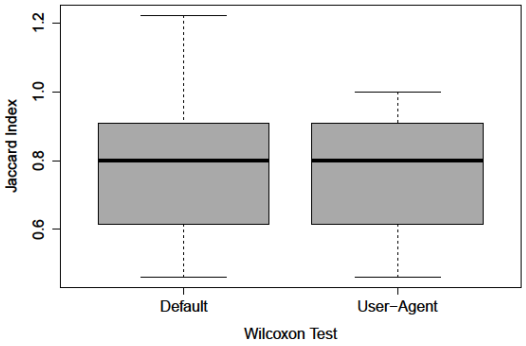


Figure A4.129– 3rd Run-Google-Randomised- Scenario 4-Jaccard Index-Wilcoxon Test- Default vs. Location

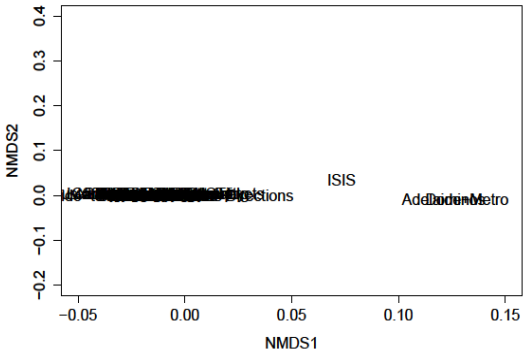


Figure A4.130– 3rd Run-Google-Randomised- Scenario 4-Jaccard Index-NMDS-Query Terms

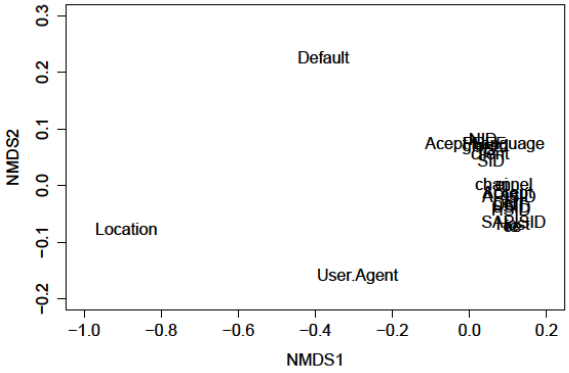


Figure A4.131– 3rd Run-Google-Randomised- Scenario 4-Jaccard Index-NMDS-Query Parameters

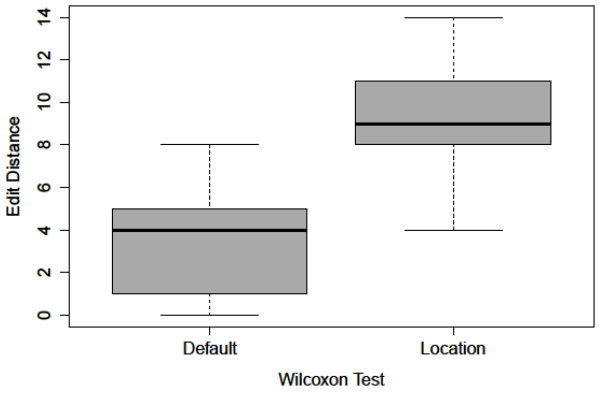


Figure A4.132– 3rd Run-Google-Randomised- Scenario 4-Edit Distance-Wilcoxon Test- Default vs. Location

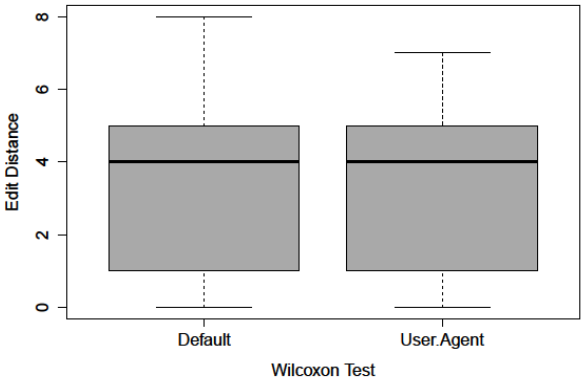


Figure A4.133– 3rd Run-Google-Randomised- Scenario 4-Edit Distance-Wilcoxon Test- Default vs. User-Agent

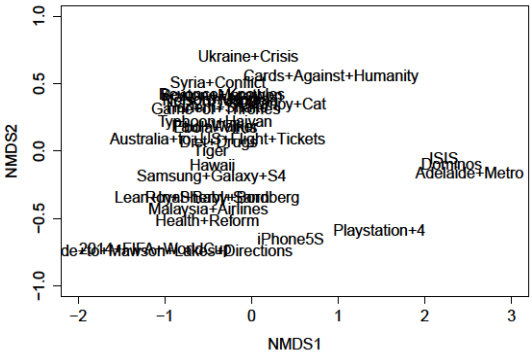


Figure A4.134– 3rd Run-Google-Randomised- Scenario 4-Edit Distance-NMDS-Query Terms

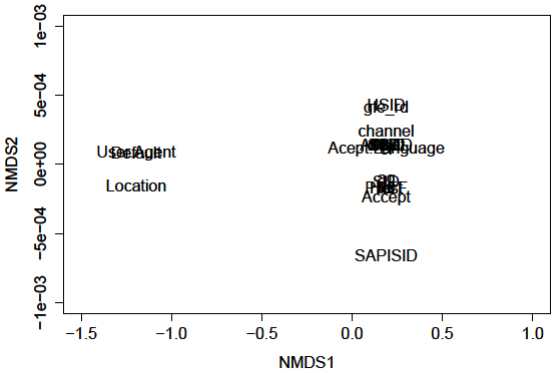


Figure A4.135– 3rd Run-Google-Randomised- Scenario 4-Edit Distance-NMDS-Query Parameters

Google: Exclusion Effect – 3rd Run

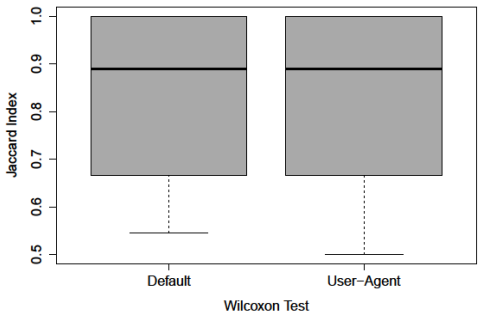


Figure A4.136a– 3rd Run-Google-Exclusion- Scenario 1- Jaccard Index-Wilcoxon Test- Default vs. User-Agent

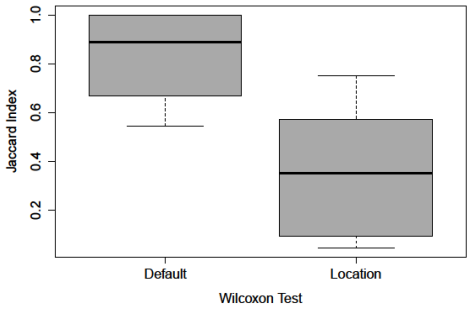


Figure A4.136b– 3rd Run-Google-Exclusion- Scenario 1- Jaccard Index- -Wilcoxon Test- Default vs. Location

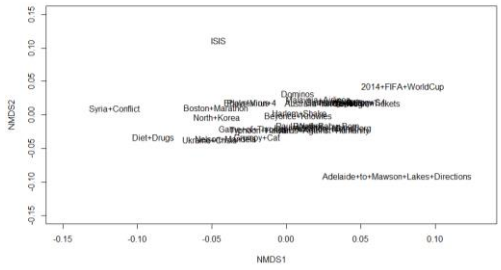


Figure A4.137a– 3rd Run-Google-Exclusion- Scenario 1- Jaccard Index-NMDS-Query Terms

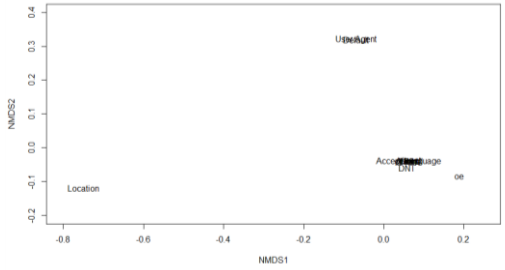


Figure A4.137b– 3rd Run-Google-Exclusion- Scenario 1- Jaccard Index-NMDS-Query Parameters

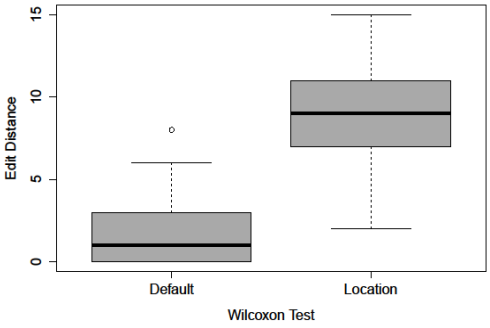


Figure A4.138– 3rd Run-Google-Exclusion- Scenario 1- Edit Distance-Wilcoxon Test-Default vs. Location

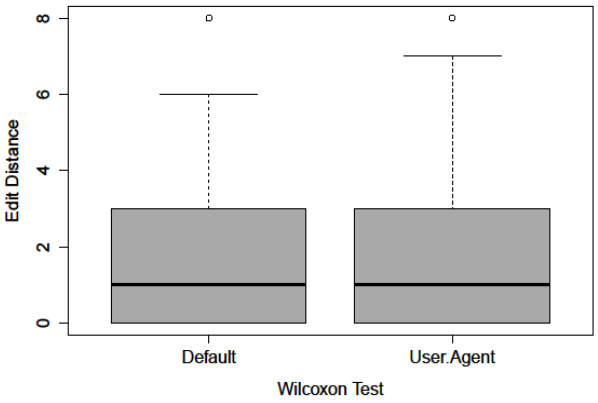


Figure A4.139– 3rd Run-Google-Exclusion- Scenario 1- Edit Distance-Wilcoxon Test-Default vs. User-Agent

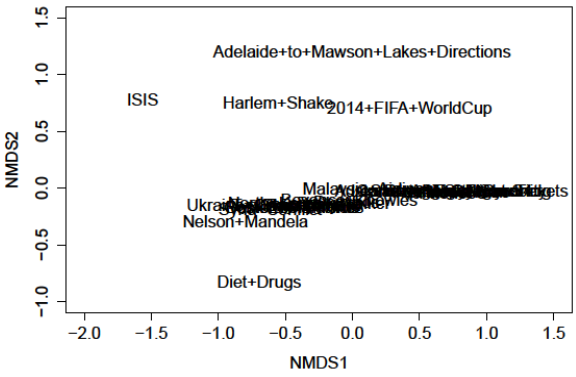


Figure A4.140– 3rd Run-Google-Exclusion- Scenario 1- Edit Distance-NMDS-Query Terms

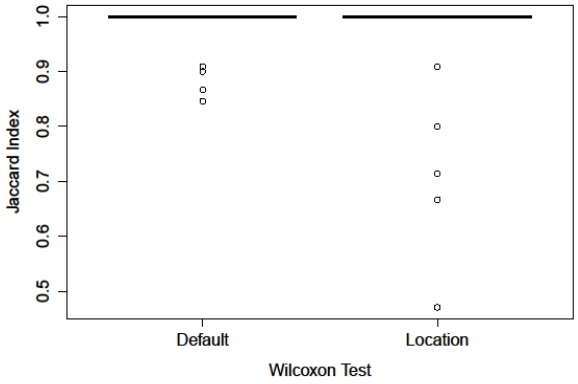


Figure A4.142– 3rd Run-Google-Exclusion- Scenario 2- Jaccard Index- Wilcoxon Test- Default vs. Location



Figure A4.143– 3rd Run-Google-Exclusion- Scenario 2- Jaccard Index- Wilcoxon Test- Default vs. User-Agent

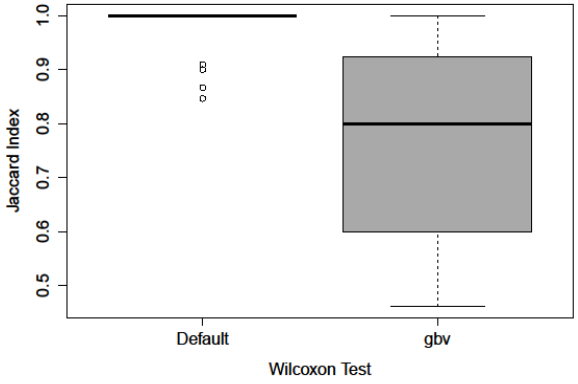


Figure A4.144– 3rd Run-Google-Exclusion- Scenario 2- Jaccard Index- Wilcoxon Test- Default vs. gbv

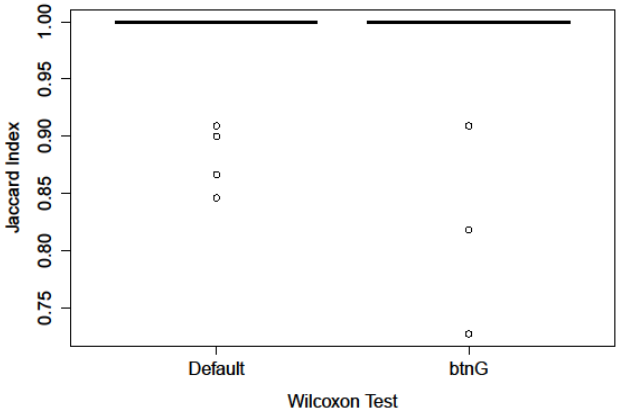


Figure A4.145– 3rd Run-Google-Exclusion- Scenario 2- Jaccard Index- Wilcoxon Test- Default vs. btnG

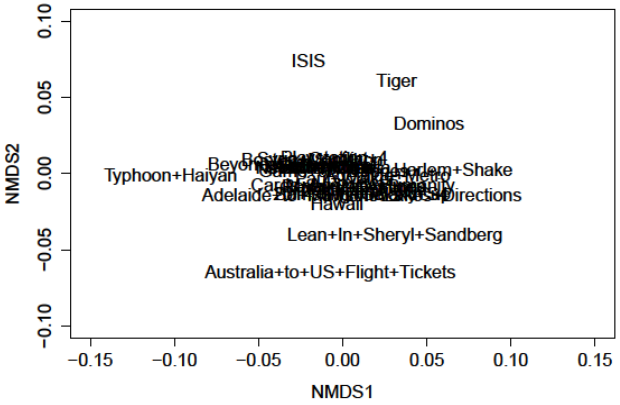


Figure A4.146– 3rd Run-Google-Exclusion- Scenario 2- Jaccard Index- Query Terms

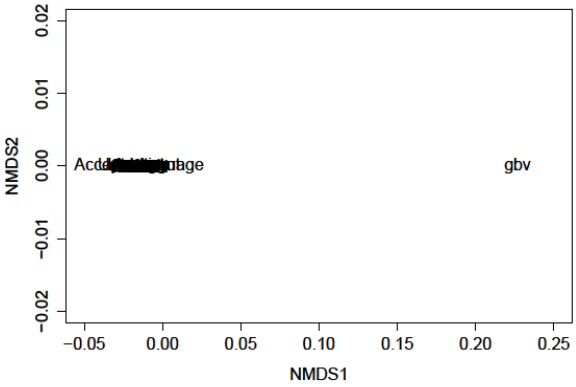


Figure A4.147– 3rd Run-Google-Exclusion- Scenario 2- Jaccard Index- Query Parameters

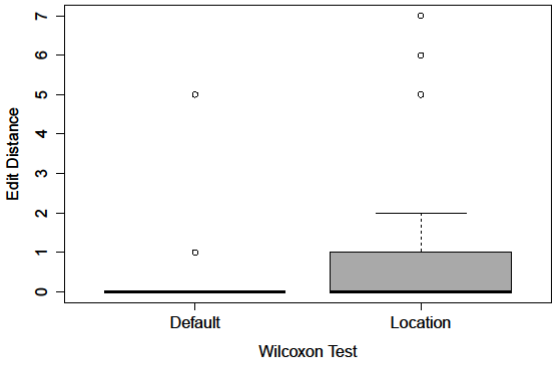


Figure A4.148– 3rd Run-Google-Exclusion- Scenario 2-Edit Distance-Wilcoxon test- Default vs. Location

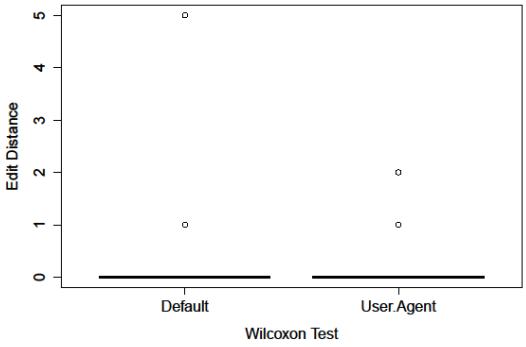


Figure A4.149– 3rd Run-Google-Exclusion- Scenario 2-Edit Distance-Wilcoxon test- Default vs. User-Agent

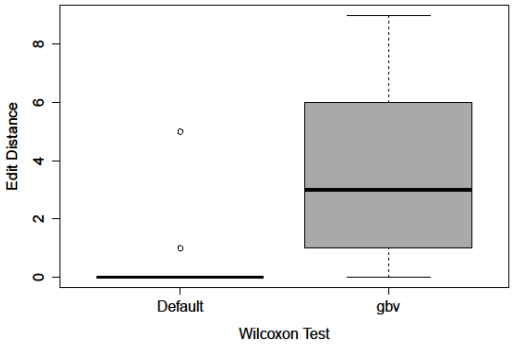


Figure A4.150– 3rd Run-Google-Exclusion- Scenario 2-Edit Distance-Wilcoxon test- Default vs. gbv

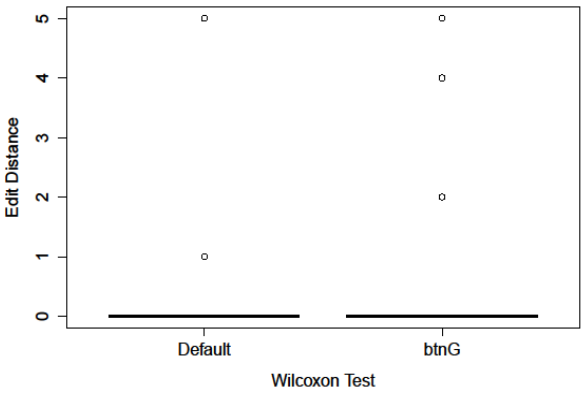


Figure A4.151– 3rd Run-Google-Exclusion- Scenario 2-Edit Distance-Wilcoxon test- Default vs. btnG

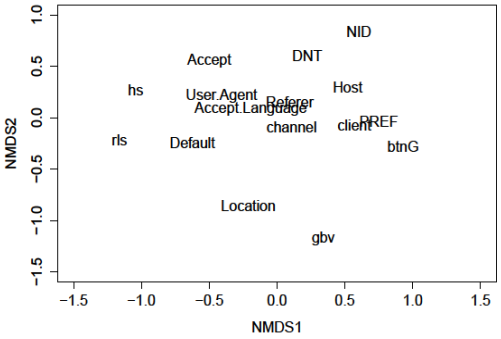


Figure A4.152– 3rd Run-Google-Exclusion- Scenario 2- Edit Distance-Query Parameters

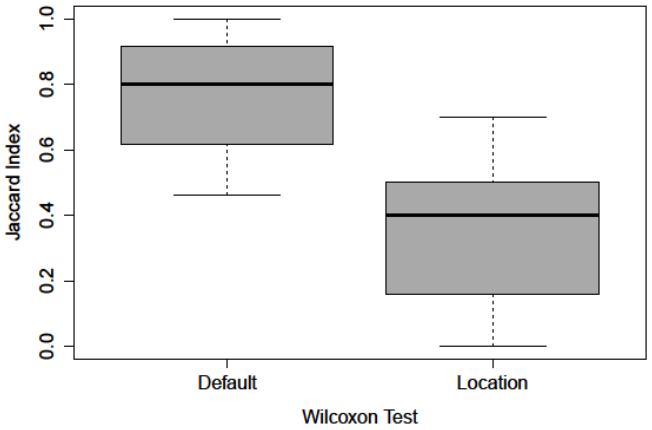


Figure A4.153– 3rd Run-Google-Exclusion- Scenario 3- Jaccard Index-Wilcoxon Test- Default vs. Location

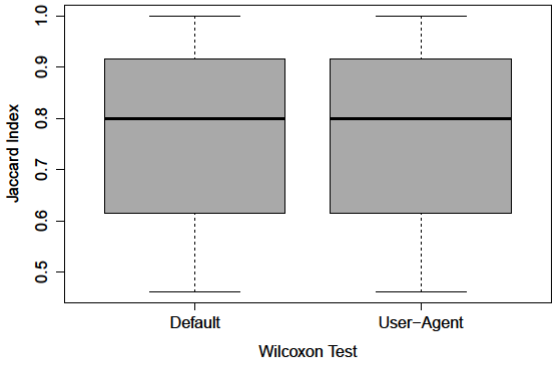


Figure A4.154– 3rd Run-Google-Exclusion- Scenario 3- Jaccard Index-Wilcoxon Test- Default vs. User-Agent

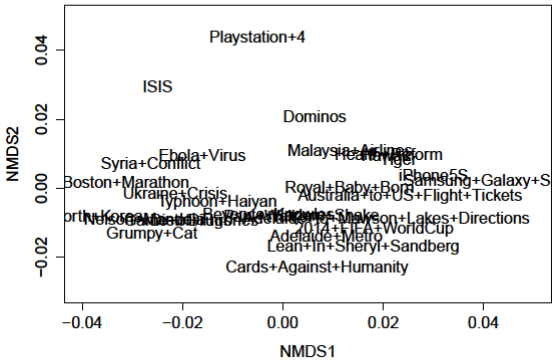


Figure A4.155– 3rd Run-Google-Exclusion- Scenario 3- Jaccard Index-NMDS- Query Terms

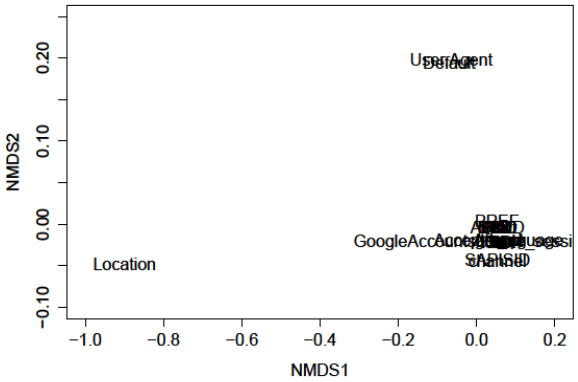


Figure A4.156– 3rd Run-Google-Exclusion- Scenario 3- Jaccard Index-NMDS- Query Parameters

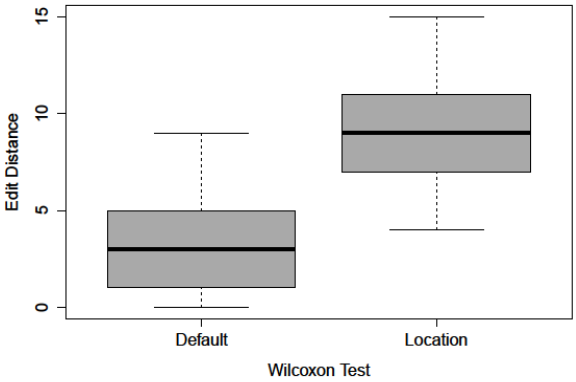


Figure A4.157– 3rd Run-Google-Exclusion- Scenario 3- Edit Distance-Wilcoxon Test- Default vs. Location

