
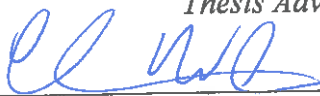

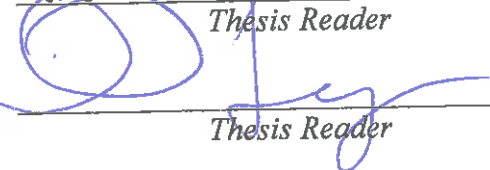


THESIS TITLE: *Personalization in Online Services
Measurement, Analysis, and Implications*

AUTHOR: *Aniko Hannak*

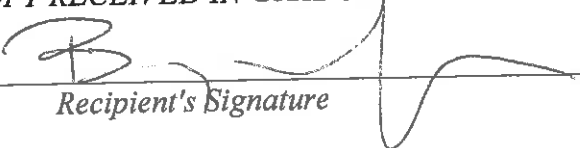
Ph.D. Thesis Approved to complete all degree requirements for the Ph.D. Degree in Computer Science.

 _____ Thesis Advisor	<u>10/15/16</u> Date
 _____ Thesis Reader	<u>10/15/16</u> Date
 _____ Thesis Reader	<u>10/15/16</u> Date
 _____ Thesis Reader	<u>10/25/16</u> Date
_____ Thesis Reader	_____ Date

GRADUATE SCHOOL APPROVAL:

 _____ Director, Graduate School	<u>12/14/16</u> Date
---	-------------------------

COPY RECEIVED IN GRADUATE SCHOOL OFFICE:

 _____ Recipient's Signature	<u>12/14/2016</u> Date
---	---------------------------

Distribution: Once completed, this form should be scanned and attached to the front of the electronic dissertation document (page 1). An electronic version of the document can then be uploaded to the Northeastern University-UMI website.

**Personalization in Online Services
Measurement, Analysis, and Implications**

A Dissertation Presented

by

Aniko Hannak

to

The College of Computer and Information Science

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Computer Science

Northeastern University

Boston, Massachusetts

April 2016

Contents

List of Figures	iv
List of Tables	vi
Acknowledgments	vii
Abstract of the Dissertation	viii
1 Introduction	1
1.1 Outline of the Presented Studies	3
1.1.1 Methodology for Measuring Personalization	4
1.1.2 Web Search Personalization	4
1.1.3 Location-based Personalization	5
1.1.4 Personalization of E-commerce	6
2 Background and Related Work	8
2.1 Search Personalization	8
2.2 Personalization of E-commerce	14
2.3 Methodology	16
3 Measuring Personalization	18
3.1 Experiment Design	19
3.2 Data Collection	21
3.3 Implementation	22
3.4 Measurement Metrics	23
4 Measuring Web Search Personalization	24
4.1 Introduction	24
4.2 Methods	26
4.2.1 Terminology	26
4.2.2 Experiment Design	28
4.2.3 Search Queries	30
4.2.4 Scope	31
4.3 Real-World Personalization	32

4.3.1	Collecting Real-World Data	32
4.3.2	Results	35
4.4	Personalization Features	36
4.4.1	Collecting Synthetic Account Data	37
4.4.2	Basic Features	37
4.4.3	Historical Features	42
4.5	Quantifying Personalization	46
4.5.1	Temporal Dynamics	47
4.5.2	Personalization of Query Categories	49
4.5.3	Personalization and Result Ranking	50
4.5.4	Personalization and Aggregated Search	52
4.6	Concluding Discussion	54
5	The Impact of Geolocation on Web Search Personalization	56
5.1	Introduction	56
5.2	Methodology	57
5.2.1	Locations and Search Terms	58
5.2.2	Data Collection and Parsing	60
5.2.3	Measuring Personalization	61
5.3	Analysis and Findings	61
5.3.1	Noise	62
5.3.2	Personalization	63
5.4	Concluding Discussion	67
6	Measuring Personalization of Ecommerce Sites	68
6.1	Introduction	68
6.2	Methodology	70
6.2.1	Definitions	70
6.2.2	E-commerce Sites	71
6.2.3	Searches	72
6.3	Real-World Personalization	72
6.3.1	Data Collection	73
6.3.2	Price Steering	74
6.3.3	Price Discrimination	76
6.3.4	Per-User Personalization	77
6.3.5	Summary	78
6.4	Personalization Features	78
6.4.1	Experimental Overview	79
6.4.2	Hotels	82
6.4.3	General Retailers	87
6.5	Concluding Discussion	88
7	Conclusion	90
	Bibliography	93

List of Figures

4.1	Example page of Google Search results.	27
4.2	Example page of Bing Search results.	27
4.3	Example of result carry-over, searching for “hawaii” then searching for “urban outfitters.”	29
4.4	Overlap of results when searching for “test” followed by “touring” compared to just “touring” for different waiting periods.	29
4.5	Results for the no-SSL versus SSL experiment on Google Search.	33
4.6	Usage of Google/Microsoft services by AMT workers.	34
4.7	% of AMT and control results changed at each rank.	34
4.8	Results for the cookie tracking experiments on Google and Bing.	38
4.9	Results for the browser experiments on Google, Bing, and DuckDuckGo.	39
4.10	Results for the geolocation experiments on Google, Bing, and DuckDuckGo.	41
4.11	Results for the User Profile: Gender experiments on Google and Bing.	42
4.12	Results for the Search History: Age Bracket experiments on Google and Bing.	43
4.13	Results for the targeted domain clicking experiments on Google and Bing.	45
4.14	Day-to-day consistency of results for the cookie tracking experiments.	47
4.15	Day-to-day consistency within search query categories for the cookie tracking experiments.	48
4.16	Differences in search results for five query categories on Google Search and Bing.	49
4.17	The percentage of results changed at each rank on Google Search and Bing.	50
4.18	Movement of results to and from rank 1 for personalized searches.	51
4.19	Rank of embedded services in search results from Google and Bing.	52
4.20	Percentage of embeddings of different services on Google and Bing.	53
5.1	Example search results from the mobile version of Google Search.	59
5.2	Average noise levels across different query types and granularities. Error bars show standard deviations.	63
5.3	Noise levels for local queries across three granularities.	63
5.4	Amount of noise caused by different types of search results for <i>local</i> queries.	63
5.5	Average personalization across different query types and granularities. Black bars shows average noise levels from Figure 5.2.	64
5.6	Personalization of each search term for <i>local</i> queries.	65
5.7	Amount of personalization caused by different types of search results.	65

5.8	Personalization of 25 locations, each compared to a baseline location, for <i>local</i> queries. The red line compares two treatments at the baseline location (i.e., the experimental control), and thus shows the noise floor.	66
5.9	Correlation of physical distance and edit distance	67
6.1	Previous usage (i.e., having an account and making a purchase) of different e-commerce sites by myAMT users.	73
6.2	Average Jaccard index (top), Kendall’s τ (middle), and nDCG (bottom) across all users and searches for each web site.	74
6.3	Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with $\geq 0.5\%$ of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).	75
6.4	Example of price discrimination. The top result was served to the AMT user, while the bottom result was served to the comparison and control.	75
6.5	AMT users that receive highly personalized search results on general retail, hotels, and car rental sites.	76
6.6	Examining the impact of user accounts and cookies on hotel searches on Cheaptickets.	80
6.7	Price discrimination on Cheaptickets. The top result is shown to users that are not logged-in. The bottom result is a “Members Only” price shown to logged-in users.	82
6.8	Home Depot alters product searches for users of mobile browsers.	85
6.9	Clearing cookies causes a user to be placed in a random bucket on Expedia.	85
6.10	Users in certain buckets are steered towards higher priced hotels on Expedia.	86
6.11	Priceline alters hotel search results based on a user’s click and purchase history.	86
6.12	Travelocity alters hotel search results for users of Safari on iOS, but not Chrome on Android.	87

List of Tables

4.1	Categories of search queries used in my experiments	30
4.2	Top 10 most/least personalized queries on Google Search and Bing.	35
4.3	User features evaluated for effects on search personalization.	37
5.1	Example <i>controversial</i> search terms.	58
6.1	The general retailers I measured in this study.	71
6.2	The travel retailers I measured in this study.	72
6.3	User features evaluated for effects on personalization.	79
6.4	Jaccard overlap between pairs of user feature experiments on Expedia.	84

Acknowledgments

None of the work presented in this thesis would have been possible without the amazing people who encouraged and supported me along the way. First and foremost, I wish to express my sincere gratitude to Alan Mislove, for teaching me about high standards in every aspect of being a researcher, for creating an extraordinarily productive and supportive environment, and for always having an open door. And for booking my flights, of course. I am tremendously thankful to Christo Wilson for all the help he provided me with in the studies that later became my thesis work, for all the great conversations (and tea) in the process, for showing me that no deadline is impossible to make and for the oxford comma. I am grateful to David Lazer (pronounced as if it were Lazar), who introduced me to interdisciplinary research and to the diverse group of postdocs surrounding him. I also want to thank the members of the Lazer-lab (actually pronounced as Lazer-lab) for the stimulating discussions that always gave me fresh perspectives on my research problems. My sincere thanks also goes to Bernardo Huberman and Markus Strohmaier, who gave me the opportunity to intern at their teams.

I thank my fellow labmates and the members of the CCIS social networks group for all the useful feedback I got over the past years, for all the foosball games, the late nights, moaning sessions, and the fun we call work. I am grateful to my friends and collaborators in academia: Piotr, Arash, Scott, Eric, Brian, Dan, Beni, Johannes, Claudia, and Ziku, whose experiences helped me avoid many mistakes, and who I could always bug with my questions. I owe you many beers! To my extended family in Boston: Dori, Gabor and the members of “El Castillo”, who constantly reminded me that there is life outside of work. I am forever indebted to Piotr for teaching me the word *indebted* and suggesting me to use it in my acknowledgements. Finally, to those who always believed in me and kept me (borderline) sane, through endless skype calls: my parents, my brother, Eszter, David, Barbi, Bori, Akos, Marci, Biba, and, basically, the rest of Budapest.

Abstract of the Dissertation

Personalization in Online Services Measurement, Analysis, and Implications

by

Aniko Hannak

Doctor of Philosophy in Computer Science

Northeastern University, April 2016

Dr. Alan Mislove, Adviser

Since the turn of the century more and more of people's information consumption has moved online. The increasing amount of online content and the competition for attention has created a need for services that structure and filter the information served to consumers. Competing companies try to keep their customers by finding the most relevant and interesting information for them. Thus, companies have started using algorithms to tailor content to each user specifically, called *personalization*. These algorithms learn the users' preferences from a variety of data; content providers often collect demographic information, track user behavior on their website or even on third party websites, or turn to data brokers for personal data. This behavior has created a complex ecosystem in which users are unaware of what data is collected about them and how it is used to shape the content that they are served.

In most cases personalization is useful for the users but there have been articles in the press with worrying examples of price discrimination or the Filter Bubble Effect. While this has triggered some awareness among the general public, it also made users realize how little control they have over their data and the form of the web they are presented. Meanwhile legal scholars and policy makers expressed concerns about algorithms' power to systematize biases and reduce accountability [15]. Unfortunately, detecting the negative consequences or measuring large-scale effects is in practice very challenging, as we still lack the tools and techniques for it.

My work starts with developing a methodology that will allow me to investigate personalization on any chosen content-based web service. With the help of this methodology I measure personalization on several services in two large sectors, search engines and e-commerce sites.

In my investigation about search engines I find that, on average, 11.7% of results show differences due to personalization on Google, while 15.8% of results are personalized on Bing, but

that this varies widely by search query and by result ranking. I also investigate the user features used to personalize on Google Web Search and Bing. Surprisingly, I only find measurable personalization as a result of searching with a logged-in account and the IP address of the searching user.

Next, to further investigate location-based personalization, I design a new tool that is able to send queries to the Google Search API appearing to come from any given GPS coordinate. Assessing the relationship between location and personalization is crucial, since users' geolocation can be used as a proxy for other demographic traits, like race, income, educational attainment, and political affiliation. Using this methodology, I collect 30 days of search results from Google Search in response to 240 different queries. By comparing search results gathered from 59 GPS coordinates around the US at three different granularities (county, state, and national), I am able to observe that differences in search results due to personalization grow as physical distance increases. However these differences are highly dependent on what a user searches for: queries for local establishments receive 4-5 different results per page, while more general terms exhibit essentially no personalization.

Finally, I turn my attention to personalization on e-commerce sites. Personalization on e-commerce sites may be used to the user's disadvantage by manipulating the products shown (price steering) or by customizing the prices of products (price discrimination). I use the accounts and cookies of over 300 real-world users to detect price steering and discrimination on 16 popular e-commerce sites. I find evidence for some form of personalization on nine of these e-commerce sites. I also create fake accounts to simulate different user features including web browser/OS choice, owning an account, and history of purchased or viewed products and identify numerous instances of price steering and discrimination on a variety of top e-commerce sites.

Tied together, these results present the first steps towards quantifying the prevalence of personalization in web-based content services and understanding the algorithms behind them. My work also provides a novel methodology that can easily be adapted by researchers who want to study content-based web services, or regulators whose goal is to audit algorithms.

Chapter 1

Introduction

Since the turn of the century, computers and smartphones have become an essential part of our lives. People use them both for their work and in their personal lives. As a result, information consumption has moved online; we rely on the Internet for the news, education, and for keeping in touch with friends and family; many institutions now promote “going paperless” by asking users to interact online when banking, finding employment, dealing with healthcare, etc.

The increasing amount of online content competing for attention has created a need for services that structure and filter the information served to consumers. While some web services reorganize already-existing information — e.g., search engines index existing content and make information more accessible — others work with data created in the context of the service itself: users log into LinkedIn or Facebook, create profiles, and place the content that becomes the service. Other examples include content recommendation systems, freelance marketplaces, and online retailers. Throughout my study I will refer to this class of services as *web-based content services*. What all of these services have in common is that they serve as a gateway to information for consumers, presented through the platform they operate.

To understand this complicated ecosystem, it is important to understand the main participants and their motivations. First, there are the *consumers*, whose main goal is to find content or participate in online communities as conveniently as possible. Second, there are the *owners* or *producers* of the content essentially competing for attention. For example, musicians earn money proportional to user click-rates and, as such, try to gain popularity through music-streaming services. Finally, there are the *service providers* who, beyond wanting to keep their customers content, have a clear economic incentive to figure out the content their users are most likely to consume. Music-streaming services make money proportional to user click-rates, online stores earn a percentage of

CHAPTER 1. INTRODUCTION

transactions they sell, etc. To maximize profit, providers try to figure out what users are most likely to consume, and place such content more prominently on their websites.

To tailor content to each user, web services have started implementing personalization algorithms [58,125,153]. For example, Google Search returns results tailored to the user’s IP ¹address or GPS ²coordinates [68] and Amazon recommends products to a user based on “similar” users’ shopping histories [94]. These two examples are fairly straightforward, but the more data the companies have about a user, the better they can predict their preferences. This leads to content providers collecting demographic information, tracking user behavior on their website or even on third party websites, or turning to data brokers for personal data. It has become easier to come by user data since users’ phones and computers essentially act as sensors that gather data every minute we use them. Most phone applications track our locations even if that is not their main functionality [138,175], a lot of them see our media, contacts, the websites we browse, our searches, our purchases, etc. Since these devices also have personal data about us, it is easy to put the pieces together into a large detailed user profile. To deal with the vast amounts of data, operators use machine learning algorithms for “Big Data” [104,111], where the input is information about the users and the output is a prediction about their preferences. Unfortunately, these algorithms often have the drawback of acting as black boxes—even the operators might not know what combination of features will be used to make the best prediction model [152].

In most cases, this effort to match the right content with the right consumers leads to satisfied users [50,56,97,147]. It makes our life easier if our Google search for restaurants returns local results because our search efforts are simpler, and when Amazon bundles similar products based on similar past purchases. However, in systems driven by user data, there is a danger that the algorithms will learn and reproduce the biases present in the population that uses them. Algorithms may reinforce common social phenomena such as discrimination or homophily. For example, in her study [157] about Google’s online ad delivery, Latanya Sweeney found that so-called black-identifying names were significantly more likely to be accompanied by text suggesting that person had an arrest record, regardless of whether a criminal record existed or not. In this case racism, results unintentionally from the technology design. Since these systems are mostly proprietary and non-transparent; users can not know what information is collected about them and how it is used to alter the content they see. Moreover, it is not even clear that the designers of the algorithms necessarily

¹Internet Protocol Address—a numerical identifier of a device (e.g., computer, printer) participating in a computer network that uses the Internet Protocol for communication.

²Global Positioning System (GPS)—is a space-based navigation system that provides location and time information through satellite connection.

CHAPTER 1. INTRODUCTION

know about aggregate effects and long term impacts. The more sophisticated the algorithms are, the harder it gets to predict the long-term consequences and to detect potential harm.

One of the first times worries about the potential harm of big data and algorithms appeared in the popular press regarded the Filter Bubble effect [148, 154], which describes the case when users are shown content that the personalization algorithm thinks they want, while other, potentially important, informations remain hidden [61]. Eli Pariser demonstrated that during the 2010 Egyptian revolution, different users searching for “Tahrir Square” on Google Search received either links to news reports of protests, or links to travel agencies [118]. Clearly, someone planning a trip to Egypt would find links to travel websites useful, but might be quite surprised on arrival to find themselves amidst a revolution. The worrisome implication is that users may have no idea that critical information has been withheld, and may not even be aware that their search results are personalized at all. Another such example was then unbeknownst to users, Orbitz (a large travel website) “steered” Apple users towards more expensive hotels placing them at higher ranks in search results [103].

Unfortunately, many people believe that algorithms are, by design, impartial and there is a false sense of objectivity associated with them [116]. However, companies to date have no legal liability measuring or controlling the possible negative impacts of their algorithms.

Researchers are working on new Machine Learning (ML) techniques that support different notions of fairness [22,23,66,84,121,183], but we currently lack the tools to verify whether these new techniques are being adopted. Similarly, legal scholars [10] have identified shortcomings in existing anti-discrimination laws that hinder the effective regulation of algorithmic systems. However, even strong regulations cannot be enforced if we do not have the ability to accurately measure algorithms.

In my thesis work, I present the first steps towards developing methods that help us detect the above described negative impacts of big data algorithms. I collect data from various web-based content services to investigate the presence of systematic biases and harmful practices. Next, I will describe in more detail the specific studies that constitute my thesis.

1.1 Outline of the Presented Studies

The basis of all three studies is a measurement methodology which I introduce first. Then I describe the contribution of each study in more detail, which is both the data collection and the analysis and implications in each case.

CHAPTER 1. INTRODUCTION

1.1.1 Methodology for Measuring Personalization

The first contribution of my thesis focuses on developing a novel methodology so I can investigate personalization on several different web-based content services. Measuring personalization is conceptually simple: one needs to compare the content different people receive when they search for information in the same way. However there are many challenges I encounter once I closely inspect this problem. Accurately attributing differences in returned content to personalization requires accounting for a number of confounding factors, including temporal changes, consistency issues in distributed search indices, and A/B tests being run by the provider, collectively referred to as *noise*. There are further difficulties that are specific to the system of interest and my methodology has to be both easily adaptable to these specifics but also high-level enough to allow investigation over a wide range of online services. In this study, I build a tool that fits these requirements and later prove its usefulness and generality by analyzing several service providers in multiple different types of content-based service providers. The details of the measurement methodology can be found in Chapter 3.

1.1.2 Web Search Personalization

With the methodology in my hand I first inspect personalization algorithms that operate on search engines. I pick search engines as the first target for two reasons. First, because of their popularity: Google has been the most visited site on the Internet for several years in a row [4], receiving more than 48,000 queries every second. Second, because of the growing concerns and coverage of the Filter Bubble Effect in the media [118], specifically in the context of search engines. Despite the worries, these concerns have not been backed up by scientific evidence or large-scale studies before my work.

First, I measure the extent of personalization by analyzing real-world data collected from users of these services. I recruit 300 users with active Google and Bing accounts from Amazon’s Mechanical Turk to run a list of web searches and I measure the differences in search results that they are given.³ By controlling for differences in time, location, distributed infrastructure, and other sources of noise, I can attribute the remaining differences observed to personalization. *Second*, I investigate the user features used for personalization, covering user-provided profile information,

³This study was conducted under Northeastern University IRB protocol #13-04-12; all personally identifiable information was removed from our collected data.

CHAPTER 1. INTRODUCTION

web browser and operating system choice, search history, search-result-click history, and browsing history.

I find that 11.7% of search results on Google and 15.8% on Bing show differences due to personalization. I see the highest personalization for queries related to political issues, news, and local businesses. When investigating specific user features, I observe that measurable personalization is mostly triggered by (1) being logged in to a user account and (2) making requests from different geographic areas. I also investigate is DuckDuckGo, which claims not to track users or personalize content. This aligns with the results of my measurements on this site and I do not observe any personalization. Therefore, the measurements of DuckDuckGo can be thought of as a baseline to compare Google Web Search and Bing Search against.

My results are the first step in providing transparency for users of web search—since users are not aware of these practices, informing them is an important duty. Moreover the methodology that I developed can be used to conduct “algorithm audits” on similar web-based services, and help regulators uncover potentially harmful practices. The data collection and analysis of search engine personalization is presented in Chapter 4.

1.1.3 Location-based Personalization

In the previous study, I found that Google varies users’ results based on their IP address, which suggests location-based personalization. Moreover, I found that political and news related queries exhibit most personalization. Motivated by these findings, I take a closer look at geo-location-based personalization of Google search, with a special focus on new and politics.

The previous study used the users’ IP address as a proxy for their location. IP address-based localization is very coarse-grained but fortunately modern web browsers—especially on mobile devices—offer APIs⁴ that can query a user’s precise location via GPS; spoofing GPS coordinates to the HTML5 Geolocation API [75] will allow me to “fake” a user’s precise location.

With the help of this newly designed tool, I am able to collect search results appearing to be from any location around the globe and answer some questions that remained open after my first study about Google search, such as: does location-based personalization impact all types of queries (e.g., politics vs. news) equally? At what distance do users begin to see changes in search results due to location? Answering these questions is crucial, since users’ geolocation can be used as a proxy for other demographic traits, like race, income-level, educational attainment, and political affiliation. In

⁴In computer programming an Application programming interface (API) is a set of routines, protocols, and tools for building software and applications

CHAPTER 1. INTRODUCTION

other words, does location-based personalization trap users in geolocal Filter Bubbles? Given the increasing penetration of mobile devices, this precise geolocation data is likely used to personalize content; thus, adding this capability to my measurement suite is of paramount importance.

I collect 30 days of search results from Google Search in response to 240 different queries. By selecting 75 GPS coordinates around the US at three granularities (county, state, and national), I am able to examine the relationship between distance and location-based personalization, as well as the impact of location-based personalization on different types of queries. One group of query terms focuses on news and politicians. The key advantage of politics as a substrate for studying personalization is that there are well-developed methodologies to study ideological/political tilt; I can “map” the personalization of web content across a given state, providing unprecedented insight into how location is used to tailor political and news-related content.

I find that the differences between search results grow as physical distance between the locations of the users increases and that queries for local terms (“airport”) are highly personalized. However, the impact of location-based personalization changes depending on the query type. Queries for politicians’ names (e.g. “Joe Biden”) and controversial topics (e.g. “abortion”) see minor changes. Surprisingly, only 20-30% of differences are due to Maps embedded in search results. Additional content analysis on the search results may help us uncover the specific instances where personalization algorithms reinforce demographic biases and finding these instances is crucial since they can serve as a bases for designing protective laws. It is important to add that while the presented study focuses on Google Search in the US, the methodology is general, and could easily be applied to other countries or search engines like Bing. The detailed study is presented in Chapter 5.1.

1.1.4 Personalization of E-commerce

Next, I take a closer look at another branch of important web-based content services: online stores. Unlike Web Search, on these sites the benefits of personalization for users are less clear; e-commerce sites have an economic incentive to use personalization to induce users into spending more money. For example, the travel website Orbitz was found to personalize the results of hotel searches. Unbeknownst to users, Orbitz “steered” Mac OS X users towards more expensive hotels in select locations by placing them at higher ranks in search results. Orbitz [103] states that they discontinued the use of this personalization algorithm after one month [31] but we lack the tools to verify the truthfulness of such statements. There may easily be similar “unfair” side effects but as long as they remain undetected the companies do not have to take responsibility for them.

CHAPTER 1. INTRODUCTION

The basic structure of an online store is very close to a search engine: people search for products in the search box and the list of results contains products with prices. Thus, I can easily adapt the methodology I used to measure Web Search personalization. In this study, I address two main questions: first, how widespread is personalization on today's e-commerce web sites? This includes *price discrimination* (customizing prices for some users) as well as *price steering* (changing the order of search results to highlight specific products). Second, how are e-commerce retailers choosing to implement personalization?

My data collection focuses on 16 top e-commerce sites in the US covering general goods retailers as well as hotel and rental car booking sites. I examine price discrimination and price steering for 300 real-world users⁵, as well as synthetically-generated fake accounts. The real-world data indicates that eight of these sites implement personalization, including cases where sites altered prices by hundreds of dollars for the same products. I also run controlled tests based on fake accounts that allow me to identify specific user features that trigger personalization. Here I found cases of sites altering results based on the user's operating system, browser, account on the site, and history of clicked/purchased products. I also observe two travel sites conducting A/B tests that steer users towards more expensive hotel reservations.

In addition to positively identifying price discrimination and price-steering on several well-known e-commerce sites, this study demonstrates the generality of the methodology I introduce in my previous study about search personalization. It can be adopted to measure various other online markets and help regulators detect illegal practices. Without quickly and accurately detecting violations, regulators have difficulties enforcing the existing consumer protection laws. The detailed study is presented in Chapter 6.1.

In the remainder of this thesis I will first give an overview of the background and literature related to my work, then describe in detail the measurement methodology I developed to measure personalization. This is followed by three studies I have conducted. I conclude with a discussion about my findings and their implications.

⁵My study is conducted under Northeastern Institutional Review Board protocol #13-04-12.

Chapter 2

Background and Related Work

This section aims to cover related work to my two main topics, personalization of search engines and e-commerce sites. Its recency and complexity makes this research area very unique, it is still at the beginning of its evolution, quickly reacting to the changes in the online world. I try to mirror this dynamic in that I first introduce the historical context in which personalization first appeared and then the new challenges that personalization in an online context introduces for researchers. I also discuss the often contradicting reactions in the popular media which is essential here, since a lot of new research was inspired by the uncertainty that the media and companies' reaction to the media has stirred. Researchers try to create more transparency by investigating the pervasiveness of personalization practices, building the necessary tools to measure it, raise attention about it and build mechanisms around it. I also cover more traditional related research areas that arise from the question of how to build successful personalized systems. This question motivates a lot of problems ranging from completely theoretical (e.g., data mining methods) to very practical (building specific personalized systems). Finally I discuss the related literature in information retrieval which helped me find metrics for measuring differences in search results.

2.1 Search Personalization

History of Search Engines The very first service we refer to as a search engine was Whois, which was made even before the debut of the Web, in 1990. Whois was a centralized system to look up domains, people and other resources related to domain and number registrations on the emerging internet [70]. The first well documented search engine was Archie. Archie emerged from a project at McGill university, when the school of computer science connected to the Internet and wanted to

CHAPTER 2. BACKGROUND AND RELATED WORK

make information available for every student. They crawled a list of FTP archives on the Web (about once a month each) and save the content. Then they made the information available and searchable for the students at the university [39, 140].

A little later, in 1993, the first web crawler was created at MIT which automatically wandered the Web (not based on a list of destinations like everything else before), allowing to measure the size of the Web and creating an index of all existing sites called Wandex [67]. JumpStation, a similar web robot used a web form as an interface to its query program thus it was the first WWW resource-discovery tool to combine the three essential features of a web search engine: crawling, indexing, and searching. Starting in 1993 a lot of search engines appeared competing for popularity. These services were already available to the public and integrated the same features that search engines present today. Among these early search engines were Altavista and Yahoo which are still major players in the search engine field. Google came around in 1998 and rose to prominence very quickly in 2000.

Search Engine Bias The main objective of search engines is to find the most relevant content to every search query [20]. Most search engines base their ranking on some modified version of the Page Rank algorithm [115], which takes both relevancy of a web page to the search query, and the absolute popularity of the web page into account when assigning relevancy scores. Even though relevancy and popularity should be the determining factors in the placement of results, studies show that there might be bias in the results due to various factors such as economics, politics, social biases, etc. Some location based bias comes simply from the fact that most popular Search Engines are based in the U.S.. A study by Vaughan et al. [169] tested three major search engines for national bias and found significant differences in their coverage of commercial Web sites. The U.S. sites were much better covered than the sites based in other countries measured in the study.

Sometimes there are explicit blacklists to prevent certain sensitive content from being shown, and these can be different depending on the regulation of different countries [55]. For example, Google will not surface certain Neo-Nazi websites in France and Germany, where Holocaust denial is illegal [19]. Another example for this was shown by Soeller et al. [150], who noticed that there are differences in Google Maps due local political views in the way borders are drawn between certain countries. Moreover, different countries have different privacy regulatory models which means they limit online tracking and advertising to a different extent. This might be another cause for differences in the results per country.

Another form of bias comes from the fact that search engines are designed to show popular

CHAPTER 2. BACKGROUND AND RELATED WORK

results higher [79] and learn from users' clicking behaviors. Google's autocomplete feature was shown to bias people towards more popular searches. Unfortunately, in many cases this can reinforce stereotypes which can be viewed as racist, sexist or homophobic [7]. Kay et al. [85] show that search results of Google's image search for careers are not representative of gender distributions of real life. Women are underrepresented and stereotypes are magnified. They emphasize the impact these results by also showing that shifting the representation of gender in image search results can shift people's perceptions about real-world distributions. These biases become especially worrisome in the light of studies attempting to prove that search results indeed have power on people's perception of the world and are able to influence their behavior. A recent study by White et al. [174] is investigating the inherent biases of search engines and their impact on the quality of information that reaches people. They show that the combined effect of people's preferences and the system's inherent bias results in settling on incorrect beliefs about half of the time.

A study by Epstein et al. [46] received a lot of media attention because of its political implications. They investigate how the impact of search results on the election outcomes and find that biased search results can shift voting preferences of undecided users by 20% or more. More over such rankings can be masked so that people show no awareness of the manipulation. Search engines have long term impact on how views and beliefs change in society. Several scholars have studied the cultural changes triggered by search engines, [72] and the representation of certain controversial topics in their results, such as terrorism in Ireland [132] and conspiracy theories [8].

