Doctoral Thesis Proposal

Personalization in Online Services: Measurement, Analysis, and Implications

Aniko Hannak
College of Computer and Information Science
Northeastern University
ancsaaa@ccs.neu.edu

August 21, 2014

Abstract

An increasing number of web services are now using personalization algorithms to shape the content they serve to their users in order to best meet users' tastes and needs. In many cases, personalization provides advantages for users: for example, users now expect search engines to return local results when searching for restaurants. However, the increasing level of personalization and the lack of transparency is now leading to concerns about its potential negative effects: in case of web search engines, personalization might lead to users not being able to access information that the search engine's algorithm decides is irrelevant. Or, in case of e-commerce sites, it might lead to price discrimination, where different users see different prices or products when shopping online. Despite these concerns, there has been little quantification of the extent of personalization on Web services, or the user attributes that cause it.

In this thesis proposal, I aim to develop the necessary tools to detect and quantify personalization. I first develop a general methodology for measuring personalization that can be used across multiple different web services. While conceptually simple, there are numerous details that the methodology must handle in order to accurately attribute differences in the content served to personalization. I collect data from real-world users to quantify the level of personalization observed today; I then build synthetic user accounts to investigate which user features drive personalization. This work is the first step towards understanding the extent and effects of personalization. Overall, the goal of this work is to both increase transparency into the operation of the web services as well as provide tools for the research community to further investigate the issue in other contexts.

Contents

1 Introduction

Personalization is a ubiquitous feature on today's top web destinations. Search engines such as Google, streaming-media services such as Netflix, and recommendation sites such as Yelp all use sophisticated algorithms to tailor content to each individual user. In many cases, personalization provides advantages for users. For example, when a user searches on Google with an ambiguous query such as "apple," there are multiple potential interpretations. By personalizing the search results (e.g., by taking the user's history of prior searches into account), Google is able to return results that are potentially more relevant (e.g., computer products, rather than orchards).

While the benefits of personalization are well-studied, the potential negative effects of personalization are not nearly as well-understood. Search engine operators do not typically label which of the returned results were personalized, or explain why those results were chosen; only operators themselves know the specifics of how the personalization algorithms alter the results. For example, Eli Pariser demonstrated that during the recent Egyptian revolution, different users searching for "Tahrir Square" received either links to news reports of protests, or links to travel agencies [28]. This example highlights the inherent danger in companies trying to serve the most relevant content to every user: it might result in other potentially important content being invisible to them. In Eli Pariser's example a user very interested in traveling might not hear about the news in Egypt since they are primarily served travel related content. The Filter Bubble effect is exacerbated by the dual issues that most users do not know that search results are personalized, yet users tend to place blind faith in the quality of search results [26].

Recently, researchers and Internet users have uncovered evidence that personalization is present on e-commerce sites [1, 24, 25] as well. On such sites, the benefits of personalization for users are less clear; e-commerce sites have a clear economic incentive to use personalization to induce users into spending more money. For example, the travel website Orbitz was shown to be personalizing the results of hotel searches [23]. Unbeknownst to users, Orbitz "steered" Mac OS X users towards more expensive hotels, in an attempt to extract more money from (presumably) more affluent individuals.

In this work I focus on measuring and understanding how personalization is used and implemented today. By building the tools to collect and analyse data from the top web destinations such as search engines and e-commerce sites I will increase transparency into their services and the public understanding of their operation.

I propose to first, develop a general methodology for measuring personalization on different Web Services. Measuring personalization is conceptually simple: one can run multiple searches for the same queries and compare the results. However, accurately attributing differences in returned search results to personalization requires accounting for a number of phenomena, including temporal changes in the search index, consistency issues in distributed search indices, and A/B tests being run by the search provider. The goal is' to develop a methodology that is able to successfully control for these phenomena.

Second, I will use this methodology to measure the extent of personalization on multiple popular Web search engines such as Google and Bing and some of the biggest online Purchasing Sites e.g. Home Depot or Expedia. I will implement an experiment that users can run in their browsers and that allows me to capture the pages returned to them. The infrastructure should be able to control for differences in time, location, distributed infrastructure, and noise, allowing me to attribute any differences observed to personalization. Running this experiment on populations with different traveling habits will allow me to further analyse how user history effects personalization.

Finally I plan to 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. This entails creating numerous accounts on each website and assign each a set of unique behaviors as well as developing a standard list of search queries that cover a variety of topics/products to then measure the differences in results that are returned for this list of searches. As a result I will be able to attribute differences in the returned results to the features assigned to each account. Additionally it will be interesting to investigate differences in query categories, intuitively more ambiguous query categories should result in more personalized content.

2 Related work

I briefly overview the academic literature on both implementing and measuring personalization. Most work on personalization has been in the context of web search, so much of the related work is in the context of that domain.

Personalization of search Personalized search has been extensively studied in the literature [21,30,32, 33,37,38,40,41]. Dou et al. provide a comprehensive overview of techniques for personalizing search [10]. They 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. A recent study is investigating the inherent biases of search engines and their impact on the quality of information that reaches people. [47] 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.

Other features besides click history have been used to power personalized search. Three studies leverage geographic location to personalize search [3, 51, 52]. Two studies have shown that user demographics can be reliably inferred from browsing histories, which can be useful for personalizing content [14,18]. To my knowledge, only one study has investigated privacy-preserving personalized search [50]. Given growing concerns about the Filter Bubble effects, this area seems promising for future research.

