Tracking and Tricking a Profiler

Automated Measuring and Influencing of Bluekai's Interest Profiling

Martin Degeling
Ruhr University Bochum
Bochum, Germany
martin.degeling@ruhr-uni-bochum.de

Jan Nierhoff
Ruhr University Bochum
Bochum, Germany
jan.nierhoff@ruhr-uni-bochum.de

ABSTRACT

Online advertising services infer interest profiles based on users browsing behavior, but little is known about the extent of these profiles and how they can be influenced. In this paper we describe and evaluate a system to analyze online profiling as a black box by simulating web browsing sessions based on links posted to Reddit. The study utilizes Oracle's Bluekai Registry[1] to gain insights into profiles created through online tracking and evaluates how they can be obfuscated. We report on the extent of Bluekai's tracking network and taxonomy, analyze how profiles are shown to users, and observe how they develop for sessions of up to 3,000 website visits. Our results show that only a fraction of websites influence the interests assigned to a session's profile, that the profiles themselves are very noisy, and that identical browsing behavior results in different profiles. We evaluate two simple obfuscation schemes that effectively alter interest profiles by selectively adding 5% targeted obfuscation traffic.

ACM Reference Format:
Martin Degeling and Jan Nierhoff. 2018. Tracking and Tricking a Profiler: Automated Measuring and Influencing of Bluekai's Interest Profiling. In *2018 Workshop on Privacy in the Electronic Society (WPES'18), October 15, 2018, Toronto, ON, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3267323.3268955

1 INTRODUCTION

Online advertising has seen massive changes and innovation since the first banner ad was launched in 1994 [51]. The market has evolved not only in terms of revenue and competition but also with respect to the amount of data that is collected about internet users, often referred to as *audience*. Nowadays data management platforms support marketers in managing advertising campaigns on multiple channels and analyzing their effectiveness [18]. Those platforms help to automate advertising campaigns and make use of mechanisms like programmatic advertising to direct campaigns at relevant target groups. Ad spaces on websites are traded on high frequency bidding platforms to allow marketers to direct their campaigns to the "right" audience [35, 60]. This automated process involves various actors including supply and demand side platforms between marketers that want to place an ad and publishers that provide ad space on their websites. The algorithms take various factors into consideration, e.g. the general audience of a website, the topic of the article or content, where the ad will be placed, and, of course, the profile associated with a browser requesting the website.

Generating these profiles is crucial, since marketers want to target specific groups of people based on demographics, psychographic traits, and behavioral data [19]. Profiling is easier in closed platforms like social networks where users voluntarily provide parts of the information that comprises a profile and link them to one single account used on different devices. Outside of these platforms profiling is a more complex task that requires tracking and uniquely identifying individual users, mostly using cookies and properties of the browser [42]. Based on the websites visited, attributes describing personality traits and interests are assigned without users' explicit interaction (or consent). All this happens in a scattered market landscape with thousands of competing services [58], where most services are only able to track a fraction of a user's movements on the internet fostering cooperation in tracking networks.

Instead of analyzing the network of tracking services in breadth this paper focuses in depth on one specific online profiling service, Oracle's *Bluekai*. Bluekai is worth studying as the service is not only widely used directly, but known to exchange data with a large number of other tracking services. Moreover, it belongs to the few services that offer some transparency about the profiles created. For our analysis we simulated web browsing sessions based on links users have posted to Reddit and found that an average interest profile consisted of about 5 out of 22 high-level interest categories. We then compared profiles that were created based on the same websites and found them to be rather inconsistent. Even in long sessions, in which up to 3,000 websites were visited in the same order, the resulting profiles overlap only by 75%. We then evaluated two different obfuscation strategies where obfuscation traffic was generated by observing the profile created. We were able to reliably alter or extend the interest profile created by Bluekai, either by doubling the number of interests in a profile or by changing the interests assigned. In both cases only 5% of the original number of visited pages was needed to alter or extend the profile.

The rest of the paper is organized as follows: We first summarize what is known about online tracking and profiling from previous research (3.1refsec:relatedwork). We then describe how Bluekai profiles individuals (3) and our methodology to observe this profiling (4). In the following sections we analyze the profiles that are

[1] See https://datacloudoptout.oracle.com/registry

WPES'18, October 15, 2018, Toronto, ON, Canada

ACM ISBN 978-1-4503-5989-4/18/10…$15.00
https://doi.org/10.1145/3267323.3268955

created (5), evaluate what influences their extent and stability (5.4) and test how they can be obfuscated (6).

2 RELATED WORK

Online tracking and consecutive profiling help to create a personalized web experience and, as part of the advertising business, is crucial to finance the majority of free web services. However, as the tracking ecosystem evolved it has drawn criticism, because a large amount of data about internet users is collected and aggregated without offering sufficient transparency and choice for consumers. Research has shown how tracking data is collected technically and how the use of that data influences individual privacy. We want to extend the knowledge on online tracking by focussing on the results of tracking: the high-level personality profiles created from clickstreams.

2.1 Online Tracking and Advertising

Although new technologies have emerged to track users' web browsing behavior, the majority of services, including Oracle's Bluekai, rely on cookies to track users. Cookies are small text tokens stored in the web browser that are send along all HTTP(S) requests to a specific host domain. Advertising networks that are included on multiple websites use Cookie IDs together with the HTTP-Referer Header to identify users across different websites. Researchers have frequently measured the extent of web tracking [29, 32, 43, 46, 59], the lack of disclosure of the practices in privacy policies [22], and uncovered new trends in how users' web navigation can be tracked with methods like browser fingerprinting and cookie syncing [2, 5], on multiple devices [7] and how third party tracking services leak information [16].

Online profiling, the process of generating high level information like demographics or psychograms based on the tracking data, has seen slightly less attention. Researchers have shown that profiling for advertisement is largely biased [13], and many users feel uncomfortable with being profiled [45] requesting more control over their data [33]. Studies on profiling in social networks have also revealed that the profiles created are often not as correct as users think [48, 50]. However, there is limited knowledge about how online tracking companies create profiles, as the algorithms are kept secret. With respect with the Bluekai, Rao et al. have shown in a small user study that users consider the interest profiles created by Bluekai to be inaccurate [41]. Google has been shown to be quite good at guessing demographic information [53], but there is evidence that its interest profiling lacks accuracy [15].

2.2 Privacy Impact of Online Profiling

User tracking and profiling has been claimed to harm privacy rights as they are often conducted without (explicit) consent and only a fraction of users are aware of the extent and influence profiles can have [26]. Although the data collected through tracking is often not directly linked to an individual by name, the mere number of details about which websites were visited, and when, makes users identifiable [4, 55].

In addition, profiling, and the personalization that it is driving, has many drawbacks to individual privacy and autonomy. First, those that perform profiling are trying to single out those recipients (e.g. for advertisements) that are considered relevant. While it can be beneficial to not be targeted with irrelevant ads, this often comes hand in hand with reproducing stereotypes [13, 49] or can be used to offer dynamic prices [25][2] to different groups, ultimately leading to exclusion from markets [21].

Second, it has been suggested that, if profiling and personalization is used excessively and without user notice, it can lead to a filter bubble [37], which creates informational spaces with a lack of serendipity. Research has shown that many users are aware that algorithms influence what is shown to them, e.g. on Facebook [40].