Personalization of Web Search Since the relevancy of a result might depend on specific user preferences or the users' context at the time of the search, search engines place a lot of emphasis on personalizing content. With the increasing amount of information available about users it becomes an interesting question which of these is worth taking into account. It is also not clear to what extent content should be personalized for the most success. As an attempt to increase the quality of results both companies and researchers investigate the best strategies for personalizing content [56,96,112,124,125,127,143,155,159,161].

The services have many ways to acquire information about their users' interests, either explicitly or implicitly. For example interest can be taken into account explicitly, when users either specify their interests or they are asked whether they are satisfied with the results they have received. But more commonly user interests are provided implicitly; the way users search, click or browse can say a lot about them. Micarelli et al. [105] describe and contrast the most popular explicit and implicit techniques for using user data in search personalization. Most of the literature focuses on

CHAPTER 2. BACKGROUND AND RELATED WORK

the implicit techniques given that it is a very complex machine learning problem to predict user preferences or to build user interest profiles using various sources of user data [71, 101, 181, 184].

Even the “simplest” kind of information such as gender or age can reveal a lot about a user’s interests but completing this with behavioral data can result in surprisingly comprehensive user profiles. Dou et al. [43] evaluate many strategies for personalizing search, and conclude that mining user click-histories leads to the most accurate results. In contrast, user profiles have low utility. The authors also note that personalization is not useful for all types of queries. Similarly to click histories browsing history is a good proxy for a user’s interest and several studies infer interests and habits by “watching over a user’s shoulder” while surfing [56, 125].

Beside user interests the local context of the search can be also very important. Several studies have focused on the importance of location in search personalization: Yi et al. [179] and Bennet et al. [14] use linguistic tools to infer geo-intention from search queries, while study by Yo et al. [182] focuses on location relevance of webpage content to the given search query. Two studies have also shown that user demographics can be reliably inferred from browsing histories, which can be useful for personalizing content [57, 77].

Comparing Search Engines The specific search engine one chooses to use will likely also influence the results they end up seeing. There are multiple large search engines competing for users, each implementing different strategies, thus serving different content. Several studies have examined the differences between results from different search engines. Two studies have performed user studies to compare search engines [9, 168]. Sun et al. [156] propose a method for visualizing different results from search engines that is based on expected weighted Hoeffding distance. Agarwal et al. [3] come up with a novel approach to search by mining Twitter data and comparing the information received with information found on Google or Bing. Although all the above mentioned studies uncover differences between competing search engines, neither study examines the impact of personalization. Since different search engines use different personalization algorithms, personalization might add an additional layer of inconsistency to the results between users. My study investigates the personalization strategies of three different search engines side-by-side and I find that indeed there are large differences in both the coverage of domains and in the amount of personalization implemented by different services.

Exploiting Personalization Work by Xing et al. [177] also shows that it is possible to exploit the mechanisms that create personalization for nefarious purposes. The authors demonstrated that repeatedly clicking on specific search results can cause search engines to rank those results higher;

CHAPTER 2. BACKGROUND AND RELATED WORK

thereby influencing the personalization observed by others. Thus, fully understanding the presence and extent of personalization today can aid in understanding the potential impact of these attacks.

Filter Bubble The concept of the Filter Bubble effect can be attributed to Eli Pariser, who described it as “the personal ecosystem of information that’s been created by these algorithms” [118, 119]. The term describes a phenomenon in which websites use algorithms to selectively guess what information a user would like to see, based on information that they collected about the user (such as location, past click behavior and search history) [88]. Others have described the phenomenon as “ideological frames” [173] or a figurative sphere surrounding you as you search the Internet [88].

Pariser’s book about the Filter Bubble [118] warns us about potential downsides such as being closed off from new ideas, subjects and important information. He says that through personalized web services users get less exposure to conflicting viewpoints and will be isolated intellectually in their own informational bubble. He showed an example [119] in which one user searched Google for “BP” and got investment news about British Petroleum while another searcher got information about the Deepwater Horizon oil spill and that the two search results pages were “strikingly different”.

Pariser’s concerns are somewhat similar to the one made by Tim Berners-Lee in a 2010 report in *The Guardian* along the lines of a Hotel California Effect. “You can check-out any time you like, but you can neverleave” refers to the recent phenomena of Internet Social Networking sites hosting content rather than linking out to external pages and walling off information posted by their users from the rest of the web. As Berners-Lee puts it, the internet becomes a closed silo of content with the risk of fragmenting the World Wide Web.

Concerns about the Filter Bubble effect created a lot of media attention, mostly warning users about the potential downsides of personalization. However the reports about the extent to which personalization is happening and whether such activity is harmful are conflicting. Several people in the media tried to recreate Pariser’s shocking examples. They ran identical searches with accounts that had vastly different search histories but the results received were still almost identical [18, 173]. Moreover reports say [166] that users are able to control personalization in their searches by deleting history or explicitly turning it off. A study by Wharton [73] analyzed music recommendations and found that users use these filters to broaden their taste rather than to limit it. According to their observations over time recommendation creates commonality rather than fragmentation in online music taste. Google’s spokesperson says that they deliberately added algorithms to “limit personalization and promote variety” [173].

CHAPTER 2. BACKGROUND AND RELATED WORK

Nevertheless it is clear that Google and other search engines possess large amounts of data about every user and can choose to personalize content any time. Studies show that due to their large ad-network Google can keep track of user past histories even if they don't have a personal Google account or are not logged into one [126, 135]. One report says that Google has collected "10 years worth" of information amassed from varying sources, such as Gmail, Google Maps, and other services besides its search engine, [62] although a contrary report was that trying to personalize the Internet for each user was technically challenging for an Internet firm to achieve despite the huge amounts of available web data. But clearly most content providers are pushing towards creating personalized information engines, which means if not now, in a couple of years they will be able to use all the personal data they collect about us and tailor content to those that users are likely to agree with [173].

Preserving Privacy At the time of my publication about Google's search personalization only a couple of studies have focused on privacy-preserving personalized search [178]. However given growing concerns about the Filter Bubble effect and bid data algorithms, after 2014 more and more focus shifted towards investigating what data is collected about users to personalize content [24, 44, 51, 52, 90] as well as building systems that protect user privacy [83, 89, 102].

Reverse-engineering the bubble Given all the conflicting reports and news articles about this topic, it is no surprise that researchers started investigating the Filter Bubble effect as well. After 2010 a new line of research started to emerge with similar goals to my studies i.e., quantifying personalization on deployed web services. Latanya Sweeney [157] examined Google AdSense and uncovered that the system serves ads in a racially biased manner. My first study about Google search [68] was the first one to systematically collect data from Google Search and reverse-engineer their personalization strategies. A little later two studies investigate how personal data is collected and used to shape search results and both find that location has the largest impact among the examined features. Fernando et al. [52] show that this effect is even larger than the effect of turning personalization off completely. This aligns with my findings as well and gave motivation for my second study focusing only on location based personalization. Even though the Filter Bubble issue was first brought up in the context of Web Search it effects a much wider pool of web services. Online Social Platforms, E-Commerce Sites, News Sources, etc are also known to personalize content. Many studies discuss the possible negative effects of algorithmic personalization in these contexts [41, 137]. Flexion et al. [53, 54] argue that personalization can highly effect news consumption and can lead to ideological segregation. Several studies investigate personalization in online ads [64, 89]. Further

CHAPTER 2. BACKGROUND AND RELATED WORK

studies have examined the effects of algorithmic personalization on the Facebook News Feed [47,48], e-commerce [69, 106, 107], and online ads [24, 64, 89].

2.2 Personalization of E-commerce

Direct Marketing Personalization commonly occurs in contexts other than search as well. In fact its practice started even before online services existed. The first form of personalization appeared when companies in the retail industry, healthcare and credit card business switched to direct marketing from mass targeting, around 1990 [17,95]. This meant that instead of equally advertising to everyone, companies tried to identify likely buyers of certain products and promoted the products accordingly. For example in 1990 AT&T made a very successful move [17] by entering the credit card business, since they already owned a lot of information on their customers, they could easily target the ones likely to want their credit cards.

The switch to direct marketing had a large effect on companies' coupon distribution practices. Between 1979 and 1984 the number of coupons distributed has grown from 81 billion to 163 billion. (J. O. Peckham 1985, vice president of nielsen). Later in 1993 Nielsen Clearing House reports that more than 75% of households use coupons in some product category. In 1994 more than 327 billion coupons were issued and the average face value of coupons increased by 7% that year [122]. The sudden change can be attributed to a couple of different things. Companies realized that directly communicating with customers helps understanding each others' needs and customers are more likely to cooperate. Which of course more often results in an exchange [145]. This was also the time when service companies realized that the quality of service is important and started to measure it. Striving for "zero defections" they tried to keep every customer the company could profitably serve [133]. Later in the 90s some big companies worked on creating automated systems for coupon generation and distribution [33, 114].

Following this big change researchers started reflecting on the new trends and searching for new paradigms that can better describe marketing and economics processes. Researchers in economics investigated redemption rates and patterns [78, 131] and the impact of coupons on market share [109]. In marketing most studies were concerned with the success of different targeting strategies [11, 92] and methods to identify the consumer groups that are more likely to redeem their coupons [110, 160]. Rossi et al. [136] look at the importance of purchasing history information of house holds in direct targeting. They compare the success of demographic data with additional

CHAPTER 2. BACKGROUND AND RELATED WORK

purchasing history data and find that even rather short purchase histories can produce a net gain in revenue from target coupling which is 2.5 times the gain from blanket couponing. Even information about one purchase can boost by 50%. Bawa et al. [12] analyze specific characteristics of households that make them more likely to adapt to a brand after coupon redemption. They find that for the product tested, coupons produced greater incremental sales among households that were larger, more educated, and were homeowners. They conclude that directing coupons to the most responsive market segments can increase profits significantly.

It is widely believed that the success of direct marketing correlates with the amount of information companies possess about users and the specificity with which they are able to target customers using this data [113]. Thus both companies and researchers heavily investigate how data mining, machine learning and statistical techniques can be used on user data to target customers [2, 32, 162]. Sarwar et al. [139] several techniques for analyzing large scale purchase and preference data for the purpose of producing useful recommendations to customers. They compare several common data mining methods on two data sets of customer transactions.

Naturally with the growing amount of personal data collection a lot of privacy concerns arise. In their study Nowak et al. [113] raise awareness about important issues that come with direct marketers' growing reliance on computerized databases, customized persuasion and other consumer intensive strategies. "How far can companies go in learning about customers without hurting users' privacy", they ask. They address privacy concerns by developing a framework for identifying the underlying dimensions of the privacy construct and examining the relationships between those dimensions and direct marketers' consumer information practices. There are several studies investigating the possible negative effects of direct marketing in specific sensitive segments of the retail industry. For example pharmaceutical companies argue that direct to-consumer advertising can raise awareness about health and educate patients. But in fact studies show that it has been largely ineffective in educating patients with medical conditions about the medications for those conditions and it only helps them to increase revenue [100]. Study by Lewis et al. [93] investigates direct advertising in tobacco industry and concludes that its potential to increase consumption and impede cessation is unquestionable.

E-commerce Given the long history of personalization practices in the retail industry, it is natural that companies transferred this mindset to their online platforms as well. In a lot of ways the digital age makes it even more convenient for the companies. They can easily track purchase histories, provide personalized recommendations, or get hold of personal data on their users from data brokers.

CHAPTER 2. BACKGROUND AND RELATED WORK

Study by Lee et al. [91] shows that item browsing patterns and cart usage patterns are the important predictors of the actual purchases and their prediction model on user purchases achieves over 80% accuracy based on features about user behavior on the site. This supports the idea that it is worth investing in user tracking and incorporating the collected data into their algorithms.

Only around 2012 did researchers start to investigate personalization on e-commerce sites. The first study focusing specifically on e-commerce sites was done by J. Mikians et al. [106] who established the terminology for measuring inconsistencies on e-commerce sites. They investigated the effect of location, OS/browser settings and browsing history as features and found examples for price discrimination based on location and price steering based on browsing history. In their later paper [107] they extended this study to use crowdsourcing to help detect instances of price variation and such identify a set of online vendors where price variation is more pronounced. Both studies served as inspiration for my own work. I improve the methods they used in these studies by introducing a control to every experiment I run (both crowdsourcing and the synthetic tests). This allows me to specifically differentiate between difference due to the inherent noise in the systems and actual personalization. A close relative of personalization on ecommerce sites is algorithmic pricing. It is similar to personalization in that there are automatic self-learning algorithms that help maximize profit margins for companies but the differences in prices occur over time and not between people. A 2016 study by Chen et al. [29] develops the methodology to measure algorithmic pricing and analyzes a large data set collected from Amazon to uncover strategies of over 500 sellers.

2.3 Methodology

Comparing Search Results and Engines Comparing ranked lists, such as search engine results, is an active topic in the Information Retrieval (IR) community. Several studies improve the classic Kendall's τ metric by adding per-rank weights [49, 146], and by taking item similarity into account [87, 141]. DCG and NDCG use a logarithmic scale to reduce the scores of bottom ranked items [80]. The cascade model [35] and its successor ERR [27] extend DCG by taking the relatedness of subsequent list items into account. Two studies develop probabilistic metrics that penalizing errors at top ranks without requiring explicit weights [25, 180].

Classical metrics such as Spearman's footrule and ρ [40, 151] and Kendall's τ [86] both calculate pairwise disagreements between ordered lists. However, these classic metrics do not take into account that when evaluating results from search engines, disagreements among low ranking items (e.g., rank 1) are more costly. Numerous new metrics have been proposed to address this

CHAPTER 2. BACKGROUND AND RELATED WORK

shortcoming. Fagin et al. [49] and Shieh et al. [146] both propose modifications of Kendall's τ that incorporate per-rank weights. Discounted Cumulative Gain (DCG) and Normalized Discounted Cumulative Gain (NDCG) are both metrics designed to score search engine results by taking rank order into account [81]. Yilmaz et al. [180] and Carterette [25] both propose new probabilistic metrics that penalize errors at low ranks without requiring explicit rank weights. Sculley [141] and Kumar et al. [87] extend Kendall's τ by taking item similarity into account.

Unfortunately, these metrics are not suitable for use in my study. The above metrics are designed to streamline the comparison of ranked lists down to a single value that is suitable for evaluating and optimizing IR systems. In my work, I aim for a more nuanced understanding of the differences between lists of search results, i.e., I want to examine overlap and position switching separately. Sun et al. [156] propose a method for visualizing different results from search engines that is based on expected weighted Hoeffding distance. Although this technique is very promising, it does not scale to the size of my experiments.

Learning to Rank A large amount of research has been conducted on using implicit user information to improve the ranking of search engines. Studies have examined clickthrough of search results [28, 45, 76, 82, 129, 130], dwell time on webpages [1], and even cursor tracking [65] as signals to solve the learning to rank problem. However, these studies are focused on improving the overall quality of the search index, not leveraging user information to personalize search results.

Chapter 3

Measuring Personalization

The first contribution of my thesis is the design of a methodology that allows me to measure personalization on the web services I am interested in. When measuring personalization I want to see whether different people receive different content when they use the services in identical contexts. To establish these identical contexts I will leverage the search functionality of web-based content services. Most web-based content services have search implemented on top of their interface as the immediate tool for people to filter content and find what they are looking for. In the case of search engines of course it is the main functionality of the site but if we look at other commonly used services, it is hard to find one without a search box on the main interface. With this idea, measuring personalization becomes simple: let's compare what different users are shown when searching for the same concepts.

My main goal when designing my measurement methodology was to make it easily adaptable to any system I want to investigate. Since there are new services appearing with each day and old ones periodically introducing new techniques, any particular finding may only be accurate for a small time window. Thus it is especially important to design tools that are easily reusable on a variety of systems from time to time.

In the later parts of my thesis I will demonstrate how I used my methodology to measure personalization on Search Engines as well as Online Stores. Thus most of my examples in this section will refer to Search Engines or e-commerce sites nevertheless my methodology could be applied to any system that operates with a search functionality.

In the following I will first, give the high-level intuition that guides the design of my experiments, and identify sources of noise that can lead to errors in data collection. Second, I introduce the goals and specifics of my two data collection methods. Third, I describe the implementation of

my experiments and lastly, I define the measurement metrics used to quantify personalization.

3.1 Experiment Design

Search As mentioned above the key to observing personalization will be to run searches on the target web-based content service. For example when I measure personalization on an e-commerce site, I can run searches for the same products from different machines and compare the resulting pages with products. Thus it is important to first define the specific set of terms I will use throughout the study when referring to *search*. Each *query* to the search of a web-based content service is composed of one or more *keywords*. In response to a query, the site returns a page of *results*. For example in the case of Web Search this is a page containing 10 results (URLs), or in the case of online stores it is a list of products and associated prices.

Personalization or Noise? Personalization on web services comes in many forms (e.g., “localization”, per-account customization, etc.), and it is not entirely straightforward to declare that an inconsistency between the search results observed by two users is due to personalization. For example, two users’ search queries may have been directed to different data centers, and the differences are a result of data center inconsistency rather than intentional personalization. For the purposes of this study, I define personalization to be taking place when an inconsistency in the search results is due to a piece of client-side state associated with the request. For example, a client’s request often includes tracking cookies, a User-Agent identifying the user’s browser and Operating System (OS), and a source IP address where the client’s request originated. If any of these lead to an inconsistency in the results, I declare the inconsistency to be personalization. In the different-datacenter example from above, the inconsistency between the two results is not due to any client-side state, and I therefore declare it not to be personalization. These inconsistencies or essentially, noise, can be caused by a variety of factors:

- **Updates to the Search Index:** Search services constantly update their search indices. This means that the results for a query may change over time.
- **Distributed Infrastructure:** Large-scale web search or e-commerce services are spread across geographically diverse data centers. My tests have shown that different data centers may return different results for the same queries. It is likely that these differences arise due to inconsistencies in the search index across data centers.

CHAPTER 3. MEASURING PERSONALIZATION

- **Geolocation:** Web services use the user’s IP address to provide localized results [179]. E-commerce sites might also account for price differences due to shipping costs, local taxes and currency conversions. Movie streaming services often have different movies available depending on the country the consumer is streaming from.
- **A/B testing:** Sites may conduct A/B testing [117], where the results are altered to measure whether users click on them more often. I do not consider such testing as personalization as long as it is performed randomly, independent of any client-side state.
- **Updates specific to the measured service:** For example e-commerce services are known to update their inventory often, as products sell out, become available, or prices are changed or music streaming services change the pool of available songs based on their contracts with the artists. This means that the results for a query may change even over short timescales.

Controlling Against Noise To control for all sources of noise, in each experiment I will include a control, that is configured in an identical manner to one other treatment (i.e., I run one of the experimental treatments twice). Doing so allows me to measure the noise as the level of inconsistency between the control account and its twin; since these two treatments are configured identically, any inconsistencies between them must be due to noise, not personalization. Then, I can measure the level of inconsistency between the different experimental treatments; if this is higher than the baseline noise, the increased inconsistencies are due to personalization. As a result, I cannot declare any particular inconsistency to be due to personalization (or noise), but I can report the overall rate.

To see why this works, suppose I want to determine if Firefox users receive different prices than Safari users on a given site. The naive experiment would be to send a pair of identical, simultaneous search—one with a Firefox User-Agent and one with a Safari User-Agent—and then look for inconsistencies. However, the site may be performing A/B testing, and the differences may be due to requests given different A/B treatments. Instead, I run an additional control (say, a third request with a Firefox User-Agent). The differences I see between the two Firefox treatments will then measure the frequency of differences due to noise. Of course, running a single query is insufficient to accurately measure noise and personalization. Instead, I run a large set of searches on each site over multiple days and report the aggregate level of noise and personalization across all results.

To control for temporal effects all of my machines execute searches for the same query at the same time (i.e., in lock-step). To eliminate differences that might arise from inconsistencies

CHAPTER 3. MEASURING PERSONALIZATION

between different data centers, I use static DNS entries to direct all of query traffic to one specific IP address for each website. Finally, unless otherwise stated, I send all of the search queries for a given experiment from the same /24 subnet which ensures that any geolocation would affect the results equally.

3.2 Data Collection

My study seeks to answer two broad questions. First, *to what extent does personalization actually affect the content users are shown?* Although it is known that web services use personalization algorithms, it is not clear how much they actually alter the content. If the delta between “normal” and “personalized” content is small, then concerns over the Filter Bubble effect may be misguided. Second, *what user features influence personalization algorithms on web-based content services?* This question is fundamental: outside of the companies themselves, nobody knows the specifics of how personalization works.

Real-World data collection To answer my first question, I begin by measuring the extent of personalization that users are seeing today. Doing so requires obtaining access to the search results observed by real users. Comparing the results for the same queries received by different people is a good proxy for the amount of personalization they experience as they interact with web services. I therefore design a user study that allows me to collect search results from a given web service and that can be crowdsourced to users of the service. The intuition behind the experiment is to have many users of a service run the same set of searches (in the same exact setting they usually use the service) while allowing me to record the results that they receive. In more detail, first the participants are instructed to configure their web browser to use a Proxy Auto-Config (PAC) file provided by me. The PAC file routes all traffic to the sites under study to an HTTP proxy controlled by me. Then, users are directed to visit a web page containing JavaScript that performs my set of searches in an `iframe`. After each search, the Javascript grabs the HTML in the `iframe` and upload it back to the server, allowing me to view the results of the search. By having the users run the searches within their own browsers, any cookies that the users’ browser had previously assigned would automatically be forwarded in my searches. This allows me to examine the exact results that the user would have received. It is a good idea to wait 15-20 second between searches to not overload the browser and the service being queried.

Besides allowing me to observe the results received by the users, the proxy serves another

CHAPTER 3. MEASURING PERSONALIZATION

important function. Whenever it observes a search request, it fires off *two* identical searches using PhantomJS (with no cookies) and saves the resulting pages. The results from PhantomJS serve as a *comparison* and a *control* result: the comparison query allows me to compare results served to the users with results served to “blank” users and the control will show the underlying noise when compared to the (identical) comparison query.

Collecting Synthetic Account Data My second question is about investigating which specific user features effect personalization. At a high-level, my methodology is to execute carefully controlled queries on the target web service to identify what user features trigger personalization. Each experiment follows a similar pattern: first, create x accounts to the web service that each vary by one specific feature. Second, execute q identical queries from each account, once per day for d days. Save the results of each query. Finally, compare the results of the queries to determine whether the same results are being served in the same order to each account. If the results vary between accounts, then the changes can be attributed to personalization linked to the given experimental feature. Note that certain experimental treatments are run *without* accounts (i.e., to simulate users without accounts).

For example let’s suppose I want to test the effect of Gender on the search results on Google Search. I would create a “Male” and a “Female” account and one more “Female” account as a control. In every other aspect these accounts would be exactly identical. By comparing search results that I run via the two Female accounts will give me an idea about the underlying noise in Google’s search. If I see greater differences between the Male and Female account, I can safely attribute it to gender based personalization.

3.3 Implementation

My experiments are implemented using custom scripts for PhantomJS [123]. I chose PhantomJS because it is a full implementation of the WebKit browser, i.e., it executes JavaScript, manages cookies, *etc.* Thus, using PhantomJS is significantly more realistic than using custom code that does not execute JavaScript, and it is more scalable than automating a full Web browser (e.g., Selenium [142]).

On start, each PhantomJS instance logs in to the specified account (e.g., a Google or Microsoft account) using separate credentials, and begins issuing queries to the Web search engine. The script downloads a specified number of pages of search results for each query. (In my studies I mainly focused on the first page of results.) The script waits a specified amount of time in-between

CHAPTER 3. MEASURING PERSONALIZATION

searches for subsequent queries to avoid the queries effecting each other.

During execution, each PhantomJS instance remains persistent in memory and stores all received cookies. After executing all assigned queries, each PhantomJS instance closes and its cookies are cleared. The various cookies are recreated during the next invocation of the experiment when the script logs in to its assigned account.

All instances of PhantomJS are run on a single machine. I modified the `/etc/hosts` file of this machine so that DNS queries to Web search services resolve to specific IP addresses. I use SSH tunnels to forward traffic from each PhantomJS instance to a unique IP address in the same /24 subnet.

3.4 Measurement Metrics

When comparing the lists of search results between two accounts, there are essentially two question I can ask. Are the results the same across the two pages? And are the results presented in the same order?

Jaccard Index To measure the overlap of results across the two pages, I use Jaccard Index, which views the result lists as sets and is defined as the size of the intersection over the size of the union. A Jaccard Index of 0 represents no overlap between the lists, while 1 indicates they contain the same results (although not necessarily in the same order).

Edit Distance Next, to measure reordering between the lists, I use edit distance. To calculate edit distance, I compute the number of list elements that must be inserted, deleted, substituted, or swapped (i.e., the Damerau-Levenshtein distance [37]) to make one test list

Chapter 4

Measuring Web Search Personalization

4.1 Introduction

I now move on to the second contribution of my thesis; measuring personalization of web search. My choice to study personalization on web search engines is two-fold. First, web search services like Bing and Google Web Search (Google Search) are an integral part of our daily lives; Google Search alone receives 17 billion queries per month from U.S. users [34]. People use web search for a number of reasons, including finding authoritative sources on a topic, keeping abreast of breaking news, and making purchasing decisions. The search results that are returned, and their order, have significant implications: ranking certain results higher or lower can dramatically affect business outcomes (e.g., the popularity of search engine optimization services), political elections (e.g., U.S. Senator Rick Santorum’s battle with Google [158]), and foreign affairs (e.g., Google’s ongoing conflict with Chinese Web censors [176]).

The second reason to study search engines has to do with user expectations. On some services users expect to see personalized content, for example music or movie streaming services, where more of the emphasis lies on recommendation. On those services users expect the system to learn their taste and give recommendations accordingly [120]. They often willingly “teach” the algorithms by providing personal data or rating the content that they have consumed [74].

However people turn to search engines to find factual information. They use search engines to find answers to their questions, or learn about the world and it is crucial that when doing so they expect some sort of objectivity. (For example, if someone these days is not familiar with a concept, instead of reaching for a lexicon they will likely reach for a laptop or phone and type it into a search box.) Even if they might question the validity of content from certain sources they likely do not

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

question that other people are presented the same information. Thus I think it is extremely important to educate people about the prevalence of personalization and its possible negative effects. Beyond speculations in the popular media based the few instances where personalization was detected, until my study there has been almost no scientific quantification of the basis and extent of search personalization in practice.

In this chapter, I make two contributions towards remedying this situation. *First*, using the methodology described in Section 3 I measure the extent of personalization on multiple popular web search engines: Google Web Search, Bing Search, and DuckDuckGo.¹ I recruit 300 users with active Google and Bing accounts from Amazon's Mechanical Turk to run a list of web searches, and I measure the differences in search results that they are given. I control for differences in time, location, distributed infrastructure, and noise, allowing me to attribute any differences observed to personalization. Although my results are only a lower bound, I observe significant personalization: on average, 11.7% of Google Web Search results and 15.8% of Bing Search results show differences due to personalization, with higher probabilities for results towards the bottom of the page. I see the highest personalization for queries related to political issues, news, and local businesses. I do not observe any noticeable personalization on DuckDuckGo.

Second, I investigate the user features used to personalize, covering user-provided profile information, web browser and operating system choice, search history, search-result-click history, and browsing history. I create numerous Google and Bing accounts and assign each a set of unique behaviors. I develop a standard list of 120 search queries that cover a variety of topics pulled from Google Zeitgeist [59] and WebMD [172]. I then measure the differences in results that are returned for this list of searches. Overall, I find that while the level of personalization is significant, there are very few user properties that lead to personalization. Contrary to my expectations, for both Google and Bing, I find that only being logged in to the service and the location (IP address) of the user's machine result in measurable personalization. All other attributes do not result in a level of personalization beyond the baseline noise level.

The results presented in this study regarding Google Search were published in the proceedings of the World Wide Web conference 20113 under the title Measuring the Personalization of Web Search [68].

¹DuckDuckGo is a relatively new search engine that claims to not track users or personalize results. As such, I do not expect to see personalized results, and I include my measurements of DuckDuckGo primarily as a baseline to compare Google Web Search and Bing Search against.

Roadmap The remainder of this chapter is organized as follows: in Section 4.2, I describe my experimental methodology. In Section 4.3, I quantify real-world search personalization using results from crowdsourced workers, while in Section 4.4, I perform controlled experiments to ascertain what features search engines use to personalize results. Next, in Section 4.5, I examine how the personalization varies over time, across query categories, and by result rank. I conclude with a discussion of results and limitations in Section 4.6.

4.2 Methods

In this section, I describe how I adapt the experimental methodology from section ref:sec:method to collect data from search engines.

4.2.1 Terminology

In this study, I use a specific set of terms when referring to web search. Each *query* to a web search engine is composed of one or more keywords. In response to a query, the search engine returns a *page of results*. Figure 4.1 shows a truncated example page of Google Search results for the query “coughs”, and Figure 4.2 shows a truncated example page of Bing Search results for the query “tornado.” Each page contains ≈ 10 results (in some cases there may be more or less). I highlight three results with red boxes in both figures. Most results contain ≥ 1 links. In this study, I only focus on the *primary link* in each result, which I highlight with red arrows in Figures 4.1 and 4.2.

In most cases, the primary link is *organic*, i.e., it points to a third-party website [26]. The WebMD result in Figure 4.1 falls into this category. However, the primary link may point to another Google or Microsoft service. For example, in Figure 4.1 the “News for coughs” link directs to Google News, and the “More news about Tornado” link in Figure 4.2 directs to Bing News. Search engines often include links to other *services* offered by the same company; this strategy is sometimes referred to as “aggregated search.”

A few services inserted in web search results do not include a primary link. The “Related Searches” result in Figure 4.1 falls into this category. Another example is Google Dictionary, which displays the definition of a search keyword. In these cases, I treat the primary link of the result as a descriptive, static string, e.g., “Related” or “Dictionary.”



Figure 4.1: Example page of Google Search results.

