Detecting price and search discrimination on the Internet

Jakub Mikians†, László Gyarmati⋆, Vijay Erramilli⋆, Nikolaos Laoutaris⋆
Universitat Politecnica de Catalunya†, ⋆Telefonica Research
jmikians@ac.upc.edu,{laszlo,vijay,nikos}@tid.es

ABSTRACT

Price discrimination, setting the price of a given product for each customer individually according to his valuation for it, can benefit from extensive information collected online on the customers and thus contribute to the profitability of e-commerce services. Another way to discriminate among customers with different willingness to pay is to steer them towards different sets of products when they search within a product category (*i.e.*, search discrimination). Our main contribution in this paper is to empirically demonstrate the existence of signs of both price and search discrimination on the Internet, and to uncover the information vectors used to facilitate them. Supported by our findings, we outline the design of a large-scale, distributed watchdog system that allows users to detect discriminatory practices.

Categories and Subject Descriptors

I2 [**Information Systems**]: World wide web

General Terms

Measurement

Keywords

Economics, Privacy, Search, E-Commerce, Price Discrimination, Search Discrimination

1. INTRODUCTION

The predominant economic model behind most Internet services is to offer the service for free, attract users, collect information about and monitor these users, and monetize this information. The collection of personal information is done using increasingly sophisticated mechanisms [12] and this has attracted the attention of privacy advocates, regulators, and the mainstream media. A natural question to ask is: what is done with all the collected information? And the popular answer is, the information is being used increasingly to drive targeted advertising.

Another hypothesis put forward for the wide-scale collection of information, and the related "erosion of privacy" is to facilitate price discrimination [14]. Price discrimination[1] is defined as the ability to price a product on a per customer basis, mostly using personal attributes of the customer. The collected information can be used to estimate the price a customer is willing to pay. Thus, it can have a huge impact on the e-commerce business, whose estimated market size is $961B [9]. The question we deal with in this paper is, "*does price discrimination, facilitated by personal information, exist on the Internet?*". In addition to price discrimination, users can also be subjected to search discrimination, when users with a particular profile are steered towards appropriately priced products.

Detecting price or search discrimination online is not trivial. First, we need to decide which information vectors are relevant and can cause or trigger discrimination, if it exists. We look into three distinct vectors: technological differences, geographical location, and personal information (Sec. 3). For system-based differences, the question is whether the underlying system used to query for prices make a difference? For location, we check whether the price for exactly the same product, sold by the same online site at the same time, differs based on the location of the *originating* query. And for personal information, we are interested if there is a difference in prices shown to users who have certain traits (affluent vs budget conscious). Second, we need to be able to finely *control* the information that is exposed while searching for price or search discrimination, to claim *causality*. In order to uncover price/search discrimination while addressing these concerns, we develop a comprehensive methodology and build a distributed measurement system based on the methodology.

Using our distributed infrastructure, we collect data from multiple vantage points over a period of 20 days (early July 2012), on a set of 200 online vendors. Our

Hotnets '12, October 29–30, 2012, Seattle, WA, USA.

[1]Price discrimination is an established term of economics literature and we use it as such. It does not imply any opinions of the authors regarding price setting policies of any third parties.

main results are:
• We find *no* evidence of price/search discrimination for system based differences, *i.e.*, different OS/Browser combinations do not seem to impact prices.
• We find price differences based on the geographical location of the customer, primarily on digital products, up to 166%—e-books and video games. In addition, we also see price differences for products on a popular office supplies vendor site, when the queries originate from different locations within the same state (MA, USA). However, we cannot claim with certainty that these differences are due to price discrimination, since digital rights costs or competition could offer alternative interpretations.
• When we use trained personas that possess certain attributes (affluent, budget conscious), we find evidence of search discrimination. For some products, we observe prices of products that were shown to be up to 4 times higher for affluent than for budget conscious customers. We also observe this on a popular online hotels/tickets vendor.
• We find signs of price discrimination when we consider the origin URL of the user. For some product categories, when a user visits a vendor site via a discount aggregator site, the prices can be 23% lower as compared to visiting the same vendor site directly.

2. BACKGROUND

Price Discrimination. Price discrimination is the practice of pricing the same product differently to different buyers, depending on the maximum price (reservation price) that each respective buyer is willing to pay. For example, Alice and Bob want to buy the same type of computer monitor and visit the same e-commerce site at approximately the same time. Alice receives $179 as the price while Bob gets $199. The seller offers different prices to them by profiling them (see Sec. 3.4 for details) and realizing that Alice has already visited many electronics' websites and therefore might be more price sensitive than Bob.