Third, from an philosophical perspective individualized marketing is criticized as part of a *panoptic sort* [20]. Those that perform profiling have power over which categorizations are created and who is categorized for what purpose. The power relation between tracking services and users is asymmetric as few insights and options are offered to control a system that mostly operates as a black box. Instead of informed choice the panoptic sort produces cybernetic feedback loops [14, 31] that, while trying to personalize and offer choice, in fact reduce serendipity and re-enforce potentially discriminating norms.

Discussions about the negative effects of online tracking on privacy have also had impact on regulation, most recently the European General Data Protection Regulation (GDPR) the upcoming Eprivacy Directive are supposed to bring change to the industry. The study described in this paper was conducted before the GDPR went into effect in May 2018. By the time of writing it seems that the new rules have only little impact on the extent of tracking and profiling, but lead to more transparency about the practices [11].

2.3 Simulating Internet Usage

When trying to automate the analysis of online profiling, one challenge to simulate browsing behavior. Previous studies on online tracking have mostly used publicly available toplists to visit sites that have a large user base (e.g. [2, 43, 46]). Toplists are largely biased towards platforms and sites that themselves are hubs for users to get to other sites on the web (e.g. Google or Facebook). While toplists are useful to assess how widely used certain tracking technologies are, they can't be used to simulate individual browsing behavior. In addition, ranking websites by traffic is difficult and even the largest companies struggle to provide good data [30, 44]. Especially considering how the use of the internet changes over time, be it with the rise of social media, mobile browsing or apps, the way the web is used is constantly in flux.

Another study used released data on search terms [16] to discover and measure tracking websites. A drawback of this approach is that the search term list originated in 2006 and therefore does not cover current events and topics that are also often used in online marketing campaigns.

Although there are numerous metrics of which websites are visited most often or which search terms and topics are most relevant to web users in general, only few scientific publications offer insights into which websites are visited in which order during a browsing session. Goel et al. present a breakdown of internet usage categories and demographics showing that those that spent the

[2] Another study could not confirm price discrimination [54], though the fact that *dynamic pricing* or *surge pricing* is happening is undisputed [1].

most time online do so on a few websites, mostly related to social media [24]. Papadikis et al. report that between 45% and 81% of the websites users visited, they visited multiple times [38].

In this study we used a different approach that is presented in detail section 4.3. We make use of links posted to Reddit as it lists links to specific articles instead of front pages of websites and leverage the subreddit structure to identify websites related to similar interests and use these to create long running sessions. This allows us to automatically study tracking and profiling, but also comes with some limitations when compared to qualitative studies in this area.

3 BLUEKAI

Our study focuses on Bluekai's services, whose profiling capabilities we analyzed. Bluekai is part of Oracle, one of the largest companies engaged in the business of online advertising. Bluekai started as a independent tracking service and was acquired by Oracle in 2014. It is now integrated into Oracle's Data Management Platform, which combines tracking data from multiple tracking services and offers technology for marketers to plan and evaluate advertising campaigns [10]. Until today several of Oracle's services, including the tracking services and the registry described in this section, are available under the Bluekai brand.

3.1 Tracking by Bluekai and Partners

According to its privacy policy[4] Bluekai relies on cookie tracking and Pixel Tags. Thus Bluekai is not able to make use of evercookies or browser fingerprinting, techniques that used to circumvent tracking blockers [2]. Still, Bluekai claims it is performing cross-device tracking, used to link browsing sessions on multiple devices to one user profile. The cookie set by Bluekai's server (named BKU) contains an ID, and its retention time in the browser is 180 days. This is also the time frame for which, according to the privacy policy, interest categories are assigned to a user.

The extent of Bluekai's services was assessed in multiple recent studies. Yu et al., based on the data from 200.000 users of their browser add-on, found that Bluekai is tracking on 1.3% of all websites [59] while Engelhardt et al. found its trackers on 10% of the top 1 million websites in Alexa's Index[17]. The latter result is in line with another study [15] that found Bluekai to be able to directly track users on 10 to 20% of regularly visited sites.

Besides generating data from its own tracking service, Bluekai listed 86 other companies like AddThis (acquired by Oracle in January 2016) or Acxiom and Lotame (see "Branded Data" in 1). While some companies provide additional consumer data other have their own tracking services and can theoretically combine their data by using techniques like cookie syncing [2] or internal exchange data and user IDs to merge their data sets.

3.2 Bluekai Taxonomy

Bluekai combines data from a large number of tracking services and data providers in its data management platform. An overview of the data sources marketers can use to plan and target their campaigns is provided in the Bluekai taxonomy report [6]. Table 1 summarizes the main categories from which marketers can select to create target audiences. Each user can potentially be put in any of these categories.

In addition to basic information that is related to the device and browser used to surf the web (device data) or that can be easily inferred (geographic data), Bluekai also provides data in three categories: demographic, interests, business, and in-market. The largest number of subcategories is related to *branded data*, which contain information from 86 data providers that sell to or cooperate with Bluekai. Many of those are specialized in different areas and provide fine grained categorizations from behavior (e.g. "Prefer Picking Up Quick Meals") to product preferences ("Maple/Pancake & Waffle Syrup") and political orientation ("Politics & Society - Liberal").

The taxonomy itself is dynamic and Bluekai publishes updates irregularly. For example, between January 2015 and May 2017 eighteen change reports were issued each of them listing between 1,500 and 4,000 changes for the list of 50,000 categories[5]. While the majority of changes are based on changes partner companies made to their taxonomies, this also affects Bluekai's own data set. During our study, eight interests were added and one was deleted.

Unlike other advertising and webmetric companies, Bluekai intentionally does not provide or allow targeting by information about ethnicity (which is provided for example by Quantcast or Alexa) or religion. However, Bluekai's taxonomy includes categories that might be used as proxies for more sensitive information, like a persons religion from the categorization as"Consumers of Christian Television Network (Broadcast)" as well as ethnic affiliation for households categorized as "Spanish Language Spoken".

3.3 Bluekai Registry

What distinguishes Bluekai from other services is that the company maintains a website where users can review the profile created about them (see Fig. 1). Only few other services, including Google and Yahoo[6], offers similar transparency mechanisms, though the number increased after the introduction of the GDPR.[7] We focussed on Bluekai since Google has been studied before [13, 15] and Yahoo's interest profiling is very limited in size and reach[8].

In the latest version, the purpose of the registry is to increase transparency for the users about what data is collected. In previous versions, including the one we studied, it also allowed users to delete certain entries. At no time was there a way for a user to add information to the profile. Although the registry displays categories from the hierarchy of the taxonomy, category titles are re-labeled, presumably to be more understandable. But the categorization itself sometimes produces incomprehensible information. For example the rather abstract information "Branded Data >Forbes >Performance Segments", a category assigned by forbes.com to guide advertisers, is displayed in the category "What others know about you", is hard to understand without additional knowledge.

[4] https://www.oracle.com/legal/privacy/marketing-cloud-data-cloud-privacy-policy.html

[5] Reports are available at https://docs.oracle.com/cloud/latest/marketingcs_gs/OMCDA/Help/AudienceDataMarketplace/bluekai_taxonomy_report.html

[6] Registered Google users can review their ad preferences and Yahoo offers an ad interest manager.

