Measuring Personalization of Web Search

Aniko Hannak
Northeastern University
ancsaaa@ccs.neu.edu

Piotr Sapieżyński
Technical University of Denmark
sapiezynski@gmail.com

Arash Molavi Kakhki
Northeastern University
arash@ccs.neu.edu

Balachander Krishnamurthy
AT&T Labs–Research
bala@research.att.com

David Lazer
Northeastern University
d.lazer@neu.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

Web search is an integral part of our daily lives. Recently, there has been a trend of personalization in Web search, where different users receive different results for the same search query. The increasing personalization is leading to concerns about *Filter Bubble* effects, where certain users are simply unable to access information that the search engines' algorithm decides is irrelevant. Despite these concerns, there has been little quantification of the extent of personalization in Web search today, or the user attributes that cause it.

In light of this situation, we make three contributions. First, we develop a methodology for measuring personalization in Web search results. While conceptually simple, there are numerous details that our methodology must handle in order to accurately attribute differences in search results to personalization. Second, we apply our methodology to 200 users on Google Web Search; we find that, on average, 11.7% of results show differences due to personalization, but that this varies widely by search query and by result ranking. Third, we investigate the causes of personalization on Google Web Search. Surprisingly, we only find measurable personalization as a result of searching with a logged in account and the IP address of the searching user. Our results are a first step towards understanding the extent and effects of personalization on Web search engines today.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

Keywords

Personalization; Web search; Measurement

1. INTRODUCTION

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 [52]. 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 [18]), and foreign affairs (e.g., Google's ongoing conflict with Chinese Web censors [46]).

Recently, major search engines have implemented *personalization*, where different users searching for the same terms may observe different results [1, 34]. For example, users searching for "pizza" in New York and in Boston may receive different, but locally relevant restaurant results. Personalization provides obvious benefits to users, including disambiguation and retrieval of locally relevant results.

However, personalization of Web search has led to growing concerns over the *Filter Bubble* effect [9], where users are only given results that the personalization algorithm thinks they want (while other, potentially important, results remain hidden). 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 [26]. 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 [25].

Concerns about the Filter Bubble effects are now appearing in the popular press [35, 38], driving growth in the popularity of alternative search engines that do not personalize results (e.g., `duckduckgo.com`). Unfortunately, to date, there has been little scientific quantification of the basis and extent of search personalization in practice.

In this paper, we make three contributions towards remedying this situation. *First*, we develop a methodology for measuring personalization in Web search results. 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. We develop a methodology that is able to control for these phenomena and create a command line-based implementation that we make available to the research community.

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

Second, we use this methodology to measure the extent of personalization on Google Web Search today. We recruit 200 users with active Google accounts from Amazon's Mechanical Turk to run a list of Web searches, and we measure the differences in search results that they are given. We control for differences in time, location, distributed infrastructure, and noise, allowing us to attribute any differences observed to personalization. Although our results are only a lower bound, we observe significant personalization: on average, 11.7% of search results show differences due to personalization, with higher probabilities for results towards the bottom. We see the highest personalization for queries related to political issues, news, and local businesses.

Third, we investigate the causes of personalization, covering user-provided profile information, Web browser and operating system choice, search history, search-result-click history, and browsing history. We create numerous Google accounts and assign each a set of unique behaviors. We develop a standard list of 120 search queries that cover a variety of topics pulled from Google Zeitgeist [14] and WebMD [48]. We then measure the differences in results that are returned for this list of searches. Overall, we find that while the level of personalization is significant, there are very few user properties that lead to personalization. Contrary to our expectations, we find that only being logged in to Google and the location (IP address) of the user's machine result in measurable personalization. All other attributes do not result in level of personalization beyond the baseline noise level.

We view our work as a first step towards measuring and addressing the increasing level of personalization on the Web today. All Web search engines periodically introduce new techniques, thus any particular findings about the level and causes of personalization may only be accurate for a small time window. However, our methodology can be applied periodically to determine if search services have changed. Additionally, although we focus on Google Search in this paper, our methodology naturally generalizes to other search services as well (e.g., Bing, Google News).

2. BACKGROUND

We now provide background on Google Search and overview the terminology used in the remainder of the paper.

2.1 A Brief History of Google

Personalization on Google Search. Google first introduced "Personalized Search" in 2004 [17], and merged this product into Google Search in 2005 [1]. In 2009, Google began personalizing search results for all users, even those without Google accounts [15]. Recently, Google started including personalized content from the Google+ social network into search results [33]. For example, users may see Web pages which were shared or "+1'd" by people in their Google+ circles alongside normal Google search results.

There is very little concrete information about how Google personalizes search results. A 2011 post on the official Google blog states that Google Search personalizes results based on the user's language, geolocation, history of search queries, and their Google+ social connections [32]. However, the specific uses of search history data are unclear: the blog post suggests that the temporal order of searches matters, as well as whether users click on results. Similarly, the specific uses of social data from Google+ are unknown.

Figure 1: Example page of Google Search results.