Personalization of e-commerce Very recent work has has begun to measure the personalization present in existing systems, such as web search engines [17,22] or recommender systems [4]. The first study focusing specifically on e-commerce sites was done by J. Mikians et al. [24] 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 [25] 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. I improve the method 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.

Improving Personalization Personalizing search results to improve Information Retrieval (IR) accuracy has been extensively studied in the literature [30, 40]. While these techniques typically use click histories for personalization, other features have also been used, including geographic location [3, 51, 52] and user demographics (typically inferred from search and browsing histories) [18, 46]. To my knowledge, only one study has investigated privacy-preserving personalized search [50]. Dou et al. provide a comprehensive overview of techniques for personalizing search [10]. Several studies have looked at improving personalization on systems other than search or e-commerce. Studies have examined personalization of targeted ads on the Web [16, 48] and news aggregators [9, 20] and found that location, behavior, profile information and social network information is used in targeting users.

Comparing Search Engines Several studies have examined the differences between results from different search engines. Two studies have performed user studies to compare search engines [5, 44]. Although both studies uncover significant differences between competing search engines, neither study examines the impact of personalization. Sun et al. propose a method for visualizing different results from search engines that is based on expected weighted Hoeffding distance [39]. Although this technique is very promising, it does not scale to the size of my experiments.

Exploiting Personalization Recent work has also shown that it is possible to exploit the mechanisms that create personalization for nefarious purposes. Xing et al. [49] demonstrated that repeatedly clicking on specific search results can cause search engines to rank those results higher; 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.

3 Methodology

This study seeks to answer two broad questions. First, to what extent does personalization affect the information users see? Although it is known that both search engines and online purchasing sites use personalization algorithms, it is not clear to what extent they alter the results they show to their users. Second, what are the user features that influence personalizations? This question is fundamental as outside of the companies themselves, nobody knows the specifics of how these algorithms work.

3.1 Terminology

Search Web Search and online purchasing sites have very similar mechanisms: they are both web services where people can search for certain keywords and this search returns a set of ordered results. In this study I use a specific set of terms when referring to *search*. Each *query* to a Web search engine or an e-commerce site is composed of one or more *keywords*. In response to a query, the site returns a page of *results*, which in case of Google or Bing is a page containing 10 results (URLs) and in the case of purchasing sites a list of products and associated prices.

Sources of Noise The main challenge when designing my experiments is noise, or inconsistencies in search results that are not due to personalization. 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:** Web search services constantly update their search indices. This means that the results for a query may change over time.

- **Updates to the E-commerce site:** E-commerce services are known to update their inventory often, as products sell out, become available, or prices are changed. This means that the results for a query may change even over short timescales.

- **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 datacenters.

- **Geolocation:** Web services use the user's IP address to provide localized results [51]. E-commerce sites might also account for price differences due to shipping costs, local taxes and currency conversions.

- **A/B testing:** Sites may conduct A/B testing [27], 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.

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 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 Implementation

To answer my two initial questions, I will use two different methods to collect data. To determine the extent of personalization in the "real-world", I collect data from real people, since they already have histories with the measured websites. To measure the effect of separate features leading to personalization, I will conduct controlled experiments with fake accounts.

All of the experiments are implemented using custom scripts for PhantomJS [29]. 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 [36]).

Real-World data collection 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 conduct a user study. The first part of the experiment asks the users to fill out a short survey with basic demographic questions as well as questions about their past interactions with the measured websites. This will be important information once I try to correlate the level of personalization I observe with their past behavior.

The second part of the experiment is about collecting users' search results. Participants were 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 were directed to visit a web page containing JavaScript that performed my set of searches in an `iframe`. After each search, the

Javascript would grab the HTML in the `iframe` and upload it back to the server, allowing me to view the results of the search. By having the user run the searches within their own browser, any cookies that the user's browser had previously been assigned would automatically be forwarded in my searches. This allows me to examine the results that the user would have received. I waited 15 seconds between each search, and the overall experiment took about 45 minutes to complete.

The HTTP proxy serves two important functions. *First*, 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: the comparison query allows me to compare results served to the AMT users with results served to "blank" users and the control will show the underlying noise when compared to the (identical) comparison query. *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.[1]

Collecting Synthetic Account Data To examine which user features web search services use to personalize results, I conduct controlled experiments with fake accounts created by me. The idea is to have a separate experiment with a set of fake accounts to examine the impact of each specific feature on personalization.

To assess the impact of feature X that can take on values $x_1, x_2, \ldots, x_n$, 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. PhantomJS downloads the first page of results for each query. For example, if I want assess the effect of gender on personalization, I will create 4 "personas": one "male," one "female," one "other," and one additional "female" as a control. I will execute $x + 1$ instances of the PhantomJS script for each experiment, and forward the traffic to $x + 1$ unique endpoints via SSH tunnels. Unless otherwise specified, PhantomJS persists all cookies between experiments. All of my experiments are designed to complete in <24 hours so that I can repeat them daily.

4 Personalization of Web Search

My first project focuses on measuring personalization on Google Search as it is not just the most popular search engine but the number one web destination according to Alexa [2]. The study seeks to answer two broad questions. First, *what user features influence Google's search personalization algorithms?* Second, *to what extent does search personalization actually affect search results?* Although it is known that Google personalizes 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.