[7] TheTradeDesk and Lotame have added subject access requests options to their websites in May 2018.

[8] During a pre-study we found that Yahoo only created profiles for 30% of our sessions and profiles were less informative, presumably due to a smaller tracking network.

Information	#	*Subcategories*
Geographic	2540	Locations in the US are organized in Census Areas[3] derived from IP address range
Device Data	193	Browser, Browser Version and Operating System retrieved from user-agent strings
Demographic	241	Age, Education (including "recently graduated"), Family Composition (if there are children in the household and if yes, their age group. also "Family Position"), Financial attributes like household income, net worth, Gender, housing attributes, languages, military and martial status
Business Information	760	Company age and size, employment status, Groups (e.g. High income or "decision maker"), Industry & Occupation (199 subgroups), Roles (Manager or Board Member), sales volume.
Branded Data	54176	Acxiom (1499), AddThis (1138), AdAdvisor bei Neustar (1082), IRI (1532), IXI (262), Lotame (1196), Mastercard (267), Navegg (868) Specialist Marketing Services (764),TruSignal (1020) ... 86 companies in total
In- Market	3264	Autos, Consumer Packaged Goods, Education, Financial Products, Real Estate, Retail, Services, Travel (Departure and Destinations, Car Rentals and Lodging Information)
Interest Categories	786	Animals (9), Arts & Entertainment (49), Autos (299), Business & Finance (12), Education (12), Food & Drink (28), Health, Beauty & Style (13), Hobbies, Games & Toys (17), Home & Garden (7), Internet & Online Activities (13), Lifestyles (16), News & Current Events (9), Other Vehicles (11), Parenting & Family (9), Personal Finance (19), Politices & Society (20), Science & Humanities (13), Shopping (25), Sports & Recreation (44), Technology & Computers (57), Travel (51), Video Games (31)
Mobile	367	Platform, Carrier, Genres, and Games
Past-Purchases	619	Travel, Services, Retail, Consumer Packaged Goods, Financial Products, Vehicles
Predictors	52	Autos, Retail Travel
Television	371	Shows, Viewing Frequency and Time

Table 1: Summary of the Bluekai taxonomy; "Branded Data" row lists a subselection of companies, number of categories in brackets

Consequently user-oriented studies have shown that regular internet users have a hard time to understand the information provided and asses the privacy impact of their profile [41].

4 METHODOLOGY TO STUDY PROFILING

To study profiling in an automated fashion requires the generation of web traffic that is similar to real users browsing behaviour. As described above many previous studies used toplists provided by Alexa to generate this traffic, but this approach has some limitations because of which we decided to use Reddit as a source of links connected to real users interests.

4.1 Website Categories

We first replicated the approach of previous studies that used the rankings published by Alexa. We visited multiple sites of a single category[9] within one session, concluded with a visit to Bluekai's registry. The 22,130 sessions created with this method consisted of 7.41 urls in average and 808 (3.65%) of these resulted in a profile being shown on the registry. Profiles showed up to 16 interests with two being the median. Due to these low rate of profiles we decided to evaluate a different approach and found Reddit to be a good source as it offers a glimpse at websites a user has visited, without knowing his or her whole browser history. We constructed browsing "histories" based on the websites a user has posted and additional links posted to the same interest subreddits.

4.2 Using Data from Reddit

Reddit is a free online service where users post and discuss links organized by topics (subreddits). While the majority of users posts only a few links and comments [23] to a small number of sites [47], there are others that use the service more actively, highlighting the websites they visit and discussing topics they find interesting. We use data from these users to construct lists of websites that are likely to be visited by a person. To do so we made use of Reddit's public API[10] for our selection process by requesting random posts, extracting the user that made that post, and then request this user's posting history going back up to a 100 entries - the maximum provided through the API. To single out those that use Reddit to discuss topics related to their broader personal interests, we excluded post histories that primarily listed pictures and videos to a limited number of sites. The selection criteria were that the requested history: (a) contained at least 80 links, where these links

[9] See http://www.alexa.com/topsites/category

[10] A detailed description of the API is available at https://www.reddit.com/dev/api

Figure 1: Screenshot of the Bluekai registry available as of December 2017 https://datacloudoptout.oracle.com/registry/

pointed to (b) at least 40 different subreddits and (c) at least 70 links did not contain blacklisted words (to exclude heavily meme related content)[11]. About 3% of the user accounts matched these criteria, and we created a set of nearly 8,000 users that, in average, posted 99.81 (standard deviation sd = 1.48) links to 20.82 (sd=13.07) domains in 55.32 different (sd=9.90) subreddits.

To see how different our sample of users is from the general Reddit population, we compared how often the most read subreddits (by subscribers) occurred in our set. The Reddit community as a whole seems most interested in community related issues, followed by entertainment, and news[12] But we also found topics that cause more discussions (/r/environment and /r/politics) in the list of subreddits that the users we studied posted to. Subreddits related to visual entertainment (movies, videos, pics) were similarly ranked in the top 20, while music, sports, and gaming were less popular in our subset than the general Reddit audience.

Our selection process successfully reduces the amount of users primarily involved in Reddit community issues and entertainment, which was our primary concern to be able to create sessions covering a wider range of topics. The high ranking of subreddits related to science and technology supports the notion that our subset does not reflect the general internet population, though this does not influence the fact that this group is profiled as such and our approach can be expanded to different data sources in future work.

[11]URLs were omitted when they contained the strings imgur, giphy, youtube.com, youtu.be, self., redd.it, reddituploads.com, reddit.com, me.me, meme, .gif, .pdf, .jpg , .jpeg

[12]Based on http://redditlist.com/ the top ten subreddits by subscribers are 1. AskReddit, 2. funny, 3. todayilearned, 4. science, 5. worldnews, 6. pics, 7. IAmA, 8. gaming, 9. videos, 10. movies. The subset we sampled is more involved in issues centered around news as well as science and technology as the top subreddits by number of posts from our data set are: 1. worldnews, 2. news, 3. politics, 4. technology, 5. science.

4.3 Constructing browsing sessions

As described above Reddit's API allowed us to request only 100 posts per user. To better observe tracking and profiling on long browsing sessions we leveraged the subreddit structure of Reddit to imitate web browsing behavior as it happens outside of closed platforms. We assumed that the 40 to 60 subreddits users have posted to reflect a reasonable amount of their actual interests. Therefore, we collected additional URLs posted to the same subreddits by different users, assuming that it is likely that a user, who has posted to a subreddit, also visits the links others post there. When selecting additional links from subreddits to extend browsing sessions, we maintained the same proportions, e.g. if someone has posted 30% of links to "news" and 70% to "anarchy" we selected 30 and 70 new links from these subreddits to continue the session. To do so, we continuously monitored all subreddits in our set and stored any new link that appeared. In addition we automatically stored new links posted to the front page of Reddit (those lists did not overlap) and created a list of roughly 70 million links from which we selected to increase session length.