From an economics point of view, price discrimination is the optimal method of pricing and increases *social welfare* [19, 3, 13]. Despite its theoretical merits, buyers generally dislike paying different prices than their peers for the same product/service. From a legal point of view, the Robinson-Patman Act prohibits price discrimination in the US under certain circumstances [2] but the possibility is largely open in the current largely unregulated cross-border electronic retail market on the Internet. Recently, a new congress bill aims to make price discrimination on the Internet transparent to end users [16].

Historically, price discrimination has been practiced in myriad industries such as the US railways in the 19th century, flight tickets, personal computers and printers, and college fees [14]. Besides these examples, some minor instances of price discrimination have emerged in the last decade on the Internet as well, *e.g.*, Amazon showed different prices to customers [17], and more recently, Orbitz displayed search results in different orders to some groups of customers [18]. We emphasize that price discrimination and price dispersion[2] are different concepts. Price dispersion occurs when the same product has different prices across different stores for reasons other than the intrinsic value of the product, *e.g.*, because one store wants to reduce its stock or has had a better deal with the manufacturer.

Search Discrimination. Another way to extract more revenue from buyers with a higher willingness to pay is to return more expensive products when they search within a product category. Search discrimination is different from price discrimination because instead of operating on one product, it operates on multiple products trying to *steer* buyers towards an appropriate price range. Ranking of search results greatly impacts the result eventually chosen by the user; users seldom go beyond the first page of results [11]. Hence the search provider, whether a generic search engine or search on e-commerce sites, is in a position enable such discrimination. For example, Alice and Bob are searching for a hotel in Redmond during the same days and for the same type of room. Their searches are launched at approximately the same time. A booking site offers Alice three hotels with prices $180, $200, and $220, while Bob receives quotes from a slightly different set of hotels with prices $160, $180, and $200. This can happen if the site has access to historic data that indicates that Alice tends to stay in more expensive hotels, or by other means such as system information [18]. While search personalization is not entirely new[3], in this paper we draw attention to the *economic* ramifications of it, and in particular study if the information vectors that cause price discrimination also play a role in search discrimination.

Information leading to discrimination. In order to detect discrimination—price or search—we first need to fix the different axes along which the discrimination can take place. We consider three distinct sources of information:
• *Technological/System based differences:* Does the combination of OS and/or browser lead to being offered different prices?
• *Geographic Location:* Does the location of the originating query for the same product and from the same vendor/site play a role? Note that we are *not* interested in the same product sold via local affiliates—for instance Amazon has sites in multiple countries, often selling the same products.
• *Personal Information:* Does personal information, collected and inferred via behavioral tracking methods, impact prices? For instance, does an 'affluent' user see higher prices for the same product than a 'budget-conscious' user?

[2] `http://en.wikipedia.org/wiki/Price_dispersion`
[3] With new implications being discovered, for instance the Filter Bubble concept [6]

Requirements of the system. Based on the definition of price and search discrimination, as well as the axes along which we seek to uncover discrimination, we set the following requirements for our methodology:

• *Sanitary and controlled system:* In order to attribute *causality*, we need to have clean, sanitary, and controlled systems. We should be able to test for one of the axes described above, while keeping the others fixed. For all our measurements, we keep time fixed, *i.e.*, request all price quotations at nearly the same time.

• *Distributed system:* In order to have indicative results, we need a distributed system where we can collect measurements from multiple vantage points.

• *Automated:* To scale the study in terms of customers and vendors, we need to automate the process.

3. METHODOLOGY

The test that we employ while searching for price discrimination is to select a website, an associated product, and then study whether the website returns dynamic prices based on who the potential buyer is. In all the experiments, we compare the results (price or search) retrieved simultaneously to exclude the impact of time from the analyses, *i.e.*, all measurements for a single product happen within a small time window.

3.1 Generic measurement framework

We have developed a measurement framework that uses three components: browsers, a measurement server, and a proxy server.[4] The browser(s) run on separate clean local machines, with the possibility to run over different OSes. To access the pages, we use a JavaScript (JS) application that loads the pages in separate IFrames. We use browsers and JS to ensure we can browse sites that need full features (as opposed to issuing `wget`'s) and to ensure cross-browser compliance. The measurement server controls the JS robot.