Google Accounts. As the number and scope of the services provided by Google grew, Google began unifying their account management architecture. Today, Google Accounts are the single point of login for all Google services. Once a user logs in to one of these services, they are effectively logged in to all services. A tracking cookie enables all of Google's services to uniquely identify each logged in user. As of May 2012, Google's privacy policy allows between-service information sharing across all Google services [45].

Advertising and User Tracking. Google is capable of tracking users as they browse the Web due to their large advertising networks. Roesner et al. provide an excellent overview of how Google can use cookies from DoubleClick and Google Analytics, as well as widgets from YouTube and Google+ to track users' browsing habits [31].

2.2 Terminology

In this study, we use a specific set of terms when referring to Google Search. Each *query* to Google Search is composed of one or more keywords. In response to a query, Google Search returns a *page* of *results*. Figure 1 shows a truncated example page of Google Search results for the query "coughs." Each page contains ≈10 results (in some cases there may be more or less). We highlight three results with red boxes in Figure 1. Most results contain ≥ 1 links. In this study, we only focus on the *primary link* in each result, which we highlight with red arrows in Figure 1.

In most cases, the primary link is *organic*, i.e., it points to a third-party website. The WebMD result in Figure 1 falls into this category. However, the primary link may point to another Google service. For example, in Figure 1 the "News for coughs" link directs to Google News.

A few services inserted in Google Search results do not include a primary link. The Related Searches result in Figure 1 falls into this category. Another example is Google Dictionary, which displays the definition of a search keyword. In these cases, we treat the primary link of the result as a descriptive, static string, e.g., "Related" or "Dictionary."

3. METHODOLOGY

In this section, we describe our experimental methodology. First, we give the high-level intuition that guides the design of our experiments, and identify sources of noise that can

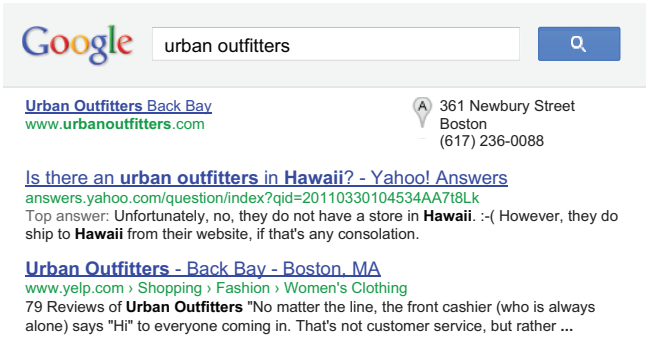


Figure 2: Example of result carry-over, searching for "hawaii" then searching for "urban outfitters."

lead to errors in data collection. Second, we describe the implementation of our experiments. Lastly, we introduce the queries we use to test for personalization.

3.1 Experiment Design

Our study seeks to answer two broad questions. First, *what user features influence Google's search personalization algorithms?* This question is fundamental: outside of Google, nobody knows the specifics of how personalization works. 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.

In this paper, we focus on measuring Google Search, as it is the most popular search engine. However, our methodology is Web service agnostic, and could be repeated on other search engines like Bing or Google News Search.

At a high-level, our methodology is to execute carefully controlled queries on Google Search to identify what user features trigger personalization. Each experiment follows a similar pattern: first, create x Google accounts 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 we run some experimental treatments *without* Google accounts (e.g., to simulate users without Google accounts).

Sources of Noise. Despite the simplicity of the high-level experimental design, there are several sources of noise that can cause identical queries to return different results.

- **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.
- **Distributed Infrastructure:** Large-scale Web search services are spread across geographically diverse datacenters. Our tests have shown that different datacenters 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:** Search engines use the user's IP address to provide localized results [51]. Thus, searches from different subnets may receive different results.
- **A/B Testing:** Web search services sometimes conduct A/B testing [24], where certain results are altered to measure whether users click on them more often. Thus, there may be a certain level of noise independent of all other factors.

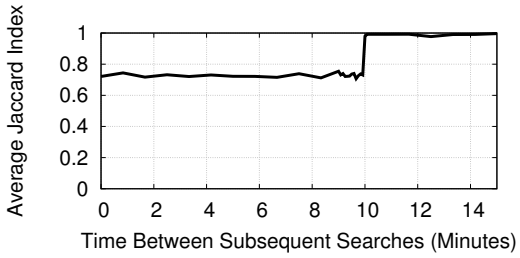


Figure 3: Overlap of results when searching for "test" followed by "touring" compared to just "touring" for different waiting periods.

The Carry-Over Effect. One particular source of noise comes from the influence of one search on subsequent searches. In other words, 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. We term this phenomenon the *carry-over effect*. Prior research on user intent while searching has shown that sequential queries from a user are useful for refining search results [5, 40], so it is not surprising that Google Search leverages this feature.

