Trinity College Dublin

Doctoral Thesis

Personal Privacy and Online Systems

Author:
Pól Mac Aonghusa

Supervisor:
Prof. Douglas Leith

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

School of Computer Science and Statistics

April 5, 2019

Declaration of Authorship

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Acknowledgements

I would like to express my sincere appreciation to Professor Douglas Leith for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously is very much appreciated.

Finally, I wish to thank my wife, Ann, and my family for their support and encouragement throughout.

When inspiration fails, remember, *Ansin, bhí timpiste …*

TRINITY COLLEGE DUBLIN

Abstract

School of Computer Science and Statistics

Doctor of Philosophy

Personal Privacy and Online Systems

by Pól Mac Aonghusa

A significant portion of the modern internet is funded by commercial return from customised content such as advertising where user interests are learned from users' online behaviour and used to display personalised content. Privacy becomes a concern when personalisation reveals evidence of learning about sensitive topics a user would rather keep private. Examples of potentially sensitive topics we consider include health, finance and sexual orientation.

In this thesis we develop novel technologies allowing users to improve control over their personal privacy. We consider three aspects of privacy protection here: i) detecting evidence of unwanted profiling, ii) assessing the potential impact of a threat, and, iii) a flexible framework to help users to take control the flow of information used in personalisation.

We model online systems as black-box adversaries with unknown internal workings but with an objective to maximise commercial utility. In a black-box environment absolute measures of privacy are problematic and so our formalism builds on a notion of privacy relative to a baseline. The relative models we develop have the advantage of being learn-able from observation of the black-box system and so can be readily implemented as practical technologies for privacy threat detection, analysis and privacy defence which we validate against data from well-known, real-world online systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"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."*
>
> Eric Schmidt, then Executive Chairman Google LLC, and now of Alphabet, (Jenkins Jr., 2010)

1.1 Context

In the interview quoted above, Eric Schmidt goes on to suggests that because Google knows "roughly who you are, roughly what you care about, roughly who your friends are", its algorithms could helpfully remind you what groceries you need to buy when passing a shop, (Jenkins Jr., 2010). At first glance this seems like a very useful trade-off; user data collection allows Google Search to be helpful in a personally aware way.

In reality, profiling also allows Google to learn about possible user interests, preferences and behaviours and so display targeted content tuned to attract user attention. Web search users seem to prefer a degree of personalised content, (Panjwani et al., 2013). Personalisation can, however, also reveal evidence of bias in the machine learning algorithms used to recommend content. Studies have shown that Google displayed ads for high-income jobs preferentially to men much more to women; and that advert related to arrest records were significantly more likely to appear when searching for names or college fraternities associated with African Americans, (Damm, 2019; Amit Datta et al., 2015).

Privacy becomes a concern when personalised content displays evidence of a preference towards topics a user considers sensitive and so wishes to keep private. The central question we ask in this thesis is – how can we detect, assess and control machine learning inference threatening privacy in web search? Our approach is to analyse inputs to, and outputs from, a search engine for evidence of inference toward sensitive topics. Rather than ask *how* a search engine generates personalised content, we treat search engines as black-boxes with hidden internal workings so that our work can be compared with verification of fairness and detection of unwanted inference in machine learning.

We take the view that personal privacy is not fixed, but rather it is fundamentally an ongoing risk management exercise where there are no absolute guarantees. Just as search engines adapt and learn, individuals must take ongoing responsibility to adapt, evaluate and manage their own balance between privacy and utility. Protecting user privacy is a practical trade-off between the needs of users to avoid unwanted personalisation and an Internet business model underpinned by personalisation. In answering our central question a crucial balance must be struck between empowering users with capabilities to detect, assess and limit privacy threats arising from personalisation while recognising the the necessity of maintaining a level of personalised content to sustain the free-to-use Internet ecosystem.

1.2 Motivation

Our specific interest is in privacy in web search. We are interested in evidence of unwanted learning only with respect to topics of interest defined by an individual user. We seek evidence of specific biases in choosing personalised content with respect to specific, private topics rather than in detection of general learning in the underlying algorithms employed by the search engine. When a topic is not regarded as private by a user it is not of interest.

We organise user interests into two categories for the purposes of our analysis and illustrated in Figure 1.1. The green box on the right of Figure 1.1 represents the subset of interests that an individual is happy to discuss in public. Privacy is not a concern when personalised content refers to public topics. There are also topics where an individual would rather keep her interests private, represented as the red box on the right of Figure 1.1. When online services respond with personalised content related to topics in the subset of private interest then privacy is a concern.

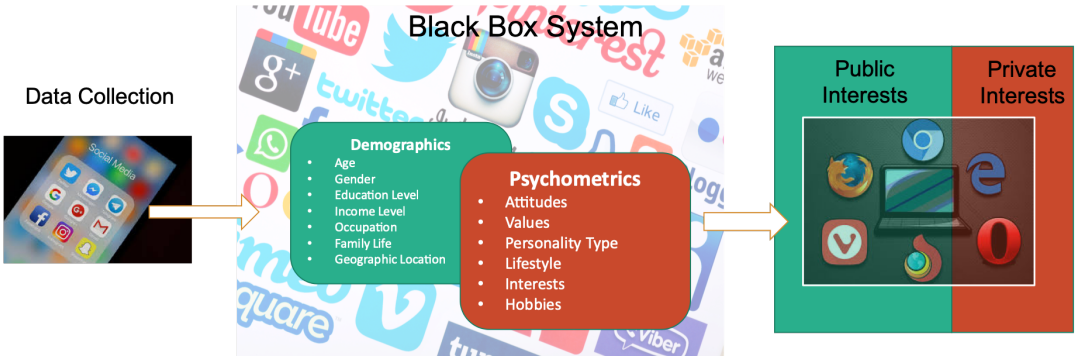


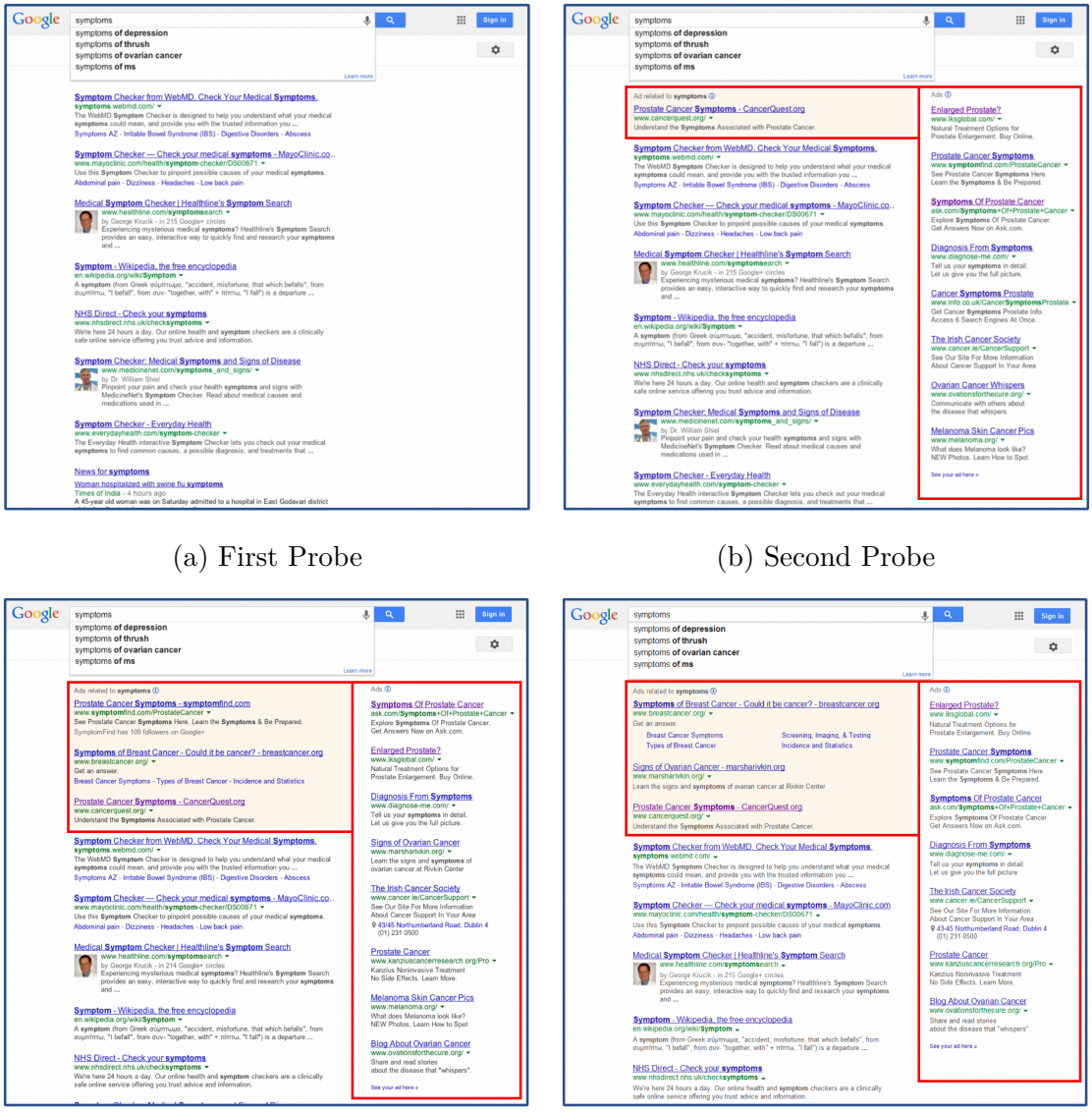
FIGURE 1.1: Balancing Personalisation with Privacy.

For illustration purposes, we have included examples of profiling by the online system in the middle section of Figure 1.1. Our approach, however, is to treat online services as black-boxes that do not reveal their internal workings except through the outputs produced by the black-box in response to individual user inputs.

The following informal example of an interaction illustrates potential concerns with personalisation during a Google Search. The interaction is performed on a laptop through a standard web browser. Before beginning, we remove obvious traces of local state such as browser history, cookies and caches. An anonymous user is used and IP address is constant through the interaction. In this way, observed changes in personalised content can be reasonably associated with active profiling during the interaction by Google search. During the interaction we ask a range of uninteresting queries about everyday topics such as weather, traffic and music to represent public interests. We mix occasional queries among the public queries about a specific private topic – "cancer" at intervals of 1 private query to every 2–3 public, uninteresting queries.

After every fourth private query about cancer, we issue a fixed "probe" query – it is the query "symptoms" in the example. We choose symptoms because it is sufficiently generic that Google could associate many medical conditions with it besides cancer. We compare what adverts appear as a result of the probe query in Figure 1.2.

An inspection of the search results in each of the sub-figures confirms that the probe query retrieves generic results – mostly related to symptom checkers. By contrast, personalised adverts build up through the interaction until adverts related to



(a) First Probe (b) Second Probe

(c) Third Probe (d) Fourth Probe

FIGURE 1.2: Changes in personalised adverts at successive probe queries for the topic "Cancer".

cancer pervade each response page. Even though queries about cancer are interleaved with uninteresting queries about everyday topics, the search engine has successfully identified a sensitive topic among the noise of general queries. Comparing the non-specific nature of the search results with the specific nature of the personalised, advert content suggests that by analysing personalised content, such as adverts, we might reasonably hope to spot evidence of profiling with respect to interests we regard as private.

1.3 Scope and Limitations

Our objective in this research is to demonstrate feasibility and utility of approaches with the minimum technical overhead. Our choice of underpinning techniques and technologies is chosen deliberately to be simple to comprehend and to implement. Accordingly we have chosen a PC-based browser platform for implementation and experimental evaluation. We elected to avoid mobile devices because of added complexities of implementation and to avoid concerns with hidden tracking and sharing, (Razaghpanah et al., 2018). We leave consideration of mobile web search to future research.

It seems reasonable to assume that a for-profit commercial search engine selects page content to maximise its expected revenue. This means that when a search engine infers that a particular advertising topic is likely to be of interest to a user, and so more likely to generate click through and sales, it is obliged to use this information when selecting which adverts to display. Since a revenue maximising search engine acts to display adverts associated with topics it detects are most interesting to the user, the potential exists to detect search engine learning via analysis of changes in the choice of displayed adverts and to inform the user of this learning.

Conversely, our work excludes the situation where the system does not reveal its hand through personalised content. The latter could happen when the system is not capable or is unwilling to personalise its output - for example when the real motive is data collection for undisclosed background processing or security analysis. These specific situations are left as cases best addressed by the law and through strong and active governance. We also exclude situations where hyper-personalisation at individual level is required, such as systems to support medical consultations, security or defence.

Our focus here is on privacy concerns arising from inference by the search engine resulting from explicit user web search interactions. Online systems, including search engines, gather data from many sources to produce personalised content. We consider direct identification techniques, such as IP tracing or browser finger-printing to be outside the scope of our current analysis. Implicit profiling effects due to Geo-location,

for example, also effect personalised content. Simple changes from one IP range to another can change the advert content of search results as Figure 6.1 in Chapter 6 illustrates. When creating personas for testing purposes we have tried to provide the minimum possible profiling information required for registration. We specifically avoided potentially revealing demographic information, such as age and gender for example, to isolate system learning effects from variations in profile demographics in so far as possible.

All of the technologies used here were implemented with open source tools and written in the Python language. The Natural Language Processing Bag-of-Words model we use is among the simplest possible that facilitates obtaining useful experimental results. More sophisticated language models, for example using n-grams or word embedding, will likely improve the capabilities of the tools. We focus on text-based advert content appearing on web search result pages. Our approach is to spot changes in frequency of occurrence of keyword features associated with topics we have defined as sensitive. By comparing keyword frequencies with baseline values learned from training data we hope to detect evidence of bias towards sensitive topics. We describe the training and verification setup used to learn baseline values of keyword frequencies in detail in experimental sections.

Exact reproducibility of results from experiment to experiment is difficult in a dynamic learning environment such as web search so that our results are presented as average effects over several experimental iterations. Personalised content varies from iteration to iteration, and in some cases personalised content may not appear at all. Programatic interaction with a web search engine is technically challenging. Web search engines have developed a sophisticated array of tools to to detect automated users. When inspected, a portion of search engine result pages with no personalised content occurred when our programs activated a search engine feature such as a CAPTCHA or other challenge. These events happened irregularly, depending on machine and network load, and so could not be reliably controlled programatically. When detected, we excluded the corresponding result page from the subsequent analysis phase as we observed such challenges to be relatively rare in real-world human

interactions.

Search engines are a convenient and openly available source of personalised content, but not the only online service that profiles users. A significant portion of modern online systems profile users to boost commercial return through improved personalised content. In Chapter 6 we extend our analysis to include openly available examples data sources for TripAdvisor and Amazon to illustrate how the techniques we develop can be extended beyond search engines.

1.4 Contributions and Structure of this Thesis

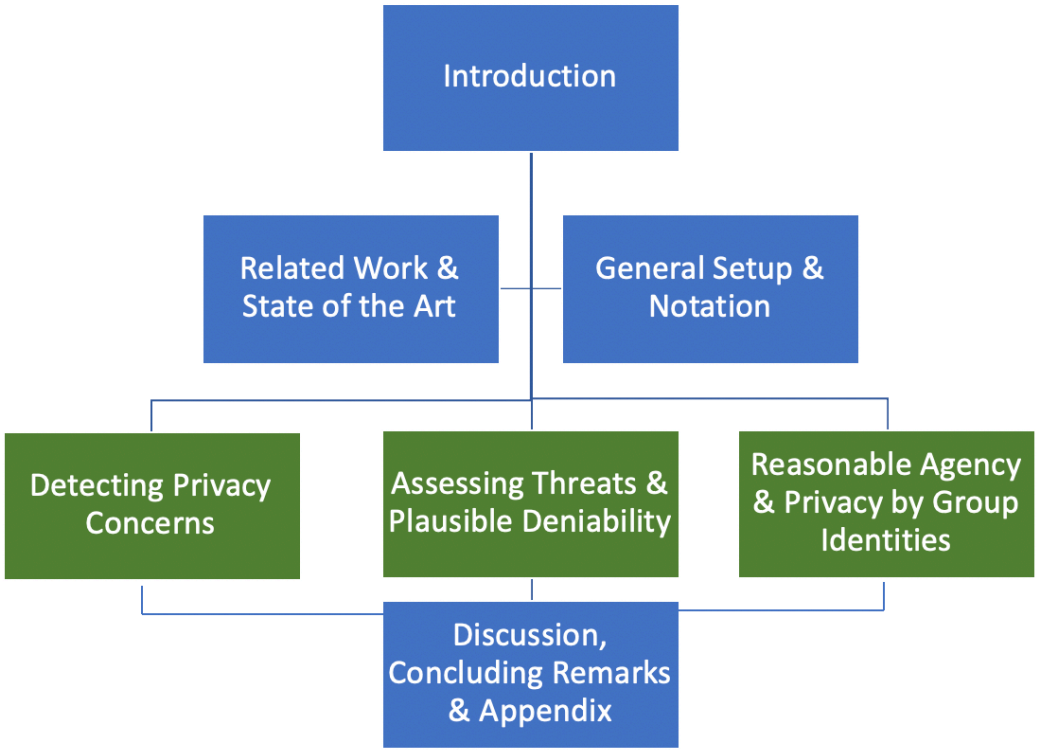


FIGURE 1.3: Structure of this Thesis.

The main technical content in this thesis are contained in Chapters 4-6. Preliminary chapters contain introductory material, a perspective of the contribution of this thesis in the context of related work and state of the art, and a chapter introducing the common notation to be used in the formal sections of Chapters 4-6.

In Chapter 4 we investigate how a user who, knowing what topics they deem to be sensitive, can construct a classifier to detect potential privacy concerns by analysing personalised content appearing on output from a search engine. The formalism we develop identifies possible privacy concerns relative to a baseline level. The baseline is learned from training examples such as historical results. In this chapter we develop the mathematical formalism from which we implement and test a classifier using both Google and Bing as examples of search engines. The main contributions in this chapter are

A novel definition of individual privacy we call ϵ-*Indistinguishability* that is compatible with existing privacy models and readily implementable as a practical user technology

An effective method for detection of privacy threats across a sequence of observations by collecting and comparing responses to a sub-sequence of preselected *probe queries*

A fast, scalable estimator of ϵ-*Indistinguishability*, we call **PRI** ("**PR**ivacy for **I**ndividuals").

An extensive measurement campaign showing that evidence of adaptation is easy to detect for a wide range of sensitive topics.

In Chapter 5 we extend the work in the previous chapter on detecting privacy concerns to address how to assess the degree of threat associated with a detection. The privacy model we use is based on plausible deniability of interest in topics. The formalism we develop is implemented and tested against the Google search engine in this case. Contributions in this chapter include

A formal definition of Plausible Deniability in web search allowing users to test if they can reasonably deny their interest in topics they regard as sensitive.

A tool called **PDE** ("**P**lausible **D**eniability **E**stimator") implementing our formal definition of Plausible Deniability.

Extensive experimental validation that the **PDE** tool is effective in detecting threats to Plausible Deniability even when user queries are obfuscated through injection of high levels of noise.

A novel defence for Plausible Deniability, called the Proxy Topic Defence in this paper, that is observed to provide protection in 100% of tests.

In Chapter 6 we develop a prototype system demonstrating how a search engine might provide privacy preserving services to users with minimal disruption. To show that the prototype can be applied across a broad range of systems we test the prototype using openly available data containing hotel reviews from TripAdvisor and product reviews from Amazon - in addition to testing with Google Search. Contributions in this chapter include

A novel *proxy agent* framework we call *3PS* for **P**rivacy **P**reserving **P**roxy **S**ervice, where a user may protect their interests in sensitive topics from unwanted personalisation by submitting queries though a pool of group identities called *Proxy Agents*.

A formal definition of personalisation utility and privacy detection in a plausible deniability compatible with the 3PS setting. We show that user privacy need not come at the cost of reduced utility in personalised services when aggregated group information represented by the proxy agent pool is sufficient for personalisation.

Privacy preserving algorithm for selecting group membership of proxy agents users can run locally to find the group identity best matching their interests without revealing their interests.

An extensive campaign of experimental verification using openly available datasets to confirm the privacy guarantees provided by 3PS and that our method of selecting group membership is both accurate and converges rapidly.

Material contained in this thesis has been published separately as follows:

Pól Mac Aonghusa et al. (2016). "Don't Let Google Know I'm Lonely". In: *ACM Transactions on Privacy and Security* 19.1, pp. 1–25

P Mac Aonghusa et al. (2018). "Plausible Deniability in Web Search; From Detection to Assessment". In: *IEEE Transactions on Information Forensics and Security* 13.4, pp. 874–887

Pól Mac Aonghusa et al. (2018). "3PS - Online Privacy through Group Identities". In: *Submitted - IEEE Transactions on Information Forensics and Security*

Chapter 2

Related Work

2.1 Data Collection in the Online World

In 1965 the US Government decided to build the first ever "Data Center" to organise over 750 million tax returns, 175 million fingerprints of citizens, 14 million records of civilian security clearance vetting by the Defence Department and 8 million records of people who applied for Government jobs, into a single search-able database, (Alle et al., 1966).

In the intervening years since 1965, our ability to collect individual data has exploded, so that the 1965 project seems modest compared to the scale of today's online data collection. It is estimated that in 2016 as much data was produced as in the *entire* history of humankind up to 2016. By 2026 it is estimated that there will be 150 billion networked measuring sensors – approximately 20 sensors for each person on the planet – and that by 2026 the amount of data generated on the Internet will double every 12 hours, (Helbing et al., 2017).

The financial investment required to sustain the levels technology required to store, organise and process data at this scale has resulted in a situation where a small number of companies can dominate in specific activities. For example, in 2018, over 70% of all web searches worldwide are estimated to have been through Google Search, the nearest competitor, Bing, is estimated to have 7% of web search volumes, (NetApplications, 2018).

There is pressure on commercial companies to evolve increasingly sophisticated data collection capabilities - and to respond when their ability to collect is threatened,

(Sivakorn et al., 2016). In the case of Google, major algorithm changes such as *Caffeine*, (Peng et al., 2010), *Social Search*, (Heymans, 2009) and *Search Plus, Your World*, (Singhal, 2012), included additional sources of background knowledge from Social Media, improved filtering of content such as *Panda*, (Slegg, 2015), to counter spam and content manipulation. In 2018, an estimated 52.2% of all website traffic worldwide came from mobile devices, (Statistica, 2019). Google introduced its "Speed Update" for web search on mobile to all users in July 2018, (Wang et al., 2018), and within days, announced the inclusion of Mobile Landing Page Speed Score in the Google Ads, (Osmani et al., 2018).

Much of the free–to–use Internet is free largely because it is under-pinned by a de facto business model of gathering and analysing data about user interests and behaviours to produce targeted commercial content. Some numbers help illustrate how important targeted commercial content is. Facebook earned an average of US$4.65 per user from personalised content such as advertising and promoted posts in the second quarter of 2017, according to the Economist (Economist, 2017). By comparison, an average of just US$0.08 per user came from direct fees such as payments for games.

We classify data collection as either implicit or explicit for our purposes here. Implicit collection includes data gathered without requiring direct user input. For example, capturing details of the underlying hardware and software by fingerprinting a device. Implicit data collection activities also include activities such as network packet inspection and tracing. Mobile devices in particular provide opportunities for enhancing Geo-location based data collection. Data collected implicitly has been covered extensively in the research literature and is known to be highly revealing of individual behaviours, (Bielova, 2017; Binns et al., 2018; Englehardt et al., 2016; Narayanan et al., 2017). We adopt the approach that implicit data collection is an inevitable consequence of being online. When implicit data collection effects personalisation we assume it manifests through effects observed in personalised output.

Explicit data collection arises through deliberate user actions – such as submitting a search query to Google Search. The importance of explicit data collection is

evidenced by practices such as the use of intrusive pop-ups to force users to consent to accept cookies since the introduction of GDPR, (Burgess, 2018). We take explicit data collection arising from deliberate actions of the user as a controllable aspect of the interaction between user and system. We are interested in understanding how a user can control explicit data collection to quantify and ultimately regulate personal profiling by during interactions with learning systems.

2.2 Privacy and Societal Risks of Profiling

Privacy as a normative concept is deeply rooted in economic, legal and philosophical discussion, (Nissim et al., 2018; Solove, 2006; Swanson, 1992). The literature is vast and, though interesting, is beyond the scope of our focus here on privacy in the context of online data collection and subsequent profiling through machine learning. Our concern is that data collected may be misunderstood, contain errors or be sensitive to an individual. Processing may introduce errors or be less exact than intended. Researchers and practitioners are increasingly warning against the naive usage of internet data collection for profiling, (Olteanu et al., 2018), and documenting the consequences such as adverts connecting socio-economic status with race, (Speicher et al., 2018), or adverts associating criminal behaviour with individual ethnicity, (Sweeney, 2013).

By 1966, the US Government project to collect and process data was dropped because of concerns about invasion of privacy on what, in 1965, was regarded as a ”vast” trove of data, (Alle et al., 1966). Today, a small number of commercial companies dominate data collection and processing on the Internet resulting in so-called “digital oligarchies”, (Andriole, 2017). When personal data is collected without transparent purpose it erodes the ability to define the boundaries between what is, and is not, private. Purposes which, when revealed, can cross the line from “helpful” to “unwanted”. This was highlighted in 2017, when an Australian news website revealed that Facebook offered advertisers the ability to target teenagers suffering “moments of psychological vulnerability”. By monitoring posts, photographs and interactions Facebook was profiling teenagers who felt “worthless, insecure, stressed, defeated,

anxious and like a failure", (Whigham, 2017). Transparency of purpose and processing employed in complex profiling algorithms used by online systems speaks directly to topical concerns with transparency of machine learning in general.

Societal concerns resulting from over-collection and lack of transparency in collection and subsequent processing of data are reflected privacy legislation, such as the EU GDPR (European Union, 2016). GDPR specifically mandates minimisation in the collection of, access to, and transparent processing of data as basic principles of privacy protection. The practical consequences of GDPR are now being felt with overly broad collection and non-transparent data processing risking exposure to legislative action for commercial companies. In January 2019 the first GDPR fine was announced in Portugal, (Monteiro, 2019). The majority of the fine was imposed for breaches of principles of data minimisation and subsequent data processing. Also in January 2019, the French Data Regulator (CNIL) announced a EUR 50M fine against Google under GDPR consent rules, (Ram et al., 2019). In this case, lack of transparency was a principle concern, with CNIL stating that "It is not possible to be aware of the plurality of services, websites and applications involved in these processing operations (Google search, YouTube, Google Home, Google Maps, Playstore, Google Pictures . . .) and therefore of the amount of data processed and combined".

Concerns with potential bias in personalisation by machine learning fall into two broad categories in the literature.

Discrimination Concerns over negative consequences associated with personalisation on Google Search adverts have been identified over several years (Guha et al., 2010; Sweeney, 2013). Our work identifies potential grounds for discrimination in Chapter 5 where we find strong evidence of profiling with respect to health status and sexual orientation. Our findings correspond with recent results identifying potential bias in online advertising by analyzing explanations provided by systems for selection of advertising content, (Andreou et al., 2018; Speicher et al., 2018).

Restriction Restricting access to content via a so-called filter bubble,(Pariser, 2011), and explored in the case of Google Search in (Hannák et al., 2017). In a filter

bubble, a user cannot access subsets of information because the recommender system algorithm has decided it is irrelevant for that user. Recent work has explored algorithmic frameworks to reduce filter bubbles, (Celis et al., 2019), and towards formalizing notions of fairness in machine learning, (Naudts, 2018). Detection and assessment of privacy concerns by analysing changes in content are similar to the techniques used to analyse filter bubbles. Our work here offers, therefore, a potential, additional perspective on filter bubbles for future research.

Privacy concerns can be viewed in terms of two major factors – awareness of a sensitive social situation, and, the ability of an individual to control the social situation, (Boyd, 2012). The importance of reasonable agency or control over appropriate flow of information is discussed extensively in the legal and social science fields. The importance of individual agency over personal information flow was discussed in a critique of the *nothing to hide* defence for widespread surveillance in (Solove, 2007). Individual privacy and its social consequences are discussed in (Bennett, 2011; Boyd, 2012), where agency or control over appropriate disclosure is identified as a key concern. Recent legislation, such as the General Data Protection Regulation (GDPR), requires that *personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which those data are processed*, (European Union, 2016). In this context, broad collection of user data without transparent purpose in online interactions with everyday online systems is a particular concern for individual privacy.

Given concerns with transparency and purpose of data collection, assisting individuals to detect and assess privacy risks is fundamental to protection. Users are concerned about their privacy on the Web but do not always reflect this concern in their online behaviours, (Alessandro Acquisti et al., 2015). In (Pujol et al., 2015), in-the-wild measurements of user interactions with Ad blocking technologies suggest that users overwhelmingly accept default settings and do not install updates such as white-lists. Consequently technologies for user privacy must be effective, but also unobtrusive and simple to maintain. In comparison with users, online systems have

proven alert and adaptable in responding to attempts to protect privacy at individual user level. Stateful (cookie) and stateless (fingerprinting) tracking are widespread on the web. In (Bielova, 2017; Binns et al., 2018; Englehardt et al., 2016; Narayanan et al., 2017) separate studies of 1 million websites reveal widespread data exchange among third parties, stateful tracking from third-party cookie spawning and stateless fingerprint-based tracking. In (Binns et al., 2018) users are observed to be tracked by multiple entities in tandem on the web.

We ascribe to the view that personal privacy requires active engagement from users. In (Ramakrishnan et al., 2001) in concluding remarks, the authors state that "the ideal deterrents are better awareness of the issues and more openness in how systems operate in the marketplace. In particular, individual sites should clearly state the policies and methodologies they employ with recommender systems".

We propose three principles of personal privacy desirable in interactions with online systems, comparable with the "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms" from the ACM FAT/ML website, (FATML, 2019), and intended to be agnostic with respect to implementation choices.