Our approach allowed us to extent sessions to up to 3,000 URLs which, taking into account previous work, accounts for four to six months of a person's internet traffic, which is also the retention time for Bluekai's cookies. In 2006, a study on web browsing behavior of 25 persons reported an average of 55 website visits per day which resulted in the visit of 390 different domains on average over the course of 105 days [56]. More recently, in 2012. Goel et al. reported 5100 page views per user per year, resulting in an average of 425 page views per month [24]. While the number of websites visited per user per day has most likely increased over the past few years and individual usage patterns vary, the increased internet use does not necessarily mean that the number of different sites visited increased. The more time is spent online, the more it is also spend on the same websites[24, 57]. Our way of constructing sessions reflects this fact. While in the links originally posted to Reddit by a user 20% of the domains make up about 47.9% of the links this factor increases to 58% for constructed session of 1000 URLs, meaning the diversity of websites is decreasing the longer a session is.

4.4 Simulating a web session

To automate a web session, we used a method similar to Engelhardt et al. [17] and deployed selenium webdriver[13], a tool to automate web browsers, in combination with Firefox Extended Support Release (based on in Firefox 45). The systems were set up on multiple distinct Ubuntu/Debian-Linux virtual machines located in the United States. Results were stored in a central mongoDB database server. We collected data between October 2016 and March 2017. A subset of these systems also made use of the browsermob proxy[14] to monitor all requests that were made during a session in order to assess the amount of third party tracking that occurred, we also simulated scrolling when a website was loaded to measure possible effects of the interaction.

Each machine called one website after another based on lists created as described above. After every 100 pages, the browser was

[13]http://seleniumhq.github.io/

[14]https://github.com/lightbody/browsermob-proxy

directed to the Bluekai registry page where the profile was automatically extracted and, after the images were processed with OCR software[15], stored in the database. The OCR process only produced minor but consistent errors, for example detecting a vertical bar instead of a capital I.

The setup has some limitations that resulted in sessions being aborted. First, there are technical limitations related to the automated browsing process that cannot react to certain events, like requests that do not load correctly, scripts that produce to much load or crash the browser, as well as non-standard dialogs in pop-ups or alert messages. While 2.2% of website were not fully loaded before the timeout that we had set to 30 seconds, other errors lead to a crash of the browser destroying the session profile created so far. As a result, the number of data points we could gather for longer sessions decreased.

Second, sessions ended before a specific number of sites were visited because our database lacked additional URLs to extend the session. Though users posted to an average of 55 subreddits, most of these subreddits were not very active. Our strategy to add additional URLs from the same subreddits to extend the session was therefore limited by the number of additional URLs available for each subreddit. Especially for less busy subreddits, there were less additional links, while those with a large number of frequent posts (e.g. /r/worldnews) continuously provided new data to continue sessions. For sessions with up to 2,000 URLs, our database contained new links for 75% of the subreddits the user had posted to. Sessions were ended when our data set returned less than 5 URLs we had not visited before, which lead to an average session length of 1,100 URLs and the cutoff at 3,000.

5 ANALYZING PROFILES

In line with previous work we found Bluekai's tracking scripts to be present on a rather small number of websites, but their practice of sharing data with other services allows them to extend their reach and create profiles of 7.57 interests in average being assigned in browsing sessions of 800 URLs. Our study of longer sessions shows that these profiles are inherently noise and the assigned interests change over time.

5.1 Extent of Tracking

As described above, we monitored all requests to third party services in a subset of our long running sessions (69 sessions visiting 103274 URLs). Figure 2 shows the percentage of a browsing session that could be observed by a selected list of tracking services. The list is based on the cooperation partners that are mentioned in Bluekai's taxonomy (see Table 1) and compared to big competitors in the online advertising market that are not part of Bluekai's network, namely Google, Facebook, Krux and Yahoo. To map third party domains to tracking company, we used the NAI and DAA websites or individual company websites to draw conclusions on the relation between tracking domains and companies.

Again we found that Bluekai itself is only able to directly track on a limited number of sites. But several Bluekai partners like AddThis (also owned by Oracle), Lotame (tracking domain: crwdctrl.net) and Liveramp (owned by Axciom, rcdntl.net) are each able to track more than 20% of a session. Assuming that those companies are able to share information (e.g. by cookie syncing or direct data exchange between the servers) the conglomerate would be able to track nearly half of each session (48%). For comparison figure 2 also shows the numbers for selected competitors to Bluekai. While google (with google analytics, 71%) and facebook (through the like button, 74%) are leading the tracking services, Krux, a direct competitor to Bluekai, is able to track approximately the same amount of websites (50%).

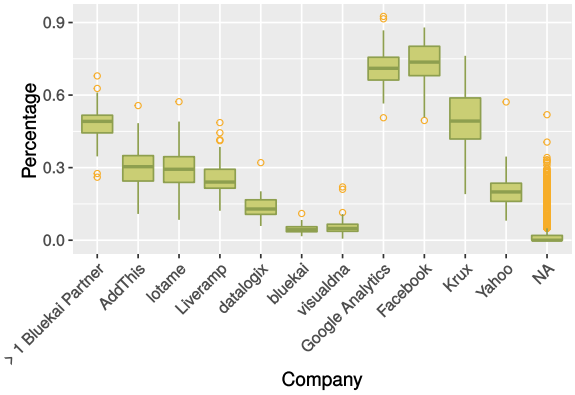

Figure 2: Box-and-whisker plot with outliers. Percentage websites in sessions individual trackers can observe. ">1 Bluekai Partner" = websites that include at least one Bluekai partner. "NA" = smaller tracking services.

5.2 Profiles in Long Browsing Sessions

In total we visited the registry over 85,000 times and observed 3,772 different attributes over all the categories. After geographic and device information, which are independent from the visited websites, the interests were most often reported (47% of all profiles). Within the interest reports we observed a large variety, 450 out of the 786 possible interested were assigned at least once. In the category *Branded Data* (What Others Know About You) that was part of 42% of the measured profiles, we observed 2724 attributes. Figure 3 shows how many of the top level categories of each attribute were observed, compared to the session length. Because of the richness of the reported interests we focused on this category for further analysis.

Our data confirms that the more websites are visited, the more interests are added to a profile. There is a significant ($p<0.001$) and strong (person-r = 0.55) correlation between these two factors. Although, as figure 3 shows, the number of interests increases at a greater rate for the first 500 websites it stays relatively stable afterwards.

Figure 3 and the following graphs show the number of top level *interest categories* present in a profile rather than the actual number of interests. While there are 768 interests of which e.g. after 800 URLs 7.57 were assigned (sd=7.69) on average. The total number of available interest categories listed in Table 1 is 22 of which a

[15] We used Tesseract OCR https://github.com/tesseract-ocr/tesseract

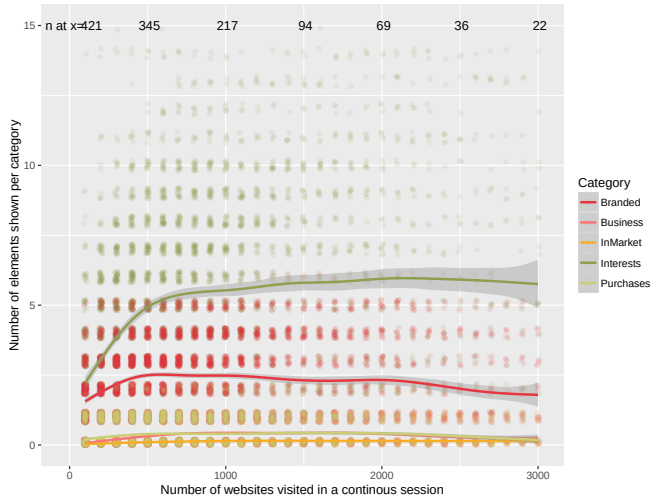

Figure 3: Number of attributes in different categories in relation to the session length. In this and the following charts dots represents a single measurement (with a small jitter to indicate the number of measurements), numbers on top show the number of sessions that reached a specific session length, lines shows linear smooth between averages at each step, grey area represent the .95 confidence interval.

maximum of 15 (sd=3.22) was assigned. Figure 3 also shows the observed profile measured approximately every 100th URL. For some sessions where measurements were done more frequent we omitted the measurements in-between.