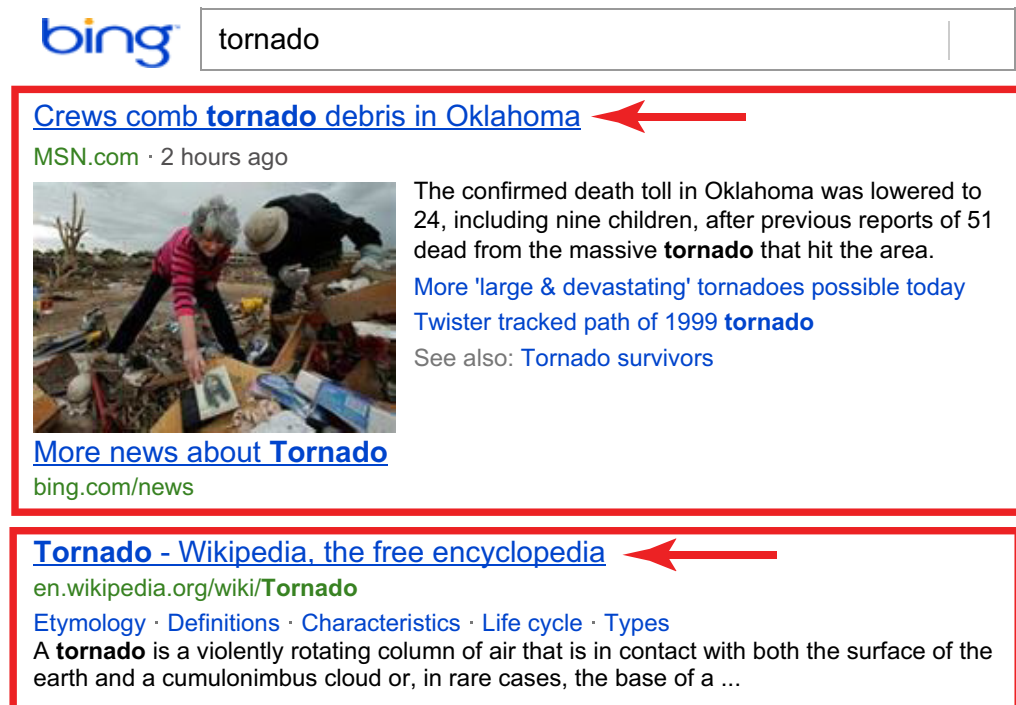


Figure 4.2: Example page of Bing Search results.

DuckDuckGo Search results from DuckDuckGo follow a different format from Google and Bing. On DuckDuckGo, the top of the search result page is dominated by a box of contextual information related to the query. For example, after searching for “barack obama” the contextual box contains information about the president taken from Wikipedia, and links to recent news articles. Below the contextual box is the list of organic search results. Unlike Google and Bing, DuckDuckGo does not return multiple different pages of search results. Instead, the page continually loads more results as the user scrolls down the page.

In this study, I focus on the search results returned by DuckDuckGo, and ignore links in the contextual box. On DuckDuckGo, results are presented in a simple ordered list, so there is no problem of having multiple links in one result. I focus on the top 10 results returned by DuckDuckGo, so that the analysis is comparable across the three search engines.

4.2.2 Experiment Design

My study seeks to answer two broad questions. First, *what user features influence web search personalization algorithms?* This question is fundamental: outside of web search companies, nobody knows the specifics of how personalization works. Second, *to what extent does search personalization actually affect search results?* Although it is known that web search companies personalize search results, it is not clear how much these algorithms actually alter the results. If the delta between “normal” and “personalized” results is small, then concerns over the Filter Bubble effect may be misguided.

As described in section 3 there is many sources of noise that can cause differences in the received results that I have to carefully separate from the effects of personalization in my experiments. I send all queries at the same time, I fix the IP address to Google and all machines sending the queries are in the same /24 IP range.

The Carry-Over Effect Besides the sources of noise mentioned in Section 3.1 that my methodology controls for Google search presents an extra challenge. This particular source of noise comes from the dependency of searches within one “browsing session.” For example, if a user searches for query *A*, and then searches for query *B*, the results for *B* may be influenced by the previous search for *A*. Prior research on user intent while searching has shown that sequential queries from a user are useful for refining search result [6, 30, 108, 144, 149]. Thus, it is not surprising that some search

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

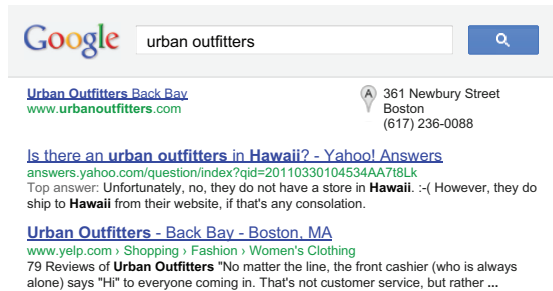


Figure 4.3: Example of result carry-over, searching for “hawaii” then searching for “urban outfitters.”

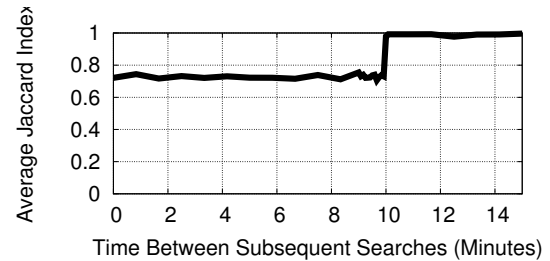


Figure 4.4: Overlap of results when searching for “test” followed by “touring” compared to just “touring” for different waiting periods.

engines implement *query refinement* using consecutive keywords within a user’s browsing session. I term the effect of query refinement on subsequent searches as the *carry-over effect*.

An example of carry-over on Google Search is shown in Figure 4.3. In this test, I search for “hawaii” and then immediately search for “urban outfitters” (a clothing retailer). I conducted the searches from a Boston IP address, so the results include links to the Urban Outfitters store in Boston. However, because the previous query was “hawaii,” results pertaining to Urban Outfitters in Hawai’i are also shown.

To determine how close in time search queries must be to trigger carry-over, I conduct a simple experiment. I first pick different pairs of queries (e.g., “gay marriage” and “obama”). I then start two different browser instances: in one I search for the first query, wait, and then for the second query, while in the other I search only for the second query. I repeat this experiment with different wait times, and re-run the experiment 50 times with different query pairs. Finally, I compare the results returned in the two different browser instances for the second term.

The results of this experiment on Google Search are shown in Figure 4.4 for the terms “test” and “touring” (other pairs of queries show similar results). The carry-over effect can be clearly observed: the results share, on average, seven common results (out of 10) when the interval between the searches is less than 10 minutes (in this case, results pertaining to Turing Tests are included). After 10 minutes, the carry-over effect disappears. Thus, in all Google-focused experiments in the following sections, I wait at least 11 minutes between subsequent searches in order to avoid any carry-over effects. In my testing, my observed carry-over for both logged in users and users without Google accounts.

I performed the same experiments on Bing and DuckDuckGo, but did not observe any carry-over effects. Thus, I conclude that the carry-over effect is unique to Google Search (at least in

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

Table 4.1: Categories of search queries used in my experiments

Category	Examples	No.
Tech	Gadgets, Home Appliances	20
News	Politics, News Sources	20
Lifestyle	Apparel Brands, Travel Destinations, Home and Garden	30
Quirky	Weird Environmental, What-Is?	20
Humanities	Literature	10
Science	Health, Environment	20
Total		120

fall 2012, when I was conducting measurements).

To avoid measurement errors due the carry-over effect, I wait 11 minutes in-between subsequent queries. As shown in Figure 4.4, an 11 minute wait is sufficient to avoid the majority of instances of carry-over. For consistency, I use this same methodology for Google Search, Bing and DuckDuckGo, even though the latter two do not exhibit carry-over.

Finally, I include a *control account* in each of my experiments. The control account is configured in an identical manner to one other account in the given experiment (essentially, I run one of the experimental treatments twice). By comparing the results received by the control and its duplicate, I can determine the baseline level of noise in the experiment (e.g., noise caused by A/B testing). Intuitively, the control should receive exactly the same search results as its duplicate because they are configured identically, and perform the same actions at the same time. If there is divergence between their results, it must be due to noise.

Accounts Unless otherwise specified, each Google and Microsoft account I create has the same profile: 27 year old, female. The default User-Agent is Chrome 22 on Windows 7. As shown in Section 4.4.2, I do not observe any personalization of results based on these attributes.

I manually crafted each of my accounts to minimize the likelihood of being automatically detected. Each account was given a unique name and profile image. I read all of the introductory emails in each account’s email inbox (i.e., in GMail or Hotmail). To the best of my knowledge, none of my accounts were banned or flagged by Google or Microsoft during my experiments.

4.2.3 Search Queries

In my experiments, each account searches for a specific list of queries. It is fundamental to my research that I select a list of queries that has both breadth and impact. Breadth is vital, since I do

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

not know which queries web services personalize results for. However, given that I cannot test all possible queries, it is important that I select queries that real people are likely to use. ‘

Traditionally, search queries are classified into three different classes according to their intent: navigational, informational and transactional [21]. Navigational queries are not interesting from the perspective of personalization, since navigational queries tend to have a single, “correct” answer, i.e., the URL of the desired website. In contrast, the results of informational and transactional queries could be personalized; in both cases, the user’s intent is to seek out information or products from a potentially large number of websites. Thus, in my experiments I focus on informational and transactional queries.

As shown in Table 4.1, I use 120 queries divided equally over 12 categories in my experiments. These queries were chosen from the 2011 Google Zeitgeist [59], and WebMD [172]. Google Zeitgeist is published annually by Google, and highlights the most popular search queries from the previous calendar year. I chose these queries for two reasons: first, they cover a broad range of categories (breadth). Second, these queries are popular by definition, i.e., they are guaranteed to impact a large number of people.

The queries from Google Zeitgeist cover many important areas. 10 queries are political (e.g., “Obama Jobs Plan”, “2012 Republican Candidates”) and 10 are related to news sources (e.g., “USA Today News”). Personalization of political and news-related searches are some of the most contentious issues raised in Eli Pariser’s book on the Filter Bubble effects [118]. Furthermore, several categories are shopping related (e.g., gadgets, apparel brands, travel destination). As demonstrated by Orbitz, shopping related searches are prime targets for personalization [103].

One critical area that is not covered by Google Zeitgeist is health-related queries. To fill this gap, I chose ten random queries from WebMD’s list of popular health topics [172].

4.2.4 Scope

All of my experiments were conducted in fall of 2012 and spring of 2013. Although my results are representative for this time period, they may not hold in the future, since web search engines are constantly tweaking their personalization algorithms.

4.3 Real-World Personalization

I begin by measuring the extent of personalization that users are seeing today. Doing so requires obtaining access to the search results observed by real users; I therefore conducted a simple user study.

4.3.1 Collecting Real-World Data

I posted two tasks on Amazon’s Mechanical Turk (AMT), explaining my study and offering each user \$2.00 to participate.² In the first task, participants were required to 1) be in the United States, 2) have a Google account, and 3) be logged in to Google during the study. The second task was analogous to the first, except it targeted users with Bing accounts. Users who accepted either task were instructed to configure their web browser to use a HTTP proxy controlled by us. Then, the users were directed to visit a web page hosted on my research server. This page contained JavaScript that automatically performed the same 80 searches on Google or Bing, respectively.³ 50 of the queries were randomly chosen from the categories in Table 4.1, while 30 were chosen by us.

The HTTP proxy serves several functions. *First*, the proxy records the search engines’ HTML responses to the users’ queries so that I can observe the results returned to the user. I refer to these results as the *experimental results*. *Second*, each time the proxy observes a user making a query, it executes two PhantomJS scripts. Each script logs in to the respective search engine and executes the same exact query as the user. I refer to the results observed by these two scripts as the *control results*, and they allow us to compare results from a real user (who Google/Bing has collected extensive data on) to fresh accounts (that have minimal Google/Bing history). *Third*, the proxy controls for noise in two ways: 1) by executing user queries and the corresponding scripted queries in parallel, and 2) forwarding all search engine traffic to hard-coded IP addresses for Google and Bing.

SSL versus no-SSL Although the proxy is necessary to control for noise, there is a caveat to this technique when it is applied to Google Search. Queries from AMT users must be sent to `http://google.com`, whereas the controls use `https://google.com`. The reason for this issue is that HTTPS Google Search rejects requests from proxies, since they could indicate a man-

²This study was conducted under Northeastern University IRB protocol #12-08-42; all personally identifiable information was removed from the dataset.

³I make the source code for this page available to the research community so that my experiment can easily be replicated.

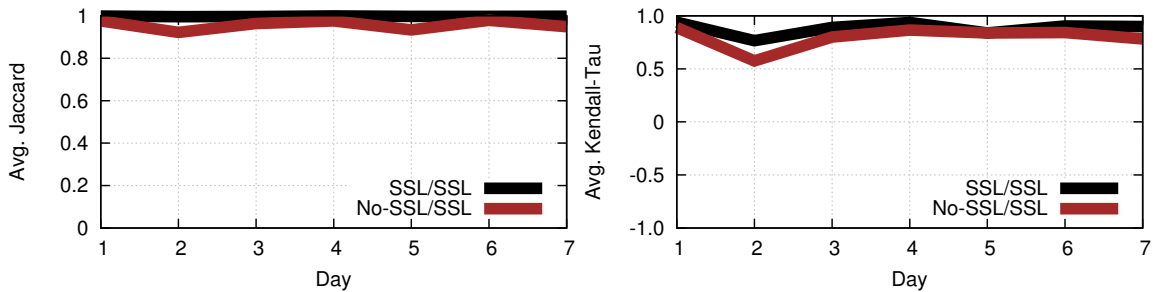


Figure 4.5: Results for the no-SSL versus SSL experiment on Google Search.

in-the-middle attack. Unfortunately, result pages from HTTP Google Search include a disclaimer explaining that some types of search personalization are disabled for HTTP results.

To understand if the differences between SSL and no-SSL Google Search are significant, I conducted a simple pilot study. I automated three Firefox browsers to execute our 120 search queries every day for seven days. Two of the browsers searched using `https://google.com`, and the third searched on `http://google.com` (i.e., SSL search serves as the control for this experiment). The three browsers were sandboxed so they could not influence each other (e.g., via cached files or cookies), and all cookies and history were cleared from the browsers before beginning the experiment.

Figure 4.5 shows the average Jaccard Index and average Kendall’s Tau for each day of test results. Both quantities are averaged over all 120 queries. The “SSL/SSL” line compares the results received by the two accounts that searched using `https://google.com`. As expected, the results received by the accounts have the same composition (i.e., Jaccard Index is 0.998 on average), although the order of results is somewhat noisy (i.e., Kendall’s Tau is 0.88 on average). The “No-SSL/SSL” line compares the results received by the account that searched using `http://google.com` to an account that searched using `https://google.com`. The results show that there are consistent, but minor, differences between the composition and ordering of the two search results. Average Jaccard and Kendall’s Tau are 0.95 and 0.79 for the the no-SSL/SSL experiments, respectively.

The takeaway from Figure 4.5 is that there are slight differences in the search results from SSL and no-SSL Google Search. However, the variation induced by noise is greater than the variation induced by the presence or absence of encryption. Thus, I feel that the experimental methodology used in this section is sound overall, because I am able to control for changes in search results due to noise.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

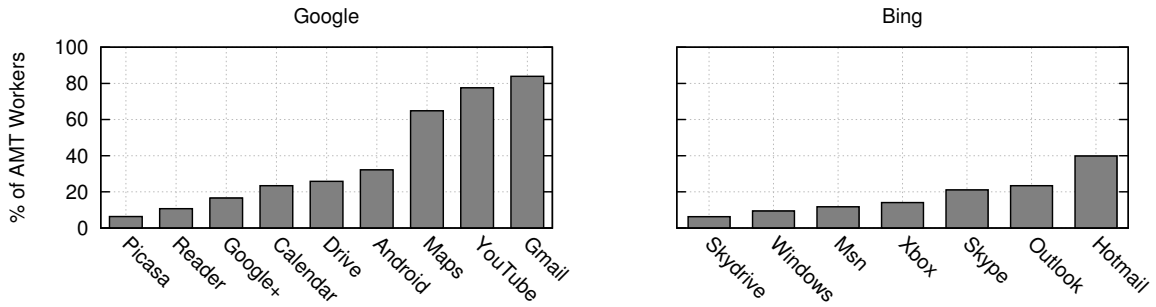


Figure 4.6: Usage of Google/Microsoft services by AMT workers.

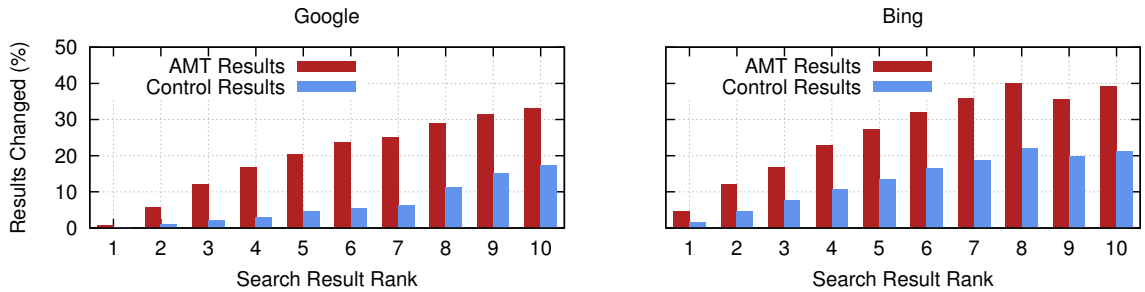


Figure 4.7: % of AMT and control results changed at each rank.

Alternate Methodologies Other researchers have developed alternate techniques to compare search results across users. For example, the authors of the “Bing it On” study [16] had users take screenshots of search results and uploading them to the experimenters. I found such an approach to be a poor fit for our experimental goals, as requesting users to submit screenshots for every search would (a) significantly reduce the coverage of search terms (since users would have to manually upload screenshots, instead of the searches being automatic) and (b) make it more difficult to control for noise (since it would not be possible to run the user query and the control query in lock-step).

AMT Worker Demographics In total, I recruited 300 AMT workers, 200 for our Google Search experiment and 100 for our Bing experiment. The reason for fewer users in the Bing experiment is that I were only able recruit 100 AMT workers who hold Bing accounts (it appears that Bing accounts are much less common). In both experiments, the participants first answered a brief demographic survey. Our participants self-reported to residing in 43 different U.S. states, and range in age from 18 to 66 (with a bias towards younger users). Figure 4.6 shows the usage of Google and Microsoft

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

Table 4.2: Top 10 most/least personalized queries on Google Search and Bing.

Most Personalized		Least Personalized	
Google	Bing	Google	Bing
1. gap	harry	what is gout	what is vegan
2. hollister	2008 crysis	dance with dragons	theadvocate
3. hgtv	nuclear weapon	what is lupus	arash molavi
4. boomerang	witch	gila monster facts	hollister
5. home depot	job creation	what is gluten	osteoporosis
6. greece	tax cuts	ipad 2	what is gluten
7. pottery barn	issue	cheri daniels	hot to dispose of paint
8. human rights	abortion	psoriatic arthritis	wild kratts
9. h2o	iran and isreal	keurig coffee maker	gap
10. nike	obama	maytag refrigerator	amana refrigerator

services by our participants. For Google, 84% are Gmail users, followed by 76% that use YouTube, while for Bing 40% are Hotmail users. These survey results demonstrate that our participants 1) come from a broad sample of the U.S. population, and 2) use a wide variety of Google and Microsoft services. The low usage of Microsoft Windows may be due to issues experienced by Internet Explorer users: written feedback from several of our participants indicated that Internet Explorer users found it difficult to set up the necessary proxy settings for our tasks.

4.3.2 Results

I now pose the question: *how often do real users receive personalized search results?* To answer this question, I compare the results received by AMT users and the corresponding control accounts. Figure 4.7 shows the percentage of results that differ at each rank (i.e., result 1, result 2, etc.) when I compare the AMT results to the control results, and the control results to each other. Intuitively, the percent change between the controls is the noise floor; any change above the noise floor when comparing AMT results to the control can be attributed to personalization.

There are three takeaways from Figure 4.7. First, I observe extensive personalization of search results. On average, across all ranks, AMT results showed an 11.7% *higher* likelihood of differing from the control result than the controls results did from each other on Google Search, and 15.8% higher likelihood on Bing. This additional difference can be attributed to personalization. To make sure these differences between the AMT and the control results are in fact statistically significant (and not just a reflection of the sampling), I perform the Chi squared test. I calculate the *p*-value for each rank for both Bing and Google; I find all of the *p*-values to be lower than

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

0.0001, indicating the the results are statistically significant. Second, as already indicated, I observe more personalization on Bing than on Google Search. Third and finally, top ranks tend to be less personalized than bottom ranks on both search engines.

To better understand how personalization varies across queries, I list the top 10 most and least personalized queries on Google Search and Bing in Table 4.2. The level of personalization per query is calculated as the probability of AMT results equaling the control results, minus the probability of the control results equaling each other. Large values for this quantity indicate large divergence between AMT and control results, as well as low noise (i.e., low control/control divergence).

As shown in Table 4.2, the most personalized queries on Bing tend to be related to important political issues (e.g., “job creation” and “tax cuts”) whereas on Google the most personalized queries tend to be related to companies and politics (e.g., “greece”, “human rights,” and “home depot”). In contrast, the least personalized results on both search engines are often factual (“what is”) and health related queries.

I manually examined the most personalized results and observed that most of the personalization on Google is based on location. Even though all of the AMT users’ requests went through my proxy and thus appeared to Google as being from the same IP address, Google Search returned results that are specific to other locations. This was especially common for company names, where AMT users received results for different store locations.

4.4 Personalization Features

In the previous section, I observed significant personalization for real users on Google Search and Bing. I would now like to explore which user *features* (i.e., aspect of the users’ profile or activity) are most likely to lead to personalized results. To do so, I am unable to use existing real user accounts as I did before, as the history of profile attributes and activity of these accounts are unknown to us. Instead, I create new, synthetic accounts under my control, and use these accounts (whose entire history I do know) to determine which features are most influential.

Although I cannot possibly enumerate and test all possible user features, I can investigate likely candidates. To do so, I enumerated the list of user features that (a) have been suggested in the literature as good candidates for personalization and (b) are possible to emulate given the constraints of my experimental methodology; I discuss the user features I was not able to explore in Section 4.6. Table 6.3 lists the different user features that my experiments emulate, as well as which search engines each user feature was evaluated on.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

Table 4.3: User features evaluated for effects on search personalization.

Category	Feature	Tested On	Tested Values
Tracking	Cookies	G, B	Logged In, Logged Out, No Cookies
User-Agent	OS	G, B, D	Win. XP, Win. 7, OS X, Linux
	Browser	G, B, D	Chrome 22, Firefox 15, IE 6, IE 8, Safari 5
Geolocation	IP Address	G, B, D	MA, PA, IL, WA, CA, UT, NC, NY, OR, GA
User Profile	Gender	G, B	Male, Female, Other
	Age	G, B	15, 25, 35, 45, 55, 65
	Zip Code	B	MA, CA, FL, TX, WA
Search History, Click History, and Browsing History	Gender	G, B	Male, Female
	Age	G, B	<18, 18-24, 25-34, 35-44, 45-54, 55-64, ≥65
	Income	G, B	\$0-50K, \$50-100K, \$100-150K, >\$150K
	Education	G, B	No College, College, Grad School
	Ethnicity	G, B	Caucasian, African American, Asian, Hispanic

4.4.1 Collecting Synthetic Account Data

For each user feature I wish to examine, I create $x + 1$ fresh user accounts, where x equals the number of possible values of the feature I are testing in that experiment, plus one additional *control account*. I refer to all non-control accounts as *test accounts*. For example, in the Gender experiment, I create four accounts in total: three test accounts (one “male,” one “female,” one “other”) and one control account (“female”). We execute $x + 1$ instances of my PhantomJS script for each experiment (one for each account), and forward the traffic to $x + 1$ unique endpoints via SSH tunnels. Each account searches for all 120 of my queries, and I repeat this process daily for 30 days. This complete treatment is conducted on Google, Bing, and DuckDuckGo (depending on the particular feature under analysis). As before, I compare the differences in the results between the control account and its counterpart (in my example above, the two “female” accounts) to measure the baseline noise; I then compare the differences in the results between the test accounts and the control to measure personalization.

It is important to note that I can not compare results across search engines given that their coverage on different topics might vary; thus, my measurements aim for capturing the personalization level within each search engine.

4.4.2 Basic Features

I begin my experiments by focusing on features associated with a user’s browser, their physical location, and their user profile.

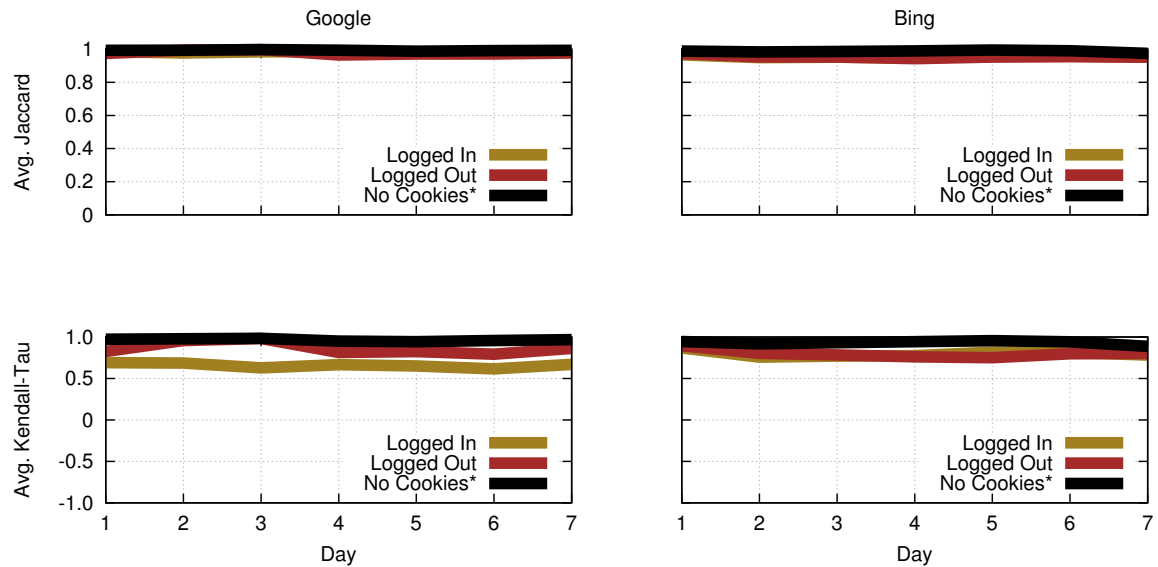


Figure 4.8: Results for the cookie tracking experiments on Google and Bing.

Basic Cookie Tracking In this experiment, the goal is to compare the search results for users who are logged in to a Google/Bing account, not logged in, and who do not support cookies at all. Google and Bing are able to track the logged in and logged out users, since both search engines place tracking cookies on all users, even if they do not have a user account. The user who does not support cookies receives a new tracking cookie after every request, and I confirm that the identifiers in these cookies are unique on every request. However, it is unknown whether Google or Bing are able to link these new identifiers together behind-the-scenes (e.g., by using the user’s IP address as a unique identifier).

To conduct this experiment, I use four instances of PhantomJS per search engine. The first two completely clear their cookies after every request. The third account logs in to Google/Bing and persists cookies normally. The fourth account does not log in to Google/Bing, and also persists cookies normally.

Figure 4.8 shows the results of my experiments. The upper left plot shows the average Jaccard Index for each account type (logged in/logged out/no cookies) across all search queries on Google when compared to the control (no cookies). In all of my figures, I place a * on the legend entry that corresponds to the control test, i.e., two accounts that have identical features. The figure reveals that the results received by users are not dependent on whether they support cookies, or their login state with Google. However, just because the results are the same, does not mean that they are returned in the same order.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

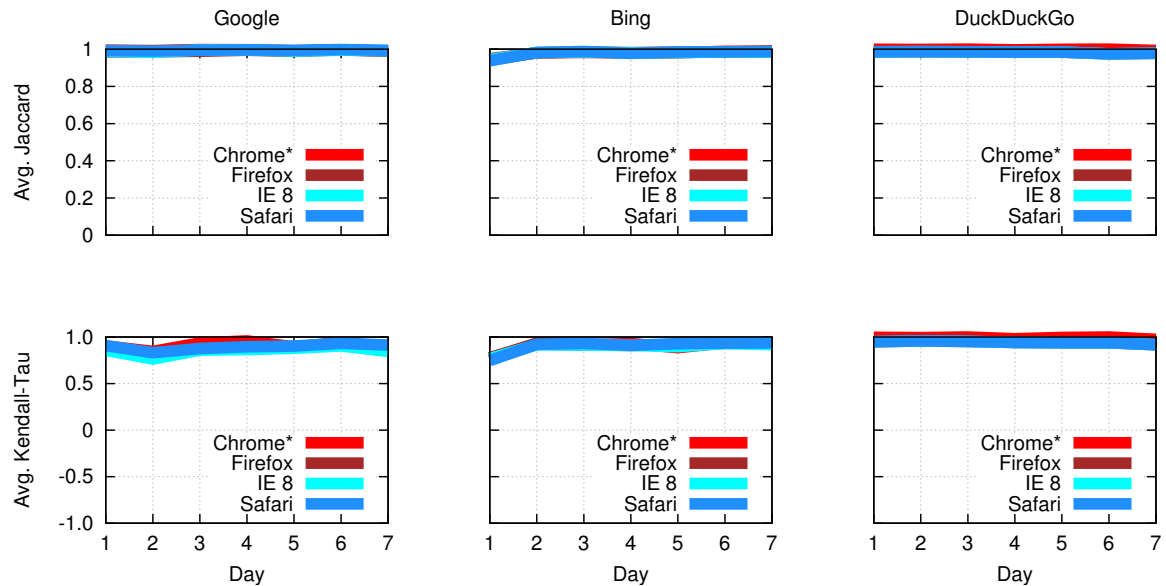


Figure 4.9: Results for the browser experiments on Google, Bing, and DuckDuckGo.

To examine how the order of results changes, I plot the average Kendall’s tau coefficient between each account type versus the control on Google in the lower left plot of Figure 4.8. I observe that a user’s login state and cookies do impact the order of results from Google Search. The greatest difference is between users who are logged in versus users that clear their cookies. Logged in users receive results that are reordered in two places (on average) as compared to users with no cookies. Logged out users also receive reordered results compared to the no cookie user, but the difference is smaller. The results in this figure are consistent with the techniques that search engines are likely to use for personalization (i.e., per-user cookie tracking), and give the first glimpse of how Google alters search results for different types of users.