An example of carry-over is shown in Figure 2. In this test, we search for "hawaii" and then immediately search for "urban outfitters" (a clothing retailer). We 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, we conduct a simple experiment. We first pick different pairs of queries (e.g., "gay marriage" and "obama"). We then start two different browser instances: in one we search for the first query, wait, and then for the second query, while in the other we search only for the second query. We repeat this experiment with different wait times, and re-run the experiment 50 times with different query pairs. Finally, we compare the results returned in the two different browser instances for the second term.

The results of this experiment are shown in Figure 3 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 experiments in the following sections, we wait at least 11 minutes between subsequent searches in order to avoid any carry-over effects. In our testing, we observed carry-over for both logged in users and users without Google accounts.

Controlling Against Noise. In order to mitigate measurements errors due to these factors, we perform a num-

ber of steps (some borrowed from [10]): *First*, all of our queries are executed by the normal Google Search webpage, rather than Google's Search API. It has been shown that search engine APIs sometimes return different results than the standard webpage [4]. *Second*, all of our machines execute searches for the same query at the same time (i.e., in lock-step). This eliminates differences is query results due to temporal effects. This also means that each of our Google accounts has exactly the same search history at the same time. *Third*, we use static DNS entries to direct all of our query traffic to a specific Google IP address. This eliminates errors arising from differences between datacenters. *Fourth*, we wait 11 minutes in-between subsequent queries to avoid carry-over. As shown in Figure 3, an 11 minute wait is sufficient to avoid the majority of instances of carry-over. *Fifth*, unless otherwise stated, we send all of the search queries for a given experiment from the same /24 subnet. Doing so ensures that any geolocation would affect all results equally.

Sixth, we include a *control account* in each of our experiments. The control account is configured in an identical manner to one other account in the given experiment (essentially, we run one of the experimental treatments twice). By comparing the results received by the control and its duplicate, we 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.

3.2 Implementation

Our experiments are implemented using custom scripts for PhantomJS [28]. We chose PhantomJS because it is a full implementation of the WebKit browser, meaning that 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 [42]).

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 our experiments are designed to complete in ≈24 hours.

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

All of our experiments were conducted in fall of 2012. Although our results are representative for this time period, they may not hold in the future, since Google is constantly tweaking their personalization algorithms.

Google Accounts. Unless otherwise specified, each Google account we create has the same profile: 27 year old, female. The default User-Agent we use is Chrome 22 on Windows 7. As shown in Section 5.2, we do not observe any personalization of results based on these attributes.

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

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

3.3 Search Queries

In our experiments, each Google account searches for a specific list of queries. It is fundamental to our research that we select a list of queries that has both breadth and impact. Breadth is vital, since we do not know which queries Google personalizes results for. However, given that we cannot test all possible queries, it is important that we select queries that real people are likely to use.

As shown in Table 1, we use 120 queries divided equally over 12 categories in our experiments. These queries were chosen from the 2011 Google Zeitgeist [14], and WebMD [48]. Google Zeitgeist is published annually by Google, and highlights the most popular search queries from the previous calendar year. We 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 [26]. 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 [21].

One critical area that is not covered by Google Zeitgeist is health-related queries. To fill this gap, we chose ten random queries from WebMD's list of popular health topics [48].

4. REAL-WORLD PERSONALIZATION

We 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; we therefore conducted a simple user study.

4.1 Collecting Real-World Data

We posted a task on Amazon's Mechanical Turk (AMT), explaining our study and offering each user $2.00 to participate. Participants were required to 1) be in the United

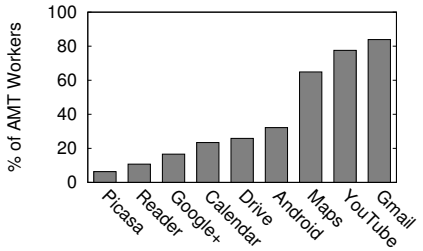


Figure 4: Usage of Google services by AMT workers.

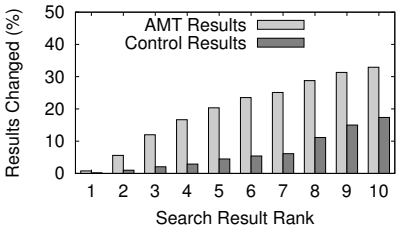


Figure 5: % of AMT and control results changed at each rank.

Most Personalized	Least Personalized
gap	what is gout
hollister	dance with dragons
hgtv	what is lupus
boomerang	gila monster facts
home depot	what is gluten
greece	ipad 2
pottery barn	cheri daniels
human rights	psoriatic arthritis
h2o	keurig coffee maker
nike	maytag refrigerator

Table 2: Top 10 most/least personalized queries.

States, 2) have a Google account, and 3) be logged in to Google during the study.[1] Users who accepted the 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 that automatically performed 80 Google searches. 50 of the queries were randomly chosen from the categories in Table 1, while 30 were chosen by us.

The HTTP proxy serves several functions. *First*, the proxy records Google Search's HTML responses to the users' queries. *Second*, each time the proxy observes a user making a query, it executes two PhantomJS scripts. Each script logs in to Google and executes the same exact query as the user. The results served to the scripts act as the control, allowing us to compare results from a real user (who Google has collected extensive data on) to fresh accounts (that have minimal Google 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 Google Search traffic to a hard-coded Google IP address.