To make sure the frequent visits of the registry page do not influence the outcome - Bluekai could punish those observing their profile too closely - we also checked for a correlation between the average number of interests measured and the number of URLs visited between observations. We did so by analyzing all observations taken after the session length exceeded 1,000 URLs and found no significant correlation. Therefore we assume that visiting the registry more frequently does not lead to a different treatment by Bluekai.

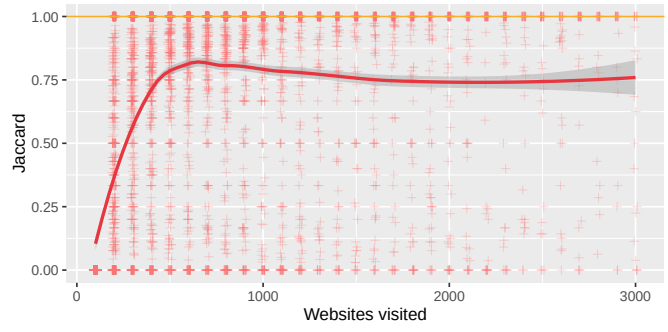

Figure 4: Jaccard Coefficient between two observations on interest categories. Changes are calculated between two steps in one session.

Looking at the relation between sessions and the number of assigned interests we found a significant ($p < 0.001$) correlation of 0.31 between the number of domains visited and the number of interest categories. At first it seems as if a profile stabilizes once a user has been tracked on a number of websites, as the average number of interest categories has a median of five. But a closer look shows that, within a session, the profiles still change. The overall average of interest categories for each session, counting the number of interest categories that were observed in a profile at least once, was 7.5 with a standard of deviation 3.8. To quantify the difference between two profiles we use the Jaccard index (1) that calculates the ratio between intersection and union of two samples.

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B} \quad (1)$$

Figure 4 shows the Jaccard index between the interest categories listed for two consecutive observations. Although the absolute number of assigned interests does not change much after 1000 URLs were visited the profile itself keeps changing, leading to an average Jaccard index of .75 between two consecutive observations. For the average of five interest categories this means that between two measurements about two categories change, either they are removed or added between observations.

Taking in to account the possibility that the registry might not display all available categories at each observation, we binned the interest categories observed in five consecutive steps, counting all interest categories that appeared once within five observations. While the Jaccard index is .67 when comparing observations 1 to 5 (100-500 URLS visited) with 6 to 10 (600-1,000 URLs visited) it raises to 0.77 when comparing observations 6 to 10 with 11 to 15 and increases to 0.81 in the next section. Afterwards it slowly decreases again to .72 and then .69 for session length longer than 2,000 URLs. We therefore consider the fluctuations within the profiles to be a feature of Bluekai's profiling that quickly adapts to recently observed traffic.

5.3 Reviewing the Profiles Created

To review what information exactly was generated about the sessions, we again looked at those sessions in which more than 1000 websites were visited.

Regarding the location, Bluekai positioned all sessions correctly in the category *country >United States*, where our servers were located. A slightly lower number of profiles (97%) reported correct information on the browser, browser language, operating system and device type. Regarding the "professional interests" for 43% of the sessions, a result in the category *industry and occupation* was returned. The majority of these (>90%) were categorized as *Software Designers & Programmers*, while only 2-3% were labeled as working in *Business & Finance*, *Human Resources* or *Sales*. Also, a number of attributes listed in the taxonomy as past purchases and called "things you might have bought" on the registry website were reported in 44% of the sessions, although we did not make any attempts to buy anything. Within this group the most reported past purchases were those of *Luxury Cars* (44%), *clothing* (38%) and *Spices and Seasoning* (25%). Demographics attributes were only reported for 31% of the sessions. Within this group *financial attributes* were reported frequently (64%), followed by gender (32%).

Among the 86 companies that share data with Bluekai, only a small number are included in the registry reports. Some companies simply indicate that they somehow contribute to the profile without any information. For example, Advisor by Neustar is reporting *element 090* as a uninterpretable profile attribute, Forbes lists the category *ads* and Profound assigns an attribute named *webanalytics* that also shows which users where tracked by other trackers. Other data contributors reveal additional information. For example Lotame (present in 40% of the profiles) made use of 51 subcategories while IXI categorized two sessions as detailed as *Economic Cohorts >50K−100K Income, Age, Retired (65+), Nest Egg Elders >Older Retirees*. Additional observed entries in *branded data* were found for 33Across, AddThis, Affinity Answers, Bambora, Conexity, Datax-pand, Datacratic, Skimlinks, VisualDNA, and Ziff Davis (all in less than 1% of the profiles).

The profiles created do not spread evenly across all interest categories. While the peak at *technology & computers* could very well be explained by the specifics of the Reddit audience, other simply show more common online activities, insofar they can be reflected as interests, like *News & Current Events* and *Arts & Entertainment* [61]. The data could also be interpreted in a way that Bluekai is not actually able to evenly measure web traffic, because only a limited number of websites that use Bluekai's tracking network. The comparison with interests measured in relation to sites categorized by Alexa shows a similar bias: although we visited 996 websites from the category "society", an interest in *Politics & Society* was never measured for these sessions.

5.4 Reproducibility of Profiles

Besides the general analysis of how profiles are created and how many items they contain, we also wanted to test if the results are reproducible. Previous work on Google's profiling [15] has shown on a smaller scale that the same website visits do not necessarily lead to the same profile created. Repeated visits of the same URLs only had an overlap of 60% for sessions covering 100 websites.

We define the reproducibility of an interest profile as the average Jaccard index of all indexes per observation in each session. Since the profile reveals most of its information in the "interest" category we only compare the interest profiles:

$$Reproducibility(A, B) = \sum_{\forall i} \frac{A_i \cap B_i}{A_i \cup B_i} / i \tag{2}$$

For Bluekai we repeated 160 sessions. We visited the same URLs in the same order and profile was requested at the same point in the browsing session. Here the Jaccard index was only 0.51 on average, meaning that the profiles resulting from the same web site visits overlapped only by about 50%. A comparison of any two profiles has an average Jaccard of 0.28, that increases to 0.35 for sessions longer than 1,000 URLs.

To learn more about why repeated visits lead to different profiles, we tested two assumptions. First, the shorter the time between two sessions, the closer the profiles. And second, the longer the sessions are, the more likely it is that the profiles overlap.