Detection of Privacy Concerns An individual user should be able to detect evidence of unwanted personalisation with respect to topics *they* regard as sensitive. In particular a user should be allowed to define what they regard as sensitive or non-sensitive topics without having to share details of their interests with other parties.

Plausible Deniability of Interests When presented with content regarded as inappropriate or discreditable, a user may wish to deny their interest in the content. A user should be able to assess their ability to plausibly deny interest in sensitive topics they have defined.

Reasonable Agency A user should be able to exercise reasonable agency over privacy choices during interactions. Here the requirement is for *reasonable* control meaning that the expectation of privacy must be appropriate to the context. A person registering with an online dating site must accept a reasonable conclusion is they have *some* interest in dating. Without further information it is not

reasonable to say whether their interest is in academic research or in seeking a date. So that this person can reasonably expect content related to the dating site - but not related to their personal dating preferences.

2.3 Privacy Models and Risk

We consider a setup where a search engine does not seek to identify users as individuals, but rather it seeks to determine likely user interest in topics it deems commercially valuable. We model machine learning of behaviours, interests and preferences as a process of labelling an individual with respect to topic categories. Privacy provides a formal framework to investigate the strength of association between labels and individuals without requiring knowledge of how the association has arisen. When evidence of association between labels and individuals is detected it corresponds to detection of learning in the underlying algorithms employed by the search engine. A privacy concern is *detected* when labels are associated with an individual the individual regards as sensitive. Quantifying the strength of association with sensitive labels corresponds to *assessing* the degree of risk in potential privacy threats. Detection and assessment of privacy concerns are therefore related to inference in machine learning, and more generally to fairness, accountability and transparency in machine learning, (FATML, 2019), and our contributions can be framed within the broader context of verification fairness and analysis of inference in machine learning, (Olhede et al., 2016, 2018).

Our privacy model is based on the notion of plausible deniability. Informally, user activity observed by the search engine exhibits plausible deniability when, with high probability, it is consistent with the user being interested in any one of several topics at least one of which is not sensitive for the user. That is, the patterns of user activity supports reasonable doubt about the user's actual interest in a given sensitive topic. Plausible deniability to counteract the impact of personalisation is examined in (Cummings et al., 2014) for the case of a privacy aware user who knows they are being observed. The authors show that, no matter what the behaviour of the user is, it is always compatible with some concern over privacy. In this way the user

can offer their awareness of privacy concerns as a general alibi to justify any range of preferences. Technologies enabling plausible deniability for web search are addressed in the literature. In (Avi Arampatzis et al., 2013) alternative, less revealing queries are mixed with sensitive topic queries to obfuscate true user interest. In (Arampatzis et al., 2011) queries with generalised terms are used to approximate the search results of a true query, which is never revealed.

We will compare privacy models we use to other privacy models in current use. Two examples we will draw on are k–anonymity and differential privacy. Privacy as a form of hiding in the crowd, where an attacker cannot associate an individual with less than k records in a data set, was first formalised as k-anonymization in (Sweeney, 2000). Since its original introduction a variety of refinements such as l-diversity, (Machanavajjhala et al., 2006) and t-closeness, (Li et al., 2007), have addressed weaknesses with the original definition. Differential privacy, (Dwork, 2006), is a formal framework for privacy preserving statistical queries over databases. Differential privacy has been criticised, (Bambauer et al., 2013), and implementations have been criticised for being opaque, (Tang et al., 2017). Differential privacy has been included in commercial products by Apple, (Apple, 2017) and Google, (Erlingsson et al., 2014).

2.4 Privacy and Web-search Profiling

Mechanisms for privacy protection from web-search profiling have been extensively covered in the literature. An early approach is to obfuscate or mask queries from the system by injecting non-sensitive query terms as "noise" within which to hide or distort sensitive queries. The essential challenge in this type of approach is to define a practical method of selecting "noise" query terms to provide a verifiable level of anonymity while not overly upsetting overall utility,(Domingo-Ferrer et al., 2009; Howe et al., 2009; Peddinti et al., 2011; Sánchez et al., 2013). Query obfuscation and masking is addressed in (Ahmad et al., 2016), where user queries are hidden within a stream of at least k 'cover queries' to provide a form of k-anonymity. PEAS, (Petit et al., 2014, 2015), combines obfuscation and a proxy to also provide unlink-ability

between user and query. In (Ahmad et al., 2016) user queries are hidden within a stream of at least k 'cover queries' to provide a form of k-anonymity. PWS, (Balsa et al., 2012), and TrackMeNot, (Howe et al., 2009; Peddinti et al., 2011), inject distinct noise queries into the stream of true user queries during a user query session, seeking to achieve acceptable privacy while not overly upsetting overall utility.

An alternative approach is to apply encryption and multi-party computation techniques to process sensitive user queries, leveraging techniques from the privacy preserving data mining domain. Protecting users from individual re-identification often combines encryption, hashing and noise addition on the local user machine. A common challenge in this type of approach is that it can be computationally prohibitive and require substantial user management for locally maintained dictionaries of queries, features or URLS accessed by the user. For example, in (Z. Erkin et al., 2011, 2010), the authors propose to encrypt privacy sensitive data and generate recommendations by processing them under encryption. Approaches of this type typically rely on a user, or a learning algorithm, being able to identify which queries are sensitive, and trust in the service provider to perform query processing under secure encryption.

In the recommendation systems literature privacy technologies have largely focused on how to incorporate privacy into the recommendation process itself. In (Batmaz et al., 2016), random perturbation of data is used to develop privacy-preserving frameworks for collaborative filtering methods. In (Boutet et al., 2016), profile obfuscation together with a randomised dissemination protocol are employed. Another approach is to distribute the recommendation process by including a trusted intermediate agent between user and back-end system, such as (Aïmeur et al., 2008). In (McSherry et al., 2009), differential privacy is incorporated into the algorithms used in the Netflix prize competition to produce privacy preserving recommendations. Our work compliments this body of research in that provide technologies that can verify the effectiveness of privacy embedded in the recommendation process by monitoring the outputs.

Raising user awareness has been extensively investigated in the literature. A significant body of research exists to capture user activity and then provide feedback on where information is flowing. Popular browser add-ons, such as Mozilla Lightbeam, (Mozilla, 2016), and PrivacyBadger, (EFF, 2018), facilitate active user awareness of possible privacy and consent issues by helping understand where user data is shared with third parties through the sites they visit. XRay, (Lecuyer et al., 2014), reports high accuracy in identifying which sources of user data such as email or web search history might have triggered particular results from online services such as adverts. Active consensual sharing of personal data is investigated in (Fredrikson et al., 2011) through an in-browser capability, called RePriv, allowing a user to select which portions of their personal data they wish to share with requesters. Data collection on mobile devices is also a concern. Tracking or profiled advertising without consent on the Android platform is addressed in (Razaghpanah et al., 2018), with undocumented services, previously unknown to mainstream advertising and white-listing services constituted over 10% of third party tracking.

Website proxy services offer privacy preserving access to mainstream search engines on the Internet. Two of the better known are DuckDuckGo hosted in the US on Amazon Web Services, (Inc, 2018), and StartPage hosted privately in the Netherlands, (Holding BV, 2018). Functionally both are similar, encrypting traffic via https, and employing POST and re-direct techniques to obfuscate requests. Both claim to relieve so-called filter-bubbles, (Pariser, 2011), by aggregating results from several source systems. In both DuckDuckGo and StartPage the proxy user profile adopted by users of both systems is global. Personalised content such as advertising that is displayed on search result pages is correspondingly generic. The 3PS prototype we develop in Chapter 6 also provides proxy access to web search, but differs in allowing a user to dynamically adopt a group profile that is closest in interest to their personal interests.

In our approach, we regard online systems as black-boxes with unknown internal workings and state. Modelling a system as a black-box is well established in system testing, (Limaye, 2009), and is mentioned in the context of privacy, (Anupam Datta,

2014; Hannák et al., 2017). Grouping users behind intermediate or proxy layers is a well studied privacy technique. Protecting the sensitivity of user data, and particularly of user profiles exposed to the online system, by grouping users behind a proxy layer is defined as *Level 2 Privacy* in the classification scheme of online privacy approaches in (Shen et al., 2007).

Evaluation of privacy technologies in the wild is surprisingly underrepresented in the literature. Our choice of Google Search as a live target for our work was motivated by this imbalance by using a live setup where possible. Where evaluations have been performed, it appears to be largely for direct evaluation of browser privacy plugins, perhaps reflecting the difficulty of performing these experiments where more sophisticated setup is required. Effectiveness of privacy defences in the wild was evaluated by (Peddinti et al., 2011) in the case of *TrackMeNot* where the authors demonstrate that by using only a short-term history of search queries it is possible to break the privacy guarantees of TrackMeNot. In (Ling et al., 2012), the authors demonstrate an effective attack to detect the communication relationships between TOR users. The importance of background information in user profiling is explored in (Petit et al., 2016) where a similarity metric between *known* background information and queries is shown to identify 45.3% of TrackMeNot and 51.6% of GooPIR queries. Anti-tracking is an ongoing area of research and recently in (Pan et al., 2015) an anti-tracking browser called TrackingFree was reported to be effective at disrupting all of the trackers in the Alexa top-500 list. Self-regulation has also proven problematic, in (al, 2012), six different privacy tools, intended to limit advertising due to behavioural profiling, are assessed. The tools assessed implement a variety of tactics including cookie blocking, site blacklisting and Do-Not-Track (DNT) headers. DNT headers were found to be ineffective in tests at protecting against adverts based on user profiling.

We conclude with a cautionary word to the reader. In questions of online privacy, the adage *caveat emptor* ("let the buyer beware") applies. Examples of unsubstantiated and misleading claims of enhanced individual privacy by providers of technology are unfortunately all too common, (Blue, 2016; Day, 2018). Concerns about objective

evaluation of the claims by providers of such technologies have attracted the attention of Government, where the need for *"Awareness and education of the users …"* is identified in (Santa, 2010) as a key step to building trust and acceptance of privacy technologies. Our contribution in this work is deliberately structured as formal followed by experimental. In this way we aim to provide a firm foundation underpinning ensuing experimental results, avoiding confusing claims, albeit at the risk of mathematical density.

Chapter 3

General Setup

3.1 Formal Setup

3.1.1 Black-box Models

We consider a setup where users interact with a system $\mathscr{S}$, such as a search engine, by issuing a query as input and receiving an output in response. Each interaction between a user and $\mathscr{S}$ consists of an input–output pair, referred to as an *input–output interaction*. We gather a sequence of consecutive input-output interactions between a user and $\mathscr{S}$ into a *session*. We sometimes refer to the input–output interactions in a session as *steps*. To improve readability, set operator notation is sometimes used to indicate operations on sequences where there is no scope for confusion.

We assume that user inputs and system outputs are each decomposable into *features*. For example, when modelling a user querying movies or hotels the input features might consist of keywords, or if assigning ratings the features might consist of integers. An ordered list of features with no duplicate entries is called a *dictionary*. We let D^X and D^Y denote the dictionary containing valid input features to $\mathscr{S}$, and valid output features generated by $\mathscr{S}$ respectively. Individual features are indicated thus, θ_i^X, $i = 1, \ldots |D^X|$ and θ_j^Y, $j = 1, \ldots |D^X|$ so that θ_i^X indicates the i^{th} feature in D^X and θ_j^Y the j^{th} feature in D^Y. We let $\mathcal{X}$ and $\mathcal{Y}$ denote the sets of possible valid inputs and outputs comprised of combinations of features from D^X and D^Y respectively, and the set of valid input–output interactions is $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Input–output interactions may repeat during a session and so sessions are represented as *sequences* of input–output interactions.

The system $\mathscr{S}$ is treated as a black-box with internal state unknown to users. Our assumption is that $\mathscr{S}$ uses its internal state, which includes knowledge of user interests, when producing personalised outputs for individual users, thereby potentially revealing something about its internal state. Given a sequence of user inputs we observe corresponding system outputs and try to spot evidence of learning of topics the user considers private.

The Basic Black-box Model for user–system interaction is illustrated in Figure 3.1.

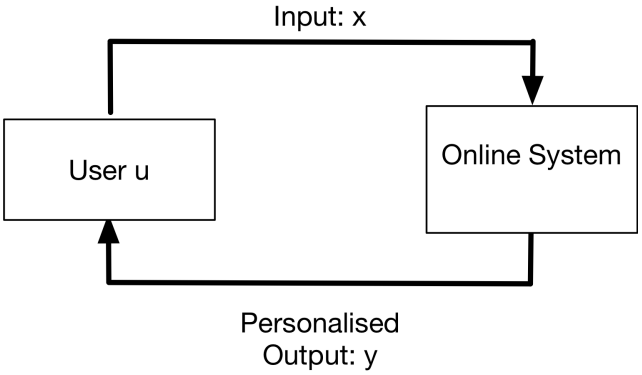


FIGURE 3.1: Overview of the Basic Black-box Model

Definition 1 (Basic Black-box Model) *The Basic Black-box Model consists of two interacting components* $\{\mathscr{U}, \mathscr{S}\}$

- *An* online system $\mathscr{S}$ *for which only inputs to, and outputs from,* $\mathscr{S}$ *are observable to users, while details of the internal workings of* $\mathscr{S}$ *are hidden.*
- *A set* $\mathscr{U}$ *of users who can submit input to, and receive corresponding output responses from,* $\mathscr{S}$*.*

We also define a *Proxy Black-box Model* by extending the Basic Black-box Model whereby users access the system through a pool of group identities referred to as *proxy agents*. This is illustrated schematically in Figure 3.2.

Definition 2 (Proxy Black-box Model) *The Proxy Black-box Model consists of three interacting parties denoted* $\{\mathscr{U}, \mathscr{P}, \mathscr{S}\}$ *as follows:*

- *An online* system $\mathscr{S}$ *for which only inputs to, and outputs from,* $\mathscr{S}$ *are observable while details of the internal workings of* $\mathscr{S}$ *are hidden.*
- *A set* $\mathscr{P}$ *of* Proxy Agents*. Proxy agents function as* Group Identities*, routing user queries to, and output responses from* $\mathscr{S}$*.* $\mathscr{P}$ *is sometimes referred to as the* Proxy Agent Pool*.*
- *A set* $\mathscr{U}$ *of* users *can submit input to, and receive corresponding output responses from,* $\mathscr{S}$ *via the group identities provided by the proxy agents in* $\mathscr{P}$*.*

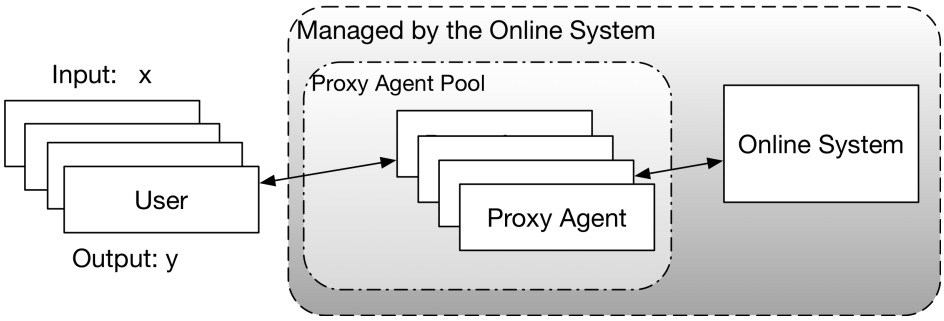


FIGURE 3.2: Overview of the Proxy Black-box Model

In the Proxy Black-box Model the proxy agent pool is assumed controlled by the back-end service. One key reason for doing this is to ensure that proxy agents are recognised as genuine users by the back-end system. If not recognised as bona fide users the proxy agents may be flagged as a bot or robot and so trigger defences, such as "captchas", or even be blocked. Other than acknowledging the proxy agents as legitimate users, the Proxy Black-box Model is intended to be backwards compatible not requiring significant engineering changes in the back-end system.

3.1.2 Modelling Interactions

It is convenient to discretise the time variable and a non-negative integer $k \in \{1, 2, \ldots\}$ is used to track input–output interactions in both the Basic Black-box Model and Proxy Black-box Model and refer to input–output interactions at *step* k. The sequence of input–output interactions generated by the users in $\mathscr{U}$ to step k is denoted

$$\mathcal{Z}_k = (z_1, \ldots z_k) \cup \mathcal{Z}_0 \tag{3.1}$$

where z_1 is the first input–output interaction between users in $\mathscr{U}$ and $\mathscr{S}$ during the session and z_k is the k^{th} input–output interaction between users and $\mathscr{S}$ for $k \geq 1$. The quantity $\mathcal{Z}_0$ denotes prior background knowledge gathered from $\mathscr{U}$ at the beginning of the session.

For an individual user labelled with $u \in \mathscr{U}$ the sequence of input–output interactions between u and $\mathscr{S}$ is denoted by

$$\mathcal{Z}_{u,k} = (z_{u,1}, \ldots z_{u,k}) \cup \mathcal{Z}_{u,0} \tag{3.2}$$

where $\mathcal{Z}_{u,0}$ is the background knowledge available about user u at the beginning of the session.

The sequence of input-output interactions submitted to a proxy agent labelled by $p \in \mathscr{P}$ from the set of users in $\mathscr{U}$ is denoted $\mathcal{Z}_{p,k}$.

$$\mathcal{Z}_{p,k} = (z_{p,1}, \ldots z_{p,k}) \cup \mathcal{Z}_{p,0} \tag{3.3}$$

where $\mathcal{Z}_{p,0}$ is the background knowledge available to proxy agent p at the beginning of the session.

Let $\mathscr{P}_u \subset \mathscr{P}$ denote the subset of proxy agents used by user u up to step k. The input-output interactions submitted by user u through proxy agent $p \in \mathscr{P}_u$ to step k are contained in the sub-sequence

$$\mathcal{Z}_{u,p,k} = (z \in \mathcal{Z}_{u,k} : \iota_p(z) = 1) \tag{3.4}$$

where indicator function ι_p equals 1 for input-output pairs submitted via proxy p and 0 otherwise.

3.1.3 Topic Labelling

Each user u has a private *labelling function* $l_u : \mathcal{Z} \longrightarrow \mathcal{C}$ which associates input-output interactions in $\mathcal{Z}$ with topic labels selected from a private, user-defined set of labels $\mathcal{C} = \{c_0, c_1, \ldots c_K\}$. Each user can have different labels so that $\mathcal{C}$ can change by user. When necessary we will write $\mathcal{C}_u$ when we need to emphasise which user has chosen a particular set of topic. We omit the subscript u from $\mathcal{C}_u$ otherwise for notational clarity.

We adopt the convention that the label c_0 is identified with a catch-all "non-private" category while the remaining elements in $\mathcal{C} \setminus \{c_0\}$ label individual "sensitive"

topics such as "health" or "finances". The user labelling function l_u is itself private and labels every input–output interaction in $\mathcal{Z}_{u,k}$, with at least one topic from $\mathcal{C}$. Often we are simply interested in whether an input-output interaction is private or not for a user, and so we define the indicator function $I_u : \mathcal{Z} \longrightarrow \{0,1\}$ with $I_u(z) = 1$ when $l_u(z) = c$, $c \in \mathcal{C} \setminus \{c_0\}$, i.e. when an input–output interaction is labelled with a private topic by user u, and $I_u(z) = 0$ otherwise. We assume that user labelling functions are well-behaved in the following sense.

Assumption 1 (Meaningful Labelling) *An input-output interaction which is labelled as non-private by a user is truly non-private for that user e.g. the user would be content for it to be shared publicly.*

Assumption 1 requires users to strike their own balance between utility and privacy as discussed in the third principle introduced in Section 1. The low risk strategy of simply labelling every input-output interaction as private implies that the user may not be able to use the system at all. For example, if the system is a dating service, the knowledge that a person uses the system necessarily reveals their interest in such a service. A user choosing to use the system cannot include such system-level topics in their set of private topics. The implicit statement in Assumption 1 is that users form an individual judgement regarding the inference capabilities of observers and to accept a degree of risk associated with this judgement call proving incorrect.

The sequence $\mathcal{Z}_{u,k}^{c} := (z \in \mathcal{Z}_{u,k} : l_u(z) = c)$ denotes the sub-sequence of input–output interactions the user labels with topic $c \in \mathcal{C}$, and, $\mathcal{Z}_{u,k}^{sens} := \bigcup_{c \in \mathcal{C} \setminus \{c_0\}} \mathcal{Z}_{u,k}^{c}$ denotes the sub-sequence of input–output interactions labelled as private.

Let $\mathcal{Z}_{k}^{u,c} := (z \in \mathcal{Z}_k : l_u(z) = c)$ and $\mathcal{Z}_{k}^{u,sens} := \bigcup_{c \in \mathcal{C} \setminus \{c_0\}} \mathcal{Z}_{k}^{u,c}$. The sequence $\mathcal{Z}_{k}^{u,sens}$ contains items from users other than u. Consequently, while $\mathcal{Z}_{u,k}^{u,sens} \subseteq \mathcal{Z}_{k}^{u,sens}$, it is not generally the case that $\mathcal{Z}_{k}^{u,sens}$ is a sub-sequence of $\mathcal{Z}_{u,k}$.

When a user decides a topic is private it is a personal decision. We want to capture the flexibility a user has to choose their own private topics while avoiding subjective words like "embarrassing" or "awkward" in assigning topics as private. We will, however, use the term "sensitive topic" to describe topics that the user regards as private. Our intention is to avoid having to repeat phrases like "topics the user

regards as sensitive" in favour of the simpler term "sensitive topic". Describing a topic as a sensitive topic should be interpreted accordingly, rather than attaching any subjective or emotional meaning to the term "sensitive". We will also use terms like "sensitive query" to indicate a query about a topic labelled as private by a user.

3.1.4 Bag–of–Words Text Model

We take inputs and outputs to be comprised of text from a natural language, such as English, and provide some brief background on tools from Natural Language Processing we will use later. For a full treatment, using the Python language, see (Bird et al., 2009).

The first preprocessing step is to extract personalised text appearing on a result page as a single block of text and tokenise it into individual words by using white-spaces and punctuation as token separators. Common, uninformative, high-frequency stop-words are removed and stemming is performed to remove common prefixes and suffixes. The result is a sequence of tokens or fragments of words occurring in the original text. We will sometimes refer to tokens as *keywords* or *features*. Tokens extracted from the input are gathered into a dictionary, denoted $D^X := (\theta_1^X, \theta_2^X, \dots)$ where θ_j^X is the j^{th} token in D^X. Tokens extracted from the output are added to a dictionary $D^Y := (\theta_1^Y, \theta_2^Y, \dots)$ and θ_j^Y denotes the j^{th} token in D^Y.

We adopt a standard bag–of–words language model (Weikum, 2002) where features in an input–output interaction are modelled as being drawn independently with replacement and ignoring order according to the mixture model, (Hofmann et al., 1998),

$$\begin{aligned} &\mathrm{P}(z \in \mathcal{A}_k | z \in \mathcal{B}_k) \\ &= \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} \mathrm{P}(z \in \mathcal{A}_k | \{\theta_i^X, \theta_j^Y\} \in z) \mathrm{P}(\{\theta_i^X, \theta_j^Y\} \in z | z \in \mathcal{B}_k) \end{aligned} \tag{3.5}$$

where $\mathcal{A}_k$ and $\mathcal{B}_k$ are sequences of input–output observations.

In the bag–of–words model, an input x is mapped to a count-vector $\phi^X(x) \in \mathbb{Z}^{|D^X|}$ where the i^{th} component $\phi_i^X(x)$ of $\phi^X(x)$ is equal to the number of occurrences of i'th

feature θ_i^X in input x. Similarly, output y returned by $\mathscr{S}$ in response is mapped to count-vector $\phi^Y(y) \in \mathbb{Z}^{|D^Y|}$. An input–output interaction $z = (x, y)$ is mapped to count-vector $(\phi^X(x), \phi^Y(y))$.

3.2 Experimental Setup

3.2.1 General Setup

In this section we describe the common experimental setup with Google Search. Google Search is used as the principal source of data experiments. We also describe supplementary data sources we make use of occasionally. In Chapter 4 we also report measurements and experimental results taken from Bing Search. We use an identical setup for data collection from both Google and Bing Search for direct comparison purposes. The intention is to provide additional context in Chapter 4 to illustrate concepts and confirm experimental findings. In Chapter 6 we use supplementary data to show how the techniques there can be extended to systems other than search engines. We describe these supplementary data sources in more detail in Section 3.2.3.

Data was collected using Linux virtual machines located in a University domain supporting several thousand users. Custom scripts were written to automate query execution and response collection. These scripts used Python, BeautifulSoup, (Richardson, 2016), for HTML processing and Phantomjs, (Friesel, 2014) for browser automation. The Python SciKit toolkit, (Pedregosa et al., 2012), was used for text preprocessing. Numeric processing was performed using the NumPy numerical processing toolkit, (Idris, 2012; Oliphant, 2006).

3.2.2 Web Search Assigning Topic Categories and Queries

For web search, we select twelve user interest categories to study, detailed in Table 3.1. Of the eleven sensitive topics, (i) ten are sensitive categories associated with subjects generally identified as causes of discrimination (medical condition, sexual orientation *etc*) or sensitive personal conditions (gambling addition, financial problems *etc*),

see for example (Equal Opportunity Commission, 2014; European Union, 2010) (ii) a further sensitive topic is related to "London" as a specific destination location, providing an obviously interesting yet potentially sensitive topic that a search engine might track, (iii) the last topic is a non-sensitive category, labeled *other*, which is based on the top-50 queries taken from Google Trends (Google Trends, 2018), providing the catch-all *other* topic representing topics that are not sensitive. The queries selected from Google Trends for the non-sensitive topic do not contain terms appearing in any of the sensitive topic queries.

Category	**Keywords**
anorexia	nerves eating disorder body image binge diet weight lose fat
bankrupt	bankrupt insolvent bad credit poor credit clear your debts insolvency payday insolvent any purpose quick cash benefits low income
diabetes	diabetes mellitus hyperglycaemia blood sugar insulin resistance
disabled	disabled special needs accessibility wheelchair
divorce	divorce separation family law
gambling addiction	uncontrollable addiction compulsive dependency problem support counselling advice therapist therapy help treatment therapeutic recovery anonymous
gay (homosexuality)	gay queer lesbian homosexual bisexual transgender LGBT dyke queen homo
location (London)	london england uk
payday loan	default unsecured debt consolidate advice payday cheap
prostate cancer	prostate cancer PSA male urethra urination
unemployed	job seeker recruit search position cv work employment
other	Select the top-50 queries on Google Trends as examples of non-sensitive queries, excluding terms appearing in sensitive topics.

TABLE 3.1: Categories and associated keyword terms

For each category, apart from *other*, a keyword list is created by extracting associated terms from curated sources including Wikipedia (common terms co-occurring on the category page) and Open Directory Project (pages and sub-topics associated with a category). These are detailed in Table 3.1. Candidate search queries are then generated for each category by drawing groups of one or more keywords uniformly

at random with replacement from the keyword lists. These candidate queries are manually augmented with common words (and, of *etc*) to yield queries resembling the English language. In this way a keyword such as fat, for example, might be transformed into a query "why am i so fat". Non-sensical or overly robotic queries are removed by manual inspection. For the *other* category, queries are taken from the top-50 on Google Trends.

[01] ! keywords: london england uk	[16] ! wait 4
[02] ! probe: help and advice	[17] help and advice
[03] help and advice	[18] ! wait 7
[04] ! wait 7	[19] things to do london next week
[05] weather forecast for london	[20] ! wait 5
[06] ! wait 5	[21] regents park hotels
[07] find hotels in london city	[22] ! wait 7
[08] ! wait 3	[23] get cheap london show tickets
[09] help and advice	[24] ! wait 7
[10] ! wait 7	[25] shows on london now
[11] cheap hotels in london	[26] ! wait 5
[12] ! wait 10	[27] tickets london shows
[13] hotels in regents park cheap	[28] ! wait 7
[14] ! wait 7	[29] help and advice
[15] marriott courtyard regents park	

TABLE 3.2: Example query script. Numbers in square brackets indicate line numbers for readability. The command `!wait n` instructs the Python script to wait n seconds. The script is run sequentially and is split into two columns here to save space.

According to (Lioma et al., 2018) a search session consists on average of three query submissions. To construct sequences of queries for use in user sessions, a predefined probe query is inserted at intervals of $1-5$ queries so that there are 3 topic-specific queries on average between instances of a probe query. In this way we obtain twelve "scripts" of queries. Each script consists of between $25-40$ queries

including the inserted probe queries. A user session then consists of a single iteration of a single script run from beginning to end. An example script is shown in Table 3.2.

3.2.3 Supplementary Data Sources

In Chapter 6, in addition to using experimental data from Google Search, the following experimental data from supplementary real-world sources are used in experiments.

Hotels Tripadvisor hotel reviews containing hotel review titles, review bodies and lowest price per room downloaded from, (Hongning Wang et al., n.d.), and consisting of over 1.6 million hotel reviews. Queries consisting of words extracted from review titles are used as inputs and detailed review bodies represent outputs.

Products Product review titles, review bodies and overall rating scores downloaded from, (Hongning Wang et al., n.d.), containing Amazon product reviews for 6 types of merchandise and consisting of over 2.2 million product reviews. Words appearing in product review titles are used as query inputs and outputs review bodies.

Default topics for experiments were defined as follows from each of the supplementary experimental data-sets.

Hotels Five topic categories are defined by dividing the *lowest price per room* into equally spaced ranged, namely $0 := [0, 110], 1 := (110, 220], 2 := (220, 330], 3 := (330, 440], 4 := (440, 550], 5 := (550, \infty)$. Reviews are then labeled according to the lowest price.

Products The *overall rating score* is used to define topic categories, namely very dissatisfied (Topic 1) to very satisfied (Topic 4). Topic 0 is used to indicate no rating was given so there are 5 topic categories in total.