The right column of Figure 4.8 examines the impact of login cookies on Bing. From the upper right figure (which plots the average Jaccard Index), I see that, unlike Google Search, having Bing cookies does impact the results returned from Bing. The lower right plot in Figure 4.8 (which plots the average Kendall’s tau coefficient) demonstrates that cookies also influence the order of results from Bing.

I did not run my cookie-based experiments against DuckDuckGo because it does not place cookies on users’ browsers.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

Browser User-Agent Next, I examine whether the user’s choice of browser or Operating System (OS) can impact search results. To test this, I created 22 user accounts (11 for Google, 11 for Bing) and assigned each one a different “User-Agent” string. As shown in Table 6.3, I encoded user-agents for 5 browsers and 4 OSs. Chrome 22 and Windows 7 serve as the controls. For DuckDuckGo, I conduct the same experiment sans user accounts, since DuckDuckGo does not have support for user accounts.

Figure 4.9 shows the results for my browser experiments on Google, Bing, and DuckDuckGo. Unlike the cookie tracking experiment, there is no clear differentiation between the different browsers and the control experiment. The results for different OSs are similar, and we omit them for brevity. Thus, I do not observe search personalization based on user-agent strings for Google, Bing, or DuckDuckGo.

IP Address Geolocation Next, I investigate whether the three target search engines personalize results based on users’ physical location. To examine this, I create 22 user accounts (11 Google, 11 Bing) and run our test suite while forwarding the traffic through SSH tunnels to 10 geographically diverse PlanetLab machines. These PlanetLab machines are located in the US states shown in Table 6.3. Two accounts forward through the Massachusetts PlanetLab machine, since it is the control. As before, I conduct this experiment against DuckDuckGo sans user accounts.

Figure 4.10 shows the results of our location tests. There is a clear difference between the control and all the other locations on both Google and Bing. On Google Search, the average Jaccard Index for non-control tests is 0.91, meaning that queries from different locations generally differ by one result. The same is true on Bing, where the average Jaccard Index is 0.87. The difference between locations is even more pronounced when I consider result order: the average Kendall’s tau coefficient for non-control accounts is 2.12 and 1.94 on Google and Bing, respectively.

These results reveal that Google Search and Bing do personalize results based on the user’s geolocation. One example of this personalization can be seen by comparing the MA and CA Google Search results for the query “pier one” (a home furnishing store). The CA results include a link to a local news story covering a store grand opening in the area. In contrast, the MA results include a Google Maps link and a CitySearch link that highlight stores in the metropolitan area.

In contrast to Google and Bing, the search results from DuckDuckGo are essentially identical regardless of the user’s IP address. This result is not surprising, since it fits with DuckDuckGo’s stated policy of not personalizing search results for any reason.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

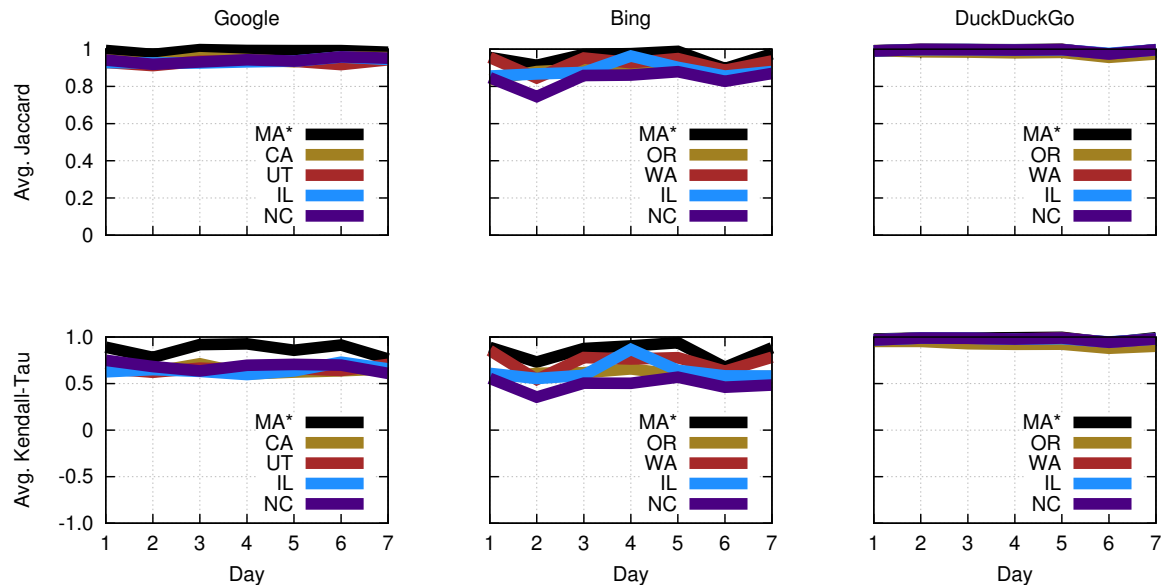


Figure 4.10: Results for the geolocation experiments on Google, Bing, and DuckDuckGo.

Inferred Geolocation During my experiments, I observed one set of anomalous results from experiments that tunneled through Amazon EC2. In particular, 9 machines out of 22 rented from Amazon’s North Virginia datacenter were receiving heavily personalized results, versus the other 13 machines, which showed no personalization. Manual investigation revealed that Google Search was returning results with `.co.uk` links to the 9 machines, while the 13 other machines received zero `.co.uk` links. The 9 machines receiving UK results were all located in the same /16 subnet.

Although I could not determine why Google Search believes the 9 machines are in the UK (I believe it is due to an incorrect IP address geolocation database), I did confirm that this effect is independent of the Google account. As a result, I did not use EC2 machines as SSH tunnel endpoints for any of the results in this paper. However, this anomaly does reveal that Google returns dramatically different search results to users who are in different countries (or in this case, users Google believes are in different countries).

User Profile Attributes In my next set of tests, I examine whether Google Search and Bing uses demographic information from users’ profiles to personalize results. Users must provide their gender and age when they sign up for a Google account, which means that Google Search could leverage this information to personalize results. Bing, on the other hand, collects gender, age, and zip code.

To test this hypothesis, I created Google and Bing accounts with specific demographic

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

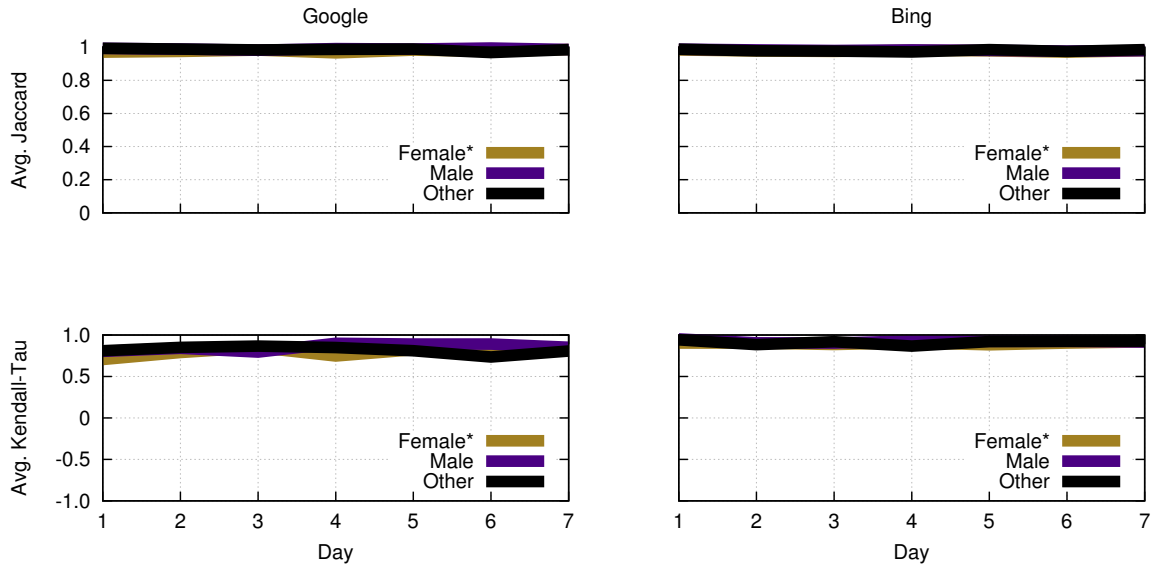


Figure 4.11: Results for the User Profile: Gender experiments on Google and Bing.

qualities. As shown in Table 6.3, I created “female,” “male,” and “other” accounts (these are the 3 choices Google and Bing give during account sign-up), as well as accounts with ages 15 to 65, in increments of 10 years. On Bing, we also create accounts from five different zip codes. The control account in the gender tests is female, the control in the age tests is 15, and the control in the zip code test is in Massachusetts.

The results for the gender test are presented in Figure 4.11 I do not observe user profile gender-based personalization on Google or Bing. Similarly, I do not observe personalization based on profile age or zip code, and I omit the results for brevity. DuckDuckGo does not allow users to create user accounts, so I do not run these tests on DuckDuckGo.

4.4.3 Historical Features

I now examine whether Google Search and Bing use an account’s history of activity to personalize results. I consider three types of historical actions: prior searches, prior searches where the user clicks a result, and web browsing history.

To create a plausible series of actions for different accounts, I use data from Quantcast, a web analytics and advertising firm. Quantcast publishes a list of top websites (similar to Alexa) that includes the *demographics* of visitors to sites [128], broken down into the 20 categories shown in Table 6.3. Quantcast assigns each website a score for each demographic, where scores >100 indicate

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

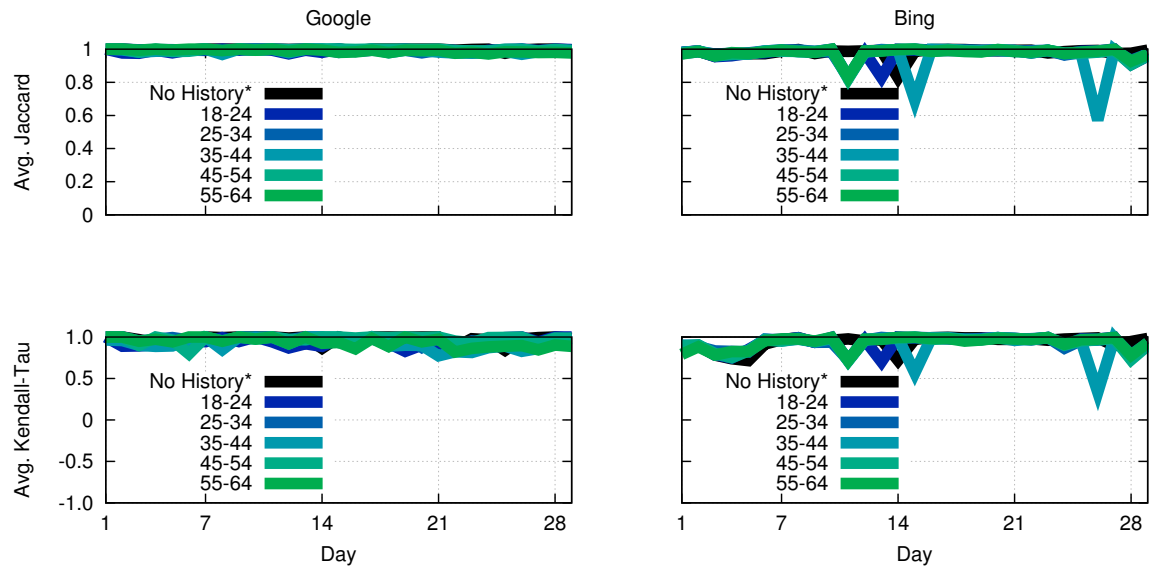


Figure 4.12: Results for the Search History: Age Bracket experiments on Google and Bing.

that the given demographic visits that website more frequently than average for the web. The larger the score, the more heavily weighted the site’s visitors are towards a particular demographic.

I use the Quantcast data to drive my historical experiments. In essence, my goal is to have different accounts “act” like a member of each of Quantcast’s demographic groups. The choice of my features was motivated by other web services and online advertisement services that use similar demographic categorizations to personalize content. Studies have shown that user demographics can be reliably inferred from browsing histories, which can be useful for personalizing content [57, 77]. Thus, for each of my experiments, I create 22 user accounts, two of which only run the 120 control queries, and 20 of which perform actions (i.e., searching, searching and clicking, or web browsing) based on their assigned demographic before running the 120 control queries. For example, one account builds web browsing history by visiting sites that are frequented by individuals earning $> \$150k$ per year. Each account is assigned a different Quantcast demographic, and chooses new action targets each day using weighted random selection, where the weights are based on Quantcast scores. For example, the $> \$150k$ browsing history account chooses new sites to browse each day from the corresponding list of URLs from Quantcast.

I execute all three experimental treatments (searching, searching and clicking, and web browsing) on Google Search, but only execute two (searching, and searching and clicking) on Bing. As previous studies have shown, Google is a ubiquitous presence across the web [135], which gives

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

Google the ability to track user's as they browse. In contrast, Microsoft and Bing do not have a widespread presence: out of 1500 top sites ranked by Quantcast, <1% include cookies from Microsoft or its subsidiaries (e.g., Live.com, Outlook.com), versus 63% for Google and its subsidiaries (e.g., YouTube, Doubleclick). Therefore, it is not feasible for Bing to track users' browsing behavior or personalize search results based on browsing history.

DuckDuckGo does not use cookies, and thus has no way to track users or build up history. Thus, I do not execute any of my historical experiments on DuckDuckGo.

Search History First, I examine whether Google Search and/or Bing personalize results based on search history. Each day, the 40 test accounts (20 for Google, 20 for Bing) search for 100 demographic queries before executing the standard 120 queries. The query strings are constructed by taking domains from the Quantcast top-2000 that have scores >100 for a particular demographic and removing subdomains and top level domains (e.g., `www.amazon.com` becomes "amazon").

Figure 4.12 shows the results of the search history test for five different age brackets. The "No History" account does not search for demographic queries, and serves as the control. The vast majority of the time, all accounts receive almost identical search results across both search engines (except for a few, random outliers in the Bing results). If Google or Bing was personalizing search results based on search history, I would expect the results for the age bracket accounts to diverge from the control results over time. However, I do not observe this over the course of 30 days of experiments. This observation holds for all of the demographic categories I tested, and I omit the results for brevity. Thus, I do not observe personalization based on search history, although it is possible that longer experiments could show larger differences.

Search-Result-Click History Next, I examine whether Google Search and/or Bing personalizes results based on the search results that a user has clicked on. I use the same methodology as for the search history experiment, with the addition that accounts click on the search results that match their demographic queries. For example, an account that searches for "amazon" would click on a result linking to `amazon.com`. Accounts will go through multiple pages of search results to find the correct link for a given query.

The results of the click history experiments are the same as for the search history experiments. There is little difference between the controls and the test accounts, regardless of demographic. Thus, I do not observe personalization based on click history, and I omit the results for brevity.

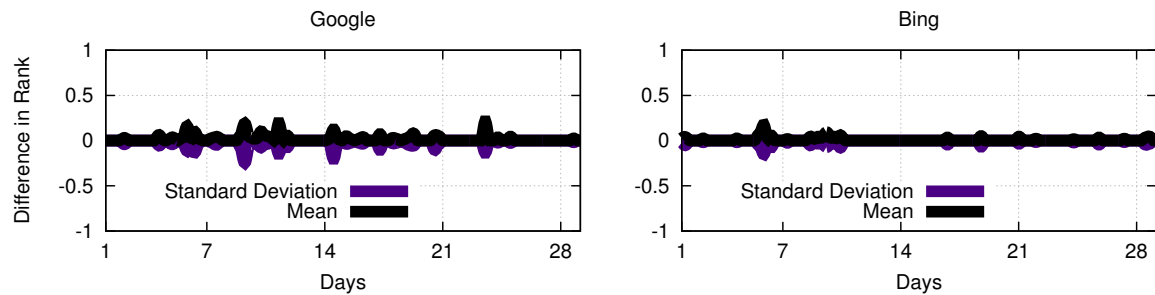


Figure 4.13: Results for the targeted domain clicking experiments on Google and Bing.

Browsing History Next, I investigate whether Google Search personalizes results based on web browsing history (i.e., by tracking users on third-party web sites). In these experiments, each account logs into Google and then browses 5 random pages from 50 demographically skewed websites each day. I filter out websites that do not set Google cookies (or Google affiliates like DoubleClick), since Google cannot track visits to these sites. Out of 1,587 unique domains in the Quantcast data that have scores >100 , 700 include Google tracking cookies.

The results of the browsing history experiments are the same as for search history and click history: regardless of demographic, I do not observe personalization. I omit these results for brevity.

Targeted Domain Clicking Finally, I conduct a variant of my click history experiment. In the previous search-result-click experiment, each account executed 100 “demographic” searches and 120 standard test queries per day. However, it is possible that this methodology is too complex to trigger search personalization, i.e., because each account creates such a diverse history of searches and clicks, the search engines may have trouble isolating specific features to personalize on.

Thus, in this experiment, I simplify my methodology: I create 10 accounts, each of which is assigned a specific, well-known news website. Each account executes 6 news-related queries 4 times on each day (so, 24 searches each day, evenly spaced throughout the day). After searching the account clicks on the link that is its assigned news website in the list of results. For example, one account was assigned `www.foxnews.com`; 24 times per day this account executed news-related queries, and always clicked on results pointing to `www.foxnews.com` (if they appeared in the top 10 results). In theory, this creates a very strong signal for personalization, i.e., a search engine could trivially observe that this user favors a specific website, and increase the rank of results pointing to this website.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

I conduct the targeted domain clicking test on both Google Search and Bing. We created 10 experimental accounts on each search engine, each of which was assigned a unique target domain, as well as 1 control account that searches but does not click on any links.

Figure 4.13 shows the results of my targeted domain clicking experiments. To quantify my results, I plot the average difference in rank between the targeted domains as seen by the experimental accounts and the control account. Difference of zero means that a particular domain appears at the same rank for both the experimental account (which clicks on the domain) and the control (which clicks on nothing). Positive difference in rank means the domain appears at higher ranks for the experimental account, while a negative difference means that the domain appears at higher ranks for the control.

As shown in Figure 4.13, on average, there is close to zero difference between the ranks of domains, regardless of whether they have been clicked on. This result holds true across Google Search and Bing. As shown by the standard deviation lines, even when the rank of the domains differs, the variance is very low (i.e., less than a single rank difference). Furthermore, although this test was run for 30 days, I do not observe divergence over time in the results; if the search engines were personalizing results based on click history, I would expect the difference in rank to increase over time as the experimental accounts build more history. Thus, I conclude that clicking on results from particular domains does not cause Google or Bing to elevate the rank of that domain.

Discussion I was surprised that the history-driven tests did not reveal personalization on Google Search or Bing. One explanation for this finding is that account history may only impact search results for a brief time window, i.e., carry-over is the extent of history-driven personalization on these search engines.

4.5 Quantifying Personalization

In the previous section we demonstrate that Google Search personalization occurs based on 1) whether the user is logged in and 2) the location of the searching machine. In this section, we dive deeper into the data from our synthetic experiments to better understand how personalization impacts search results. First, we examine the temporal dynamics of search results. Next, we investigate the amount of personalization in different categories of queries. Finally, we examine the rank of personalized search results to understand whether certain positions are more volatile than others.

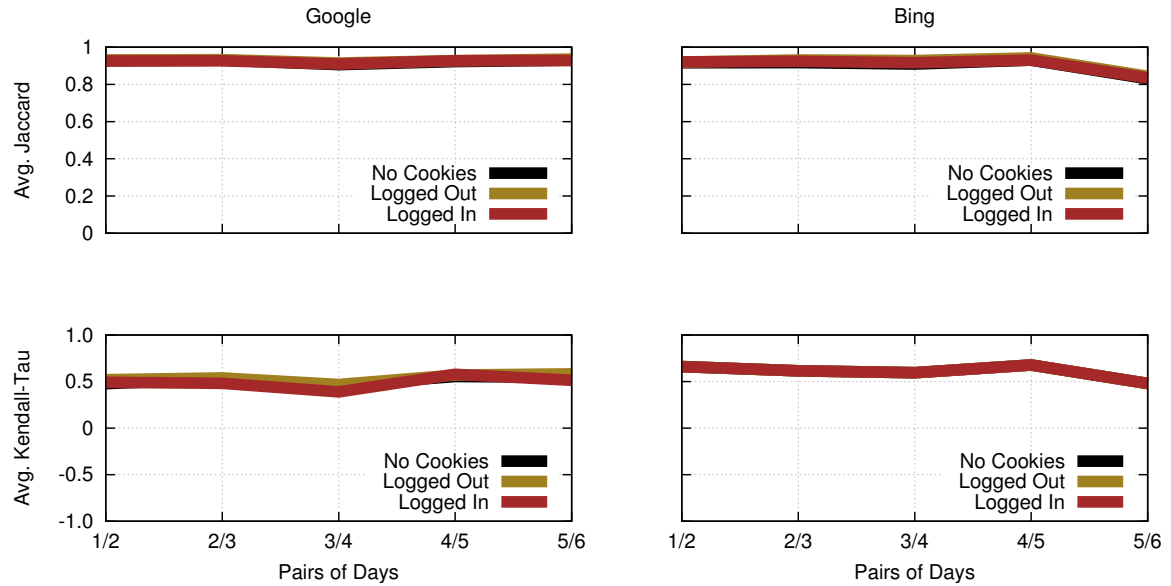


Figure 4.14: Day-to-day consistency of results for the cookie tracking experiments.

4.5.1 Temporal Dynamics

In this section, we examine the temporal dynamics of results from Google Search and Bing to understand how much search results change day-to-day, and whether personalized results are more or less volatile than non-personalized search results. To measure the dynamics of search engines over time, we compute the Jaccard Index and Kendall Tau coefficient for search results from subsequent days. Figure 4.14 shows the day-to-day dynamics for our cookie tracking experiment (i.e., the accounts are logged in, logged out, and do not support cookies, respectively). The x-axis shows which two days of search results are being compared, and each line corresponds to a particular test account.

Figure 4.14 reveals three facts about Google Search and Bing. First, the lines in Figures 4.14 are roughly horizontal, indicating that the rate of change in the search indices is roughly constant. Second, we observe that there is more reordering over time than new results: average Jaccard Index on Google and Bing is ≈ 0.9 , while average Kendall Tau coefficient is 0.5 for Google and 0.7 for Bing. Third, we observe that both of these trends are consistent across all of our experiments, irrespective of whether the results are personalized. This indicates that personalization does not increase the day-to-day volatility of search results.

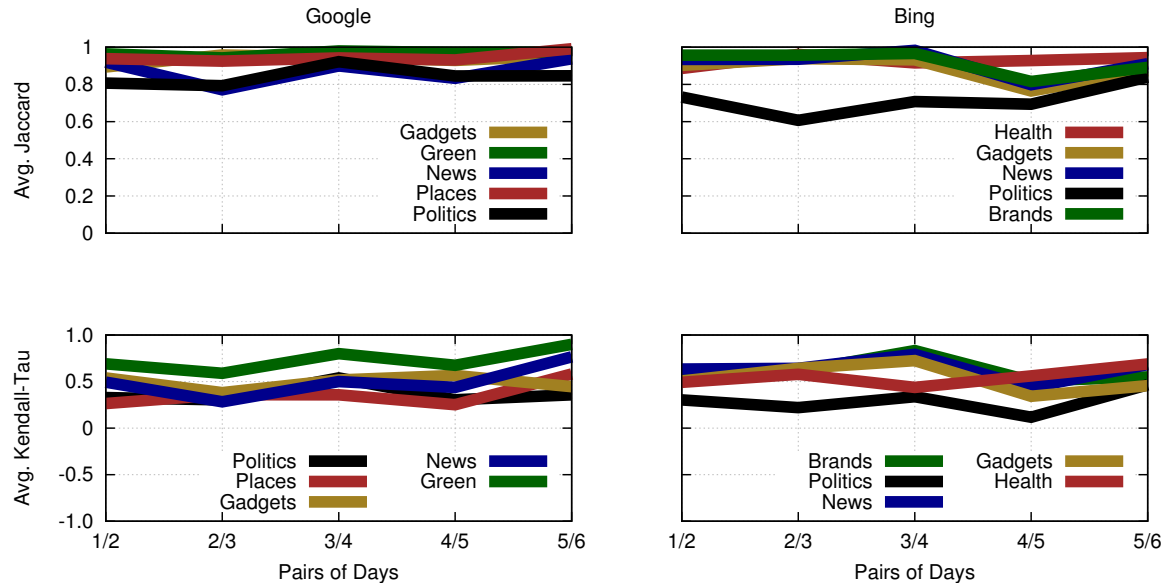


Figure 4.15: Day-to-day consistency within search query categories for the cookie tracking experiments.

Dynamics of Query Categories We now examine the temporal dynamics of results across different categories of queries. As shown in Table 4.1, we use 12 categories of queries in our experiments. Our goal is to understand whether each category is equally volatile over time, or whether certain categories evolve more than others.

To understand the dynamics of query categories, we again calculate the Jaccard Index and Kendall Tau coefficient between search results from subsequent days. However, instead of grouping by experiment, we now group by query category. Figure 4.15 shows the day-to-day dynamics for query categories during our cookie tracking experiments. Although we have 12 categories in total, Figure 4.15 only shows the 1 least volatile, and 4 most volatile categories, for clarity. The results for all other experiments are similar to the results for the cookie tracking test, and we omit them for brevity.

Figure 4.15 reveals that the search results for different query categories change at different rates day-to-day. For example, there are more new results per day for “politics” related-queries on both Google Search and Bing. Similarly, “politics” and “gadgets” related-queries both exhibit above average reordering each day. This reflects how quickly information in these categories changes on the web. In contrast, search queries in factual categories like “what is” and “green” (environmentally friendly topics) are less volatile over time (and are omitted from Figure 4.15 for clarity).

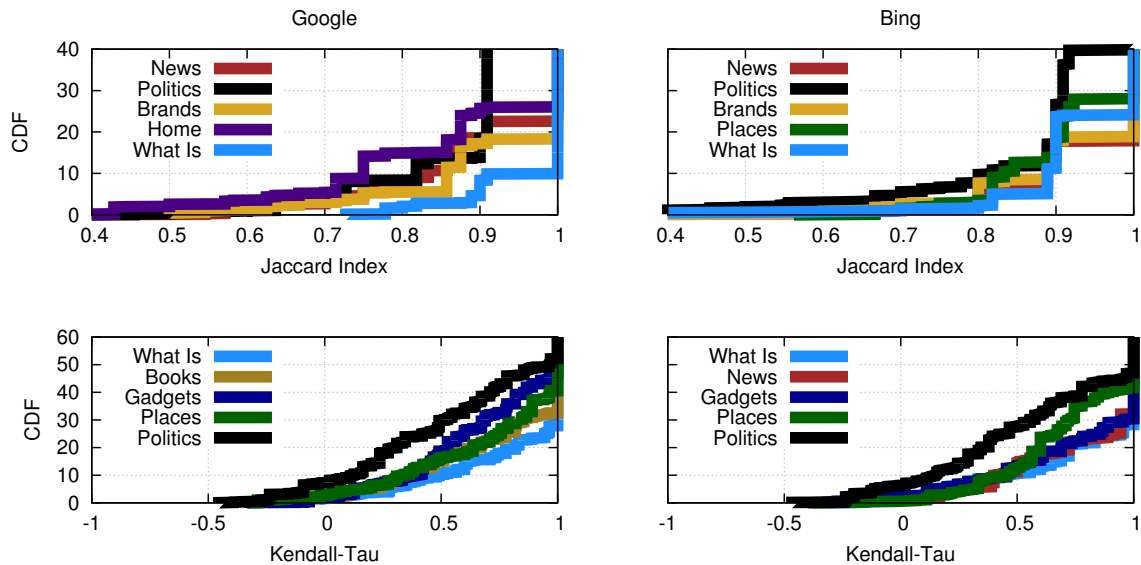


Figure 4.16: Differences in search results for five query categories on Google Search and Bing.

4.5.2 Personalization of Query Categories

We now examine the relationship between different categories of search queries and personalization. In Section 4.4, we demonstrate that Google Search and Bing do personalize search results. However, it remains unclear whether all categories of queries receive equal amounts of personalization.

To answer this question, we plot the cumulative distribute of Jaccard Index and Kendall Tau coefficient for each category in Figure 4.16. These results are calculated over all of our experiments (i.e., User-Agent, Google Profile, geolocation, *etc.*) for a single day of search results. For clarity, we only include lines for the 1 most stable category (i.e., Jaccard index and Kendall Tau are close to 1), and the 4 least stable categories.

Figure 4.16 demonstrates that Google Search and Bing personalize results for some query categories more than others. For example, 88% of results for “what is” queries are identical on Google, while only 66% of results for “gadgets” are identical on Google. Overall, “politics” is the most personalized query category on both search engines, followed by “places” and “gadgets.” CDFs calculated over other days of search results demonstrate nearly identical results.

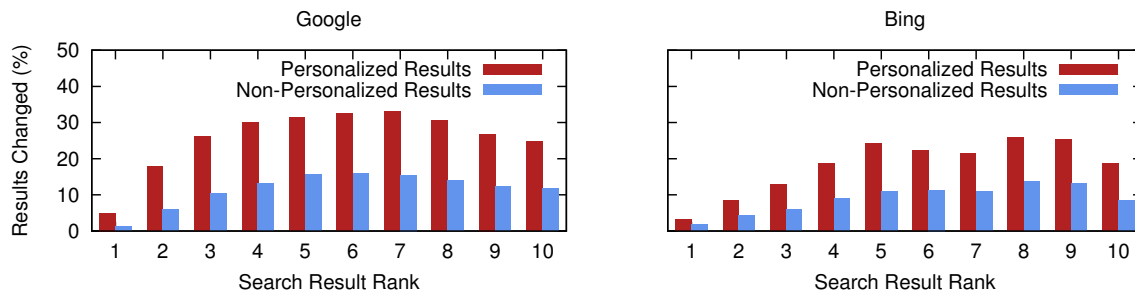


Figure 4.17: The percentage of results changed at each rank on Google Search and Bing.

4.5.3 Personalization and Result Ranking

In this section, we focus on the volatility of results from Google Search and Bing at each rank, with rank 1 being the first result on the page and rank 10 being the last result. Understanding the impact of personalization on top ranked search results is critical, since eye-tracking studies have demonstrated that users rarely scroll down to results “below the fold” [36, 60, 63, 99]. Thus, we have two goals: 1) to understand whether certain ranks are more volatile in general, and 2) to examine whether personalized search results are more volatile than non-personalized results.

