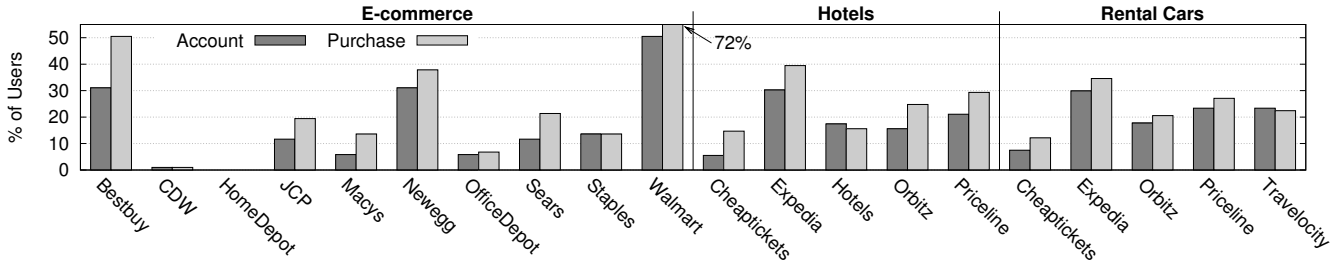

Figure 1: Previous usage (i.e., having an account and making a purchase) of different e-commerce sites by our AMT users.

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

Second, the proxy reduces the amount of noise by sending the experimental, comparison, and control searches to the web site at the same time and from the same IP address. As stated in § 3.2, sending all queries from the same IP address controls for personalization due to geolocation, which we are specifically not studying in this paper. Furthermore, we hard-coded a DNS mapping for each of the sites on the proxy to avoid discrepancies that might come from round-robin DNS sending requests to different data centers.

In total, we recruited 100 AMT users in each of our retail, hotel, and car experiments. In each of the experiments, the participants first answered a brief survey about whether they had an account and/or had purchased something from each site. We present the results of this survey in Figure 1. We observe that many of our users have accounts and a purchase history on a large number of the sites we study.[3]

4.2 Price Steering

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

To measure price steering, we use three metrics:

Jaccard Index. To measure the overlap between two different sets of results, we use Jaccard index, which is the size of the intersection over the size of the union. A Jaccard Index of 0 represents no overlap between the results, while 1 indicates they contain the same results (although not necessarily in the same order).

[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".

Kendall's τ. To measure reordering between two lists of result, we use Kendall's τ rank correlation coefficient. This metric is commonly used in the Information Retrieval (IR) literature to compare ranked lists [23]. The metric ranges between -1 and 1, with 1 representing the same order, 0 signifying no correlation, and -1 being inverse ordering.

nDGG. To measure how strongly the ordering of search results is correlated with product prices, we use Normalized Discounted Cumulative Gain (nDCG). nDCG is a metric from the IR literature [20] for calculating how "close" a given list of search results is to an ideal ordering of results. Each possible search result r is assigned a "gain" score $g(r)$. The DCG of a page of results $R = [r_1, r_2, \ldots, r_k]$ is then calculated as $DCG(R) = g(r_1) + \sum_{i=2}^{k}(g(r_i)/\log_2(i))$. Normalized DCG is simply $DCG(R)/DCG(R')$, where R' is the list of search results with the highest gain scores, sorted from greatest to least. Thus, nDCG of 1 means the observed search results are the same as the ideal results, while 0 means no useful results were returned.

In our context, we use product prices as gain scores, and construct R' by aggregating all of the observed products for a given query. For example, to create R' for the query "ladders" on Home Depot, we construct the union of the results for the AMT user, the comparison, and the control, then sort the union from most- to least-expensive. Intuitively, R' is the most expensive possible ordering of products for the given query. We can then calculate the DCG for the AMT user's results and normalize it using R'. Effectively, if $nDCG(AMTuser) > nDCG(control)$, then the AMT user's search results include more-expensive products towards the top of the page than the control.

For each site, Figure 2 presents the average Jaccard index, Kendall's τ, and nDCG across all queries. The results are presented comparing the comparison to the control searches (Control), and the comparison to the AMT user searches (User). We observe several interesting trends. *First*, Sears, Walmart, and Priceline all have a lower Jaccard index for AMT users relative to the control. This indicates that the AMT users are receiving different products at a higher rate than the control searches (again, note that we are *not* comparing AMT users' results to each other; we only compare each user's result to the corresponding comparison result). Other sites like Orbitz show a Jaccard of 0.85 for Control and User, meaning that the set of results shows inconsistencies, but that AMT users are not seeing a higher level of inconsistency than the control and comparison searches.

Second, we observe that on Newegg, Sears, Walmart, and Priceline, Kendall's τ is at least 0.1 lower for AMT users, i.e., AMT users are consistently receiving results in a differ-