When experiments are performed requiring a larger number of topics than those above, the Hotels data-set is divided into a larger number of topic categories by specifying different lowest price ranges. In this way it is possible to create a variety of topic categories automatically by re-grouping the data into finer price categories to create more topic categories. The Hotel data-set was chosen for convenience since the

categories are defined by numeric, price-per-room, ranges and so it is straightforward to programatically define more categories by changing the numeric ranges.

Chapter 4

Detecting Privacy Concerns

4.1 Introduction

How far should we accept promises of privacy in the face of personalised profiling? In particular, we ask how can we improve detection of sensitive topic profiling by web search engines.

We consider the Basic Black-box Model described in Section 3.1.1 with users directly accessing a search engine $\mathscr{S}$ such as Google or Bing. Inputs are web-search queries submitted to the search engine and outputs are the corresponding response pages containing several components including personalised adverts. We investigate how a user who, having a number of sensitive topics they wish to keep private, can detect potential privacy concerns by analysing personalised content appearing on output from a search engine. The formalism we develop identifies possible privacy concerns relative to a baseline level. The baseline is learned from training examples such as historical results.

We begin by developing a formal privacy model for a novel definition of individual privacy we call ϵ-*Indistinguishability* that is compatible with existing privacy models. To show ϵ-*Indistinguishability* is readily implementable as a practical user technology we implement a fast, scalable estimator of ϵ-*Indistinguishability*, we call **PRI** ("**PR**ivacy for **I**ndividuals"). The **PRI** estimator looks for changes relative to a baseline to detect privacy threats. Consequently, we introduce an effective method for detection of privacy threats across a sequence of observations by collecting and comparing responses to a sub-sequence of preselected *probe queries*.

We end this chapter by showing that evidence of adaptation is easy to detect for a wide range of sensitive topics in an extensive measurement campaign. Google is the main search engine we consider, however we also show results using Bing for comparison in this Chapter to illustrate that our techniques are applicable more broadly than to a single search engine.

4.2 Privacy Model

We adopt an indistinguishability definition of disclosure risk, tailored to the context of the Basic Black-box Model:

Definition 3 (ϵ-*Indistinguishability*) *ϵ-Indistinguishability is satisfied by a user session $\mathcal{Z}_{u,k}$ with respect to a sensitive topic $c \in \mathcal{C}$, if there exists a privacy parameter $\epsilon > 0$ such that*

$$e^{-\epsilon} \leq M_{u,k}(c) \leq e^{\epsilon} \tag{4.1}$$

where

$$M_{u,k}(c) := \frac{\mathrm{P}(I_u^{u,c} = 1 | \mathcal{Z}_{u,k})}{\mathrm{P}(I_u^{u,c} = 1 | \mathcal{Z}_{u,0})} \tag{4.2}$$

where $I_u^{u,c}$ is an indicator random variable with value 1 when u is interested in topic c and 0 otherwise So that the evidence available for interest in sensitive topic c at step k having knowledge of the full history $\mathcal{Z}_{u,k}$ cannot differ from the evidence at the beginning of the session by more than an amount determined by ϵ.

Given the sequence of observations $\mathcal{Z}_{u,k}$ associated with a user, our aim is to (1) estimate whether ϵ-*Indistinguishability* has been violated for any of the sensitive categories in $\mathcal{C}$, and (2) identify which of these sensitive categories have been learned, with high probability in both cases.

4.3 Using Probe Queries to Simplify Estimation

Estimating $M_{u,k}(c)$ is challenging since it depends on the full user session history up to step k. To simplify the task we assume that the user issues a pre-defined probe query at intervals during the session. In brief, a probe query should be plausible in relation to a sensitive topic so that it does not suggest a change of topic to the search engine; a probe query should also be ambiguous so that the search engine has several possible adaptations to the probe query. In Section 4.6.1 experimental probe query

selection is discussed, where selecting high-frequency terms appearing on multiple result pages, while taking care to avoid obviously revealing terms, is shown to be a practical method of probe selection. In practice, a probe query might be issued in an automated manner by the user's browser and the response processed in the background so as not to disturb the user.

We make the following assumptions when using probe queries.

Assumption 2 (Sufficiently Informative Responses) *At each step k at which a probe query is issued*

$$\frac{\mathrm{P}(I_u^{u,c}=1|\mathcal{Z}_{u,k})}{\mathrm{P}(I_u^{u,c}=1|\mathcal{Z}_{u,0})}=\frac{\mathrm{P}(I_u^{u,c}=1|Z_k=z_k,\mathcal{Z}_{u,0})}{\mathrm{P}(I_u^{u,c}=1|\mathcal{Z}_{u,0})} \tag{4.3}$$

where $Z_k=z_k$ denotes the event that input–output interaction z_k is observed at step k of the session. So that it is not necessary to explicitly use the full search history up to step k as background knowledge during the current session when estimating $M_{u,k}(c)$ as the current session history is fully reflected in the response to the probe query at step k.

Assumption 2 greatly simplifies estimation as it means we do not have to take account of the full search history, but requires that the response to the probe query reveals any search engine learning of interest in sensitive category c which has occurred. Methods for the selection of an appropriate probe query that tends to elicit revealing responses are discussed in detail in Section 4.6.1.

Our next assumption follows from the observation in Section 4.1 that a commercial system, such as a search engine, is obliged to use information about user interests when selecting which adverts to display. For an input–output observation $z:=(x,y)\in\mathcal{Z}_{u,t}$ let $\omega(z)=a$ where $a\in y$ is the advert content on the output y. Let $\mathcal{A}_{u,k}:=(\omega(z):z\in\mathcal{Z}_{u,k})$ denote the sequence of advert content appearing on input–output interactions in $\mathcal{Z}_{u,k}$. Let $\mathcal{A}_{u,k}^{u,c}:=(\omega(z):z\in\mathcal{Z}_{u,k}^{u,c})$ denote the sequence of advert content appearing on input–output interactions in $\mathcal{Z}_{u,k}^{u,c}$. Let $\mathcal{A}_{u,0}:=(\omega(z):z\in\mathcal{Z}_{u,0})$ be the sequence of advert content in the background knowledge $\mathcal{Z}_{u,0}$. Let $\mathcal{A}_{u,0}^{u,c}:=(\omega(z):z\in\mathcal{Z}_{u,0}^{u,c})$ denote the sequence of advert content appearing on input–output

interactions in $\mathcal{Z}_{u,0}^{u,c}$. We assume that the labelling function l_u is consistent with the function ω in the sense that $l_u(z) = l_u(\omega(z)) = l_u(a)$, with the obvious abuse of notation where there is no scope for confusion, so that advert content appearing in input–output interactions can be labelled using l_u.

Assumption 3 (Revealing Adverts) *In the response by $\mathscr{S}$ to the probe query at step k it is the adverts on the response page at step k which primarily reveal learning of sensitive categories by $\mathscr{S}$. Therefore, since the probe query input is fixed,*

$$M_{u,k}(c) := \frac{\mathrm{P}(I_u^{u,c} = 1 | \Omega_k = a_k, \mathcal{A}_{u,0})}{\mathrm{P}(I_u^{u,c} = 1 | \mathcal{A}_{u,0})} \tag{4.4}$$

where $\Omega_k = a_k$ denotes the event that advert content a_k is observed in response at step k. Consequently, only the advert content on the output in response to a probe query needs to be analysed.

When $M_{u,k}(c) > e^{\epsilon}$ or $M_{u,k}(c) < e^{-\epsilon}$ for *any* $k \in \mathcal{K}$ then ϵ-*Indistinguishability* is violated. To ensure that the converse holds, namely that when $e^{-\epsilon} \leq M_{u,k}(c) \leq e^{\epsilon}$ for all $k \in \mathcal{K}$ so that ϵ-*Indistinguishability* is satisfied, we also need the following assumption.

Assumption 4 (Sufficiency of Sampling) *When $e^{-\epsilon} \leq M_{u,k}(c) \leq e^{\epsilon}$ for the subset of probe query steps $\mathcal{K}$ in a session then $e^{-\epsilon} \leq M_{u,k}(c) \leq e^{\epsilon}$ for every step $k \in \{1, 2, \cdots\}$ in that session. That is, when ϵ-Indistinguishability is satisfied at the sub-sequence of steps $\mathcal{K}$ at which the probe query is issued then it is satisfied at all steps and ϵ-Indistinguishability holds.*

In practice it can be difficult to verify whether Assumption 4 holds or not. When we cannot rely on Assumption 4 then, as already noted, violations where $M_{u,k}(c) > e^{\epsilon}$ or $M_{u,k}(c) < e^{-\epsilon}$ for $k \in \mathcal{K}$ are still informative of disclosure risk, and so measurements taken at $k \in \mathcal{K}$ should be regarded as an underestimate, or lower bound, of disclosure risk for the user.

4.4 Bayesian Estimator

Empirical estimators for quantities in (4.7) can be defined in the following way. Assume the availability of a training data set $\mathcal{T}$ consisting of labelled advert content from input–output interactions. Approximate the prior evidence at the beginning of the query session empirically with $\mathcal{T}$ - that is $\hat{\mathcal{Z}}_0 = \mathcal{T}$. By applying the natural language processing techniques described in Section 3.1.4 to $\mathcal{T}$ we produce a dictionary of advert keywords D^A. From the definition of ϵ-*Indistinguishability* in (4.4):

$$M_{u,k}(c) := \frac{\mathrm{P}(I_u^{u,c} = 1 | \Omega_k = a_k, \mathcal{A}_{u,0})}{\mathrm{P}(I_u^{u,c} = 1 | \mathcal{A}_{u,0})} \tag{4.5}$$

$$\overset{(a)}{=} \sum_{j=1}^{|D^A|} \frac{\mathrm{P}(I_u^{u,c} = 1 \,|\, \theta_j^A \in a_k, \mathcal{A}_{u,0}) \mathrm{P}(\theta_j^A \in a_k \,|\, \Omega_k = a_k, \mathcal{A}_{u,0})}{\mathrm{P}(I_u^{u,c} = 1 | \mathcal{A}_{u,0})} \tag{4.6}$$

$$\overset{(b)}{=} \sum_{j=1}^{|D^A|} \frac{\mathrm{P}(\theta_j^A \in a_k \,|\, I_u^{u,c} = 1, \mathcal{A}_{u,0}) \mathrm{P}(\theta_j^A \in a_k \,|\, \Omega_k = a_k, \mathcal{A}_{u,0})}{\mathrm{P}(\theta_j^A \in a_k | \mathcal{A}_{u,0})} \tag{4.7}$$

where equality (a) follows from applying the discrete bag–of–words model, (3.5), using the output features in the dictionary D^A since we consider adverts contained in outputs only in the case of a probe query by Assumption 3. Equality (b) follows from Bayes' Theorem.

With D^A the block of advert content a appearing on a result page is mapped to its count-vectorised form $\phi^Y(a)$. The i'th component of the count-vectorised form, $\phi_i^Y(a)$, is equal to the number of occurrences of the keyword feature $\theta_i^A \in D^A$ in the advert output a. We apply regular Laplace Smoothing to the count-vectorised form $\phi^Y(a)$, (Manning et al., 2008), to avoid divide by zero under-flows in subsequent computations when there are sparse occurrences of keywords in a training sequence. Laplace smoothing resolves this problem by adding a factor $\lambda > 0$ to each keyword count so that $\phi_i^Y(a) \longrightarrow \phi_i^Y(a) + \lambda$.

Let $n_i(a)$ denote the frequency with which keyword θ_i^A occurs in count-vectorised advert output a. That is,

$$n_i(a) = \frac{\phi_i^Y(a)}{\sum_{j=1}^{|D^A|} \phi_j^Y(a)} \tag{4.8}$$

and we can define estimators for the quantities in (4.7) as follows

$$\widehat{\mathrm{P}}(\theta_j^A \in a_k \,|\, \Omega_k = a_k, \mathcal{A}_{u,0}) = n_j(a_k) \tag{4.9}$$

$$\widehat{\mathrm{P}}(\theta_j^A \in a_k \,|\, I_u^{u,c} = 1, \mathcal{A}_{u,0}) = \frac{\sum_{a \in \mathcal{T}^{u,c}} n_j(a)}{N^{\mathcal{T}^c}}, \qquad N^{\mathcal{T}^c} = \sum_{j=1}^{|D^A|} \sum_{a \in \mathcal{T}^{u,c}} n_j(a) \tag{4.10}$$

$$\widehat{\mathrm{P}}(\theta_j^A \in a_k | \mathcal{A}_{u,0}) = \frac{\sum_{a \in \mathcal{T}} n_j(a)}{N^{\mathcal{T}}}, \qquad N^{\mathcal{T}} = \sum_{j=1}^{|D^A|} \sum_{a \in \mathcal{T}} n_j(a) \tag{4.11}$$

where $\mathcal{T}^{u,c} \subseteq \mathcal{T}$ is the subset of training data labelled with topic $c \in \mathcal{C}$ by the labelling function l_u. Combining these estimators with (4.7) results in the following estimator for $M_{u,k}(c)$:

$$\widehat{\mathbb{M}}_{u,k}(c) = \frac{N^{\mathcal{T}}}{N^{\mathcal{T}^c}} \sum_{j=1}^{|D^A|} \left(\frac{\sum_{a \in \mathcal{T}^{u,c}} n_j(a)}{\sum_{a \in \mathcal{T}} n_j(a)} n_j(a_k) \right) \tag{4.12}$$

where a_k is the advert content from a probe query at step k.

We refer to the expression for $\widehat{\mathbb{M}}_{u,k}(c)$ as the **PRI** estimator.

4.5 Example

θ_i^A	$\widehat{\mathrm{P}}(\theta_j^A \in a \,\|\, a \in \mathcal{A}_{u,0}, \mathcal{A}_{u,0})$	$\widehat{\mathrm{P}}(\theta_j^A \in a \,\|\, a \in \mathcal{A}_{u,0}^{u,c}, \mathcal{A}_{u,0})$	
		$c = prostate$	$\bar{c} = other$
prostat, cancer	$\frac{5}{12}$	$\frac{5}{12}$	0
possibl, learn, here	$\frac{1}{6}$	$\frac{1}{6}$	0
treat, suffer	$\frac{5}{12}$	$\frac{1}{4}$	$\frac{1}{6}$
risk	$\frac{5}{12}$	$\frac{1}{6}$	$\frac{1}{4}$
revers, natur, lifetim	$\frac{1}{6}$	0	$\frac{1}{6}$

TABLE 4.1: Illustrative example estimator values.

Consider the following illustrative example. Let $\mathcal{C} = \{prostate\}$ (*i.e.* we have a single sensitive category), label non-sensitive category $\bar{c}$ as *other* and suppose the

training set (after text pre-processing) is,

$$\begin{aligned}\mathcal{T} = \{&(\textit{prostate},\ \{\text{prostat cancer possibl risk learn here}\}),\\ &(\textit{prostate},\ \{\text{prostat cancer suffer treat}\}),\\ &(\textit{other},\ \{\text{diabet treatment suffer discov revers natur}\}),\\ &(\textit{other},\ \{\text{discov lifetim risk diabet}\})\}\end{aligned}$$

Dictionary D^A therefore consists of the terms {prostat, cancer, diabet, discov, possibl, learn, here, treat, risk, suffer, revers, natur, lifetim}. Values of the associated probability estimators are given in Table 4.1.

An advert with text terms (after filtering)

$$y = \{\text{patient choos safer treat here}\}$$

is observed. Since the terms patient, choos, safer do not appear in the training data set – only the terms treat, here contribute to $\widehat{\mathbb{M}}_{u,k}(c)$. We have $n_i(y) = \frac{1}{5}$ for $\theta_i^A \in \{\text{treat, here}\}$ and so $\widehat{\mathbb{M}}_{u,k}(c) = \frac{8}{25} = 0.32$ for $c = \textit{prostate}$. For comparison, $\widehat{\mathbb{M}}_{u,k}(c) = \frac{2}{25} = 0.08$ for $c = \textit{other}$. The advert in this example is in fact taken from the Google result page for a probe query during a session where the user is carrying out searches related to prostate cancer. The high value for $\widehat{\mathbb{M}}_{u,k}(c)$ when $c = \textit{prostate}$ is therefore as expected.

4.6 Experimental Setup

4.6.1 Selecting Informative Probe Queries

Pre-defined "Probe Queries" are issued during a query session as a way to gather responses from the search engine for comparison. The first query in any session is always a probe query so that we have a baseline for comparison. Responses to subsequent instances of a probe query are then compared to the responses obtained from the initial, baseline, probe query to look for changes. A probe query is required to be sufficiently informative that it could reveal adaptation in the user-search engine

interaction (Assumption 1), but should not overly disturb the search engines responses to user queries (so as to preserve the utility of the search engine for the user). To meet these requirements we propose that a good probe query should possess the following general characteristics:

Ambiguity It should be meaningful with respect to the sensitive topic but allow more than one interpretation, so allowing the search engine to choose from a variety of plausible topics.

Consistency It should be consistent with the user's information requirement so as not to disturb search engine learning. The probe should not "surprise" the search engine.

Candidate probe query keywords were identified as follows. Each of the scripts in Table 3.1 was executed three times, without probe queries, and collecting the response pages. We filtered the text in the response pages by stemming terms and removing stop-words. Next term frequency analysis of the filtered terms was performed and the top 10 terms identified, see Table 4.2. For comparison, we also report the same results for Bing Search.

It can be seen that the top-4 words appearing in both Google and Bing search results are {help, advice, symptom, cause} and that these are significantly more frequent than lower ranked terms. Additionally these terms are in the top-5 for each of Google and Bing individually. We use these keywords to form two probe queries: "symptoms and causes" for disease and medical topics and "help and advice" for non-medical topics.

As a rough test of the ambiguity requirement for a probe query discussed in Section 4.3, we used the number of results indicator provided by each search engine. We recorded the number of results $N(c)$ returned from querying for sensitive topic c and also the number of results $N(c, p_j), j = 1, 2$ returned when each of the candidate probe queries is appended to the queries for topic c (with p_1="symptoms and causes" and p_2 ="help and advice"). We expect $N(c, p_j) < N(c)$ since the extra query text will narrow the query to some extent. However, we would like to avoid this narrowing being too great, *e.g.* we would certainly like to avoid $N(c, p_j) = 0$. The

	Google		Bing		Both	
Rank	**Term**	**TF**	**Term**	**TF**	**Term**	**TF**
1	help	4.37	help	4.62	help	4.49
2	advice	4.32	advice	3.45	advice	4.02
3	symptom	1.81	symptom	2.38	symptom	2.04
4	cause	0.90	check	0.77	cause	0.82
5	homecare	0.60	cause	0.68	person	0.53
6	offer	0.54	person	0.60	checker	0.49
7	person	0.48	plan	0.58	check	0.48
8	answer	0.48	checker	0.58	sign	0.45
9	gamble	0.44	sign	0.57	offer	0.43
10	checker	0.43	hiv	0.56	homecare	0.37

TABLE 4.2: Top-10 candidate probe terms with term frequency (TF) of occurrence.

values measured are reported in Table 4.3 for Google. Also reported in this table is the ratio $\hat{P}(c \mid p_j) = \frac{N(c,p_j)}{N(c)}$. It can be seen that $\hat{P}(c \mid p_1) = 0$ for the *bankrupt* topic and has a low value for *gambling*, *gay* and *unemployed*. In contrast, for these topics $\hat{P}(c \mid p_2)$ has a fairly high value. This therefore indicates the use of the "help and advice" probe query for non-medical topics rather than the "symptoms and causes" probe query, which seems intuitive. Based on Table 4.3 the "help and advice" probe query also seems reasonable for use with medical topics, and $\hat{P}(c \mid p_1)$ (corresponding to the "symptoms and causes" probe) is also reasonable for these topics. Again, this is as might be expected.

4.6.2 User Click Emulation

To reduce the appearance of robotic interaction, the script automation program inserts a random pause of 1 to 10 seconds between queries, see Table 3.2 for an example. After remaining 5 seconds on a clicked link page, the browser "back" button is invoked to navigate back to the search result page.

To emulate user clicking, we adopt the following user click model. Given the response page generated in response to a query, for each search result and advert we

		p_1 = 'symptoms and causes'		p_2 = 'help and advice'	
Topic = c	N(c)	N(c,p_1)	$\hat{P}(c\|p_1)$	N(c, p_2)	$\hat{P}(c\|p_2)$
anorexia	28.5	0.834	3%	1.78	6%
bankrupt	86.9	0.434	0%	48.6	56%
diabetes	267	66.5	25%	114	43%
disabled	506	26	5%	159	31%
divorce	185	11.1	6%	79.7	43%
gambling	103	0.526	1%	30.6	30%
gay	782	9.53	1%	119	15%
location (London)	1930	72.2	4%	373	19%
payday	70.3	45.9	65%	6.57	9%
prostate	83.3	14.7	18%	12.5	15%
unemployed	54.8	0.619	1%	48.1	88%

TABLE 4.3: Approximate result numbers returned by Google on different topics and for different choices of probe query. Counts are in units of millions.

calculate the Term-Frequency (TF) of the visible text with respect to the keywords associated with session interest category, see Table 3.1. When any keyword term associated with a topic is present the item is clicked, otherwise it is not clicked. We automate this by clicking when score $TF > 1.0$, indicating that a keyword term is present. As mentioned in Section 4.3, search results in response to probe queries are not clicked.

4.6.3 Web Search Data Collection

Scripts were executed daily in the morning and evening over 28 days. We took a number of precautions to minimise interactions between runs of each script – cleaning cookies, history and cache before and after scripts, terminating the session and logging the user out, and waiting for a minimum of twenty minutes between runs to ensure connections are reset or timed out. All scripts were run for 3 registered users and 1 anonymous user, and for both the Google and Bing search engines, yielding a data set consisting of $37,134$ queries and response. Registered users were created with the

	Training Data Sets		Test Data Sets	
Name	$N_{queries}$	N_{probes}	$N_{queries}$	N_{probes}
Bing	1,051	367	10,970	3,795
Google	1,343	451	14,669	4,488

TABLE 4.4: Summary of training and test data sets. $N_{queries}$ is the number of user search queries and N_{probes} the number of probe queries for which data was collected.

minimum profile information required by Google and Bing so that demographic data such as gender and date of birth were not provided.

The data was partitioned into training and test data sets, see Table 4.4. The test data contains 28 separate runs of each of the 12 test scripts. For training and performance evaluation we labeled all queries in a session with the intended topic of the session as given by the query script used. For example, all queries from a session about *prostate* are labeled as *prostate* or *sensitive*, including probe queries. In this respect the labels capture the intended behaviour, rather than attempting an individual interpretation of specific query keywords during a user session.

4.6.4 Feature Selection: Adverts or Links?

Search result pages contain multiple content types, in particular search links and adverts. For the collected data sets Table 4.5 summarises the percentage change in the text of search links and adverts for each of the interest categories and for each search engine. Also shown is $\pm$ the standard error in the mean. It can be seen that link text changes very little, less than 3% for Google and 5% for Bing. In contrast it can be seen that the advert text is much more dynamic with $12.4\% - 65.5\%$ of the advert text changing for Bing and $17.3\% - 39\%$ for Google.

This supports Assumption 3, namely that it is the adverts which primarily reveal personalised learning by the search engine and are the most discriminating element for probe comparison.

	Bing		Google	
Topic	**Advert**	**Link**	**Advert**	**Link**
anorexia	65.4% ± 7.7%	3.6% ± 0.3%	34.8% ± 1.5%	0.9% ± 0.2%
bankrupt	15.8% ± 1.5%	5.0% ± 0.3%	39.0% ± 2.5%	2.0% ± 0.3%
diabetes	49.4% ± 12.5%	3.9% ± 0.3%	39.5% ± 1.7%	0.9% ± 0.2%
disabled	12.4% ± 1.0%	3.5% ± 0.2%	17.3% ± 1.7%	2.1% ± 0.3%
divorce	15.8% ± 1.7%	4.7% ± 0.4%	22.1% ± 2.5%	2.9% ± 0.5%
gambling	15.7% ± 1.3%	4.0% ± 0.2%	34.2% ± 1.7%	1.8% ± 0.3%
gay	13.8% ± 1.3%	4.0% ± 0.2%	34.3% ± 1.8%	2.4% ± 0.3%
location	16.3% ± 1.5%	4.8% ± 0.3%	25.3% ± 2.1%	2.4% ± 0.4%
payday	17.4% ± 1.4%	3.9% ± 0.2%	29.7% ± 1.7%	1.4% ± 0.3%
prostate	52.6% ± 6.8%	3.7% ± 0.3%	34.6% ± 1.4%	0.9% ± 0.2%
unemployed	14.3% ± 1.2%	4.5% ± 0.3%	22.8% ± 1.8%	2.9% ± 0.5%
other	17.8% ± 27.9%	3.7% ± 0.2%	27.5% ± 1.5%	1.4% ± 0.2%

TABLE 4.5: Average percentage content change per instance of probe query, grouped by topic and search engine.

4.7 Experimental Results

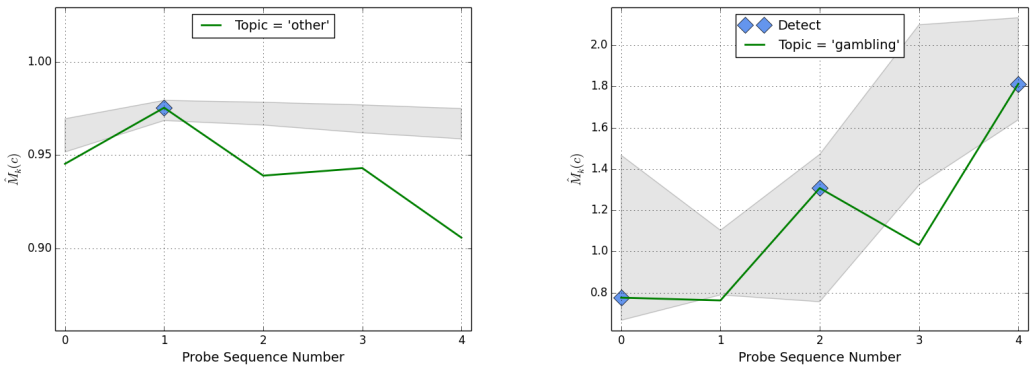
As already discussed, our approach is to issue a sequence of probe queries interleaved at steps $k \in \mathcal{K}$ amongst the user queries. We then use the **PRI** estimator to estimate $\hat{M}_{u,k}(c)$, for $k \in \mathcal{K}$ based on the response to each probe query and then look for significant changes in these $\hat{M}_{u,k}(c)$ values. To determine whether changes are significant, for each topic $c \in \mathcal{C}$, we use the mean ± three standard deviations to define a confidence interval (the mean and standard deviation are estimated using the training data). The choice of three standard deviations is taken after performing verification testing on the training data before testing. Choosing the number of standard deviations to use is a balance – too small a number of standard deviations generates excessive "False Negatives" while too large a number of standard deviations results in a larger number of "False Positives".

We use the Google Search experimental data collection setup as Chapter 3. We also collect experimental data from Bing Search, for this chapter only, using the same

setup as for Google Search to demonstrate our approach in action for search engines.

4.7.1 Sensitive – Non-sensitive Detection

We begin by evaluating the performance of this approach for detecting whether learning of any sensitive topics has taken place or not during a query session, without trying to specify which sensitive topics are involved. For this we use the catch-all other topic $\bar{c}$. Namely, when the estimate $\hat{M}_{u,k}(\bar{c})$ lies outside its confidence interval during a user session we take this as rejecting the hypothesis that no learning of sensitive topics has occurred during that session. We standardise a query session to consist of the first 5 probe queries in a run for the purposes of analysis.

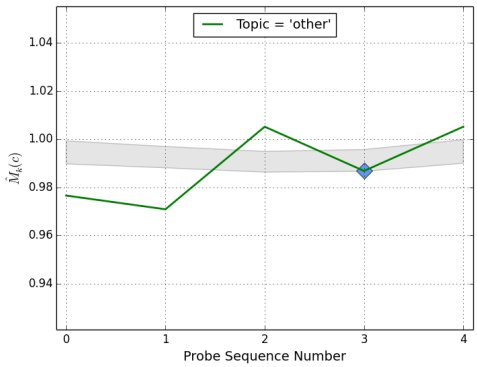


(A) Expect *gambling*, detect *other*.

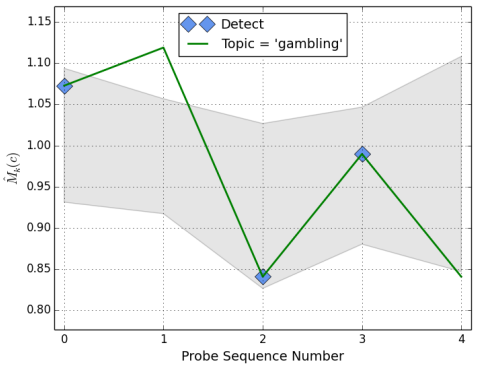
(B) Expect *gambling*, detect *gambling*.

FIGURE 4.1: Illustrating detection of learning for a user session on topic *gambling*. Shaded areas indicate the confidence interval for $\hat{M}_{u,k}$ for the *other* topic in the upper figure, and for the *gambling* topic in the lower figure. Google search engine.

The plots in Figure 4.1 illustrates this procedure for a user session on the topic *gambling* with the Google search engine. It can be seen from Figure 4.1(a) that $\hat{M}_{u,k}$ for the *other* topic (i.e. $\bar{c}$) quickly leaves its confidence interval as the session progresses (probe 1 is detected as *other*, however the other probe queries $\{0, 2, 3, 4\}$ lie outside the *other* confidence interval). In comparison, it can be seen from Figure 4.1(b) that $\hat{M}_{u,k}$ for the *gambling* topic (i.e. the topic which matches the user session) stays close to the confidence interval throughout the user session. The corresponding results for the Bing search engine are shown in Figure 4.2 and exhibit



(A) Expect *gambling*, detect *other*.



(B) Expect *gambling*, detect *gambling*.