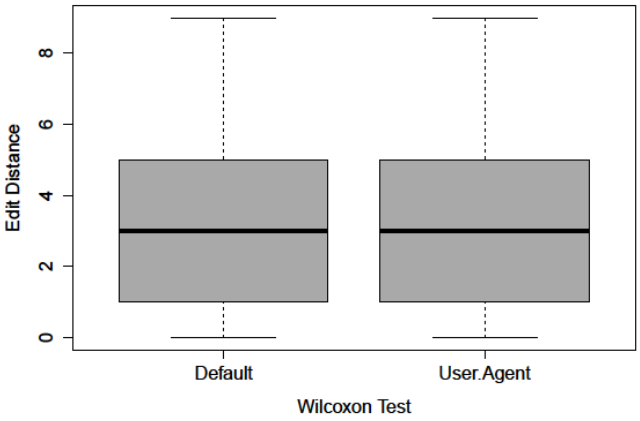


Figure A4.158– 3rd Run-Google-Exclusion- Scenario 3- Edit Distance-Wilcoxon Test- Default vs. User-Agent

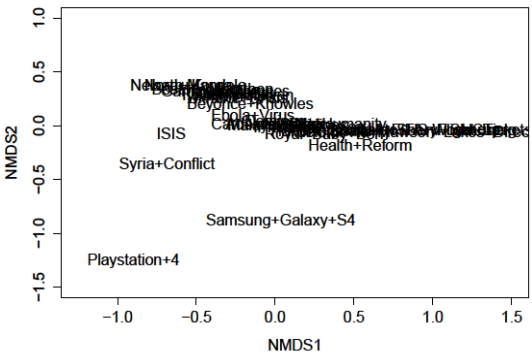


Figure A4.159– 3rd Run-Google-Exclusion- Scenario 3- Edit Distance-NMDS-Query Terms

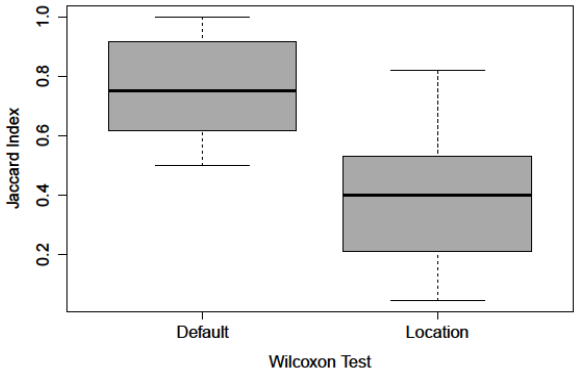


Figure A4.160– 3rd Run-Google-Exclusion- Scenario 4- Jaccard Index-Wilcoxon Test- Default vs. Location

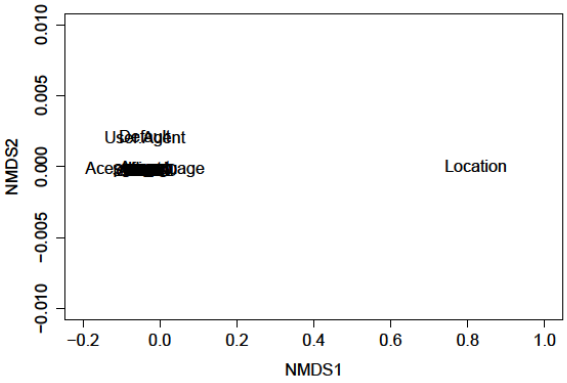


Figure A4.163– 3rd Run-Google-Exclusion- Scenario 4- Jaccard Index-NMDS-Query Parameters

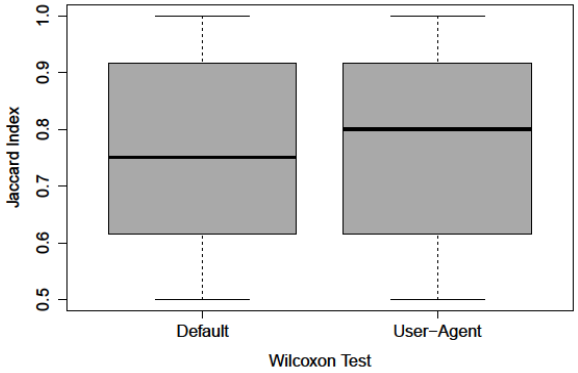


Figure A4.161– 3rd Run-Google-Exclusion- Scenario 4- Jaccard Index-Wilcoxon Test- Default vs. User-Agent

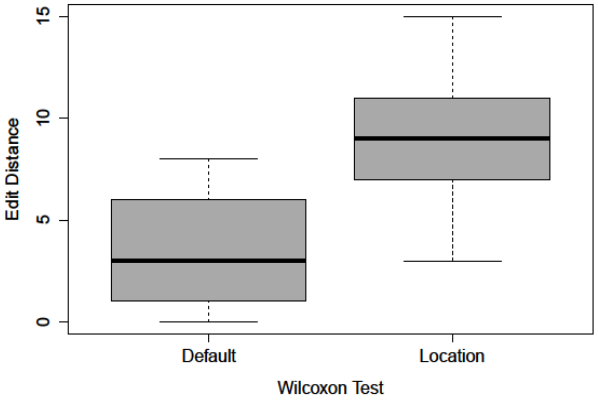


Figure A4.164– 3rd Run-Google-Exclusion- Scenario 4-Edit Distance-Wilcoxon Test-Default vs. Location

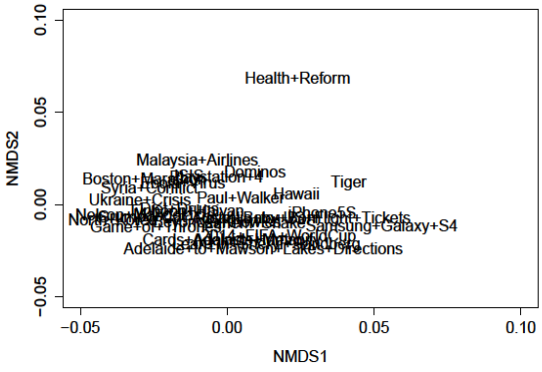


Figure A4.162– 3rd Run-Google-Exclusion- Scenario 4- Jaccard Index-NMDS-Query Terms

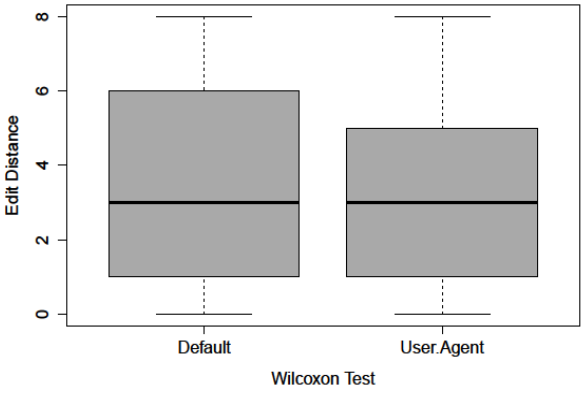


Figure A4.165– 3rd Run-Google-Exclusion- Scenario 4-Edit Distance-Wilcoxon Test-Default vs. User-Agent

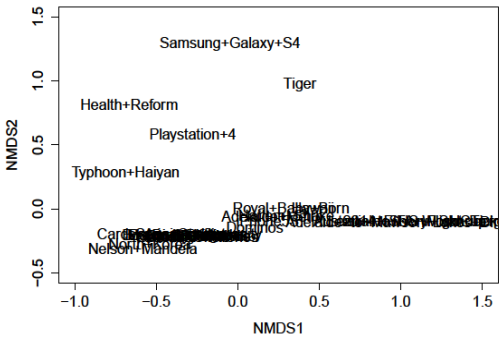


Figure A4.166– 3rd Run-Google- Exclusion- Scenario 4-Edit Distance- NMDS-Query Terms

Google - 2nd Run (May 2015)

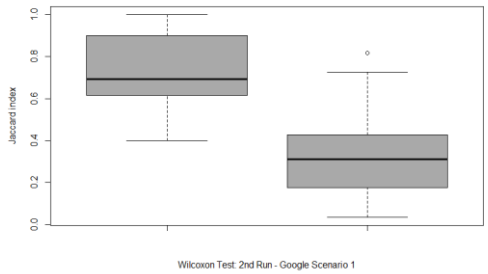


Figure A4.167– 2nd Run-Google- Scenario 1-Jaccard Index- Wilcoxon Test- Default vs. Location

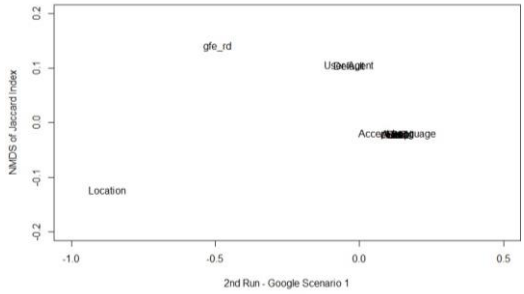


Figure A4.168– 2nd Run-Google- Scenario 1-Jaccard Index- NMDS- Query Parameters

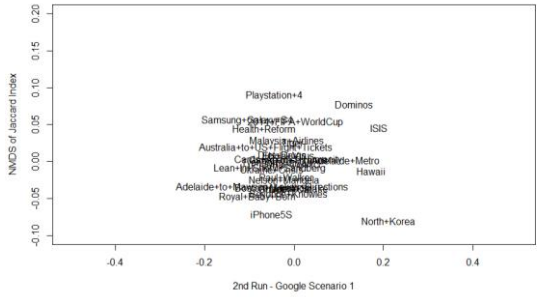


Figure A4.169– 2nd Run-Google- Scenario 1-Jaccard Index- NMDS- Query Terms

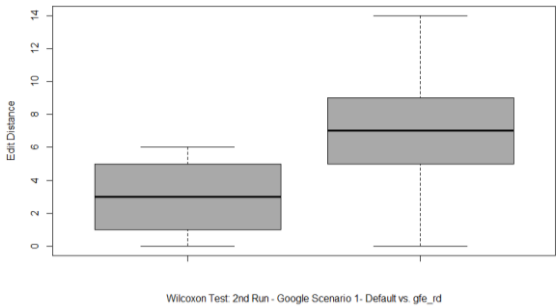


Figure A4.170– 2nd Run-Google- Scenario 1-Edit Distance- Wilcoxon Test- Default vs. gfe_rd

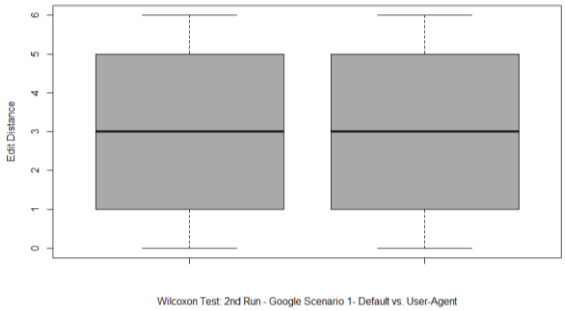


Figure A4.171– 2nd Run-Google- Scenario 1-Edit Distance- Wilcoxon Test- Default vs. User-Agent

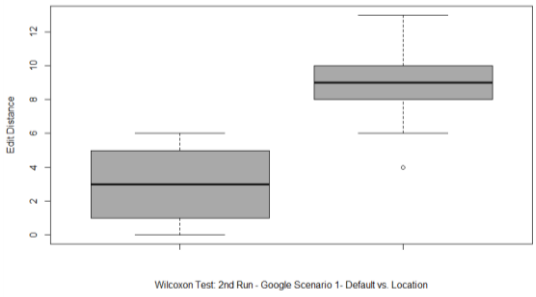


Figure A4.172– 2nd Run-Google- Scenario 1-Edit Distance- Wilcoxon Test- Default vs.Location

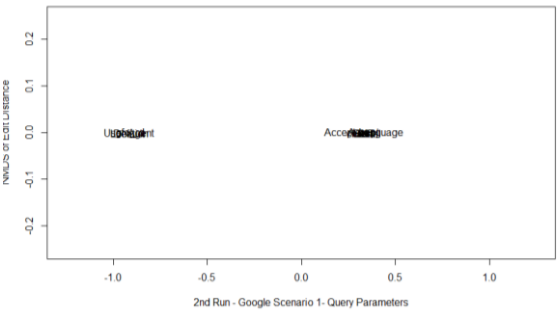


Figure A4.173– 2nd Run-Google- Scenario 1-Edit Distance-NMDS-Query Parameters

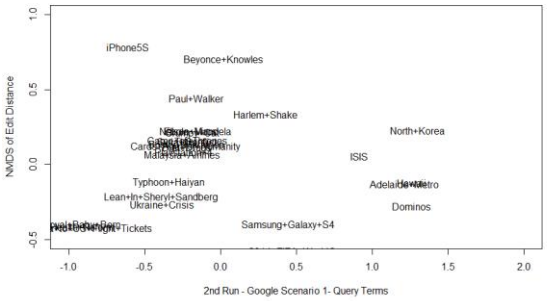


Figure A4.174– 2nd Run-Google- Scenario 1-Edit Distance-NMDS-Query Terms

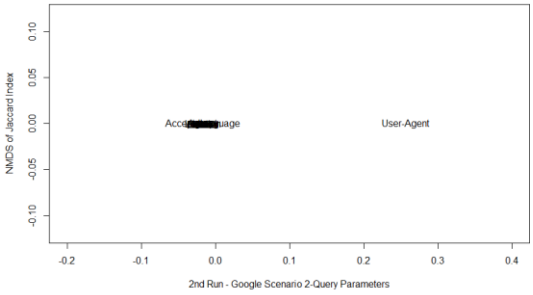


Figure A4.175a– 2nd Run-Google-Scenario 2-Jaccard Index-NMDS-Query Parameters

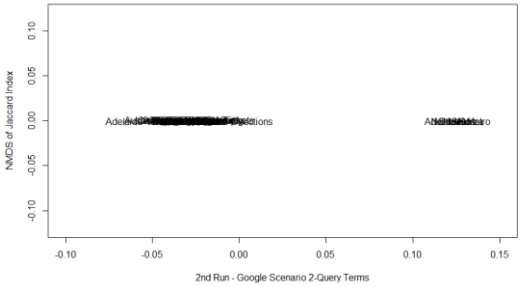


Figure A4.175b– 2nd Run-Google-Scenario 2-Jaccard Index-NMDS-Query Terms

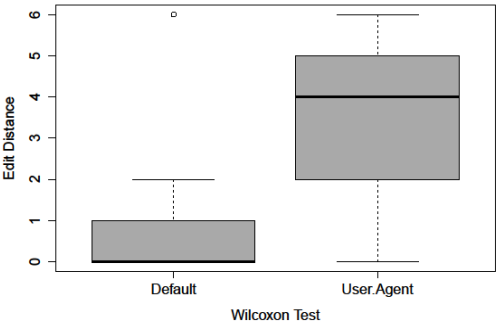


Figure A4.176a– 2nd Run-Google-Scenario 2-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

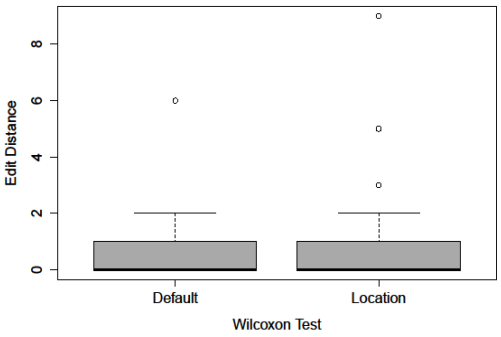


Figure A4.176b– 2nd Run-Google-Scenario 2-Jaccard Index-NMDS-Query Terms-Wilcoxon Test- Default vs. Location

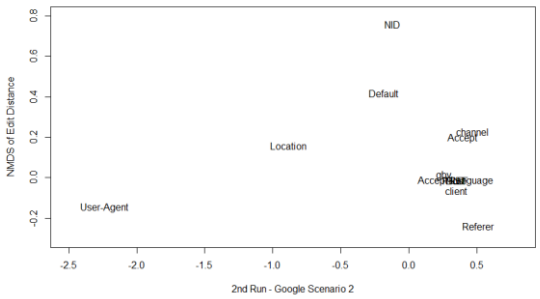


Figure A4.177– 2nd Run-Google- Scenario 2-Edit Distance-NMDS-Query Parameters

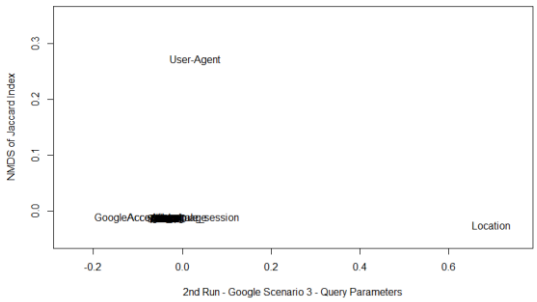


Figure A4.178– 2nd Run-Google- Scenario 3-Jaccard Index-NMDS-Query Parameters

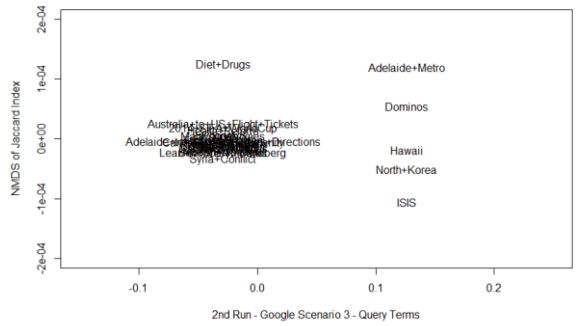


Figure A4.179– 2nd Run-Google- Scenario 3-Jaccard Index-NMDS-Query Terms

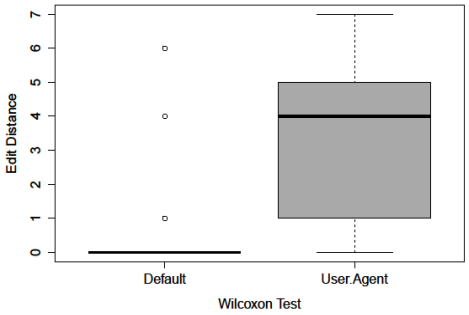


Figure A4.180a– 2nd Run-Google-Scenario 3-Edit Distance-Wilcoxon Test-Default vs. User-Agent

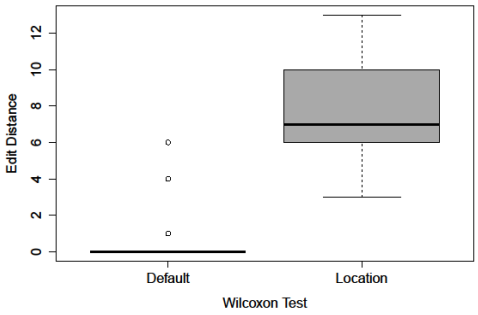


Figure A4.180b– 2nd Run-Google-Scenario 3-Edit Distance-Wilcoxon Test-Default vs. Location

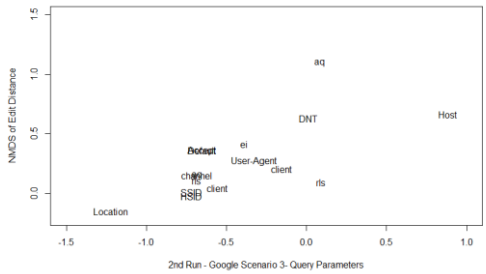


Figure A4.181a– 2nd Run-Google-Scenario 3-Edit Distance-NMDS-Query Parameters

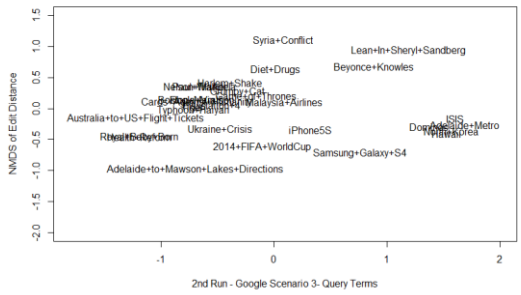


Figure A4.181b– 2nd Run-Google-Scenario 3-Edit Distance-NMDS-Query Terms

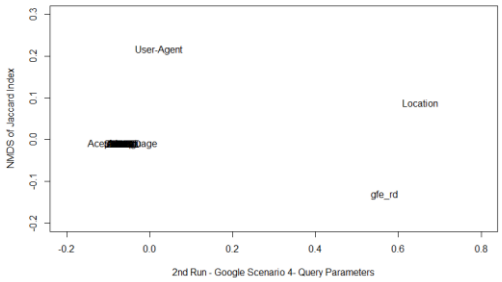


Figure A4.182a– 2nd Run-Google-Scenario 4-Jaccard Index-NMDS-Query Parameters

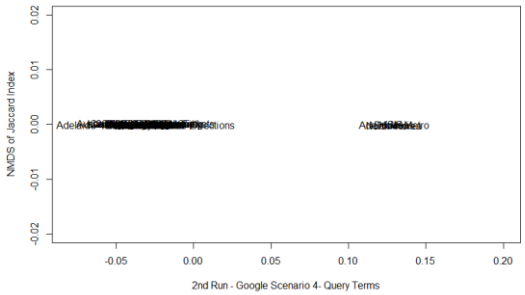


Figure A4.182b– 2nd Run-Google-Scenario 4-Jaccard Index-NMDS-Query Terms

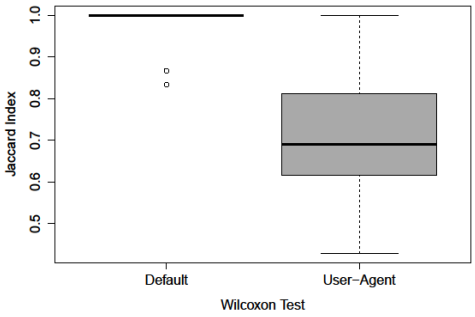


Figure A4.183a– 2nd Run-Google-Scenario 4-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

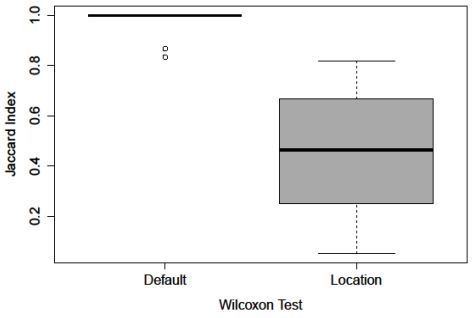


Figure A4.183b– 2nd Run-Google-Scenario 4-Jaccard Index-Wilcoxon Test-Default vs. Location

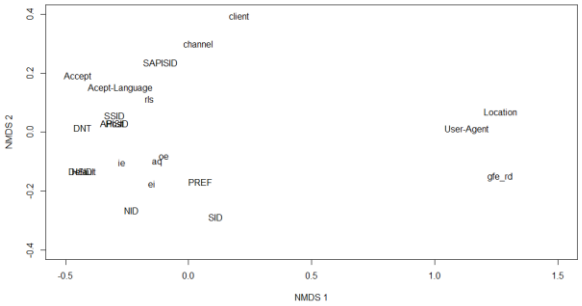


Figure A4.184– 2nd Run-Google- Scenario 4-Edit Distance-NMDS-Query Parameters

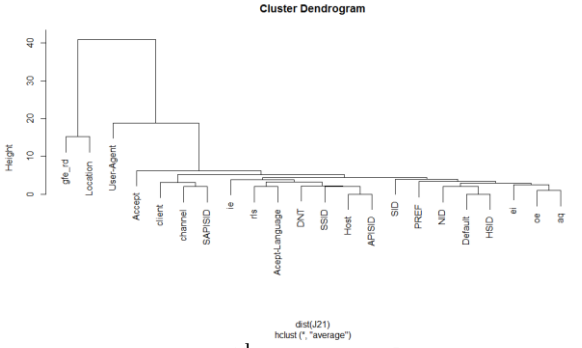


Figure A4.185– 2nd Run-Google- Scenario 4-Edit Distance-NMDS-Query Parameters – Cluster Dendrogram using average

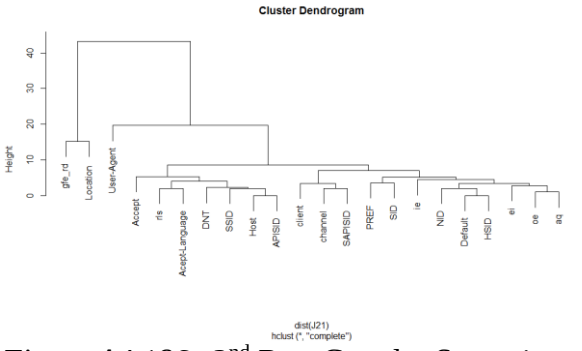


Figure A4.186– 2nd Run-Google- Scenario 4-Edit Distance-NMDS-Query Parameters – Cluster Dendrogram using complete