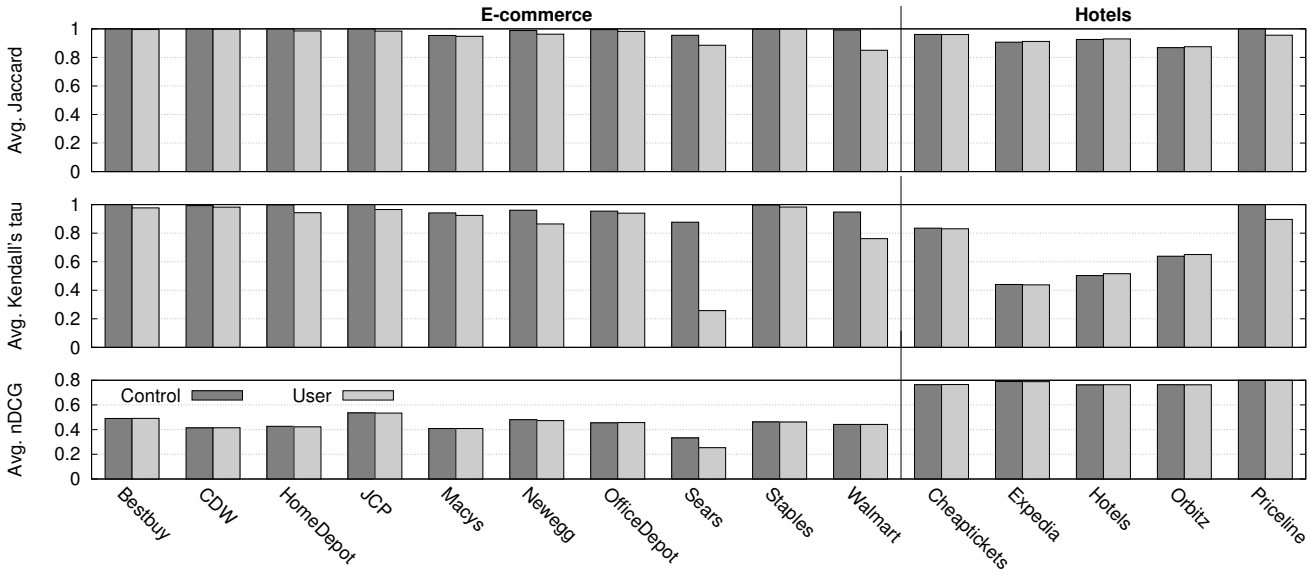

Figure 2: Average Jaccard index (top), Kendall's τ (middle), and nDCG (bottom) across all users and searches for each web site.

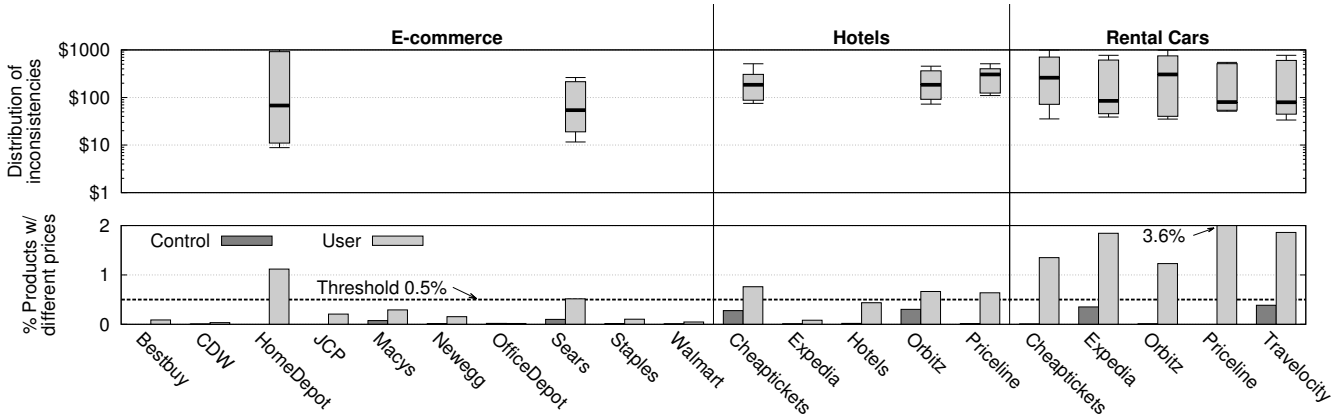

Figure 3: Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with $\geq$0.5% of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).

ent order than the controls. This observation is especially true for Sears, where the ordering of results for AMT users is markedly different. *Third*, we observe that Sears alone appears to be ordering products for AMT users in a price-biased manner. The nDCG results show that AMT users tend to have *cheaper* products near the top of their search results relative to the controls. Note that the results in Figure 2 are only useful for uncovering price steering; we examine whether the target sites are performing price discrimination in § 4.3.

Besides Priceline, the other four travel sites do not show significant differences between the AMT users and the controls. However, these four sites do exhibit significant noise: Kendall's τ is $\leq$0.83 in all four cases. On Cheaptickets and Orbitz, we manually confirm that this noise is due to randomness in the order of search results. In contrast, on Expedia and Hotels.com this noise is due to systematic A/B testing on users (see § 5.2 for details), which explains why we see equally low Kendall's τ values on both sites for all users. Unfortunately, it also means that we cannot draw any conclusions about personalization on Expedia and Hotels.com from the AMT experiment, since the search results for the comparison and the control rarely match.

4.3 Price Discrimination

So far, we have only looked at the set of products returned. We 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 3, we 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, we 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 3, we plot the distribution of price differentials for all sites where >0.5% of the products show inconsistency. We plot the mean price differential (thick line), 25th and

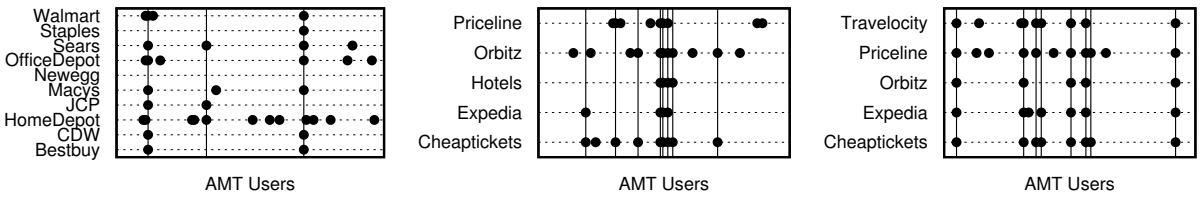

Figure 5: AMT users that receive highly personalized search results on general retail, hotels, and car rental sites.