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 not know which queries Web search engines 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.

As shown in Table 1, I use 120 queries divided equally over 12 categories in my experiments. These queries were chosen from the 2011 Google Zeitgeist [15], and WebMD [45]. 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.

[1] I also hard-coded the DNS mapping for each of the sites on the proxy to avoid discrepancies that might come from round-robin DNS hitting different data centers.

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

Table 1: Categories of search queries used in my experiments.

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 [45].

4.1 Collecting Real-World Data

I posted a task on Amazon's Mechanical Turk (AMT), explaining my study and offering each user $2.00 to participate. 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.[2] Users who accepted the task were instructed to configure their Web browser to use a HTTP proxy controlled by me. Then, the users were directed to visit a Web page that automatically performed 80 Google searches. 50 of the queries were randomly chosen from the categories in Table 1, while 30 were chosen by me.

AMT Worker Demographics. In total, I recruited 200 AMT workers, each of whom answered a brief demographic survey. My participants self-reported to residing in 43 different U.S. states, and range in age from 12 to >48 (with a bias towards younger users). Figure 1 shows the usage of Google services by my participants: 84% are Gmail users, followed by 76% that use Google Maps. These survey results demonstrate that my participants 1) come from a broad sample of the U.S. population, and 2) use a wide variety of Google services.

Results To answer the question of *how often do real users receive personalized search results*, I compare the results received by AMT users and the corresponding control accounts. Figure 2 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 two takeaways from Figure 2. 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. This additional difference can be attributed to personalization. Second, top ranks tend to be less personalized than bottom ranks.

4.2 Personalization Features

In the previous section, I observed personalization for real users on Google Search. Now I examine which user features Google Search uses to personalize results. Although I cannot possibly enumerate and test all possible features, I can investigate likely candidates. Table 3 lists the different demographic profiles that my experiments emulate during experiments.

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

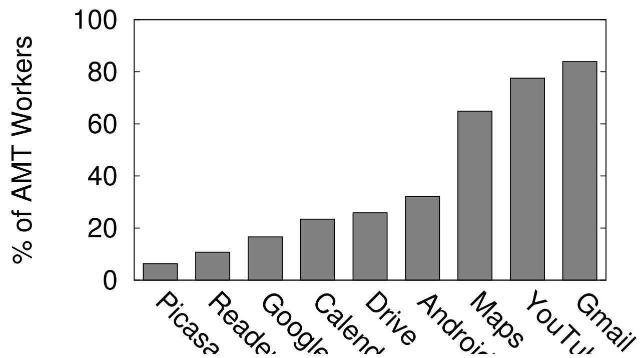


Figure 1: Usage of Google services by AMT workers.

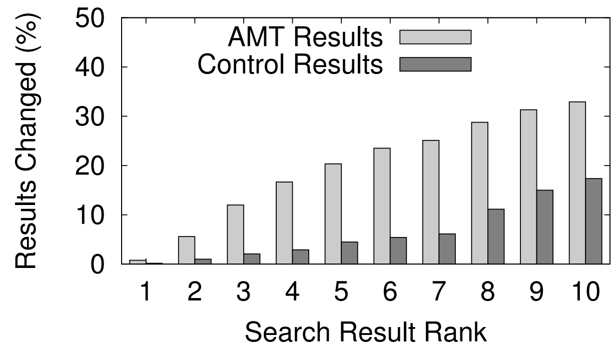


Figure 2: % of AMT and control results changed for each rank.

Implementation On start, each PhantomJS instance logs in to Google using a separate Google account, and begins issuing queries to Google Search. The script downloads the first page of search results for each query. The script waits 11 minutes in-between searches for subsequent queries.

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 Google cookies are recreated during the next invocation of the experiment when the script logs in to its assigned Google account. All of my experiments are designed to complete in ≈24 hours.

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

Google Accounts. Unless otherwise specified, each Google account I create has the same profile: 27 year old, female. The default User-Agent I use is Chrome 22 on Windows 7. As shown in Section 4.2.1, I do not observe any personalization of results based on these attributes.

I manually crafted each of my Google accounts to minimize the likelihood of Google automatically detecting them. Each account was given a unique name and profile image. I read all of the introductory emails in each account's Gmail inbox, and looked at any pending Google+ notifications. To the best of my knowledge, none of my accounts were banned or flagged by Google during my experiments.

Measuring Personalization When comparing the list of search results for test and control accounts, I use two metrics to capture two different ways personalization can happen. First, 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).

To measure reordering, 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 [8]) to make the test list identical to the control list.

4.2.1 Basic Features

I begin by focusing on features associated with a user's browser, their physical location, and their Google profile. For each experiment, I create $x + 1$ fresh Google accounts, where x equals the number of possible values of the feature I am testing in that experiment, plus one additional control account. For example, in the Gender experiment, I create 4 accounts: one "male," one "female," one "other," and one additional "female" as a control. I execute $x + 1$ instances of the PhantomJS script for each experiment, 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 seven days. For brevity I will only present results associated with one feature (geo-location) now but the analysis for all other features follows a very similar structure. Results for all the measured features can be found in my paper [17].

In this experiment I create 11 Google accounts and run my 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 3. Two accounts forward through the Massachusetts PlanetLab machine, as it is the control.