Google - 1st Run

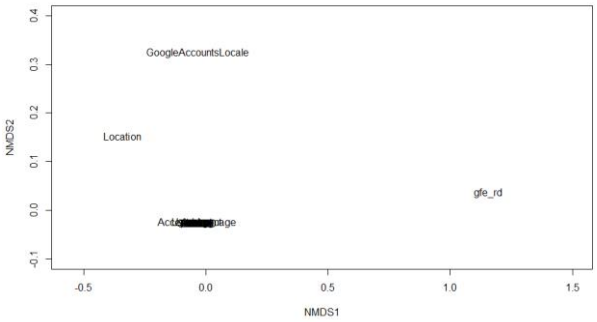


Figure A4.187– 1st Run-Google-Jaccard Index-Query Parameters

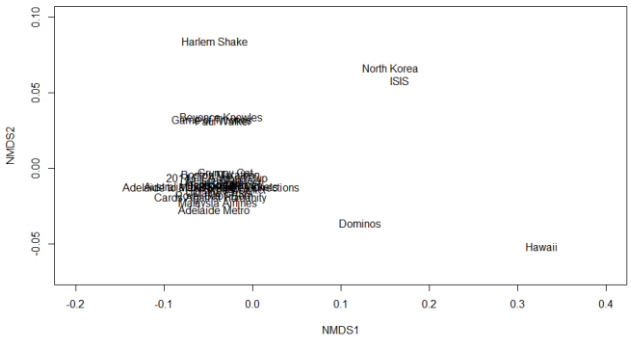


Figure A4.188– 1st Run-Google-Jaccard Index-Query Terms

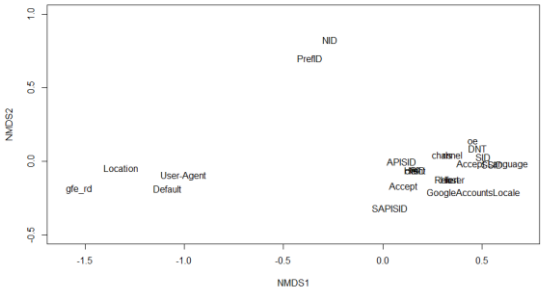


Figure A4.189– 1st Run-Google-Edit Distance-Query Parameters

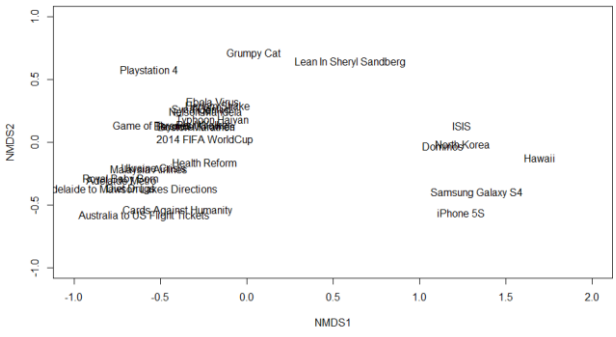


Figure A4.190– 1st Run-Google-Edit Distance-Query Terms

Google - User-Agent Experiment

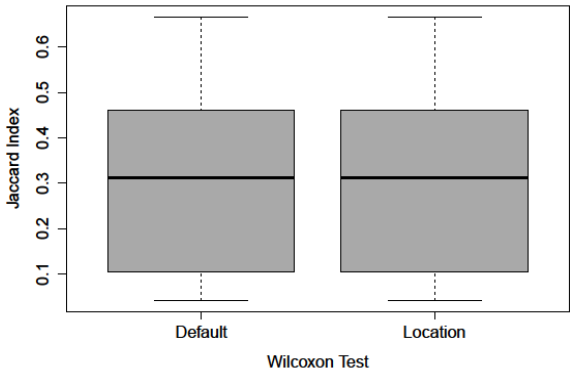


Figure A4.191– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Jaccard Index-Wilcoxon Test-Default vs. Location

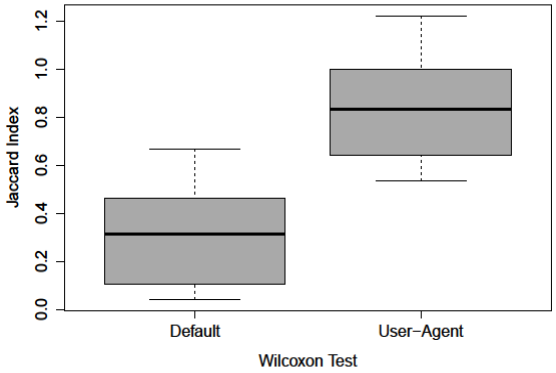


Figure A4.192– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

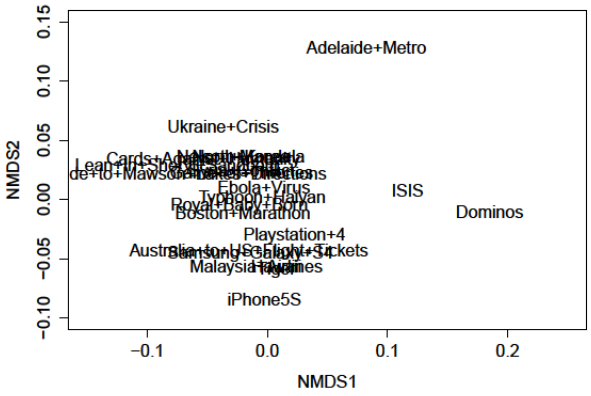


Figure A4.193– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Jaccard Index-NMDS-Query Terms

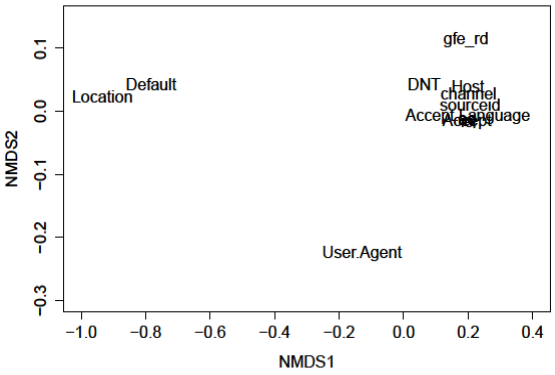


Figure A4.194– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Jaccard Index-NMDS-Query Parameters

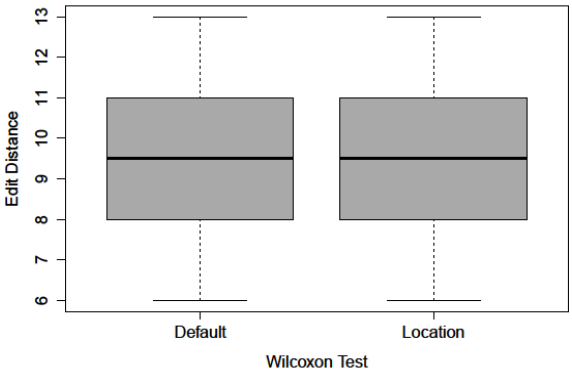


Figure A4.195– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Edit Distance-Wilcoxon Test-Default vs. Location

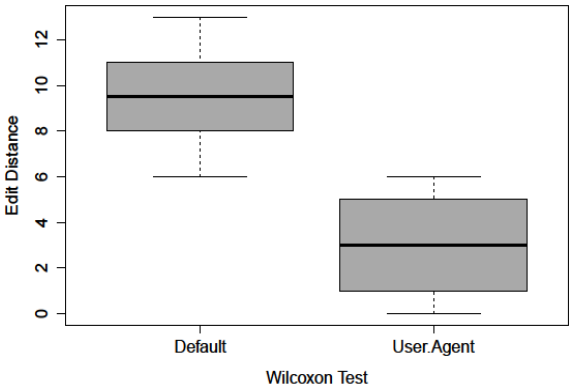


Figure A4.196– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Edit Distance-Wilcoxon Test-Default vs. User-Agent

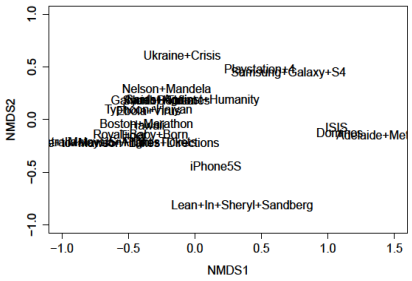


Figure A4.197– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Edit Distance-NMDS-Query Terms

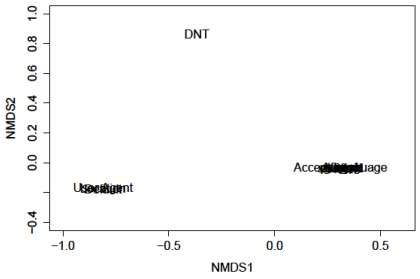


Figure A4.198– 3rd Run-Google-User-Agent Experiment-Google Chrome-Scenario1-Edit Distance-NMDS-Query Parameters

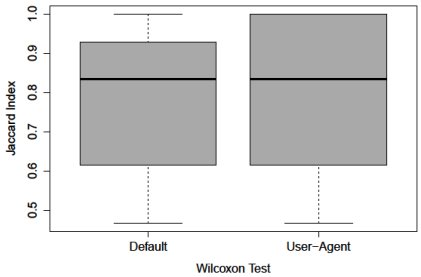


Figure A4.199– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

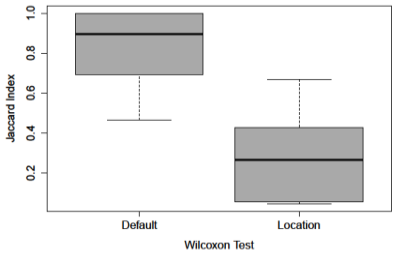


Figure A4.200– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. Location

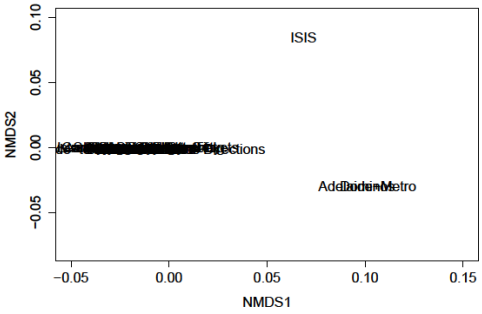
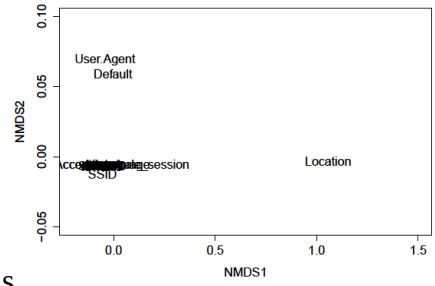


Figure A4.201– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Jaccard Index-NMDS-Query



Terms

Figure A4.202– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Jaccard Index-MMDS-Query Parameters

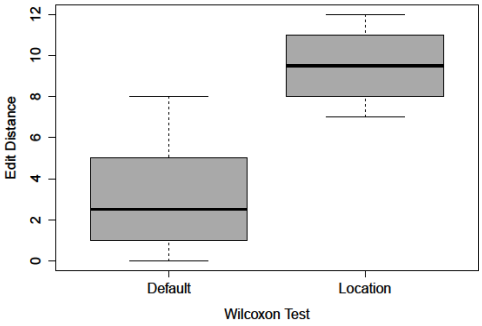


Figure A4.203– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Edit Distance-Wilcoxon Test-Default vs. Location

Figures A4.207 – A4.261 Bing

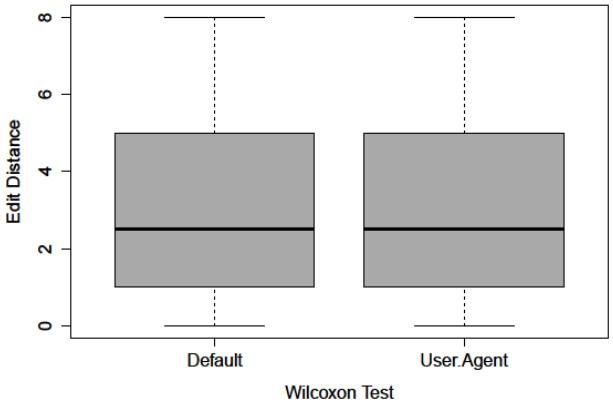


Figure A4.204– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3- Edit Distance-Wilcoxon Test-Default vs. User-Agent

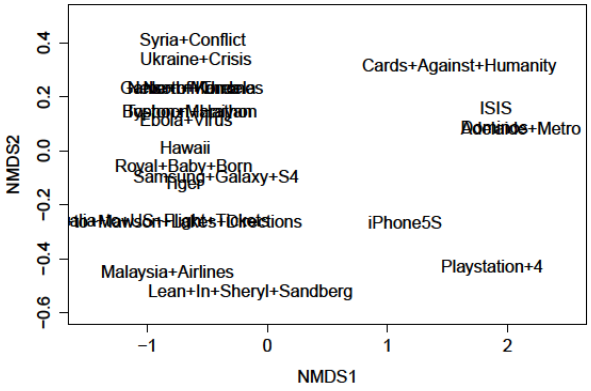


Figure A4.205– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Edit Distance-MMDS-Query Terms

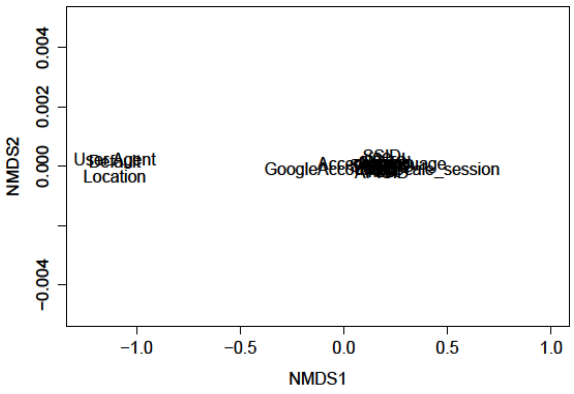


Figure A4.206– 3rd Run-Google-User-Agent Experiment-Internet Explorer-Scenario3-Jaccard Index-MMDS-Query Parameters

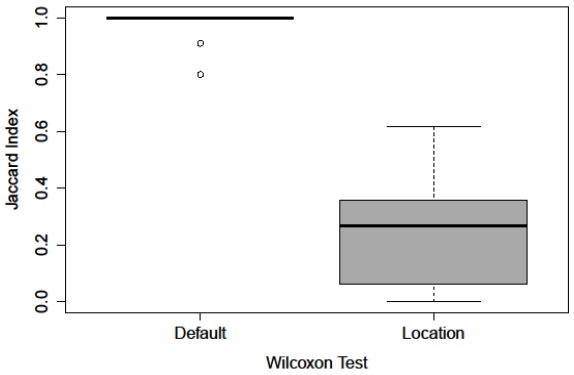


Figure A4.207– 2nd Run-Bing-Scenario1-Jaccard Index-Wilcoxon Test- Default vs. Location



Figure A4.208– 2nd Run-Bing-Scenario1-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

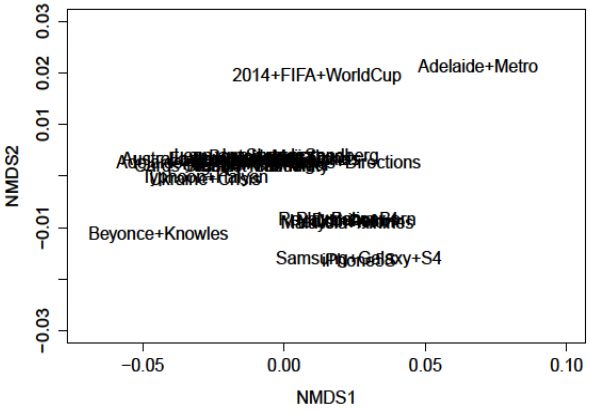


Figure A4.209– 2nd Run-Bing-Scenario1-Jaccard Index-NMDS-Query Terms

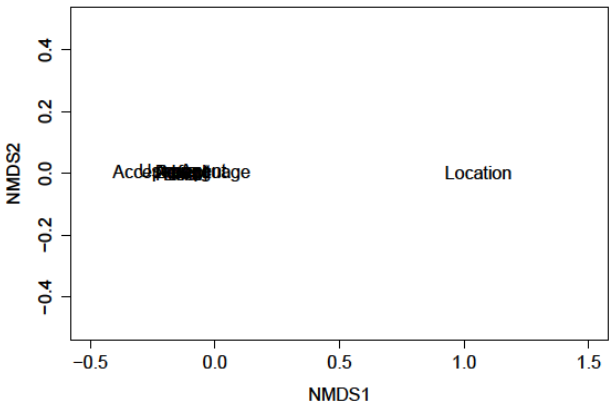


Figure A4.210– 2nd Run-Bing-Scenario1-Jaccard Index-NMDS-Query Parameters

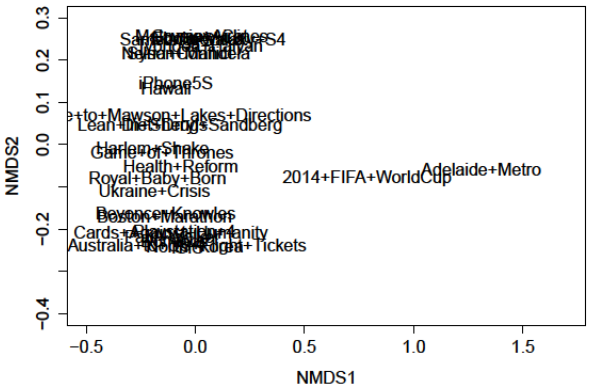


Figure A4.213– 2nd Run-Bing-Scenario1-Edit Distance-NMDS-Query Terms

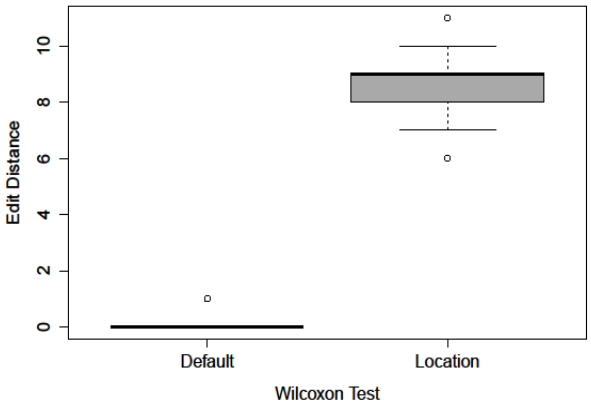


Figure A4.211– 2nd Run-Bing-Scenario1-Edit Distance-Wilcoxon Test- Default vs. Location

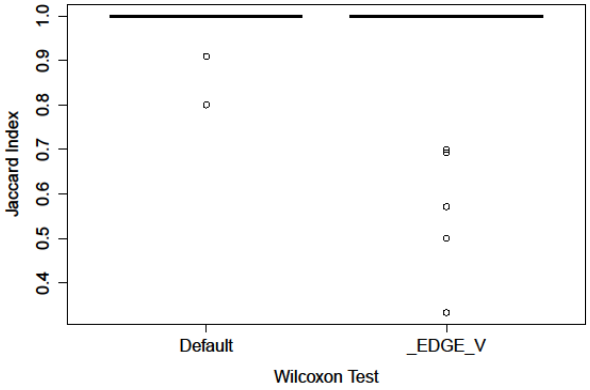


Figure A4.214– 2nd Run-Bing-Scenario2-Jaccard Index-Wilcoxon Test- Default vs. _EDGE_V

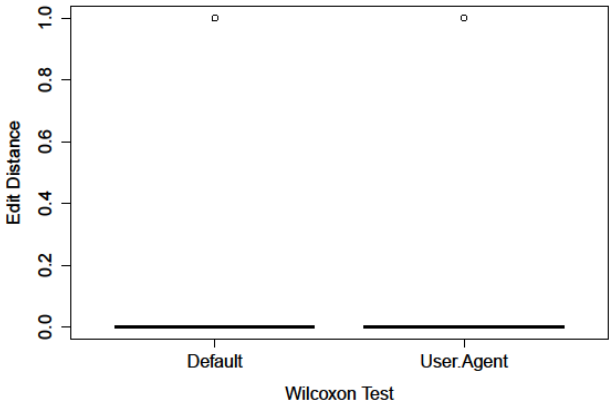


Figure A4.212– 2nd Run-Bing-Scenario1-Edit Distance-Wilcoxon Test- Default vs. User-Agent

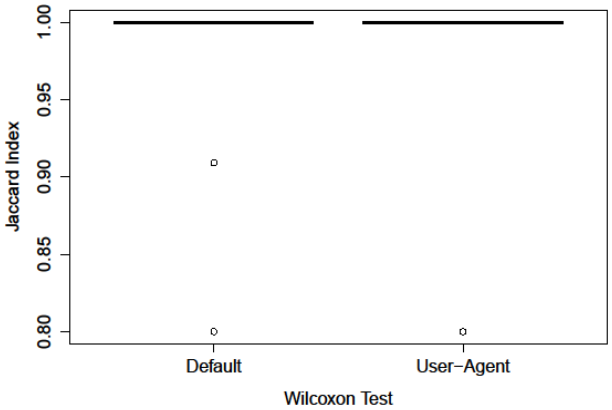


Figure A4.215– 2nd Run-Bing-Scenario2-Jaccard Index-Wilcoxon Test- Default vs. User-Agent

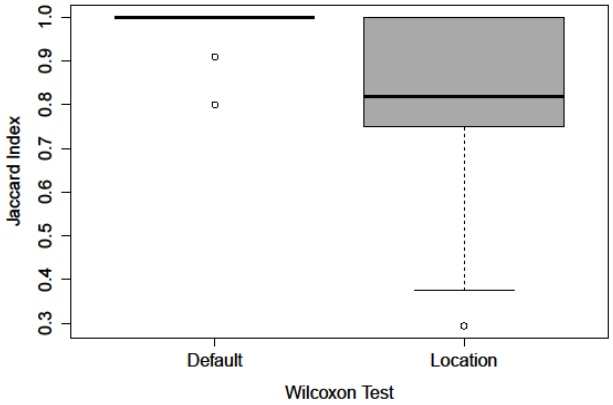


Figure A4.216– 2nd Run-Bing-Scenario2-Jaccard Index-Wilcoxon Test- Default vs. Location

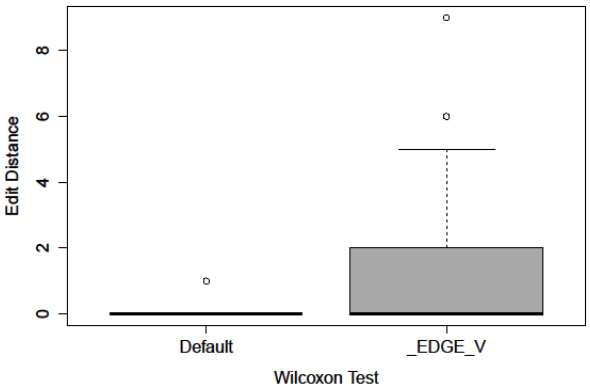


Figure A4.219– 2nd Run-Bing-Scenario2-Edit Distance- Wilcoxon Test- Default vs. _EDGE_V