FIGURE 4.2: Illustrating detection of learning for a user session on topic *gambling*. Shaded areas indicate the confidence interval for $\hat{M}_{u,k}$ for the *other* topic in the upper figure, and for the *gambling* topic in the lower figure. Bing search engine.

similar behaviour.

Table 4.6 summarises the detection performance on a full set of Bing and Google test data. We declare a positive detection when at least one probe query in a session of 5 probes is detected as sensitive. For user sessions on sensitive topics it can be seen that the detection accuracy is high. For Google, 100% of user sessions on a sensitive topic reject the hypothesis that no learning of the sensitive topic by the search engine has taken place and so are identified as sensitive. For Bing the corresponding detection rate is 91%. Recall that this hypothesis testing is being carried out based purely on the adverts in the response pages to user queries, and the queries themselves are not being used. We manually inspected a sample of the user sessions, confirming the results of Table 4.5, that the displayed adverts consistently change significantly over the course of user sessions on sensitive topics. It is therefore reasonable to conclude that learning by the search engine has indeed occurred. That is, the rejection of the hypothesis that no learning has occurred that is reported in Table 4.6 appears to be justified.

Table 4.6 also shows the percentage of user sessions which are sensitive but which are flagged as non-sensitive, which can be interpreted as the false negative rate. For Google, no sensitive sessions are classed as non-sensitive, and for Bing 9% are classes

		Predicted			
		Bing		Google	
		Sens.	Non-sens.	Sens.	Non-sens.
Expected	Sensitive	91%	9%	100%	0%
	Non-sensitive	1%	99%	1%	99%

TABLE 4.6: Measured detection rate of search engine learning of at least one occurrence of one or more sensitive topics during a 5 probe session.

as non-sensitive. Also shown in the table is the percentage of user sessions which are non-sensitive but are flagged as sensitive, which can be interpreted as the false positive rate. This is low at 1% for both search engines. A manual inspection of the data shows that the first probe in a session can be misdetected sometimes, demonstrating a *topic lag* effect after there is a change in topic. The influence of the first probe makes it difficult to distinguish sensitive/non-sensitive based on observation of a single step. We will discuss mis-detection in detail in Section 4.7.5.

Overall, the results in Table 4.6 indicate that the proposed approach can correctly identify potential privacy concerns for sensitive topics while keeping noise levels from false positive detection low.

We comment briefly on the difference in Table 4.6 in the measured False Negative rates for the two search engines. This difference is at least partially explained by two factors. The first is that Bing seems to be slower at adapting to changes in session topic than Google, see Section 4.7.5. This apparent difference in adaptation rate is also observable by comparing Figures 4.1(b) and 4.2(b), noting the differences in behaviour of the confidence intervals for the gambling topic. The second factor is differences between the search engines in the range and diversity of the available adverts across the various topics. For example, analysis of our test data shows that Google has on average 3.3 unique adverts per probe across all topics whereas Bing has a lower average of 1.7 unique adverts per probe. This suggests that Google's dominant position in the search market means it may have a larger advert pool allowing more finely tuned fitting of adverts to detected topics of interest.

	Reference Topic										
	anorexia	bankrupt	diabetes	disabled	divorce	gambling	gay	location	payday	prostate	unemployed
True Detect	100%	98%	100%	99%	99%	99%	98%	99%	99%	99%	99%
True Other	100%	91%	93%	93%	98%	95%	100%	87%	92%	96%	97%
False Detect	0%	9%	7%	7%	2%	5%	0%	13%	8%	4%	3%
False Other	0%	2%	0%	1%	1%	1%	2%	1%	1%	1%	1%

(A) Bing

	Reference Topic										
	anorexia	bankrupt	diabetes	disabled	divorce	gambling	gay	location	payday	prostate	unemployed
True Detect	100%	100%	96%	100%	100%	100%	100%	99%	99%	99%	100%
True Other	96%	96%	92%	100%	100%	100%	100%	100%	100%	100%	100%
False Detect	4%	4%	8%	0%	0%	0%	0%	0%	0%	0%	0%
False Other	0%	0%	4%	0%	0%	0%	0%	1%	1%	1%	0%

(B) Google

TABLE 4.7: Measured detection rate of search engine learning of individual sensitive topics.

4.7.2 Individual Sensitive Topic Detection

We now evaluate the detection performance for individual sensitive topics. For each sensitive topic c studied, when (i) the estimated $\hat{M}_{u,k}(c)$ lies inside the confidence interval for that topic and (ii) $\hat{M}_{u,k}(\bar{c})$ lies outside the confidence interval for the catch-all other topic (i.e. $\bar{c}$), then we say that we cannot reject the hypothesis that learning of topic c has occurred.

Table 4.7 summarises the detection performance for the Bing and Google test data for each of the sensitive topics studied. When evidence of learning of sensitive topic c is detected and the user session is on topic c then we label this a "True Detect", otherwise we label this a "False Detect". Conversely, when no evidence is found of topic c then when the user session is in fact on topic c we label this a "False Other", otherwise we label this a "True Other". Again, recall that the hypothesis testing here is being carried out based purely on the adverts in the response pages to probe queries.

In the Google results in Table 4.7(b), it can be seen that "True Detect" and

"True Other" results range from $96-100\%$ across all sensitive topics. "False Detect" results, corresponding to false positives, lie in a range of $0-8\%$. "False Other" results, corresponding to false negatives, are in the range $0-4\%$. We note that topics such as *bankrupt* and *payday* tended to share adverts related to financial services, see next section, making these topics harder to distinguish from one another. This data therefore provides strong support for the assertion that detection of individual sensitive topics is indeed feasible with Google.

Table 4.7(a) presents the corresponding results for Bing. The "False Detect" results, corresponding to false positives, tend to be higher than for the Google data. We note that the responses for some sensitive topics overlap in terms of advert content and are not readily differentiated in our data for Bing search (as already noted, in our data set we find that Bing displays fewer unique adverts than Google). Since our test classifies all non-sensitive topics as *other* then sensitive topics that share adverts with *other* may increase the number of false positives. Overall, the detection rate for individual sensitive topics is notably high (exceeding 98%) and the false positive rate remains below 10% except for the *location* topic.

We next test whether probe queries can themselves generate significant levels of false positive sensitive topic detection. We constructed a test script consisting of randomly selected queries from Google Trends into which we injected the previously selected probe queries. This randomised script was executed for both Bing and Google and for each of our user configurations. Relevant result items appearing on non-probe queries were clicked. In total $1,264$ probe queries were tested for both Bing and Google using the **PRI** framework. Tests yielded a 0% sensitive topic detection rate for any sensitive topic in combinations of search engine and users. We conclude that the selected probe queries do not themselves generate a significant amount of false sensitive topic detection.

4.7.3 Topic Similarity and Topic Confusion

Intuitively, we expect that some sensitive topics are similar in the sense that similar adverts tend to be associated with each. For example, the adverts prompted by the

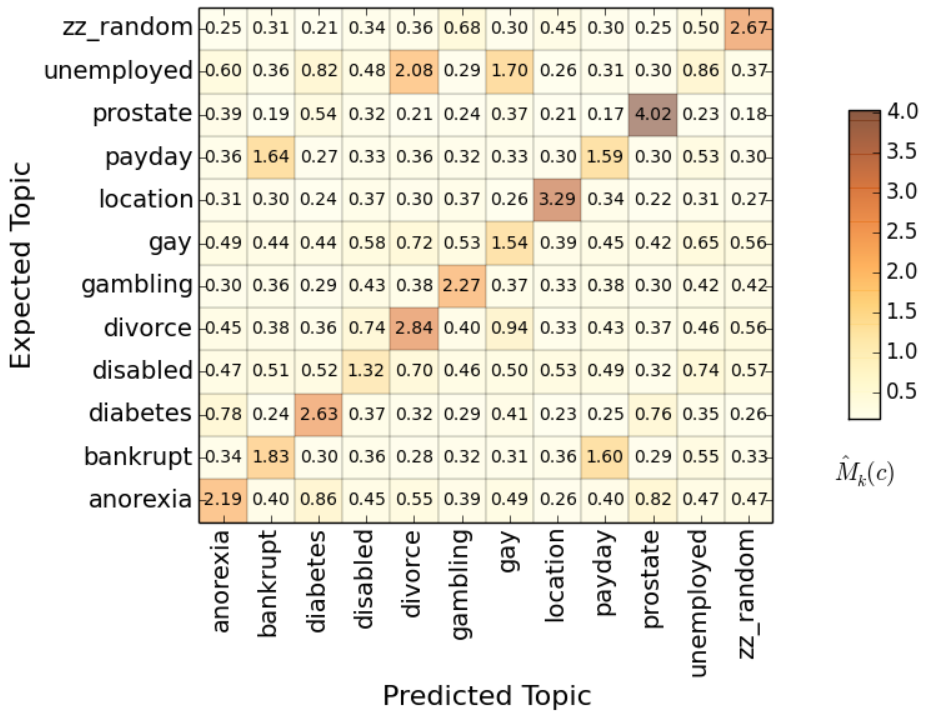
bankrupt topic, which relates to insolvency, might be expected to have some overlap with the *payday* topic, which relates to short-term loans.

We can gain some insight into this via the $\widehat{\mathbb{M}}_{u,k}(c)$ estimates for each topic. Figure 4.3 shows the average $\widehat{\mathbb{M}}_{u,k}(c)$ measured for each topic c vs the user session topic. That is, cell (i, j) shows the average $\widehat{\mathbb{M}}_{u,k}(c)$ measured value attained by topic j when running query scripts for reference topic i. Each cell is heat-mapped within its row, from brightest for maximum value to darkest for lowest value per row, to improve readability. Figure 4.3(a) shows results for the Google data and Figure 4.3(b) for the Bing data.

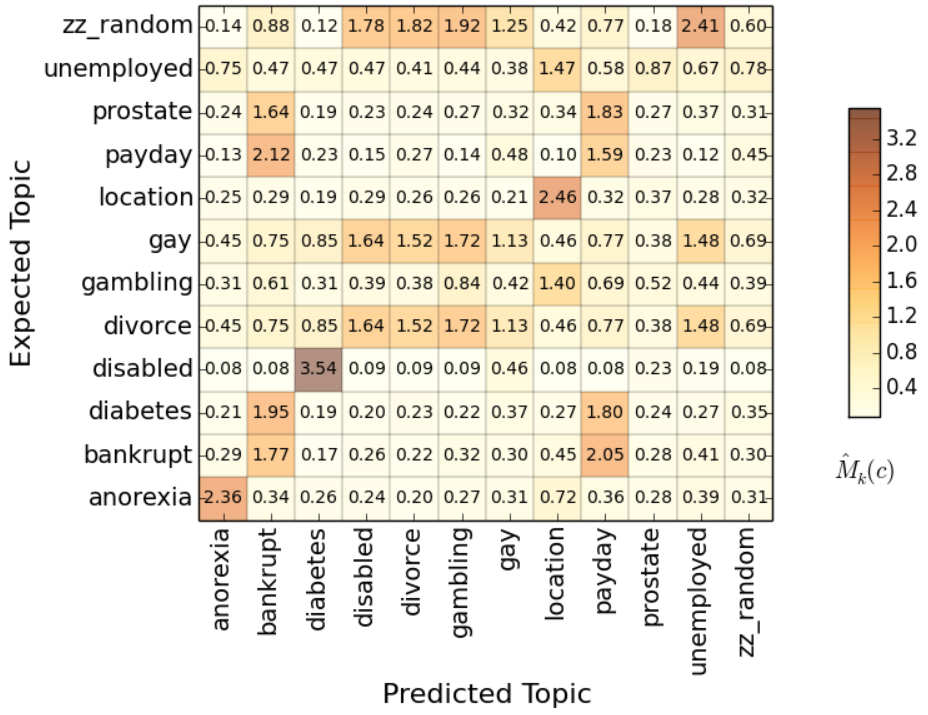
For the Google data, it can be seen that the maximum element in each row and column is the diagonal element, as expected from the results presented in the previous section. However, it can also be seen that the *payday* topic has a significantly higher $\widehat{\mathbb{M}}_{u,k}(c)$ value than other topics for user sessions on the *bankrupt* topic. Similarly, the *bankrupt* topic has a significantly higher $\widehat{\mathbb{M}}_{u,k}(c)$ value for user sessions on the *payday* topic. Less pronounced, but still evident, is that all health related topics tend have a higher $\widehat{\mathbb{M}}_{u,k}(c)$ value whenever the user session is on a health topic. For example, *diabetes* and *prostate* have elevated $\widehat{\mathbb{M}}_{u,k}(c)$ values for user sessions on *anorexia*.

For the Bing data in Figure 4.3(b) it can be seen that the results are more complicated. As with Google, the adverts for the *payday* and *bankrupt* topics show correlated behaviour. Similarly, the adverts for health-related topics tend to be correlated. However, the Bing adverts for the *disabled*, *divorce*, *gambling*, *gay* and *unemployed* topics also exhibit significant correlation. This is consistent with the results in Section 4.7.2 where it was observed that topics for Bing appear less readily distinguishable, possibly due to the smaller size of the pool of available adverts.

While the existence of correlation among topics is itself unsurprising, the fact that the proposed approach for detecting search engine learning is able to uncover this correlation provides additional support for the effectiveness of the approach. It also suggests that the potential exists to use the approach to infer additional information from displayed adverts. We explore this further in the following sections.



(A) Google



(B) Bing

FIGURE 4.3: Average $\widehat{\mathbb{M}}_{u,k}(c)$ measured by topic.

Topic – % Increase in $\widehat{\mathbb{M}}_{u,k}(c)$					
anorexia	49%	*divorce*	153%	*payday*	62%
bankrupt	30%	*gambling*	108%	*prostate*	451%
diabetes	417%	*gay*	158%	*unemployed*	62%
disabled	57%	*location*	63%	*other*	233%

TABLE 4.8: Percentage increase in $\widehat{\mathbb{M}}_{u,k}(c)$ by topic for click versus non-click. Google search data.

4.7.4 You click – therefore – I learn!

In addition to entering queries, users provide feedback to the search engine via the links that they click. Since clicking is an active step, we might expect it to influence search engine learning. Separate sets of non-click data were collected by running a single iteration of all of the test scripts on both search engines with user clicking turned off. Table 4.8 shows the percentage change in the average $\widehat{\mathbb{M}}_{u,k}(c)$ score for each test topic with and without user clicking of relevant search results. It can be seen that all topics had higher $\widehat{\mathbb{M}}_{u,k}(c)$ values when the user clicks on relevant links, suggesting that user clicks are actively used by the search engine for learning.

4.7.5 Time to Learn?

Inspection of the test data reveals that correct topic identification sometimes lags by one to two probes at the start of a new user session. This accounts for approximately 70% of cases where "False Detects" and "False Other" results are encountered in testing. Examination of these cases provides insight into the observed *speed* of learning, and the potential consequences for noise based privacy defences. Letting X denote the random variable counting the number of consecutive mis-classifications occurring together, then dividing by the total number of mis-classifications we can estimate the probability that $X = 1$, $X = 2$, *etc.* This data is shown in the first column of Table 4.10. It can be seen that there are no runs of more than two mis-classifications

Number of Consecutive Mis-classifications (X)			**Probe ID of First Mis-classification (Y)**		
	Bing	**Google**		**Bing**	**Google**
$\Pr(X = 1)$	0.23	0.95	$\Pr(Y = 1)$	0.92	0.98
$\Pr(X = 2)$	0.77	0.05	$\Pr(Y = 2)$	0.03	0.01
$\Pr(X = 3)$	0.00	0.00	$\Pr(Y = 3)$	0.04	0.01
$\Pr(X = 4)$	0.00	0.00	$\Pr(Y = 4)$	0.00	0.00
$\Pr(X = 5)$	0.00	0.00	$\Pr(Y = 5)$	0.00	0.00

TABLE 4.9: Recall rate by probe query excluding successive probe queries – Google.

and the average length of a run of mis-classifications is,

$$\mathscr{E}[X;\ \text{Bing}\ ] = 1.77 \tag{4.13}$$

$$\mathscr{E}[X;\ \text{Google}\ ] = 1.05 \tag{4.14}$$

Letting Y be a random variable indicating the probe sequence number where a "False Detects" or "False Other" event *first* occurs, Table 4.10 reports the estimated probability that $Y = 1$, $Y = 2$, *etc.* As expected the overwhelming majority for "False Detects" and "False Other" events happen on the first probe in a session, with $\Pr(Y = 1) > 0.90$ for both Bing and Google.

The data in Table 4.10 therefore suggests that Google search takes an average of 1.05 probe queries and Bing takes an average of 1.77 probe queries to re-calibrate learning after a topic change. On average probe queries in the test data were issued after 4 user queries. Hence, Google appears to adapt to a new topic in approximately 4 queries, while Bing requires approximately 7 queries. Rapid re-calibration can also be seen in Table 4.9 by looking at sensitive topic classification recall for Google when successive probe queries are excluded from the calculation. When every probe query is included true positive recall is 62%. True positive accuracy improves once the first probe query is excluded and stabilises at 66% thereafter. The false positive rates are low in all cases, falling to 0% when the first three probes are excluded.

This means that a privacy defence based on random topic changes achieved, for

	Include All	**Exclude $k = 1$**	**Exclude $k = 1, 2$**	**Exclude $k = 1, 2, 3$**
True Positive	62%	66%	66%	66%
False Positive	1%	1%	1%	0%

TABLE 4.10: Estimated probabilities of mis-classification of various lengths and probe number of first mis-classification in a session.

		Predicted			
		Bing		**Google**	
		Sensitive	**Non-sensitive**	**Sensitive**	**Non-sensitive**
Expected	**Sensitive**	83%	0%	100%	0%
	Non-sensitive	17%	100%	0%	100%

TABLE 4.11: Measured detection rate of search engine learning for an anonymous user.

example, by injecting spurious queries, could prove to be ineffective unless the spurious queries are repeated at intervals of less than every 4 real queries for Google and 7 for Bing. This is a considerable overhead.

4.7.6 Logged-in vs Anonymous

We collected data for user sessions both when the user is logged-in and when the user is anonymous. As already noted, we clean local caches and user session data between each user session.

Figure 4.4 shows the average $\widehat{\mathbb{M}}_{u,k}(c)$ measured for each topic for the Google search engine when the user is not logged in. It can be seen that this shows a similar overall pattern to Figure 4.3(a), suggesting the search engine is successful at identifying sensitive topics even in the case of an anonymous user.

Table 4.11 shows the corresponding measured rates for sensitive/non-sensitive topic detection, which can be compared to Table 4.6. Table 4.12 shows the detection rate for individual topics, which can be compared to Table 4.7. It can be seen that the detection rates are similar to the results presented previously for logged-in users.

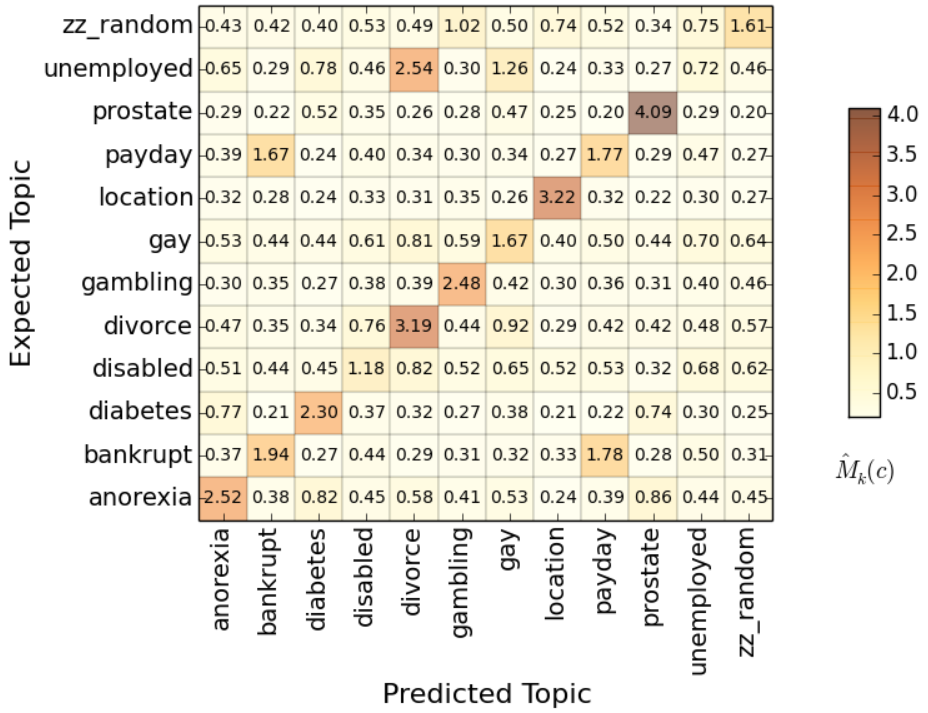


FIGURE 4.4: Average $\widehat{\mathbb{M}}_{u,k}(c)$ by topic. Anonymous user, Google test data

In particular the True Detection rate for individual topics is high e.g. $97-100\%$ for Google.

We conclude that anonymity seems to provide little protection within an individual query session. The results of Section 4.7.5 show that the users previous search history is not really required to infer the topic of a sessions, the session itself is enough.

4.7.7 Comparison with Other Estimators

We also compare the performance of **PRI** with alternative implementations using Naive Bayes and Support Vector Machine as sensitive topic detectors. Comparison of **PRI** with alternative implementations was performed by taking results from Multinomial Naive Bayes (NB) and Linear SVM (SVM) classifiers to estimate the probabilities in (4.2) and so provide alternative estimations of $\widehat{\mathbb{M}}$. The intent of the comparison is to determine which of the NB, **PRI** and SVM estimators detect privacy threats, using the definition of $\widehat{\mathbb{M}}$, for test items previously labeled as sensitive or non-sensitive.

	Reference Topic										
	anorexia	bankrupt	diabetes	disabled	divorce	gambling	gay	location	payday	prostate	unemployed
True Detect	100%	95%	100%	98%	100%	100%	96%	100%	100%	98%	100%
True Other	100%	83%	86%	86%	100%	100%	100%	75%	100%	100%	100%
False Detect	0%	17%	14%	14%	0%	0%	0%	25%	0%	0%	0%
False Other	0%	5%	0%	2%	0%	0%	4%	1%	0%	2%	0%

(A) Bing

	Reference Topic										
	anorexia	bankrupt	diabetes	disabled	divorce	gambling	gay	location	payday	prostate	unemployed
True Detect	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
True Other	100%	100%	92%	100%	100%	100%	100%	100%	100%	100%	100%
False Detect	4%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%
False Other	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

(B) Google

TABLE 4.12: Measured detection rate of search engine learning of individual sensitive topics for an anonymous user.

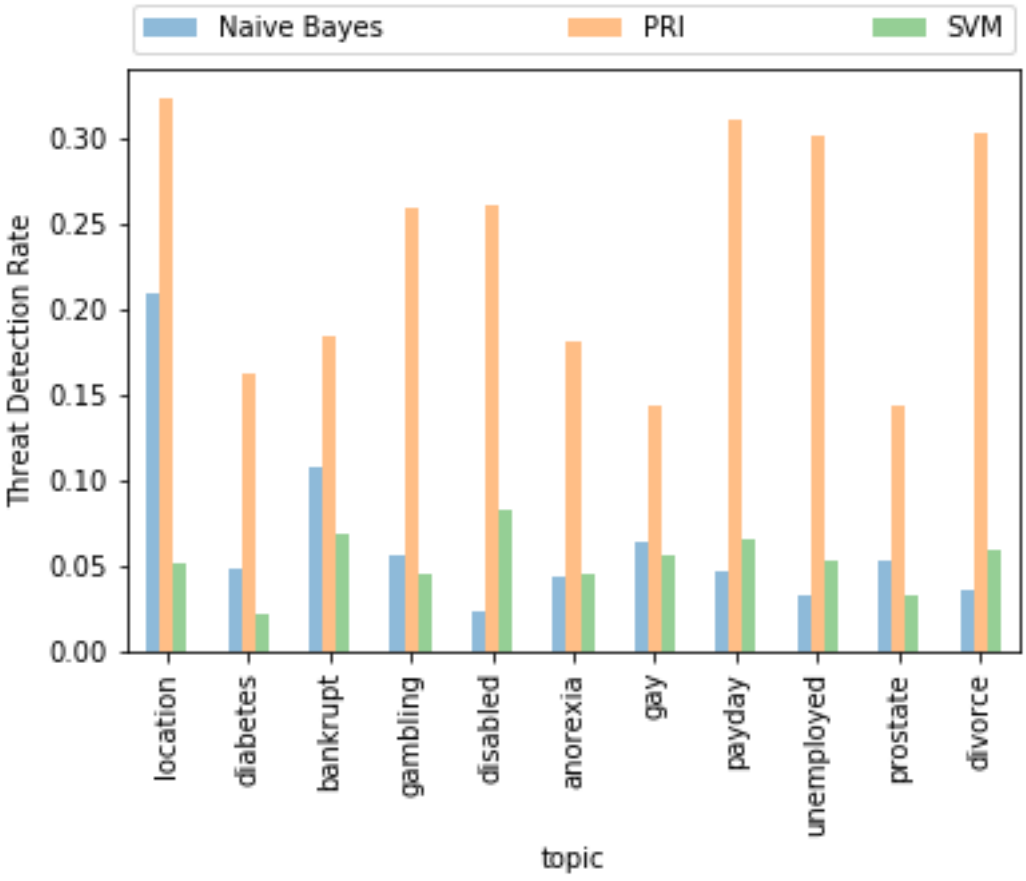


FIGURE 4.5: Comparison of Naive Bayes, **PRI** and Support Vector Machine estimators. (as Threat Detection Rate by Topic)

All inputs and calculations of $\widehat{\mathbb{M}}$ were performed in an identical manner for all classifiers. A common test data set was constructed by selecting 5,500 result pages for each sensitive topic and then randomly selecting an additional 5,500 result pages labeled for the non-sensitive topic from the Search data set. In this way each sensitive topic had a balanced verification data set of 11,000 labeled items. Each verification data-set was divided randomly into $20\% - 80\%$ test–training sets and calculations repeated 5 times for 5-fold verification of each of the NB, **PRI** and SVM estimators. The Multinomial Naive Bayes and Linear SVC modules from the Python Sklearn package were used to construct the NB and SVM estimators, (Pedregosa et al., 2012). After common preprocessing each of the NB, **PRI** and SVM classifiers were trained and probability estimates captured for the 5-fold test data-sets. A threat is declared "detected" if the calculated value of $\widehat{\mathbb{M}}$ for the sensitive topic exceeds 1.0. Precision of sensitive topic threat detection is shown by topic in Figure 4.5 for the NB, **PRI** and SVM approaches.

The results Figure 4.5 indicate that that the **PRI** estimator detects significantly more true-positive detection results than either of the NB or SVM estimators for all sensitive topics tested. The initial detection sensitivity of each of these estimators is influenced by the labelling assigned to examples in the training set. We adopt the perspective that privacy tools should err on the side of caution so that high detection sensitivity in the initial "out of the box" stage is a prudent approach. In a real-world application of **PRI** the user would provide incremental training examples over time reflecting their tolerance of privacy risk and so tune **PRI**.

4.8 Conclusion

With ϵ-*Indistinguishability* as a practical model for detection of user privacy risk, we show that this is readily implementable with available open tools that are simple to apply and provide highly accurate results. An appealing aspect is the use of openly available resources – Bing and Google search – a feature often missing in traditional privacy research where concerns over data disclosure limit access to potentially sensitive test data sources.

Using Bing and Google Search, we demonstrated that by monitoring changes in the adverts displayed in the response to probe queries we are able to accurately detect evidence of learning for a range of sensitive topics in over 98% of cases. Topics studied include medical conditions (cancer, anorexia *etc*), sexual orientation, disability, bankruptcy and unemployment. Our method is accurate, with typical false detection rates of less than 10% (and less than 1% for many sensitive topics).

We also show that detection rates remain high for anonymous users, suggesting that search engines learn quickly; even without search history as background knowledge. Our experiments suggest that search engines have an ability to learn user interests quickly. Our estimation of search engine adaptation rates indicate that sensitive topic learning is detectable after as few as $3-4$ queries on average.

Finally we compare **PRI** with privacy detectors constructed using common, openly available, machine learning classifiers. We show that **PRI** is more accurate in correctly detecting privacy concerns.

Chapter 5

Assessing Threats - Plausible Deniability

5.1 Introduction

When unwanted personalisation suggests we are interested in sensitive topics a natural reaction is to deny interest. To be credible, our denials should be plausible – so that with high probability, our action is consistent with interest in any of several topics. When is denial of our interest in topics plausible? What defences can we deploy to protect our ability to deny interests in a plausible manner? In this chapter we extend the work of Chapter 4 on detecting privacy concerns to address the question of how to assess the degree of threat associated with a detection.

We begin this chapter by formalising the notion of plausible deniability of interest in topics during web search we call (ϵ, m)–Plausible Deniability. Our intention is to allow users to test if they can reasonably deny their interest in topics they regard as sensitive. To show this we implement aan estimator called **PDE** ("**P**lausible **D**eniability **E**stimator") implementing (ϵ, m)–Plausible Deniability. We also show that (ϵ, m)–Plausible Deniability is compatible with ϵ-*Indistinguishability* so that much of the machinery of **PRI** from Chapter 4 can be reused to implement **PDE**.

In the experimental section we show that **PDE** effectively detects threats to plausible deniability even when user queries are obfuscated through injection of high levels of noise. A particular concern uncovered during experiments is that profiling with respect to topics such as secual orientation and financial status are least plausibly deniable in our tests. We also introduce a novel defence for plausible deniability, called the Proxy Topic Defence here, that is observed to provide protection in 100% of tests. We report result in this chapter for the Google search engine only for conciseness and clarity as our results are similar for Bing.