75th percentile (box), and 5th and 95th percentile (whisker). Note that in our data, AMT users always receive higher prices than the controls (on average), thus all differentials are positive. We observe that the price differentials on many sites are quite large (up to hundreds of dollars). As an example, in Figure 4, we show a screenshot of a price inconsistency that we observed. Both the control and comparison searches returned a price of $565 for a hotel, while our AMT user was returned a price of $633.

4.4 Per-User Personalization

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

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

Next, we examine the AMT users' browser and OS choices. We are able to infer their platform based on the HTTP headers sent by their browser through our proxy. Again, we find no correlation between browser/OS choice and high personalization. In § 5, we do uncover personalization linked to the use of mobile browsers, however none of the AMT users in our study did the HIT from a mobile device.

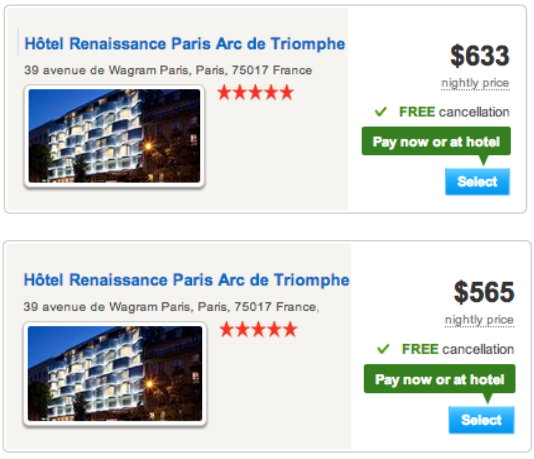

Figure 4: Example of price discrimination. The top result was served to the AMT user, while the bottom result was served to the comparison and control.

Finally, we ask the question: are there AMT users who receive personalized results on multiple e-commerce sites? Figure 5 lists the 100 users in our experiments along the x-axis of each plot; a dot highlights cases were a site personalized search results for a particular user. Although some dots are randomly dispersed, there are many AMT users that receive personalized results from several e-commerce sites. We highlight users who see personalized results on more than one site with vertical bars. More users fall into this category on travel sites than on general retailers.

The takeaway from Figure 5 is that we observe many AMT users who receive personalized results across multiple sites. This suggests that these users share feature(s) that all of these sites use for personalization. Unfortunately, we are unable infer the specific characteristics of these users that are triggering personalization.

Cookies. Although we logged the cookies sent by AMT users to the target e-commerce sites, it is not possible to use them to determine why some users receive personalized search results. First, cookies are typically random alphanumeric strings; they do not encode useful information about a user's history of interactions with a website (e.g., items clicked on, purchases, *etc.*). Second, cookies can be set by content embedded in third-party websites. This means that a user with a cookie from e-commerce site S may never have consciously visited S, let alone made purchases from S. These reasons motivate why we rely on survey results (see Figure 1) to determine AMT users' history of interactions with the target e-commerce sites.

4.5 Summary

To summarize our findings in this section: we 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, we observe that many AMT users experience personalization across multiple sites.

5. PERSONALIZATION FEATURES

In § 4, we demonstrated that e-commerce sites personalize results for real users. However, we cannot determine *why* results are being personalized based on the data from real-world users, since there are too many confounding variables attached to each AMT user (e.g., their location, choice of browser, purchase history, *etc.*).

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.

In this section, we conduct controlled experiments with fake accounts created by us to examine the impact of specific features on e-commerce personalization. Although we cannot test all possible features, we examine five likely candidates: browser, OS, account log-in, click history, and purchase history. We chose these features because e-commerce sites have been observed personalizing results based on these features in the past [1, 28].

We begin with an overview of the design of our synthetic user experiments. Next, we highlight examples of personalization on hotel sites and general retailers. None of our experiments triggered personalization on rental car sites, so we omit these results.

5.1 Experimental Overview

The goal of our synthetic experiments is to determine whether specific user features trigger personalization on e-commerce sites. To assess the impact of feature X that can take on values $x_1, x_2, \ldots, x_n$, we execute $n + 1$ PhantomJS instances, with each value of X assigned to one instance. The $n + 1$th instance serves as the control by duplicating the value of another instance. All PhantomJS instances execute 20 queries (see § 3.4) on each e-commerce site per day, with queries spaced one minute apart to avoid tripping security countermeasures. PhantomJS downloads the first page of results for each query. Unless otherwise specified, PhantomJS persists all cookies between experiments. All of our experiments are designed to complete in <24 hours.

To mitigate measurements errors due to noise (see § 3.1), we perform three steps (some borrowed from previous work [16, 17]): *first*, all searches for a given query are executed at the same time. This eliminates differences in results due to temporal effects. This also means that each of our treatments has exactly the same search history at the same time. *Second*, we use static DNS entries to direct all of our query traffic to specific IP addresses of the retailers. This eliminates errors arising from differences between datacenters. *Third*, although all PhantomJS instances execute on one machine, we use SSH tunnels to forward the traffic of each treatment to a unique IP address in a /24 subnet. This process ensures that any effects due to IP geolocation will affect all results equally.

Static Features. Table 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 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.

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

To conduct our historical experiments, we 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, we manually logged-in to the remaining four accounts on each site and performed specific actions. Two accounts searched for a hotel/car and clicked on the highest and lowest priced results, respectively. The remaining two accounts searched for the same hotel/car and booked the highest and lowest priced results, respectively. Separate credit cards were used for high- and low-priced reservations, and neither card had ever been used to book travel before. Although it is possible to imagine other treatments for account history (e.g., a person who always travels to a specific country), price-constrained (inelastic) and unconstrained (elastic) users are a natural starting point for examining the impact of account history. Furthermore, although these treatments may not embody realistic user behavior, they do present unambiguous signals that could be observed and acted upon by personalization algorithms.