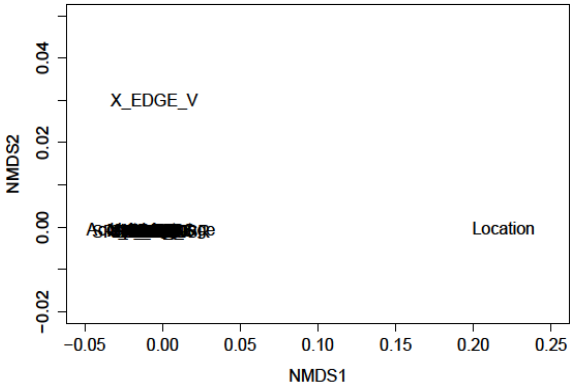


Figure A4.217– 2nd Run-Bing-Scenario2-Jaccard Index-NMDS-Query Parameters

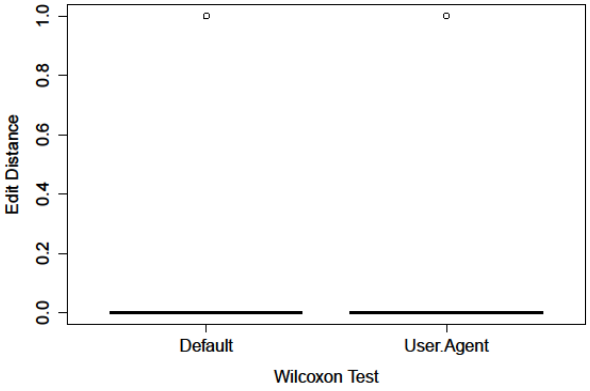


Figure A4.220– 2nd Run-Bing-Scenario2-Edit Distance- Wilcoxon Test- Default vs. User-Agent

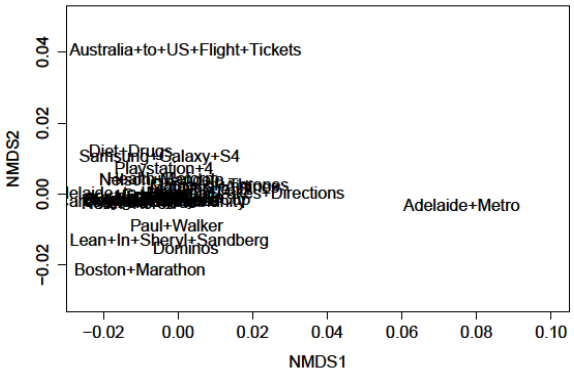


Figure A4.218– 2nd Run-Bing-Scenario2-Jaccard Index-NMDS-Query Terms

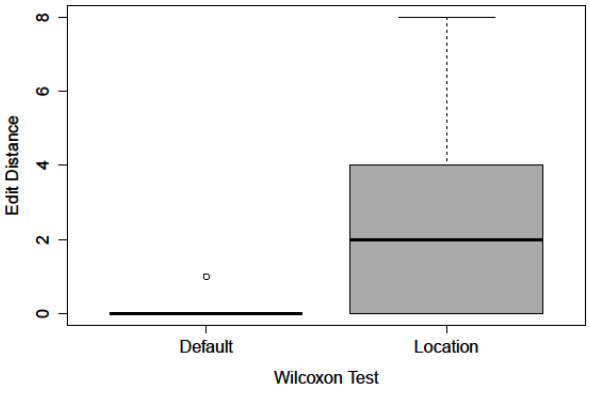


Figure A4.221– 2nd Run-Bing-Scenario2-Edit Distance- Wilcoxon Test- Default vs. Location

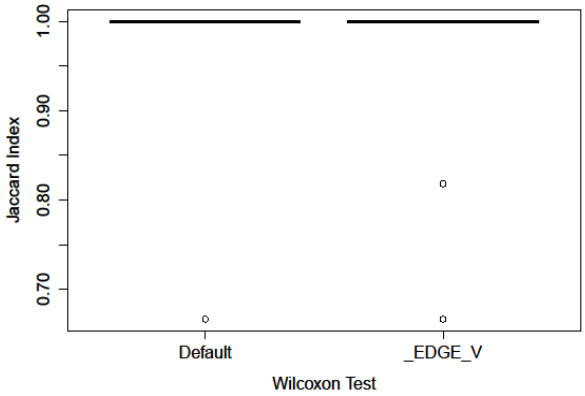


Figure A4.222– 2nd Run-Bing-Scenario3- Jaccard Index- Wilcoxon Test- Default vs. _EDGE_V

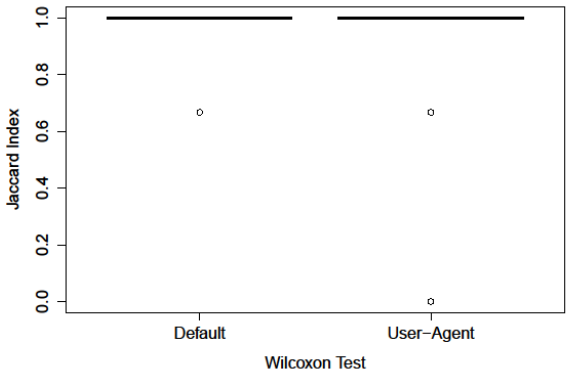


Figure A4.223– 2nd Run-Bing-Scenario3- Jaccard Index- Wilcoxon Test- Default vs. User-Agent

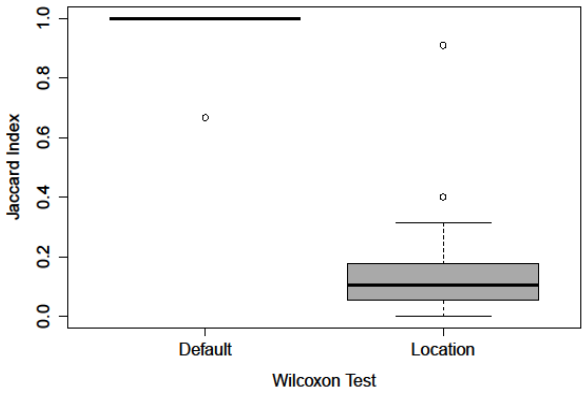


Figure A4.224– 2nd Run-Bing-Scenario3- Jaccard Index- Wilcoxon Test- Default vs. Location

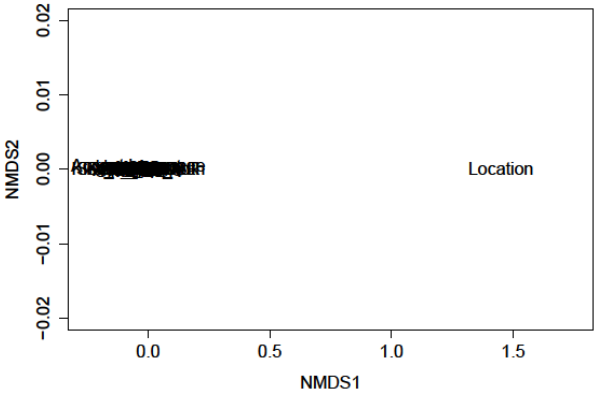


Figure A4.225– 2nd Run-Bing-Scenario3- Jaccard Index- NMDS-Query Parameters

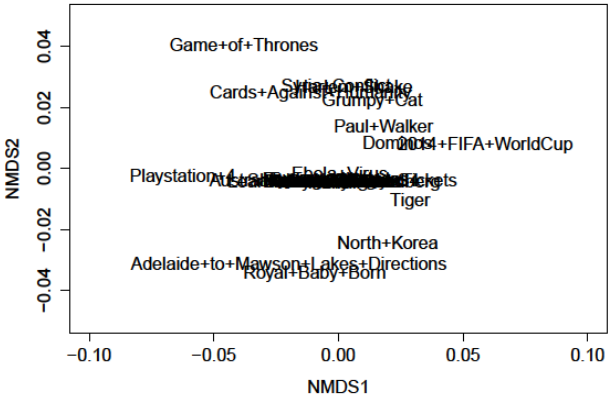


Figure A4.226– 2nd Run-Bing-Scenario3- Jaccard Index- NMDS-Query Terms

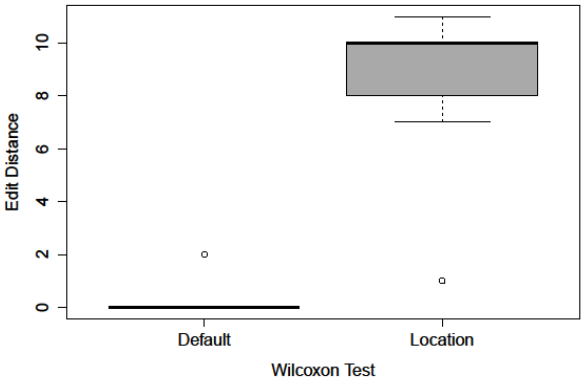
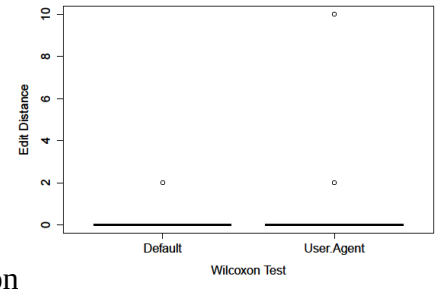


Figure A4.227– 2nd Run-Bing-Scenario3— Edit Distance- Wilcoxon Test- Default vs.



Location

Figure A4.228– 2nd Run-Bing-Scenario3—Edit Distance- Wilcoxon Test- Default vs. User-Agent

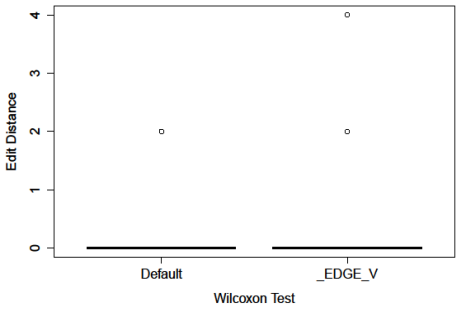


Figure A4.229– 2nd Run-Bing-Scenario3—Edit Distance- Wilcoxon Test- Default vs. _EDGE_V

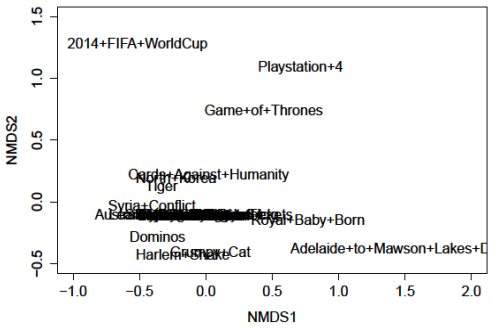


Figure A4.230– 2nd Run-Bing-Scenario3—Edit Distance-NMDS-Query Terms

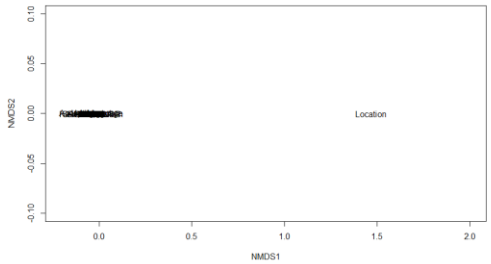


Figure A4.231– 2nd Run-Bing-Scenario4—Jaccard Index-NMDS-Query Parameters

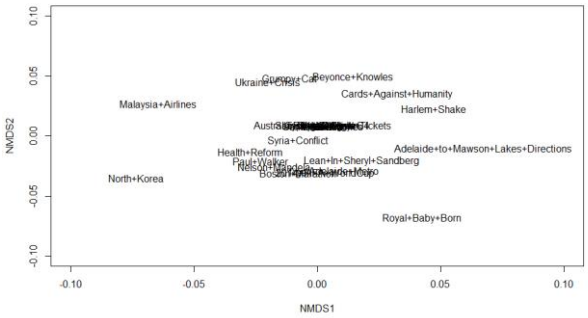


Figure A4.232– 2nd Run-Bing-Scenario4—Jaccard Index-NMDS-Query Terms

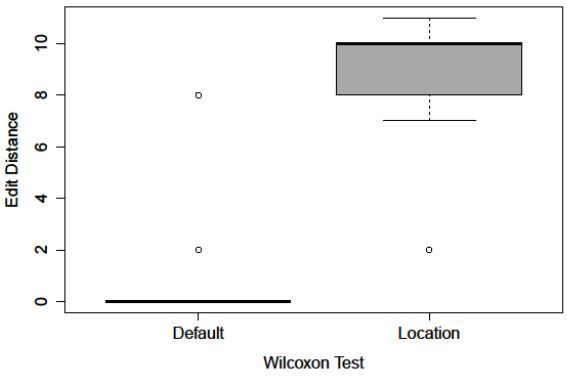


Figure A4.233– 2nd Run-Bing-Scenario4—Edit Distance-Wilcoxon Test- Default vs. Location

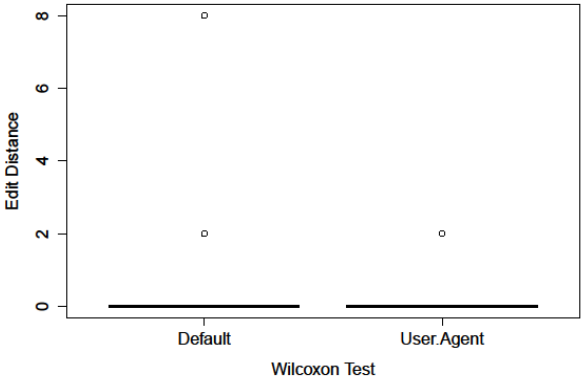


Figure A4.234– 2nd Run-Bing-Scenario4—Edit Distance-Wilcoxon Test- Default vs. User-Agent

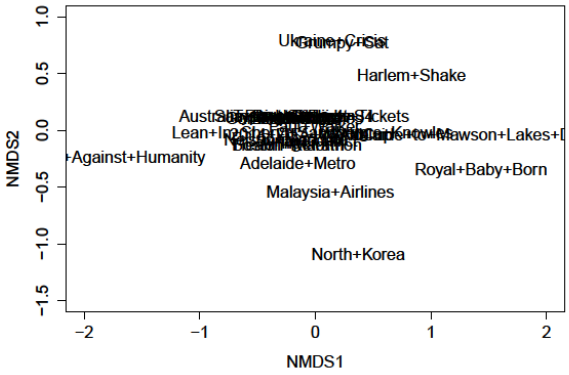


Figure A4.235– 2nd Run-Bing-Scenario4—Edit Distance-NMDS-Query Terms
3rd Run – January 2016

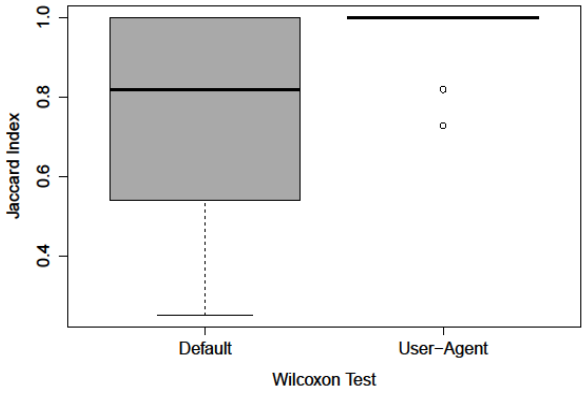


Figure A4.236– 3rd Run-Bing-Scenario1—Jaccard Index-Wilcoxon Test-Default vs. User-Agent

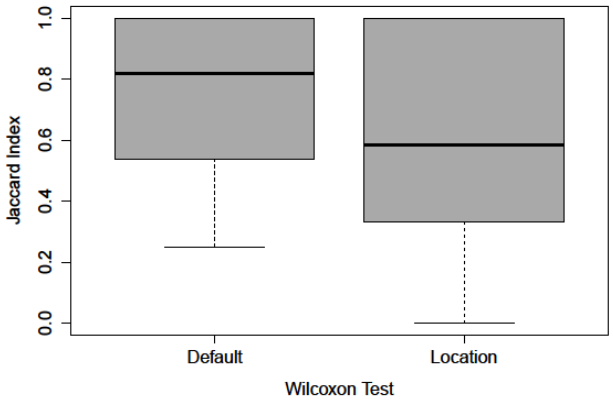


Figure A4.237– 3rd Run-Bing-Scenario1—Jaccard Index-Wilcoxon Test-Default vs. Location

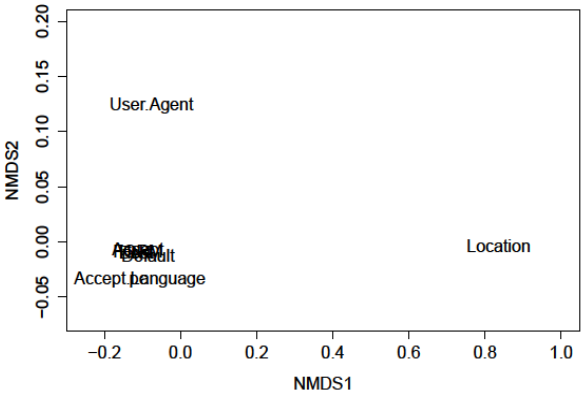


Figure A4.238– 3rd Run-Bing-Scenario1—Jaccard Index-NMDS-Query Parameters

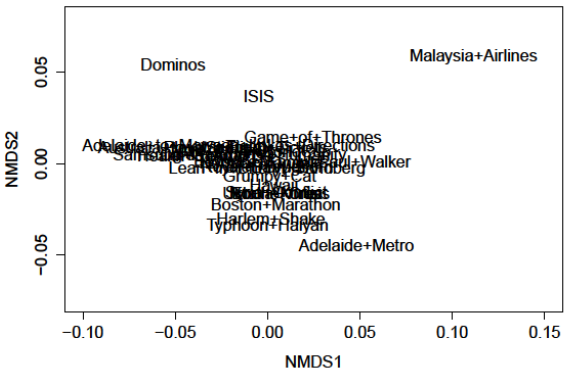


Figure A4.239– 3rd Run-Bing-Scenario1—Jaccard Index-NMDS-Query Terms

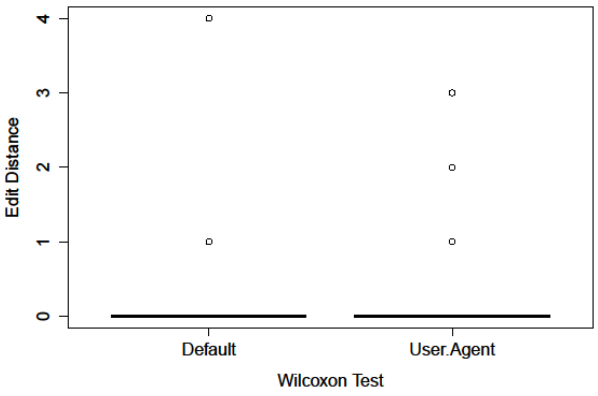


Figure A4.240– 3rd Run-Bing-Scenario1—Edit Distance-Wilcoxon Test-Default vs. User-Agent

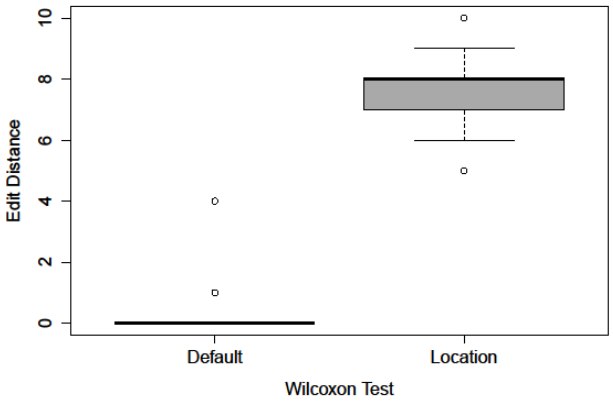


Figure A4.241– 3rd Run-Bing-Scenario1—Edit Distance- Wilcoxon Test-Default vs. Location

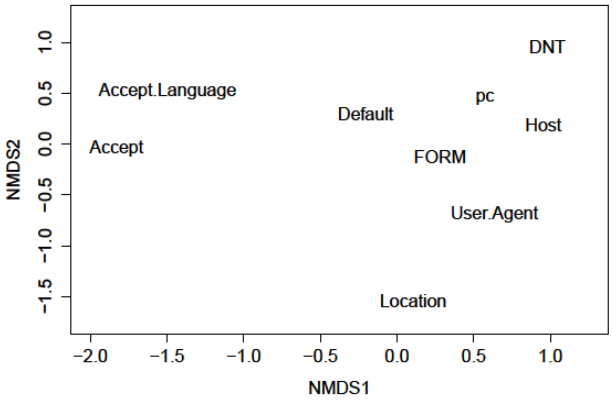


Figure A4.242– 3rd Run-Bing-Scenario1—Edit Distance-NMDS-Query Parameters

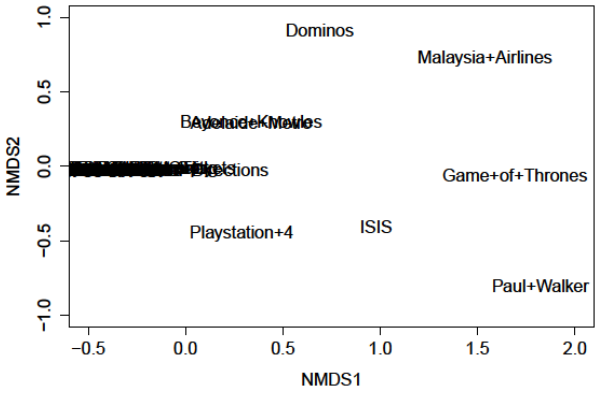


Figure A4.243– 3rd Run-Bing-Scenario1—Edit Distance-NMDS-Query Terms

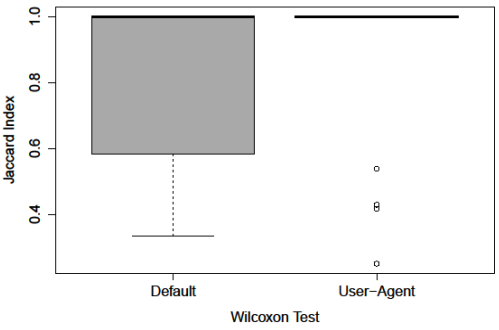


Figure A4.244– 3rd Run-Bing-Scenario3—Jaccard Index-Wilcoxon Test-Default vs. User-Agent

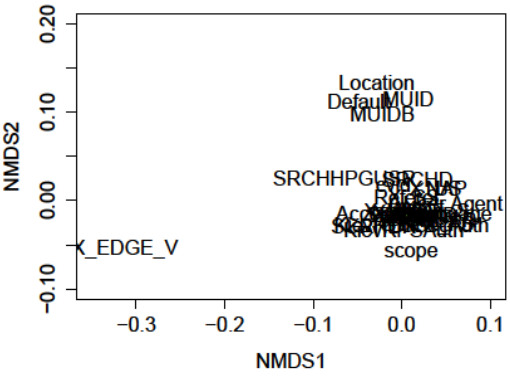


Figure A4.245a– 3rd Run-Bing-Scenario3—Jaccard Index-NMDS-Query Parameters

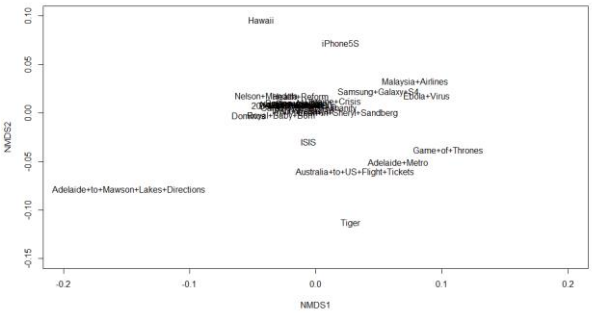


Figure A4.245b– 3rd Run-Bing-Scenario3—Jaccard Index-NMDS-Query Terms

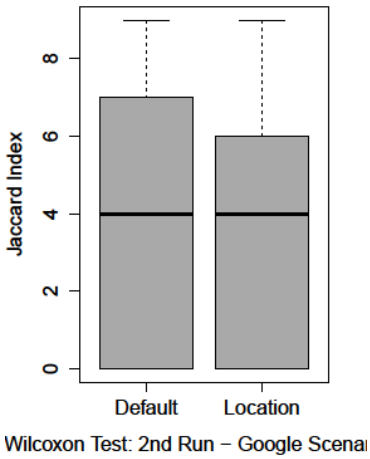


Figure A4.246– 3rd Run-Bing-Scenario3-Edit Distance-Wilcoxon Test-Default vs. Location