5.2 Plausible Deniability

The setup we consider is that of general users of a commercial, for–profit search engine. The relationship between the users and the search engine is based on mutual utility where both parties obtain something useful from the interaction users get useful

information and recommendations – while the search engine gets an opportunity to "up- sell" to users through targeted content such as advertising. As a commercial business, the search engine recognises cost per user interaction and responsiveness of service are critical to competitiveness. Accordingly content based on user profiling is intended to adapt dynamically to the changing interests of users. When a user detects threats to her privacy we assume she wishes to assess her ability to plausibly deny her interest in compromising content and so avoid awkward social implications.

We consider the Basic Black-box Model where users access a search engine $\mathscr{S}$ directly. Inputs are web-search queries submitted to the search engine and outputs are the corresponding response pages containing personalised content such as adverts. For a subset $\mathcal{A} \subseteq \mathcal{C}$ let $\mathcal{Z}_{u,k}^{u,\mathcal{A}} = \{z \in \mathcal{Z}_{u,k} : l_u(z) \in \mathcal{A}\}$ denote the subsequence of observations user u has labelled for any of the topics in $\mathcal{A}$ at step k.

We formalise the notion of (ϵ, m)–Plausible Deniability tailored to the context of the Basic Black-box Model as follows

Definition 4 ((ϵ, m)–Plausible Deniability) *A user with interest in a topic $c \in \mathcal{C}$ is said to have* (ϵ, m)–Plausible Deniability *for the sequence of interactions $\mathcal{Z}_{u,k} := (z_1, \dots z_k) \cup \mathcal{Z}_{u,0}$, if there is a set of topics $\mathcal{A} \subseteq \mathcal{C} \setminus \{c\}$, such that*

$$e^{-\epsilon} < \mathbb{D}_k(c, \mathcal{A}) < e^{\epsilon} \tag{5.1}$$

with $|\mathcal{A} \cup \{c\}| = m$, and

$$\mathbb{D}_k(c, \mathcal{A}) = \frac{\mathrm{P}(Z_k = z_k, \dots Z_1 = z_1 | I_u^{u,c} = 1, \mathcal{Z}_{u,0})}{\mathrm{P}(Z_k = z_k, \dots Z_1 = z_1 | I_u^{u,\mathcal{A}} = 1, \mathcal{Z}_{u,0})} \tag{5.2}$$

where $I_u^{u,c}$ is an indicator random variable with value 1 when u is interested in topic c and 0 otherwise, $I_u^{u,\mathcal{A}}$ is an indicator random variable with value 1 when u is interested in any topic in $\mathcal{A}$ and 0 otherwise, and $Z_j = \zeta$, $j = 1, 2, \dots$ denotes the event that input–output interaction $\zeta \in \mathcal{Z}_{u,j}$ is observed at step j of the session.

The privacy parameters $\epsilon > 0$ and $m > 1$ in Definition 4 are chosen by the user. For (5.2) to be well-defined, all probabilities are assumed to be non-zero. In practice, this

is not a significant restriction since categories with zero probability can be gathered into the catch-all topic c_0.

Given the sequence of observations associated with a user, our aim is to (1) estimate whether (ϵ, m)–Plausible Deniability has been violated for one or more of the sensitive categories in $\mathcal{C}$ (and so the user cannot reasonably deny interest in one or more of these categories), and (2) identify which of these sensitive categories has been affected, with high probability in both cases.

Expression (5.2) in Definition 4 can be rewritten as

$$\mathbb{D}_k(c, \mathcal{A}) = \prod_{j=0}^{k-1} \frac{\mathrm{P}(Z_{k-j} = z_{k-j} \mid I_u^{u,c} = 1, \mathcal{Z}_{u,k-j-1})}{\mathrm{P}(Z_{k-j} = z_{k-j} \mid I_u^{u,\mathcal{A}} = 1, \mathcal{Z}_{u,k-j-1})} \tag{5.3}$$

$$= \prod_{j=0}^{k-1} \mathbf{d}_{k-j}(c, \mathcal{A}) \tag{5.4}$$

where the step (5.3) results from applying the chain-rule for conditional probability to the RHS of (5.2), and

$$\mathbf{d}_{k-j}(c, \mathcal{A}) := \prod_{j=0}^{k-1} \frac{\mathrm{P}(Z_{k-j} = z_{k-j} \mid I_u^{u,c} = 1, \mathcal{Z}_{u,k-j-1})}{\mathrm{P}(Z_{k-j} = z_{k-j} \mid I_u^{u,\mathcal{A}} = 1, \mathcal{Z}_{u,k-j-1})} \tag{5.5}$$

is the incremental change in (ϵ, m)–Plausible Deniability at step $k - j$.

5.2.1 Comparison with Other Anonymity Measures

Intuitively, Definition 4 is similar to k–anonymity in that an observer can only explain observations to within a set consisting of at least $k := m$ topics with probability bounded by the choice of ϵ. Definition 4 differs from regular k–anonymity in requiring both upper and lower bounds on (5.2) since evidence of *loss of interest* in a sensitive topic may be as revealing as evidence of *increase of interest*.

Definition 4 can also be compared with a slightly weaker form of Differential Privacy. Informally, making an observation should not make $\mathscr{S}$ significantly more, or less, confident of user interest in a particular sensitive topic.

From (5.5) the incremental change due to a single observation $Z_j = z_j$ is

$$\frac{\mathrm{P}(Z_j = z_j \,|\, I_u^{u,c} = 1, \mathcal{Z}_{u,j-1})}{\mathrm{P}(Z_j = z_j \,|\, I_u^{u,\mathcal{A}} = 1, \mathcal{Z}_{u,j-1})} = \frac{\mathrm{P}(I_u^{u,c} = 1 \,|\, Z_j = z_j, \mathcal{Z}_{u,j-1})}{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \,|\, Z_j = z_j, \mathcal{Z}_{u,j-1})} \tag{5.6}$$

by applying Bayes Theorem. Since (5.5) is bounded above and below for at least $m-1$ other topics in $\mathcal{A}$ when Definition 4 holds, it follows that

$$e^{-\epsilon} < \frac{\mathrm{P}(I_u^{u,c} = 1 \,|\, Z_j = z_j, \mathcal{Z}_{u,j-1})}{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \,|\, Z_j = z_j, \mathcal{Z}_{u,j-1})} < e^{\epsilon} \tag{5.7}$$

for at least $m-1$ other topics in $\mathcal{A}$ – but not necessarily for *all* topic vectors. In which case we say that m–Differential Privacy holds for $\epsilon > 0$ whenever Definition 4 holds, meaning that for any topic c it is impossible to distinguish it from at least $m-1$ other topic vectors in in $\mathcal{A}$. This is a slightly weaker statement of Differential Privacy from the usual global definition.

5.2.2 Testing for Plausible Deniability

The next result provides the necessary connection to apply ϵ-*Indistinguishability* from Chapter 4 to (ϵ, m)–Plausible Deniability.

Proposition 5.2.1 *If ϵ-Indistinguishability holds for a topic c and on a subset $\mathcal{A} \subseteq \mathcal{C}$ for $\epsilon > 0$ at step k and at the initial step 1, then $(4\epsilon, m)$–Plausible Deniability holds so that*

$$e^{-4\epsilon} < \mathbb{D}_k(c, \mathcal{A}) < e^{4\epsilon}$$

for $m \leq |\mathcal{A}|$. Furthermore,

$$\mathbb{D}_k(c, \mathcal{A}) = \frac{\mathbb{M}_k(c)}{\mathbb{M}_k(\mathcal{A})} \frac{\mathbb{M}_1(\mathcal{A})}{\mathbb{M}_1(c)} \tag{5.8}$$

where $\mathbb{M}_k(\mathcal{A})$ and $\mathbb{M}_1(\mathcal{A})$ denote values of (4.2) taken over the set of topics in $\mathcal{A}$.

Proof *For a topic $c \in \mathcal{C}$, assume ϵ-Indistinguishability holds at c and on $\mathcal{A} \subseteq \mathcal{C} \setminus \{c\}$ for $\epsilon > 0$. From* (5.5)

$$\begin{aligned}
\mathbf{d}_k(c, \mathcal{A}) &:= \frac{\mathrm{P}(Z_k = z_k \mid I_u^{u,c} = 1, \mathcal{Z}_{u,k-1})}{\mathrm{P}(Z_k = z_k \mid I_u^{u,\mathcal{A}} = 1, \mathcal{Z}_{u,k-1})} \\
&= \frac{\mathrm{P}(I_u^{u,c} = 1 \mid Z_k = z_k, \mathcal{Z}_{u,k-1})}{\mathrm{P}(I_u^{u,c} = 1 \mid \mathcal{Z}_{u,k-1})} \frac{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \mid \mathcal{Z}_{u,k-1})}{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \mid Z_k = z_k, \mathcal{Z}_{u,k-1})} && (5.9) \\
&= \underbrace{\frac{\mathrm{P}(I_u^{u,c} = 1 \mid Z_k = z_k, \mathcal{Z}_{u,k-1})}{\mathrm{P}(I_u^{u,c} = 1 \mid \mathcal{Z}_{u,0})}}_{(a)} \underbrace{\frac{\mathrm{P}(I_u^{u,c} = 1 \mid \mathcal{Z}_{u,0})}{\mathrm{P}(I_u^{u,c} = 1 \mid \mathcal{Z}_{u,k-1})}}_{(b)} \\
&\qquad \times \underbrace{\frac{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \mid \mathcal{Z}_{u,k-1})}{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \mid \mathcal{Z}_{u,0})}}_{(c)} \underbrace{\frac{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \mid \mathcal{Z}_{u,0})}{\mathrm{P}(I_u^{u,\mathcal{A}} = 1 \mid Z_k = z_k, \mathcal{Z}_{u,k-1})}}_{(d)} && (5.10) \\
&= \frac{\overbrace{\mathbb{M}_k(c)}^{(a)} \overbrace{\mathbb{M}_{k-1}(\mathcal{A})}^{(c)}}{\underbrace{\mathbb{M}_{k-1}(c)}_{(b)} \underbrace{\mathbb{M}_k(\mathcal{A})}_{(d)}} && (5.11)
\end{aligned}$$

where (5.9) *follows from Bayes Theorem. Expressions (a) – (d)* (5.10) *and* (5.11) *come directly from the definition of $\mathbb{M}_k$ in* (4.2)*.*

Applying (5.4) *results in*

$$\mathbb{D}_k(c, \mathcal{A}) = \frac{\mathbb{M}_k(c)}{\mathbb{M}_k(\mathcal{A})} \frac{\mathbb{M}_1(\mathcal{A})}{\mathbb{M}_1(c)} \tag{5.12}$$

So that (5.8) *holds.*

Since individual elements in (5.12) *satisfy ϵ-Indistinguishability for $\epsilon > 0$ it follows that $(4\epsilon, m)$–Plausible Deniability holds as required.* ■

By establishing a value of ϵ for which a collection of topics $\mathcal{A}$ that satisfies ϵ-*Indistinguishability*, $(4\epsilon, m)$–Plausible Deniability follows with, at least, $m = |\mathcal{A}|$. This is a *minimum* guarantee, as there may be topics for which ϵ-*Indistinguishability* fails but $(4\epsilon, m)$–Plausible Deniability holds.

In later experiments we test whether the user can plausibly deny whether or not observed actions can be uniquely associated with interest in a given sensitive topic c_1 versus interest in "any other" topic $c_0 = \mathcal{C} \setminus \{c_1\}$ so that $m = 2$. From (5.2) the

expression for (ϵ, m)–Plausible Deniability at step k is

$$\mathbb{D}_k(c_1, c_0) = \frac{\mathrm{P}(Z_k = z_k, \dots Z_1 = z_1 | I_u^{u,c_1} = 1, \mathcal{Z}_{u,0})}{\mathrm{P}(Z_k = z_k, \dots Z_1 = z_1 | I_u^{u,c_0} = 1, \mathcal{Z}_{u,0})} \tag{5.13}$$

Proposition 5.2.2 *If (ϵ, m)–Plausible Deniability holds for $\{c_1, c_0\}$ with $\epsilon > 0$ and $m = 2$ then*

$$\epsilon_* := \left| \log \left(\frac{\mathbb{M}_k(c_1)}{\mathbb{M}_k(c_0)} \frac{\mathbb{M}_1(c_0)}{\mathbb{M}_1(c_1)} \right) \right| \tag{5.14}$$

is a lower bound for the best possible achievable level of (ϵ, m)–Plausible Deniability.

Proof *If (ϵ, m)–Plausible Deniability holds for $\epsilon > 0$ then*

$$|\log(\mathbb{D}_k(c_1, c_0))| < \epsilon \tag{5.15}$$

and $|\log(\mathbb{D}_k(c_1, c_0))|$ is a lower bound for all $\epsilon > 0$ where (ϵ, m)–Plausible Deniability holds. ■

Proposition 5.2.2 will be used later to create an estimator for ϵ_* that can be measured in experiments. From now on we simplify our discussion to the case $m = 2$ and so experimental results are reported for the two-topic case accordingly.

The following result connects (ϵ, m)–Plausible Deniability to variation in probabilities

Proposition 5.2.3 *If (ϵ, m)–Plausible Deniability holds for $\{c_1, c_0\}$ with $\epsilon > 0$ and $m = 2$ then*

$$\begin{aligned} &|\mathrm{P}(Z_k = z_k, \dots Z_1 = z_1 | I_u^{u,c_1} = 1, \mathcal{Z}_{u,0}) - \mathrm{P}(Z_k = z_k, \dots Z_1 = z_1 | I_u^{u,c_0} = 1, \mathcal{Z}_{u,0})| \\ &\quad \leq \left| \log \left(\frac{\mathbb{M}_k(c_1)}{\mathbb{M}_k(c_0)} \frac{\mathbb{M}_1(c_0)}{\mathbb{M}_1(c_1)} \right) \right| \end{aligned} \tag{5.16}$$

Proof *If (ϵ, m)–Plausible Deniability holds for $\epsilon > 0$ the result follows from Proposition 5.2.1, Proposition 5.2.2 and by applying Lemma 1 in the Appendix to* (5.2).

■

5.3 Implementation

5.3.1 The PDE Estimator

Substituting the **PRI** estimator $\widehat{\mathbb{M}_k}$, from Chapter 4, into (5.14) gives the **PDE** estimator

$$\widehat{\epsilon}_{*,k} = \left| \log \left(\frac{\widehat{\mathbb{M}_k}(c)}{\widehat{\mathbb{M}_k}(c_0)} \frac{\widehat{\mathbb{M}_1}(c_0)}{\widehat{\mathbb{M}_1}(c)} \right) \right| \tag{5.17}$$

From Proposition 5.2.2, the **PDE** estimator in (5.17) can be interpreted directly as the best possible level of (ϵ, m)–Plausible Deniability a user can claim in the case $m = 2$. We report the maximum value of **PDE** measured by probe step in our experiments to show the worst possible (ϵ, m)–Plausible Deniability scenario for the user. We also report the median value of **PDE** as a representative bound for approximately 50% of the samples. An example of reporting is shown in Table 5.1 for the reference topic "gay".

TABLE 5.1: Measured $\widehat{\epsilon}_{*,k}$ for Reference Topic versus Any Other Topic, reported as "max (median)", by Probe Query Sequence

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
gay	64 (33)	47 (5)	72 (25)	48 (25)	48 (19)

For example, from Table 5.1, a reported maximum value of **PDE** of 47% in the second column indicates that the difference in probabilities that the user is uniquely interest in the reference topic versus being interested in any other topic is *at least* 47% in the worst case by probe step 5. The median value of 25% in parentheses in the Probe 3 and 4 columns indicates that the difference in probabilities can be expected to be at least 25% in 50% of cases by probes 3 and 4. Overall the results suggest that (ϵ, m)–Plausible Deniability is unlikely to constitute a reasonable defence in this case.

Reported values of **PDE** may increase, or decrease, during a session as individual queries are judged as more, or less, revealing by the **PDE** estimator. Inspection of the query scripts generated for the topic c_i = Gay, for example, shows that the queries

associated with probe step 3 are *same sex relationships* and *how do i know if I'm gay*, both of which appear revealing. The queries from the test script corresponding to probe steps 4 and 5 are *HIV symptoms*, *HIV treatment*, *HIV men* and *aids men* which may not point as distinctly to specific interest in the c_i = Gay as they could reasonably be associated with health concerns.

The zeroth probe in a session is always run first, before any other query, to establish a baseline **PRI** score for the session. As a result the measured **PDE** values for the zeroth probe is always 0 for both maximum and median values and is not reported in our results.

One popular approach to designing defences of (ϵ, m)–Plausible Deniability is to attempt to *hide in the crowd*. For example, by injecting varying degrees of noise in the stream of observations in the hope that $\mathscr{S}$ will not detect the true sub-stream of sensitive events. In Chapter 4 it was observed that varying click patterns is seen to change the absolute volume of adverts appearing on a page. As both user clicks and queries are potential indicators of user interest for an observer we test injected noise from both queries and clicks as possible defence strategies.

An alternative tactic is to invert the previous approach by instead attempting to *hide in plain sight*. By choosing a non-sensitive *proxy topic*, chosen to attract personalised content the user can then carefully hide true, sensitive queries in a stream of proxy topic queries. By demonstrating clear interest in a *proxy* non-sensitive topic the user may tip the balance of probability toward the proxy topic by drawing the attention of the observer $\mathscr{S}$.

5.4 Experimental Results

We use the same experimental data collection setup as Chapter 3.

5.4.1 Establishing a Baseline

We begin with a sequences of queries, interleaved with probe queries, in what we term a "no click, no noise" model. Here there is no injected noise and no items are clicked on any of the search results pages. This model provides a baseline, where the queries

alone are available to the recommender to learn about a user session as it progresses. Measurements of **PDE** for all topics using the "no click, no noise" model are shown in Table 5.2. For the topics *Anorexia, Diabetes, Prostate, Bankrupt, Divorced, Gay* the

TABLE 5.2: Measured $\widehat{\epsilon}_{*,k}$for Reference Topic versus Any Other Topic, reported as "max (median)", by Probe Query Sequence

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	56 (52)	56 (52)	56 (52)	56 (52)	56 (52)
bankrupt	1 (1)	55 (43)	55 (39)	58 (48)	56 (48)
diabetes	40 (38)	40 (38)	40 (38)	40 (38)	40 (38)
disabled	9 (9)	9 (9)	9 (9)	40 (40)	40 (33)
divorce	41 (31)	75 (65)	56 (46)	79 (68)	79 (68)
gambling	16 (12)	18 (16)	66 (4)	57 (17)	18 (3)
gay	64 (33)	47 (5)	72 (25)	48 (25)	48 (19)
location	10 (2)	11 (3)	11 (10)	18 (7)	18 (9)
payday	2 (2)	2 (2)	21 (2)	2 (2)	2 (2)
prostate	52 (17)	52 (17)	52 (17)	52 (17)	52 (17)
unemployed	7 (5)	7 (6)	7 (6)	13 (7)	7 (7)

(A) No Click, No Noise

reported results are high, indicating lack of plausible deniability for each of these topics. It is concerning that personal circumstances, health status and sexual orientation appear to be the most revealing topics according to our experiments. In the case of the topic *Disabled* there is more cause of concern about (ϵ, m)–Plausible Deniability as the session progresses. On inspection of the associated query script this appears to be again related to the specificity of the queries at each probe step. At the beginning of this script the queries are related to availability of services – for example, *locations of disabled parking* – while later queries are more specific to named conditions – for example, *treatment for spina bifida.*

The topics $\{Location, Payday, Unemployed\}$ appear among the topics of least concern from the perspective of (ϵ, m)–Plausible Deniability. Both of the topics Payday and Unemployed asked queries about availability of social support services whereas queries for the topic Bankrupt asked about availability of paid professional services

such as lawyers and accountants. It is perhaps an illustration of the motivations of a for-profit service where users seeking social supports are of less interest than users seeking expensive paid services.

Overall, measurements of **PDE** in experiments appear to agree with expectations from inspection of the underlying queries. Our results suggest that queries are a strong signal to the observer of user interest, and that estimates from **PDE** appear to distinguish queries that are strongly revealing of specific topic interest from more generic queries where plausible deniability is clearer.

5.4.2 The Effect of Random Noise Injection

We now consider the impact of injecting non-informative queries chosen at random from our popular query list into a user session. We simply refer to these as "random noise" queries. We consider three levels of random noise queries for testing purposes:

"Low Noise" The automation scripts select uninteresting queries uniformly at random from the top-query list and inject a single random noise query after every topic-specific query so that the "signal-to-noise ratio" of sensitive to noise queries in this case is $1:1$.

"Medium Noise" Here the automation scripts inject two randomly selected queries after each topic-specific query for a signal to noise ration of $1:2$.

"High Noise" In this noise-model with the highest noise setting, three random noise queries are injected, resulting in a signal-to-noise ratio of $1:3$.

Note also that the automation scripts were configured to ensure the relevant number of noise queries was always injected *immediately before* each probe query. Our intention was to construct a "worst case" for detection of learning, where probe queries are always separated from sensitive user queries by the specified number of noise queries.

Table 5.3(a-c) shows the measured **PDE** values for Low, Medium and High levels of noise respectively for the "no click" model. The **PDE** values for all levels of noise are similar to the "no click, no noise" baseline values in Table 5.2.

TABLE 5.3: Measured $\widehat{\epsilon}_{*,k}$ for Reference Topic versus Any Other Topic, reported as "max (median)", by Probe Query Sequence

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	54 (45)	54 (45)	54 (45)	54 (45)	54 (45)
bankrupt	16 (9)	56 (50)	52 (39)	54 (45)	56 (45)
diabetes	46 (35)	46 (35)	46 (35)	46 (35)	46 (35)
disabled	9 (3)	9 (8)	9 (7)	33 (7)	40 (32)
divorce	13 (7)	123 (8)	54 (8)	85 (6)	85 (6)
gambling	18 (16)	18 (16)	52 (18)	18 (10)	18 (18)
gay	73 (61)	73 (70)	76 (46)	79 (74)	79 (70)
location	18 (16)	18 (10)	18 (10)	18 (10)	18 (10)
payday	3 (2)	3 (2)	4 (3)	4 (3)	4 (3)
prostate	21 (16)	21 (16)	21 (16)	21 (16)	21 (16)
unemployed	7 (3)	7 (3)	13 (9)	13 (9)	13 (9)

(A) No Click, Low Noise

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	55 (53)	53 (53)	53 (53)	53 (53)	53 (53)
bankrupt	11 (8)	48 (33)	51 (43)	52 (38)	52 (38)
diabetes	38 (38)	38 (38)	38 (38)	38 (38)	38 (38)
disabled	4 (4)	8 (7)	1 (1)	40 (36)	40 (36)
divorce	19 (9)	65 (31)	44 (31)	72 (50)	72 (50)
gambling	18 (16)	18 (17)	18 (18)	31 (3)	18 (10)
gay	89 (68)	89 (69)	88 (64)	93 (73)	93 (64)
location	18 (10)	18 (10)	18 (7)	18 (7)	10 (7)
payday	6 (3)	6 (3)	6 (3)	6 (2)	6 (1)
prostate	32 (14)	32 (14)	18 (13)	18 (13)	18 (13)
unemployed	13 (5)	13 (10)	13 (7)	13 (9)	7 (4)

(B) No Click, Med Noise

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	48 (48)	48 (48)	48 (48)	48 (48)	48 (48)
bankrupt	16 (10)	65 (51)	65 (48)	65 (49)	65 (49)
diabetes	41 (38)	41 (38)	41 (38)	41 (38)	41 (38)
disabled	9 (9)	9 (9)	9 (5)	9 (7)	9 (8)
divorce	41 (27)	75 (38)	56 (22)	75 (29)	75 (29)
gambling	21 (16)	21 (3)	21 (4)	29 (16)	18 (4)
gay	86 (64)	86 (64)	80 (43)	94 (59)	94 (59)
location	10 (10)	8 (8)	8 (8)	18 (13)	18 (13)
payday	3 (2)	4 (2)	4 (2)	4 (2)	3 (1)
prostate	17 (15)	17 (15)	17 (15)	17 (15)	17 (15)
unemployed	10 (7)	13 (7)	13 (7)	13 (7)	13 (7)

(C) No Click, High Noise

Overall, there is no consistent reduction in values across all topics for all noise levels, indicating that injecting random noise queries does not have a consistent effect. In some cases, such as topic Gay, measured values of **PDE** increase for all noise levels indicating that noise injection *worsens* the user's ability to assert (ϵ, m)–Plausible Deniability.

These results indicate that even the "High Noise" model fails to reduce the measured values of **PDE** in a coherent way, so that injecting random noise has not improved plausible deniability significantly with any consistency. We conclude that injection of random noise, even at substantial levels, is not observed to provide a useful defence for plausible deniability in our experiments.

5.4.3 The Effect of Click Strategies

We now consider whether it is possible to disrupt search engine learning by careful clicking of the links on response pages. Intuitively, from the search engine's point of view, clicking on links is a form of active feedback by a user and so potentially informative of user interests. This is especially true when, for example, a user is carrying out exploratory search where their choice of keywords is not yet well-tuned to their topic of interest. Previous studies have also indicated that there is good reason to believe that user clicks on links are an important input into recommender system learning. In Chapter 4, user clicks emulated using the "Click Relevant" click-model were reported to result in increases of 60% – 450% in the advert *content*, depending on the "Sensitive' topic tested.

We consider four different click strategies to emulate a range of user click behaviours:

"No Click" No items are clicked on in the response page to a query. This user click-model does not provide additional user preference information to the recommender system due to click behaviour. This click model is used in the baseline measurements presented in Sections 5.4.1.

"Click Relevant" Given the response page to a query, for each search result and advert we calculate the Term-Frequency (TF) of the visible text with respect to the

keywords associated with the test session topic of interest. When $TF > 0.1$ for an item, the item is clicked, otherwise it is not clicked. This user click-model provides relevant feedback to the recommender system about the information goal of the user.

"Click Non-relevant" TF is calculated for each item with respect to the category of interest for the session in question as for the "Click Relevant" click-model, *except* that items are clicked when the TF score is below the threshold and so they are deemed non-relevant to the topic, that is when $TF \leq 0.1$. This user click-model attempts to confuse the recommender system by providing feedback that is not relevant to the true topic of interest to the user.

"Click All" All items on the response page for a query are clicked. This user click-model gives the recommender system a "noisy" click signal, including clicks on items relevant and non-relevant to the user's information goal.

"Click 2 Random Items" Two items appearing on the response page for a query are selected uniformly at random with replacement and clicked.

In all cases, when uninteresting, noise queries are included in a query session, the relevant user click-strategy is also applied to the result pages of these queries. In this way we hope to avoid providing an obvious signal to the recommender system that might differentiate uninteresting queries from queries related to sensitive topics. Items on the result page in response to probe queries are *not* clicked so that the probe query does not provide any additional information to the recommender system. Measured values of **PDE** are shown in Table 5.4. As random noise injection had no observable effect on measurements of **PDE** for different click models in experiments, only the "No Noise" results are presented here for space reasons.

Taken overall, the results in Table 5.4(a) for the "non-relevant click, no noise" model suggest clicking on non-relevant advert items is the best strategy of the click models tested. The only difference between the "non-relevant click" model and other click models is that non-relevant items *only* are clicked, whereas in other click models it is possible that relevant items are clicked. It seems reasonable to postulate that clicking on relevant items provides "fine-tuned" feedback about user interests which

TABLE 5.4: Measured Plausible Deniability versus any other tested topics as probability of interest, by Probe Query Sequence when the true topic of interest is "Other" with range ($\mu \pm 3\sigma$)

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	59 (50)	59 (50)	59 (50)	59 (50)	59 (50)
bankrupt	16 (8)	65 (42)	65 (36)	59 (40)	54 (38)
diabetes	36 (36)	36 (36)	36 (36)	36 (36)	36 (36)
disabled	7 (4)	7 (4)	9 (9)	40 (4)	40 (7)
divorce	30 (24)	30 (9)	30 (9)	30 (8)	30 (8)
gambling	6 (0)	18 (16)	32 (16)	18 (16)	18 (5)
gay	92 (51)	92 (77)	78 (51)	94 (72)	94 (80)
location	18 (18)	10 (10)	10 (10)	18 (10)	18 (10)
payday	2 (2)	2 (2)	3 (2)	3 (2)	2 (2)
prostate	17 (17)	17 (17)	17 (17)	17 (17)	17 (17)
unemployed	13 (2)	13 (4)	13 (7)	13 (7)	7 (6)

(A) Click Relevant, No Noise

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	18 (5)	22 (12)	26 (5)	31 (13)	32 (6)
bankrupt	57 (3)	53 (36)	50 (34)	43 (33)	48 (36)
diabetes	4 (2)	13 (8)	11 (8)	5 (3)	11 (2)
disabled	5 (2)	6 (2)	9 (3)	29 (10)	26 (8)
divorce	49 (25)	51 (33)	49 (30)	43 (29)	43 (29)
gambling	6 (2)	18 (4)	36 (24)	35 (13)	31 (13)
gay	36 (33)	75 (33)	51 (32)	39 (20)	31 (27)
location	9 (2)	11 (1)	7 (2)	6 (2)	9 (1)
payday	3 (3)	3 (1)	4 (2)	3 (2)	4 (3)
prostate	55 (38)	68 (36)	65 (48)	61 (48)	64 (42)
unemployed	9 (1)	6 (6)	7 (1)	9 (4)	5 (2)

(B) Click Non-relevant, No Noise

TABLE 5.4: (Continued) Measured Plausible Deniability versus any other tested topics as probability of interest, by Probe Query Sequence when the true topic of interest is "Other" with range ($\mu \pm 3\sigma$)

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	66 (57)	66 (57)	66 (57)	66 (57)	66 (57)
bankrupt	51 (42)	51 (42)	51 (42)	55 (46)	56 (46)
diabetes	35 (35)	35 (35)	35 (35)	35 (35)	35 (35)
disabled	9 (9)	9 (9)	9 (9)	31 (31)	31 (31)
divorce	30 (8)	73 (54)	54 (34)	100 (49)	100 (49)
gambling	3 (1)	16 (16)	53 (11)	16 (6)	6 (2)
gay	69 (65)	77 (73)	70 (60)	82 (75)	81 (71)
location	18 (10)	10 (6)	10 (6)	14 (10)	18 (7)
payday	2 (2)	2 (2)	2 (2)	2 (2)	2 (2)
prostate	17 (17)	17 (17)	17 (17)	17 (17)	17 (17)
unemployed	4 (4)	7 (7)	7 (7)	7 (7)	7 (6)

(C) Click All, No Noise

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	50 (12)	27 (9)	26 (9)	36 (10)	33 (11)
bankrupt	5 (3)	43 (33)	39 (37)	36 (35)	38 (35)
diabetes	38 (6)	18 (7)	17 (5)	17 (7)	11 (5)
disabled	2 (1)	4 (1)	5 (3)	39 (25)	40 (25)
divorce	24 (17)	37 (31)	37 (31)	35 (25)	35 (25)
gambling	24 (0)	7 (4)	54 (23)	33 (23)	68 (20)
gay	68 (68)	68 (65)	54 (52)	46 (36)	47 (42)
location	8 (8)	8 (8)	8 (8)	8 (8)	8 (8)
payday	4 (1)	2 (2)	4 (2)	4 (3)	4 (4)
prostate	59 (57)	67 (62)	58 (56)	60 (54)	51 (44)
unemployed	4 (3)	8 (3)	10 (4)	3 (2)	10 (1)

(D) Click 2 Random Items, No Noise

is more informative for the observer. Clicking on non-relevant items may divert attention to a modest degree, but not to the extent of masking the sensitive topic revealed by the query.