To answer these questions, we plot Figure 4.17, which shows the percentage of results that change at each rank. To calculate these values, we perform a pairwise comparison between the result at rank $r \in [1, 10]$ received by a test account and the corresponding control. We perform comparisons across all tests in all experiments, across all seven days of measurement. This produces a total number of results that are changed at each rank r , which we divide by the total number of results at rank r to produce a percentage. The personalized results come from the cookie tracking and geolocation experiments; all others experimental results are non-personalized.

Figure 4.17 reveals two interesting features. First, the results on personalized pages are significantly more volatile than the results on non-personalized pages. The result changes on non-personalized pages represent the noise floor of the experiment; at nearly every rank, there are more than twice as many changes on personalized pages. Second, Figure 4.17 shows that the volatility at each rank is not uniform. Rank 1 exhibits the least volatility on Google Search and Bing. The volatility increases until it peaks at 33% in rank 7 on Google Search, and at 26% in rank 8 on Bing. This indicates that both search engines are more conservative about altering results at top ranks.

Given the extreme importance placed on rank 1 search results, we now delve deeper into the rare cases where the result at rank 1 changes during personalized searches (5% of personalized

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

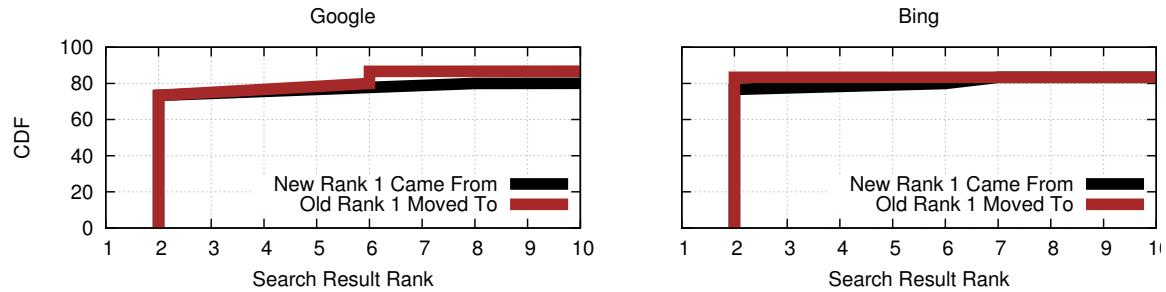


Figure 4.18: Movement of results to and from rank 1 for personalized searches.

rank 1 results change on Google, while 3% change on Bing). In each instance where the rank 1 result changes, we compare the results for the test account and the control to determine 1) what was the *original rank* of the result that moved to rank 1, and 2) what is the *new rank* of the result that used to be at rank 1.

Figure 4.18 plots the results of this test. In the vast majority of cases, the rank 1 and 2 results switch places: on Google, 73% of new rank 1 results originate from rank 2, and 58% of old rank 1 results move to rank 2. On Bing, 77% of new rank 1 results originate from rank 2, and 83% of old rank 1 results move to rank 2. Overall, on Google, 93% of new rank 1 results come from the first page of results, while 82% of old rank 1 results remain somewhere on the first result page. On Bing, 83% percent of rank 1 results come from or move to somewhere on the first page of results. However, none of the CDFs sum to 100%, i.e., there are cases where the new rank 1 result does not appear in the control results and/or the old rank 1 result disappears completely from the test results. The latter case is more common on Google, with 18% of rank 1 results getting evicted completely from the first page of results. Both cases, are equally likely on Bing.

Figure 4.18 reveals similarities and differences between Google Search and Bing. On one hand, both search engines are clearly conservative about changing rank 1 search results, i.e., the vast majority of changes are simply swaps between rank 1 and 2. On the other hand, when the rank 1 result does change, Google and Bing leverage different strategies: Google Search prefers to elevate results already on the first page, while Bing prefers to insert completely new links. We manually examined instances where Bing inserted new results at rank 1, and found that in most cases these new links were to Bing services, e.g., a box of links Bing News results.

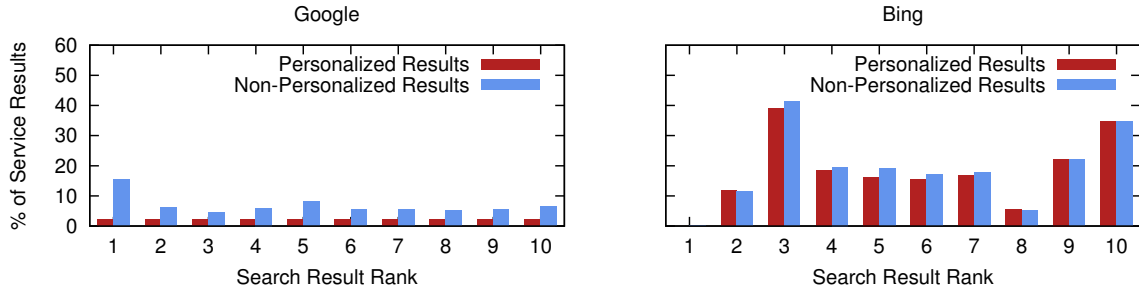


Figure 4.19: Rank of embedded services in search results from Google and Bing.

4.5.4 Personalization and Aggregated Search

In Section 4.2.1, we noted that some of the results from search engines do not point to third-party websites. Instead, some results embed links and functionality from *services* maintained by the search engine provider. The inclusion of links to other services in search results is sometimes referred to as “aggregated search.” For example, Google often embeds results from Google News, Google Maps, YouTube, and Google+ into pages of search results. Figure 4.1 shows an example of an embedded Google service: a box of queries that are “Related” to the given search query. Bing offers an array of similar services and embeds to them in search results.

In this section, we examine links to provider services in search results. Specifically, we are interested in whether personalization impacts the placement and amount of links to provider services. These are important questions, given that regulators have questioned the placement of provider services in search results within the context of antitrust regulation [134], i.e., do providers promote their own services at the expense of third-party websites?

First, we examine the percentage of results at each rank that embed provider services. Figure 4.19 shows the percentage of results at each rank that embed provider services on Google and Bing. We split our data into personalized and non-personalized pages of results, where results from the logged-in/out and location experiments are considered to be personalized. We aggregate across all 120 experimental queries and all 30 days of experiments.

Figure 4.19 demonstrates that Google and Bing exhibit different behavior when it comes to embedding provider services. Overall, Bing embeds its own services 19% of the time, whereas Google embeds its own services 9% of the time. However, Google embeds services at rank 1 on $\approx 15\%$ of result pages, whereas Bing *never* embeds services at rank 1. Instead, Google tends to embed services uniformly across ranks 2-10, whereas Bing favors embedding services at ranks 3 and

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

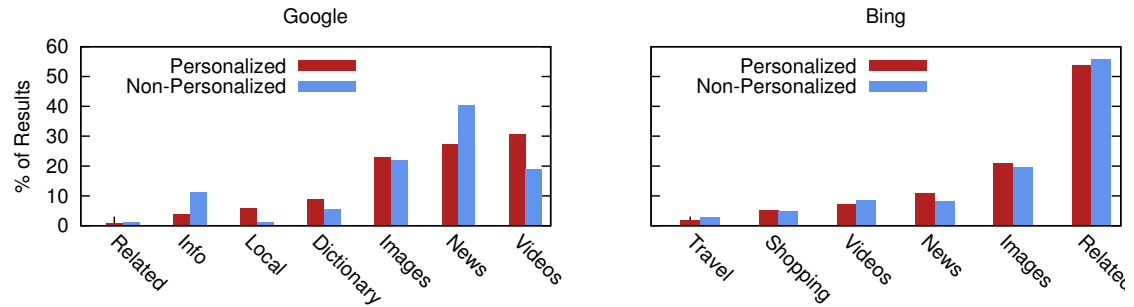


Figure 4.20: Percentage of embeddings of different services on Google and Bing.

10. On Bing, the rank 3 results point to a variety of different services (e.g., Bing Images, Bing Video, Bing News, *etc.*), while the service at rank 10 is almost always Related Searches.

Figure 4.19 also shows that personalization only appears to influence service embedding on Google. On Bing, the amount and placement of embedded services does not change between personalized and non-personalized search results. However, on Google, 12% of links on personalized pages point to services, versus 8% on non-personalized pages. This trend is relatively uniform across all ranks on Google. This demonstrates that personalization does increase the number of embedded services seen by Google Search users.

Next, we seek to understand whether personalization impacts which services are embedded by search engines. To answer this question, we plot Figure 4.20, which shows the percentage of results that can be attributed to different services (we only consider links to services, so the bars sum to 100%). As before, we examine results on personalized and non-personalized pages separately, and aggregate across all 120 search queries and all 30 days of experiments.

On Google, the top three most embedded services are Google Images, Google News, and Google Video (which also includes links to YouTube). Bing also generates many links to its equivalent image, news, and video services, however the most embedded service by a large margin is Related Searches (which is almost always placed at rank 10). In contrast, Google only embeds Related Searches into 1% of results pages.

Figure 4.20 reveals that Google does personalize the types of services it embeds, whereas Bing does not. On Google, “Info” and Google News results tend to be served on more non-personalized pages. Info results present information from Google’s Knowledge Graph, which are usually answers to questions, e.g., “Madrid” if the user searches for “Capital of Spain.” Conversely, “Local,” Dictionary, and Google Video results are served more frequently on personalized pages.

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

Local results present lists of geographically close stores and restaurants that are related to the user’s query, e.g., a list of pizza parlors if the user searches for “pizza.” In contrast to Google, Bing embeds different services at roughly static rates, regardless of personalization.

Google and Bing News As shown in Figure 4.20, pages of search results from Google and Bing often include embedded boxes of news from their respective news portals (Google News and Bing News). Each box is characterized by 1-4 links to news stories on third-party websites (e.g., CNN, New York Times, Fox News, *etc.*) that are relevant to the given query. Typically, one or more of the links will be enhanced with a thumbnail image also taken from a third-party website.

It is known that services like Google News use personalization to recommend news stories that the service believes are relevant to each user [38]. This raises the question: are the news links embedded in pages of search results also personalized? To answer this question, we went through our dataset and located all instances where an experimental account and the corresponding control account were both served embedded news boxes. In these cases, we compared the links in each embedded news box to examine whether news results are being personalized.

The results of this experiment show that search results from Google News and Bing News are not personalized, even if other results on the page are personalized. On both Google and Bing, the Jaccard Index when comparing news links from experimental and control accounts is consistently ≈ 1 , and the Kendall Tau coefficient is consistently ≈ 1 . These results are the same regardless of the characteristics of the experimental account (i.e., location, logged in/out, search history, *etc.*). Thus, it appears that Google and Bing only personalize news directly on their respective news sites, but not in their search results.

4.6 Concluding Discussion

Over the past few years, we have witnessed a trend of personalization in numerous Internet-based services, including web search. While personalization provides obvious benefits for users, it also opens up the possibility that certain information may be unintentionally hidden from users. Despite the variety of speculation on this topic, to date, there has been little quantification of the basis and extent of personalization in web search services today.

In this study, I introduce a robust methodology for measuring personalization on web search engines. My methodology controls for numerous sources of noise, allowing me to accurately measure the extent of personalization. I applied the methodology to real Google and Bing accounts

CHAPTER 4. MEASURING WEB SEARCH PERSONALIZATION

recruited from AMT and observe that 11.7% of search results on Google and 15.8% on Bing show differences due to personalization. Using artificially created accounts, I observe that measurable personalization on Google Search and Bing is triggered by 1) being logged in to a user account and 2) making requests from different geographic areas.

It is important to add that as a result of my methodology, I am only able to identify positive instances of personalization; I cannot claim the absence of personalization, as I may not have considered other dimensions along which personalization could occur and I can only test a finite set of search terms. However, the dimensions that I chose to examine in this paper are the most obvious ones for personalization (considering how much prior work has looked at demographic, location-based, and history-based personalization).

Given that any form of personalization is a moving target, this study aims to show that my methodology for measuring personalization can be used as a tool to successfully capture the effects of personalization algorithms at any given moment as well as to track their changes over time.

Chapter 5

The Impact of Geolocation on Web Search Personalization

5.1 Introduction

Motivated by the the findings of my first study, next I will focus on location based personalization in Google’s Web Search. In the first study I showed that Google infers users’ geolocation based on their IP address, and that *location-based personalization* caused more differences in search results than any other single feature 4.4. However, while these initial findings are intriguing, many questions remain, such as: does location-based personalization impact all types of queries (e.g., politics vs. news) equally? At what distance do users begin to see changes in search results due to location? Answering these questions is crucial, since users’ geolocation can be used as a proxy for other demographic traits, like race, income-level, educational attainment, and political affiliation. In other words, *does location-based personalization trap users in geolocal Filter Bubbles?*

In this study, I propose a novel methodology to explore the impact of location on Google Search results. I use the JavaScript `Geolocation` API [75] to present arbitrary GPS coordinates to the mobile version of Google Search. Google personalizes the search results based on the location we specified, giving me the ability to collect search results from any location around the globe. Although I focus on Google Search in the US, my methodology is general, and could easily be applied to other search engines like Bing.

Using my methodology, I collect 30 days of search results from Google Search in response to 240 different queries. By selecting 75 GPS coordinates around the US at three granularities

CHAPTER 5. THE IMPACT OF GEOLOCATION ON WEB SEARCH PERSONALIZATION

(county, state, and national), I am able to examine the relationship between distance and location-based personalization, as well as the impact of location-based personalization on different types of queries. I make the following observations:

- As expected, the differences between search results grows as physical distance between the locations of the users increases.
- However, the impact of location-based personalization changes depending on the query type. Queries for *politicians*' names (e.g., "Joe Biden") and *controversial* topics ("abortion") see minor changes, while queries for *local* terms ("airport") are highly personalized.
- Surprisingly, only 20-30% of differences are due to Maps embedded in search results. The remainder are caused by changes in "normal" search results.
- Also surprisingly, the search results for *local* terms are extremely noisy, i.e., two users making the same query from the same location at the same time often receive substantially different search results.

The content of this study was published in IMC2015 under the title "Location, Location, Location: The Impact of Geolocation on Web Search Personalization".

Outline. The rest of this chapter is organized as follows: in Section 5.2, I give an overview of my data collection methodology, and then present analysis and findings in Section 5.3 and finally I conclude in Section 5.4.

5.2 Methodology

my goal is to explore the relationship between geolocation and personalization on Google Search. Thus, I require the ability to send identical queries to Google Search, at the same moment in time, from different locations. In this section, I explain my methodology for accomplishing these goals. First, I introduce the locations and search terms used in my study. Next, I explain my technique for querying Google Search from arbitrary locations, and how I parsed Google Search results. Finally, I discuss how I quantify differences between pages of search results.

Progressive Tax
Impose A Flat Tax
End Medicaid
Affordable Health And Care Act
Fluoridate Water
Stem Cell Research
Andrew Wakefield Vindicated
Autism Caused By Vaccines
US Government Loses AAA Bond Rate
Is Global Warming Real
Man Made Global Warming Hoax
Nuclear Power Plants
Offshore Drilling
Genetically Modified Organisms
Late Term Abortion
Barack Obama Birth Certificate
Impeach Barack Obama
Gay Marriage

Table 5.1: Example *controversial* search terms.

5.2.1 Locations and Search Terms

Locations. First, I must choose the locations in which to execute queries. I decided to focus my study on Ohio, since it is known to be a “battleground” state in US politics. This property is important, since I want to examine whether demographics like political affiliation correlate with location-based personalization.

Overall, I picked 59 locations for my study spread across three *granularities*. For *nation*-level, I chose the centroids of 22 random states in the United States. For *state*-level, I chose the centroids of 22 random counties within Ohio. On average, these counties 100 miles apart. Finally, for *county*-level, I chose the centroids of 15 voting districts in Cuyahoga County, which is the most populous county in Ohio. On average, these voting districts are 1 mile apart. By examining locations in different granularities, I will be able to observe changes in search results across small, medium, and large-scale distances. This also gives us the ability to compare search results served in places with different demographics characteristics.

Search Terms. Next, I must select search terms for my study. I built a corpus of 240 queries that fall into three categories: 33 *local* queries, 87 *controversial* queries, and 120 names of *politicians*. *Local* queries correspond with physical establishments, restaurants, and public services such as

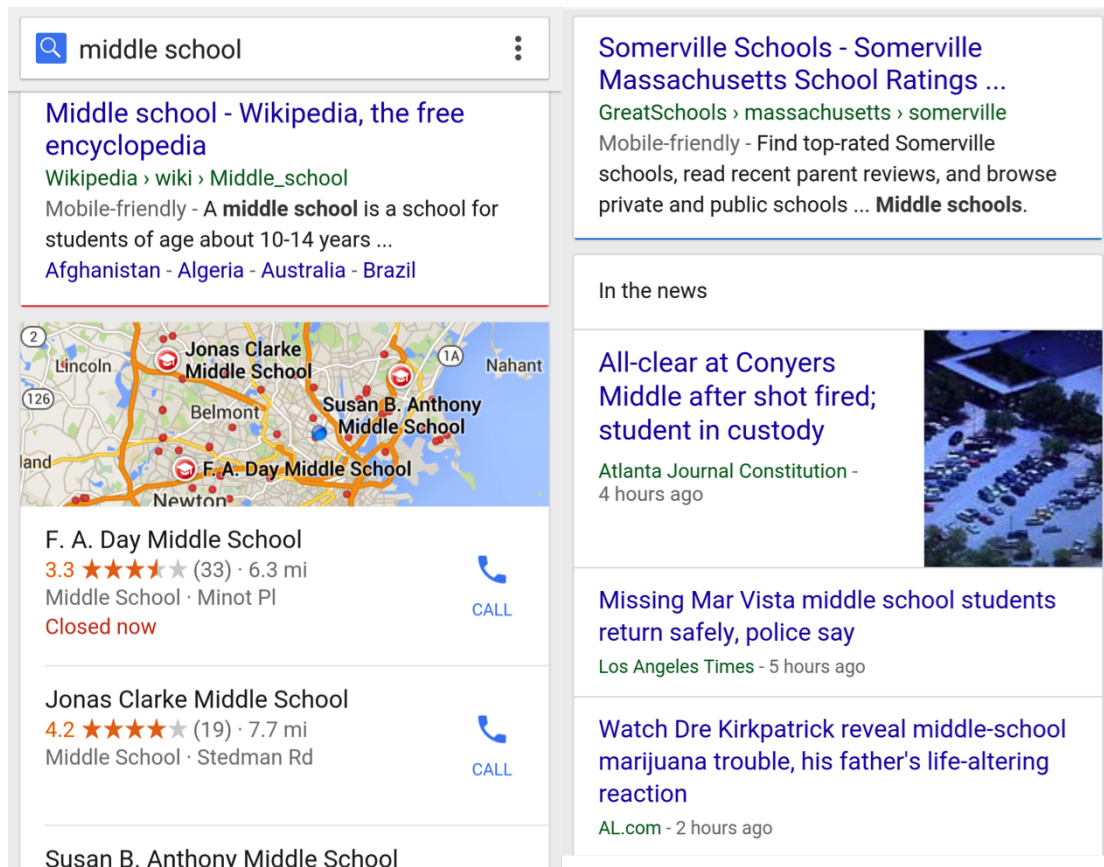


Figure 5.1: Example search results from the mobile version of Google Search.

“bank”, “hospital”, and “KFC”. I chose these terms because I expect them to produce search results that are heavily personalized based on location, i.e., I treat them as an upper-bound on location-based personalization. For *politicians*, I selected 11 members of the Cuyahoga County Board, 53 random members of the Ohio House and Senate, all 18 members of the US Senate and House from Ohio, 36 random members of the US House and Senate not from Ohio, Joe Biden, and Barack Obama. For national figures like Barack Obama, I do not expect to see differences in search results due to location; however, it is not clear how Google Search handles queries for state- and county-level officials inside and outside their home territories.

Finally, my *controversial* terms are news or politics-related issues like those shown in Table 5.1. I chose these terms because it would be concerning if Google Search personalized search results for them based on location. To avoid possible external bias, I picked search terms that, to the best of my knowledge, were not associated with specific news-worthy events at the time of my experiments. Although I cannot completely rule out the possibility that exonegous events impacted

the search results, I note that such an event would impact each treatment equally, and thus would likely not impact my findings.

5.2.2 Data Collection and Parsing

my methodology for gathering data from Google Search is based on the techniques presented in my prior work et al. [68, 69], with one key difference. As in prior work, I use PhantomJS [123] to gather data, since it is a full implementation of a WebKit browser. I wrote a PhantomJS script that takes a search term and a latitude/longitude pair as input, loads the mobile version of Google Search, executes the query, and saves the first page of search results.

Unlike prior work [68], I targeted the mobile version of Google Search because it uses the JavaScript `Geolocation` API [75] to query the user’s precise location. By overriding the `Geolocation` API in my PhantomJS script, I can feed the coordinates specified on the command line to Google Search, thus giving us the ability to run queries that appear to Google as if they are coming from any location of my choosing. I distributed my query load over 44 machines in a single /24 subnet to avoid being rate-limited by Google. Finally, all of my experimental treatments were repeated for 5 consecutive days to check for consistency over time.

Validation. To make sure that Google Search personalizes search results based on the provided GPS coordinates rather than IP address, I conducted a validation experiment. I issued identical controversial queries with the same exact GPS coordinate from 50 different Planet Lab machines across the US, and observe that 94% of the search results received by the machines are identical. This confirms that Google Search personalizes search results largely based on the provided GPS coordinates rather than the IP address. Furthermore, Google Search reports the user’s precise location at the bottom of search results, which enabled us to manually verify that Google was personalizing search results correctly based on my spoofed GPS coordinates.

Browser State. To control for personalization effects due to the state of the browser, all of my treatments were configured and behaved identically. The script presented the User-Agent for Safari 8 on iOS, and all other browser attributes were the same across treatments, so each treatment should present an identical browser fingerprint. Furthermore, I cleared all cookies after each query, which mitigates personalization effects due to search history, and prevents Google from “remembering” a treatments prior location. Lastly, I note that prior work has shown that Google Search does not personalize search results based on the user’s choice of browser or OS [68].

Controlling for Noise. Unfortunately, not all differences in search results are due to personalization; some may be due to noise. As in my prior work [68, 69], I take the following precautions to minimize noise:

1. All queries for term t are run in lock-step, to avoid changes in search results due to time.
2. I statically mapped the DNS entry for the Google Search server, ensuring that all my queries were sent to the same datacenter.
3. Google Search personalizes search results based on the user’s prior searches during the last 10 minutes [68]. To avoid this confound, I wait 11 minutes between subsequent queries.

However, even with these precautions, there may still be noise in search results (e.g., due to A/B testing). Thus, for each search term and location, I send two identical queries at the same time. By comparing each result with its corresponding *control*, I can measure the extent of the underlying noise. When comparing search results from two locations, any differences I see above the noise threshold can then be attributed to location-based personalization.

Parsing. As shown in Figure 5.1, Google Search on mobile renders search results as “cards”. Some cards present a single result (e.g., “Somerville Schools”), while others present a meta-result (e.g., locations from Google Maps or a list of “In the News” articles). In this study, I parse pages of search results by extracting the first link from each card, except for Maps and News cards where I extract all links. Thus, I observe 12–22 search results per page.

5.2.3 Measuring Personalization

As in my prior work [68], I use two metrics to compare pages of search results. First, I use *Jaccard Index* to examine the overlap: a Jaccard Index of 0 represents no overlap between the pages, while 1 indicates they contain the same search results (although not necessarily in the same order). Second, I use *edit distance* to measure reordering of search results. Edit distance calculates the number of additions, deletions, and swaps necessary to make two lists identical.

5.3 Analysis and Findings

Using the methodology described in Section 5.2, I collected 30 days of data from Google Search. I executed the 120 *local* and *controversial* queries once per day for five straight days in the county, state, and national locations (so, 15 days total). I then repeated this process with the 120

politicians. Using this dataset, I analyze the impact of location-based personalization on Google Search results.

5.3.1 Noise

To start, I examine whether there is noise in my search results. To calculate noise, I compare the search results received by treatments and their controls, i.e., two browsers that are running the same queries at the same time from the same locations.

Unlike prior work [68], I find that Google Search results are noisy. Figure 5.2 shows the average Jaccard Index and edit distance for all treatment/control pairs broken down by granularity and query types (values are averaged over all queries of the given type over 5 days). I make three observations. First, I see that *local* queries are much noisier than *controversial* and *politician* queries, in terms of result composition (shown by Jaccard) and reordering (shown by edit distance). Second, not only do *local* queries have more differences on average, but I also see that they have more variance (indicated by the standard deviation error bars). Third, I observe that noise is independent of location, i.e., the level of noise is uniform across all three granularities.

Search Terms. Given the high standard deviations for *local* queries, I pose the question: *do certain search terms exhibit more noise than others?* To answer this, I calculate the Jaccard Index and edit distance for each search term separately. Figure 5.3 shows the *local* queries along the x -axis, with the average edit distance for each query along the y -axis. The three lines correspond to search results gathered at different granularities; for clarity, I sort the x -axis from smallest to largest based on the *national* locations.

Figure 5.3 reveals a divide between the queries: brand names like “Starbucks” tend to be less noisy than generic terms like “school”. I observe similar trends for Jaccard Index. I examine this observation further next, when I look at the impact of different types of search results.

Search Result Types. To isolate the source of noise, I analyze the types of search results returned by Google Search. As described in Section 5.2.2, Google Search returns “typical” results, as well as Maps and News results. I suspect that Maps and News results may be more heavily impacted by location-based personalization, so I calculate the amount of noise that can be attributed to search results of these types separately. Intuitively, I simply calculate Jaccard and edit distance between pages after filtering out all search results that are not of type t .

Figure 5.4 shows the amount of noise contributed by Maps and News results for each query, along with the overall noise. Figure 5.4 focuses on the edit distance for *local* queries at *county*

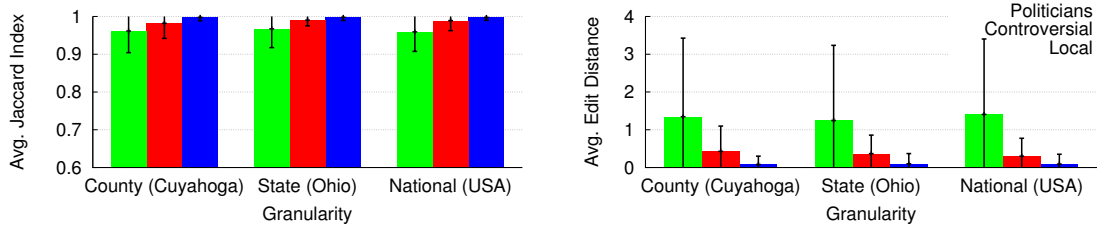


Figure 5.2: Average noise levels across different query types and granularities. Error bars show standard deviations.

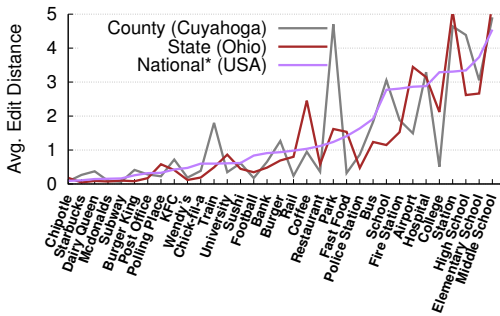


Figure 5.3: Noise levels for local queries across three granularities.

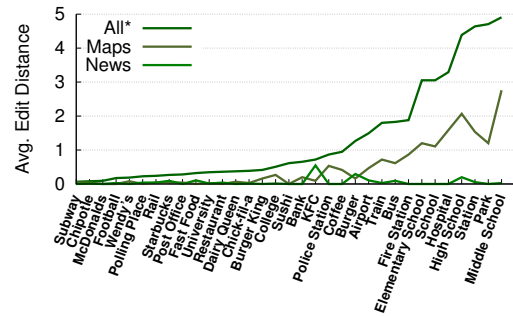


Figure 5.4: Amount of noise caused by different types of search results for local queries.

granularity, but I see similar trends at other granularities, and for Jaccard values. I observe that Maps results are responsible for around 25% of noise (calculated as the total number of search result changes due to Maps, divided by the overall number of changes), while News results cause almost zero noise. After some manual investigation I found that most differences due to Maps arise from one page having Maps results and the other having none. However, I also found cases where both queries yield Maps that highlight a different set of locations. Surprisingly, searches for specific brands typically do not yield Maps results, hence the low noise levels for those search terms.

Although I do not show the findings here due to space constraints, I observe the reverse effect for *controversial* queries: 6-17% of noise in such queries is due to News, while close to 0 is due to Maps. However, as Figure 5.2 shows, the level of noise in *controversial* queries is low overall.

5.3.2 Personalization

Now that I have quantified the noise in my dataset, I focus on answering the following two questions. First, *do certain types of queries trigger more personalization than others?* Second, *how does personalization change as the distance between two locations grows?*

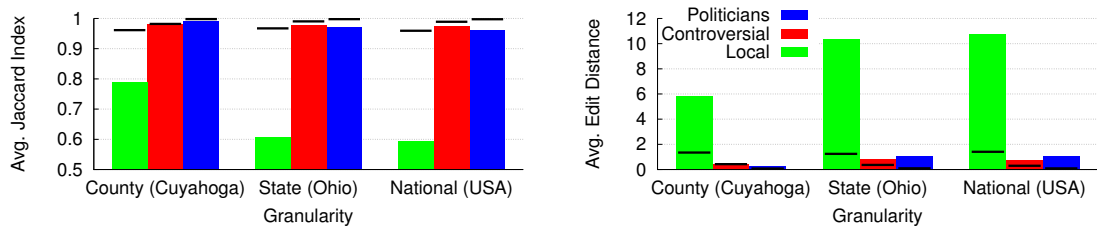


Figure 5.5: Average personalization across different query types and granularities. Black bars shows average noise levels from Figure 5.2.

Figure 5.5 shows the average Jaccard Index and edit distance values for each query category at each granularity. Values are averaged across all queries of the given types across 5 days. Recall that in the previous section, I were comparing treatments to their controls in order to measure noise; in this section, I are comparing all pairs of treatments to see if search results vary by location. For the sake of comparison, the average noise levels seen in Figure 5.2 are shown as horizontal black lines in Figure 5.5.