Figure 3 shows the results of my location tests. There is a clear difference between the control and all the other locations. The average Jaccard Index for non-control tests is 0.91, meaning that queries from different locations generally differ by one result. The difference between locations is even more pronounced when I consider result order: the average edit distance for non-control accounts is 2.12.

These results reveal that Google Search does personalize results based on the user's geolocation. One example of this personalization can be seen by comparing the MA and CA 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.

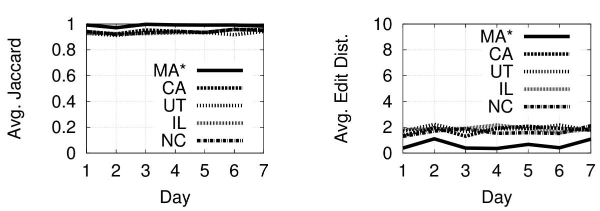
4.2.2 Historical Features

I now examine whether Google Search uses 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

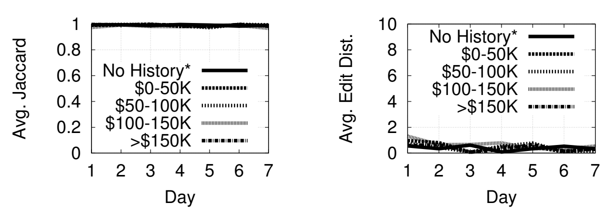
Category	**Feature**	**Tested Values**
Tracking	Cookies	Logged In, Logged Out, No Cookies
User-Agent	OS	Win. XP, Win. 7, OS X, Linux
	Browser	Chrome 22, Firefox 15, IE 6, IE 8, Safari 5
Geo-location	IP Address	MA, PA, IL, WA, CA, UT, NC, NY, OR, GA
Google Account	Gender	Male, Female, Other
	Age	15, 25,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

Table 2: User features evaluated for effects on search personalization.



(a) Jaccard Index (b) Edit Distance

Figure 3: Results for the Google Profile: Geo-location experiments.



(a) Jaccard Index (b) Edit Distance

Figure 4: Results for the search history: income level experiments.

and advertising firm. Quantcast publishes a list of top websites (similar to Alexa) that includes the *demographics* of visitors to sites [34], broken down into the 20 categories shown in Table 3. Quantcast assigns each website a score for each demographic, where scores >100 indicate 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. Thus, for each of my three experiments,I create 22 Google 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.

Search History First, I examine whether Google Search personalizes results based on search history. Each day, the 20 test accounts 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 shows the results of the search history test for four income demographics. The "No History" account does not search for demographic queries, and serves as the control. All accounts receive approximately the same search results, thus I do not observe personalization based on search history. This observation holds for all of the demographic categories I tested, and I omit the results for brevity.

Search-Result-Click History To examine whether Google Search 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 the result for `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, we do not observe personalization based on click history, and we omit the results for brevity.

Browsing History Finally, 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, we do not observe personalization. We omit these results for brevity.

5 Measuring Price Discrimination and Steering on E-commerce Web Sites

Online businesses have long used personalized recommendations as a way to boost sales. Retailers like Amazon and Target leverage the search and purchase histories of users to identify products that users may be interested in. In some cases, companies go to great lengths to obfuscate the fact that recommendations are personalized, because users sometimes find these practices to be creepy [11].

In several cases, e-commerce sites have been caught performing *price discrimination*: the practice of showing different prices to different people for the same item. Several years ago, Amazon was caught personalizing prices for frequent shoppers [1]. Similarly, e-commerce sites have been observed performing *price steering*: the practice of re-ordering search results to place expensive items towards the top of the page. An example of price steering is the travel Web site Orbitz, which was found to be promoting high-value hotels specifically to Apple users [23]. While price discrimination is not illegal [7], it is (arguably) an anti-consumer practice, and studies have shown that users are deeply against it [35].

In this project I adapt the methodology introduced in Section 4 to measure price discrimination and steering on a variety of e-commerce sites.

5.1 Experimental Overview

Implementation The first challenge I face when trying to measure personalization on e-commerce sites is dealing with the variety of sites that exist. Few of these sites offer programmatic APIs for collecting prices, and each uses substantially different HTML markup to implement their site. As a result, 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.

The second challenge is controlling for all sources of noise, just as it was when measuring personalization on Web search engines. Just as before in each experiment I include a *control* that is configured in an identical manner to one other treatment (i.e., I run one of the experimental treatments twice).

E-commerce Sites 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 evidence of price steering among such sites [23]. Of course, my methodology can be applied to other categories of e-commerce sites as well. I select 10 of the largest e-commerce retailers, according to the Top500 e-commerce database [42], for my study as well as five of the most popular web-based travel retailers [43]. For the travel retailers, I focus on searches for *hotels* and *rental cars*. I specifically do not include airline tickets, as airlines ticket pricing is done transparently through a set of Global Distribution Systems (GDSes) [6], and because commissions on airline tickets are significantly lower (providing less incentive for retailers to perform price discrimination and steering).

Searches For each e-commerce site, I select 20 searches to send to the 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 for 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).