Although the proxy is necessary to control for noise, there is a caveat to this technique. 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-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. Thus, our results from AMT users should be viewed as a lower bound on possible personalization.

AMT Worker Demographics. In total, we recruited 200 AMT workers, each of whom answered a brief demographic survey. Our 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 4 shows the usage of Google services by our participants: 84% are Gmail users, followed by 76% that use Google Maps. 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 services.

4.2 Results

We now pose the question: *how often do real users receive personalized search results?* To answer this question, we compare the results received by AMT users and the corresponding control accounts. Figure 5 shows the percentage of results that differ at each rank (i.e., result 1, result 2, *etc.*) when we 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 5. First, we 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.

To better understand how personalization varies across queries, we list the top 10 most and least personalized queries in Table 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 2, the most personalized queries tend to be related to companies and politics (e.g., "greece", "human rights" or "home depot"). Digging into the individual results, we observe a great deal of personalization based on location. Even though all of the AMT users' requests went through our 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 different store locations. In contrast, the least personalized results in Table 2 tend to be factual and health related queries.

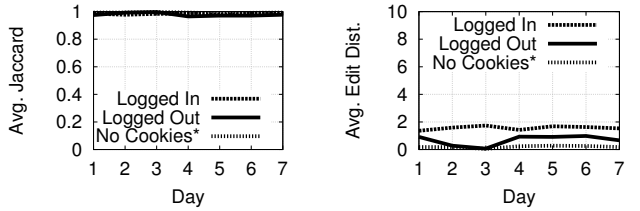
5. PERSONALIZATION FEATURES

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

5.1 Measuring Personalization

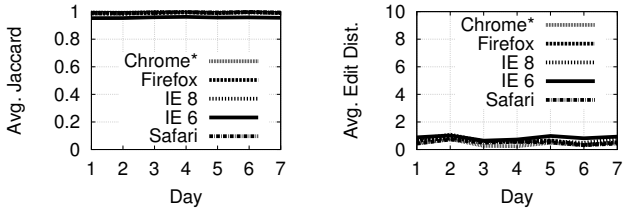
When comparing the list of search results for test and control accounts, we use two metrics to measure personalization. First, we use Jaccard Index, which views the result

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



(a) Jaccard Index (b) Edit Distance

Figure 6: Results for the cookie tracking experiments.



(a) Jaccard Index (b) Edit Distance

Figure 7: Results for the browser experiments.

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, we use edit distance. To calculate edit distance, we compute the number of list elements that must be inserted, deleted, substituted, or swapped (i.e., the Damerau-Levenshtein distance [7]) to make the test list identical to the control list. For example, if the test account receives the result list [`a.com`, `b.com`, `c.com`] and the control receives the list [`c.com`, `b.com`] for the same query, then the edit distance is 2 (one insertion and one swap).

5.2 Basic Features

We begin our experiments by focusing on features associated with a user's browser, their physical location, and their Google profile. For each experiment, we create $x + 1$ fresh Google accounts, where x equals the number of possible values of the feature we are testing in that experiment, plus one additional control account. For example, in the Gender experiment, we create 4 accounts: one "male," one "female," one "other," and one additional "female" as a control. We execute $x+1$ instances of our PhantomJS script for each experiment, and forward the traffic to $x + 1$ unique endpoints via SSH tunnels. Each account searches for all 120 of our queries, and we repeat this process daily for seven days.

Basic Cookie Tracking. In this experiment, the goal is to compare the search results for users who are logged in to a Google account, not logged in to Google, and who do not support cookies at all. Google is able to track the logged in and logged out users, since Google Search places tracking cookies on all users, even if they do not have a Google account. The user who does not support cookies receives a new tracking cookie after every request to Google, and we confirm that the identifiers in these cookies are unique on every request. However, it is unknown whether Google is 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, we use four instances of PhantomJS. The first two completely clear their cookies after every request. The third account logs in to Google and persists cookies normally. The fourth account does not log in to Google, and also persists cookies normally.

Figure 6(a) shows the average Jaccard Index for each account type (logged in/logged out/no cookies) across all search queries when compared to the control (no cookies). In all of our figures, we place a * on the legend entry that corresponds to the control test, i.e., two accounts that have identical features. We see from Figure 6(a) 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.

To examine how the order of results changes, we plot the average edit distance between each account type versus the control in Figure 6(b). We 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 Figure 6 give the first glimpse of how Google alters search results for different types of users.

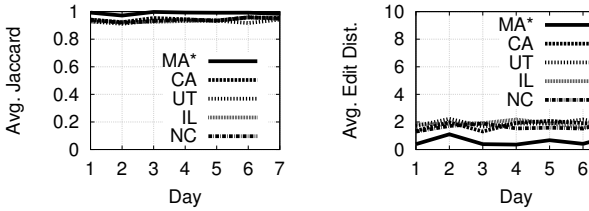
Category	Feature	Tested Values
Tracking	Cookies	Logged In, Logged Out, No Cookies
User-Agent	OS	Win. XP, Win. 7, OS X, Linux
	Browser	Chrome 22, Firefox 15, IE 6, IE 8, Safari 5
Geo-location	IP Address	MA, PA, IL, WA, CA, UT, NC, NY, OR, GA
Google Account	Gender	Male, Female, Other
	Age	15, 25, 35, 45, 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 3: User features evaluated for effects on search personalization.