The first takeaway from Figure 5.5 is that *local* queries are much more personalized than *controversial* and *politicians* queries. The Jaccard index shows that 18-34% of the search results vary based on location for *local* queries, while the edit distance shows that 6-10 URLs are presented in a different order (after subtracting the effect of noise). *Controversial* and *politician* queries also exhibit small differences in Figure 5.5, but the Jaccard and edit distance values are very close to the noise-levels, making it difficult to claim that these changes are due to personalization.

The second takeaway from Figure 5.5 is that personalization increases with distance. The change is especially high between the *county*- and *state*-levels, with 2 additional search results changed and 4 reordered. As expected, this indicates that differences due to location-based personalization grow with geographic distance.

Search Terms. my next step is to examine how personalization varies across search terms. As before, I focus on *local* queries since they are most impacted by personalization. Figure 5.6 shows the edit distances for each *local* search term at each granularity (with the *x*-axis sorted by the *national*-level values). The significant increase in personalization between *county*- and *state*-level search results is again apparent in this figure.

Overall, I see that location-based personalization varies dramatically by query. The number of search results that change is between 5 and 17, where 17 is essentially all search results on the page. I also notice that (similar to my observations about noise) general terms such as “school” or “post office” exhibit higher personalization than brand names.

CHAPTER 5. THE IMPACT OF GEOLOCATION ON WEB SEARCH PERSONALIZATION

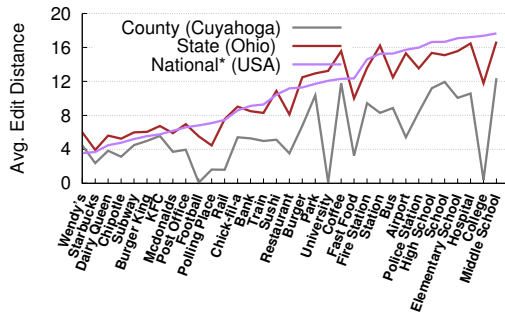


Figure 5.6: Personalization of each search term for *local* queries.

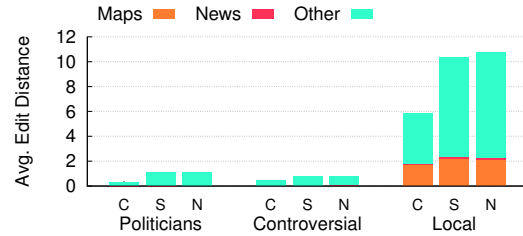


Figure 5.7: Amount of personalization caused by different types of search results.

The analogous plots for *politicians* and *controversial* queries show similar trends as Figure 5.6, but with much lower overall personalization. However, there are a few exceptional search terms. In the case of *politicians*, these exceptions are common names such as “Bill Johnson” or “Tim Ryan”, so it is likely that the differences stem from ambiguity. In the case of *controversial* terms, the most personalized queries are “health”, “republican party”, and “politics”.

Search Result Types. It is not terribly surprising that Google personalizes Maps and News results based on location. However, I find that personalization of Maps and News results only explains a small portion of the differences I observe.

Figure 5.7 breaks down the overall edit distance values into components corresponding to News, Maps, and all other search results, for each granularity and query type. For *controversial* queries, 6-18% of the edit distance can be attributed to News results, and interestingly, this fraction increases from *county* to *nation* granularity. A different composition is seen for *local* queries: 18-27% of differences are caused by Maps results. The takeaway is that, surprisingly, the vast majority of changes due to location-based personalization impact “typical” results.

Consistency Over Time. Thus far, all of my plots have presented values averaged over 5 days. To determine whether personalization is consistent over time, I plot Figure 5.8. In this figure, I choose one location in each granularity to serve as the *baseline*. The red line plots the average edit distance when comparing the baseline to its control (i.e., the red line shows the noise floor); each black line is a comparison between the baseline and another location at that granularity. I focus on *local* queries since they are most heavily personalized.

Figure 5.8 shows that the amount of personalization is stable over time. *Politicians* and *controversial* terms show the same trend but with lower personalization overall (findings not shown).

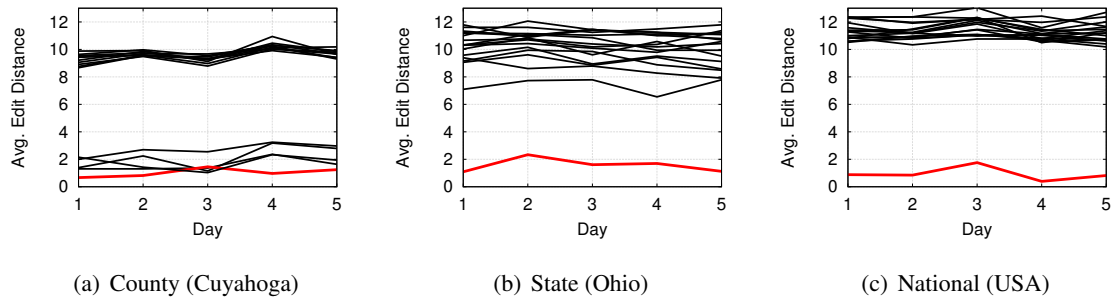


Figure 5.8: Personalization of 25 locations, each compared to a baseline location, for *local* queries. The red line compares two treatments at the baseline location (i.e., the experimental control), and thus shows the noise floor.

As expected, I see a wide gulf between the baseline and other locations at *state* and *nation* granularity, since search results are extremely different at these long distances. However, interestingly, I see that some locations “cluster” at the *county*-level, indicating that some locations receive similar search results to the baseline.

Physical Distance. My county level locations are often less than a mile apart, so my next question is whether the edit distance values correlate with physical distance. Figure 5.9 shows the edit distances and physical distances between each pair of locations within the county level. Here I consider all three search topics and see the clustering effect for both *local* terms and *politicians*. While Figure 5.9 shows this for a specific county as reference, the same trends are observed when taking any location as a reference. While there is no clear monotonic effect between the two variables, there is a clear dependence on physical distance. Namely, there is a physical distance threshold below which all edit distances belong to the low edit distance cluster, and a threshold above which all belong to the high edit distance cluster.

Demographics Since location is correlated with important demographic traits like income levels, education, race, political views, etc., Google’s personalization algorithm might result in discriminatory effects. To measure correlation between various demographic features and search results, I use the US census data and gather information about all US counties. I look at the relationship between my search results and 25 socio-demographic features that address poverty, education level, percentage of various races, level of English fluency, and population size. For each feature, I correlate the edit distance and the jaccard index between each pair of two points and the difference of the specified feature for those points. Unfortunately for the features I tested at these particular locations

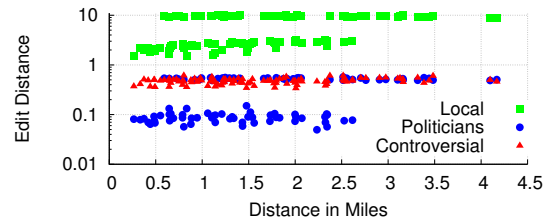


Figure 5.9: Correlation of physical distance and edit distance

I did not find any correlation.

5.4 Concluding Discussion

In this section I present a detailed analysis of location-based personalization on Google Search. I develop a novel methodology that allows me to query Google from any location around the world. Using this technique I sent 3,600 distinct queries to Google Search over a span of 30 days from 59 locations across the US.

My findings show that location does indeed have a large impact on search results, and that the differences increase as physical distance grows. However, I observe many nuances to Google’s implementation of location-based personalization. First, not all types of queries trigger the algorithm to the same degree: *politicians* are essentially unaffected by geography; *controversial* terms see small changes due to News; and *local* terms see large differences due to changes in Maps and normal results. Second, not all queries expected to trigger location-personalization do: for example, search results for brand names like “Starbucks” do not include Maps.

Finally, and most surprisingly, I also discover that Google Search returns search results that are very noisy, especially for *local* queries. This non-determinism is puzzling, since Google knows the precise location of the user (during our experiments), and thus should be able to quickly calculate the closest set of relevant locations.

My methodology can easily be extended to other countries and search engines and this provides a useful tool for uncovering location based personalization. This can be especially useful in case of applications or websites commonly used on mobile devices since they heavily use exact GPS coordinates.

Chapter 6

Measuring Personalization of Ecommerce Sites

6.1 Introduction

Shortly after worries about the Filter Bubble effect started appearing in the media, researchers and Internet users have uncovered evidence of personalization on e-commerce sites [106, 107, 165] as well. E-commerce sites have an economic incentive to convince users to spend more money thus their personalization strategies are likely not in the users' interest. Multiple retail companies were called out on their price discrimination practices, such as Staples, Orbitz or Amazon [103, 165, 167]. Since these examples were detected by chance, it is hard to know how common these practices really are among online retailers.

Luckily, e-commerce sites are similar in structure to search engines. Users have accounts which they log into before interacting with the sites. This allows sites to keep track of their users' behavior. Moreover searching and finding the right product is a similar process to searching for information using a search engine as well. The site provides a search box for the users to type their keywords into. I will adapt and reuse the methodology developed in chapter 3 to measuring e-commerce sites now. Just as in the case of search engines the biggest challenge in the measurement process is accurately attributing the observed differences to personalization. Results may differ due to changes in product inventory, regional tax differences, or inconsistencies across data centers and I need to make sure to separate these from differences stemming from actual personalization.

I investigate personalization along two questions: first, *how widespread is personalization*

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

on today's e-commerce web sites? This includes *price discrimination* (customizing prices for some users) as well as *price steering* (changing the order of search results to highlight specific products). Second, *how are e-commerce retailers choosing to implement personalization?* Although there is anecdotal evidence of these effects [167, 171] and specific instances where retailers have been exposed doing so [103, 165], the frequency and mechanisms of e-commerce personalization remain poorly understood.

To date, this is the first comprehensive study of e-commerce personalization that examines price discrimination and price steering for 300 real-world users¹, as well as synthetically generated fake accounts. I develop a measurement infrastructure that is able to distinguish genuine personalization of e-commerce sites from other sources of noise; this methodology is based on previous work on measuring personalization of web search services [68]. Using this methodology, I examine 16 top e-commerce sites covering general retailers as well as hotel and rental car booking sites. My real-world data indicates that eight of these sites implement personalization, while my controlled tests based on fake accounts allow me to identify specific user features that trigger personalization on seven sites. Specifically, I observe the following personalization strategies:

- Cheaptickets and Orbitz implement price discrimination by offering reduced prices on hotels to “members”.
- Expedia and Hotels.com engage in A/B testing that steers a subset of users towards more expensive hotels.
- Home Depot and Travelocity personalize search results for users on mobile devices.
- Priceline personalizes search results based on a user's history of clicks and purchases.

In addition to positively identifying price discrimination and steering on several well-known e-commerce sites, I also make the following four specific contributions. *First*, I introduce control accounts into all of my experiments, which allows me to differentiate between inherent noise and actual personalization. *Second*, I develop a novel methodology using information retrieval metrics to identify price steering. *Third*, I examine the impact of purchase history on personalization by reserving hotel rooms and rental cars, then comparing the search results received by these users to users with no history. *Fourth*, I identify a never-before-seen form of e-commerce personalization based on A/B testing, and show that it leads to price steering.

¹My study is conducted under Northeastern Institutional Review Board protocol #13-04-12.

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

The contents of this study appeared in the proceedings of the Internet Measurement Conference 2014 under the title “Measuring Price Discrimination and Steering on E-Commerce Web Sites”.

6.2 Methodology

In this section I first introduce the concepts I will use to describe e-commerce sites and measuring personalization on them. Then I will describe the specifics of the data collection such as the retailers and products I will focus on which are again used as part of the methodology described in Section 3.

6.2.1 Definitions

More so than other web services [68], e-commerce retailers have a number of different dimensions available to personalize on. In this study, I focus on two of the primary vectors for e-commerce personalization:

Price steering occurs when two users receive different product results (or the same products in a different order) for the same query (e.g., Best Buy showing more-expensive products to user *A* than user *B* when they both query for “laptops”). Price steering can be similar to personalization in web search [68], i.e., the e-commerce provider may be trying to give the user more relevant products (or, they may be trying to extract more money from the user). Steering is possible because e-commerce sites often do not sort search results by an objective metric like price or user reviews by default; instead, results can be sorted using an ambiguous metric like “Best Match” or “Most Relevant”.

Price discrimination occurs when two users are shown inconsistent prices for the same product (e.g., Travelocity showing a select user a higher price for a particular hotel). Contrary to popular belief, price discrimination in general is not illegal in the United States [42], as the Robinson–Patman Act of 1936 (*a.k.a.* the Anti-Price Discrimination Act) is written to control the behavior of product manufacturers and distributors, not consumer-facing enterprises. It is unclear whether price discrimination targeted against protected classes (e.g., race, religion, gender) is legal.

Retailer	Site	Category
Best Buy	http://bestbuy.com	Electronics
CDW	http://cdw.com	Computers
HomeDepot	http://homedepot.com	Home-improvement
JCPenney	http://jcp.com	Clothes, housewares
Macy's	http://macys.com	Clothes, housewares
Newegg	http://newegg.com	Computers
Office Depot	http://officedepot.com	Office supplies
Sears	http://sears.com	Clothes, housewares
Staples	http://staples.com	Office supplies
Walmart	http://walmart.com	General retailer

Table 6.1: The general retailers I measured in this study.

6.2.2 E-commerce Sites

Throughout the study, I survey a wide variety of e-commerce web sites, ranging from large-scale retailers like Walmart to travel sites like Expedia. To make the results comparable, I only consider products returned via *searches*—as opposed to “departments”, home page offers, and other mechanisms by which e-commerce sites offer products to users—as searching is a functionality supported by most large retailers. Additionally, I use *products* and their advertised price on the search result page (e.g., a specific item on Walmart or hotel on Expedia) as the basic unit of measurement.

I focus on two classes of e-commerce web sites: general e-commerce retailers (e.g., Best Buy) and travel retailers (e.g., Expedia). I choose to include travel retailers because there is anecdotal evidence of price steering among such sites [103]. Of course, my methodology can be applied to other categories of e-commerce sites as well.

General Retailers. I select 10 of the largest e-commerce retailers, according to the Top500 e-commerce database [163], for my study, shown in Table 6.1. I exclude Amazon, as Amazon hosts a large number of different merchants, making it difficult to measure Amazon itself. I also exclude sites like `apple.com` that only sell their own brand.

Travel Retailers. I select six of the most popular web-based travel retailers [164] to study, shown in Table 6.2. For these retailers, I focus on searches for *hotels* and *rental cars*. I do not include airline tickets, as airline ticket pricing is done transparently through a set of Global Distribution Systems (GDSes) [13]. Furthermore, a recent study by Vissers et al. looked for, but was unable to find, evidence of price discrimination on the websites of 25 major airlines [170].

Retailer	Site	Hotels	Cars
Cheaptickets	http://cheaptickets.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Expedia	http://expedia.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Hotels.com	http://hotels.com	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Orbitz	http://orbitz.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Priceline	http://priceline.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Travelocity	http://travelocity.com	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Table 6.2: The travel retailers I measured in this study.

Since each retailer uses substantially different HTML markup to implement their site I collect data by visiting the various sites' web pages, and I write custom HTML parsers to extract the products and prices from the search result page for each site that I study. In all cases, the prices of products returned in search results are in US dollars, are pre-tax, and do not include shipping fees. Examining prices before taxes and shipping are applied helps to avoid differences in pricing that are due to the location of the business and/or the customer.

6.2.3 Searches

I select 20 searches to send to each target e-commerce site; it is the results of these searches that I use to look for personalization. I select the searches to cover a variety of product types, and tailor the searches to the type of products each retailer sells. For example, for JCPenney, my searches include “pillows”, “sunglasses”, and “chairs”; for Newegg, my searches include “flash drives”, “LCD TVs”, and “phones”.

For travel web sites, I select 20 searches (location and date range) that I send to each site when searching for hotels or rental cars. I select 10 different cities across the globe (Miami, Honolulu, Las Vegas, London, Paris, Florence, Bangkok, Cairo, Cancun, and Montreal), and choose date ranges that are both short (4-day stays/rentals) and long (11-day stays/rentals).

6.3 Real-World Personalization

I begin by addressing my first question: *how widespread are price discrimination and steering on today's e-commerce web sites?* To do so, I have a large set of real-world users run my experimental searches and examine the results that they receive.

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

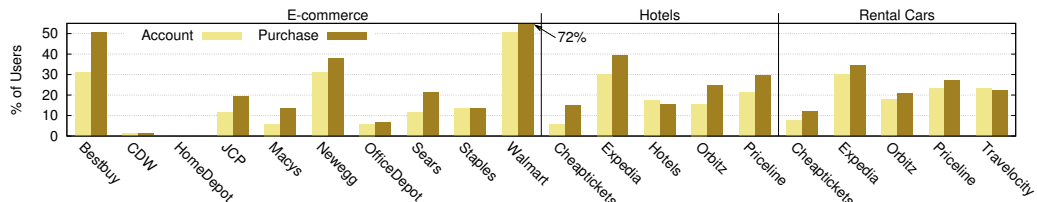


Figure 6.1: Previous usage (i.e., having an account and making a purchase) of different e-commerce sites by myAMT users.

6.3.1 Data Collection

To obtain a diverse set of users, I recruited from Amazon’s Mechanical Turk (AMT) [5]. I posted three Human Intelligence Tasks (HITs) to AMT, with each HIT focusing on e-commerce, hotels, or rental cars. In the HIT, I explained my study and offered each user \$1.00 to participate.² Users were required to live in the United States, and could only complete the HIT once.

Users who accepted the HIT were instructed to configure their web browser to use a Proxy Auto-Config (PAC) file provided by us. The PAC file routes all traffic to the sites under study to an HTTP proxy controlled by us. Then, users were directed to visit a web page containing JavaScript that performed my set of searches in an `iframe`. After each search, the Javascript grabs the HTML in the `iframe` and uploads it to my server, allowing us to view the results of the search. By having the user run the searches within their browser, any cookies that the user’s browser had previously been assigned would automatically be forwarded in my searches. This allows us to examine the results that the user would have received. I waited 15 seconds between each search, and the overall experiment took ≈ 45 minutes to complete (between five and 10 sites, each with 20 searches).

The HTTP proxy serves two important functions. *First*, it allows us to quantify the baseline amount of noise in search results. Whenever the proxy observes a search request, it fires off *two* identical searches using PhantomJS (with no cookies) and saves the resulting pages. The results from PhantomJS serve as a *comparison* and a *control* result. As outlined in Section 3.1, I compare the results served to the comparison and control to determine the underlying level of noise in the search results. I also compare the results served to the comparison and the real user; any differences between the real user and the comparison above the level observed between the comparison and the control can be attributed to personalization.

Second, the proxy reduces the amount of noise by sending the experimental, comparison, and control searches to the web site at the same time and from the same IP address. As stated in § ??,

²This study was conducted under Northeastern University IRB protocol #13-04-12; all personally identifiable information was removed from my collected data.

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

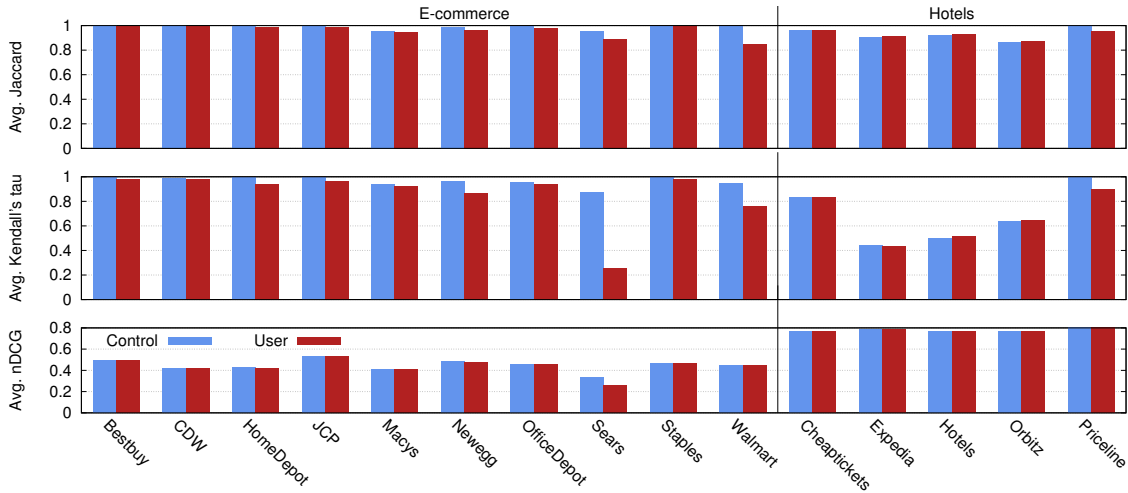


Figure 6.2: Average Jaccard index (top), Kendall’s τ (middle), and nDCG (bottom) across all users and searches for each web site.

sending all queries from the same IP address controls for personalization due to geolocation, which I am specifically not studying in this paper. Furthermore, I hard-coded a DNS mapping for each of the sites on the proxy to avoid discrepancies that might come from round-robin DNS sending requests to different data centers.

In total, I recruited 100 AMT users in each of my retail, hotel, and car experiments. In each of the experiments, the participants first answered a brief survey about whether they had an account and/or had purchased something from each site. I present the results of this survey in Figure 6.1. I observe that many of my users have accounts and a purchase history on a large number of the sites I study.³

6.3.2 Price Steering

I begin by looking for *price steering*, or personalizing search results to place more- or less-expensive products at the top of the list. I do not examine rental car results for price steering because travel sites tend to present these results in a deterministically ordered grid of car types (e.g., economy, SUV) and car companies (with the least expensive car in the upper left). This arrangement prevents travel sites from personalizing the order of rental cars.

³Note that the fraction of users having made purchases can be higher than the fraction with an account, as many sites allow purchases as a “guest”.

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

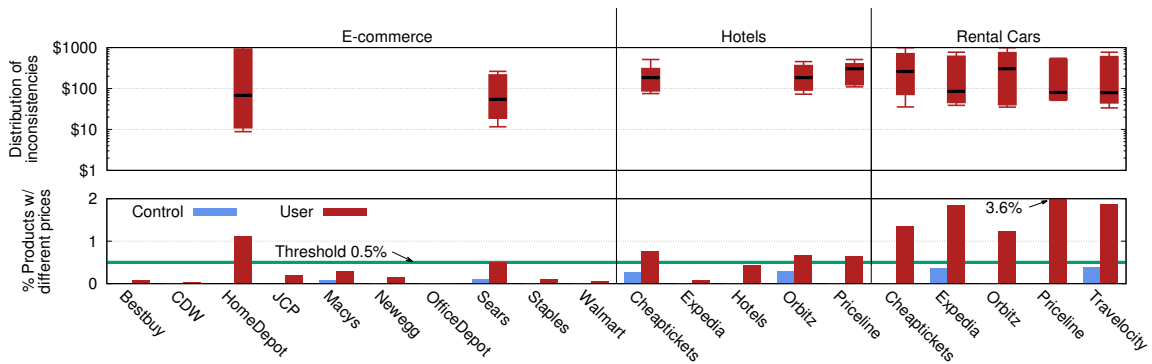


Figure 6.3: Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with $\geq 0.5\%$ of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).

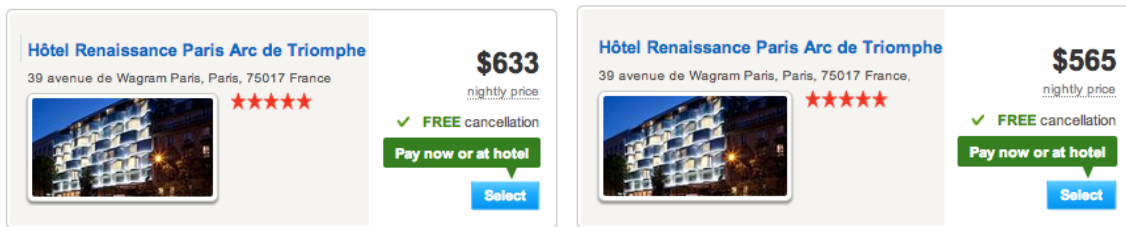


Figure 6.4: Example of price discrimination. The top result was served to the AMT user, while the bottom result was served to the comparison and control.

To measure price steering, I use three metrics which I introduced in Section ??: Jacquard Index, Kendall’s τ and nDGG.

For each site, Figure 6.2 presents the average Jaccard index, Kendall’s τ , and nDCG across all queries. The results are presented comparing the comparison to the control searches (Control), and the comparison to the AMT user searches (User). I observe several interesting trends. *First*, Sears, Walmart, and Priceline all have a lower Jaccard index for AMT users relative to the control. This indicates that the AMT users are receiving different products at a higher rate than the control searches (again, note that I are *not* comparing AMT users’ results to each other; I only compare each user’s result to the corresponding comparison result). Other sites like Orbitz show a Jaccard of 0.85 for Control and User, meaning that the set of results shows inconsistencies, but that AMT users are not seeing a higher level of inconsistency than the control and comparison searches.

Second, I observe that on Newegg, Sears, Walmart, and Priceline, Kendall’s τ is at least

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

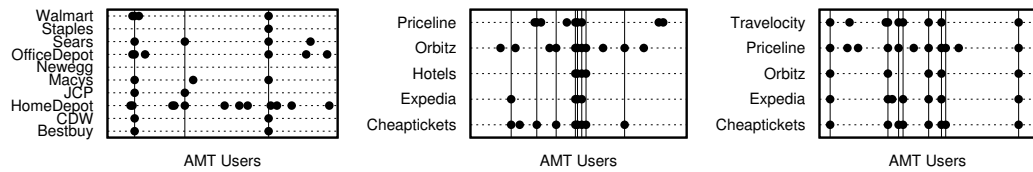


Figure 6.5: AMT users that receive highly personalized search results on general retail, hotels, and car rental sites.

0.1 lower for AMT users, i.e., AMT users are consistently receiving results in a different order than the controls. This observation is especially true for Sears, where the ordering of results for AMT users is markedly different. *Third*, I observe that Sears alone appears to be ordering products for AMT users in a price-biased manner. The nDCG results show that AMT users tend to have *cheaper* products near the top of their search results relative to the controls. Note that the results in Figure 6.2 are only useful for uncovering price steering; I examine whether the target sites are performing price discrimination in Section 6.3.3.

Besides Priceline, the other four travel sites do not show significant differences between the AMT users and the controls. However, these four sites do exhibit significant noise: Kendall's τ is ≤ 0.83 in all four cases. On Cheaptickets and Orbitz, I manually confirm that this noise is due to randomness in the order of search results. In contrast, on Expedia and Hotels.com this noise is due to systematic A/B testing on users (see Section 6.4.2 for details), which explains why I see equally low Kendall's τ values on both sites for all users. Unfortunately, it also means that I cannot draw any conclusions about personalization on Expedia and Hotels.com from the AMT experiment, since the search results for the comparison and the control rarely match.

6.3.3 Price Discrimination

So far, I have only looked at the set of products returned. I now turn to investigate whether sites are altering the prices of products for different users, i.e., price discrimination. In the bottom plot of Figure 6.3, I present the fraction of products that show price inconsistencies between the user's and comparison searches (User) and between the comparison and control searches (Control). Overall, I observe that most sites show few inconsistencies (typically $< 0.5\%$ of products), but a small set of sites (Home Depot, Sears, and many of the travel sites) show both a significant fraction of price inconsistencies *and* a significantly higher fraction of inconsistencies for the AMT users.

To investigate this phenomenon further, in the top of Figure 6.3, I plot the distribution of

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

price differentials for all sites where $>0.5\%$ of the products show inconsistency. I plot the mean price differential (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker). Note that in my data, AMT users always receive higher prices than the controls (on average), thus all differentials are positive. I observe that the price differentials on many sites are quite large (up to hundreds of dollars). As an example, in Figure 6.4, I show a screenshot of a price inconsistency that I observed. Both the control and comparison searches returned a price of \$565 for a hotel, while myAMT user was returned a price of \$633.

6.3.4 Per-User Personalization

Next, I take a closer look at the subset of AMT users who experience high levels of personalization on one or more of the e-commerce sites. my goal is to investigate whether these AMT users share any observable features that may illuminate why they are receiving personalized search results. I define highly personalized users as the set of users who see products with inconsistent pricing $>0.5\%$ of the time. After filtering I am left with between 2-12% of myAMT users depending on the site.

First, I map the AMT users' IP addresses to their geolocations and compare the locations of personalized and non-personalized users. I find no discernible correlation between location and personalization. However, as mentioned above, in this experiment all searches originate from a proxy in Boston. Thus, it is not surprising that I do not observe any effects due to location, since the sites did not observe users' true IP addresses.

Next, I examine the AMT users' browser and OS choices. I am able to infer their platform based on the HTTP headers sent by their browser through my proxy. Again, I find no correlation between browser/OS choice and high personalization. In Section 6.4, I do uncover personalization linked to the use of mobile browsers, however none of the AMT users in my study did the HIT from a mobile device.

Finally, I ask the question: are there AMT users who receive personalized results on multiple e-commerce sites? Figure 6.5 lists the 100 users in my experiments along the x -axis of each plot; a dot highlights cases where a site personalized search results for a particular user. Although some dots are randomly dispersed, there are many AMT users that receive personalized results from several e-commerce sites. I highlight users who see personalized results on more than one site with vertical bars. More users fall into this category on travel sites than on general retailers.

The takeaway from Figure 6.5 is that I observe many AMT users who receive personalized

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

results across multiple sites. This suggests that these users share feature(s) that all of these sites use for personalization. Unfortunately, I am unable to infer the specific characteristics of these users that are triggering personalization.

Cookies. Although I logged the cookies sent by AMT users to the target e-commerce sites, it is not possible to use them to determine why some users receive personalized search results. First, cookies are typically random alphanumeric strings; they do not encode useful information about a user's history of interactions with a website (e.g., items clicked on, purchases, *etc.*). Second, cookies can be set by content embedded in third-party websites. This means that a user with a cookie from e-commerce site S may never have consciously visited S , let alone made purchases from S . These reasons motivate why I rely on survey results (see Figure 6.1) to determine AMT users' history of interactions with the target e-commerce sites.

6.3.5 Summary

To summarize my findings in this section: I find evidence for price steering and price discrimination on four general retailers and five travel sites. Overall, travel sites show price inconsistencies in a higher percentage of cases, relative to the controls, with prices increasing for AMT users by hundreds of dollars. Finally, I observe that many AMT users experience personalization across multiple sites.