5.2 Real-World Personalization

I begin by addressing my first question: *how widespread are price discrimination and steering on today's e-commerce web sites?*

Data collection For this experiment I posted 3 separate HITs on Amazon's Mechanical Turk with each HIT focusing on e-commerce, hotels, or rental cars. This time the users did not have to log in to their accounts own account as in the case of Google but by having the user run the searches within their browser,

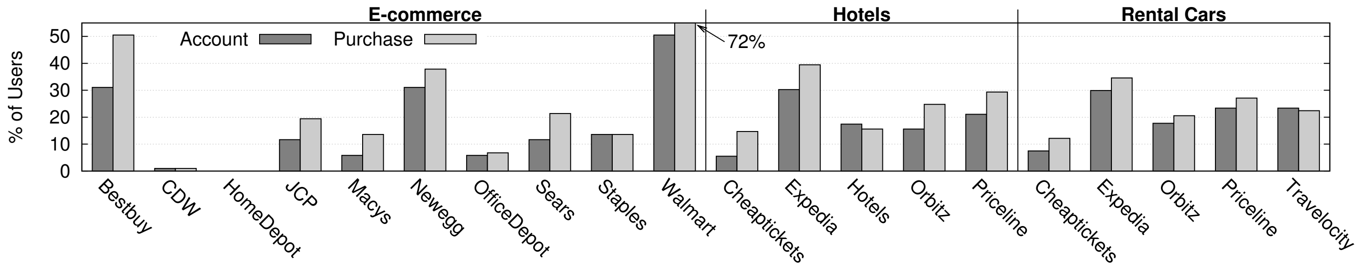


Figure 5: Previous usage (having an account and making a purchase) of different e-commerce sites by AMT users.

any cookies that the user's browser had previously been assigned would automatically be forwarded in my searches. This allows me to examine the results that the user would have received.

In total, I recruited 300 AMT users, with 100 in each of my experiments. In all three 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 5. I observe that many of my users have accounts and a purchase history on a large number of the sites I study.[3]

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. To measure price steering, I use three metrics: Jaccard Index measures the overlap between two different set of results. The Kendall's Tau rank correlation measures the reordering between the two lists. And finally, I use Normalized Discounted Cumulative Gain (nDCG). nDCG is a metric from the IR literature [19] for calculating how "close" a given list of search results is to an ideal ordering of results. nDCG of 1 means the observed search results are the same as the ideal results, while 0 means no useful results were returned. In this context the "ideal" list will be a list of products sorted from most- to least-expensive, the intuition being that the site would want to place the more expensive products first.

For each site, Figure 6 presents the average Jaccard index, Kendall's tau, 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 a number of interesting trends. *First*, I observe that Sears, Walmart, and Priceline all have a lower Jaccard index for AMT users relative to the control. This indicates that the AMT users are being returned different products at a higher rate than the control searches. 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.

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 7, 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 7, I plot the distribution of 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 **??**,

[3] Note 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".

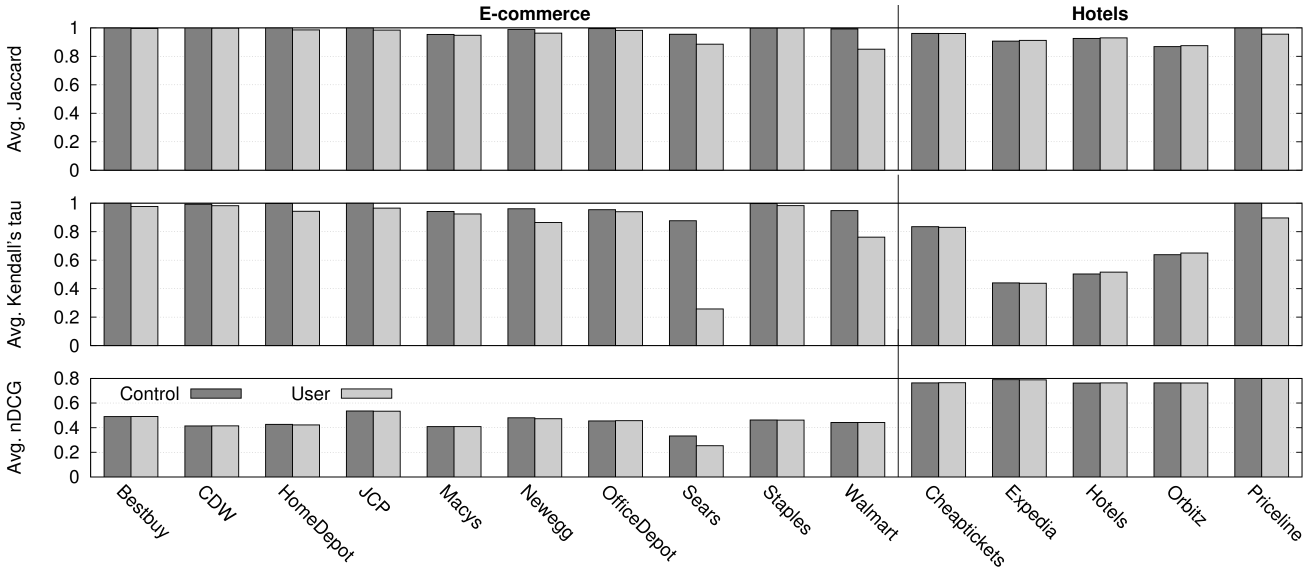


Figure 6: Average Jaccard index (top), Kendall's tau (middle), and nDCG (bottom) across all users and searches for each web site.