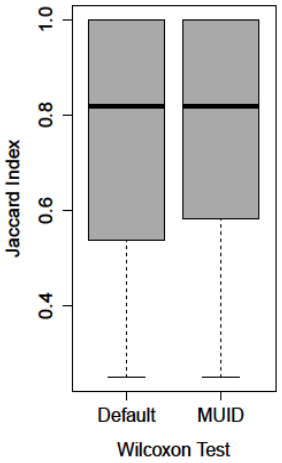


Figure A4.249– 3rd Run-Bing-Scenario4-Jaccard Index -Wilcoxon Test-Default vs. MUID

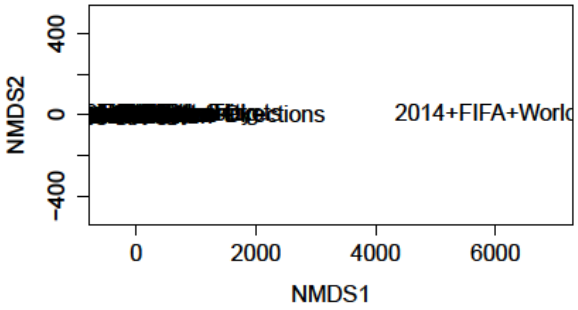


Figure A4.247– 3rd Run-Bing-Scenario3-Edit Distance-NMDS-Query Terms

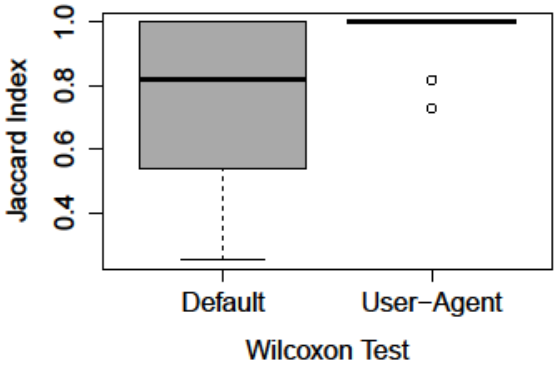


Figure A4.250– 3rd Run-Bing-Scenario4-Jaccard Index -Wilcoxon Test-Default vs. User-Agent

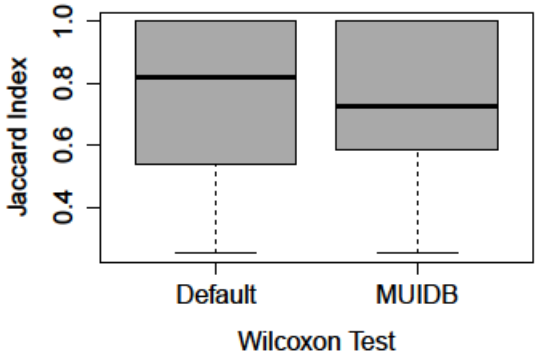


Figure A4.248– 3rd Run-Bing-Scenario4-Jaccard Index-Wilcoxon Test-Default vs. MUIDB

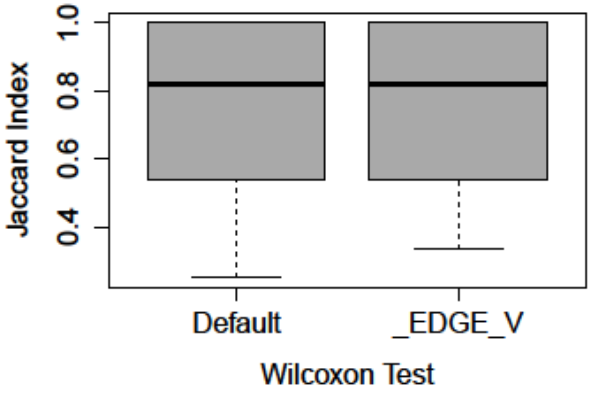


Figure A4.251– 3rd Run-Bing-Scenario4-Jaccard Index -Wilcoxon Test-Default vs. _Edge_V

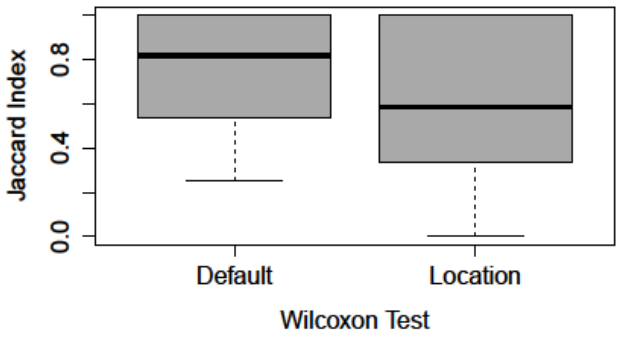


Figure A4.252– 3rd Run-Bing-Scenario4-Jaccard Index -Wilcoxon Test-Default vs. Location

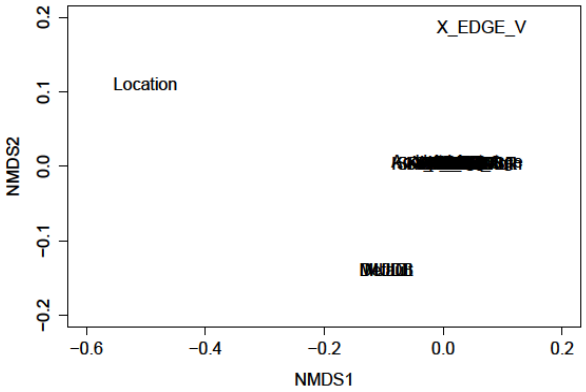


Figure A4.253– 3rd Run-Bing-Scenario4-Jaccard Index –NMDS-Query Parameters

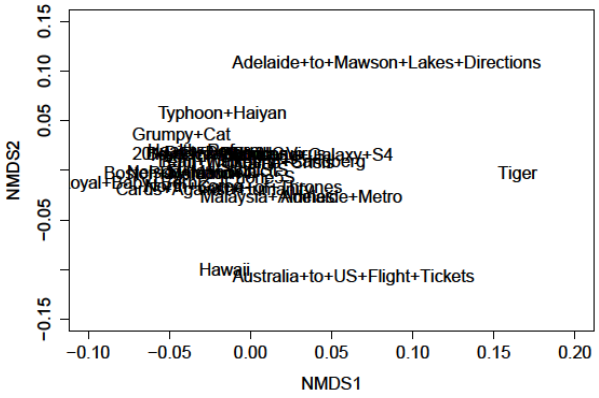


Figure A4.254– 3rd Run-Bing-Scenario4-Jaccard Index –NMDS-Query Terms

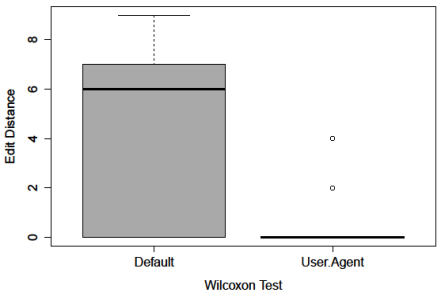


Figure A4.255– 3rd Run-Bing-Scenario4-Edit Distance -Wilcoxon Test-Default vs. User-Agent

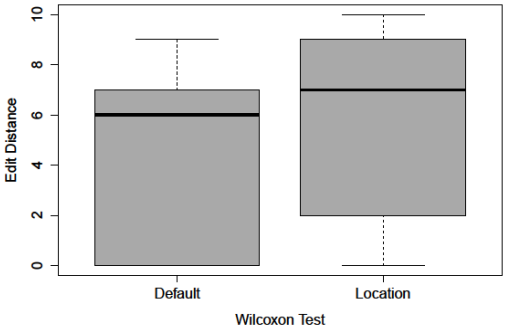


Figure A4.256– 3rd Run-Bing-Scenario4-Edit Distance –Wilcoxon Test-Default vs. Location

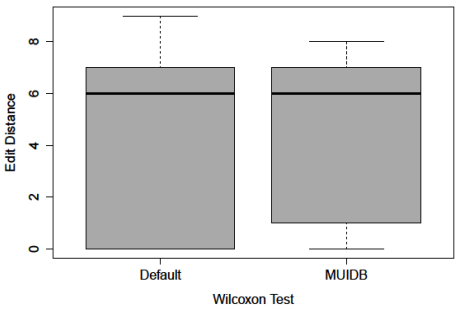


Figure A4.257– 3rd Run-Bing-Scenario4-Edit Distance –Wilcoxon Test-Default vs. MUIDB

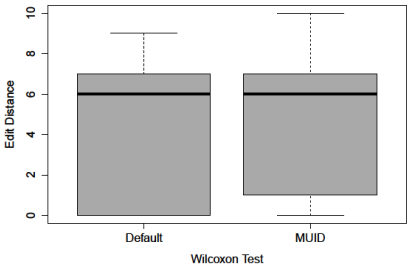


Figure A4.258– 3rd Run-Bing-Scenario4-Edit Distance –Wilcoxon Test-Default vs. MUID

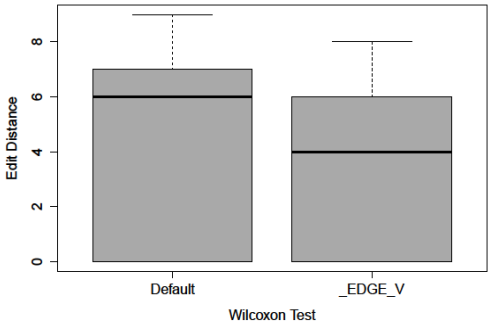


Figure A4.259– 3rd Run-Bing-Scenario4-Edit Distance –Wilcoxon Test-Default vs. _EDGE_V

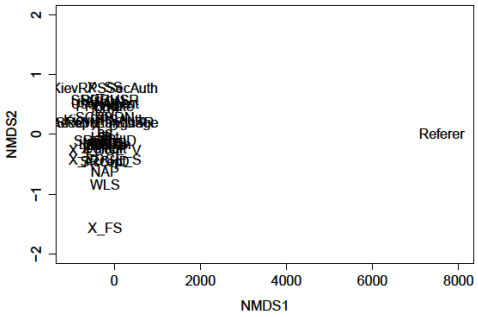


Figure A4.260– 3rd Run-Bing-Scenario4-Edit Distance –Wilcoxon Test-Default vs. Query Parameters

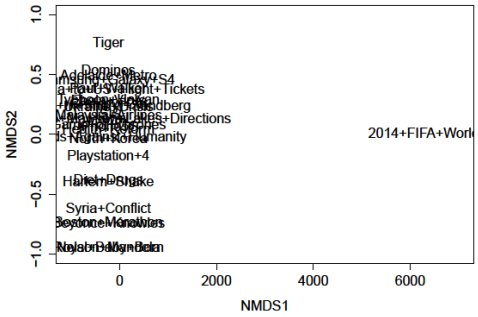


Figure A4.261– 3rd Run-Bing-Scenario4-Edit Distance –Wilcoxon Test-Default vs. Query Terms

Figures A4.262 – A4.325 Yahoo

2nd Run – May 2015

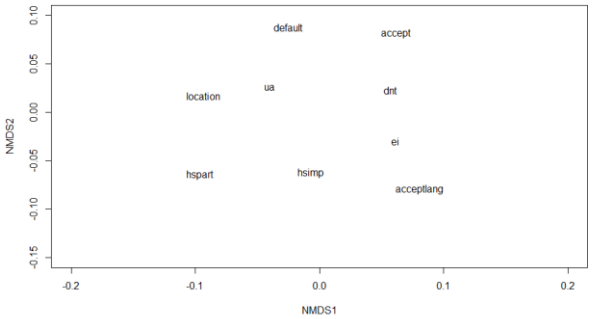


Figure A4.262– 2nd Run-Yahoo-Scenario1-Jaccard Index-NMDS-Query Parameters

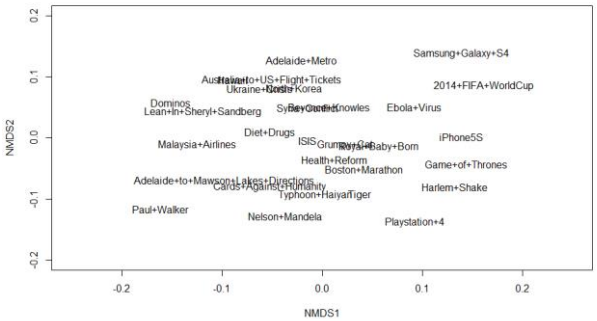


Figure A4.263– 2nd Run-Yahoo-Scenario1-Jaccard Index-NMDS-Query Terms

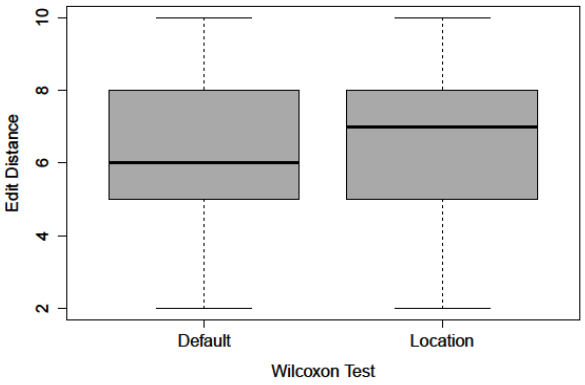


Figure A4.264– 2nd Run-Yahoo-Scenario1-Edit Distance-Wilcoxon Test-Default vs. Location

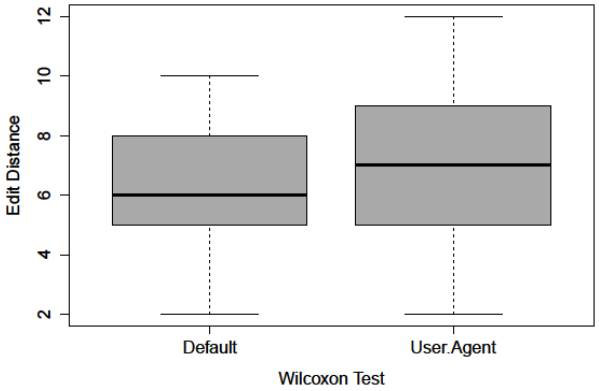


Figure A4.265– 2nd Run-Yahoo-Scenario1-Edit Distance-Wilcoxon Test-Default vs. User-Agent

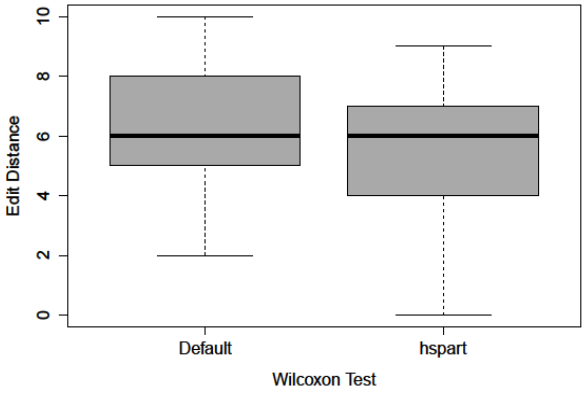


Figure A4.266– 2nd Run-Yahoo-Scenario1-Edit Distance-Wilcoxon Test-Default vs. hspart

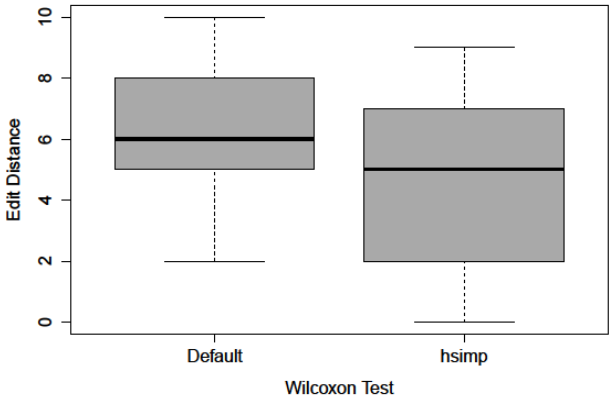


Figure A4.267– 2nd Run-Yahoo-Scenario1-Edit Distance-Wilcoxon Test-Default vs. hsimp

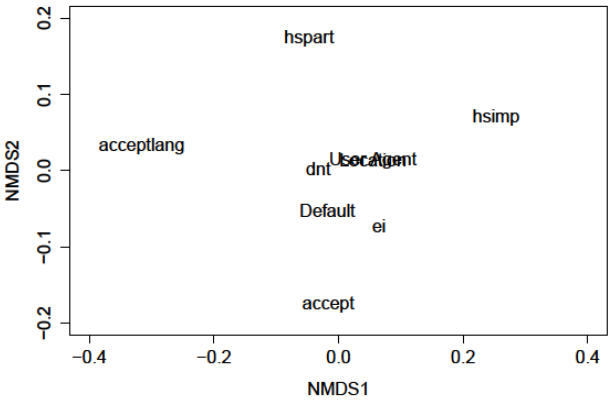


Figure A4.268– 2nd Run-Yahoo-Scenario1-Edit Distance-NMDS- Query Parameters

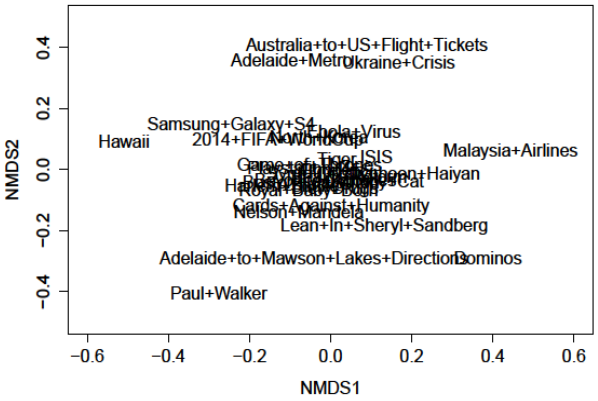


Figure A4.269– 2nd Run-Yahoo-Scenario1-Edit Distance-NMDS- Query Terms

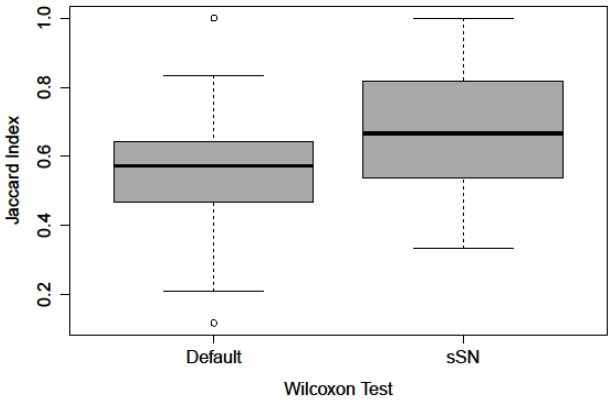


Figure A4.270– 2nd Run-Yahoo-Scenario2-Jaccard Index-Wilcoxon Test-Default vs. sSN

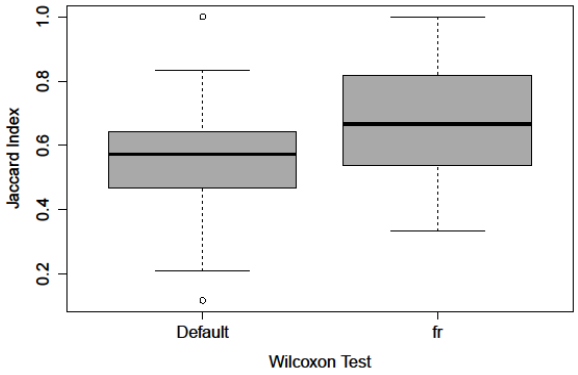


Figure A4.271– 2nd Run-Yahoo-Scenario2-Jaccard Index-Wilcoxon Test-Default vs.fr

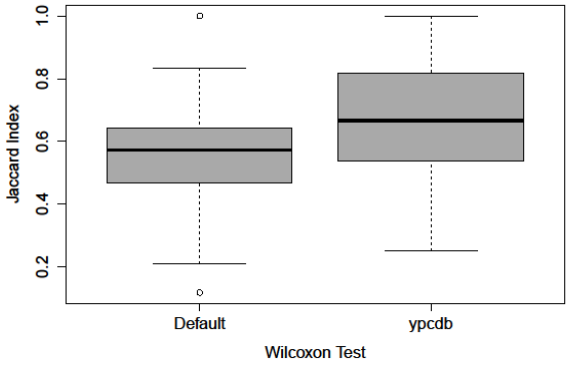


Figure A4.272– 2nd Run-Yahoo-Scenario2-Jaccard Index-Wilcoxon Test-Default vs. ypcdb

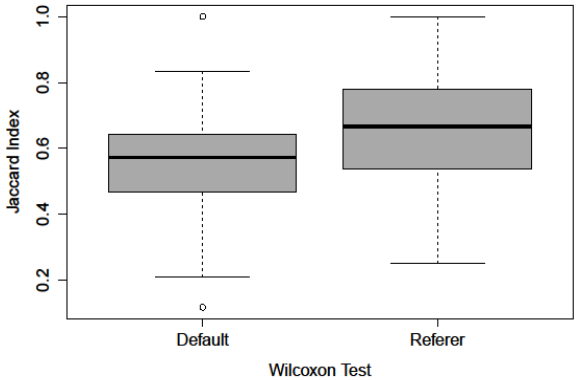


Figure A4.273– 2nd Run-Yahoo-Scenario2-Jaccard Index-Wilcoxon Test-Default vs. Referer

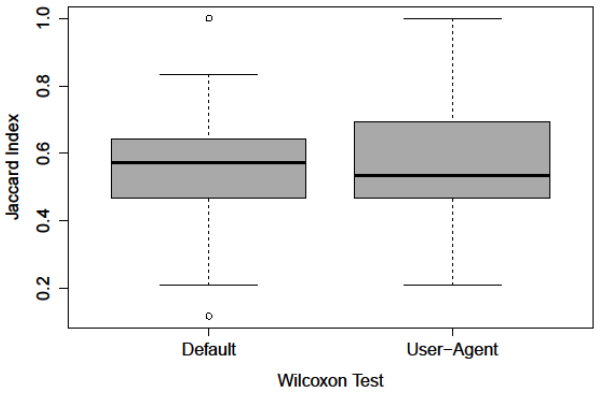


Figure A4.274– 2nd Run-Yahoo-Scenario2-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

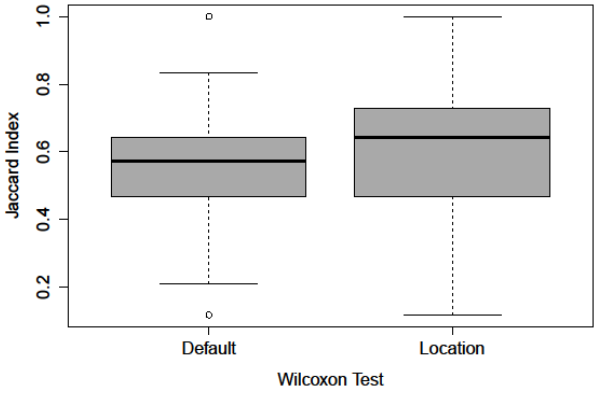


Figure A4.275– 2nd Run-Yahoo-Scenario2-Jaccard Index-Wilcoxon Test-Default vs. Location