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

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

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

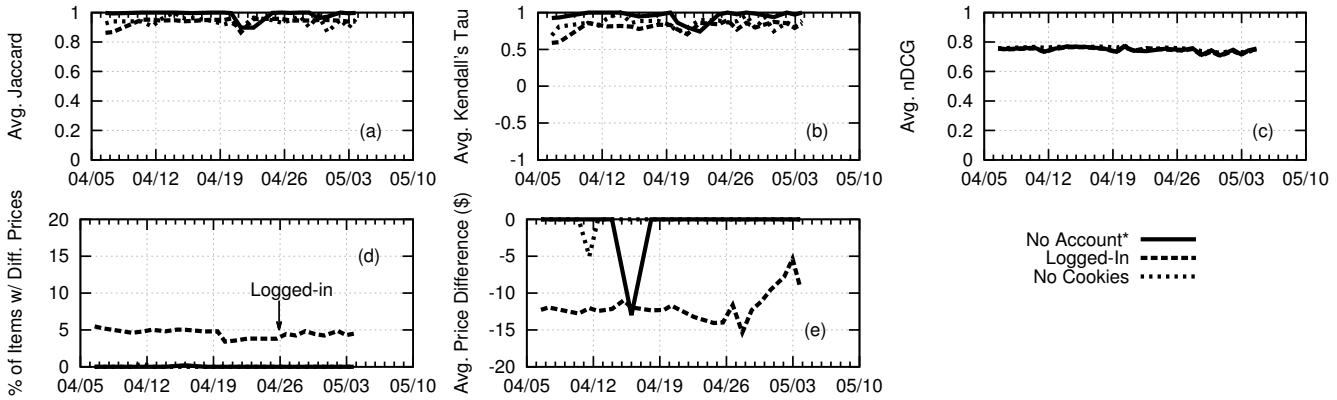

Figure 6: Examining the impact of user accounts and cookies on hotel searches on Cheaptickets.

Analyzing Results. To analyze the data from our feature experiments, we leverage the same five metrics used in § 4. Figure 6 exemplifies the analysis we conduct for each user feature on each e-commerce site. In this example, we examine whether Cheaptickets personalizes results for users that are logged-in. The x-axis of each subplot is time in days. The plots in the top row use Jaccard Index, Kendall's τ, and nDCG to analyze steering, while the plots in the bottom row use percent of items with inconsistent prices and average price difference to analyze discrimination.

All of our analysis is always conducted relative to a control. In all of the figures in this section, the starred (*) feature in the key is the control. For example, in Figure 6, all analysis is done relative to a PhantomJS instance that does not have a user account. Each point is an average of the given metric across all 20 queries on that day.

In total, our analysis produced >360 plots for the various features across all 16 e-commerce cites. Overall, most of the experiments do not reveal evidence of steering or discrimination. Thus, for the remainder of this section, we focus on the particular features and sites where we do observe personalization. None of our feature tests revealed personalization on rental car sites, so we omit them entirely.

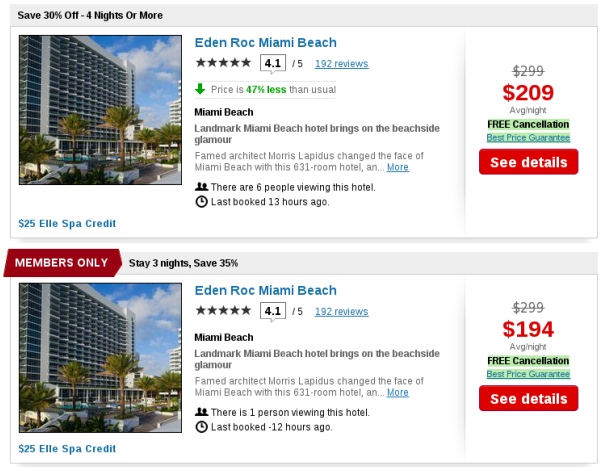

Figure 7: Price discrimination on Cheaptickets. The top result is shown to users that are not logged-in. The bottom result is a "Members Only" price shown to logged-in users.

5.2 Hotels

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

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

Figures 6(a) and (b) reveal that Cheaptickets serves slightly different sets of results to users who are logged-in to an account, versus users who do not have an account or who do not store cookies. Specifically, out of 25 results per page, ≈2 are new and ≈1 is moved to a different location on average for logged-in users. In some cases (e.g., hotels in Bangkok and Montreal) the differences are much larger: up to 11 new and 11 moved results. However, the nDCG analysis in Figure 6(c) indicates that these alterations do not have an appreciable impact on the price of highly-ranked search results. Thus, we do not observe Cheaptickets or Orbitz steering results based on user accounts.

However, Figure 6(d) shows that logged-in users receive different prices on ≈5% of hotels. As shown in Figure 6(e), the hotels with inconsistent prices are $12 cheaper on average. This demonstrates that Cheaptickets and Orbitz implement price discrimination, in favor of users who have accounts on these sites. Manual examination reveals that these sites offer "Members Only" price reductions on certain hotels to logged-in users. Figure 7 shows an example of this on Cheaptickets.

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

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

Table 4: Jaccard overlap between pairs of user feature experiments on Expedia.