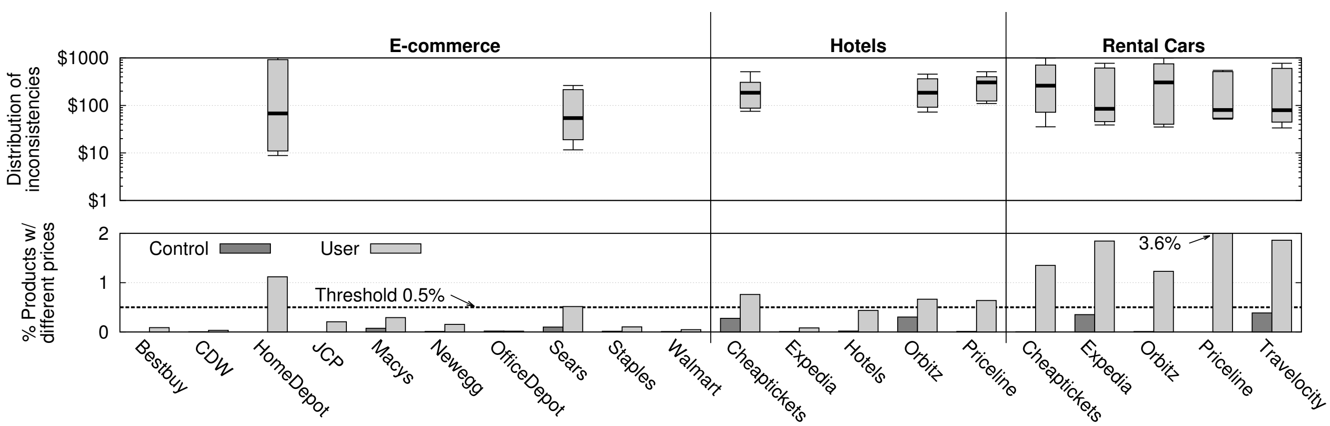


Figure 7: Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with at least 0.5% of products showing a difference (top), across all users and searches for each web site. In the top plot, I show the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).

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 my AMT user was returned a price of $633.

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.

5.3 Personalization Features

5.3.1 Static Features

Table 3 lists the five features that we evaluate in our 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 to an account before querying, and the final account clears its cookies after every HTTP request.

In two sets of experiments, we 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. We ran all of the cookie and User-Agent experiments on all 16 e-commerce sites.

Here I will highlight only the most interesting findings since the 16 companies were each tested for the 3 static features. In short I observed cases of sites altering results based on all 3 static features. Cheaptickets and Orbitz serve slightly different sets of results to users who are logged-in to an account, as compared to users who do not have an account or who do not store cookies. In practice, this means out of 25 total 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. Travelocity alters hotel search results for users who browse from iOS devices. Users browsing with Safari on iOS receive slightly different hotels, and in a much different order, than users browsing from Chrome on Android, Safari on OS X, or other desktop browsers. Interestingly the price discrimination is consistently in favor of iOS users and 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. Among the 10 retailers I ran experiments on, I only discovered evidence of personalization on Home Depot. Similar to my findings on Travelocity, Home 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. I observed both price that Home Depot is steering users on mobile browsers towards more expensive products and that they discriminate against Android users. Users of Android receive slightly higher prices.

5.3.2 Purchase History

The goal of the purchase history experiment 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 me 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. Similarly, all five travel sites allow rental cars to be reserved without up-front payment. These no-payment reservations allow me 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 priced and lowest priced

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 3: User features evaluated for effects on personalization.

results, respectively. The remaining two accounts searched for the same hotel/car and reserved the highest priced 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 before.

Out of the five measured travel sites Priceline shows the most interesting effects. 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 confirm that the reordering is not correlated with prices, i.e., all treatments receive pages with similarly priced items at the top.

6 Research plan

In Section 4 and 5, I introduced work that has already been completed and published. In this section I describe work I propose to do in the next 18 months to serve as part of my doctoral thesis.

6.1 Measuring Personalization on Bing

Since the methodology developed in Section 4 is not specific to Google, it gives me the opportunity to measure other popular search engines. Bing is the most popular search engine after Google (and the 11th most popular website according to Alexa over all [2]); thus, it is very important to understand how its personalization mechanisms work. This analysis is not only going to give us an understanding of how Bing alters the results that their users see, but it will also give me an opportunity to compare my findings across multiple search engines. I hope to be able to see the differences and commonalities in the way the biggest search engines use personal data to alter the content they serve.

Similar to Google, Microsoft provides user accounts for its various services (e.g., Windows Live, Outlook.com, Xbox LIVE, Skype). These various accounts have been consolidated into a single "Microsoft account" (previously called a Windows Live ID). When a user is signed-in to this account, Bing searches are tracked and saved. Just as described in 4.2.2, I plan to create fake accounts and run controlled experiments with them to measure the effect of each user feature. Since we know that Microsoft is also tracking users as they browse the Web through their large advertising networks (e.g., MSN advertising) I will be able to recreate the browsing history experiment. Finally, I aim collect real-world data on personalization as described in Section 4.1 by recruiting a group of AMT users.

6.2 Measuring Price Discrimination among FlyerTalk users

My results described in Section 5.2 indicate that user history is a primary driver of personalization algorithms on online purchasing sites. As shown in work by Ipeirotis [13], AMT users are mostly low-income users and thus probably do not have extensive histories with travel websites. I plan to conduct a similar experiment to the one described in Section 5.2 on a population that is likely to have an extensive history with traveling and reward programs. I will post the experiment on FlyerTalk [12], which is an Internet forum for discussion of airline frequent-flyer programs, hotel loyalty programs, and other issues related to travel. Users of the forum are an ideal population for for this experiment both because they travel frequently and manage their travels online but also because they are interested in the issue of price discrimination and thus are more likely to be helpful in the process of collecting data.