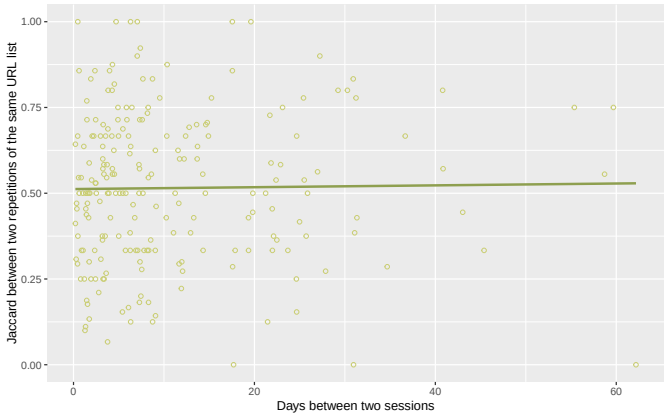

Figure 5: Comparison of two runs in relation to the time between them

5.5 Timing and Session Length

One assumption is that the Jaccard index decreases the more time passes between two sessions with the same URL list. The assumption is based on the knowledge that online advertisement industry makes use of real time bidding where advertisers and tracking services quickly decide who will place an ad on a website. This results in different tracking services being present on the same site on different visits. Since timing is an important issue (advertisers wants to show ads related to current needs, not to those that are probably already fulfilled), one could assume that websites that have "aged" are less interesting and therefore other advertisers and trackers might be present on a website, resulting in a different profile. Our data does not support this hypothesis. We found no significant correlation between profile reproducibility and the time between two repeated session, when looking at repetition of sessions after some time, see fig. 5 for details.

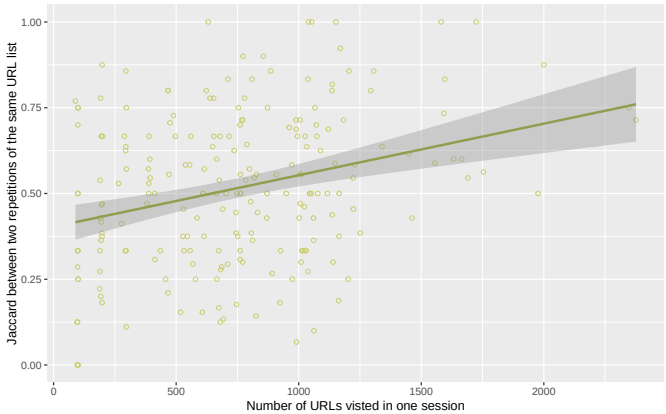

Figure 6: Comparison of two runs in relation to number of URLs visited

Another assumption could be that longer session result in a higher reproducibility. We found a slight, but significant ($p < 0.001$)

positive correlation (pearson r = 0.29) between the length of a session, i.e. the number of URLs visited) and the reproducibility of a profile, though the variance is very high and for the longest sessions with more 3,000 URLs the Jaccard index was still about 0.75 (see fig. 6). The reproducibility of a profile improves the more websites are visited, meaning the more visits can be tracked.

6 INFLUENCING A PROFILE

In addition to analyzing the profiles for longer browsing sessions, we wanted to explore if a profile can be influenced for the purpose of obfuscation as theorized by Brunton and Nissenbaum. In contrast to just blocking trackers and opting-out of profiling, the purpose of obfuscation is to blend in [8, 9]. It is a technique that allows participation in a system without standing out using *expressive privacy* [28]. Instead of being detectable as a privacy aware user that e.g. uses an ad-blocker, obfuscation would make a privacy aware user not stand out but be as trackable as everyone, while still protecting privacy by adding noise to the data. Mechanisms could be used to create specific profiles for specific purposes, modify an existing profile or broaden the profile to render make it (more) inaccurate.

We tested two obfuscation schemes, modifying an existing profile (by adding dummy traffic once, see 6.2) and broaden a profile to render it inaccurate (by continuously adding dummy traffic, see 6.3).

Obfuscation has been successfully implemented for web search with TrackMeNot [27, 52], although there are challenges regarding the theoretical attack of reidentifying those that make use of obfuscation tools [3, 12]. In contrast to TrackMeNot, which can only make assumptions about the capabilities of the attacker (in this case Google) to counter obfuscation, we reviewed the profiles that are created at Bluekai through the registry and use it as feedback for our obfuscation mechanisms.

Our goal was to obfuscate interest categories of profiles in comparison to our baseline described above (see fig. 3). We tested two ways to obfuscate a profile. First we added additional *dummy traffic* at a given point and second we recursively added a smaller amount of traffic every 100 sites. In both cases, 50 URLs were used to obfuscate the profile, either at one point in the middle of the session, or spread over the second half.

To measure the effect of obfuscation against a more or less stable profile, we started adding dummy traffic after 1,000 websites were visited as described above. At that point, the average Jaccard index between two steps was around .75 when comparing single observations and .88 when comparing multiple observations before and after 1,000 URLs.

6.1 Generating Obfuscation Traffic

To created dummy traffic we used a number of websites which we had found to have a distinct impact on profile, although the previously described profiles can not be explained just based on visits to these sites. We created this list by randomly selecting URLs from our data set of links posted to Reddit, visited them one by one in individual sessions, and checked whether the Bluekai registry showed a profile just based on one website visit. If a profile was created, we repeated the visit of that website in five additional sessions to confirm the result. The average Jaccard index of five visits to one of these pages was 0.91.

Of the 7,692 different domain checked, only 2.2% (170 in total) led to an observable interest profile. The interests observed where part of 12 different interest categories. While the profiles created over multi-page sessions were subject to constant change, these one-page session profiles were rather stable. One page visit lead to an average of 1.14 interests (see fig. 10). Mismatches between multiple sessions occurred mostly because for 30% of domains one of the re-visits did not result in an observable profile at all. Only for 2.5% of domains different interests were assigned.

6.2 One time obfuscation

With the first obfuscation scheme we tried to measure the impact of introducing dummy traffic at a specific point in a session. Comparable to a user that consciously decides to alter a profile. To do so we used the same methods described above and checked the current interest profile after every 100th website visit. Dummy traffic was added after 1,000 URLs were visited since the majority of profiles had stabilized in size at this point. The dummy traffic was selected so that only websites that were associated with interests not present in the last observed profile were used for obfuscation. After visiting these URLs, the session continued as before, selecting URLs from subreddits to which the original user had initially posted to.

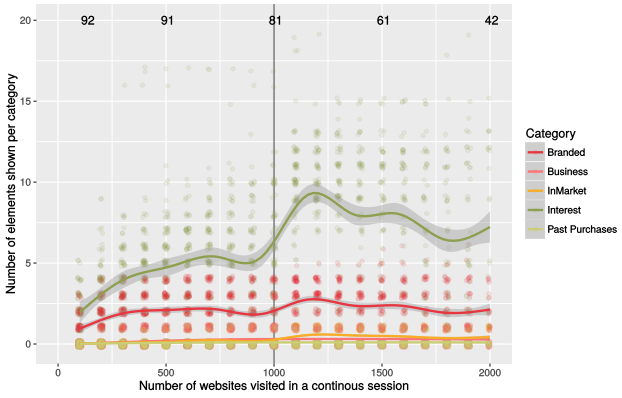

Figure 7: Average profile breadth with one time obfuscation (added 50 URLs) after 1000 site visits

Figure 7 shows how the profile changed over the course of the session. The impact of the dummy traffic is immediately effective, but the effect starts to wear off over the rest of the session, at least with regard to the number of interests added to the profile. While each session had a median of four interest categories assigned before the obfuscation, this increased to 10 after dummy traffic was added. After visiting another 950 websites related to the original profile, the profile breadth decreases to five.