Role of the Proxy. We used a proxy for three reasons: (i) We are interested in extracting prices embedded in the pages. Unfortunately JS cannot access and store the content of the opened pages due to its internal *Same Origin Policy*. Hence we configured the browsers to use the proxy server. The proxy then monitored and stored all the traffic going through it. (ii) Some of the destination sites (*e.g.* `amazon.com`) did not open in an iFrame by setting `X-Frame-Options` in the HTTP response headers. The proxy modified the headers on the fly so the option was removed before the page reached the browser. (iii) The proxies allowed us to add additional privacy features, *e.g.*, set the *Do Not Track* option in HTTP headers. In order to mimic behavior of users for sites that need interaction, we used iMacro [10].

Ensuring a Sanitary Environment. We made an effort to prevent any permanent data from being stored in the browser, and thus allowing tracking of the user. The proxy layer allowed us to remove the "Referer" field in the HTTP header that would point to the measurement server, and block pixel bugs [1]. All the browsers were configured to block 3rd party cookies, commonly used for tracking, and we also dealt with flash cookies. Additionally, after each measurement round we deleted the files that might have stored the browsers' state. This restrictive configuration was used for both the system- and the location-based studies.

[4]We modified Privoxy [1].

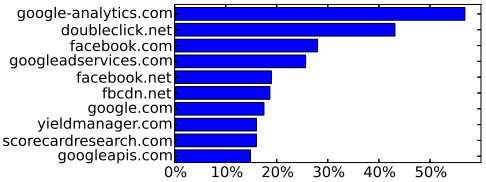

Figure 1: Presence of third party resources on the sites used for training personas.

3.2 System-based measurement specifics

We compared prices of various products accessed from different browsers running on different OSes, from a single geographical location (Barcelona, Spain). We used three systems: Windows 7 Professional, Ubuntu Linux 12.04 and Mac OS X 10.7 Lion with browsers: Firefox 14.0, Google Chrome 20.0 (for all the systems), Safari 5.1 (for OS X) and Internet Explorer 9.0 (Windows). Since we have fixed time and location and prevented identity information leakage, we attribute price difference to the employed system.

3.3 Location measurement specifics

To investigate the impact of a customer's geographical location on the prices she receives, we deployed several proxy servers at different Planetlab nodes. We chose 6 distinct sites: two sites in US (east and west coast), Germany, Spain, Korea, and Brazil. For this experiment, we used 6 separate, identical virtual machines with Windows 7 and Firefox. With this configuration, the only information that distinguished the browsers externally was their IP. We assume that the IP address is enough to identify the geographical location of the originating query and is enough for price discrimination to take place. We fixed time when we conducted our measurements across sites, syncing various sites using NTP.

3.4 Personal info measurement specifics

In order to uncover discrimination based on personal information, we follow two methods that differ in the amount of information that they employ. In the first we train "personas" that conform to two extreme customer segments: *affluent customers* and *budget conscious customer*. The two profiles are quite distinct. The budget conscious customer visits price aggregation and discount sites (like `nextag.com`). The affluent customer visits sites selling high-end luxury products. The customers might be tracked by third party aggregators (*e.g.*, DoubleClick) that have presence on many sites around the web and can chain such visits to construct a profile of the user.

We train personas as follows. We obtain the generic traits followed by an affluent consumer and a budget conscious consumer from [4]. An affluent consumer is more likely to visit "Retail–Jewelry/Luxury Goods/Accessories" sites as well as "Automotive resources" and "Community Personals" sites than the average user. For each of these categories, we use Alexa.com and Google to select the top 100 popular sites, and configure a freshly installed system to visit these sites, and to train the profile. In order to mimic a real human, we train only between 9AM-12PM and use an exponential distribution (mean: 2 min) between requests. We do the same to train the "budget conscious" consumer by using the relevant sites. We train both profiles for 7 days, and we permit tracking and disable all blocking. Note that we can train multiple personas resembling different segments—this is left for future work. We show the distribution of third party trackers on the sites we used for the training in Fig. 1.

The second method that we use to test for discrimination based on personal information uses the "Referer" header that reveals where a request came from. Therefore, if you come from a discount site or a luxury site the e-commerce site where you land knows about it and can use it as indication of your willingness to pay. We fix one location—Los Angeles, USA—and fix one system—Windows 7 with Firefox—to run the personal information related measurements.