Users will run the experiment described in Section 5.2. I plan to extend the survey to include more detailed questions about the users' travel and booking habits or their price elasticity since these are known to be important for e-commerce websites [31]. Once I have data with the results the FlyerTalk users are returned, I can correlate the differences I see to their purchasing and traveling habits. Comparing results from the FlyerTalk population with those from AMT will give me interesting insights into the decisions websites make in the selection they offer to users based on their previous histories.

6.3 Hyper-local Personalization of Google

In my prior work, I identified IP address geo-location as a key feature used by search engines to personalize content [17]. However IP location is very coarse-grained. Fortunately, modern Web browsers (especially on mobile devices) offer APIs that allow websites to query a user's precise location via GPS. Spoofing GPS coordinates to the HTML5 Geolocation API will allow me to "fake" a user's precise location. Using this feature, I can collect very fine grained data on a grid all over the United States and compare the search results users get according this very precisely defined location. 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. One focus of this data collection could be politics, since existing literature has identified personalization of political content as one of the key concerns of Filter Bubbles. My prior measurements of Google Search show that political queries exhibit more personalization than any other category of query, suggesting that these fears may be valid. Using this technique and querying political terms, names of candidates or politically divisive topics, I could discover interesting political biases based on location.

The key advantage of politics as a substrate for studying personalization is that there are well-developed methodologies to study ideological/political tilt. At the individual level, there are reliable survey methods to capture left to right orientation, and at the collective level there are electoral data that allow the calculation of political tilt of areas down to a few thousand individuals. The first steps are to take a politically polarized state (e.g., Pennsylvania or Florida), divide it into a fine grid, and then execute political search queries from web browsers that are located within each section of the grid. This will allow me to "map" the personalization of Web content across a given state, giving me unprecedented insight into how location is used to tailor political and news-related content.

6.4 Plan for completion of the research

Table 4 presents my plan for completion of the research.

Timeline	Work	Progress
	Measurement Methodology	completed
	Personalization of Google Search	completed
	Personalization of E-commerce	completed
Spring 2015	Measuring Personalization on Bing	ongoing
Summer 2015	Flyertalk Survey	ongoing
Winter 2016	Hyperlocal Google Personalization	ongoing
May 2016	Thesis writting	
Summer 2016	Thesis defense	

Table 4: Plan for completion of my research

Thus, I plan to defend my thesis in the summer of 2016.

References

[1] Bezos Calls Amazon Experiment 'a Mistake'. Puget Sound Business Journal, 2000. `http://www.bizjournals.com/seattle/stories/2000/09/25/daily21.html`.

[2] Alexa Top 500 Global Sites. `http://www.alexa.com/topsites`.

[3] Leonardo Andrade and Mário J. Silva. Relevance Ranking For Geographic Ir. *ACM Workshop On Geographic Information Retrieval (GIR)*, Seattle, Washington, August 2006.

[4] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Satel, RenéWittand Greg Butler, and Adrian Tsang. An Approach To Controlling User Models And Personalization Effects In Recommender Systems. *International Conference on Intelligent User Interfaces (IUI)*, Santa Monica, California, March 2013.

[5] Judit Bar-Ilan, Kevin Keenoy, Eti Yaari, and Mark Levene. User Rankings Of Search Engine Results. *Journal of the American Society for Information Science and Technology*, 58(9):1254–1266, July 2007.

[6] Peter Belobaba, Amedeo Odoni, and Cynthia Barnhart. *The Global Airline Industry*. Wiley, 2009.

[7] Ryan Calo. Digital Market Manipulation. *The George Washington Law Review*, 82, 2014.

[8] Fred J. Damerau. A Technique For Computer Detection And Correction Of Spelling Errors. *Communications of the ACM*, 7(3):171–176, 1964.

[9] Abhinandan Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. *International World Wide Web Conference (WWW)*, Banff, Canada, May 2007.

[10] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A Large-scale Evaluation And Analysis Of Personalized Search Strategies. *International World Wide Web Conference (WWW)*, Banff, Canada, May 2007.

[11] Charles Duhigg. How Companies Learn Your Secrets. The New York Times, 2012. `http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html`.

[12] Flyertalk Forum. `http://www.flyertalk.com/forum/`.

[13] Ipeirotis, Panagiotis G. Demographics Of Mechanical Turk. NYU, Stern School of Business, Technical Report CeDER-10-01, 2010.

[14] Sharad Goel, Jake M. Hofman, and M Irmak Sirer. Who Does What On The Web: A Large-scale Study Of Browsing Behavior. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, Dublin, Ireland, June 2012.

[15] Google Zeitgeist. 2012. `http://www.googlezeitgeist.com`.

[16] Saikat Guha, Bin Cheng, and Paul Francis. Challenges In Measuring Online Advertising Systems. *ACM Internet Measurement Conference (IMC)*, Melbourne, Victoria, Australia, November 2010.

[17] Aniko Hannak, Piotr Sapiezyński, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring Personalization Of Web Search. *International World Wide Web Conference (WWW)*, Rio de Janeiro, Brazil, May 2013.