Browser User-Agent. Next, we examine whether the user's choice of browser or Operating System (OS) can impact search results. To test this, we created 11 Google accounts and assigned each one a different "User-Agent" string. As shown in Table 3, we encoded user-agents for 5 browsers and 4 OSs. Chrome 22 and Windows 7 serve as the controls.

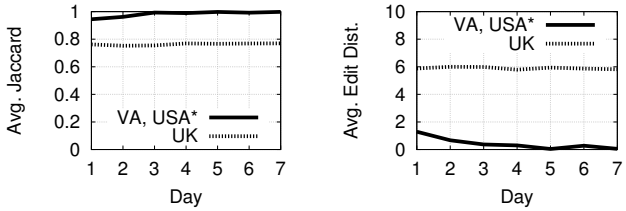
Figure 7 shows the results for our browser experiments. Unlike the cookie tracking experiment, there is no clear differentiation between the different browsers and the control experiment. The results for different OSs are similar, and we omit them for brevity. Thus, we do not observe search personalization based on user-agent strings.

IP address Geolocation. Next, we investigate whether Google Search personalizes results based on users' physical location. To examine this, we create 11 Google accounts and run our test suite while forwarding the traffic



(a) Jaccard Index (b) Edit Distance

Figure 8: Results for the geolocation experiments.



(a) Jaccard Index (b) Edit Distance

Figure 9: Results for inferred location experiments.

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 8 shows the results of our 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 we 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.

Inferred Geolocation. During our experiments, we 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.

Figure 9 shows some of the results for this anomaly. Although we could not determine why Google Search believes the 9 machines are in the UK (we believe it is due to an incorrect IP address geolocation database), we did confirm that this effect is independent of the Google account. As a result, we did not use EC2 machines as SSH tunnel endpoints for any of the results in this paper.

Google Account Attributes. In our next pair of tests, we examine whether Google Search uses demographic information from users' Google accounts 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.

To test this hypothesis, we created Google accounts with specific demographic qualities. As shown in Table 3, we created "female," "male," and "other" accounts (these are the 3 choices Google gives during account sign-up), as well as accounts with ages 15 to 65, in increments of 10 years. The control account in the gender tests is female, while the control in the age tests is 15.

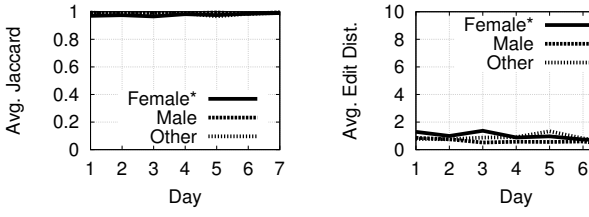
The results for the gender test are presented in Figure 10. We do not observe personalization based on gender in our experiments. Similarly, we do not observe personalization based on profile age, and we omit the results for brevity.

5.3 Historical Features

We now examine whether Google Search uses an account's history of activity to personalize results. We consider three types of historical actions: prior searches, prior searches where the user clicks a result, and Web browsing history.

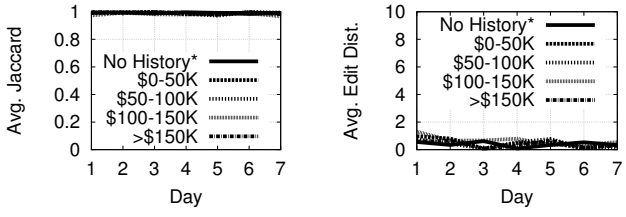
To create a plausible series of actions for different accounts, we use data from Quantcast, a Web analytics and advertising firm. Quantcast publishes a list of top websites (similar to Alexa) that includes the *demographics* of visitors to sites [30], broken down into the 20 categories shown in Table 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.

We use the Quantcast data to drive our historical experiments. In essence, our goal is to have different accounts "act" like a member of each of Quantcast's demographic groups. Thus, for each of our three experiments, we create 22 Google



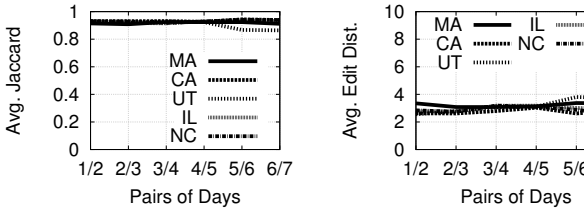
(a) Jaccard Index (b) Edit Distance

Figure 10: Results for the Google Profile: Gender experiments.