6.4 Personalization Features

In Section 6.3, I demonstrated that e-commerce sites personalize results for real users. However, I cannot determine *why* results are being personalized based on the data from real-world users, since there are too many confounding variables attached to each AMT user (e.g., their location, choice of browser, purchase history, *etc.*).

In this section, I conduct controlled experiments with fake accounts created by us to examine the impact of specific features on e-commerce personalization. Although we cannot test all possible features, I examine five likely candidates: browser, OS, account log-in, click history, and purchase history. I chose these features because e-commerce sites have been observed personalizing results based on these features in the past [103, 165].

I begin with an overview of the design of my synthetic user experiments. Next, I highlight examples of personalization on hotel sites and general retailers. None of my experiments triggered

Category	Feature	Tested Values
Account	Cookies	No Account, Logged In, No Cookies
User-Agent	OS	Win. XP, Win. 7, OS X, Linux
	Browser	Chrome 33, Android Chrome 34, IE 8, Firefox 25, Safari 7, iOS Safari 6
Account History	Click	Low Prices, High Prices
	Purchase	Low Prices, High Prices

Table 6.3: User features evaluated for effects on personalization.

personalization on rental car sites, so I omit these results.

6.4.1 Experimental Overview

The goal of my synthetic experiments is to determine whether specific user features trigger personalization on e-commerce sites. To assess the impact of feature X that can take on values x_1, x_2, \dots, x_n , I execute $n + 1$ PhantomJS instances, with each value of X assigned to one instance. The $n + 1$ th instance serves as the control by duplicating the value of another instance. All PhantomJS instances execute 20 queries (see § 6.2.3) on each e-commerce site per day, with queries spaced one minute apart to avoid tripping security countermeasures. PhantomJS downloads the first page of results for each query. Unless otherwise specified, PhantomJS persists all cookies between experiments. All of my experiments are designed to complete in <24 hours.

To mitigate measurements errors due to noise (see § 3.1), we perform three steps (some borrowed from previous work [64, 68]): *first*, all searches for a given query are executed at the same time. This eliminates differences in results due to temporal effects. This also means that each of my treatments has exactly the same search history at the same time. *Second*, I use static DNS entries to direct all of my query traffic to specific IP addresses of the retailers. This eliminates errors arising from differences between datacenters. *Third*, although all PhantomJS instances execute on one machine, I use SSH tunnels to forward the traffic of each treatment to a unique IP address in a /24 subnet. This process ensures that any effects due to IP geolocation will affect all results equally.

Static Features. Table 6.3 lists the five features that I evaluate in my experiments. In the cookie experiment, the goal is to determine whether e-commerce sites personalize results for users who are logged-in to the site. Thus, two PhantomJS instances query the given e-commerce site without logging-in, one logs-in before querying, and the final account clears its cookies after every HTTP request.

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

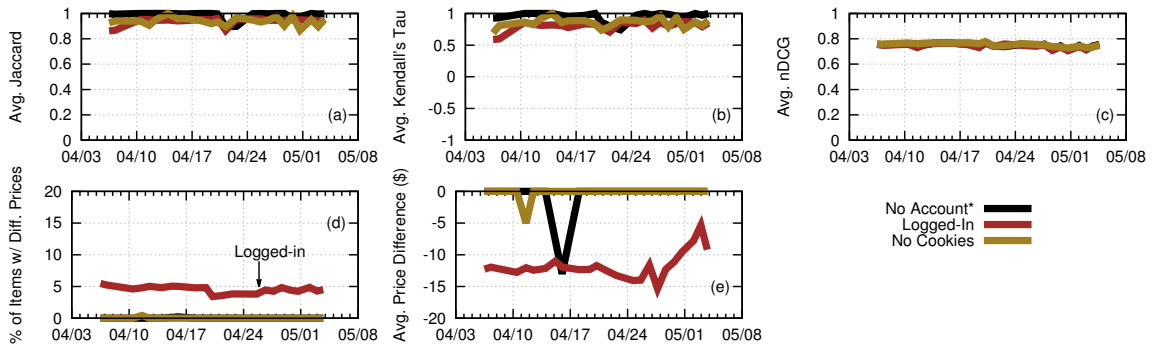


Figure 6.6: Examining the impact of user accounts and cookies on hotel searches on Cheaptickets.

In two sets of experiments, I vary the `User-Agent` sent by PhantomJS to simulate different OSes and browsers. The goal of these tests is to see if e-commerce sites personalize based on the user's choice of OS and browser. In the OS experiment, all instances report using Chrome 33, and Windows 7 serves as the control. In the browser experiment, Chrome 33 serves as the control, and all instances report using Windows 7, except for Safari 7 (which reports OS X Mavericks), Safari on iOS 6, and Chrome on Android 4.4.2.

Historical Features. In my historical experiments, the goal is to examine whether e-commerce sites personalize results based on users' history of viewed and purchased items. Unfortunately, I am unable to create purchase history on general retail sites because this would entail buying and then returning physical goods. However, it is possible for us to create purchase history on travel sites. On Expedia, Hotels.com, Priceline, and Travelocity, some hotel rooms feature "pay at the hotel" reservations where you pay at check-in. A valid credit card must still be associated with "pay at the hotel" reservations. Similarly, all five travel sites allow rental cars to be reserved without up-front payment. These no-payment reservations allow us to book reservations on travel sites and build up purchase history.

To conduct my historical experiments, I created six accounts on the four hotel sites and all five rental car sites. Two accounts on each site serve as controls: they do not click on search results or make reservations. Every night for one week, I manually logged-in to the remaining four accounts on each site and performed specific actions. Two accounts searched for a hotel/car and clicked on the highest and lowest priced results, respectively. The remaining two accounts searched for the same hotel/car and booked the highest and lowest priced results, respectively. Separate credit cards were used for high- and low-priced reservations, and neither card had ever been used to book travel

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

before. Although it is possible to imagine other treatments for account history (e.g., a person who always travels to a specific country), price-constrained (elastic) and unconstrained (inelastic) users are a natural starting point for examining the impact of account history. Furthermore, although these treatments may not embody realistic user behavior, they do present unambiguous signals that could be observed and acted upon by personalization algorithms.

I pre-selected a destination and travel dates for each night, so the click and purchase accounts all used the same search parameters. Destinations varied across major US, European, and Asian cities, and dates ranged over the last six months of 2014. All trips were for one or two night stays/rentals. On average, the high- and low-price purchasers reserved rooms for \$329 and \$108 per night, respectively, while the high- and low-price clickers selected rooms for \$404 and \$99 per night. The four rental car accounts were able to click and reserve the exact same vehicles, with \$184 and \$43 being the average high- and low-prices per day.

Each night, after I finished manually creating account histories, I used PhantomJS to run my standard list of 20 queries from all six accounts on all nine travel sites. To maintain consistency, manual history creation and automated tests all used the same set of IP addresses and Firefox.

Ethics. I took several precautions to minimize any negative impact of my purchase history experiments on travel retailers, hotels, and rental car agencies. I reserved, at most, one room from any specific hotel. All reservations were made for hotel rooms and cars at least one month into the future, and all reservations were canceled at the conclusion of my experiments.

Analyzing Results. To analyze the data from my feature experiments, I leverage the same five metrics used in § 6.3. Figure 6.6 exemplifies the analysis I conduct for each user feature on each e-commerce site. In this example, I examine whether Cheaptickets personalizes results for users that are logged-in. The x -axis of each subplot is time in days. The plots in the top row use Jaccard Index, Kendall's τ , and nDCG to analyze steering, while the plots in the bottom row use percent of items with inconsistent prices and average price difference to analyze discrimination.

All of my analysis is always conducted relative to a control. In all of the figures in this section, the starred (*) feature in the key is the control. For example, in Figure 6.6, all analysis is done relative to a PhantomJS instance that does not have a user account. Each point is an average of the given metric across all 20 queries on that day.

In total, my analysis produced >360 plots for the various features across all 16 e-commerce sites. Overall, most of the experiments do not reveal evidence of steering or discrimination. Thus, for the remainder of this section, I focus on the particular features and sites where I do observe

The image shows two identical hotel listings for Eden Roc Miami Beach on Cheaptickets, illustrating price discrimination. The top listing is for non-logged-in users, showing a price of \$299/night. The bottom listing is for logged-in users, showing a 'Members Only' price of \$194/night. Both listings include a 4.1 rating, 192 reviews, and a 'See details' button. The top listing also features a 'Save 30% Off - 4 Nights Or More' banner and a '\$25 Elle Spa Credit' offer.

Figure 6.7: Price discrimination on Cheaptickets. The top result is shown to users that are not logged-in. The bottom result is a “Members Only” price shown to logged-in users.

personalization. None of my feature tests revealed personalization on rental car sites, so I omit them entirely.

6.4.2 Hotels

I begin by analyzing personalization on hotel sites. I observe hotel sites implementing a variety of personalization strategies, so I discuss each case separately.

Cheaptickets and Orbitz. The first sites that I examine are Cheaptickets and Orbitz. These sites are actually one company, and appear to be implemented using the same HTML structure and server-side logic. In my experiments, I observe both sites personalizing hotel results based on user accounts; for brevity I present the analysis of Cheaptickets and omit Orbitz.

Figures 6.6(a) and (b) reveal that Cheaptickets serves slightly different sets of results to users who are logged-in to an account, versus users who do not have an account or who do not store

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

cookies. Specifically, out of 25 results per page, ≈ 2 are new and ≈ 1 is moved to a different location on average for logged-in users. In some cases (e.g., hotels in Bangkok and Montreal) the differences are much larger: up to 11 new and 11 moved results. However, the nDCG analysis in Figure 6.6(c) indicates that these alterations do not have an appreciable impact on the price of highly-ranked search results. Thus, I do not observe Cheaptickets or Orbitz steering results based on user accounts.

However, Figure 6.6(d) shows that logged-in users receive different prices on $\approx 5\%$ of hotels. As shown in Figure 6.6(e), the hotels with inconsistent prices are \$12 cheaper on average. This demonstrates that Cheaptickets and Orbitz implement price discrimination, in favor of users who have accounts on these sites. Manual examination reveals that these sites offer “Members Only” price reductions on certain hotels to logged-in users. Figure 6.7 shows an example of this on Cheaptickets.

Although it is not surprising that some e-commerce sites give deals to members, my results on Cheaptickets (and Orbitz) are important for several reasons. First, although members-only prices may be an accepted practice, it still qualifies as price discrimination based on direct segmentation (with members being the target segment). Second, this result confirms the efficacy of my methodology, i.e., I am able to accurately identify price discrimination based on automated probes of e-commerce sites. Finally, my results reveal the actual differences in prices offered to members, which may not otherwise be public information.

Hotels.com and Expedia. my analysis reveals that Hotels.com and Expedia implement the same personalization strategy: randomized A/B tests on users. Since these sites are similar, I focus on Expedia and omit the details of Hotels.com.

Initially, when I analyzed the results of my feature tests for Expedia, I noticed that the search results received by the control and its twin never matched. More oddly, I also noticed that 1) the control results did match the results received by other specific treatments, and 2) these matches were consistent over time.

These anomalous results led us to suspect that Expedia was randomly assigning each of my treatments to a “bucket”. This is common practice on sites that use A/B testing: users are randomly assigned to buckets based on their cookie, and the characteristics of the site change depending on the bucket you are placed in. Crucially, the mapping from cookies to buckets is deterministic: a user with cookie C will be placed into bucket B whenever they visit the site unless their cookie changes.

To determine whether my treatments are being placed in buckets, I generate Table 6.4, which shows the Jaccard Index for 12 pairs of feature experiments on Expedia. Each table entry is averaged over 20 queries and 10 days. For a typical website, I would expect the control (*Ctrl*) in

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

		Account			Browser				OS			
		In	No*	Ctrl	FX	IE8	Chr*	Ctrl	OSX	Lin	XP	Win7*
OS	Ctrl	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	Win7*	1.0	1.0	1.0	0.9	1.0	1.0	0.9	0.9	0.9	0.9	
	XP	0.9	0.9	0.9	1.0	0.9	0.9	1.0	1.0	1.0		
	Lin	0.9	0.9	0.9	1.0	0.9	0.9	1.0	1.0			
	OSX	0.9	0.9	0.9	1.0	0.9	0.9	1.0				
Browser	Ctrl	0.9	0.9	0.9	1.0	0.9	0.9					
	Chr*	1.0	1.0	1.0	0.9	1.0						
	IE8	1.0	1.0	1.0	0.9							
	FX	0.9	0.9	0.9								
Acct	Ctrl	1.0	1.0									
	No*	1.0										

Table 6.4: Jaccard overlap between pairs of user feature experiments on Expedia.

each category to have perfect overlap (1.0) with its twin (marked with a *). However, in this case the perfect overlaps occur between random pairs of tests. For example, the results for Chrome and Firefox perfectly overlap, but Chrome has low overlap with the control, which was also Chrome. This strongly suggests that the tests with perfect overlap have been randomly assigned to the same bucket. In this case, I observe three buckets: 1) {Windows 7, account control, no account, logged-in, IE 8, Chrome}, 2) {XP, Linux, OS X, browser control, Firefox}, and 3) {OS control}.

To confirm my suspicion about Expedia, I examine the behavior of the experimental treatment that clears its cookies after every query. I propose the following hypothesis: if Expedia is assigning users to buckets based on cookies, then the clear cookie treatment should randomly change buckets after each query. my assumption is that this treatment will receive a new, random cookie each time it queries Expedia, and thus its corresponding bucket will change.

To test this hypothesis I plot Figure 6.9, which shows the Jaccard overlap between search results received by the clear cookie treatment, and results received by treatments in other buckets. The x -axis corresponds to the search results from the clear cookie treatment over time; for each page of results, I plot a point in the bucket (y -axis) that has >0.97 Jaccard overlap with the clear cookie treatment. If the clear cookie treatment’s results do not overlap with results from any of the buckets, the point is placed on the “Unknown” row. In no cases did the search results from the clear cookie treatment have >0.97 Jaccard with more than a single bucket, confirming that the set of results returned to each bucket are highly disjoint (see Table 6.4).

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

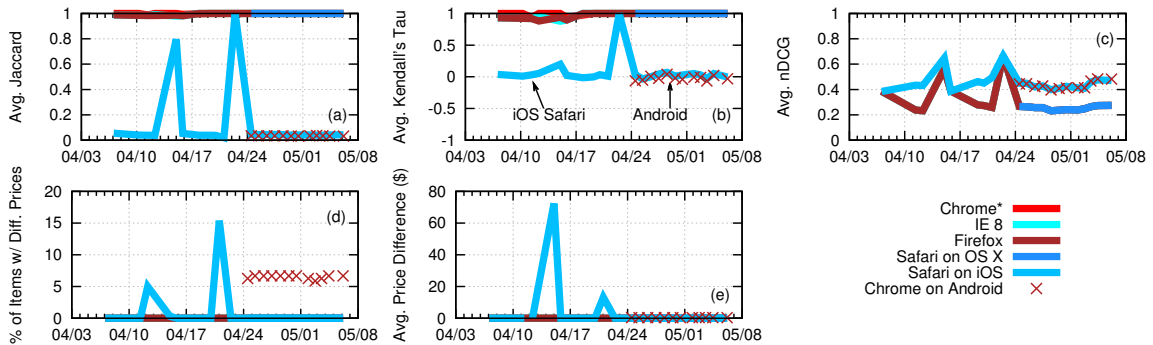


Figure 6.8: Home Depot alters product searches for users of mobile browsers.

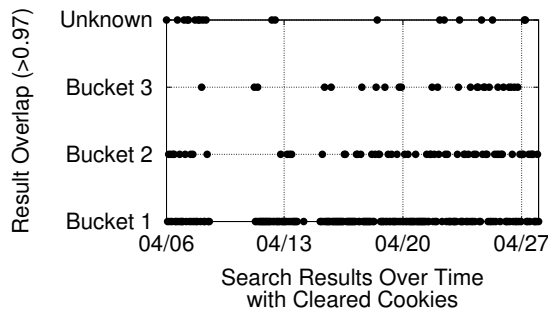


Figure 6.9: Clearing cookies causes a user to be placed in a random bucket on Expedia.

Figure 6.9 confirms that the clear cookie treatment is randomly assigned to a new bucket on each request. 62% of results over time align with bucket 1, while 21% and 9% match with buckets 2 and 3, respectively. Only 7% do not match any known bucket. These results suggest that Expedia does not assign users to buckets with equal probability. There also appear to be time ranges where some buckets are not assigned, e.g., bucket 3 in between 04/12 and 04/15. I found that Hotels.com also assigns users to one of three buckets, that the assignments are weighted, and that the weights change over time.

Now that I understand how Expedia (and Hotels.com) assign users to buckets, I can analyze whether users in different buckets receive personalized results. Figure 6.10 presents the results of this analysis. I choose an account from bucket 1 to use as a control, since bucket 1 is the most frequently assigned bucket.

Two conclusions can be drawn from Figure 6.10. First, I see that users are periodically shuffled into different buckets. Between 04/01 and 04/20, the control results are consistent, i.e.,

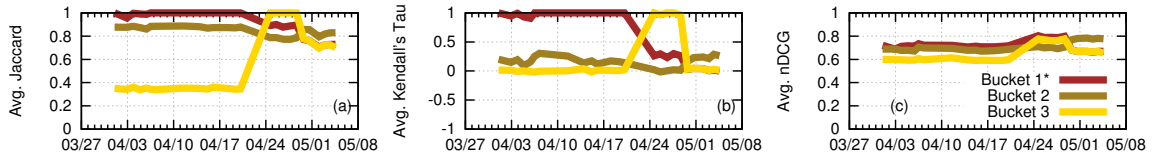


Figure 6.10: Users in certain buckets are steered towards higher priced hotels on Expedia.

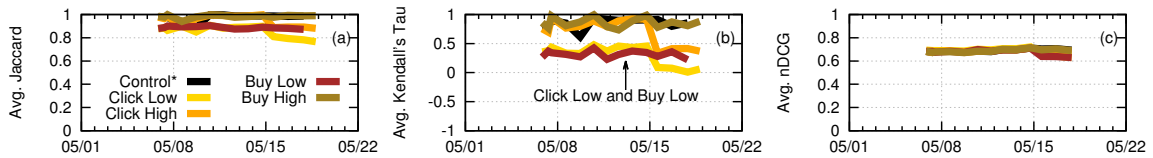


Figure 6.11: Priceline alters hotel search results based on a user's click and purchase history.

Jaccard and Kendall's τ for bucket 1 are ≈ 1 . However, on 04/21 the three lines change positions, implying that the accounts have been shuffled to different buckets. It is not clear from my data how often or why this shuffling occurs.

The second conclusion that can be drawn from Figure 6.10 is that Expedia is steering users in some buckets towards more expensive hotels. Figures 6.10(a) and (b) show that users in different buckets receive different results in different orders. For example, users in bucket 3 see $>60\%$ different search results compared to users in other buckets. Figure 6.10(c) highlights the net effect of these changes: results served to users in buckets 1 and 2 have higher nDCG values, meaning that the hotels at the top of the page have higher prices. I do not observe price discrimination on Expedia or Hotels.com.

Priceline. As depicted in Figure 6.11, Priceline alters hotel search results based on the user's history of clicks and purchases. Figures 6.11(a) and (b) show that users who clicked on or reserved low-price hotel rooms receive slightly different results in a much different order, compared to users who click on nothing, or click/reserve expensive hotel rooms. I manually examined these search results but could not locate any clear reasons for this reordering. The nDCG results in Figure 6.11(c) confirm that the reordering is not correlated with prices. Thus, although it is clear that account history impacts search results on Priceline, I cannot classify the changes as steering. Furthermore, I observe no evidence of price discrimination based on account history on Priceline.

Travelocity. As shown in Figure 6.12, Travelocity alters hotel search results for users who browse from iOS devices. Figures 6.12(a) and (b) show that users browsing with Safari on iOS receive

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

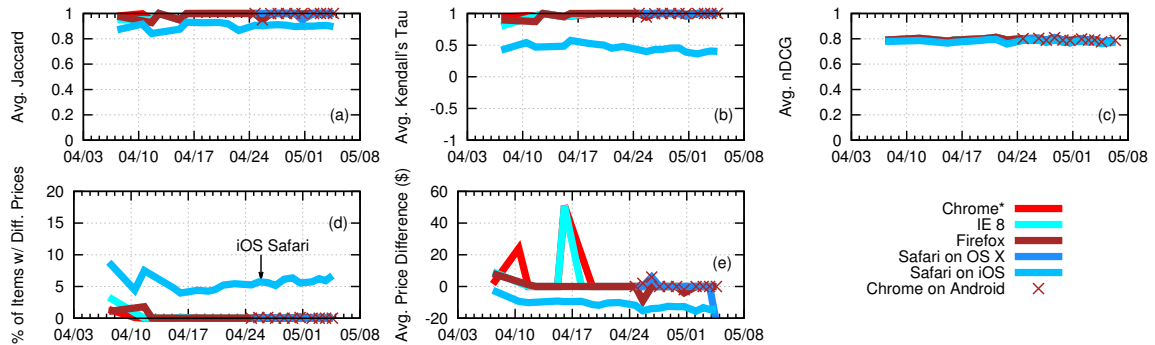


Figure 6.12: Travelocity alters hotel search results for users of Safari on iOS, but not Chrome on Android.

slightly different hotels, and in a much different order, than users browsing from Chrome on Android, Safari on OS X, or other desktop browsers. Note that I started my Android treatment at a later date than the other treatments, specifically to determine if the observed changes on Travelocity occurred on all mobile platforms or just iOS.

Although Figure 6.12(c) shows that this reordering does not result in price steering, Figures 6.12(d) and (e) indicate that Travelocity does modify prices for iOS users. Specifically, prices fall by $\approx \$15$ on $\approx 5\%$ of hotels (or 5 out of 50 per page) for iOS users. The noise in Figure 6.12(e) (e.g., prices increasing by \$50 for Chrome and IE 8 users) is not significant: this result is due to a single hotel that changed price.

The takeaway from Figure 6.12 is that I observe evidence consistent with price discrimination in favor of iOS users on Travelocity⁴. Unlike Cheaptickets and Orbitz, which clearly mark sale-price “Members Only” deals, there is no visual cue on Travelocity’s results that indicates prices have been changed for iOS users. Online travel retailers have publicly stated that mobile users are a high-growth customer segment, which may explain why Travelocity offers price-incentives to iOS users [98].

6.4.3 General Retailers

Home Depot. I now turn my attention to general retail sites. Among the 10 retailers we examined, only Home Depot revealed evidence of personalization. Similar to our findings on Travelocity, Home

⁴I spoke with Travelocity representatives in November 2014 and they explained that Travelocity offers discounts to users on all mobile devices, not just iOS devices.

Depot personalizes results for users with mobile browsers. In fact, the Home Depot website serves HTML with different structure and CSS to desktop browsers, Safari on iOS, and Chrome on Android.

Figure 6.8 depicts the results of my browser experiments on Home Depot. Strangely, Home Depot serves 24 search results per page to desktop browsers and Android, but serves 48 to iOS. As shown in Figure 6.8(a), on most days there is close to zero overlap between the results served to desktop and mobile browsers. Oddly, there are days when Home Depot briefly serves identical results to all browsers (e.g., the spike in Figure 6.8(a) on 4/22). The pool of results served to mobile browsers contains more expensive products overall, leading to higher nDCG scores for mobile browsers in Figure 6.8(c). Note that nDCG is calculated using the top k results on the page, which in this case is 24 to preserve fairness between iOS and the other browsers. Thus, Home Depot is steering users on mobile browsers towards more expensive products.

In addition to steering, Home Depot also discriminates against Android users. As shown in Figure 6.8(d), the Android treatment consistently sees differences on $\approx 6\%$ of prices (one or two products out of 24). However, the practical impact of this discrimination is low: the average price differential in Figure 6.8(e) for Android is $\approx \$0.41$. I manually examined the search results from Home Depot and could not determine why the Android treatment receives slightly increased prices. Prior work has linked price discrimination on Home Depot to changes in geolocation [107], but we control for this effect in my experiments.

It is possible that the differences I observe on Home Depot may be artifacts caused by different server-side implementations of the website for desktop and mobile users, rather than an explicit personalization algorithm. However, even if this is true, it still qualifies as personalization according to my definition (see Section 6.2.1) since the differences are deterministic and triggered by client-side state.

6.5 Concluding Discussion

Personalization has become an important feature of many web services in recent years. However, there is mounting evidence that e-commerce sites are using personalization algorithms to implement price steering and discrimination.

In this study, I build a measurement infrastructure to study price discrimination and steering on 16 top online retailers and travel websites. my method places great emphasis on controlling for various sources of noise in our experiments, since I have to ensure that the differences I see are actually a result of personalization algorithms and not just noise. *First*, I collect real-world data from

CHAPTER 6. MEASURING PERSONALIZATION OF ECOMMERCE SITES

300 AMT users to determine the extent of personalization that they experience. This data revealed evidence of personalization on four general retailers and five travel sites, including cases where sites altered prices by hundreds of dollars.

Second, I ran controlled experiments to investigate what features e-commerce personalization algorithms take into account when shaping content. I found cases of sites altering results based on the user's OS/browser, account on the site, and history of clicked/purchased products. I also observe two travel sites conducting A/B tests that steer users towards more expensive hotel reservations.

Comments from Companies. I reached out to the six companies I identified in this study as implementing some form of personalization (Orbitz and Cheaptickets are run by a single company, as are Expedia and Hotels.com) asking for comments on a draft of this study. I received responses from Orbitz and Expedia. The Vice President for Corporate Affairs at Orbitz provided a response confirming that Cheaptickets and Orbitz offer members-only deals on hotels. However, their response took issue with my characterization of price discrimination as “anti-consumer”; I removed these assertions from the final draft of this manuscript. The Orbitz representative kindly agreed to allow us to publish their letter on the web [31].

I also spoke on the phone with the Chief Product Officer and the Senior Director of Stats Optimization at Expedia. They confirmed my findings that Expedia and Hotels.com perform extensive A/B testing on users. However, they claimed that Expedia does not implement price discrimination on rental cars, and could not explain my results to the contrary (see Figure 6.3).

Chapter 7

Conclusion

Since the turn of the century, we have witnessed a trend of personalization in numerous Internet-based services, including web search and online retailers. While personalization provides obvious benefits for users, it also opens up the possibility that certain information may be unintentionally hidden from users. Moreover, the algorithms used to process, filter and recommend content might reinforce biases, which over time can lead to discrimination on a societal scale. This is worsened by the fact that these practices happen “under the hood” and users are highly unaware of them. Despite the variety of speculation on this topic, until my work, there has been little quantification of the basis and extent of personalization in web-based content services today.

In this thesis, I took the first steps towards quantifying the prevalence of personalization in web-based content services and understanding the algorithms behind them. My investigation started with developing a methodology that helped me quantify personalization on a service of my interest. To get the full picture I needed to both examine data from real user accounts as well as systematically test the algorithm. By controlling the data input to the algorithm, I could observe what triggers the changes in the results.

As I walked through my methodology, I demonstrated that measuring personalization and *only* personalization is a challenge in itself. It requires a tool that controls for numerous sources of noise allowing me to accurately identify personalization and isolate it from other sources of noise. Moreover, keeping in mind that these services often change their strategies, my objective was to make it easy for researchers to repeat my experiments at a later time or on different systems of interest.

With the new methodology in hand, I first turned my attention towards search engines. Search engines are people’s primary gateway to information thus it is important to make their operations as transparent as possible. In my study I inspected three large search engines: Google

CHAPTER 7. CONCLUSION

Search, Bing Search and DuckDuckGo. The data I collected from real user accounts showed that 11.7% of search results on Google and 15.8% on Bing show differences due to personalization. DuckDuckGo claims not to personalize content and thus serves as a baseline in my investigation about personalization. Using artificially created accounts, I observed that measurable personalization on Google Search and Bing is triggered by 1) being logged in to a user account and 2) making requests from different geographic areas. When examining trends over time, I found that changes between results are more often caused by reordering of the same results rather than new results appearing. Of course there is a difference in the volatility of results based on the topic of the search term: for politics and news related searches the changes are more frequent and new results enter the result pool more often.

Motivated by the finding that the physical location of users has the largest effect on personalization, I took a closer look at location-based personalization in Google's search algorithm. I developed a novel methodology that allowed me to query Google from any location around the world. Using this technique I sent 3,600 distinct queries to Google Search over a span of 30 days from 59 locations across the US. My findings showed that location does indeed have a large impact on search results, and that the differences increase as physical distance grows. However, I observed many nuances to Google's implementation of location-based personalization. First, not all types of queries trigger the algorithm to the same degree: politicians were essentially unaffected by geography; controversial terms see small changes due to News; and local terms see large differences due to changes in Maps and normal results. Second, not all queries expected to trigger location-personalization do: for example, search results for brand names like "Starbucks" do not include Maps. Finally, and most surprisingly, I also discovered that Google Search returns search results that are very noisy, especially for local queries. This non-determinism is puzzling, since Google knows the precise location of the user (during our experiments), and thus should be able to quickly calculate the closest set of relevant locations.

Lastly, I turned my attention to another important branch of online services; e-commerce sites. Using the methodology presented in Chapter 3 I investigated price discrimination and steering on 16 top online retailers and travel websites. *First*, I collected real-world data from 300 AMT users to determine the extent of personalization that they experience. This data revealed evidence of personalization on four general retailers and five travel sites, including cases where sites altered prices by hundreds of dollars. *Second*, I ran controlled experiments to investigate what features e-commerce personalization algorithms take into account when shaping content. I found cases of sites altering results based on the user's OS/browser, account on the site, and history of clicked/purchased products.

CHAPTER 7. CONCLUSION

I also observed two travel sites conducting A/B tests that steer users towards more expensive hotel reservations.

Hopefully my work convinced the reader that measuring the internal mechanisms and effects of big data algorithms is an important and complex challenge. It is no question that developing these tools is crucial. The demand for this is clearly apparent among all the different actors of the ecosystem: users need to understand what happens under the hood of the services they use in order to trust them, regulators require the insight to be able to create and enforce laws, and even the operators of the services often need external resources to measure the large-scale and long-term effects of their own algorithms. Addressing these needs requires the collaboration of researchers in many different disciplines: legal scholars, social scientists and researchers with the technical and computational background all contribute to the quickly growing literature of this young area. Seeing the large amount of on-going work since the start of my research program is the perfect confirmation that I indeed touched on a very important topic.