[18] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic Prediction Based On User's Browsing Behavior. *International World Wide Web Conference (WWW)*, Banff, Canada, May 2007.

[19] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-based Evaluation Of Ir Techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.

[20] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-bandit Approach To Personalized News Article Recommendation. *International World Wide Web Conference (WWW)*, Raleigh, North Carolina, April 2010.

[21] Fang Liu, Clement Yu, and Weiyi Meng. Personalized Web Search By Mapping User Queries To Categories. *ACM International Conference on Information and Knowledge Management (CIKM)*, McLean, Virginia, November 2002.

[22] Anirban Majumder and Nisheeth Shrivastava. Know Your Personalization: Learning Topic Level Personalization In Online Services. *International World Wide Web Conference (WWW)*, Rio de Janeiro, Brazil, May 2013.

[23] Dana Mattioli. On Orbitz, Mac Users Steered To Pricier Hotels. *The Wall Street Journal*, August 2012. `http://on.wsj.com/LwTnPH`.

[24] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting Price And Search Discrimination On The Internet. *Workshop on Hot Topics in Networks (HotNets)*, Seattle, Washington, October 2012.

[25] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Crowd-assisted Search For Price Discrimination In E-commerce: First Results. *International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, 2013.

[26] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In Google We Trust: Users' Decisions On Rank, Position, And Relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

[27] Ambar Pansari and Marissa Mayer. This Is A Test. This Is Only A Test. *Google Official Blog*, April 2006. `http://bit.ly/Ldbb0`.

[28] Eli Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Press, 2011.

[29] Phantomjs. 2013. `http://phantomjs.org`.

[30] James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. Personalized Search. *Communications of the ACM*, 45(9):50–55, September 2002.

[31] Ivan Png. *Managerial Economics*. Routledge, 2012.

[32] Alexander Pretschner and Susan Gauch. Ontology Based Personalized Search. *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Chicago, Illinois, November 1999.

[33] Feng Qiu and Junghoo Cho. Automatic Identification Of User Interest For Personalized Search. *International World Wide Web Conference (WWW)*, Edinburgh, Scotland, May 2006.

[34] Quantcast. Top Sites For The United States. 2012. `http://www.quantcast.com/top-sites`.

[35] Anita Ramasastry. Web Sites Change Prices Based On Customers' Habits. CNN, 2005. `http://edition.cnn.com/2005/LAW/06/24/ramasastry.website.prices/`.

[36] Selenium. 2013. `http://selenium.org`.

[37] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Implicit User Modeling For Personalized Search. *ACM International Conference on Information and Knowledge Management (CIKM)*, Bremen, Germany, November 2005.

[38] Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. Cubesvd: A Novel Approach To Personalized Web Search. *International World Wide Web Conference (WWW)*, Chiba, Japan, May 2005.

[39] Mingxuan Sun, Guy Lebanon, and Kevyn Collins-Thompson. Visualizing Differences In Web Search Algorithms Using The Expected Weighted Hoeffding Distance. *International World Wide Web Conference (WWW)*, Raleigh, North Carolina, April 2010.

[40] Bin Tan, Xuehua Shen, and ChengXiang Zhai. Mining Long-term Search History To Improve Search Accuracy. *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining (KDD)*, Philadelphia, Pennsylvania, August 2006.

[41] Jamie Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing Search Via Automated Analysis Of Interests And Activities. *Conference of the ACM Special Interest Group on Information Retrieval (SIGIR)*, Sheffield, United Kingdom, July 2005.

[42] Top 500 E-retailers. `http://www.top500guide.com/top-500/`.

[43] Top Booking Sites. `http://skift.com/2013/11/11/top-25-online-booking-sites-in-travel/`.

[44] Liwen Vaughan. New Measurements For Search Engine Evaluation Proposed And Tested. *Information Processing and Management*, 40(4):677–691, May 2004.

[45] Webmd Year In Health. 2011. `http://on.webmd.com/eBPFxH`.

[46] Ingmar Weber and Alejandro Jaimes. Who Uses Web Search For What? And How? *ACM International Conference of Web Search and Data Mining (WSDM)*, Hong Kong, China, February 2011.

[47] Ryen W. White. Beliefs And Biases In Web Search. *Conference of the ACM Special Interest Group on Information Retrieval (SIGIR)*, 2013.

[48] Craig E. Wills and Can Tatar. Understanding What They Do With What They Know. *Workshop on Privacy in the Electronic Society (WPES)*, 2012.

[49] Xinyu Xing, Wei Meng, Dan Doozan, Alex C. Snoeren, Nick Feamster, and Wenke Lee. Take This Personally: Pollution Attacks On Personalized Services. *USENIX Security Symposium (USENIX Security)*, Washington, D.C., August 2013.

[50] Yabo Xu, Benyu Zhang, Zheng Chen, and Ke Wang. Privacy-enhancing Personalized Web Search. *International World Wide Web Conference (WWW)*, Banff, Canada, May 2007.

[51] Xing Yi, Hema Raghavan, and Chris Leggetter. Discovering Users' Specific Geo Intention In Web Search. *International World Wide Web Conference (WWW)*, Madrid, Spain, April 2009.

[52] Bo Yu and Guoray Cai. A Query-aware Document Ranking Method For Geographic Information Retrieval. *ACM Workshop On Geographic Information Retrieval (GIR)*, Lisbon, Portugal, November 2007.