Comparing the similarity of the profiles measured, the average Jaccard index is .64 between 900 and 1,000 URLs and drops to .46 after dummy traffic is added. The Jaccard index decreases as the the number of interests, influenced by dummy traffic, increases. Interestingly, although the number of assigned interests decreases

when the sessions continue to browse websites related to the original profile the assigned interests do not return to those seen before dummy traffic was added. Figure 9 shows that the Jaccard index between profiles measured at the beginning and end of the sessions is just 0.53.

The influence on the rest of the profile differs between categories. While the list of secondary sources ("Branded") and assignments to market segments ("InMarket") spiked in parallel to the list of assigned interests, the list of past purchases and business categories did not change.

6.3 Frequent obfuscation

To continuously influence a profile we again checked the current interest profile every 100th website visit. We extended the session by selecting links from the same subreddits as the original set until 1,000 websites were visited. After that, the obfuscation algorithm selected five URLs that have been shown to lead to interest categories that were not part of the profile observed so far. Afterwards another 95 URLs, related to the original set, were visited. This process was repeated up to 10 times.

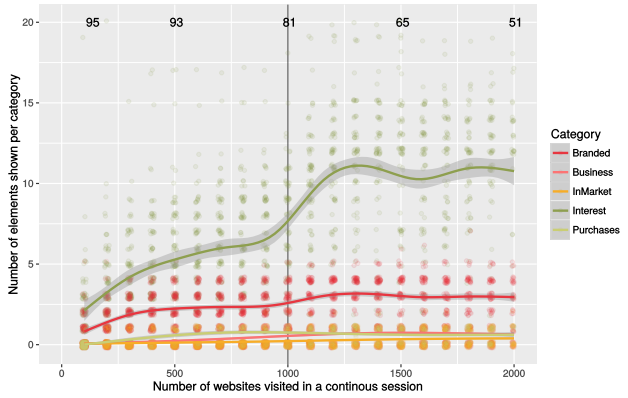

Figure 8: Average profile breadth with recurring obfuscation (visited 5 URLs after each 100) starting after 1000, n=51

Figure 8 shows how the number of interest categories increases and lead to a median of 7 out of 22 interest categories over the second half of the session, or 14 different interests in total. The similarity between the profiles is reduced from an average of .73 before the first dummy traffic, to .68 after the first dummy traffic is added, and is .53 on average for the second half of a session compared to the observation before the obfuscation.

Again, the influence on other profile categories differs. The continuous obfuscation increases the secondary data ("Branded") as well as the average amount of information given in the "In Market" and "Business" categories.

6.4 Comparison of Obfuscation Schemes

Figure 9 shows the effect of the obfuscation strategies on the Jaccard coefficient within a session. It emphasizes the effect by comparing all interests observed over a longer span of each session, reducing the effect of inner session noise. The strategy of adding a larger amount of dummy traffic once is more effective with regard to this metric. It introduces a large number of new interests, changing the profile that then stays changed although the session continues to browse websites that support the original profile. Frequent obfuscation performs less well here as it might reintroduce interests that might have already been pushed out of the profile.

Figure 10 shows a summary of different session types we tested with regard to the absolute number of interest observed, in contrast to the interest categories used before, over the whole session length. After browsing 2,000 websites related to the same topics on Reddit, the average profile reaches a ceiling at about five interests, though over the whole session about seven interests where observed. These averages were consistent between initial as well as repeated session in which the same URL lists were revisited in new sessions.

Comparing the obfuscation strategies, the frequent obfuscation leads to 14 interests in 7 of the 22 interest categories to be associated with a profile over the course of 2,000 website visits. Adding dummy traffic only once per session also increases the number of interests observed within a session, but the median of 12 is slightly smaller.

Taking both metrics into consideration, it seems like one time obfuscation is effective in *altering*, while frequent obfuscation helps to *anonymize* an interest profile. The first could therefore be useful in cases were a user wants to interact with the (advertising) system in a specific way, e.g. see if different prices are offered to users with different interests. The latter option is useful for those who want to hide their real interests within the noise of an interest profile that has a large number of interests.

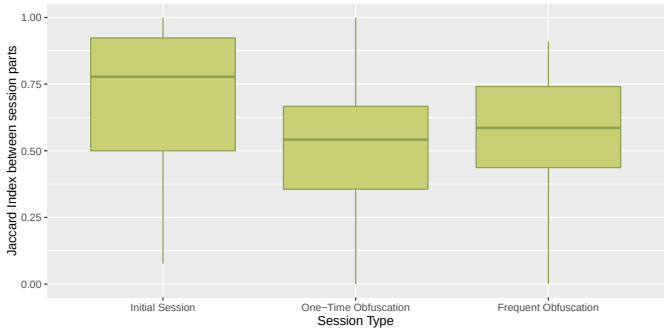

Figure 9: Change in profiles before and after obfuscation, when comparing profiles created after 500 and 1000 visits with those after 1,500-2,000.

6.5 Obfuscation and Mitigation

Both ways of obfuscation could be mitigated by Bluekai, but would require different strategies on their end. The frequent obfuscation requires adding 5% of traffic as dummy traffic. As described by Balsa et al. [3] it is theoretically possible to identify the real profile within the larger, obfuscated profile, that is to single out those traffic that is added as noise. During our study this did not happen, at least not in a way that we were able to observe. Moreover, since tracking services are not able to track 100% of website visits it might be hard to assess when the introduction of dummy traffic starts and ends.

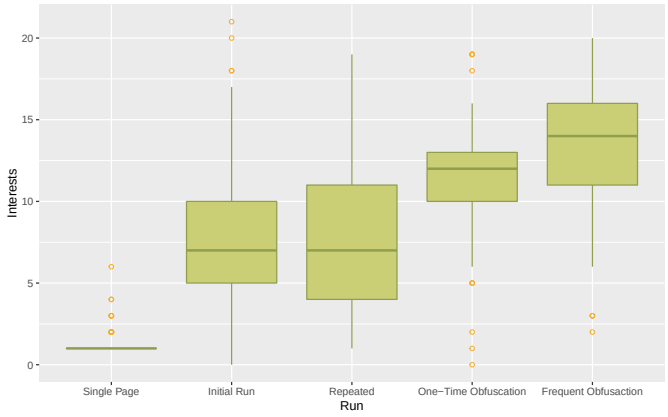

Figure 10: Comparison of the number of all observed interests per session type measuring sessions with >1,000 URLs per session (except single website visits). Single Page (n=170), Initial run (n=210), Repeated (n=92), One-Time Obfuscation (n=93), Frequent Obfuscation (n=92)

Regardless, there are multiple ways to improve the obfuscation process. First, the dummy traffic was introduced at a specific point in the browsing session (after 1,000 URLs), doing this in a randomized way could reduce the likelihood of identification. Second, for the purpose of the study, we started with creating a non-obfuscated profile. This "original" profile could be used later on to re-identify the real profile within the obfuscated one. Users that want to create specific profiles could counter this by making sure all tracking cookies are deleted or start with a fresh browser profile that includes obfuscation traffic from the beginning.