Assumptions: For the three sources of price discrimination we are studying, we assume that the information vectors we use are sufficient in isolation for price discrimination to kick-in. In reality, a composition of different vectors may be needed for price discrimination. For instance, personas and a specific type of system configuration may be needed together for price discrimination. Composing different vectors and then testing for discrimination is left for future work.

3.5 Analyzed Products

To determine the types of products to focus on, we selected the product categories from Alexa. In total, we examined 35 product categories (*e.g.*, "clothing") and we choose 200 distinct vendors (*e.g.*, `gap.com`). From the identified e-commerce sites, we selected 3 concrete products with their unique URLs (*e.g.*, specific piece of clothing). For each vendor, we selected low/mid/high price products. In case of hotels, we selected three different dates (low/mid/high season) at multiple locations. The 200 vendors we chose may appear to be a small set. However, we limit ourselves to 200 to first understand issues with scaling. In addition, these 200 vendors also account for the vast majority of user traffic as they include large vendors like `amazon.com` and `bestbuy.com`. We intend to increase these 200 vendors to 1000+ vendors to also cover long-tail sites. In the end we had a total of 600 products; we provide more details on them in the Appendix.

4. EMPIRICAL RESULTS

4.1 System based differences

We collected extensive measurements on 600 different products. We used the 8 distinct system–browser setups to examine the potential price differences. We ran the measurements for four days, and collected over 20,000 distinct measurement points in total. In addition, we queried Google and Bing to examine if the search results differ based on the systems. For this, we used 26 different phrases related to the products we analyze. The measurement did *not* reveal any price differences between the end systems. Regarding search discrimination, we did not find differences that were significant.

4.2 Geographic location

Next, we looked into the impact of geographic location from where the user accesses an e-commerce site. We issued queries through the proxies described in Sec. 3.3 on the same set of products/sites as before. In total, we accessed each product 10 times. The measurement results do *not* indicate significant differences, neither in prices nor in search results, for the majority of the products. However, the prices shown by three particular websites appeared to depend strongly on the users' location. In particular, `amazon.com` and `steampowered.com` returned prices for digital products (e-books and computer games, respectively) and `staples.com` for office products that differ between buyers at different locations.

In the case of Amazon, we observed price differences only for Kindle e-books. We queried the prices of books listed on the top 100 list of Amazon from six locations.[5] Only 27 out of these 100 books were available for purchase in their original English version from Amazon.com (US site) to customers coming from all the 6 locations we were testing. We illustrate the price differences of these products in Fig. 2, where we plot the ratio of the products' prices using the prices in New York, USA as reference. In majority of the cases, the price difference is at least 21%; however, in extreme cases it can be as high as 166%.

For the Steam site, we examined more than 300 additional products. We compared the prices of the products where their prices were displayed in the same currency to avoid the bias of currency exchange. We observed price differences for 20% of the products in case of Spain and Germany (figure not shown). Moreover, 3.5% of the products had different prices in case of US, Brazil, and Korea.

Next we analyzed the impact of location on a finer scale, *i.e.*, within the US only. We used 67 Planetlab nodes in US acting as proxy servers. We accessed 10 random products from `staples.com` using the proxies. 4 products showed different prices when accessed from different locations. In those cases, there were two dis-

[5]For both websites, results for US/LA and US/NY overlap and are not shown.

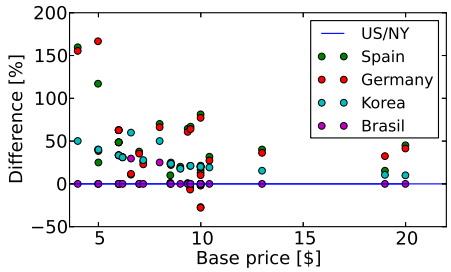

Figure 2: Price differences at Amazon based on the customer's geographic location using the prices in New York, USA as reference. For each of the considered products there exist at least two locations with different prices.

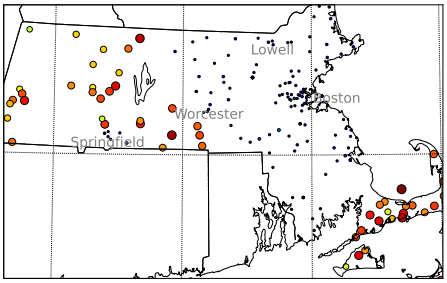

Figure 3: Price differences at staples.com. The dot sizes mark the mean price surplus for the locations, from 0% (small dots) up to 3.9% (large dots)

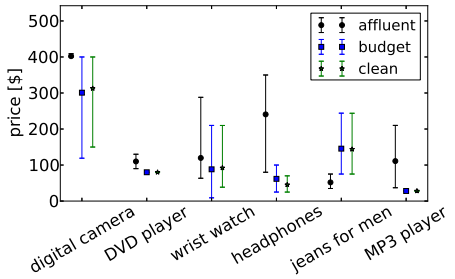

Figure 4: Prices (mean/min/max) shown by Google to the different personas. The median number of products in each category per persona is 12.