Comparing the baseline "No Click" **PDE** observations in Table 5.2 each of the subtables in Table 5.4 shows similar lack of consistency to the noise injection models. In our experiments there is no consistent change observed in **PDE** across topics due to variation in the click patterns tested. As with the noise injection case, there are sporadic increases and decreases in values of **PDE** but the lack of overall consistency makes using click models as a defence impractical.

It would appear in summary, that clicks transmit information to the observer, but not as consistently as does a revealing query. Consequently none of the user click-models tested appear to change the baseline level of plausible deniability associated with the query in a predictable way so that there is no globally discernible pattern with which to construct practical defence tools based on clicks.

5.4.4 The Effect of Proxy Topics

The next privacy protection strategy we consider is the introduction of proxy topics. In this case sequences of queries, with each sequence related to a single proxy topic which is not sensitive for the user but capable of attracting personalised advert content, are injected into a user session. The idea here is that each such sequence of queries emulates a user session where the proxy topic is the topic of interest. In this way we hope to misdirect learning by the search engine of user interests. The results in Section 5.4.2 are relevant here since they suggest that isolated, individual queries – such as randomly selected noise queries – tend not to provoke search engine learning. Our hope is that this can be exploited by inverting the notion of random noise injection so that individual *sensitive* queries are injected as the noise in proxy topic sessions. Isolated sensitive queries will hopefully not provoke learning whereas the larger number of uninteresting proxy sessions will. In this way we can misdirect learning by the observer.

In out tests the following proxy topics are used:

Tickets Searching for tickets for events in a well-known local stadium

Vacation Queries related to a vacation such as flights and accommodation.

Car Searches by a user seeking to trade in and change their car.

and related queries are constructed by selecting related keywords through the same process as was used for the sensitive topics.

TABLE 5.5: Measured Plausible Deniability versus any other tested topics as probability of interest, by Probe Query Sequence when the true topic of interest is "Other" with range $(\mu \pm 3\sigma)$

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
all topics	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

(A) All Click and Noise Models

Proxy topic query scripts where constructed by selecting a sensitive topic, and then selecting an uninteresting proxy topic from the list of 3 proxy topics. Having decided on a sensitive query we wish to issue, we select at least three and no more than four queries related to the proxy topic from a prepared list of proxy topic queries. We next randomly shuffle the order of the selected sensitive and proxy topic queries. In this way there is always a subgroup of at least two proxy topic queries next to each other in each query session. Finally, for testing purposes, we place a probe query before and after each block of 3-4 proxy + 1 sensitive queries to measure changes in **PRI** score. We repeat this exercise using the same proxy topic until a typical query session consisting of 5 probe queries is created.

Data was collected for $2,300$ such proxy topic sessions. This included each of the sensitive topics and each of the click models described in previous sections. The same **PRI** and **PDE** setup, including the same training set, as before was used to process the search results.

Measured detection rates are shown in Table 5.5. The measured probability calculated from **PDE** is 0 for all topics and for all click-models tested. That is, we find it is possible to claim full plausible deniability of interest in all of the topics tested. Since

our detection approach is demonstrated to be notably sensitive to observer learning in earlier sections, we can reasonably infer that this result is not due to a defect in the detection methodology but rather genuinely reflects successful misdirection of the search engine away from sensitive topics.

This result is encouraging, especially in light of the negative results in previous sections for other obfuscation approaches. It suggests use of sequences of queries on uninteresting proxy topics may provide a defence of plausible deniability. The trade-offs for the user include the overhead of maintaining proxy topics and associated queries and the additional resources required to issue proxy topic queries in a consistent way. However since both of these tasks were readily automated during our testing it seems reasonable that these trade-offs could be readily managed by software in a way that is essentially transparent to the user.

Chapter 6

Reasonable Agency - Privacy by Group Identity

6.1 Introduction

Limiting online data collection to the minimum required for specific purposes is mandated by modern privacy legislation such as the General Data Protection Regulation (GDPR) and the California Consumer Protection Act. This is particularly true online where broad collection of personal information represents an obvious concern for privacy. We challenge the view that broad personal data collection is required to provide personalised services.

By first developing formal models of privacy and utility, we show how users can obtain personalised content, while retaining an ability to plausibly deny their interests in topics they regard as sensitive using a system of proxy, group identities. We show that, while some utility loss is an inevitable trade-off for improved privacy, user privacy need not destroy utility when aggregated group information is sufficient for personalisation. From our formal models we implement a *proxy agent* framework we call *3PS* for **P**rivacy **P**reserving **P**roxy **S**ervice, where a user may submit queries though a pool of group identities called *Proxy Agents*. We introduce a privacy preserving algorithm for selecting group identities that users can run locally to find the group identity best matching their interests without revealing their interests.

We end with an extensive experiment on a prototype implementation, using openly accessible data sources, we show that 3PS provides personalised content to individual users over 98% of the time in our tests, while protecting plausible deniability effectively in the face of worst-case threats from a variety of attack types. We test the prototype with Google Search, maintaining consistency with previous chapters. To illustrate potential applicability beyond web search, we also test our prototype on hotel reviews from TripAdvisor and product reviews from Amazon using openly available data.

6.2 Privacy and Threat Model

Our interest is in privacy attacks where an attacker seeks to infer topics of likely interest to users of online systems. An attacker is successful when users are unable

to deny their interest in a topic on the balance of probabilities. Here attackers have access to input–output interactions $\mathcal{Z}_{att,k} \subseteq \mathcal{Z}_k$. By analysing $\mathcal{Z}_{att,k}$ the attacker attempts to estimate topics that are of likely interest to u. The privacy model here is *plausible deniability*, allowing users to reasonably deny that observations are solely associated with topics they deem sensitive. We formalise plausible deniability in our context as follows:

Definition 5 (δ-Plausible Deniability) *A user u can plausibly deny their input–output observations are associated with topics they deem sensitive if* [1]

$$\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_{att,k}) \leq \delta \tag{6.1}$$

where the deniability parameter, δ, is chosen by u and $\mathcal{Z}_{att,k}$ is the background knowledge of an attacker at step k of a session.

This differs from the (ϵ, m)–Plausible Deniability model introduced in (P Mac Aonghusa et al., 2018) where an individual user claimed plausible deniability because an input–output observation from that user could be associated with any of several topics.

Observe that

$$\begin{aligned}
&\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_{att,k}) \\
&\overset{(a)}{\leq} \frac{\mathrm{P}(z \in \mathcal{Z}_k^{u,c} \cap \mathcal{Z}_k)}{\mathrm{P}(z \in \mathcal{Z}_k)} \frac{\mathrm{P}(z \in \mathcal{Z}_k)}{\mathrm{P}(z \in \mathcal{Z}_{att,k})} && (6.2) \\
&\overset{(b)}{=} \frac{\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_k)}{\mathrm{P}(z \in \mathcal{Z}_{att,k} | z \in \mathcal{Z}_k)} && (6.3)
\end{aligned}$$

where inequality (a) follows from the facts that $\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_{att,k}) = \mathrm{P}(z \in \mathcal{Z}_k^{u,c} \cap \mathcal{Z}_{att,k}) / \mathrm{P}(z \in \mathcal{Z}_{att,k})$ and $\mathcal{Z}_{u,k}^{att} \subseteq \mathcal{Z}_k$, and equality (b) follows since $\mathcal{Z}_{att,k} \subseteq \mathcal{Z}_k$. Hence, for δ-plausible deniability to hold it is sufficient that

$$\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_k) \leq \delta \mathrm{P}(z \in \mathcal{Z}_{att,k} | z \in \mathcal{Z}_k) \tag{6.4}$$

[1] In this case $\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_{att,k})$ denotes $\mathrm{P}(\exists m : l_u(\mathcal{Z}_{att,k}(k)) = 1, m \in \{1, 2, \dots\})$.

From (6.4), when an observer has access to all of the observations in the system so that $\mathcal{Z}_{att,k} = \mathcal{Z}_k$ and $\mathrm{P}(z \in \mathcal{Z}_{att,k} | z \in \mathcal{Z}_k) = 1$ then it is sufficient to have $\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_k) \leq \delta$ for δ-plausible deniability to hold. In the case that the observer is able to make observations at a more local level, so that $\mathrm{P}(z \in \mathcal{Z}_{att,k} | z \in \mathcal{Z}_k) = \pi < 1$, then (6.4) implies that $\mathrm{P}(z \in \mathcal{Z}_k^{u,c} | z \in \mathcal{Z}_k) \leq \delta\pi$ is required for δ-plausible deniability to hold. Consequently, unless the user can plausibly deny that they contributed to $\mathcal{Z}_{att,k}$, we have

Observation 6.2.1 (Power of Observers) *Observers represent more powerful threats when they have access to more localised sequences of input–output interactions so there is some trade-off involved in locality versus deniability.*

6.2.1 Comparison with Other Privacy Models

In the group identity setup considered here, the intention is to deny interest by hiding sensitive user activity in the overall activity of users of shared group identifiers. The setup here can be compared with other privacy models. We show briefly how this is done in the cases of two common models of privacy, Differential Privacy, (Dwork, 2006), and Individual Re-identification, (Sweeney, 2000).

Re-identification

Re-identification risk occurs when an attacker, possessing observations $\mathcal{Z}_{att,k}$, can assert that sensitive input–output interactions generated by user u are identified with probability greater than $1 - \epsilon$ for $0 < \epsilon \ll 1$. In other words, when

$$\mathrm{P}(z \in \mathcal{Z}_k^{u,c} \cap \mathcal{Z}_{u,k} | z \in \mathcal{Z}_{att,k}) > 1 - \epsilon \tag{6.5}$$

for $0 < \epsilon \ll 1$.

If δ-plausible deniability holds (6.1) guarantees

$$\mathrm{P}(z \in \mathcal{Z}_k^{u,c} \cap \mathcal{Z}_{u,k} | z \in \mathcal{Z}_{att,k}) \leq \delta \tag{6.6}$$

since $\mathcal{Z}_k^{u,c} \cap \mathcal{Z}_{u,k} \subseteq \mathcal{Z}_k^{u,c}$. Consequently (6.1) prevents re-identification of those sensitive input–output interactions with probability at least $1 - \delta$.

Differential Privacy

Recall that a query mechanism $\mathbf{M} : \mathcal{D} \longrightarrow \mathcal{R}$ satisfies (ϵ, γ)-differential privacy (Dwork, 2006) if, for any two sequences $\mathcal{D}_1, \mathcal{D}_2 \in \mathcal{D}$ of length n differing in one element, and any set of output values $\mathcal{S} \subseteq \mathcal{R}$, we have

$$\mathrm{P}(\mathbf{M}(\mathcal{D}_1) \in \mathcal{S}) \leq e^{\epsilon}\mathrm{P}(\mathbf{M}(\mathcal{D}_2) \in \mathcal{S}) + \gamma \tag{6.7}$$

One important class of mechanisms are those where sequences in $\mathcal{D}$ are first perturbed, e.g., by adding noise, and then queries are answered. It is this approach which is effectively adopted here, with the perturbations being introduced by the randomness of the process generating the input–output interactions. An attacker observes a sequence of input–output interactions and seeks to associate a label with one or more input–output interactions, namely whether or not they were likely to be generated by a target user u and are sensitive for that user. Consider therefore the query $\mathbf{M}_z(\mathcal{Z}_k) = l_u(z)$ i.e. which labels input-output pair z as 1 when it is sensitive for user u and labels it 0 otherwise. This is a worst case query in the sense that it assumes the attacker knows the labelling function l_u, and when this is not the case the labelling accuracy will obviously be degraded. Let $\mathcal{D}_1, \mathcal{D}_2 \in \mathcal{D}$ be two input-output sequences such that $\mathcal{D}_1(k) = \mathcal{D}_2(k)$, $k = \{1, \ldots, n\} \setminus \{j\}$ where $\mathcal{D}_1(k)$ denotes the k'th element of sequence $\mathcal{D}_1$ and similarly for $\mathcal{D}_2(k)$ i.e. sequences $\mathcal{D}_1$ and $\mathcal{D}_2$ are identical except for the j'th element. Mechanism $\mathbf{M}_z$ is (ϵ, γ)-differentially private provided

$$p_1 \leq e^{\epsilon} p_2 + \gamma,\ p_2 \leq e^{\epsilon} p_1 + \gamma \tag{6.8}$$

$$1 - p_1 \leq e^{\epsilon}(1 - p_2) + \gamma,\ 1 - p_2 \leq e^{\epsilon}(1 - p_1) + \gamma \tag{6.9}$$

where

$$p_1 := \mathrm{P}(l_u(\mathcal{D}_1(j)) = 1),\ p_2 := \mathrm{P}(l_u(\mathcal{D}_2(j)) = 1) \tag{6.10}$$

are the probabilities that input-output pair j in sequence $\mathcal{D}_1$, respectively $\mathcal{D}_2$, is labelled sensitive by user u. For sequences satisfying the δ-plausible deniability condition (6.1) we have $p_1 \leq \delta$ and $p_2 \leq \delta$. It can be verified that the (ϵ, γ)-differential privacy conditions (6.8)-(6.9) are therefore satisfied for $\epsilon \geq 0$ and $\gamma \geq \max\{\delta, 1 - e^{\epsilon}(1-\delta)\}$.

6.2.2 Other Linking Attacks

The privacy model described here is concerned with attacks at the application layer that seek to link input–output interactions and associated topics to individual user interests. Linking attacks targeting other vectors are also possible.

One vector for attack is for the service provider to attempt to place cookies or third-party tracking content on the web pages viewed by a user. Within the EU, the GDPR rules require that users be explicitly informed of such actions and must take a positive step to opt in. Hence attempts at such tracking seem like a relatively minor concern. Outside the EU, existing tools for blocking third-party trackers can be used, leaving the setting of unique identifying first party cookies as the main concern. This can be mitigated by standard approaches e.g. by activists maintaining lists of cookies that can be safely used (similar to existing lists of malware sites, trackers and so on) and users blocking the rest.

Another possible vector of attack is to record the IP address of the user browser, and thereby try to link the ratings back to the individual user. However, due to the widespread use of techniques such as VPN or NAT, use of IP addresses as identifiers is unreliable. Users also have the option of using tools such as TOR to further conceal the link between the IP address revealed to the server and the users identity. Such tools are the subject of an extensive literature in their own right and are complementary to the present discussion.

6.2.3 Providing Personalisation

The challenge is to construct an implementation which satisfies Definition 5, thereby providing δ-plausible deniability to users, while also providing an effective personalised service. Our prototype implementation, called 3PS, is based on the Proxy Black-box Model introduced in Chapter 3. The backend system $\mathscr{S}$ is assumed to generate recommendations for a proxy agent based on profiling interests in topics as it would for any other user. In a shared proxy setup users inherit the shared profile of the proxy agent they choose. A user accessing $\mathscr{S}$ via the pool of proxy agents and wishing to obtain good recommendations should therefore choose the proxy agent whose interests most closely match their interests. As an example, Figure 6.1a and Figure 6.1b show the results of issuing the query "cheap flights" through two different proxy agent setups. The choice of query is deliberately intended to trigger commercial advertising for illustrative purposes. In Figure 6.1a the proxy agent is dedicated to Google Search users located in a single country, Ireland. In Figure 6.1b the proxy agent is a web-proxy gateway shared by Google Search users from many countries. The response via the proxy agent in Figure 6.1a contains significantly more content than the proxy agent in Figure 6.1b. Content in Figure 6.1a is also more localised to the region of the user, as illustrated by the Google flight search box outlined in red on the figure and in the Ireland ".ie" domains on other results. Content obtained from the shared proxy agent in Figure 6.1b by contrast reflects the regional settings of the proxy agent rather than the user – in this case, UK currency and websites appear in the adverts.

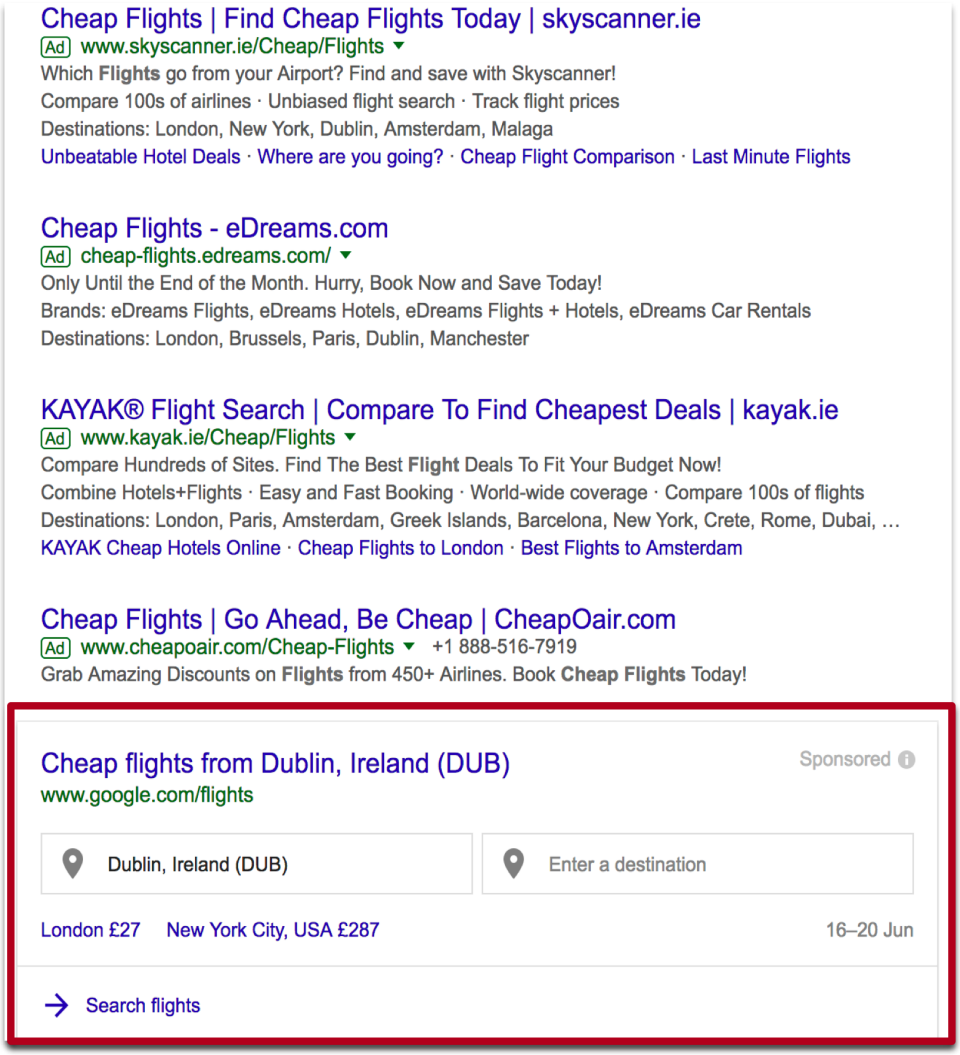
To obtain personalised content, each user chooses a proxy agent closest to their interests in the sense that it is a solution to

$$\begin{aligned} \min_{p \in \mathscr{P}} \sum_{c \in \mathcal{C}} & \left| \mathrm{P}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{u,k}) - \mathrm{P}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{p,k}) \right| \\ \text{s.t.} \quad & \mathrm{P}(z \in \mathcal{Z}^{u,c}_{k} | z \in \mathcal{Z}_{p,k}) \leq \delta \end{aligned} \tag{6.11}$$

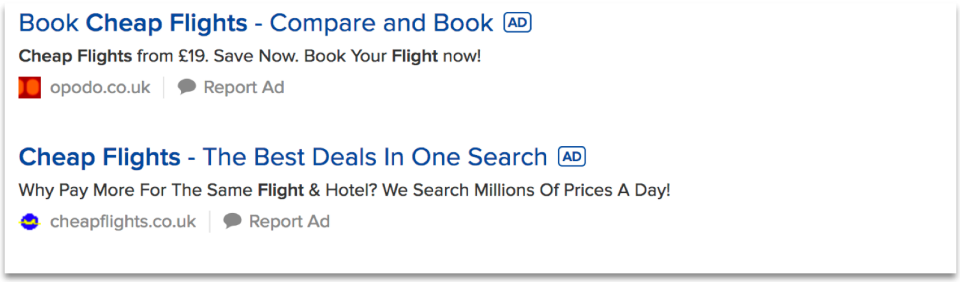
where $\mathcal{Z}_{p,k}$ denotes the input–output interactions of all users with proxy p. The constraint in (6.11) ensures that δ-plausible deniability holds for an observer with

FIGURE 6.1: Examples of Google Search adverts for individual and shared user profiles.

(A) Google Search Adverts for an individual user



(B) Google Search Adverts for a shared proxy user



access to $\mathcal{Z}_{p,k}$.

6.2.4 Threat Models

By varying the observations, $\mathcal{Z}_{att,k}$, available to an observer it is possible to model classes of attack encompassing the system itself and observers with access to more localised background knowledge. We introduce two observer classes we will use in the remainder of this paper.

Privacy Against A Global Observer

A *global observer* denotes an attacker where $\mathcal{Z}_{att,k} = \mathcal{Z}_k$. That is, with access to all of the input–output interactions for the entire system up to the present step k. A global observer does not have knowledge of the user labelling function l_u but can try to cluster the observed input–output interactions to infer topics of likely interest. This class of attacker encompasses the system itself, external parties such as advertising partners and attackers obtaining data by hacking of the system. Provided (6.1) holds for $\mathcal{Z}_{att,k} = \mathcal{Z}_k$ then a user has δ-plausible deniability against global observers.

Privacy Against A Proxy Observer

We also consider a *proxy observer*, namely a global observer who also has knowledge of the set of proxy agents $\mathscr{P}_u \subset \mathscr{P}$ used by user u. Hence, a proxy observer knows that the input–output interactions $\mathcal{Z}_{u,k}$ generated by user u are contained in the subsequence

$$\mathcal{Z}_{att,k} = (z \in \mathcal{Z}_k \,:\, \iota_p(z) = 1, p \in \mathscr{P}_u) \tag{6.12}$$

where indicator function ι_p equals 1 for input–output interactions submitted via proxy p and 0 otherwise. From Observation 6.2.1, a proxy observer is a more powerful attacker than a global observer by having access to more localised data. Provided (6.1) holds with $\mathcal{Z}_{att,k}$ given by (6.12) then a user has δ-plausible deniability against proxy observers.

6.3 Prototype Implementation

In this section we describe an experimental implementation of a backend recommender system accepting text queries as inputs and producing text-based outputs. It is not intended to be a fully working system but rather a proof of concept implemented as software that is sufficient to demonstrate the feasibility of 3PS and to illustrate how personalisation and privacy verification might be implemented. In the prototype implementation the internal state of simulated users, proxy agents and the backend system can be inspected for measurement during test. This allows us to conveniently compare probability estimators during experiments that would be private in a production system.

6.3.1 Personalisation

Expression (3.5) from the Bag–of–Words model can be applied directly to (6.11) so that

$$\begin{aligned}
&\mathrm{P}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{u,k}) - \mathrm{P}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{p,k}) \\
&= \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} \underbrace{\mathrm{P}(z \in \mathcal{Z}^{u,c}_{u,k} | \{\theta^X_i, \theta^Y_j\} \in z)}_{(a)} \\
&\quad \times \left(\underbrace{\mathrm{P}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{u,k})}_{(b)} - \underbrace{\mathrm{P}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{p,k})}_{(c)} \right)
\end{aligned} \tag{6.13}$$

and the minimisation element of (6.11) becomes a calculation over the term labelled (c) in (6.13). We will return to the constraint element of (6.11) later.

Term (6.13)(a) is the only element of the RHS of (6.13) that depends on knowledge of the user labelling function l_u. Since (6.13)(a) and (6.13)(b) do not depend on $\mathcal{Z}_{p,k}$ they can be estimated privately by u. To allow (6.13) to be *privately* by a user, it is sufficient for each proxy agent $p \in \mathscr{P}$ to release the probability distribution (6.13)(c) *publicly*. With this a user can construct (6.13).

Expression (6.13) consists of matrix multiplications of matrices of size $|D^X| \times |D^Y|$. The proxy selection condition in (6.11) can be solved efficiently in practice by

estimating the various probabilities.

6.3.2 Estimating Probabilities

To estimate probabilities in our prototype implementation, user u applies their private labelling function l_u to label each input–output pair $\{x, y\} \in \mathcal{Z}_{u,k}$ for topics in $\mathcal{C}$. Let $\mathcal{U}^c_{u,k}$ and $\mathcal{V}^c_{u,k}$ denote the labelled inputs and outputs of $\mathcal{Z}^{u,c}_{u,k}$ respectively. Apply count-vectorisation to each element of $\mathcal{U}^c_{u,k}$ and $\mathcal{V}^c_{u,k}$ and gather the result into count-matrices $\mathbf{A}_c$ and $\mathbf{B}_c$ of size $|\mathcal{U}^c_{u,k}| \times |D^X|$ and $|\mathcal{V}^c_{u,k}| \times |D^Y|$ respectively. Since $|\mathcal{U}^c_{u,k}| = |\mathcal{V}^c_{u,k}|$, the quantity $\mathbf{N}_c = \mathbf{A}_c^T \mathbf{B}_c$ is of dimension $|D^X| \times |D^Y|$. $\mathbf{N}_c$ is the count co-occurrence matrix of input–output interactions of input–output features in $\mathcal{Z}_{u,k}$ labelled for topic c. The ij–element of matrix $\mathbf{N}_c$, denoted $\mathbf{N}_{c,ij}$, is the co-occurence count of the features $\{\theta^X_i, \theta^Y_j\}$ in $\mathcal{Z}_{u,k}$ labelled for topic $c \in \mathcal{C}$. We apply regular Laplace Smoothing, (Manning et al., 2008), to avoid divide by zero underflows in subsequent computations when there are sparse occurrences of keywords in $\mathcal{Z}_{u,k}$. Laplace smoothing resolves this problem by adding a factor $\lambda_u > 0$ to each keyword count so that $N_{c,ij} \longrightarrow N_{c,ij} + \lambda_u$. The quantity

$$\begin{aligned} \widehat{\mathrm{P}}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}^{u,c}_{u,k}) &= \frac{N_{c,ij}}{N_c} \\ N_c &= \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} N_{c,ij} \end{aligned} \tag{6.14}$$

is then an estimator for $\mathrm{P}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}^{u,c}_{u,k})$. Similarly, an estimator for $\mathrm{P}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{u,k})$ is given by

$$\begin{aligned} \widehat{\mathrm{P}}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{u,k}) &= \frac{N_{ij}}{N} \\ N = \sum_{c \in \mathcal{C}} N_c, \quad N_{ij} &= \sum_{c \in \mathcal{C}} N_{c,ij} \end{aligned} \tag{6.15}$$

and

$$\widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{u,k}) = \frac{N_c}{N} \tag{6.16}$$

is an estimator for the probability of an observation being labelled for topic c.

Let $\mathbf{O}$ have components $O_{ij}(z)$ given by

$$O_{ij}(z) = \begin{cases} 1 \text{ if } \phi_i^X(x) > 0 \text{ and } \phi_j^Y(y) > 0 \text{ for } z = \{x, y\} \\ 0 \text{ otherwise} \end{cases}$$

and define

$$O_{c,ij} := \sum_{z \in \mathcal{Z}_{u,k}^{u,c}} O_{ij}(z), \qquad O_c := \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} O_{c,ij}$$

$$\text{and, } O := \sum_{c \in \mathcal{C}} O_c$$

so that an estimator for $\mathrm{P}(z \in \mathcal{Z}_{u,k}^{u,c} | \{\theta_i^X, \theta_j^Y\} \in z)$ is

$$\widehat{\mathrm{P}}(z \in \mathcal{Z}_{u,k}^{u,c} | \{\theta_i^X, \theta_j^Y\} \in z) = \frac{O_{c,ij}}{O_c} \tag{6.17}$$

and an estimator for $\mathrm{P}(z \in \mathcal{Z}_{u,k} | \{\theta_i^X, \theta_j^Y\} \in z)$

$$\widehat{\mathrm{P}}(z \in \mathcal{Z}_{u,k} | \{\theta_i^X, \theta_j^Y\} \in z) = \frac{\sum_{c \in \mathcal{C}} O_{c,ij}}{O} \tag{6.18}$$

For a proxy agent p, let $\mathcal{U}_{p,k}$ and $\mathcal{V}_{p,k}$ denote the inputs and outputs in $\mathcal{Z}_{p,k}$ respectively. Apply count-vectorisation to each element of $\mathcal{U}_{p,k}$ and $\mathcal{V}_{p,k}$ and gather the result into count-matrices $\mathbf{C}$ and $\mathbf{D}$ respectively of size $|\mathcal{U}_{p,k}| \times |D^X|$ and $|\mathcal{V}_{p,k}| \times |D^Y|$ respectively. The quantity $\mathbf{M} = \mathbf{C}^T\mathbf{D}$, of dimension $|D^X| \times |D^Y|$, is the count co-occurrence matrix of input–output interactions of input–output features in $\mathcal{Z}_{p,k}$, to which Laplace smoothing is applied. We estimate $\mathrm{P}(\{\theta_i^X, \theta_j^Y\} \in z | z \in \mathcal{Z}_{p,k})$ for each proxy agent p as

$$\widehat{\mathrm{P}}(\{\theta_i^X, \theta_j^Y\} \in z | z \in \mathcal{Z}_{p,k}) = \frac{M_{ij}}{M}, \; M = \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} M_{ij} \tag{6.19}$$

and M_{ij} denotes the ij–element of matrix $\mathbf{M}$.