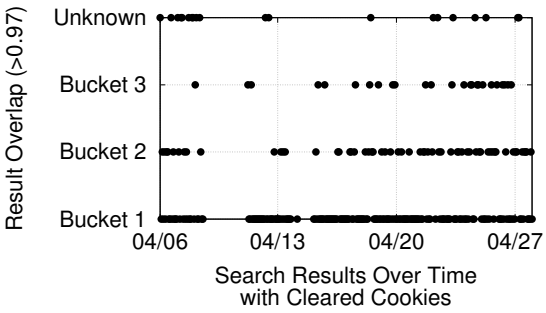

Figure 8: Clearing cookies causes a user to be placed in a random bucket on Expedia.

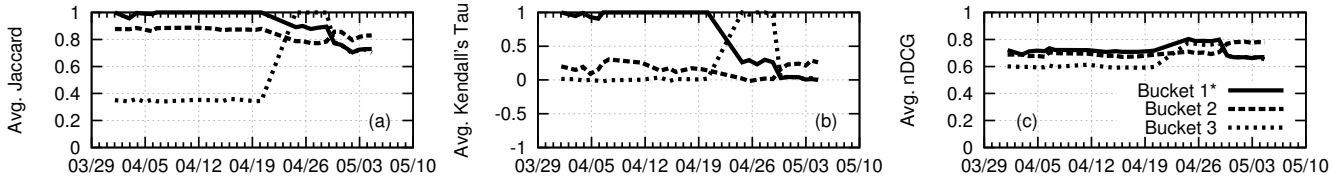

Figure 9: Users in certain buckets are steered towards higher priced hotels on Expedia.

reveal the actual differences in prices offered to members, which may not otherwise be public information.

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

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

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

To determine whether our treatments are being placed in buckets, we generate Table 4, which shows the Jaccard Index for 12 pairs of feature experiments on Expedia. Each table entry is averaged over 20 queries and 10 days. For a typical website, we would expect the control (*Ctrl*) in each category to have perfect overlap (1.0) with its twin (marked with a *). However, in this case the perfect overlaps occur between random pairs of tests. For example, the results for Chrome and Firefox perfectly overlap, but Chrome has low overlap with the control, which was also Chrome. This strongly suggests that the tests with perfect overlap have been randomly assigned to the same bucket. In this case, we observe three buckets: 1) {Windows 7, account control, no account, logged-in, IE 8, Chrome}, 2) {XP, Linux, OS X, browser control, Firefox}, and 3) {OS control}.

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

To test this hypothesis we plot Figure 8, which shows the Jaccard overlap between search results received by the clear cookie treatment, and results received by treatments in other buckets. The x-axis corresponds to the search results from the clear cookie treatment over time; for each page of results, we plot a point in the bucket (y-axis) that has >0.97 Jaccard overlap with the clear cookie treatment. If the clear cookie treatment's results do not overlap with results from any of the buckets, the point is placed on the "Unknown" row. In no cases did the search results from the clear cookie treatment have >0.97 Jaccard with more than a single bucket, confirming that the set of results returned to each bucket are highly disjoint (see Table 4).

Figure 8 confirms that the clear cookie treatment is randomly assigned to a new bucket on each request. 62% of results over time align with bucket 1, while 21% and 9% match with buckets 2 and 3, respectively. Only 7% do not match any known bucket. These results suggest that Expedia does not assign users to buckets with equal probability. There also appear to be time ranges where some buckets are not assigned, e.g., bucket 3 in between 04/12 and 04/15. We found that Hotels.com also assigns users to one of three buckets, that the assignments are weighted, and that the weights change over time.

Now that we understand how Expedia (and Hotels.com) assign users to buckets, we can analyze whether users in different buckets receive personalized results. Figure 9 presents

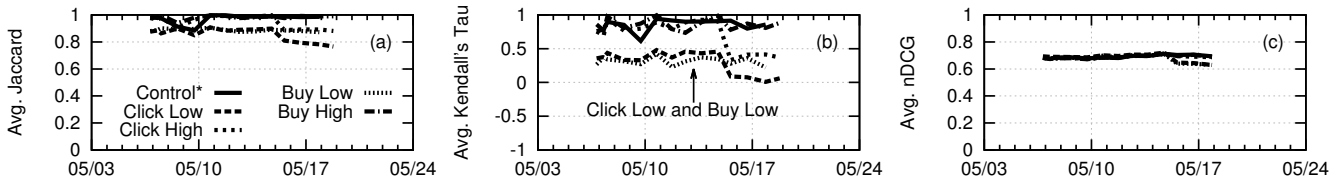

Figure 10: Priceline alters hotel search results based on a user's click and purchase history.

the results of this analysis. We choose an account from bucket 1 to use as a control, since bucket 1 is the most frequently assigned bucket.

Two conclusions can be drawn from Figure 9. First, we see that users are periodically shuffled into different buckets. Between 04/01 and 04/20, the control results are consistent, i.e., Jaccard and Kendall's τ for bucket 1 are $\approx$1. However, on 04/21 the three lines change positions, implying that the accounts have been shuffled to different buckets. It is not clear from our data how often or why this shuffling occurs.

The second conclusion that can be drawn from Figure 9 is that Expedia is steering users in some buckets towards more expensive hotels. Figures 9(a) and (b) show that users in different buckets receive different results in different orders. For example, users in bucket 3 see >60% different search results compared to users in other buckets. Figure 9(c) highlights the net effect of these changes: results served to users in buckets 1 and 2 have higher nDCG values, meaning that the hotels at the top of the page have higher prices. We do not observe price discrimination on Expedia or Hotels.com.