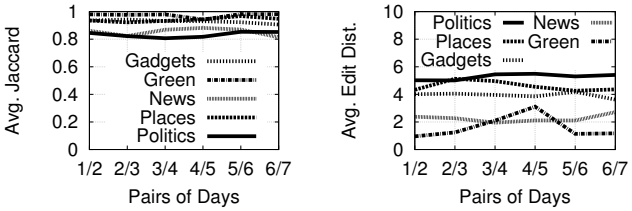
(a) Jaccard Index (b) Edit Distance

Figure 11: Results for the search history: income level experiments.



(a) Jaccard Index (b) Edit Distance

Figure 12: Day-to-day consistency of results for the geolocation experiments.



(a) Jaccard Index (b) Edit Distance

Figure 13: Day-to-day consistency within search query categories for the geolocation test.

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, we 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 11 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 we do not observe personalization based on search history. This observation holds for all of the demographic categories we tested, and we omit the results for brevity.

Search-Result-Click History. Next, we examine whether Google Search personalizes results based on the search results that a user has clicked on. We 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, we 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. We 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.

Discussion. We were surprised that the history-driven tests did not reveal personalization on Google Search. 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 Google Search. As future work, we plan on conducting longer lasting history-driven experiments to confirm our findings.

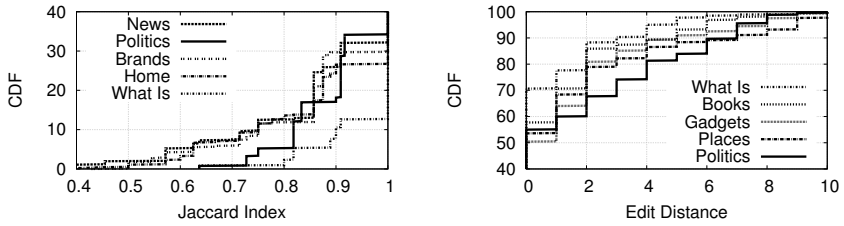
6. QUANTIFYING PERSONALIZATION

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.

6.1 Temporal Dynamics

In this section, we examine the temporal dynamics of results from Google Search to understand how much results from Google Search change day-to-day, and whether personalized results are more or less volatile than non-personalized search results. To measure the dynamics of Google Search over time, we compute the Jaccard Index and edit distance for search results from subsequent days. Figure 12 shows the day-to-day dynamics for our geolocation experiment. The x-axis shows which two days of search results are being compared, and each line corresponds to a particular test account.

Figure 12 reveals three facts about Google Search. First, the lines in Figures 12 are roughly horizontal, indicating that the rate of change in the search index is constant. Second, we observe that there is more reordering over time than new results: average Jaccard Index is 0.9, while average edit distance is 3. 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.



(a) Jaccard Index (b) Edit Distance

Figure 14: Differences in search results for five query categories.

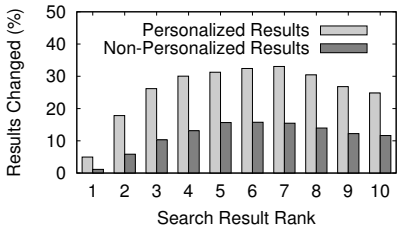


Figure 15: The percentage of results changed at each rank.

Dynamics of Query Categories. We now examine the temporal dynamics of results across different categories of queries. As shown in Table 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 edit distance between search results from subsequent days. However, instead of grouping by experiment, we now group by query category. Figure 13 shows the day-to-day dynamics for query categories during our geolocation experiment. Although we have 12 categories in total, Figure 13 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 geolocation test, and we omit them for brevity.

Figure 13 reveals that the search results for different query categories change at different rates day-to-day. Figure 13(a) shows that there are more new results per day for "politics" and "news" queries. Similarly, Figure 13(b) shows that queries for "politics," "news," and "places" all exhibit above average reordering each day. This reflects how quickly information in these categories changes on the Web. In the case of "places," the reordering is due to location specific news items that fluctuate daily. In contrast, search queries for factual categories like "what is" and "green" (environmentally friendly topics) are less volatile over time.

6.2 Personalization of Query Categories

We now examine the relationship between different categories of search queries and personalization. In Section 5, we demonstrate that Google Search does 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 edit distance for each category in Figure 14. 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 close to 1, edit distance close to 0), and the 4 least stable categories.

Figure 14 demonstrates that Google Search personalizes results for some query categories more than others. For example, 82% of results for "what is" queries are identical, while only 43% of results for "gadgets" are identical. Overall, "politics" is the most personalized query category, followed by "places" and "gadgets." CDFs calculated over other days of search results demonstrate nearly identical results.

6.3 Personalization and Result Ranking

In this section, we focus on the volatility of results from Google Search 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" [3, 12, 13, 20]. 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 15, 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 logged in/logged out and geolocation experiments; all others are non-personalized.

Figure 15 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 every rank, there are more than twice as many changes on personalized pages. Second, Figure 15 shows that the volatility at each rank is not uniform. Rank 1 exhibits the least volatility, and the volatility increases until it peaks at 33% in rank 7. This indicates that Google Search is more conservative about altering results at top ranks.