Expressions (6.15), (6.17) and (6.19) can then be combined, to estimate the RHS of (6.13) for each user u.

In our experimental setup, it is convenient to estimate plausible deniability directly from the definition (6.1) as

$$\Delta^{u,c}_{att,k} := \widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{k} | z \in \mathcal{Z}_{att,k}) = \frac{|z \in \mathcal{Z}_{att,k} \,:\, l_u(z) = c|}{|z \in \mathcal{Z}_{att,k}|} \tag{6.20}$$

The probability of user u observing an input–output pair labelled with topic c when accessing $\mathscr{S}$ through proxy agent p is $\mathrm{P}(z \in \mathcal{Z}^{u,c}_{p,k} | z \in \mathcal{Z}_{p,k})$. This is estimated in our experimental setup as

$$\widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{p,k} | z \in \mathcal{Z}_{p,k}) = \frac{|z \in \mathcal{Z}_{p,k} \,:\, l_u(z) = c|}{|z \in \mathcal{Z}_{p,k}|} \tag{6.21}$$

and $\mathrm{P}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{u,k})$, the probability of user u observing an input–output pair labelled with topic c when accessing $\mathscr{S}$ directly is estimated as

$$\widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{u,k}) = \frac{|z \in \mathcal{Z}_{u,k} \,:\, l_u(z) = c|}{|z \in \mathcal{Z}_{u,k}|} \tag{6.22}$$

We measure the estimated *utility loss* incurred by user u as a result of selecting proxy agent p, using (6.21) and (6.22), as

$$\Delta U^{u,c}_{p,k} := \frac{1}{2} \sum_{c \in \mathcal{C}} |\widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{u,k} | z \in \mathcal{Z}_{u,k}) - \widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{p,k} | z \in \mathcal{Z}_{p,k})| \tag{6.23}$$

that is, the total variation between the sensitive topic probability estimator the user would calculate if they used $\mathscr{S}$ directly and the probability estimator of the topic calculated by the proxy agent they used.

6.3.3 User Estimate of Privacy Threat

The challenge for a user in checking (6.1) is that it requires knowledge of $\mathcal{Z}^{u,c}_{k}$ by user u. So that u is required to know the history of input–output interactions for each sensitive topic c for *all* users in the 3PS system.

In the prototype implementation we use the approach that each user u has defined a set, $\Theta^{u,c}_{u,k} \subseteq D^X \times D^Y$, for each sensitive topic c, consisting of input–output keywords whose presence means an input–output observation is labelled as sensitive by u. In experiments, $\Theta^{u,c}_{u,k}$, is selected for each user u and topic c using the training data to choose the keyword pairs for which

$$\Theta^{u,c}_{u,k}(\alpha) = \left\{ \{\theta^X_i, \theta^Y_j\} \,:\, \{\widehat{\mathrm{P}}(z \in \mathcal{Z}^{u,c}_{u,k} | \{\theta^X_i, \theta^Y_j\} \in z) > \alpha \right\} \tag{6.24}$$

where $0 < \alpha \leq 1$ is a parameter chosen using cross-validation.

For each topic c define the associated indicator function over observations $z \in \mathcal{Z}_k$ and $\{\theta^X_i, \theta^Y_j\} \in \Theta^{u,c}_{u,k}(\alpha)$, as

$$\iota^c_\alpha(\{\theta^X_i, \theta^Y_j\}|z) = \begin{cases} 1 \text{ if } \{\theta^X_i, \theta^Y_j\} \in z \\ 0 \text{ otherwise} \end{cases} \tag{6.25}$$

That is, the indicator function labels an observation as sensitive if it contains an input–output keyword pair from $\Theta^{u,c}_{u,k}(\alpha)$ and non-sensitive otherwise. Using the bag-of-words model to combine this with the published estimator $\widehat{\mathrm{P}}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{p,k})$ provided by each proxy agent we get an estimator for $\mathrm{P}(z \in \mathcal{Z}^{u,c}_k | z \in \mathcal{Z}_{p,k})$ given by

$$\widehat{\mathrm{P}}_\alpha(z \in \mathcal{Z}^{u,c}_k | z \in \mathcal{Z}_{p,k}) = \frac{\sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} \iota^c_\alpha(\{\theta^X_i, \theta^Y_j\}|z) \widehat{\mathrm{P}}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{p,k})}{\sum_{c \in \mathcal{C}} \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} \iota^c_\alpha(\{\theta^X_i, \theta^Y_j\}|z) \widehat{\mathrm{P}}(\{\theta^X_i, \theta^Y_j\} \in z | z \in \mathcal{Z}_{p,k})} \tag{6.26}$$

In a real-world setup it is up to the user to decide how to select $\Theta^{u,c}_{u,k}$. For example, the **PRI** tool developed in (Pól Mac Aonghusa et al., 2016) and (P Mac Aonghusa et al., 2018) allows a user to analyse input–output observations for privacy threats and so assess which keyword pairs are more or less revealing of sensitive topics. In this way tools such as **PRI** can provide information to assist in constructing $\Theta^{u,c}_{u,k}$ in a real-world setup.

6.4 Experimental Setup

6.4.1 General Setup

In our experimental setup in this chapter we continue with Google Search as our main source of data. We also report results using the supplementary datasets described in Section 3.2.3. Topics are assigned in the usual way for Google Search and using the topics described in Section 3.2.3 for the supplementary datasets so that $\mathcal{C}$ is defined for each dataset used. Before an experimental run each user and proxy agent simulated during the experiment is allocated a topic of interest from $\mathcal{C}$. When a user or proxy agent is allocated the non-sensitive, catch-all topic c_0 we will say the user or proxy agent is *randomly initialised* meaning that they have no interest in a specific sensitive topic. We call the percentage of proxy agents in $\mathscr{P}$ or users in $\mathscr{U}$ that have been randomly initialised the *diversity* of $\mathscr{P}$ or $\mathscr{U}$. During experiments we will typically report results for 0%, 50% and 100% diversity in $\mathscr{P}$ and/or $\mathscr{U}$.

At the start of each experimental run, each user and each proxy agent is allocated initial data consisting of input–output pairs from the test dataset labelled for their allocated topic of likely interest, referred to as *background knowledge*. Each user and proxy agent in the simulation has a copy of the common dictionaries D^{X} and D^{Y} from $\mathscr{S}$. Next, each user and each proxy agent estimates initial values of the probabilities in Section 6.3.2 from the initial background knowledge using D^{X} and D^{Y}. We refer to these probabilities as the *internal state* of the user or proxy agent. An input query is a keyword in D^{X} drawn from $\Theta^{u,c}_{u,k}(\alpha = 0.5)$ at random by u.

Users select a proxy agent best matching their allocated topic of interest by solving (6.11). When a proxy agent receives an input query from a user it passes it directly to $\mathscr{S}$. Since the set of topics is known to $\mathscr{S}$ in our experiments, $\mathscr{S}$ creates a personalised response by solving $c^* = \arg\max_{c \in \mathcal{C}} \widehat{\mathrm{P}}(z \in \mathcal{Z}^{\mathscr{S},c}_{p,k} | \{\theta^{\mathrm{X}}_i\} \in z)$, to find the topic of maximum likely interest from $\mathcal{C}$ given the input it received, and then selecting an output labelled for c^*. The resulting output is returned to the proxy agent. The input–output interaction pair is added to the background knowledge of the proxy agent and its internal state is updated with new probability estimates. The output is

routed to the requesting user and the same input–output interaction is added to its background knowledge and its internal state and probability estimator are updated.

Background knowledge is not shared among users and proxy agents. When a user switches to a different proxy agent during an experimental run, the user history of input–output interactions does not transfer to the new proxy agent so that individual proxy agents see only the history of interactions from users accessing $\mathscr{S}$ through it. A full reset is performed between test runs by re-initialising the entire setup.

6.4.2 Revealing Keyword Pairs

Test data was preprocessed using the text processing described in Section 6.3.2 to produce dictionaries D^X and D^Y for each dataset. A range of dictionary sizes from 50 to 1000 features was assessed by selecting random subsequences $\mathcal{A}_k \subseteq \mathcal{Z}_k$ and choosing the dictionaries that minimise

$$|\widehat{\mathrm{P}}(z \in \mathcal{A}_k | z \in \mathcal{Z}_k) - \sum_{i=1}^{|D^X|} \sum_{j=1}^{|D^Y|} \widehat{\mathrm{P}}(z \in \mathcal{A}_k^{u,c} | \{\theta_i^X, \theta_j^Y\} \in z) \widehat{\mathrm{P}}(\{\theta_i^X, \theta_j^Y\} \in z | z \in \mathcal{Z}_k)| \tag{6.27}$$

From this we selected $|D^X| = 250$ and $|D^Y| = 500$ for our experiments.

The distribution of keyword pairs in samples drawn from each of the three test datasets is shown in Figure 6.2 by topic. Average values were calculated by taking 10 samples each of $10,000$ items from each of the test datasets. Error bars in Figure 6.2 indicates variance from sampling. In the case of all datasets and for all topics, the co-occurrence frequency of the majority of keyword pairs fall below 0.3. The rarest keyword pairs by topic, and hence the most revealing, have co-occurrence frequencies greater than 0.5. These keyword pairs comprise less than 10% of the total keyword pairs, suggesting that the most revealing keyword pairs form a small subset in the case of all datasets.

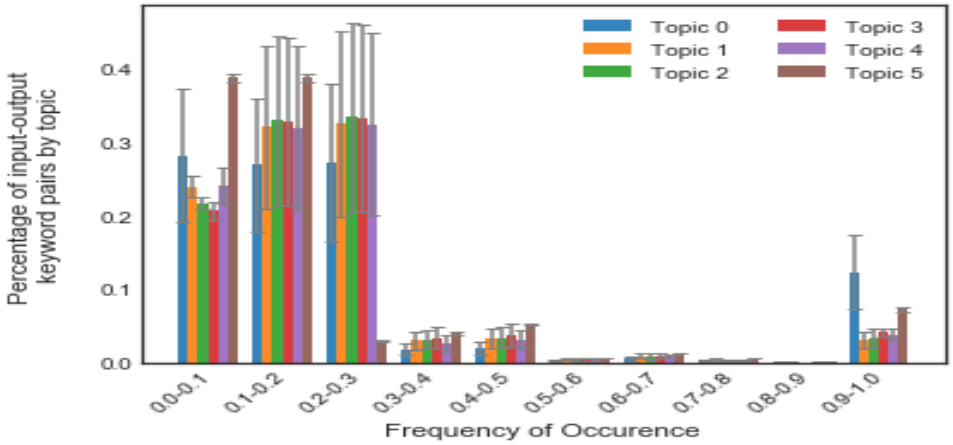


FIGURE 6.2: Frequency of co-occurence of keyword pairs by topic averaged over samples from all datasets, sample variation is shown as error per topic

6.5 Experimental Evaluation

6.5.1 Topic Diversity and User Numbers

We assess the effects of topic diversity and user numbers for the case consisting of a single proxy agent and a single sensitive topic. We denote the senstive topic c_1 so that $\mathcal{C} := \{c_0, c_1\}$ where c_0 is the catch-all topic. A single proxy agent setup means $\mathcal{Z}_k := \mathcal{Z}_{p,k}$ so that results here apply to both proxy and global observers. Tests were repeated with $0\%, 50\%$ and 100% of users having $c_u = c_0$ and the remainder having $c_u = c_1$. We report results for 10, 50 and 100 users for compactness. Results are averaged by dataset and error about the mean is shown as a shaded region. Plausible deniability, from (6.20), and utility loss, from (6.23), averaged over users, are shown in Figure 6.3. Plausible deniability is plotted in the first row and utility loss in the second row.

From (6.1), a user has better plausible deniability for lower values of δ since δ is an upper bound. Our results suggest that increasing user numbers decreases δ and so *improves* plausible deniability but *only* when users have varied interests. Once users have a diverse range of interests, increasing the number of users is observed to accelerate improvement in plausible deniability. For utility loss, increasing volumes of users without specific interests is observed to increase utility loss. When all users of

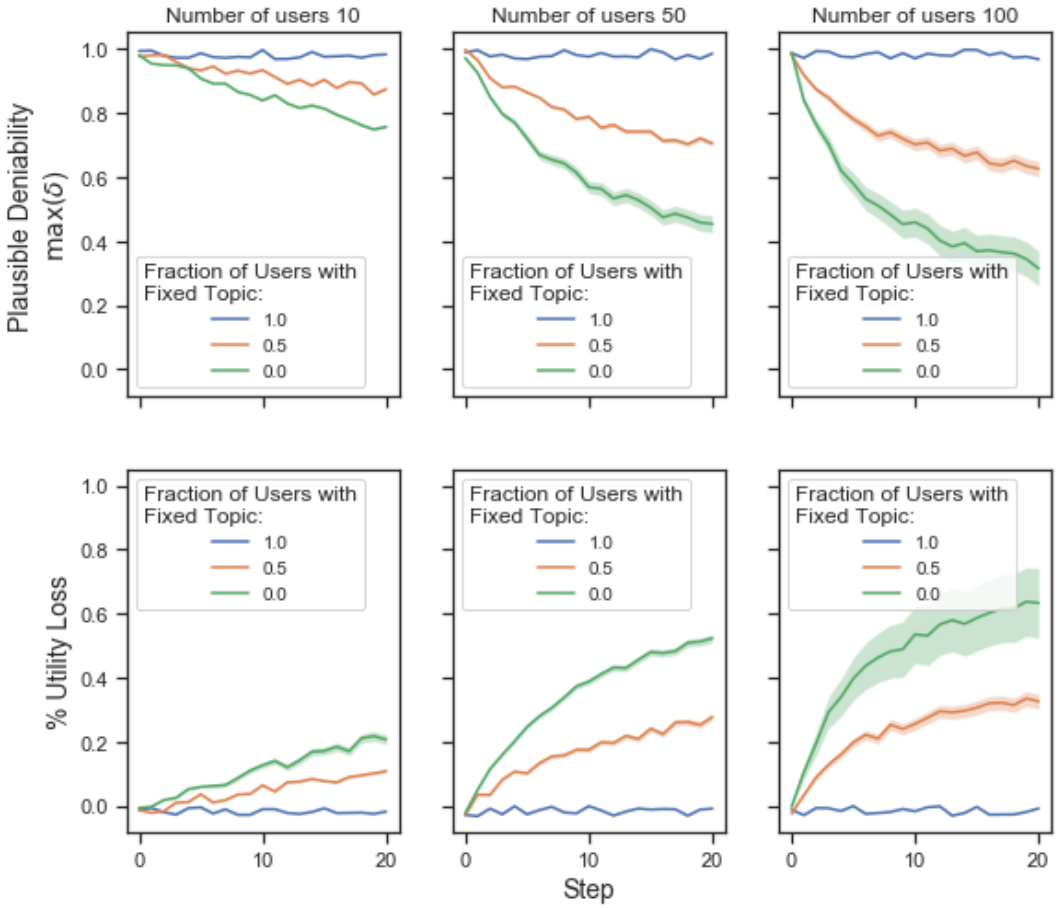


FIGURE 6.3: Effect of topic diversity among users on plausible deniability and utility loss for a single proxy agent with initial fixed topic interest by user diversity and number of users (A step is an input–output pair event)

a proxy agent have no specific topic interests so that diversity is high this is reflected in increased utility loss relative to topic c_1 as one might expect.

6.5.2 Personalisation Performance

In 3PS users select proxies closest to their interests but the responses generated by proxy agents also change as users submit queries via them. We would like this joint selection/update process to converge so as to achieve good personalisation performance. In this section we use our prototype implementation to evaluate this process. Experimental setups with proxy pools of sizes $3 \leq |\mathscr{P}| \leq 30$ and numbers of users $10 \leq |\mathscr{U}| \leq 120$ were configured for each of the test datasets. We initialise proxy agents in $\mathscr{P}$ randomly so that there is no automatic choice of best proxy agent–user

match. Users are allocated a sensitive topic as their target topic from the set of topics in each of the test datasets. Each user applies (6.11) to select a proxy agent best matching their target topic by enumerating each proxy agent in $\mathscr{P}$ in turn. Users only submit queries related to the their allocated topic of interest so that noise due to diverse topic interests of users is controlled in the setup here to focus on convergence properties. Once a proxy agent is selected a user issues a query related to their topic of interest and the internal states of users and proxy agents are updated accordingly. Results are reported as averages over $|\mathscr{P}|$ and $|\mathscr{U}|$ and topic for compactness and shown in Figure 6.4.

The measured accuracy of (6.11) for proxy agent selection is shown in the LHS plot of Figure 6.4. Proxy agent selection is deemed to be accurate when a user chooses a proxy agent whose allocated topic of most likely interest matches the allocated target topic of the user. The RHS of Figure 6.4 is the utility loss, calculated from (6.23), taken at each input–output step. For visual clarity, standard error is shown for the average utility loss over all datasets. Utility loss is high and accuracy is low

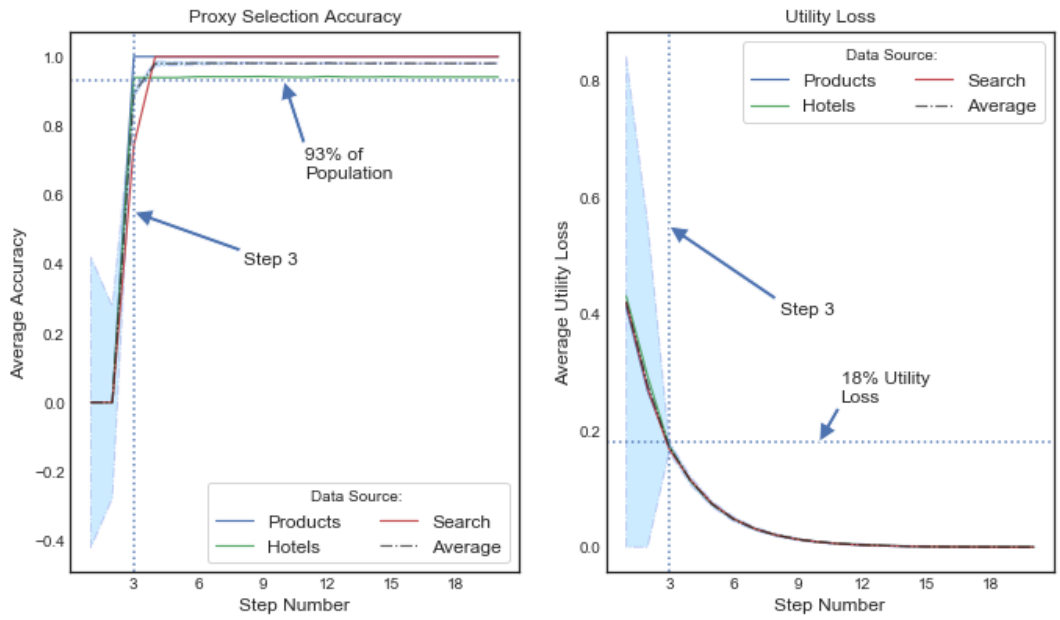


FIGURE 6.4: User to Proxy Agent Selection Accuracy (LHS) and Utility Loss (RHS) averaged over all experimental datasets

initially reflecting the fact that the initial internal state of proxy agents is randomly set. Convergence to the proxy agent with closest interests is observed to happen

quickly for all data sources, achieving at least 93% accuracy for all datasets after 3 iterations with a corresponding average utility loss of 20%. When averaged over all data sources the average accuracy is 98% after 3 input–output steps. The utility loss is also observed to decrease for all topics over time, reaching an average across all datasets of 0.18 after 3 input–output iterations and 0.0002 by iteration 20.

Users are observed to select the correct proxy agent with greater than 90% accuracy, and to reject all proxy agents with 100% accuracy if there is no suitable proxy agent available. Overall, in experiments where the ratio of users to proxy agents was increased from $1:1$ to $30:1$, the utility loss is observed to decrease more slowly as the average number of users attaching to each proxy agent increases. When the ratio of users to proxy agents was $30:1$, for example, the average utility loss on step 1 was 0.67. Convergence to a low utility loss was also observed to be rapid, even at high user to proxy agent ratios, reaching 0.18 ± 0.02 after 4 input–output steps when the user to proxy agent load factor was $30:1$.

The number of topic categories was also varied by regrouping the Hotel dataset. High proxy agent selection accuracy was consistently observed, with accuracy of greater than 90% after step 3. The utility loss was also observed to decrease rapidly to less than 0.20 ± 0.02 after 4 input–output steps, reaching minimum of less than 0.01 by iteration 20 on average over all topics.

Overall, the results suggest that the proxy agent selection method converges rapidly and accurately, providing a high degree of personalisation. Utility loss also decreases rapidly as more topic specific input–output events are observed. This is consistent across the test datasets, and for a range of user–to–proxy agent ratios, suggesting that the proxy agent selection mechanism performs well across a variety of setups.

6.5.3 Plausible Deniability

We next assess the degree of plausible deniability protection available to users with respect to a proxy observer when there are multiple proxy agents. We also assess how diversity in user topic interests influences plausible deniability and utility loss. Since

a proxy observer is at least as powerful as a global observer the results here provide worst-case bounds in the face of a global observer. Experimental setups with proxy pools of sizes $3 \leq |\mathscr{P}| \leq 30$ and numbers of users $10 \leq |\mathscr{U}| \leq 120$ were configured for each of the test datasets. Each proxy agent $p \in \mathscr{P}$ was allocated a topic $c_p \in \mathcal{C}$ as their topic of interest. Each user $u \in \mathscr{U}$ was allocated with a target topic of interest $c_u \in \mathcal{C}$ with setups of $0\%, 25\%, 50\%, 75\%$ and 100% of users having $c_u = c_0$ to model various levels of diversity of topic interests among users. Results are reported as averages over $|\mathscr{P}|$ and $|\mathscr{U}|$ and topic for compactness and shown in Figure 6.5 and Figure 6.6.

In Figure 6.5 we show measurements of estimated level of plausible deniability. We show estimates of $\Delta^{u,c}_{p,k}$ calculated directly from (6.20), together with the values of the estimator (6.26) calculated using $\Theta^c(\alpha)$ as the set of sensitive keywords. To model the situation where the user has partial or censored dictionaries D^X and D^Y in experiments, we show measurements for values for $\alpha \in \{0.25, 0.5, 0.75\}$.

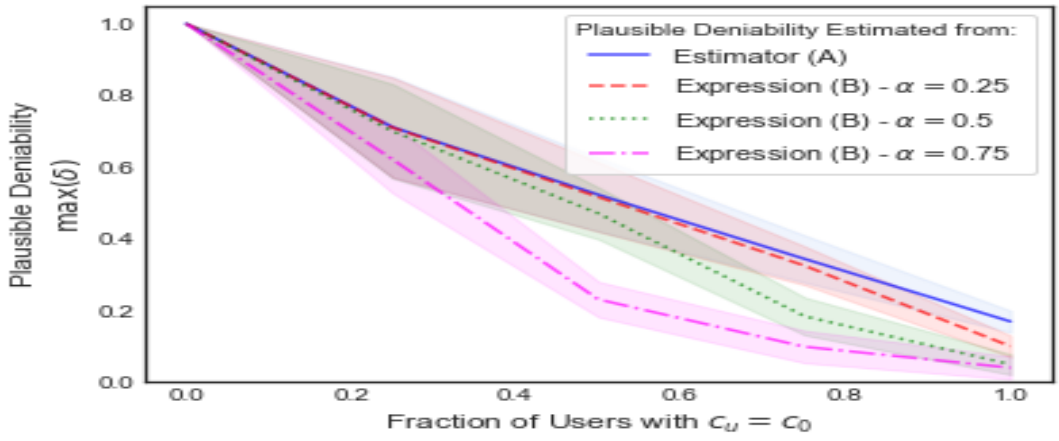


FIGURE 6.5: Plausible deniability by topic averaged over all datasets, topics, sizes of proxy agent pool and number of users. Expression (A) indicates use of (6.20), and Expression (B) use of (6.26) with value of α shown.

The results shown in Figure 6.5 indicate that plausible deniability is observed to improve monotonically as diversity of user interest in topics increases. This is true when either (6.20) or (6.26) are used as estimators, for all values of α. The estimated value using (6.26) is consistently lower than the corresponding estimation from (6.20) for all values of α tested. Figure 6.6 illustrates the trade-off between

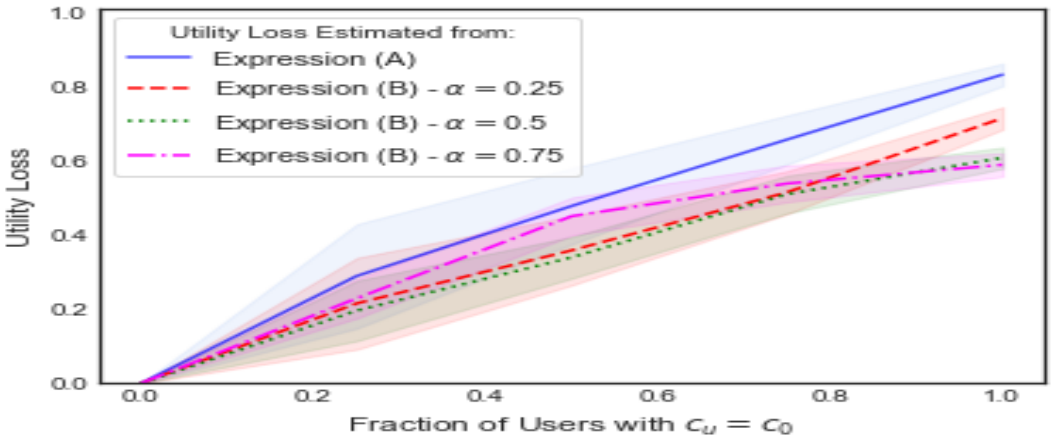


FIGURE 6.6: Utility Loss averaged over all datasets, topics, sizes of proxy agent pool and number of users. Expression (A) indicates use of (6.20), and Expression (B) use of (6.26) with value of α shown.

improved privacy and utility loss. Increasing utility loss is observed in all cases as the fraction of users with diverse topic interests increases as the "signal-to-noise" ratio of coherent interests to random interests decreases. This is observed when either (6.20) or (6.26), for all values of α, are used as estimators. Using (6.26) is observed to underestimate utility loss over all datasets tested. In this case (6.26) should be taken as a best-case guarantee of utility loss and that the actual utility loss will be higher. We note that the ultimate assessment of utility loss is up to the user - if they do not like the personalised content they receive then they can switch to another proxy agent, or stop using the system entirely.

6.5.4 Defending Privacy

We consider a proactive privacy defence strategy of injecting random queries. Between "true" queries a user issues "noise" queries to every member of the proxy agent pool *other* than their selected best matching proxy agent about topics *other* than their allocated topic of interest. This defence is motivated by the observation earlier that increased diversity of topic interests among users is reported to increase plausible deniability. By controlling the level of noise injection we hope to limit the associated utility loss. In practice this kind of injection of obfuscating, uninteresting, "noise" queries can be performed in the background by users.

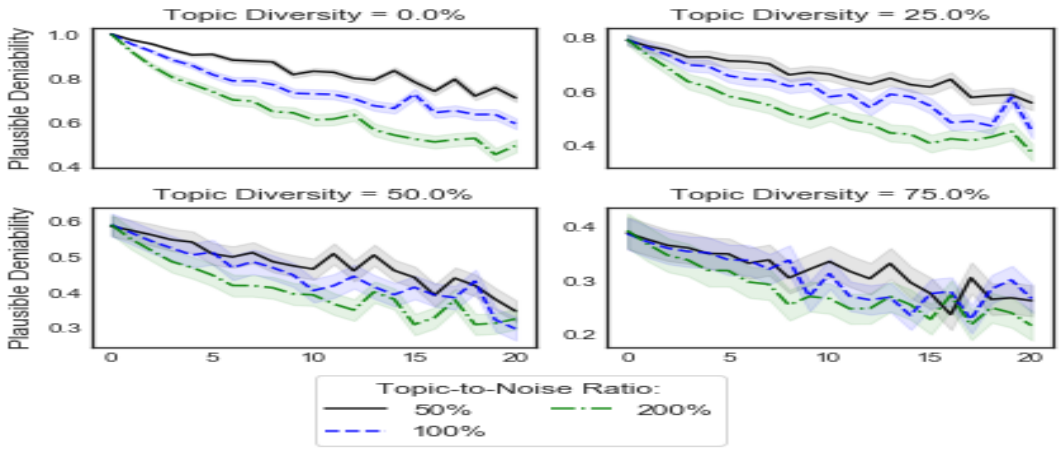


FIGURE 6.7: Plausible deniability for different diversity levels in the proxy agent pool for various topic-to-noise ratios. Results are average by topic and over all datasets.