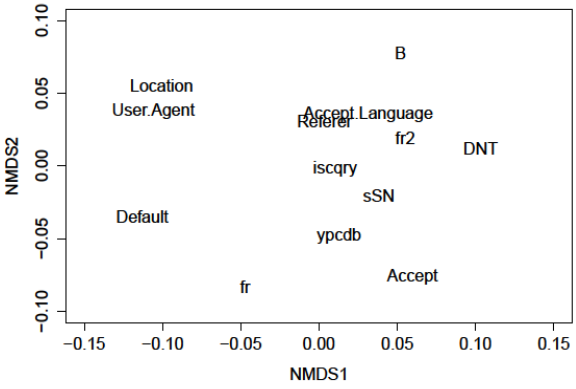


Figure A4.276– 2nd Run-Yahoo-Scenario2-Jaccard Index-NMDS-Query Parameters

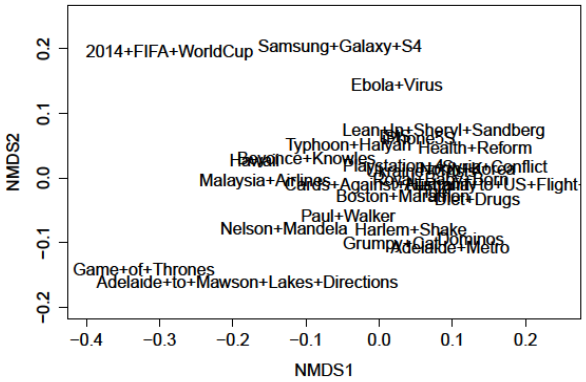


Figure A4.277– 2nd Run-Yahoo-Scenario2-Jaccard Index-NMDS-Query Terms

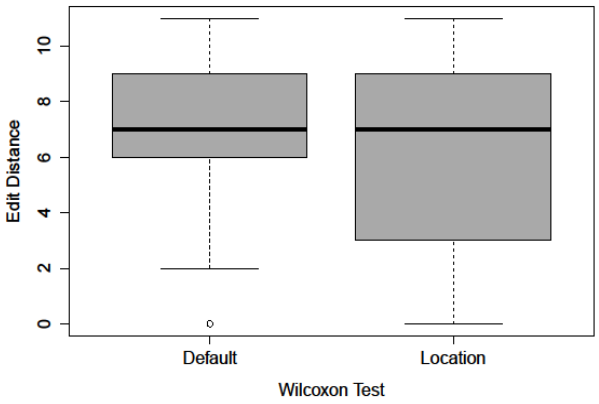


Figure A4.278– 2nd Run-Yahoo-Scenario2-Edit Distance-Wilcoxon Test-Default vs. Location

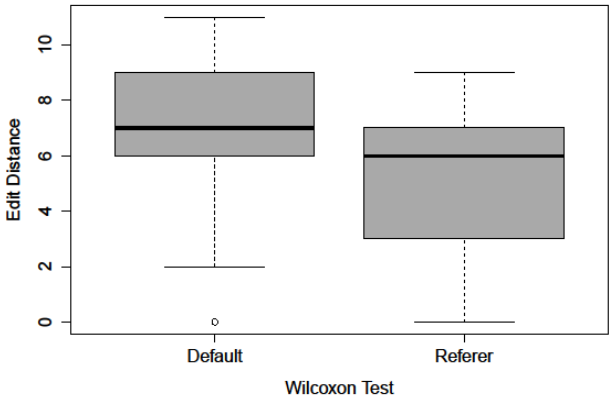


Figure A4.279– 2nd Run-Yahoo-Scenario2-Edit Distance-Wilcoxon Test-Default vs.Referer

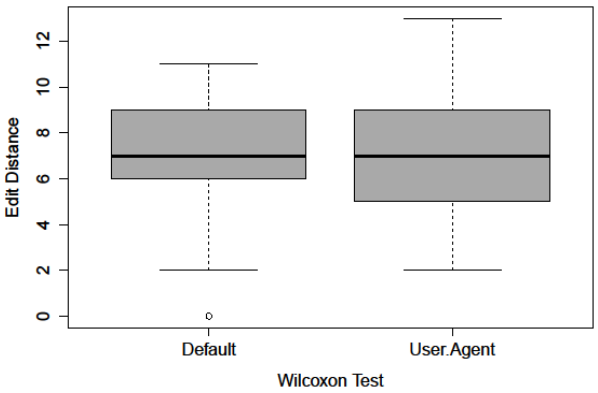


Figure A4.280– 2nd Run-Yahoo-Scenario2-Edit Distance-Wilcoxon Test-Default vs. User-Agent

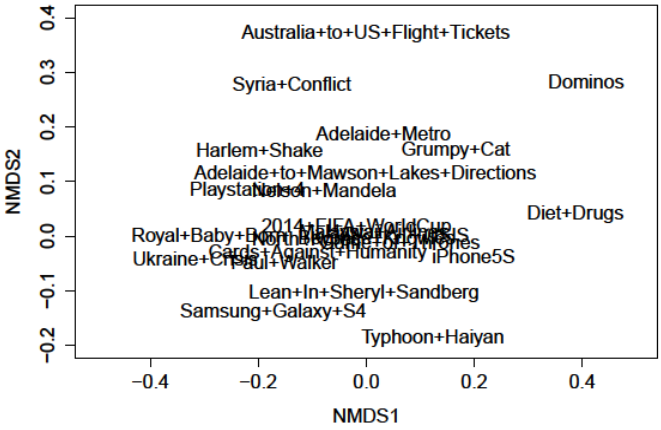


Figure A4.281– 2nd Run-Yahoo-Scenario2-Edit Distance-NMDS-Query Terms

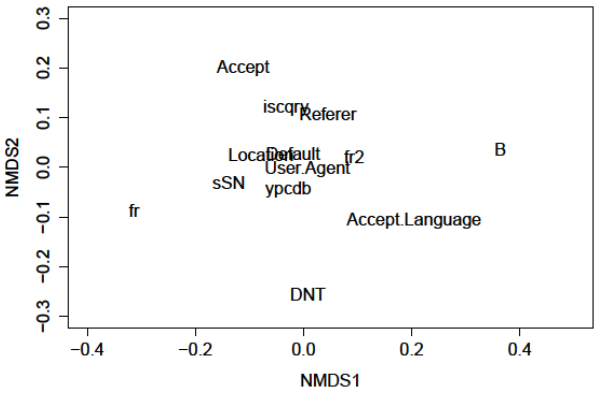


Figure A4.282– 2nd Run-Yahoo-Scenario2-Edit Distance-NMDS-Query Parameters

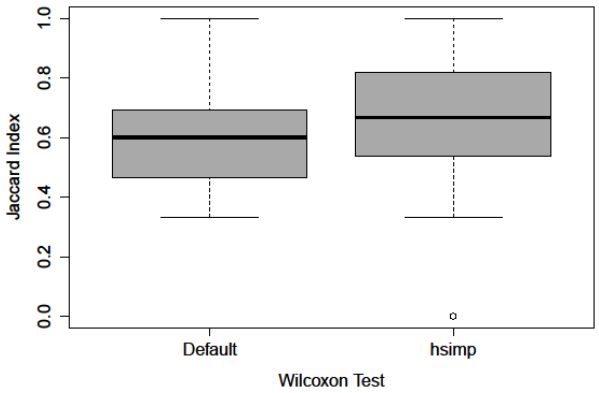


Figure A4.283– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. hsimp

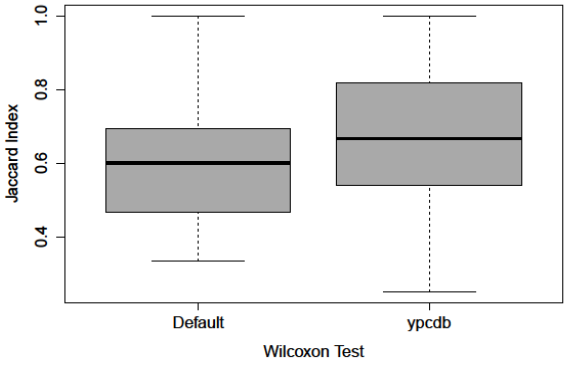


Figure A4.286– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. ypcdb

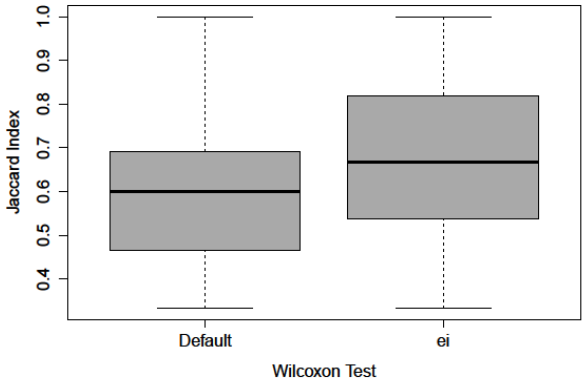


Figure A4.284– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. ei

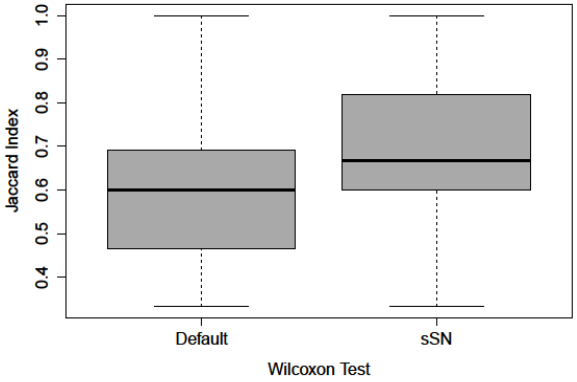


Figure A4.287– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. sSN

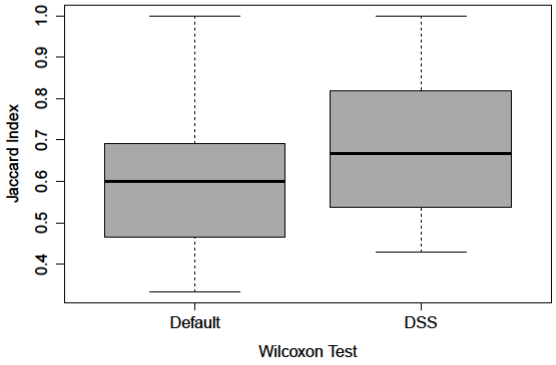


Figure A4.285– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. DSS

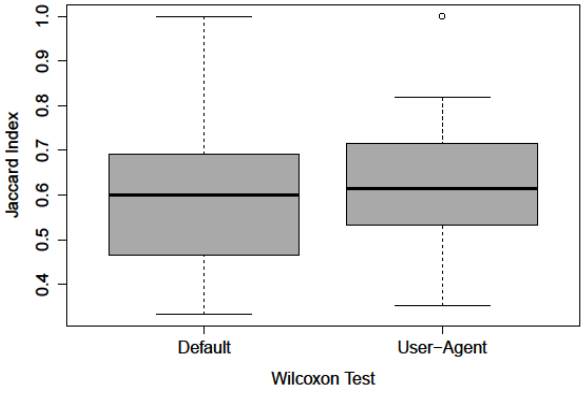


Figure A4.288– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

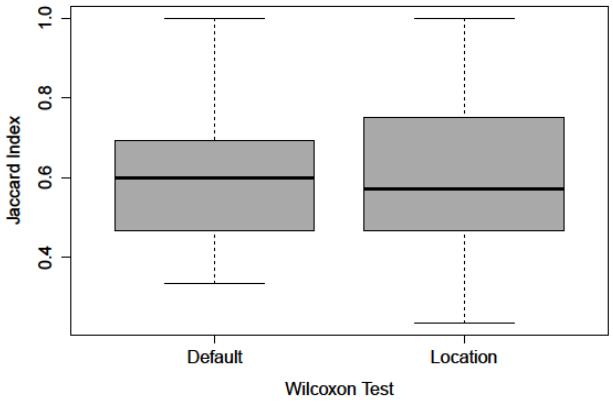


Figure A4.289– 2nd Run-Yahoo-Scenario3-Jaccard Index-Wilcoxon Test-Default vs. Location

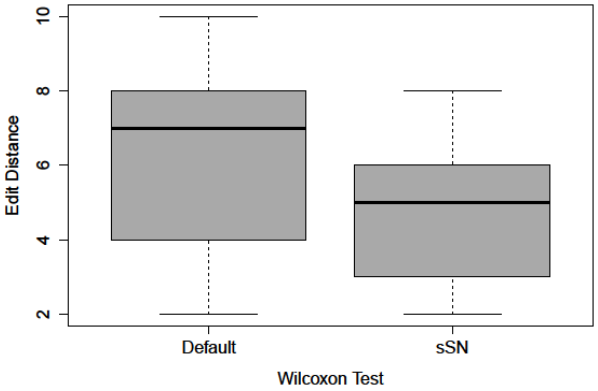


Figure A4.292– 2nd Run-Yahoo-Scenario3-Edit Distance-Wilcoxon Test-Default vs. sSN

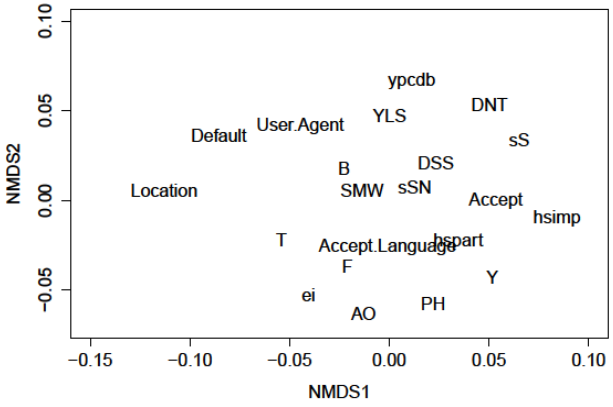


Figure A4.290– 2nd Run-Yahoo-Scenario3-Jaccard Index-NMDS-Query Parameters

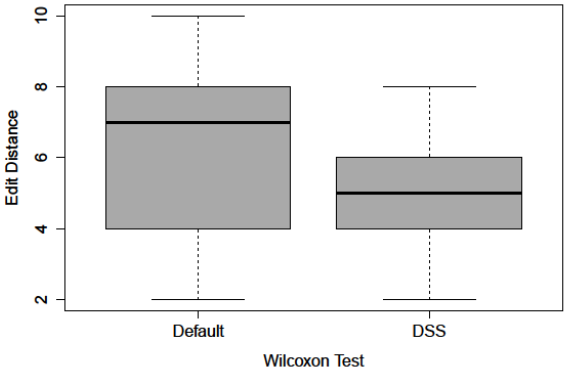


Figure A4.293– 2nd Run-Yahoo-Scenario3-Edit Distance-Wilcoxon Test-Default vs. DSS

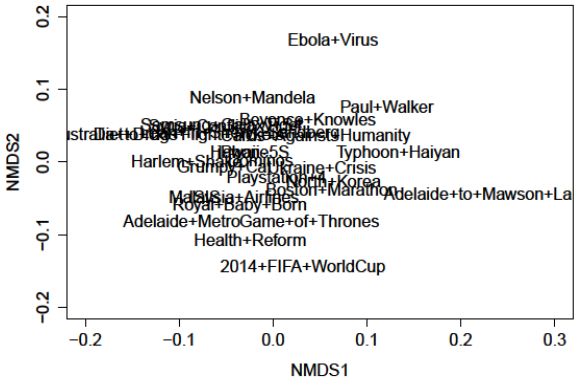


Figure A4.291– 2nd Run-Yahoo-Scenario3-Jaccard Index-NMDS-Query Terms

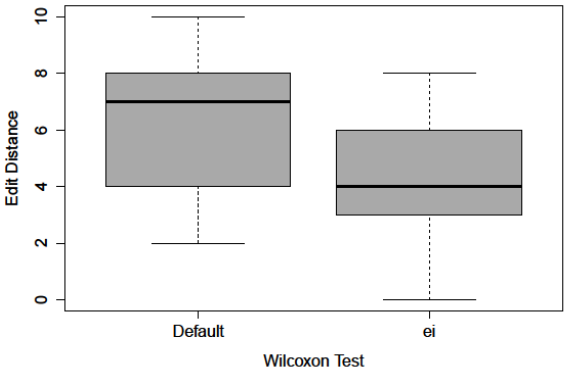


Figure A4.294– 2nd Run-Yahoo-Scenario3-Edit Distance-Wilcoxon Test-Default vs. ei

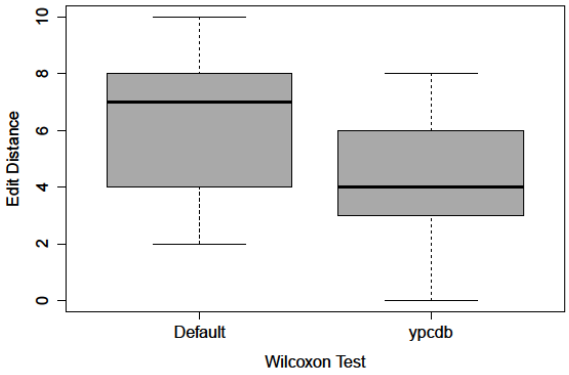


Figure A4.295– 2nd Run-Yahoo-Scenario3-Edit Distance-Wilcoxon Test-Default vs. ypcdb

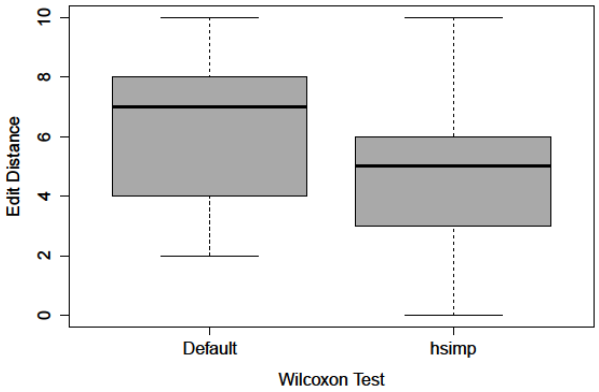


Figure A4.296– 2nd Run-Yahoo-Scenario3-Edit Distance-Wilcoxon Test-Default vs. hsimp

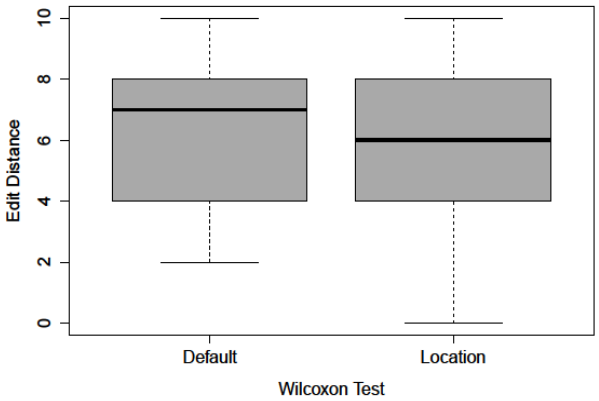


Figure A4.297– 2nd Run-Yahoo-Scenario3-Edit Distance-Wilcoxon Test-Default vs. Location

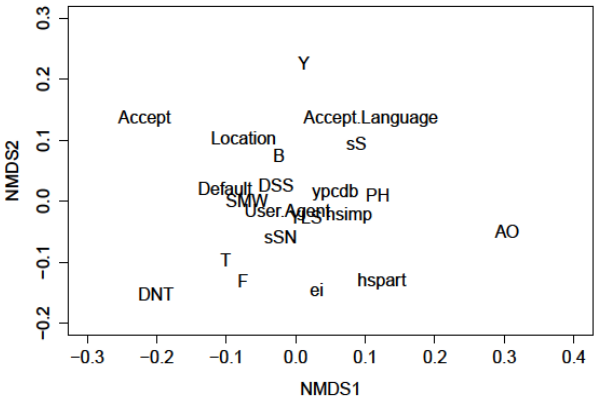


Figure A4.298– 2nd Run-Yahoo-Scenario3-Edit Distance-NMDS-Query Parameters

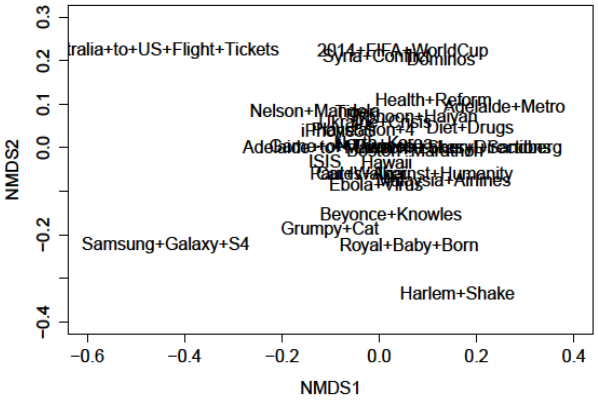


Figure A4.299– 2nd Run-Yahoo-Scenario3-Edit Distance-NMDS-Query Terms

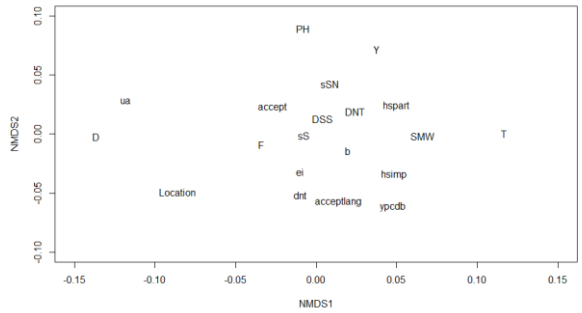


Figure A4.300– 2nd Run-Yahoo-Scenario 4- Jaccard Index-NMDS-Query Parameters

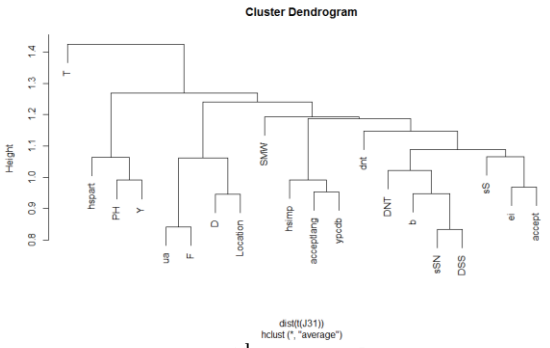


Figure A4.301– 2nd Run-Yahoo-Scenario 4- Jaccard Index-NMDS-Query Parameters – Cluster Dendrogram