Finally, I made all of the crawling and parsing code, as well as the collected data from both the search engines and e-commerce studies available to the research community at

<http://personalization.ccs.neu.edu/>

Bibliography

- [1] E. Agichtein, E. Brill, and S. Dumais. Improving Web Search Ranking By Incorporating User Behavior Information. In *Conference of the ACM Special Interest Group on Information Retrieval*, Seattle, Washington, USA, August 2006.
- [2] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami. An Interval Classifier For Database Mining Applications. In *VLDB*, 1992.
- [3] R. Agrawal, B. Golshan, and E. Papalexakis. Whither Social Networks For Web Search? In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, 2015.
- [4] Alexa Top 500 Global Sites. <http://www.alexa.com/topsites>.
- [5] Amazon Mechanical Turk. <http://mturk.com/>.
- [6] J. Andersen, A. Giversen, A. H. Jensen, R. S. Larsen, T. B. Pedersen, and J. Skyt. Analyzing Clickstreams Using Subsessions. In *ACM International Workshop On Data Warehousing and OLAP*, McLean, Virginia, USA, November 2000.
- [7] P. Baker and A. Potts. 'why Do White People Have Thin Lips?' Google And The Perpetuation Of Stereotypes Via Auto-complete Search Forms. *Critical Discourse Studies*, 10(2):187–204, 2013.
- [8] A. Ballatore. Google Chemtrails: A Methodology To Analyze Topic Representation In Search Engine Results. *First Monday*, 20(7), 2015.
- [9] J. Bar-Ilan, K. Keenoy, E. Yaari, and M. Levene. User Rankings Of Search Engine Results. *Journal of the American Society for Information Science and Technology*, 58(9):1254–1266, July 2007.

BIBLIOGRAPHY

- [10] S. Barocas and H. Nissenbaum. Big Datas End Run Around Anonymity And Consent. *Privacy, big data, and the public good: Frameworks for Engagement*, pages 44–75, 2014.
- [11] K. Bawa and R. W. Shoemaker. The Effects Of A Direct Mail Coupon On Brand Choice Behavior. *Journal of Marketing Research*, pages 370–376, 1987.
- [12] K. Bawa and R. W. Shoemaker. Analyzing Incremental Sales From A Direct Mail Coupon Promotion. *The Journal of Marketing*, pages 66–78, 1989.
- [13] P. Belobaba, A. Odoni, and C. Barnhart. *The Global Airline Industry*. Wiley, 2009.
- [14] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring And Using Location Metadata To Personalize Web Search. In *Conference of the ACM Special Interest Group on Information Retrieval*, 2011.
- [15] Big Data: Seizing Opportunities, Preserving Values, 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
- [16] Bing It On. <http://www.bingiton.com>.
- [17] R. C. Blattberg and J. Deighton. Interactive Marketing: Exploiting The Age Of Addressability. *Sloan Management Review*, 33(1):5–14, 1991.
- [18] P. Boutin. Your Results May Vary: Will The Information Superhighway Turn Into A Cul-de-sac Because Of Automated Filters? *The Wall Street Journal*, May 2011. <http://www.wsj.com/articles/SB10001424052748703421204576327414266287254>.
- [19] Google Filters Sites In France And Germany, 2002. <http://www.internetnews.com/bus-news/article.php/1488031/Google+Filters+Sites+in+France+and+Germany.htm>.
- [20] S. Brin and L. Page. Reprint Of: The Anatomy Of A Large-scale Hypertextual Web Search Engine. *Computer Networks*, 56(18):3825–3833, 2012.
- [21] A. Broder. A Taxonomy Of Web Search. *ACM SIGIR Forum*, 36(2):3–10, 2002.
- [22] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling Attribute Effect In Linear Regression. In *International Conference on Data Mining*, 2013.

BIBLIOGRAPHY

- [23] T. Calders and S. Verwer. Three Naive Bayes Approaches For Discrimination-free Classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
- [24] J. M. Carrascosa, J. Mikians, R. Cuevas, V. Erramilli, and N. Laoutaris. I Always Feel Like Somebody’s Watching Me. Measuring Online Behavioural Advertising. *Computing Research Repository*, 2014.
- [25] B. Carterette. On Rank Correlation And The Distance Between Rankings. In *Conference of the ACM Special Interest Group on Information Retrieval*, Boston, Massachusetts, USA, July 2009.
- [26] D. Chan, D. Kumar, S. Ma, and J. Koehler. Impact Of Ranking Of Organic Search Results On The Incrementality Of Search Ads. Technical Report 37731, Google, Inc., 2012.
- [27] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank For Graded Relevance. In *ACM International Conference on Information and Knowledge Management*, Hong Kong, China, November 2009.
- [28] O. Chapelle and Y. Zhang. A Dynamic Bayesian Network Click Model For Web Search Ranking. In *International World Wide Web Conference*, Madrid, Spain, April 2009.
- [29] L. Chen, A. Mislove, and C. Wilson. An Empirical Analysis Of Algorithmic Pricing On Amazon Marketplace. In *International World Wide Web Conference*, Montréal, Canada, April 2016.
- [30] Z. Cheng, B. Gao, and T.-Y. Liu. Actively Predicting Diverse Search Intent From User Browsing Behaviors. In *International World Wide Web Conference*, Raleigh, North Carolina, USA, April 2010.
- [31] C. Chiames. Correspondence with the authors, in reference to a pre-publication version of this manuscript, 2014. http://personalization.ccs.neu.edu/orbitz_letter.pdf.
- [32] V. Ciesielski and G. Palstra. Using A Hybrid Neural/expert System For Data Base Mining In Market Survey Data. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, 1996.
- [33] F. Clarke. Personalized Coupon Generating And Processing System, March 1996. US Patent 5,502,636.

BIBLIOGRAPHY

- [34] Comscore August 2012 U.S. Search Results, 2012. <http://bit.ly/ThGnOc>.
- [35] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An Experimental Comparison Of Click Position-bias Models. In *ACM International Conference of Web Search and Data Mining*, Stanford, California, USA, February 2008.
- [36] E. Cutrell and Z. Guan. What Are You Looking For?: An Eye-tracking Study Of Information Usage In Web Search. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*, San Jose, California, USA, April 2007.
- [37] F. J. Damerau. A Technique For Computer Detection And Correction Of Spelling Errors. *Communications of the ACM*, 7(3):171–176, 1964.
- [38] A. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In *International World Wide Web Conference*, Banff, Canada, May 2007.
- [39] P. Deutsch. Archie-a Darwinian Development Process. *Internet Computing*, 4(1):69–71, 2000.
- [40] P. Diaconis and R. L. Graham. Spearman’s Footrule As A Measure Of Disarray. *Journal of the Royal Statistical Society (B: Methodological)*, 39(2):262–268, 1977.
- [41] N. Diakopoulos. Algorithmic Accountability: Journalistic Investigation Of Computational Power Structures. *Digital Journalism*, 3(3):398–415, 2015.
- [42] J. H. Dorfman. *Economics And Management Of The Food Industry*. Routledge, 2013.
- [43] Z. Dou, R. Song, and J.-R. Wen. A Large-scale Evaluation And Analysis Of Personalized Search Strategies. In *International World Wide Web Conference*, Banff, Canada, May 2007.
- [44] G. Ducoffe, M. Lécuyer, A. Chaintreau, and R. Geambasu. Web Transparency For Complex Targeting: Algorithms, Limits, And Tradeoffs. In *SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, 2015.
- [45] G. E. Dupret and Benjamin Piwowarski. A User Browsing Model To Predict Search Engine Click Data From Past Observations. In *Conference of the ACM Special Interest Group on Information Retrieval*, Singapore, July 2008.

BIBLIOGRAPHY

- [46] R. Epstein and R. E. Robertson. The Search Engine Manipulation Effect (seme) And Its Possible Impact On The Outcomes Of Elections. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, 112(33):4512–4521, 2015.
- [47] M. Eslami, A. Aleyasen, K. Karahalios, K. Hamilton, and C. Sandvig. Feedvis: A Path For Exploring News Feed Curation Algorithms. In *ACM conference on Computer Supported Cooperative Work*, 2015.
- [48] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig. “i Always Assumed That I Wasn’t Really That Close To [her]”: Reasoning About Invisible Algorithms In The News Feed. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*, 2015.
- [49] R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top k Lists. In *Symposium on Discrete Algorithms*, Baltimore, Maryland, USA, January 2003.
- [50] W. Fan, M. D. Gordon, and P. Pathak. Personalization Of Search Engine Services For Effective Retrieval And Knowledge Management. In *Annual Conference on Information Systems*, Atlanta, Georgia, USA, December 2000.
- [51] A. T. Fernando. Privacy-enhanced Personalisation Of Web Search. In *User Modeling, Adaptation and Personalization*, pages 385–390. Springer, 2015.
- [52] A. T. Fernando, J. T. Du, and H. Ashman. Personalisation Of Web Search: Exploring Search Query Parameters And User Information Privacy Implications-the Case Of Google. In *Conference of the ACM Special Interest Group on Information Retrieval*, 2014.
- [53] S. Flaxman, S. Goel, and J. M. Rao. Ideological Segregation And The Effects Of Social Media On News Consumption. *Social Science Research Network Working Paper Series*, 2013.
- [54] S. R. Flaxman, S. Goel, and J. M. Rao. Filter Bubbles, Echo Chambers, And Online News Consumption. *Public Opinion Quarterly*, 2015.
- [55] B. C. for Internet & Society. Replacement Of Google With Alternative Search Systems In China, 2002. <http://cyber.law.harvard.edu/filtering/china/google-replacements/>.
- [56] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based Personalized Search And Browsing. *Web Intelligence and Agent Systems*, 1:1–3, 2003.

BIBLIOGRAPHY

- [57] S. Goel, J. M. Hofman, and M. I. Siner. Who Does What On The Web: A Large-scale Study Of Browsing Behavior. In *International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, June 2012.
- [58] Google. Personalized Search Graduates From Google Labs. News From Google Blog, 2005. http://googlepress.blogspot.com/2005/11/personalized-search-graduates-from_10.html.
- [59] Google Zeitgeist, 2012. <http://www.googlezeitgeist.com>.
- [60] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking Analysis Of User Behavior In Www Search. In *Conference of the ACM Special Interest Group on Information Retrieval*, Sheffield, United Kingdom, July 2004.
- [61] H. Green. Breaking Out Of Your Internet Filter Bubble, August 2011. <http://onforb.es/oYWbDf>.
- [62] R. Gross and A. Acquisti. Information Revelation And Privacy In Online Social Networks (the Facebook Case). In *Workshop on Privacy in the Electronic Society*, Alexandria, Virginia, USA, November 2005.
- [63] Z. Guan and E. Cutrell. An Eye Tracking Study Of The Effect Of Target Rank On Web Search. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*, San Jose, California, USA, April 2007.
- [64] S. Guha, B. Cheng, and P. Francis. Challenges In Measuring Online Advertising Systems. In *ACM Internet Measurement Conference*, Melbourne, Victoria, Australia, November 2010.
- [65] Q. Guo and E. Agichtein. Beyond Dwell Time: Estimating Document Relevance From Cursor Movements And Other Post-click Searcher Behavior. In *International World Wide Web Conference*, Lyon, France, April 2012.
- [66] S. Hajian and J. Domingo-Ferrer. A Methodology For Direct And Indirect Discrimination Prevention In Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459, 2013.
- [67] A. Halavais. *Search Engine Society*. John Wiley & Sons, 2013.

BIBLIOGRAPHY

- [68] A. Hannak, P. Sapiezynski, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring Personalization Of Web Search. In *International World Wide Web Conference*, Rio de Janeiro, Brazil, May 2013.
- [69] A. Hannak, G. Soeller, D. Lazer, A. Mislove, and C. Wilson. Measuring Price Discrimination And Steering On E-commerce Web Sites. In *ACM Internet Measurement Conference*, Vancouver, Canada, November 2014.
- [70] K. Harrenstien and V. White. Whois. <https://tools.ietf.org/html/rfc812>.
- [71] Search Personalization Using Machine Learning. http://faculty.washington.edu/hemay/search_personalization.pdf.
- [72] K. Hillis, M. Petit, and K. Jarrett. *Google And The Culture Of Search*. Routledge, 2012.
- [73] K. Hosanagar, D. Fleder, D. Lee, and A. Buja. Will The Global Village Fracture Into Tribes? Recommender Systems And Their Effects On Consumer Fragmentation. *Management Science*, 60(4):805–823, 2013.
- [74] B. E. Hosken. Automated Content And Collaboration-based System And Methods For Determining And Providing Content Recommendations, August 2002. US Patent 6,438,579.
- [75] Html5 Geolocation Api. <http://dev.w3.org/geo/api/spec-source.html>.
- [76] B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. Characterizing Search Intent Diversity Into Click Models. In *International World Wide Web Conference*, Hyderabad, India, April 2011.
- [77] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic Prediction Based On User's Browsing Behavior. In *International World Wide Web Conference*, Banff, Canada, May 2007.
- [78] J. J. Inman and L. McAlister. Do Coupon Expiration Dates Affect Consumer Behavior? *Journal of Marketing Research*, 31(3):423–428, 1994.
- [79] L. D. Introna and H. Nissenbaum. Shaping The Web: Why The Politics Of Search Engines Matters. *The Information Society*, 16(3):169–185, 2000.
- [80] K. Järvelin and J. Kekäläinen. Ir Evaluation Methods For Retrieving Highly Relevant Documents. In *Conference of the ACM Special Interest Group on Information Retrieval*, Athens, Greece, July 2000.

BIBLIOGRAPHY

- [81] K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation Of Ir Techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.
- [82] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately Interpreting Click-through Data As Implicit Feedback. In *Conference of the ACM Special Interest Group on Information Retrieval*, Sheffield, United Kingdom, July 2005.
- [83] M. Juarez and V. Torra. Dispa: An Intelligent Agent For Private Web Search. In *Advanced Research in Data Privacy*, pages 389–405. Springer, 2015.
- [84] F. Kamiran, A. Karim, and X. Zhang. Decision Theory For Discrimination-aware Classification. In *International Conference on Data Mining*, Brussels, Belgium, December 2012.
- [85] M. Kay, C. Matuszek, and S. A. Munson. Unequal Representation And Gender Stereotypes In Image Search Results For Occupations. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*, 2015.
- [86] M. G. Kendall. A New Measure Of Rank Correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [87] R. Kumar and S. Vassilvitskii. Generalized Distances Between Rankings. In *International World Wide Web Conference*, Raleigh, North Carolina, USA, April 2010.
- [88] S. Lazar. Algorithms And The Filter Bubble Ruining Your Online Experience? *Huffington Post*, June 2011. http://www.huffingtonpost.com/shira-lazar/algorithms-and-the-filter_b_869473.html.
- [89] M. Lecuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu. Xray: Enhancing The Web’s Transparency With Differential Correlation. In *USENIX Security Symposium*, San Diego, California, USA, August 2014.
- [90] M. Lecuyer, R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu. Sunlight: Fine-grained Targeting Detection At Scale With Statistical Confidence. In *ACM Conference on Computer and Communications Security*, 2015.
- [91] M. Lee, T. Ha, J. Han, and J.-Y. Rha. Online Footsteps To Purchase: Exploring Consumer Behaviors On Online Shopping Sites. *WebSci*, 2015.
- [92] R. P. Leone and S. S. Srinivasan. Coupon Face Value: Its Impact On Coupon Redemptions, Brand Sales, And Brand Profitability. *The Journal of Retailing*, 72(3):273–289, 1996.

BIBLIOGRAPHY

- [93] M. J. Lewis, C. D. Delnevo, and J. Slade. Tobacco Industry Direct Mail Marketing And Participation By New Jersey Adults. *American Journal of Public Health*, 94(2):257–259, 2004.
- [94] G. Linden, B. Smith, and J. York. Amazon.Com Recommendations: Item-to-item Collaborative Filtering. *Internet Computing*, 7(1):76–80, 2003.
- [95] C. X. Ling and C. Li. Data Mining For Direct Marketing: Problems And Solutions. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, New York, August 1998.
- [96] F. Liu, C. Yu, and W. Meng. Personalized Web Search By Mapping User Queries To Categories. In *ACM International Conference on Information and Knowledge Management*, McLean, Virginia, USA, November 2002.
- [97] F. Liu, C. Yu, and W. Meng. Personalized Web Search For Improving Retrieval Effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, January 2004.
- [98] B. D. Lollis. Orbitz: Mobile Searches May Yield Better Hotel Deals. USA Today Travel Blog, 2012. <http://travel.usatoday.com/hotels/post/2012/05/orbitz-mobile-hotel-deals/691470/1>.
- [99] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay. The Influence Of Task And Gender On Search And Evaluation Behavior Using Google. *Information Processing and Management*, 42(4):1123–1131, 2006.
- [100] A. Lyles. Direct Marketing Of Pharmaceuticals To Consumers. *Annual review of public health*, 23(1):73–91, 2002.
- [101] A. Majumder and N. Shrivastava. Know Your Personalization: Learning Topic Level Personalization In Online Services. In *International World Wide Web Conference*, Rio de Janeiro, Brazil, May 2013.
- [102] S. V. Malthankar and S. Kolte. Client Side Privacy Protection Using Personalized Web Search. *Procedia Computer Science*, 79:1029–1035, 2016.
- [103] D. Mattioli. On Orbitz, Mac Users Steered To Pricier Hotels, August 2012. <http://on.wsj.com/LwTnPH>.

BIBLIOGRAPHY

- [104] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big Data. *Harvard Business Review*, 90(10):61–67, 2012.
- [105] A. Micarelli, F. Gaspiretti, F. Sciarrone, and S. Gauch. Personalized Search On The World Wide Web. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, pages 195–230. Springer-Verlag, Berlin, Heidelberg, 2007.
- [106] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting Price And Search Discrimination On The Internet. In *Workshop on Hot Topics in Networks*, Seattle, Washington, USA, October 2012.
- [107] J. Mikians, László Gyarmati, V. Erramilli, and N. Laoutaris. Crowd-assisted Search For Price Discrimination In E-commerce: First Results. In *International Conference on Emerging Networking Experiments and Technologies*, Santa Barbara, California, USA, December 2013.
- [108] B. Mobasher, R. Cooley, and J. Srivastava. Automatic Personalization Based On Web Usage Mining. *Communications of the ACM*, 43(8):142–151, 2000.
- [109] S. A. Neslin. A Market Response Model For Coupon Promotions. *Marketing Science*, 9(2):125–145, 1990.
- [110] S. A. Neslin and D. G. Clarke. Relating The Brand Use Profile Of Coupon Redeemers To Brand And Coupon Characteristics. *Journal of Advertising Research*, 27(1):23–32, 1987.
- [111] The Age Of Big Data.
- [112] M. G. Noll and C. Meinel. Web Search Personalization Via Social Bookmarking And Tagging. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, Busan, November 2007.
- [113] G. J. Nowak and J. Phelps. Direct Marketing And The Use Of Individual-level Consumer Information: Determining How And When 'privacy' Matters. *Journal of Direct Marketing*, 9(3):46–60, 1995.
- [114] M. F. O'brien, G. W. Off, T. L. Cherney, and G. M. Katz. Method And Apparatus For Selective Distribution Of Discount Coupons Based On Prior Customer Behavior, November 1998. US Patent 5,832,457.

BIBLIOGRAPHY

- [115] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank Citation Ranking: Bringing Order To The Web. 1999.
- [116] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In Google We Trust: Users' Decisions On Rank, Position, And Relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.
- [117] A. Pansari and M. Mayer. This Is A Test. This Is Only A Test., April 2006. <http://bit.ly/Ldbb0>.
- [118] E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Press, London, UK, 2011.
- [119] L. Parramore. The Filter Bubble. The Atlantic, October 2010. <http://www.theatlantic.com/daily-dish/archive/2010/10/the-filter-bubble/181427/>.
- [120] M. J. Pazzani and D. Billsus. Content-based Recommendation Systems. In *The adaptive web*, pages 325–341. Springer, 2007.
- [121] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware Data Mining. In *sigkdd*, 2008.
- [122] J. Perrien and L. Ricard. The Meaning Of A Marketing Relationship: A Pilot Study. *Industrial Marketing Management*, 24(1):37–43, 1995.
- [123] Phantomjs, 2015. <http://phantomjs.org>.
- [124] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized Search. *Communications of the ACM*, 45(9):50–55, September 2002.
- [125] A. Pretschner and S. Gauch. Ontology Based Personalized Search. In *IEEE International Conference on Tools with Artificial Intelligence*, Chicago, Illinois, USA, November 1999.
- [126] J. Purra. Swedes Online: You Are More Tracked Than You Think. 2015.
- [127] F. Qiu and J. Cho. Automatic Identification Of User Interest For Personalized Search. In *International World Wide Web Conference*, Edinburgh, Scotland, May 2006.
- [128] Quantcast. Top Sites For The United States, 2012. <http://www.quantcast.com/top-sites>.

BIBLIOGRAPHY

- [129] F. Radlinski and T. Joachims. Query Chains: Learning To Rank From Implicit Feedback. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, Chicago, Illinois, USA, August 2005.
- [130] F. Radlinski and T. Joachims. Active Exploration For Learning Rankings From Clickthrough Data. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, San Jose, California, USA, August 2007.
- [131] D. J. Reibstein and P. A. Traver. Factors Affecting Coupon Redemption Rates. *The Journal of Marketing*, 46(4):102–113, 1982.
- [132] P. Reilly. 'googling' Terrorists: Are Northern Irish Terrorists Visible On Internet Search Engines? Springer, 2008.
- [133] F. Richhled and W. E. S. Jr. Zero Defections: Quality Comes To Services. *Harvard Business Review*, 68(5):105–11, 1990.
- [134] F. Robinson. Eu Tells Google To Offer More In Search Probe, July 2013. <http://online.wsj.com/article/SB10001424127887323993804578611362017999002.html>.
- [135] F. Roesner, T. Kohno, and D. Wetherall. Detecting And Defending Against Third-party Tracking On The Web. In *Symposium on Networked System Design and Implementation*, San Jose, California, USA, April 2012.
- [136] P. E. Rossi, R. E. McCulloch, and G. M. Allenby. The Value Of Purchase History Data In Target Marketing. *Marketing Science*, 15(4):321–340, 1996.
- [137] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing Algorithms: Research Methods For Detecting Discrimination On Internet Platforms. In *International Communication Association Conference*, 2014.
- [138] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann. Tracking Human Mobility Using Wifi Signals. *PLoS One*, 10(7), 2015.
- [139] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis Of Recommendation Algorithms For E-commerce. In *ACM Conference on Economics and Computation*, Minneapolis, Minnesota, USA, October 2000.

BIBLIOGRAPHY

- [140] C. Schwartz. Web Search Engines. *Journal of the American Society for Information Science and Technology*, 49(11):973, 1998.
- [141] D. Sculley. Rank Aggregation For Similar Items. In *SIAM International Conference on Data Mining*, Minneapolis, Minnesota, USA, April 2007.
- [142] Selenium, 2013. <http://selenium.org>.
- [143] X. Shen, B. Tan, and C. Zhai. Implicit User Modeling For Personalized Search. In *ACM International Conference on Information and Knowledge Management*, Bremen, Germany, November 2005.
- [144] Y. Shen, J. Yan, S. Yan, L. Ji, N. Liu, and Z. Chen. Sparse Hidden-dynamics Conditional Random Fields For User Intent Understanding. In *International World Wide Web Conference*, Hyderabad, India, April 2011.
- [145] J. N. Sheth and A. Parvatiyar. The Evolution Of Relationship Marketing. *International Business Review*, 4(4):397–418, 1995.
- [146] G. S. Shieh, Z. Bai, and W.-Y. Tsai. Rank Tests For Independence—With A Weighted Contamination Alternative. *Statistica Sinica*, 10:577–593, 2000.
- [147] A. Sieg, B. Mobasher, and R. Burke. Web Search Personalization With Ontological User Profiles. In *ACM International Conference on Information and Knowledge Management*, 2007.
- [148] N. Singer. The Trouble With The Echo Chamber Online, May 2011. <http://nyti.ms/jcTih2>.
- [149] B. R. Smith, G. D. Linden, and N. K. Zada. Content Personalization Based On Actions Performed During A Current Browsing Session, February 2005. U.S. Patent 6,853,982.
- [150] G. Soeller, K. Karahalios, C. Sandvig, and C. Wilson. Mapwatch: Detecting And Monitoring International Border Personalization On Online Maps. In *International World Wide Web Conference*, Montréal, Canada, April 2016.
- [151] C. Spearman. The Proof And Measurement Of Association Between Two Things. *American Journal of Psychology*, 15(1):72–101, 1904.

BIBLIOGRAPHY

- [152] J. Stoyanovich, S. Abiteboul, and G. Miklau. Data, Responsibly: Fairness, Neutrality And Transparency In Data Analysis. In *International Conference on Extending Database Technology*, 2016.
- [153] D. Sullivan. Bing Results Get Local And Personalized, February 2011. <http://selnd.com/hY4djp>.
- [154] D. Sullivan. Why Google “personalizes” Results Based On Obama Searches But Not Romney, November 2012. <http://selnd.com/PyfvvY>.
- [155] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: A Novel Approach To Personalized Web Search. In *International World Wide Web Conference*, Chiba, Japan, May 2005.
- [156] M. Sun, G. Lebanon, and K. Collins-Thompson. Visualizing Differences In Web Search Algorithms Using The Expected Weighted Hoeffding Distance. In *International World Wide Web Conference*, Raleigh, North Carolina, USA, April 2010.
- [157] L. Sweeney. Discrimination In Online Ad Delivery. In *Social Science Research Network Working Paper Series*, 2013.
- [158] L. Sydell. How Rick Santorum’s ‘google Problem’ Has Endured, 2012. <http://n.pr/wefdnc>.
- [159] B. Tan, X. Shen, and C. Zhai. Mining Long-term Search History To Improve Search Accuracy. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, Philadelphia, Pennsylvania, USA, August 2006.
- [160] J. E. Teel, R. H. Williams, and W. O. Bearden. Correlates Of Consumer Susceptibility To Coupons In New Grocery Product Introductions. *Journal of Advertising*, 9(3):31–46, 1980.
- [161] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing Search Via Automated Analysis Of Interests And Activities. In *Conference of the ACM Special Interest Group on Information Retrieval*, Sheffield, United Kingdom, July 2005.
- [162] T. Terano and Y. Ishino. Interactive Knowledge Discovery From Marketing Questionnaire Using Simulated Breeding And Inductive Learning Methods. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, 1996.
- [163] Top 500 E-retailers. <http://www.top500guide.com/top-500/>.

BIBLIOGRAPHY

- [164] Top Booking Sites. <http://skift.com/2013/11/11/top-25-online-booking-sites-in-travel/>.
- [165] Untitled. Bezos Calls Amazon Experiment 'a Mistake'. Puget Sound Business Journal, 2000. <http://www.bizjournals.com/seattle/stories/2000/09/25/daily21.html>.
- [166] Untitled. Invisible Sieve: Hidden, Specially For You. The Economist, June 2011. http://www.economist.com/node/18894910?story_id=18894910&fsrc=rss.
- [167] J. Valentino-Devries, J. Singer-Vine, and A. Soltani. Websites Vary Prices, Deals Based On Users' Information, December 2012. <http://online.wsj.com/news/articles/SB10001424127887323777204578189391813881534>.
- [168] L. Vaughan. New Measurements For Search Engine Evaluation Proposed And Tested. *Information Processing and Management*, 40(4):677–691, May 2004.
- [169] L. Vaughan and M. Thelwall. Search Engine Coverage Bias: Evidence And Possible Causes. *Information Processing and Management*, 40(4):693–707, 2004.
- [170] T. Vissers, N. Nikiforakis, N. Bielova, and W. Joosen. Crying Wolf? On The Price Discrimination Of Online Airline Tickets. In *Hot Topics in Privacy Enhancing Technologies*, Amsterdam, The Netherlands, July 2014.
- [171] T. Wadhwa. How Advertisers Can Use Your Personal Information To Make You Pay Higher Prices, January 2014. http://www.huffingtonpost.com/tarun-wadhwa/how-advertisers-can-use-y_b_4703013.html.
- [172] Webmd Year In Health, 2011. <http://on.webmd.com/eBPFxH>.
- [173] J. Weisberg. Bubble Trouble: Is Web Personalization Turning Us Into Solipsistic Twits? Slate, June 2011. http://www.slate.com/articles/news_and_politics/the_big_idea/2011/06/bubble_trouble.html.
- [174] R. W. White. Beliefs And Biases In Web Search. In *Conference of the ACM Special Interest Group on Information Retrieval*, 2013.
- [175] D. K. Wind, P. Sapiezynski, M. A. Furman, and S. Lehmann. Inferring Stop-locations From Wifi. *PLoS One*, 11(2), 2016.

BIBLIOGRAPHY

- [176] M. Wines. Google To Alert Users To Chinese Censorship, June 2012. <http://nyti.ms/JRhGZS>.
- [177] X. Xing, W. Meng, D. Doozan, A. C. Snoeren, N. Feamster, and W. Lee. Take This Personally: Pollution Attacks On Personalized Services. In *USENIX Security Symposium*, Washington, D.C., USA, August 2013.
- [178] Y. Xu, B. Zhang, Z. Chen, and K. Wang. Privacy-enhancing Personalized Web Search. In *International World Wide Web Conference*, Banff, Canada, May 2007.
- [179] X. Yi, H. Raghavan, and C. Leggetter. Discovering Users' Specific Geo Intention In Web Search. In *International World Wide Web Conference*, Madrid, Spain, April 2009.
- [180] E. Yilmaz, J. A. Aslam, and S. Robertson. A New Rank Correlation Coefficient For Information Retrieval. In *Conference of the ACM Special Interest Group on Information Retrieval*, Singapore, July 2008.
- [181] H. Yoganasimhan. Search Personalization. Technical report, Working paper, 2014.
- [182] B. Yu and G. Cai. A Query-aware Document Ranking Method For Geographic Information Retrieval. In *ACM Workshop On Geographic Information Retrieval*, Lisbon, Portugal, November 2007.
- [183] Fairness Constraints: A Mechanism For Fair Classification. http://www.mpi-sws.org/~mzafar/papers/fatml_15.pdf.
- [184] Z. Zhao, Z. Cheng, L. Hong, and E. H. Chi. Improving User Topic Interest Profiles By Behavior Factorization. In *International World Wide Web Conference*, pages 1406–1416, Florence, Italy, April 2015.