Given the extreme importance placed on rank 1 in Google Search, we now delve deeper into the 5% of cases where the result at rank 1 changes during personalized searches. In

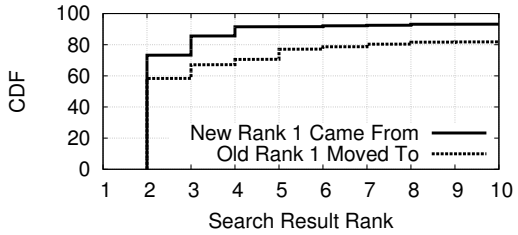


Figure 16: Movement of results to and from rank 1 for personalized searches.

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 16 plots the results of this test. In the vast majority of cases, the rank 1 and 2 results switch places: 73% of new rank 1 results originate from rank 2, and 58% of old rank 1 results move to rank 2. Overall, 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. However, neither CDF sums 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, with 18% of rank 1 results getting evicted completely from the first page of results.

7. RELATED WORK

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 [2, 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 [37]. Although this technique is very promising, it does not scale to the size of our experiments.

Personalization. Personalized search has been extensively studied in the literature [8, 19, 23, 27, 29, 36, 39, 43]. Dou et al. provide a comprehensive overview of techniques for personalizing search [6]. 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.

Other features besides click history have been used to power personalized search. Three studies leverage geographic location to personalize search [41,50,51]. Two studies have shown that user demographics can be reliably inferred from browsing histories, which can be useful for personalizing content [11,16]. To our knowledge, only one study has investigated privacy-preserving personalized search [49]. Given growing concerns about the Filter Bubble effects, this area seems promising for future research.

Several studies have looked at personalization on systems other than search. Two studies have examined personalization of targeted ads on the Web [10,47]. One study examines discriminatory pricing on e-commerce sites, which is essentially personalization of prices [22].

8. 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 paper, we take the first steps towards addressing this situation by introducing a methodology for measuring personalization on Web search engines. Our methodology controls for numerous sources of noise, allowing us to accurately measure the extent of personalization. We applied our methodology to real Google accounts recruited from AMT and observe that 11.7% of search results show differences due to personalization. Using artificially created accounts, we observe that measurable personalization is caused by 1) being logged in to Google and 2) making requests from different geographic areas.

However, much work remains to be done: we view our results as a first step in providing transparency for users of Web search and other Web-based services. In the paragraphs below, we discuss a few of the issues brought up by our work, as well as promising directions for future research.

Scope. In this paper, we focus on queries to US version of the Google Web Search. All queries are in English, and are drawn from topics that are primarily of interest to US residents. We leave the examination of Google sites in other countries and other languages to future work.

Incompleteness. As a result of our methodology, we are only able to identify positive instances of personalization; we cannot claim the absence of personalization, as we may not have considered other dimensions along which personalization could occur. However, the dimensions that we 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, we aim to continue this work by running our data collection for a longer time, looking at additional categories of Web searches, examining searches from mobile devices, and looking at other user behaviors (e.g., using services like Gmail, Google+, and Google Maps). We also plan on examining the impact of mechanisms that may disable personalization (e.g., opting-out of personalization on Google Search, and enabling Do-Not-Track headers).

Generality. The methodology that we develop is not specific to Google Web Search. The sources of noise that we control for are present in other search engines (e.g., Bing, Google News Search) as well as other Web-based services (e.g., Twitter search, Yelp recommendations, *etc.*). We plan on applying our methodology to these and other search services to quantify personalization of different types.

Impact. In this paper, we focused on quantifying literal differences in search results, e.g., `a.com` is different from `b.com`. However, we do not address the issue of semantic differences, i.e., do `a.com` and `b.com` contain different information content? If so, what is the impact of these differences? While semantic differences and impact are challenging to quantify, we plan to explore natural language processing and user studies as a first step.

Open Source. We make all of the crawling and parsing code, as well as the Google Web Search data from Section 5, available to the research community at

`http://personalization.ccs.neu.edu/`

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported by NSF grants IIS-0964465 and CNS-1054233, and an Amazon Web Services in Education Grant.

9. REFERENCES

[1] Personalized Search Graduates from Google Labs. *News From Google Blog*, 2005. http://bit.ly/Tndpgf.

[2] J. Bar-Ilan, K. Keenoy, E. Yaari, and M. Levene. User rankings of search engine results. *J. Am. Soc. Inf. Sci. Technol.*, 58(9), 2007.

[3] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. *CHI*, 2007.

[4] F. Crown and M. L. Nelson. Agreeing to Disagree: Search Engines and Their Public Interfaces. *JCDL*, 2007.

[5] Z. Cheng, B. Gao, and T.-Y. Liu. Actively Predicting Diverse Search Intent from User Browsing Behaviors. *WWW*, 2010.

[6] Z. Dou, R. Song, and J.-R. Wen. A Large-scale Evaluation and Analysis of Personalized Search Strategies. *WWW*, 2007.