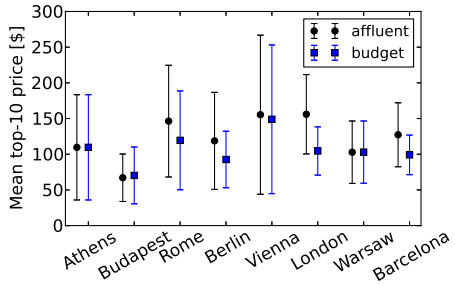

Figure 5: Mean prices (with std. deviations) of top-10 results from Cheaptickets.com returned to affluent and budget personas. The mean difference is 15%, and can be even as high as 50%.

tinct prices for the same product. We did not observe a significant correlation between the prices and population per state/city, population density per state, income per state, or tax rates per state.

We extended the study of `staples.com` by taking measurements within the same state (MA) to exclude inter-state tax differences. We selected 29 random products and 200 random ZIP codes.[6] Again, for 15 products the price varied up to 11% above the base price between the locations.[7]

Fig. 3 shows the price differences geographically. The values on the map show a mean price surplus calculated for a particular location over all the products. The map shows that the outskirts are shown higher prices than the large cities.

Discussion: Our system ensures that the only bit of information that is exposed is the IP address, hence the location. We see differences in prices for some digital goods as well as office supplies. We cannot claim to have discovered price discrimination since the differences might be attributed to other reasons such as intellectual property issues or increased competition between retailers or logistics. Further investigation is required on this issue.

4.3 Personal information

[6]When accessing `staples.com` from outside of US, the service asks for the customer's ZIP code, giving equivalent results as coming from a certain location.

[7]Base price - smallest observed price for a product.

Trained personas. We used the previously trained personas (Sec. 3.4) to examine the discrepancies of products based on the browsing behavior. We also used a clean profile as a baseline. We did *not* observe price discrimination in our results; however, we observed different search results on two sites. First, we examined 12 search queries in `google.com`, three times for each profile. For half of the queries, the results included several suggested products, together with the prices. There is a noticeable difference in the prices of these products as we show in Fig. 4. For instance, the mean price was 4 times higher in case of "headphones" for the affluent persona than for the budget one. Second, we examined the top-10 hotel offers on Cheaptickets. We searched for hotels in 8 different cities on 8 different dates. The search engine of Cheaptickets returned offers with higher prices for the affluent profile (Fig. 5).

Originating web page. Our hypothesis for studying the origin is that the site that a customer uses to reach a product site can provide valuable information for pricing purposes. For example, if the customer comes from a discount site, she will be more likely to be price sensitive than someone coming from a luxury site or a portal. Hence, we focus on price aggregator sites that provide a platform for vendors of various products and also provide discounts to users. We looked into a couple of aggregator sites (`nextag.com`, `pricerunner.co.uk`, `getprice.com.au`), but we only present results of one large site: `nextag.com`. We used a clean profile, with

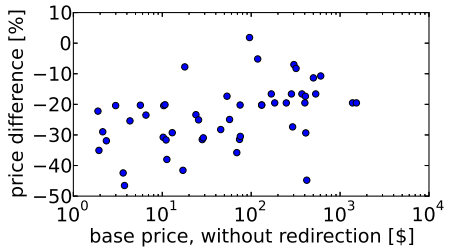

Figure 6: Price difference at the Shoplet.com online retailer site, with- and without redirection from a price aggregator.

blocking enabled but enabled first party cookies. We examined 25 different categories of products available on `nextag.com`. We found two online vendors (`shoplet.com`, `discountofficeitems.com`) who returned different prices based on the originating web page of the customers. Both retailers specialize in office equipment. In case of `shoplet.com`, users get higher prices if they access a product directly via the retailer's website than when the price aggregator (`nextag.com`) redirects the user to the store. In the latter case, the aggregator redirects the user to an intermediate site that sets a cookie, and from this point on the user starts getting lower prices. We quantify the price differences with- and without the redirection in Fig. 6. The mean difference between the prices is 23%.