Priceline. As depicted in Figure 10, Priceline alters hotel search results based on the user's history of clicks and purchases. Figures 10(a) and (b) show that users who clicked on or reserved low-price hotel rooms receive slightly different results in a much different order, compared to users who click on nothing, or click/reserve expensive hotel rooms. We manually examined these search results but could not locate any clear reasons for this reordering. The nDCG results in Figure 10(c) confirm that the reordering is not correlated with prices. Thus, although it is clear that account history impacts search results on Priceline, we cannot classify the changes as steering. Furthermore, we observe no evidence of price discrimination based on account history on Priceline.

Travelocity. As shown in Figure 11, Travelocity alters hotel search results for users who browse from iOS devices. Figures 11(a) and (b) show that 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. Note that we started our Android treatment at a later date than the other treatments, specifically to determine if the observed changes on Travelocity occurred on all mobile platforms or just iOS.

Although Figure 11(c) shows that this reordering does not result in price steering, Figures 11(d) and (e) indicate that Travelocity does modify prices for iOS users. Specifically, prices fall by $\approx$\$15 on $\approx$5% of hotels (or 5 out 50 per page) for iOS users. The noise in Figure 11(e) (e.g., prices increasing by \$50 for Chrome and IE 8 users) is not significant: this result is due to a single hotel that changed price.

The takeaway from Figure 11 is that we observe evidence consistent with price discrimination in favor of iOS users on Travelocity. Unlike Cheaptickets and Orbitz, which clearly mark sale-price "Members Only" deals, there is no visual cue on Travelocity's results that indicates prices have been changed for iOS users. Online travel retailers have publicly stated that mobile users are a high-growth customer segment, which may explain why Travelocity offers price-incentives to iOS users [26].

5.3 General Retailers

Home Depot. We now turn our attention to general retail sites. Among the 10 retailers we examined, only Home Depot revealed evidence of personalization. Similar to our findings on Travelocity, Home Depot personalizes results for users with mobile browsers. In fact, the Home Depot website serves HTML with different structure and CSS to desktop browsers, Safari on iOS, and Chrome on Android.

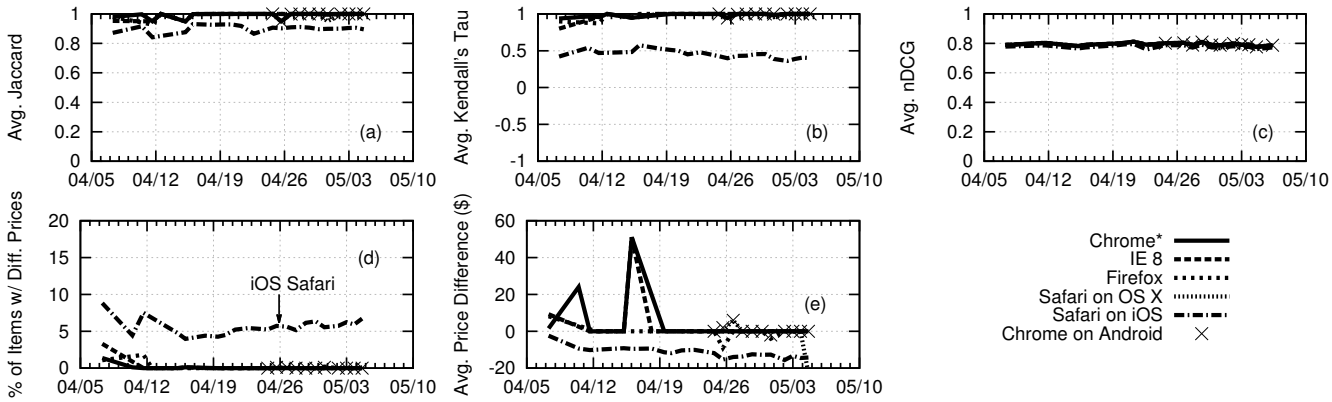

Figure 11: Travelocity alters hotel search results for users of Safari on iOS, but not Chrome on Android.

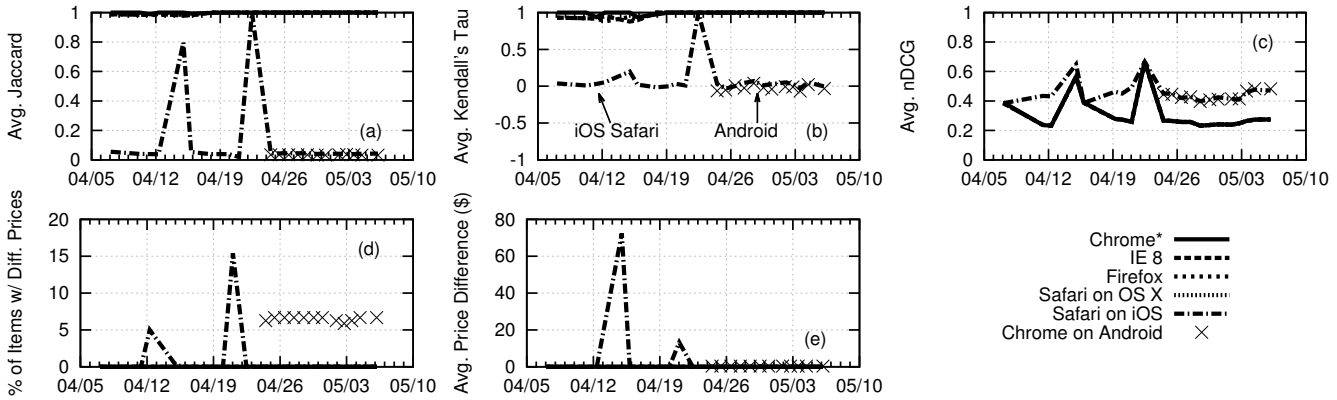

Figure 12: Home Depot alters product searches for users of mobile browsers.