[7] F. J. Damerau. A technique for computer detection and correction of spelling errors. *CACM*, 7(3), 1964.

[8] J. T. S. T. Dumais and E. Horvitz. Personalizing search via automated analysis of interests and activities. *SIGIR*, 2005.

[9] H. Green. Breaking Out of Your Internet Filter Bubble. *Forbes*, 2011. http://onforb.es/oYWbDf.

[10] S. Guha, B. Cheng, and P. Francis. Challenges in Measuring Online Advertising Systems. *IMC*, 2010.

[11] S. Goel, J. M. Hofman, and M. I. Sirer. Who Does What on the Web: A Large-scale Study of Browsing Behavior. *ICWSM*, 2012.

[12] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. *CHI*, 2007.

[13] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. *SIGIR*, 2004.

[14] Google Zeitgeist. http://www.googlezeitgeist.com.

[15] B. Horling and M. Kulick. Personalized Search for Everyone. *Google Official Blog*, 2009. http://bit.ly/71RcmJ.

[16] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic Prediction Based on User's Browsing Behavior. *WWW*, 2007.

[17] M. Hines. Google Takes Searching Personally. *CNet*, 2004. http://cnet.co/V37pZD.

[18] How Rick Santorum's 'Google Problem' Has Endured. http://n.pr/wefdnc.

[19] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. *CIKM*, 2002.

[20] 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. *Inf. Process. Manage.*, 42(4), 2006.

[21] D. Mattioli. On Orbitz, Mac Users Steered to Pricier Hotels. *Wall Street Journal*, 2012. http://on.wsj.com/LwTnPH.

[22] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting Price and Search Discrimination on the Internet. *HotNets*, 2012.

[23] A. Pretschner and S. Gauch. Ontology based personalized search. *ICTAI*, 1999.

[24] A. Pansari and M. Mayer. This is a test. This is only a test. *Google Official Blog*, 2006. http://bit.ly/Ldbb0.

[25] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *J. Comp. Med. Comm.*, 12(3), 2007.

[26] E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, 2011.

[27] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *CACM*, 45(9), 2002.

[28] PhantomJS. http://phantomjs.org.

[29] F. Qiu and J. Cho. Automatic Identification of User Interest for Personalized Search. *WWW*, 2006.

[30] Quantcast. Top Sites for the United States. 2012. http://www.quantcast.com/top-sites.

[31] F. Roesner, T. Kohno, and D. Wetherall. Detecting and Defending Against Third-Party Tracking on the Web. *NSDI*, 2012.

[32] A. Singhal. Some Thoughts on Personalization. *Google Inside Search Blog*, 2011. http://bit.ly/tJS4xT.

[33] A. Singhal. Search, Plus Your World. *Google Official Blog*, 2012. http://bit.ly/yUJnCl.

[34] D. Sullivan. Bing Results Get Local and Personalized. *Search Engine Land*, 2011. http://selnd.com/hY4djp.

[35] D. Sullivan. Why Google "Personalizes" Results Based on Obama Searches But Not Romney. *Search Engine Land*, 2012. http://selnd.com/PyfvvY.

[36] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: A Novel Approach to Personalized Web Search. *WWW*, 2005.

[37] M. Sun, G. Lebanon, and K. Collins-Thompson. Visualizing Differences in Web Search Algorithms using the Expected Weighted Hoeffding Distance. *WWW*, 2010.

[38] N. Singer. The Trouble with the Echo Chamber Online. *The New York Times*, 2011. http://nyti.ms/jcTih2.

[39] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. *CIKM*, 2005.

[40] Y. Shen, J. Yan, S. Yan, L. Ji, N. Liu, and Z. Chen. Sparse Hidden-Dynamics Conditional Random Fields for User Intent Understanding. *WWW*, 2011.

[41] L. A. M. J. Silva. Relevance Ranking for Geographic IR. *GIR*, 2006.

[42] Selenium. http://selenium.org.

[43] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. *KDD*, 2006.

[44] L. Vaughan. New measurements for search engine evaluation proposed and tested. *Inf. Process. Manage.*, 40(4), 2004.

[45] A. Witten. Google's New Privacy Policy. *Google Official Blog*, 2012. http://bit.ly/wVr4mF.

[46] M. Wines. Google to Alert Users to Chinese Censorship. *The New York Times*, 2012. http://nyti.ms/JRhGZS.

[47] C. E. Wills and C. Tatar. Understanding What They Do with What They Know. *WPES*, 2012.

[48] WebMD 2011 Year in Health. http://on.webmd.com/eBPFxH.

[49] Y. Xu, B. Zhang, Z. Chen, and K. Wang. Privacy-Enhancing Personalized Web Search. *WWW*, 2007.

[50] B. Yu and G. Cai. A query-aware document ranking method for geographic information retrieval. *GIR*, 2007.

[51] X. Yi, H. Raghavan, and C. Leggetter. Discovering Users' Specific Geo Intention in Web Search. *WWW*, 2009.

[52] comScore August 2012 U.S. Search Results. http://bit.ly/ThGnOc.