Discussion: We noticed signs of search based discrimination in case of trained personas. We stress that while we have not yet found price discrimination for trained personas, we did observe signs of discrimination via origin URL. We note that the entities who collect large amounts of information across the web (aggregators like Doubleclick)—and hence can create a more accurate representation of the user—do not actively engage in e-commerce. On the flipside, large vendors do not track users across the web. Thus, the entities who could utilize information of users for pricing are decoupled from those who collect such information. The redirection mechanism, that uses one bit of information, can be used effectively to narrow this information gap.

5. RELATED WORK

The notion of building large distributed systems to understand the effect of personal information on services obtained has been done for various reasons [8, 6]. Guha, *et al.* [8] focused on the impact of user characteristics on display advertisements. Our framework is similar; however, we focus on the differences of product prices instead of displayed ads. Our work is closely tied to online privacy, both in terms of usage of privacy preserving tools in our methodology, as well as implications of (loss of) privacy over price discrimination. For the former, we use the findings of Krishnamurthy, *et al.* [12] to block known forms of tracking, on our proxy as well as the browser. Besides cookies, other techniques can also uniquely identify users with high probability such as the properties of the browsers [5] and the browsing history [15], hence we take steps to counter such identification.

6. CONCLUSIONS

Our measurements suggest that both price and search discrimination might be taking place in today's Internet. In our ongoing efforts we are scaling by orders of magnitude both the number of sites and the product categories that we examine. Our preliminary results also point to a natural extension of our distributed system: co-opt and retrofit it as a *watchdog system* that helps users check if they are being discriminated.

7. ACKNOWLEDGEMENTS

We thank our shepherd Michael Walfish for helpful comments as well as the anonymous reviewers of Hotnets.

8. REFERENCES

[1] Privoxy, `http://www.privoxy.org/`.
[2] The Robinson-Patman Act, Pub. L. No. 74-692, 49 Stat. 1526, 1936.
[3] A. Acquisti and H.R. Varian. Conditioning Prices on Purchase History. *Marketing Science*, 24(3), 2005.
[4] AudienceScience. `http://www.audiencetargeting.com`.
[5] P Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies, LNCS 6205*, pages 1–18. 2010.
[6] Georgia Tech Information Security Center. Filter Bubble project, 2012. `http://bobble.gtisc.gatech.edu/`.
[7] Google. Adwords Keyword Tool. `https://adwords.google.com/o/Targeting/Explorer?__c=1000000000&__u=1000000000&__o=cues&ideaRequestType=KEYWORD_IDEAS`.
[8] S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. ACM IMC '10.
[9] IMRG. B2C Global e-Commerce Overview 2012.
[10] iOpus. iMacro. `https://addons.mozilla.org/en-US/firefox/addon/imacros-for-firefox/`.
[11] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. SIGIR '05.
[12] B. Krishnamurthy, D. Malandrino, and C.E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. ACM SOUPS '07.
[13] R.P. McAfee. Price Discrimination. In *Issues in Competition Law and Policy, vol. 1*. 2008.
[14] A Odlyzko. Privacy, economics, and price discrimination on the internet. ICEC '03.
[15] L. Olejnik, C. Castelluccia, and A. Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. HotPETs '12.
[16] Susan Davis. H.R. 6508: To direct the Federal Trade Commission to promulgate rules requiring an Internet merchant to disclose the use of a price-altering computer program, and for other purposes.
[17] The New York Times. Amazon's Prime Suspect. `http://www.nytimes.com/2010/08/08/magazine/08FOB-medium-t.html`.
[18] The Wall Street Journal. On Orbitz, Mac Users Steered to Pricier Hotels. `http://online.wsj.com/article/SB10001424052702304458604577488822667325882.html`.
[19] H.R. Varian. Price Discrimination and Social Welfare. *The American Economic Review*, 75(4):870–875, 1985.

APPENDIX

Examples of sites visited with products in parentheses
airlines: `aa.com` (3), `britishairways.com` (3), `easyjet.com` (3), `lufthansa.com` (3), `usairways.com` (3), digital cameras: `amazon.com` (3), `bestbuy.com` (3), `overstock.com` (3), `ritzcamera.com` (3), hotels/travel: `booking.com` (3), `expedia.com` (3), `hotels.com` (3), `cheaptickets.com` (10+), `kayak.es` (3), `orbitz.com` (3), `travelocity.com` (3)