Figure 12 depicts the results of our browser experiments on Home Depot. Strangely, Home Depot serves 24 search results per page to desktop browsers and Android, but serves 48 to iOS. As shown in Figure 12(a), on most days there is close to zero overlap between the results served to desktop and mobile browsers. Oddly, there are days when Home Depot briefly serves identical results to all browsers (e.g., the spike in Figure 12(a) on 4/22). The pool of results served to mobile browsers contains more expensive products overall, leading to higher nDCG scores for mobile browsers in Figure 12(c). Note that nDCG is calculated using the top k results on the page, which in this case is 24 to preserve fairness between iOS and the other browsers. Thus, Home Depot is steering users on mobile browsers towards more expensive products.

In addition to steering, Home Depot also discriminates against Android users. As shown in Figure 12(d), the Android treatment consistently sees differences on ≈6% of prices (one or two products out of 24). However, the practical impact of this discrimination is low: the average price differential in Figure 12(e) for Android is ≈$0.41. We manually examined the search results from Home Depot and could not determine why the Android treatment receives slightly increased prices. Prior work has linked price discrimination on Home Depot to changes in geolocation [30], but we control for this effect in our experiments.

It is possible that the differences we observe on Home Depot may be artifacts caused by different server-side implementations of the website for desktop and mobile users, rather than an explicit personalization algorithm. However, even if this is true, it still qualifies as personalization according to our definition (see § 2.3) since the differences are deterministic and triggered by client-side state.

6. RELATED WORK

In this section, we briefly overview the academic literature on implementing and measuring personalization.

Improving Personalization. Personalizing search results to improve Information Retrieval (IR) accuracy has been extensively studied in the literature [33, 42]. While these techniques typically use click histories for personalization, other features have also been used, including geolocation [2, 53, 54] and demographics (typically inferred from search and browsing histories) [18, 47]. To our knowledge, only one study has investigated privacy-preserving personalized search [52]. Dou et al. provide a comprehensive overview of techniques for personalizing search [12]. Several studies have looked at improving personalization on systems other than search, including targeted ads on the web [16,49], news aggregators [9,24], and even discriminatory pricing on travel search engines [15].

Comparing Search Results. Training and comparing IR systems requires being able to compare ranked lists of search results. Thus, metrics for comparing ranked lists are an active area of research in the IR community. Classical metrics such as Spearman's footrule and ρ [11, 38] and Kendall's τ [22] both calculate pairwise disagreement between ordered lists. Several studies improve on Kendall's τ by adding per-rank weights [14,40], and by taking item similarity into account [21,39]. DCG and nDCG use a logarithmic scale to reduce the scores of bottom ranked items [19].

The Filter Bubble. Activist Eli Pariser brought widespread attention to the potential for web personalization to lead to harmful social outcomes; a problem he dubbed The Internet Filter Bubble [31]. This has motivated researchers to begin measuring the personalization present in deployed systems, such as web search engines [17, 27, 50] and recommender systems [4].

Exploiting Personalization. Recent work has shown that it is possible to exploit personalization algorithms for nefarious purposes. Xing et al. [51] demonstrate that repeatedly clicking on specific search results can cause search engines to rank those results higher. An attacker can exploit this to promote specific results to targeted users. Thus, fully understanding the presence and extent of personalization today can aid in understanding the potential impact of these attacks on e-commerce sites.

Personalization of E-commerce. Two recent studies by Mikians et al. that measure personalization on e-commerce sites serve as the inspiration for our own work. The first study examines price steering (referred to as *search discrimination*) and price discrimination across a large number of sites using fake user profiles [29]. The second paper extends the first work by leveraging crowdsourced workers to help detect price discrimination [30]. The authors identify several e-commerce sites that personalize content, mostly based on the user's geolocation. We improve upon these

studies by introducing duplicated control accounts into all of our measurements. These additional control accounts are necessary to conclusively differentiate inherent noise from actual personalization.

7. CONCLUDING DISCUSSION

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

In this paper, we build a measurement infrastructure to study price discrimination and steering on 16 top online retailers and travel websites. Our method places great emphasis on controlling for various sources of noise in our experiments, since we have to ensure that the differences we see are actually a result of personalization algorithms and not just noise. *First*, we collect real-world data from 300 AMT users to determine the extent of personalization that they experience. This data revealed evidence of personalization on four general retailers and five travel sites, including cases where sites altered prices by hundreds of dollars.

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

Comments from Companies. We reached out to the six companies we identified in this study as implementing some form of personalization (Orbitz and Cheaptickets are run by a single company, as are Expedia and Hotels.com) asking for comments on a pre-publication draft of this manuscript. We received responses from Orbitz and Expedia. The Vice President for Corporate Affairs at Orbitz provided a response confirming that Cheaptickets and Orbitz offer members-only deals on hotels. However, their response took issue with our characterization of price discrimination as "anti-consumer"; we removed these assertions from the final draft of this manuscript. The Orbitz representative kindly agreed to allow us to publish their letter on the Web [7].

We also spoke on the phone with the Chief Product Officer and the Senior Director of Stats Optimization at Expedia. They confirmed our findings that Expedia and Hotels.com perform extensive A/B testing on users. However, they claimed that Expedia does not implement price discrimination on rental cars, and could not explain our results to the contrary (see Figure 3).