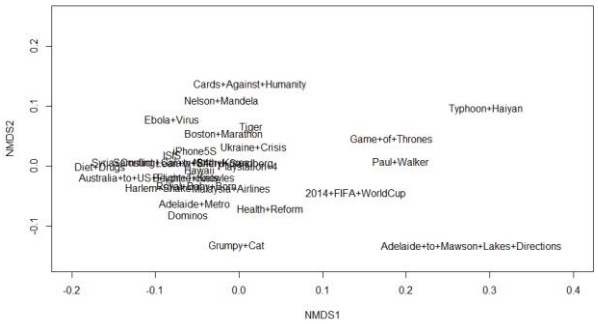


Figure A4.302– 2nd Run-Yahoo-Scenario 4- Jaccard Index-NMDS-Query Terms

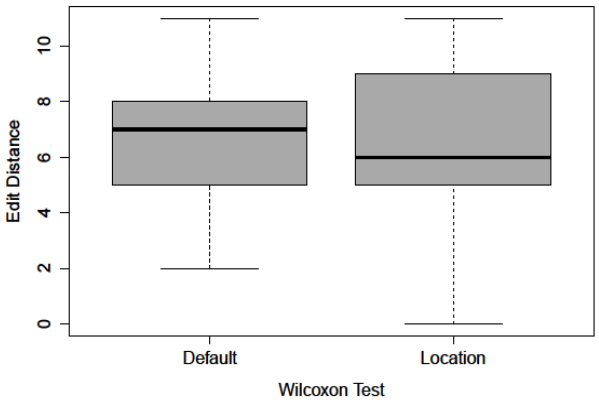


Figure A4.303– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. Location

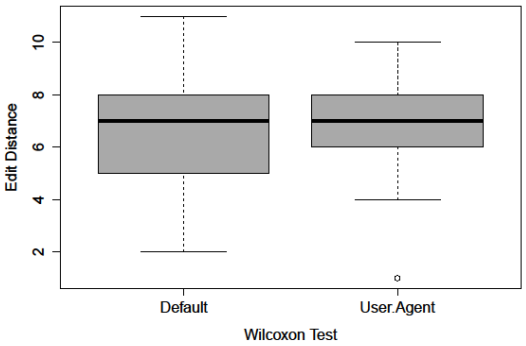


Figure A4.304– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. User-Agent

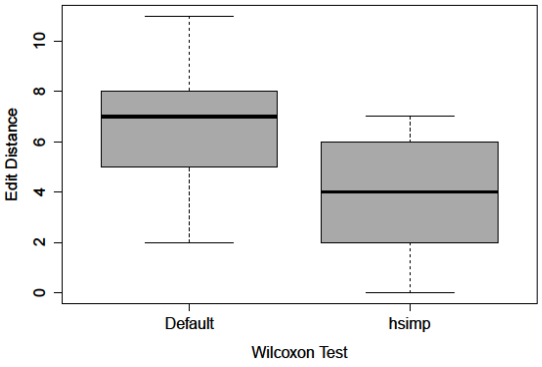


Figure A4.305– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. hsimp

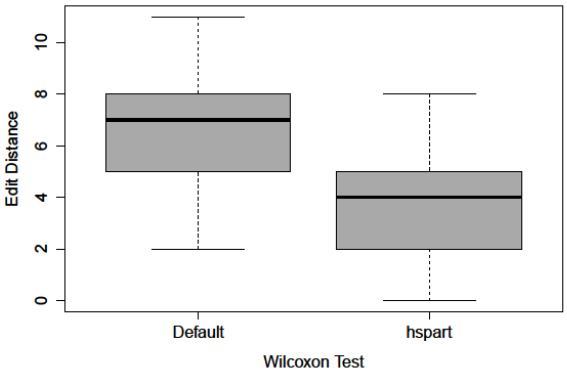


Figure A4.306– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. hspart

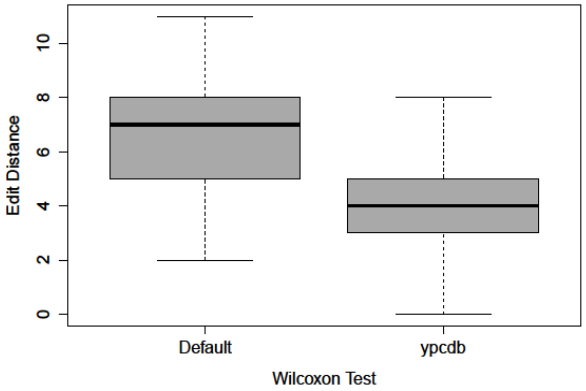


Figure A4.307– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. ypcdb

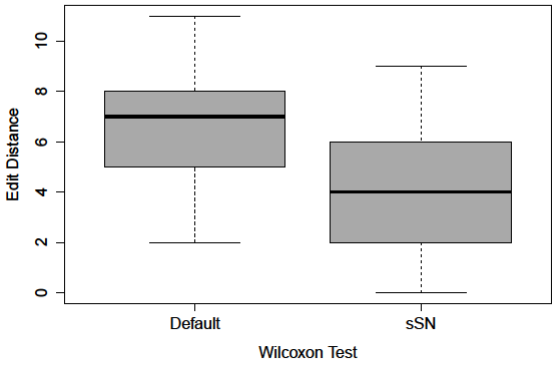


Figure A4.308– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. sSN

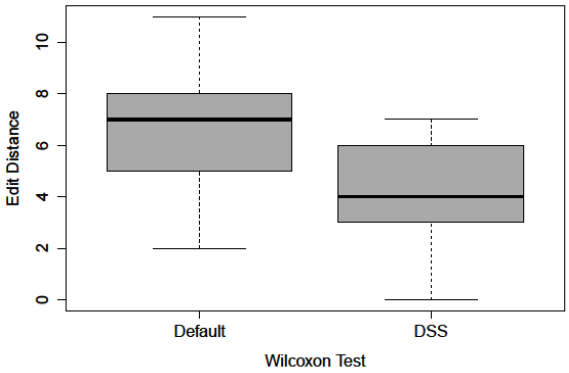


Figure A4.309– 2nd Run-Yahoo-Scenario 4-Edit Distance-Wilcoxon Test- Default vs. DSS

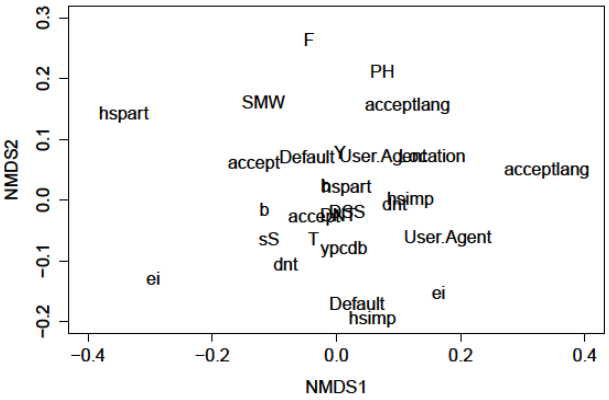


Figure A4.310– 2nd Run-Yahoo-Scenario 4-Edit Distance-NMDS-Query Parameters

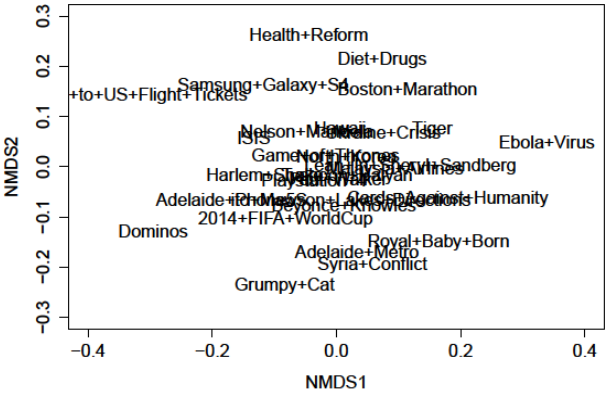


Figure A4.311– 2nd Run-Yahoo-Scenario 4-Edit Distance-NMDS-Query Terms 3rdRun – January 2016

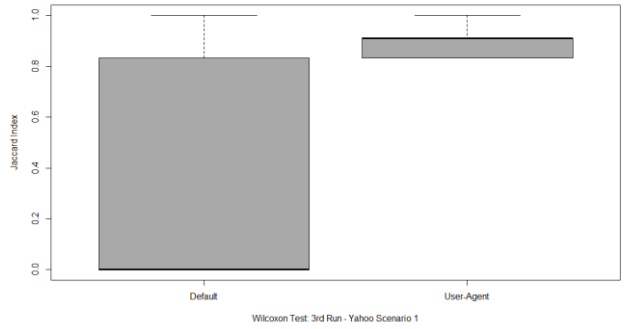


Figure A4.312– 3rd Run-Yahoo-Scenario 1-Jaccard Index- Default vs. User-Agent

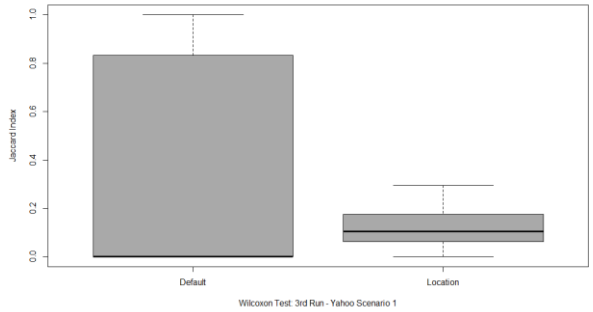


Figure A4.313– 3rd Run-Yahoo-Scenario 1-Jaccard Index- Default vs. Location

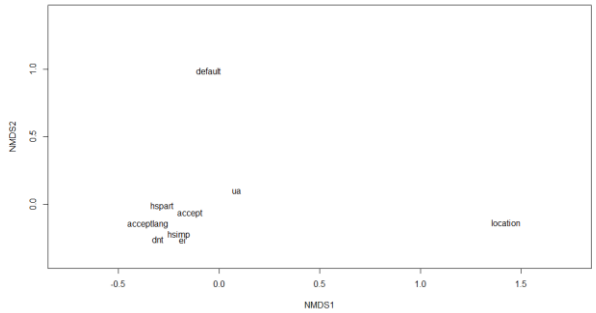


Figure A4.314– 3rd Run-Yahoo-Scenario 1-Jaccard Index-NMDS-Query Parameters

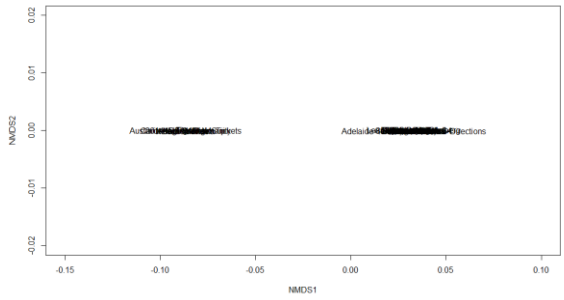


Figure A4.315– 3rd Run-Yahoo-Scenario 1-Jaccard Index-NMDS-Query Terms

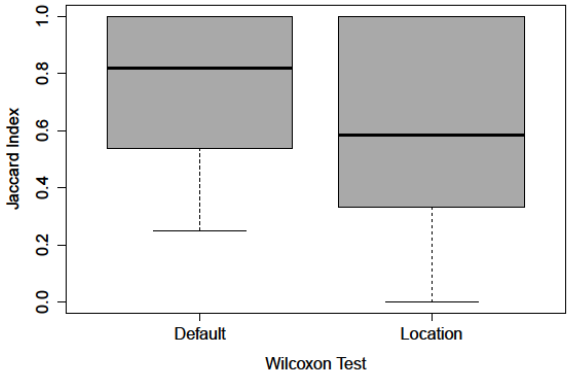


Figure A4.316– 3rd Run-Yahoo-Scenario 1-Edit Distance- Wilcoxon Test- Default vs. Location

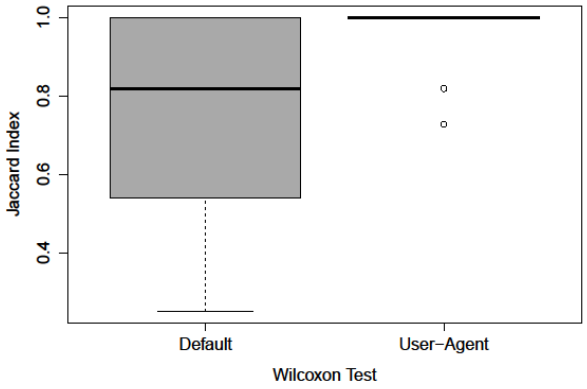


Figure A4.317– 3rd Run-Yahoo-Scenario 1-Edit Distance- Wilcoxon Test- Default vs. User-Agent

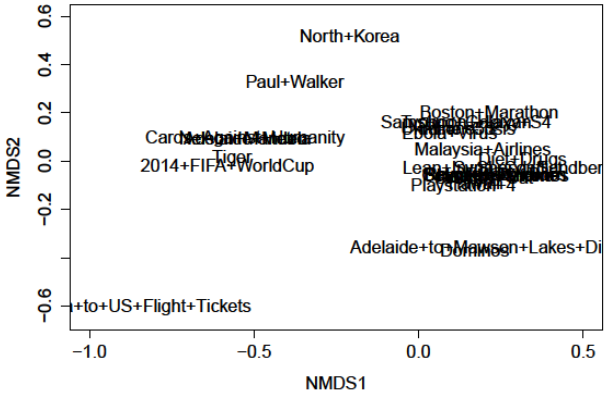


Figure A4.318– 3rd Run-Yahoo-Scenario 1-Edit Distance- NMDS-Query Terms

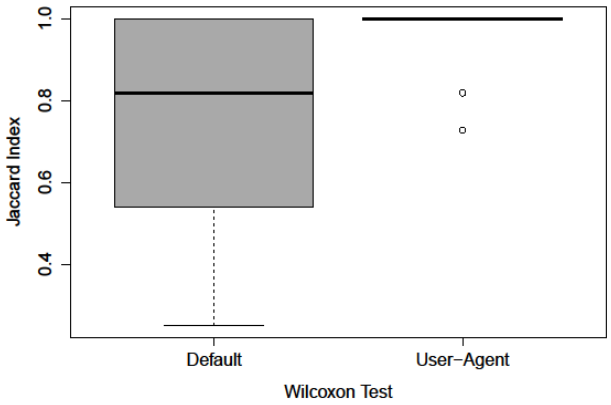


Figure A4.319– 3rd Run-Yahoo-Scenario 3-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

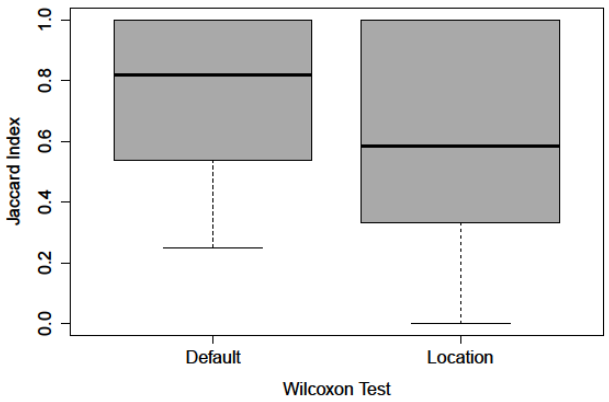


Figure A4.320– 3rd Run-Yahoo-Scenario 3-Jaccard Index-Wilcoxon Test-Default vs. User-Agent

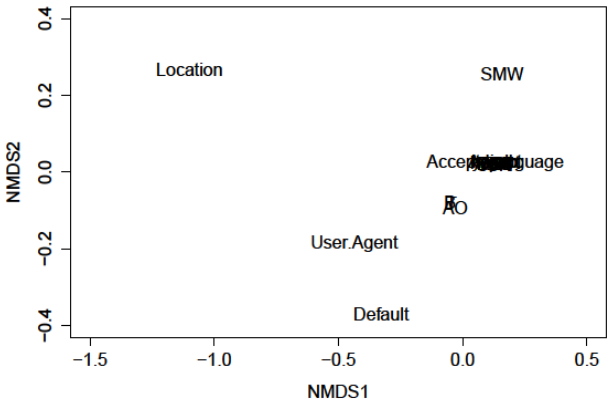


Figure A4.321– 3rd Run-Yahoo-Scenario 3-Jaccard Index-NMDS-Query Parameters

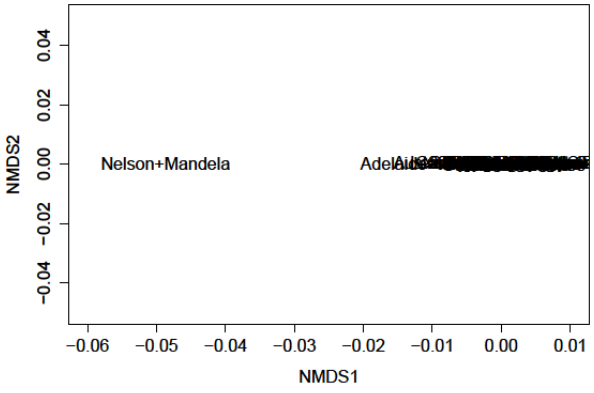


Figure A4.322– 3rd Run-Yahoo-Scenario 3-Jaccard Index-NMDS-Query Terms

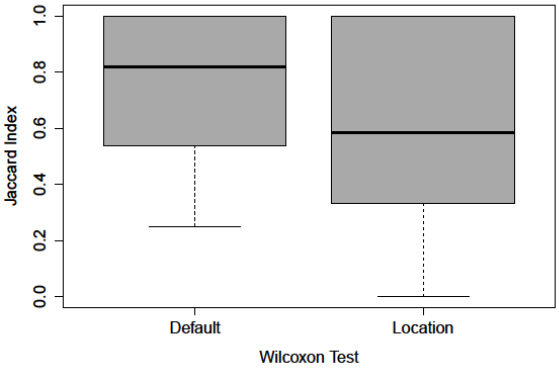


Figure A4.323– 3rd Run-Yahoo-Scenario 3-Edit Distance- Wilcoxon Test- Default vs. Location

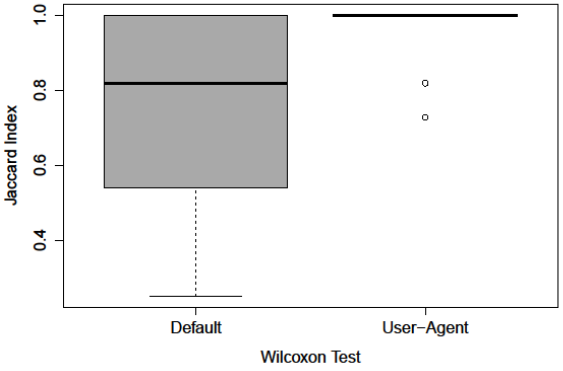


Figure A4.324– 3rd Run-Yahoo-Scenario 3-Edit Distance- Wilcoxon Test- Default vs. User-Agent

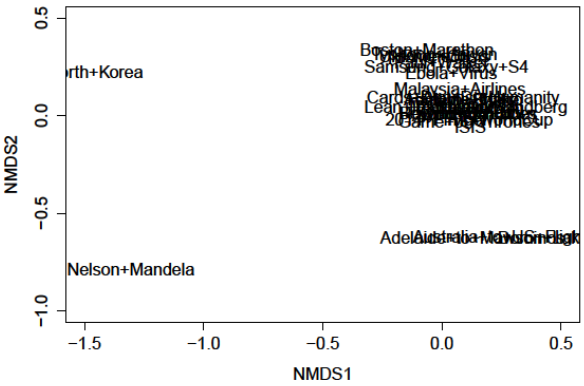


Figure A4.325– 3rd Run-Yahoo-Scenario 3-Edit Distance- NMDS-Query Terms

Google & StartPage Comparison
Scenarios (1st Run October 2014)

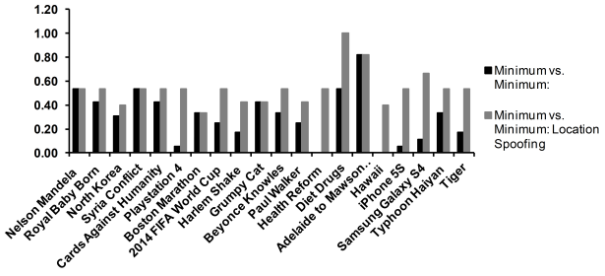


***Figure A4.327*-1st Run- Google & StartPage Comparison – Jaccard Index: Query Terms vs. Scenarios**

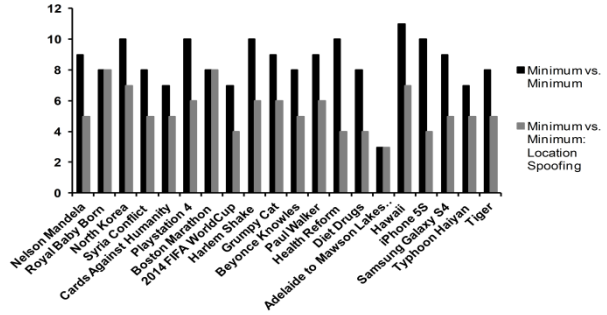


***Figure A4.327*-1st Run- Google & StartPage Comparison – Edit Distance: Query Terms vs. Scenarios**

Appendix A5.1 Survey Instrument on Web Search and User Privacy

You are invited to fill out a questionnaire as part of a research study for the Degree of Doctor of Philosophy at the University of South Australia.
The purpose of this survey is to build an understanding of how people search and their views on online privacy.
Please note:

- Your participation is voluntary and you can withdraw the information you provided at any point.
- All the information you provide will be anonymous, kept completely confidential and none of it will be released in any way that would permit identification of you.
- You can request to find out the results of the study.

Please fill out information as appropriate.

Survey Questions
Let's talk about how you search online and what you think about personal data use in search results! I appreciate you taking the time to participate in this survey!

Section 1: Demographic Info

1. First up, some **Demographic questions:**

a) What is your age?*
<text-box to enter age>

b) What is your gender?*
i. Male
ii. Female
iii. <Other option>

c) What is your current role?*
a) Student- Bachelors
b) Student- Postgraduate Coursework
c) Student - Postgraduate Research
d) Staff Academic
e) Staff Professional
f) Other

d) What is your area of study / work?*
<text box - small>

e) Which culture, heritage or ethnicity do you most identify with?
<text box – small; free-form>

Section 2: Search Behaviour

2. Your Browsing Habits:

Remember that there are no right or wrong answers. We'd just like to hear your honest opinion

a) How much do you rely on the web to search for info?*

1	2	3	4	5

Not Reliant Extremely Reliant

b) How long have you used the web to search for info?*
   i. <5 years
   ii. 5+ years
   iii. 10+ years
   iv. 15+ years
   v. 20+ years

c) Which of the following web browsers do you frequently use?*
   i. Internet Explorer
   ii. Mozilla Firefox
   iii. Google Chrome
   iv. Safari
   v. Other

d) Which of the following web search engines do you use most often?*
   i. Google
   ii. Bing
   iii. Yahoo
   iv. StartPage
   v. Ixquick
   vi. DuckDuckGo
   vii. Other

e) For <Selected Search Engine from d)> -
   i. How frequently do you use this search engine?*

1	2	3	4	5

As little as once a day Multiple times during the day

   ii. How many devices do you use to access this search engine?*

1	2	3	4	5

One device Multiple devices

f) Other than < selected answer - 2.c> and <selected answer - 2.d> - Do you have any additional comments on how you use search engines or browsers?
<free short text>

Section 3: Evaluation of privacy expectations in personal data flow scenarios

3. Web Search Contexts:

Evaluate the following scenarios – How are your expectations about the **flow of personal data** met in the following contexts?

Remember that there are no right or wrong answers. We'd just like to hear your honest opinion.

a) Do you expect your search behaviour (i.e. links you visit, search terms, search preferences) to be tracked?* <4 expectations: expect; worried> <yes/no>

b) Do you expect your search history to be made public?* <4 expectations: expect; worried> <yes/no>

c) Do you expect personal data you share or disclose to be used for subsequent searches? <yes/no/other>

1	2	3	4	5

I'm highly concerned. Doesn't affect me I'm fine with this.

For the next set of scenarios: how well are your expectations about personal data met?

d) In the context of searching, if your personal data was made publicly accessible?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

e) When you search for a specific thing (say shoes or flight tickets) and subsequently at a later time, you see ads on that item (e.g. shoe sale) or related items (e.g. if you searched for flight tickets, then ads on hotel offers)?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

f) If the links that you click on a search results page are recorded?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

g) If personally-identifiable information is used to customise search results and these personalised search results are more relevant to my needs?*

1	2	3	4	5

I prefer my data never being used in this manner

I prefer my data being used for relevant results only in specific contexts

I prefer my personal data being used for relevant results in all contexts

h) If your search history, cookies and other technologies are used to track your online behaviour?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

i) If your browsing behaviour and search preferences were aggregated over time to create a detailed profile of you?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

j) If undisclosed parties have access to your browsing behaviour and search preferences?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

k) If undisclosed parties could infer and potentially identify you based on your browsing behaviour and search preferences? *

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

l) If your browsing behaviour and search preferences were sold to undisclosed parties?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

m) If your family members or friends accessed your online search history without your permission?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

n) If your co-workers or employers accessed your online search history without your permission?*

1	2	3	4	5

Expectations Not Satisfied Doesn't really affect me Expectations Satisfied

o) Would you prefer having the option to be anonymous or use a pseudonym online?*
<Yes, No, Other>

p) Do you prefer if search engines are more transparent regarding collection, use and storage of your data? *
<Yes, No, Other>

q) Do you prefer if search engines allowed you to access and control sensitive information they hold about you?*
<Yes, No, Other>

r) Do you want to be informed as to how your data is being kept secure and accurate?*
<Yes, No, Other>

s) Would you be comfortable if you are identified by an anonymised ID as opposed to being identified and known (by your name)? *
<Yes, No, Other>

t) How often do you read terms and conditions before giving consent to using an online service?*

1	2	3	4	5

Not At All Sometimes All the time

u) Does the length of the terms and conditions impact your choice to read it?*

1	2	3	4	5

v) Does the complexity of the terms and conditions impact your choice to read it?*

1	2	3	4	5

No Sometimes Yes

w) Do you have any other expectations about the use of personal data online? Please list them if so:

4. Have you seen these items when searching?

a) Have you seen the following pages?*

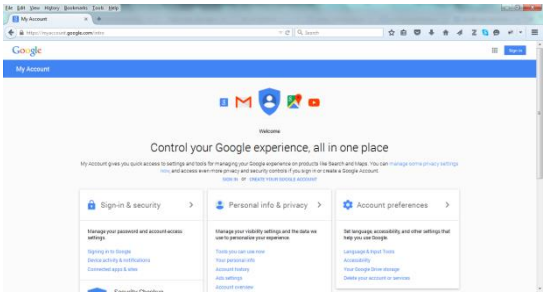
Select those that you have seen and used. Google is used as an example of the types of items. Please select the appropriate type if you've seen similar items from other search engines.