Experimental setups with proxy pools of sizes $3 \leq |\mathscr{P}| \leq 30$ and numbers of users $10 \leq |\mathscr{U}| \leq 120$ were configured for each of the test datasets. Each proxy agent $p \in \mathscr{P}$ was allocated a topic $c_p \in \mathcal{C}$ as their topic of interest. Each user $u \in \mathscr{U}$ was allocated with a target topic of interest $c_u \in \mathcal{C}$ with setups of $0\%, 25\%, 50\%, 75\%$ and 100% of users having $c_u = c_0$ to model various levels of diversity of topic interests among users. After a sensitive, true input for topic c_u was issued to a chosen proxy agent, a noise query was constructed where input keywords were drawn at random for topics other than the sensitive user topic c_u, and issued to all proxy agents in the pool, except the last chosen proxy agent. To assess the effect of issuing different amounts of noise queries mixed with true queries, "Topic–to–Noise" ratios of $50\%, 100\%$ and 200% were also used. So that, for example, in the case of a true-to-noise ratio of 200%, 2 noise queries are issued for every 1 true queries on average by a user. Results are reported as averages over $|\mathscr{P}|$ and $|\mathscr{U}|$ and topic for compactness and shown for measurements of plausible deniability in Figure 6.7, and for utility loss in Figure 6.8. The first plot in each case shows the case when there is 0% diversity of topic interest in the proxy agent pool as a baseline.

With the random noise injection strategy plausible deniability against a proxy observer improves steadily during an experimental run for all levels of topic diversity in our experiments. For all levels of topic diversity, adding more noise results in faster

improvement in plausible deniability as expected intuitively. As the topic diversity in the proxy agent pool increases, less random noise is required to produce the same changes in plausible deniability as do larger random noise levels. Intuitively this is to be expected since topic diversity is an indication of the variation in topic interests among users. Standard error in the mean, shown as shaded regions is small, indicating that improved plausible deniability is observed with high confidence for all datasets. Utility loss, shown in Figure 6.8, increases initially and achieves stable levels after

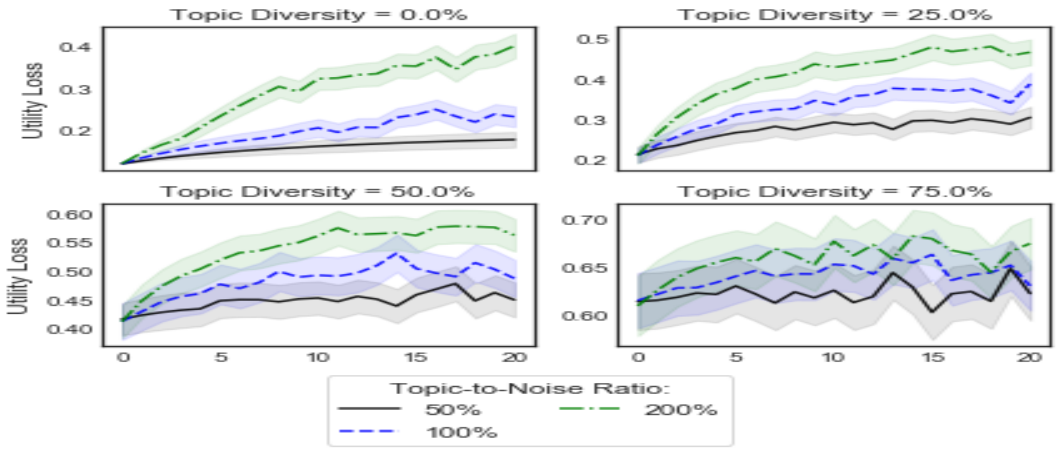


FIGURE 6.8: Utility loss for different diversity levels in the proxy agent pool for various topic-to-noise ratios. Results are average by topic and over all datasets.

$5-10$ input–output steps with the cases where topic diversity is highest reaching a stable level quickest. Standard error is small in the case of all datasets, suggesting the average values plotted reflect expected behaviour with high confidence.

The plausible deniability and utility loss results for 0% topic diversity are a worst-case. Even in this case the utility loss at levels of random noise up to 100% the utility loss is 20% after 20 steps - compared with an improvement in plausible deniability from 100% to 60% on average. As topic diversity increases the improvements in plausible deniability are larger than the associated utility losses in all cases. Taken overall, our results suggest that the benefits to privacy of adopting a strategy of random noise injection outweigh the associated utility losses, with the greatest benefits occurring when the privacy risk from low topic diversity is highest. Run as a background task, injecting random noise by all users in a controlled manner provides a mechanism for

enforcing effective topic diversity in the proxy agent pool with corresponding benefits for privacy.

6.6 Discussion

The results of the random proxy injection defence in our experiments suggest that once a user is alert to diversity, the 3PS setup can provide balance of probability plausible deniability of topic interests. The method of choosing revealing keyword pairs outlined in Section 6.3.3 provides a practical bound on plausible deniability and is straightforward to apply in practice. In a production setting a browser plug-in could automatically suggest new keywords for inclusion by the user in local keyword dictionary extensions.

To apply (6.1) in practice, a user also needs a way of confirming that proxy agents are being truthful about the probability estimators it publishes. The notion of *probe queries*, introduced in Chapter 4, allows a user to test the behaviour of black-box systems without revealing sensitive interests. By checking input–output interactions users can label the observation as sensitive or not and adjust their view of revealing keywords. The techniques introduced in Chapter 4 can be used to check for observations that vary from the values expected from (6.26), indicating possible concerns with the estimators distributed by that proxy agent.

Choosing $\Theta^{u,c}_{u,k}$ to estimate plausible deniability requires care. From (6.26) it follows that

$$\widehat{\mathrm{P}}_{\alpha}(z \in \mathcal{Z}^{u,c}_{k} | z \in \mathcal{Z}_{att,k}) < \widehat{\mathrm{P}}_{\beta}(z \in \mathcal{Z}^{u,c}_{k} | z \in \mathcal{Z}_{att,k})$$

when $0 < \alpha < \beta \leq 1$. Choosing $\alpha = 1$ to include as many keywords as possible in $\Theta^{u,c}_{u,k}$ is the safest threat detection strategy in our setup here. We have assumed here that there is no incentive for dishonesty neither is there any malicious poisoning nor accidental corruption in our setup. In a real-life, production setup when D^X or D^Y are partially complete, poisoned or deliberately censored, a user may choose any input–output keywords for $\Theta^{u,c}_{u,k}$. We note that the techniques introduced in (P Mac

Aonghusa et al., 2018; Pól Mac Aonghusa et al., 2016) provide tools to test when input–output keywords indicate privacy concerns that could be adapted to assist a user with constructing $\Theta^{u,c}_{u,k}$.

While our experiments suggest that 3PS can provide acceptable levels of plausible deniability with low utility loss, our results also emphasise the importance of maintaining adequate vigilance to prevent interests in sensitive topics from leaking and taking care to avoid overly revealing content that might compromise plausible deniability when user interests are known.

Our implementation is a prototype and so results should be taken as a first step. Scaling 3PS to a full product environment will pose engineering and business challenges. Our approach is intended to be easily integrated into the technology stack of a search engine. However introducing a group identity as an intermediary will result in disruption to personalisation since personalisation in 3PS is based on the cumulative profile of group identities rather than specific details of individual users. In effect the web search engine risks loosing utility by being able to serve less personalised content to group identities. The challenge in the wild that our prototype is unable to answer is whether a balance between the loss of utility and improved privacy of users is a viable tradeoff in a production setting.

Chapter 7

Conclusions

7.1 Discussion

In the principles proposed in Chapter 1 we introduced three aspects of personal privacy as a guiding structure for this thesis. In Chapter 4 we showed how users of search engines could detect potential privacy threats from personalisation. In Chapter 5 we built on the work in Chapter 4 to show how users could estimate risks to privacy from plausible deniability. Together the work in these chapters provides a set of tools allowing users to monitor and assess aspects of personal privacy. In Chapter 6 we addressed our third privacy principle through an architecture of group user identities, indicating that enabling users to assert agency over their personal privacy is possible.

Although the systems considered here are complex, our work suggests that personal privacy need not be difficult or opaque for users. The black-box models we employ enable a formal approach to user privacy while allowing the implementation details of systems to remain private. Our experimental results indicate that, even with the assumption of a black-box system, the **PRI** and **PDE** techniques allow users to detect and assess potential risks to their online privacy without tipping off the back-end system. The ability to detect and verify quietly, without significant impact on the back-end system, also allows our tools to work in a way that is compatible with fair-usage of these systems.

Both of the **PRI** and **PDE** techniques were implemented with readily available software tools and verified using accessible data from search engines. The 3PS prototype implementation was built from openly available software tools and was specifically designed to minimise retro-fitting impact on existing systems. We believe the approach of using open software and verifying with readily available data to be a feature often missing in traditional privacy research where concerns over data disclosure limit access to potentially sensitive test data sources.

While our work is intended to be minimally disruptive to the in situ technologies involved, there is a trade-off between utility – the degree of exact personalisation – and privacy – the degree to which a user can deny interest in private topics. In the case of 3PS, for example, users adopt group identities so that personalisation will be

based on the group profile rather than specific knowledge of the user. In this sense, privacy defences pose a risk to the underlying business model assumption that more specific personalisation implies higher click-through revenue. It remains for future research to determine if this trade-off is acceptable to systems.

Our results in Chapter 6 with 3PS show that, in fact, much less personal data collection is required for adequate personalisation than is generally believed. This has significant implications for online providers in light of legislation such as GDPR that requires data to be limited to that which is proportionate to the purpose of collection. The fast-convergence and high accuracy of the proxy agent selection method, observed in our experiments indicate that 3PS can provide a safe and scalable solution that requires little retro-fitting to work with existing systems. It suggests that Internet system providers can adapt the techniques here to provide privacy preserving services for users at little or no cost of disruption.

In beginning this work our view was that detecting personalisation threats from adaptation would require complex solutions given the complexity of the underlying systems. In fact, our results indicate that evidence of adaptation is easy to find and to assess for privacy threats. The realisation that personalisation is mandated in commercial online systems to maximise shareholder value means that such systems are forced to reveal their hand despite their black-box nature. This suggests that there is an "Elephant in the Room" for privacy in the face of sophisticated, modern, commercial internet systems. Namely, focusing on personal de-identification is to risk missing the larger threat of distinguishability. Our observation in Chapter 4 that such sensitive topic profiling persists, even for anonymous users, helps to further underline the nature of the privacy threat.

7.2 Future Research

As previously mentioned, detection and assessment of privacy concerns are related to inference in machine learning, and more generally to fairness, accountability and transparency in machine learning, (FATML, 2019). There is scope for future research

to further investigate online privacy within the broader context of verification fairness and analysis of inference in machine learning, (Olhede et al., 2016, 2018).

Our work excludes the situation where the system does not reveal its hand through personalised content. The latter could happen when the system is not capable or is unwilling to personalise its output. For example, when the real motive is data collection for undisclosed background processing or security analysis. These specific situations are left for future research, and perhaps best addressed by the law and through strong and active governance rather than through technology alone.

Our focus here is on privacy concerns arising from inference by the search engine resulting from explicit user web search interactions. Online systems, including search engines, gather data from many sources to produce personalised content. Direct identification techniques, such as IP tracing or browser finger-printing are outside the scope of our current analysis. Implicit profiling effects due to Geo-location, for example, also effect personalised content. Investigating how explicit, implicit and other data collection techniques interact, especially in growth platforms like mobile devices, is another area for future research.

Search engines are a convenient and openly available source of personalised content, but not the only online service that profiles users. A significant portion of modern online systems profile users to boost commercial return through improved personalised content. In Chapter 6 we extended our analysis by including openly available examples data sources for TripAdvisor and Amazon to illustrate how the techniques we develop can be extended beyond search engines. Our work here, though promising, is a prototype and much work remains to be done to demonstrate the feasibility of our results in a production setting.

Future avenues of research include: looking beyond search engines to other recommender systems where content types other than adverts may provide better content for adaptation detection in the case of other recommender systems; extending our techniques to incorporate more complex user interaction models; constructing effective user privacy defences by exploiting observations of topic similarity and confusion encountered in our experiments, and, investigating how our tools perform performs

for different models of contextual advert selection such as semantic or sense-based techniques that employ non-keyword based selection techniques to select adverts. The Natural Language Processing Bag-of-Words model we use is among the simplest possible that facilitates obtaining useful experimental results. More sophisticated language models, for example using n-grams or word embedding, will likely improve the capabilities of the tools. We focus on text-based advert content appearing on web search result pages. Other sources of personalised content, such as images, and pop-up suggestion cards on mobile devices, could also be investigated to provide finer-grain insight into individual privacy.

7.3 Concluding Remarks

We conclude with a note of caution to the user. Our experiments indicate that online systems such as search engines are able to identify user interests with high accuracy, exploit multiple signals, filter out uninteresting noise queries and adapt quickly when topics change. Furthermore learning appears to be sustained over the lifetime of query sessions. The power and sophistication of these systems make designing a robust defence of user privacy non-trivial.

Overall our results point towards a situation, where online system capability is continuously evolving in response to technical advances and developments in user behaviours. In this setting, even if our technologies were to become widely deployed then we can reasonably expect search engines to respond with more sophisticated learning strategies. Our results also point towards the fact that explicit input from the user, such as search queries, plays a key role in search engine learning. While perhaps obvious, this observation reinforces the user's need to be circumspect about the queries that they ask if they want to avoid search engine learning of their interests.

The importance of personal responsibility concerning privacy is pervasive in our work here. Our results strongly suggest that the need to maintain a level of engagement and alertness with respect to individual online privacy is an unavoidable feature of online existence. However good privacy technologies become, we should not become complacent. The decision to engage and to take action is, unavoidably,

a personal responsibility and personal judgements regarding risk seem intrinsic to discussions of privacy.

In conclusion we view this work as a starting point towards practical user privacy in the face of ever-evolving and more powerful online systems. The results presented here are relevant for the billions of users of everyday online systems, policy-makers and privacy watchdogs and, of course, for online system providers under increased scrutiny to demonstrate their commitment to improved user privacy.

Bibliography

Ahmad, Wasi Uddin et al. (2016). "Topic Model based Privacy Protection in Personalized Web Search". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*. SIGIR '16. New York, NY, USA: ACM, pp. 1025–1028.

Aïmeur, Esma et al. (2008). "Alambic: A privacy-preserving recommender system for electronic commerce". In: *International Journal of Information Security* 7.5, pp. 307–334.

al, Balebako et (2012). "Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising". In: researchgate.net.

Alessandro Acquisti et al. (2015). "Privacy and Human Behavior in the Information Age". In: *Science* 347.6221.

Alle, Robert S. et al. (1966). "Data Center Plan called Privacy Invasion". In: *The Lewiston Daily Sun*.

Andreou, Athanasios et al. (2018). "Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations". In: *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

Andriole, Steven (2017). "Already Too Big To Fail - The Digital Oligarchy Is Alive, Well (and Growing)". In: *Forbes.com*.

Aonghusa, P Mac et al. (2018). "Plausible Deniability in Web Search; From Detection to Assessment". In: *IEEE Transactions on Information Forensics and Security* 13.4, pp. 874–887.

Aonghusa, Pól Mac et al. (2016). "Don't Let Google Know I'm Lonely". In: *ACM Transactions on Privacy and Security* 19.1, pp. 1–25.

Aonghusa, Pól Mac et al. (2018). "3PS - Online Privacy through Group Identities". In: *Submitted - IEEE Transactions on Information Forensics and Security*.

Apple, Differential Privacy Team (2017). "Learning with Privacy at Scale". In: *Ml* 1.8, pp. 1–25.

Arampatzis, Avi et al. (2013). "A versatile tool for privacy-enhanced web search". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7814 LNCS. Springer, pp. 368–379.

Arampatzis, A et al. (2011). "Enhancing deniability against query-logs". In: *Advances in Information Retrieval*. Vol. 6611. ECIR'11. Berlin, Heidelberg: Springer-Verlag, pp. 67–76.

Balsa, Ero et al. (2012). "Obfuscation-Based Private Web Search Introduction to Obfuscation-Based Private Web Search". In: *Security and Privacy (SP), 2012 IEEE Symposium on*. 05. IEEE, pp. 491–505.

Bambauer, Jane R. et al. (2013). "Fool's Gold: an Illustrated Critique of Differential Privacy". In: *Vanderbilt Journal of Entertainment & Technology Law, 2014 (Forthcoming)* 16.4, Paper No. 13–47.

Batmaz, Zeynep et al. (2016). "Randomization-based Privacy-preserving Frameworks for Collaborative Filtering". In: *Procedia Computer Science* 96.C, pp. 33–42.

Bennett, Colin J (2011). *Privacy in Context: Policy and the Integrity of Social Life*. Vol. 8. 4. Stanford, CA, USA: Stanford University Press, pp. 541–543.

Bielova, Nataliia (2017). "Web Tracking Technologies and Protection Mechanisms". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17*. CCS '17. New York, NY, USA: ACM, pp. 2607–2609.

Binns, Reuben et al. (2018). "Measuring third party tracker power across web and mobile". In: *arXiv:1802.02507*.

Bird, Steven et al. (2009). *Natural Language Processing with Python*. 1st. January 2009. O'Reilly Media, Inc., p. 479.

Blue, Violet (2016). "Fake privacy gadgets, from Anonabox to Sever: Fighting a strange and profitable epidemic". In: *ZDNet*.

Boutet, Antoine et al. (2016). "Privacy-preserving distributed collaborative filtering". In: *Computing* 98.8, pp. 827–846.

Boyd, Dana (2012). "Debate: Networked Privacy". In: *Surveillance & Society*. Personal Democracy Forum 10.3/4, pp. 348–350.

Burgess, Matt (2018). "The tyranny of GDPR popups and the websites failing to adapt". In: *https://www.wired.co.uk/article/gdpr-cookies-eprivacy-regulation-popups*.

Celis, L Elisa et al. (2019). "Controlling Polarization in Personalization: An Algorithmic Framework". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, pp. 160–169.

Cummings, Rachel et al. (2014). "The Empirical Implications of Privacy-Aware Choice". In: *Operations Research* 64.1, pp. 67–78.

Damm, Andrew van (2019). "Google image search results for CEOs — and most jobs — dominated by men". In: *The Washington Post*.

Datta, Amit et al. (2015). "Automated Experiments on Ad Privacy Settings". In: *Proceedings on Privacy Enhancing Technologies* 2015.1.

Datta, Anupam (2014). "Privacy through accountability: A computer science perspective". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8337 LNCS. Springer, pp. 43–49.

Day, Violet (2018). "Charlatans: The new wave of privacy profiteers". In: *http://www.zdnet.com/*.

Domingo-Ferrer, Josep et al. (2009). "User-private information retrieval based on a peer-to-peer community". In: *Data and Knowledge Engineering* 68.11, pp. 1237–1252.

Dwork, Cynthia (2006). "Differential privacy". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4052 LNCS, pp. 1–12.

Economist, The (2017). "The "free" economy comes at a cost - {Free} exchange". In: *The Economist*.

EFF (2018). “Privacy Badger”. In: *Https://Www.Eff.Org/Privacybadger/*, Abruf am 2018–04–02.

Englehardt, Steven et al. (2016). “Online Tracking”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*. ACM, pp. 1388–1401.

Equal Opportunity Commission (2014). “Types of discrimination”. In:

Erkin, Z. et al. (2011). “Efficiently computing private recommendations”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, pp. 5864–5867.

— (2010). “Privacy enhanced recommender system”. In: *Thirty-first Symposium on Information Theory in the Benelux*. IEEE Benelux Information Theory Chapter, pp. 35–42.

Erlingsson, Úlfar et al. (2014). “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”. In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM, pp. 1054–1067.

European Union (2010). *Charter of Fundamental Rights of the European Union*. Vol. 53. Brussels: European Union, p. 380.

— (2016). “Regulation 2016/679 of the European parliament and the Council of the European Union of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/”. In: *Official Journal of the European Communities* 59.05, pp. 1–88.

FATML, ACM (2019). “Fairness, Accountability, and Transparency in Machine Learning”. In: *http://www.fatml.org/*.

Fredrikson, Matthew et al. (2011). “REPRIV: Re-imagining content personalization and in-browser privacy”. In: *Proceedings - IEEE Symposium on Security and Privacy*. SP '11. Washington, DC, USA: IEEE Computer Society, pp. 131–146.

Friesel, Rob (2014). “PhantomJS Cookbook”. In:

Google Trends (2018). “Google Trends”. In: *https://www.google.com/trends/*.

Guha, Saikat et al. (2010). "Challenges in measuring online advertising systems". In: *Proceedings of the 10th annual conference on Internet measurement - IMC '10*. IMC '10. New York, NY, USA: ACM, p. 81.

Hannák, Anikó et al. (2017). "Measuring Personalization of Web Search". In: *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 527–538.

Helbing, Dirk et al. (2017). "Will Democracy Survive Big Data and Artificial Intelligence?" In: *Scientific American*.

Heymans, Maureen (2009). "Introducing Google Social Search: I finally found my friend's New York blog!" In: *The Official Google Blog*.

Hofmann, Thomas et al. (1998). *Statistical Models for Co-occurrence Data*. Tech. rep. Cambridge, MA, USA.

Holding BV, Surfboard (2018). "StartPage - Privacy Policy". In: *https://www.startpage.com/*.

Hongning Wang, Yue Lu et al. (n.d.). "Data for Latent Aspect Rating Analysis". In: *The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* ().

Howe, Daniel C et al. (2009). "TrackMeNot". In: *mrl. nyu. edu/dhower/trackmenot*.

Idris, Ivan (2012). *NumPy Cookbook*. Packt Publishing, p. 226.

Inc, DuckDuckGo (2018). "DuckDuckGo Privacy Statement". In: *https://duckduckgo.com/*.

Jenkins Jr., Holman (2010). "Google and the Search for the Future, an interview with Eric Schmidt". In: *The Wall Street Journal*.

Lecuyer, Mathias et al. (2014). "XRay: Enhancing the Web's Transparency with Differential Correlation". In: *Proceedings of the 23rd USENIX Conference on Security Symposium*. SEC'14. Berkeley, CA, USA: USENIX Association, pp. 49–64.

Li, Ninghui et al. (2007). "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *IEEE 23rd International Conference on Data Engineering*. IEEE, pp. 106–115.

Limaye, M.G. (2009). *Software Testing*. McGraw-Hill Education (India) Pvt Limited. ISBN: 9780070139909.

Ling, Zhen et al. (2012). "A new cell-counting-based attack against tor". In: *IEEE/ACM Transactions on Networking* 20.4, pp. 1245–1261.

Lioma, Christina et al. (Feb. 2018). "To Phrase or Not to Phrase - Impact of User versus System Term Dependence Upon Retrieval". In: *Data and Information Management*.

Machanavajjhala, Ashwin et al. (2006). "L-Diversity: Privacy beyond k-anonymity". In: *Proceedings - International Conference on Data Engineering* 2006.1, p. 24.

Manning, Christopher D. et al. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

McSherry, Frank et al. (2009). "Differentially private recommender systems". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*. KDD '09. New York, NY, USA: ACM, p. 627.

Monteiro, Ana (IAPP) (2019). "First GDPR fine in Portugal issued against hospital for three violations". In: *The Privacy Advisor*.

Mozilla (2016). "Mozilla Lightbeam". In: *https://www.mozilla.org/en-US/lightbeam/*.

Narayanan, Arvind et al. (2017). "The Princeton Web Transparency and Accountability Project". In: *Transparent Data Mining for Big and Small Data*. Springer, pp. 45–67.

Naudts, Laurens (2018). "Towards Accountability: The Articulation and Formalization of Fairness in Machine Learning". In: *IFIP Summer School on Privacy and Identity Management" Fairness, Accountability and Transparency in the Age of Big Data"(20-24 August 2018)(submitted for pre-proceedings)*.

NetApplications (2018). "Search Engine Market Share 2018". In: *NetMarketShare.com*.

Nissim, Kobu et al. (2018). "Is privacy privacy?" In: *Digital Access to Scholarship at Harvard (DASH)*.

Olhede, SC et al. (2016). "Towards the Science of Security and Privacy in Machine Learning". In: *CoRR* abs/1611.03814. arXiv: `1611.03814`.

— (2018). "The growing ubiquity of algorithms in society: implications, impacts and innovations". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128, p. 20170364.

Oliphant, Travis (Jan. 2006). *Guide to NumPy*.

Olteanu, Alexandra et al. (2018). "A Critical Review of Online Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: ACM, pp. 785–786. ISBN: 978-1-4503-5581-0.

Osmani, Addy et al. (2018). "Speed is now a landing page factor for Google Search and Ads". In: *Google Developer Updates*.

Pan, Xiang et al. (2015). "I Do Not Know What You Visited Last Summer: Protecting users from stateful third-party web tracking with TrackingFree browser". In: *Proceedings 2015 Network and Distributed System Security Symposium*.

Panjwani, Saurabh et al. (2013). "Understanding the privacy-personalization dilemma for web search". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. CHI '13. New York, NY, USA: ACM, p. 3427.

Pariser, Eli (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group, The, p. 304.

Peddinti, Sai Teja et al. (2011). "On the limitations of query obfuscation techniques for location privacy". In: *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*. ACM, p. 187.

Pedregosa, Fabian et al. (2012). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peng, Daniel et al. (2010). "Large-scale incremental processing using distributed transactions and notifications". In: *Proceedings of the 9th USENIX conference*. Vol. 2006, pp. 1–15.

Petit, Albin et al. (2014). "Towards efficient and accurate privacy preserving web search." In: *Proceedings of the 9th ACM Workshop on Middleware for Next Generation Internet Computing*. MW4NG '14 February. New York, NY, USA: ACM, p. 1.

Petit, Albin et al. (2015). "PEAS: Private, efficient and accurate web search". In: *Proceedings - 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2015*. Vol. 1, pp. 571–580.

Petit, Albin et al. (2016). "SimAttack: private web search under fire". In: *Journal of Internet Services and Applications* 7.1, p. 1.

Pujol, Enric et al. (2015). "Annoyed Users". In: *Proceedings of the 2015 ACM Conference on Internet Measurement Conference - IMC '15*. Vol. 2015-Octob, pp. 93–106.

Ram, Aliya et al. (2019). "France fines Google €50m in test for EU's new data laws". In: *The Financial Times*.

Ramakrishnan, Naren et al. (2001). "Privacy Risks to Straddlers in Recommender Systems". In: *IEEE Internet Computing* 5.December, pp. 1–16.

Razaghpanah, Abbas et al. (2018). "Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem". In:

Richardson, Leonard (2016). *Beautiful Soup Documentation*, pp. 1–72.

Sánchez, David et al. (2013). "Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines". In: *Information Sciences* 218, pp. 17–30.

Santa, Isabella (2010). "The new users' guide: How to raise information security awareness." In: *Information Security*, pp. 1–140.

Shen, Xuehua et al. (2007). "Privacy protection in personalized search". In: *ACM SIGIR Forum*. Vol. 41. 1. ACM, pp. 4–17.

Singhal, Amit (2012). "Search, plus Your World". In: *Google Inside Search Blog*, pp. 1–5.

Sivakorn, Suphannee et al. (2016). "I'm not a human: Breaking the Google reCAPTCHA". In: *Black Hat ASIA 2016*.

Slegg, Jennifer (2015). "Google Panda Will Be Updated in 2-4 Weeks". In: *Search Engine Marketing News*.

Solove, Daniel J. (Jan. 2006). "A Taxonomy of Privacy". In: *University of Pennsylvania Law Review* 154.3, pp. 477–560.

— (2007). "'I've Got Nothing to Hide' and Other Misunderstandings of Privacy". In: *San Diego Law Review* 44.05, pp. 1–23.

Speicher, Till et al. (2018). "Potential for Discrimination in Online Targeted Advertising Till Speicher MPI-SWS MPI-SWS MPI-SWS". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. Vol. 81, pp. 1–15.

Statistica (2019). "Percentage of all global web pages served to mobile phones from 2009 to 2018". In: *The Statistics Portal*.

Swanson, Judith A. (1992). *The Public and the Private in Aristotle's Political Philosophy*. Cornell University Press.

Sweeney, Latanya (2000). "Simple demographics often identify people uniquely". In: *Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000* 671, pp. 1–34.

— (2013). "Discrimination in Online Ad Delivery". In: *Ssrn* 11.3, 10:10–10:29.

Tang, Jun et al. (2017). "Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12". In: *http://arxiv.org/abs/1709.02753*.

Wang, Zhihen et al. (2018). "Using page speed in mobile search ranking". In: *Google Webmaster Central Blog*.

Weikum, Gerhard (2002). *Foundations of statistical natural language processing*. Vol. 31. 3. MIT press, p. 37.

Whigham, Nick (2017). "Leaked document reveals Facebook conducted research to target emotionally vulnerable and insecure youth". In: *https://news.com.au*.

Appendix

Lemma 1 *For* $x, y, \epsilon \in \mathbb{R}^+$ *with* $0 < x, y < 1$

$$e^{-\epsilon} < \frac{x}{y} < e^{\epsilon} \implies |x - y| < \epsilon \tag{1}$$

Proof *Assuming the left hand side of* (1) *holds*

$$\begin{aligned}
& e^{-\epsilon} < \frac{x}{y} < e^{\epsilon} \iff ye^{-\epsilon} < x \text{ and } y > xe^{-\epsilon} \\
& \implies y(1-\epsilon) < x \text{ and } y > x(1-\epsilon) \text{ (Since } e^{-x} > 1 - x\text{)} \\
& \iff y - x < y\epsilon \quad \text{and} \quad x - y < x\epsilon \\
& \iff y - x < \epsilon \quad \text{and} \quad x - y < \epsilon \text{ (Since } x, y < 1\text{)} \\
& \iff -\epsilon < x - y < \epsilon \iff |x - y| < \epsilon
\end{aligned}$$

■