Scope. In this study we focus on US e-commerce sites. All queries are made from IP addresses in the US, all retailers and searches are in English, and real world data is collected from users in the US. We leave the examination of personalization on e-commerce 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 price discrimination and steering; we cannot claim the absence of personalization, as we may not have considered other dimensions along which e-commerce sites might personalize content. We observe personalization in some of the AMT results (e.g., on Newegg and Sears) that we cannot explain with the findings from our feature-based experiments. These effects might be explained by measuring the impact of other features, such as geolocation, HTTP Referer, browsing history, or purchase history. Given the generality of our methodology, it would be straightforward to apply it to these additional features, as well as to other e-commerce sites.

Open Source. All of our experiments were conducted in spring of 2014. Although our results are representative for this time period, they may not hold in the future, as the sites may change their personalization algorithms. We encourage other researchers to repeat our measurements by making all of our crawling and parsing code, as well as the raw data from § 4 and § 5, available to the research community at

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

Acknowledgements

We thank the anonymous reviewers and our shepherd, Vijay Erramilli, for their helpful comments. This research was supported in part by NSF grants CNS-1054233 and CHS-1408345, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education grant.

8. 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] L. Andrade and M. J. Silva. Relevance Ranking for Geographic IR. *GIR*, 2006.

[3] Amazon mechanical turk. `http://mturk.com/`.

[4] F. Bakalov, M.-J. Meurs, B. König-Ries, B. Satel, R. G. Butler, and A. Tsang. An Approach to Controlling User Models and Personalization Effects in Recommender Systems. *IUI*, 2013.

[5] K. Bhasin. JCPenney Execs Admit They Didn't Realize How Much Customers Were Into Coupons. Business Insider, 2012. `http://www.businessinsider.com/jcpenney-didnt-realize-how-much-customers-\\were-into-coupons-2012-5`.

[6] P. Belobaba, A. Odoni, and C. Barnhart. *The Global Airline Industry*. Wiley, 2009.

[7] C. Chiames. Correspondence with the authors, in reference to a pre-publication version of this manuscript, 2014. `http://personalization.ccs.neu.edu/orbitz_letter.pdf`.

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

[9] A. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. *WWW*, 2007.

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

[11] P. Diaconis and R. L. Graham. Spearman's Footrule as a Measure of Disarray. *J. Roy. Stat. B*, 39(2), 1977.

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

[13] J. H. Dorfman. *Economics and Management of the Food Industry*. Routledge, 2013.

[14] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SODA*, 2003.

[15] A. Ghose, P. G. Ipeirotis, and B. Li. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, 31(3), 2012.

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

[17] A. Hannak, P. Sapiezyński, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring Personalization of Web Search. *WWW*, 2013.

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

[19] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. *SIGIR*, 2000.

[20] K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM TOIS*, 20(4), 2002.

[21] R. Kumar and S. Vassilvitskii. Generalized Distances Between Rankings. *WWW*, 2010.

[22] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2), 1938.

[23] W. H. Kruskal. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53(284), 1958.

[24] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *WWW*, 2010.

[25] B. D. Lollis. Orbitz: Mac users book fancier hotels than PC users. USA Today Travel Blog, 2012. `http://travel.usatoday.com/hotels/post/2012/05/orbitz-hotel-booking-mac-pc-/690633/1`.

[26] B. D. Lollis. Orbitz: Mobile searches may yield better hotel deals. USA Today Travel Blog, 2012. `http://travel.usatoday.com/hotels/post/2012/05/orbitz-mobile-hotel-deals/691470/1`.

[27] A. Majumder and N. Shrivastava. Know your personalization: learning topic level personalization in online services. *WWW*, 2013.

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

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

[30] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Crowd-assisted Search for Price Discrimination in E-Commerce: First results. *CoNEXT*, 2013.

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

[32] I. Png. *Managerial Economics*. Routledge, 2012.

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

[34] Panopticlick. `https://panopticlick.eff.org`.

[35] PhantomJS. 2013. `http://phantomjs.org`.

[36] A. Ramasastry. Web sites change prices based on customers' habits. CNN, 2005. `http://edition.cnn.com/2005/LAW/06/24/ramasastry.website.prices/`.

[37] C. Shapiro and H. R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, 1999.

[38] C. Spearman. The Proof and Measurement of Association between Two Things. *Am J Psychol*, 15, 1904.

[39] D. Sculley. Rank Aggregation for Similar Items. *SDM*, 2007.

[40] G. S. Shieh, Z. Bai, and W.-Y. Tsai. Rank Tests for Independence—With a Weighted Contamination Alternative. *Stat. Sinica*, 10, 2000.

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

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

[43] Top 500 e-retailers. `http://www.top500guide.com/top-500/`.

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

[45] T. Vissers, N. Nikiforakis, N. Bielova, and W. Joosen. Crying Wolf? On the Price Discrimination of Online Airline Tickets. *HotPETs*, 2014.

[46] J. Valentino-Devries, J. Singer-Vine, and A. Soltani. Websites Vary Prices, Deals Based on Users' Information. *Wall Street Journal*, 2012. `http://online.wsj.com/news/articles/SB10001424127887323777204578189391813881534`.

[47] I. Weber and A. Jaimes. Who Uses Web Search for What? And How? *WSDM*, 2011.

[48] T. Wadhwa. How Advertisers Can Use Your Personal Information to Make You Pay Higher Prices. *Huffington Post*, 2014. `http://www.huffingtonpost.com/tarun-wadhwa/how-advertisers-can-use-y_b_4703013.html`.

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

[50] X. Xing, W. Meng, D. Doozan, N. Feamster, W. Lee, and A. C. Snoeren. Exposing Inconsistent Web Search Results with Bobble. *PAM*, 2014.

[51] X. Xing, W. Meng, D. Doozan, A. C. Snoeren, N. Feamster, and W. Lee. Take This Personally: Pollution Attacks on Personalized Services. *USENIX Security*, 2013.

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

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

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