Choose as many as you like

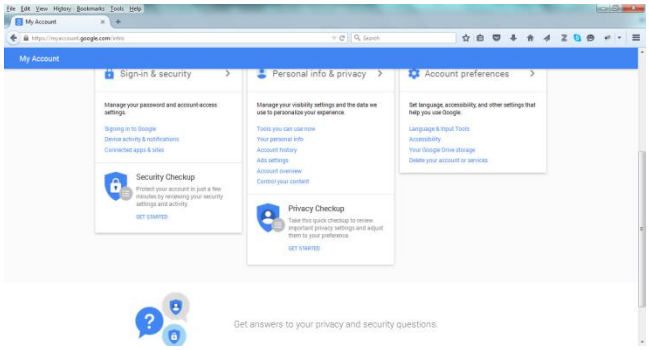
a) Google Privacy CheckUp Notice



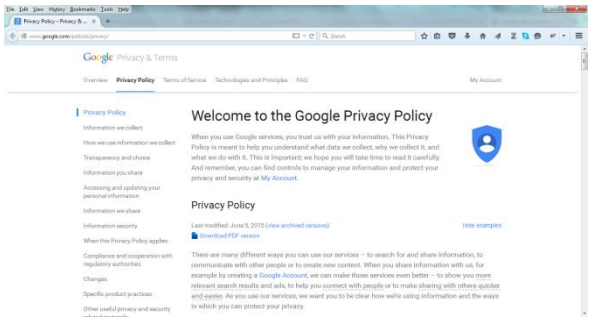
b) Google Privacy Info Center (Part 1)



c) Google Privacy Info Center (Part 2)



d) Google Privacy Policies



e) Not Seen of Used Any of These Options

f) Other – list any

b) Have you used the following technologies or tools?*

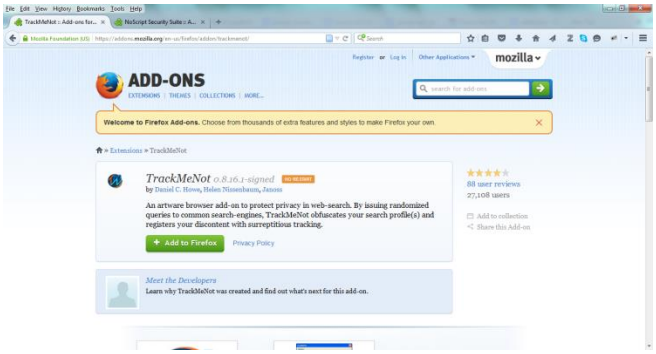
Select those that you have seen and used. The images are only examples of the type of technologies or tools you may have used.

Choose as many as you like

i. Tools that block Ads such as Adblock



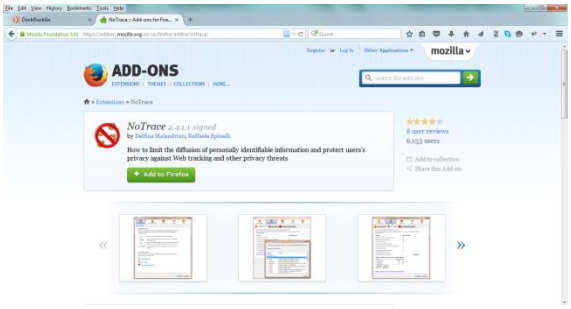
ii. Tools that prevent tracking – TrackMeNot



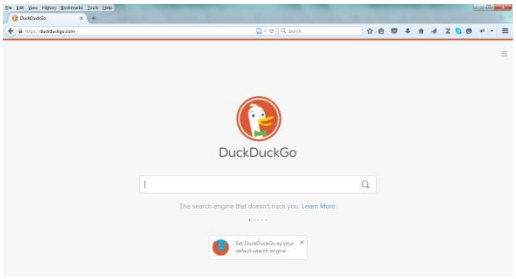
iii. Tools that prevent 3rd party Tracking



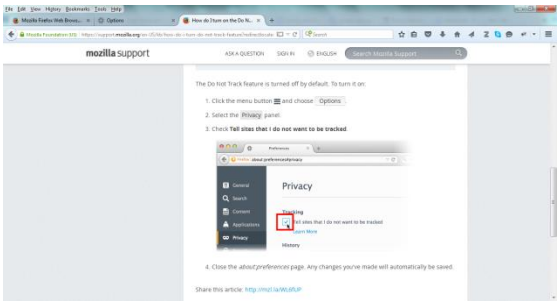
iv. Tools that let you control your identity online



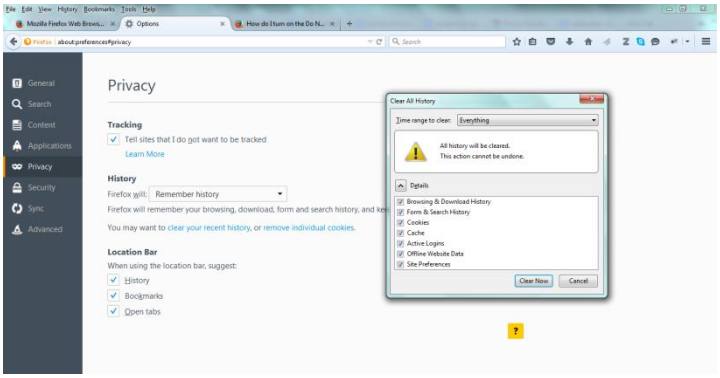
v. Privacy-enhanced Search Engine



vi. Do Not Track – Option to Limit Tracking in the Browser



vii. Your Browser Privacy Settings



viii. Not Seen or Used Any of These Options

ix. Other – Please list any other tools you use

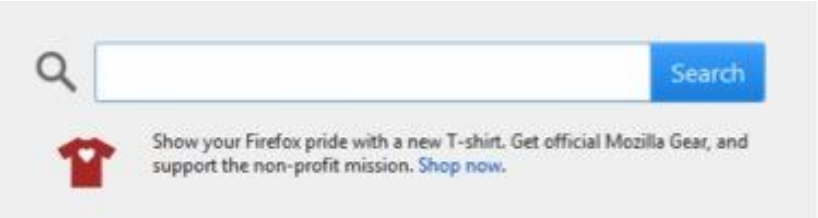
5. Give an example of an instance where you felt your expectations about privacy were not met when you searched online and your reasons why you think this is so.
   If you don't have an example of such an instance, give an example where you didn't mind your info being used to provide relevant search results.*

   <text box long>
6. Which of the following items would you like search engines to inform you about?*
   Choose as many as you like.
   a) Option to decide how my Personally Identifiable Information (PII) is used
   b) Search engines should disclose security safeguards used to protect my PII
   c) Search engines should inform me before using my data
   d) Search engines should disclose how my data is to be used
   e) I should be notified when privacy policies or practices change
   f) I should be notified when my information is disclosed, sold or shared with my own government
   g) I should be notified when my information is disclosed, sold or shared with third-parties
   h) Other

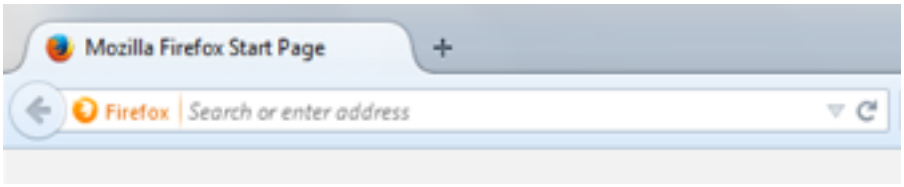
Section 4: User satisfaction of personalised search results

7. Search Task Scenarios – On your computer search for the following:
   Remember that there are no right or wrong answers. We'd just like to hear your honest opinion.
   <text box long>

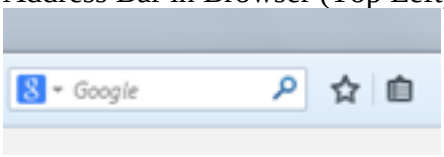
   a) Search for information about blood pressure –
      I. Briefly list the search term(s) you used.*
      II. Would this search be sensitive to you?
      III. Would you expect your personally-identifiable information to be used to tailor search results to you in this context?
      IV. Which of these search methods did you use? Google and Mozilla Firefox are only used as examples.



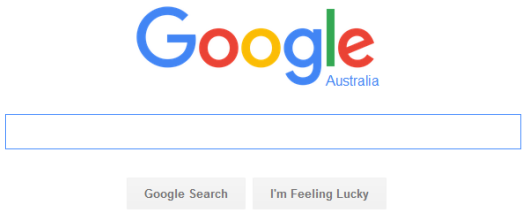
Search Bar in Browser (Middle)



Address Bar in Browser (Top Left)



Search Bar in Browser (Top Right)



Search Engine Page

V. Which search engine did you use?
   i. Google
   ii. Bing
   iii. Yahoo
   iv. StartPage
   v. Ixquick
   vi. DuckDuckGo
   vii. Other

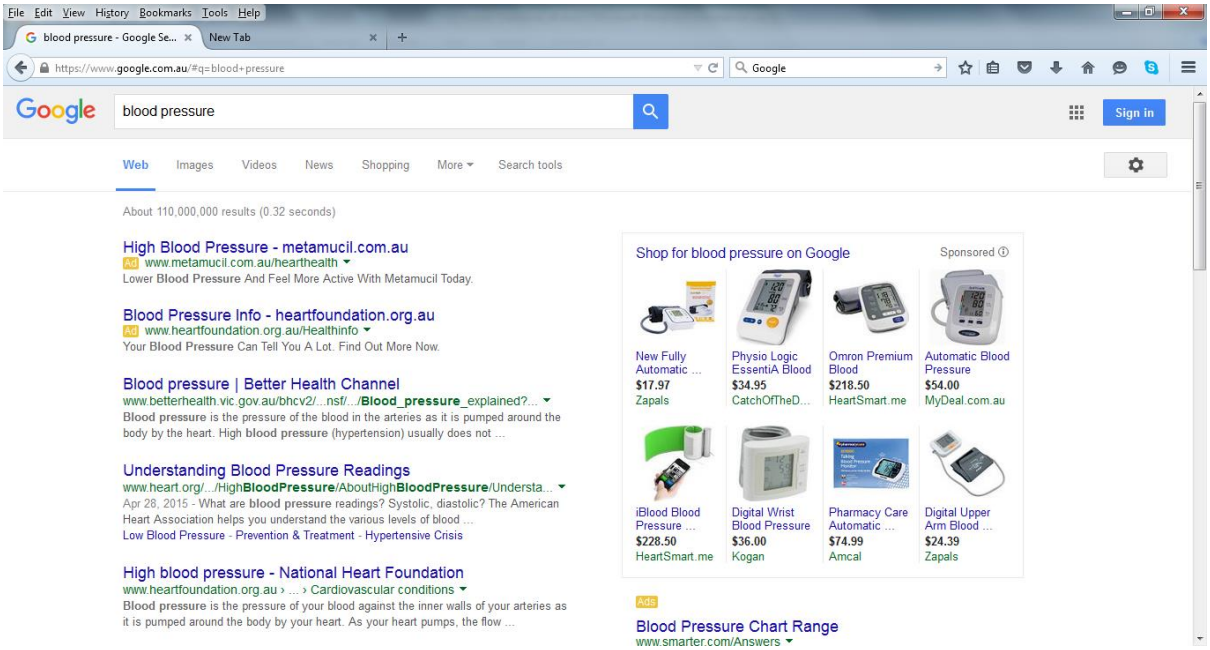
VI. Which browser did you use?
   i. Internet Explorer
   ii. Mozilla Firefox
   iii. Google Chrome
   iv. Safari
   v. Other

VII. Are you logged in to your email account?



Are you signed into your email account? For example: Gmail if using Google?

VIII. Comparing the below search result to your search result, which one do you prefer? <Yes, No, Other>

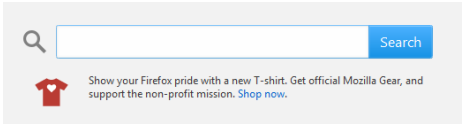


IX. How does this search result meet your expectations about the flow of personal data?

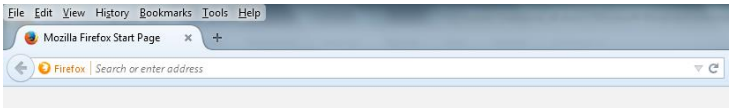
i. I can find what I would look for in the selected search result
ii. The selected search result somewhat matches the info I would look for
iii. I don't mind if identifiable information such as location is used to generate the result as long as it meets my needs
iv. I want to be able to better control how identifying information is reflected in search results
v. Both search results do not meet my expectations about the information contained in the search results
vi. I would like to prevent any type of data leakage

b) Search for flight tickets to Melbourne–

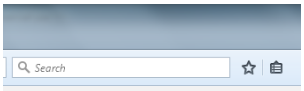
I. Briefly list the search term(s) you used.
II. Would this search be sensitive to you?
III. Would you expect your personally identifiable information to be used to tailor search results to you in this context?
IV. Which of these search methods did you use?



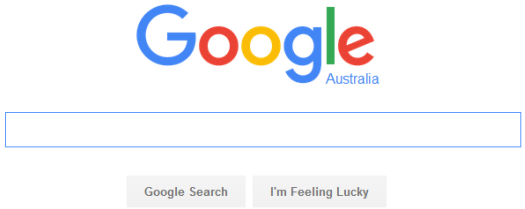
Search Bar in Browser



Address Bar in Browser



Search Bar at the Top in Browser

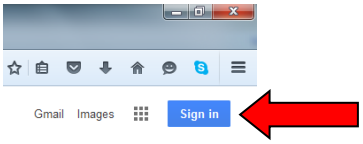


Search Engine Page

V. Which Search Engine did you use?
   i. Google
   ii. Bing
   iii. Yahoo
   iv. StartPage
   v. Ixquick
   vi. DuckDuckGo
   vii. Other

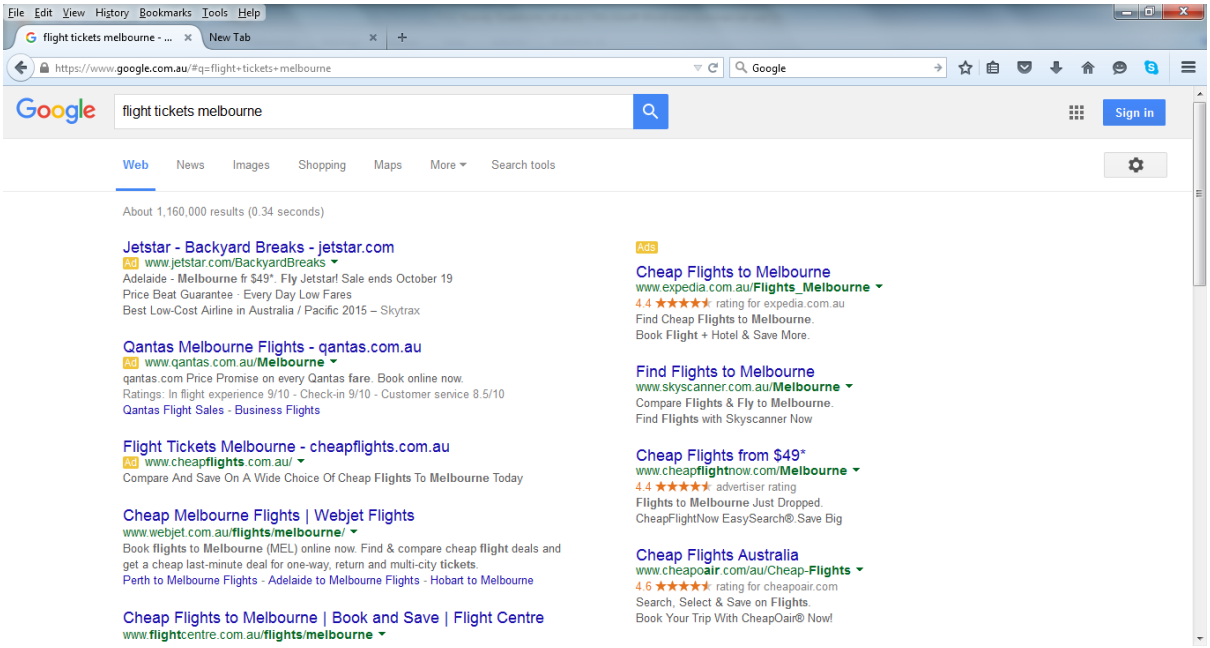
VI. Which browser did you use?
   i. Internet Explorer
   ii. Mozilla Firefox
   iii. Google Chrome
   iv. Safari
   v. Other

VII. Are you logged in to your account?



Are you signed into your account say Google account if using Google?

VIII. Comparing the below search result to your search result, which one do you prefer?



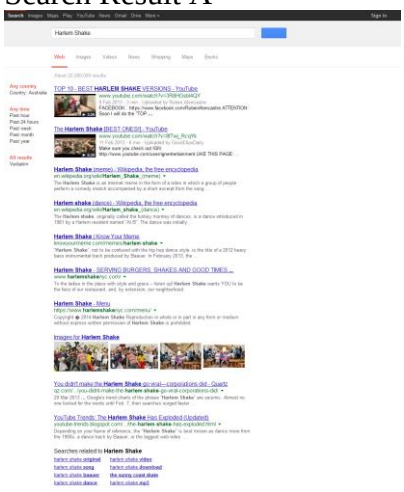
IX. How does this search result meet your expectations about the flow of personal data?

i. I can find what I would look for in the selected search result
ii. The selected search result somewhat matches the info I would look for
iii. I don't mind if identifiable information such as location is used to generate the result as long as it meets my needs
iv. I want to be able to better control how identifying information is reflected in search results
v. Both search results do not meet my expectations about the information contained in the search results
vi. I would like to prevent any type of data leakage

8. Compare these Search Results:

Say you searched about Harlem Shake and got the following results: which search result do you prefer? <image will be scaled on actual survey; Picture choice: Search Result A, Search Result B or other>

Search Result A



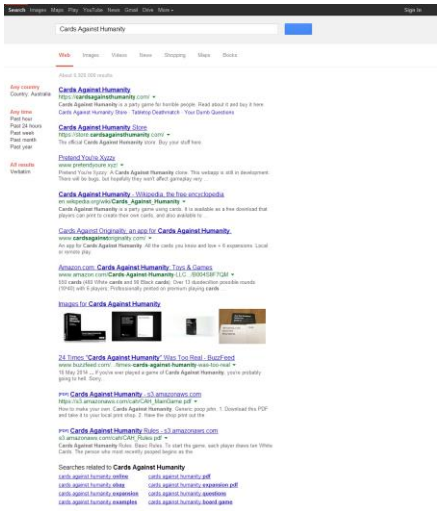
Search Result B



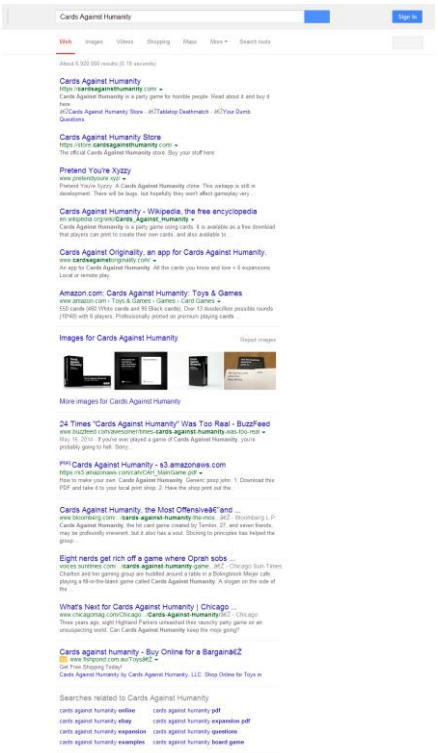
<Other>

Say you searched about 'Cards Against Humanity' and got the following results: which search result do you prefer? <Picture choice: Search Result A, Search Result B or other>

Search Result A



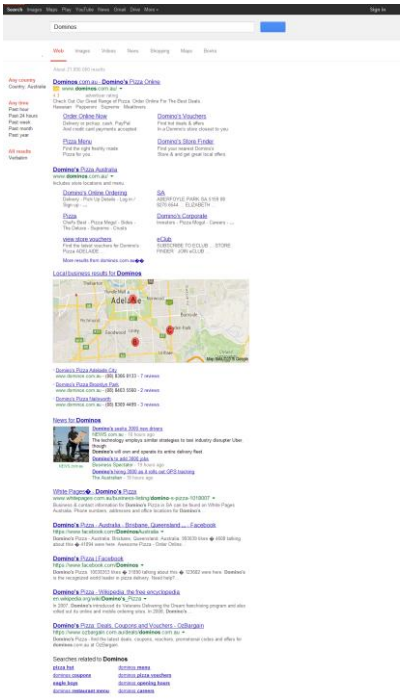
Search Result B



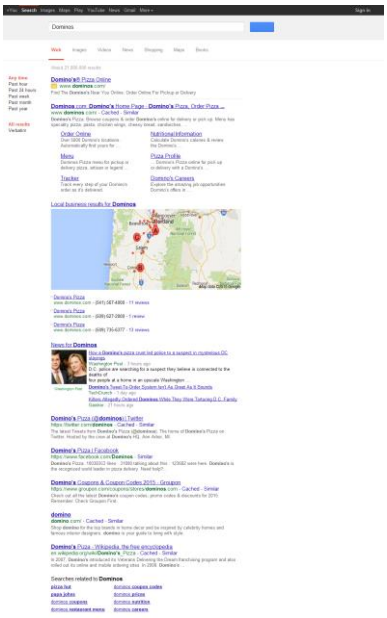
<Other>

Say you searched about 'Domino's Pizza' and got the following results: which search result do you prefer?
<Picture choice: Search Result A, Search Result B or other>

Search Result A



Search Result B



<Other>

c) How does this search result meet your expectations about the flow of personal data?

i. I can find what I would look for in the selected search result
ii. The selected search result somewhat matches the info I would look for
iii. I don't mind if identifiable information such as location is used to generate the result as long as it meets my needs
iv. I want to be able to better control how identifying information is reflected in search results
v. Neither of the search results meet my expectations about the information contained in the search results
vi. I would like to prevent any type of data leakage

Section 5: Privacy Index based on Westin (2003)

9. Your Views on Privacy

Remember that there are no right or wrong answers. We'd just like to hear your honest opinion.

a) Do you agree with this statement:
1.Consumers have lost all control over how personal information is collected and used by companies.*

1	2	3	4	5

Strongly Disagree Neutral Strongly Agree

b) Do you agree with this statement:

2. Most businesses handle the personal information they collect about consumers in a proper and confidential way.*

1	2	3	4	5

Strongly Disagree Neutral Strongly Agree

c) Do you agree with this statement:

3. Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today.*

1	2	3	4	5

Strongly Disagree Neutral Strongly Agree

END

Finally, some quick questions to wrap up the survey:

- Would you like to be invited to take part in a follow-up observation study?
  All you have to do is to perform a set of search tasks and in appreciation of your time and effort; you double your chances of winning a $50 voucher (Coles/Myer Gift card) in the prize draw.
  If yes, please type in your contact email address.
  If no, please indicate this.
- Would you like to enter the survey prize draw to stand a chance of winning one of four $50 vouchers (Coles & Myers Gift Card)?
  Please enter your contact email address (if you haven't already entered it in the above question) if you would like to enter the prize draw.
- If you would like to be notified of the results of this study, please select yes. (Yes/No)
- Any other comments
- Thank you for your time and effort in completing this survey! Press Continue to Submit your responses.

Thank you for completing this questionnaire! If you would like to be notified of the results of this study, please inform the researcher.