In addition, there are measures profiling services could take to identify obfuscation traffic as we created it. First, we did not interact with the websites, and even if implemented as a browser plugin, it might be difficult to reliably imitate real users. Second, online profiling services can use existing data to identify abnormal interest combinations and weight those less than combinations that are frequently seen within the user base. While this could reduce the effectiveness of obfuscation, it would also decrease the effectiveness of the advertising network to adapt to new trends. It would support stereotypes and might in the end not serve the interest of the advertisers. Though the idea of personalization is to reduce the audience by targeting only those identified as relevant, reducing it too far also hurts the business model. Therefore a (limited) number of falsely targeted users is preferred over not targeting users that actually demonstrate an unusual combination of interests.

7 LIMITATIONS

Besides automated studies of web tracking a number of qualitative studies has been looking at real users' browsers and browsing histories. While qualitative research can be more accurate on the individual level and offer a broader perspective especially when it comes to the perception of profiling by users (see [13, 41, 45]), automatic techniques are favored in many studies that try to measure online tracking on a larger scale. They allow researchers to examine a large number of websites with limited resources and therefore leads to quantitatively analyzable data. And even studies that look at real users' web browsing histories are limited in reflecting actual online behavior. Not only because users may use multiple devices, or change their behavior while being studied but also due to ethical restrictions on these studies. In a recent study participants were allowed to delete websites they had visited from the data they shared with researchers for ethical reasons [33].

Still, there are limitations to not study real users' full browsing histories when looking at profiling, but the same limitation applies to the tracking and profiling services themselves as they are only able to observe and create transaction or role profiles [15] based on the traffic they can observe, rather than full personal profiles. Google, for example, shows multiple profiles per user. One is based on tracking with Google Analytics on third party websites, the other is a result of search queries and does not account for the traffic websites users go to directly, have bookmarked, or visit because they received a link through a different channel. The profile Facebook creates from interactions within the platform focuses on what users like and share, which is limited to what people want others to see (e.g. Facebook might see less traffic to websites about very personal or health related issues). Even Internet Service Providers, who can observe all traffic related to a specific IP-address, are limited in their tracking abilities, because multiple devices, changing locations, networks, and addresses, as well as technologies like ad-blockers and anonymization networks, limit what traffic they can measure.

The lists we created from Reddit data has some drawbacks, too, as links publicly posted to any platform do not reflect the whole breadth of things someone does online. First, we know from previous studies and webmetrics (see 2.3) that users spend most of their time in online social networks or other platforms that offer search or emailing services. Our data set intentionally avoids most of these closed data sources, since social media websites like facebook, search websites, and related portals are often closed platforms where tracking is conducted and kept by the platform owner instead of third-party services like Bluekai. While this potentially helps to maximize their revenue, from the standpoint of our study there is no benefit in reflecting this proportion in our data set. To measure the profiles created by third party tracking services it is necessary to visit websites where these a present. Second, Reddit users do not represent a general population. While Reddit has over 230 million users worldwide, the majority are based in the US[16]. On the level of other demographics, Reddit users were reported to be mostly male in 2013 [39], but in 2015 Reddit itself claimed to have a nearly equal split of 54 to 46% (male/female). Information on Alexa.com also suggested that the general Reddit audience is slightly better educated than the average internet user[17].

Additionally, automated browsing does not impersonate a real user in all possible ways. It browses the web faster than a human user would, and waits 20 to 30 seconds to load a page, regardless of the content, while the majority of users spend less than 20 seconds on a website according to Nielson [34]. Our browser also did not interact with the websites. Though this might be a factor to detect bots, it is not a sufficient criterion, since a real internet user might

[16] See https://reddit.zendesk.com/hc/en-us/articles/205183225-Audience-and-Demographics
[17] See http://www.alexa.com/siteinfo/reddit.com

also only visit a page for a short period of time without interacting with it. In additional tests we did not find any statistically significant effects of our continuous web surfing, time of day, the amount of time we spend on a website, or additional interaction, like scrolling on the measured profiles.

It is also possible that our tests were detected and the profiles that were shown were altered to mislead us. While this is theoretically unavoidable in a study of a black box, we found no indications for this behavior. The only protective mechanism we observed is that the registry sometimes became unavailable when it was visited too frequently, we did not see a deterministic effect, or explicit blocking of our servers. Instead, the fact that online advertisement fraud [36] - setting up fake websites and automatically clicking ads - is still a big problem for the industry, supports our assumption that detecting bots is very difficult and the risk of falsely identifying real users as bots might be a higher threat to revenue than the costs of fraud are.

8 CONCLUSION

Profiling internet users and assigning categories is crucial to today's internet marketing business. Thousands of companies engage in tracking and sharing user data, while sophisticated algorithms try to estimate a large number of details about a user based on which websites they visit. Although profiling often happens without knowing the actual identity of a user, the *panoptic sort* is part of a surveillance and power structure that threatens privacy and autonomy. Therefore, it is important to learn more about how profiling works, what information is created and how it can be influenced.

In our study, we simulated web sessions of up to 3,000 URLs. We used these sessions to analyze profiling as a black-box by looking at Oracle's Bluekai Service. We showed that, though 45% of a session is tracked by at least one partner in Bluekai's network, they do not assign the same profile to sessions that contain the same websites. While this bad recall might not harm the business model of Bluekai or behavioral advertising in general, we think it is an important insight for those that are worried about their privacy online. It could be considered a good sign and support the view that profiles do not represent real users, previous studies have shown that people are often surprised by this result [41]. It can also be worrying with respect to the increasing impact of profiles that are created based on browsing behavior have (e.g. on prices) or might have in the future as these effects are based on inaccurate measures.

We also tested two strategies to obfuscate interest profiles created by Bluekai. One strategy introduced dummy traffic only once per session, while the other added a smaller number of targeted dummy traffic 5 times per session. While both strategies were able to reduce the similarity between original and obfuscated profiles, the recurring obfuscation strategy was also able to double the number of interest categories per profile.

Our research has implications on future work on improving obfuscation strategies and reducing the effectiveness of the mitigation strategies described in section 6.4. We want to analyze our data set with regard to possible interest clusters and fine-tune the obfuscation strategies to support change in assignments based on these clusters. Focusing on user interaction might also proof valuable to evaluate what, if any, strategies internet users prefer with regard to obfuscation and what possible risks are. Since online tracking and profiling are not likely to disappear, there is a need for new paradigms that support users in managing the profiles created about them.

REFERENCES

[1] 360pi. (2014) Approaches to Price Dynamism - When should you change your prices?.
[2] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," *Proceedings of the 21st ACM Conference on Computer and Communications Security*, 2014.
[3] E. Balsa, C. Troncoso, C. Diaz, "OB-PWS: Obfuscation-Based Private Web Search," *IEEE Symposium on Security and Privacy*, doi:10.1109/SP.2012.36, 2012.
[4] M. Barbaro, T. Zeller. "A Face Is Exposed for AOL Searcher No. 4417749," *The New York Times*, 2006
[5] M. Bashir, S. Arshad, W. Robertson, C. Wilson, "Tracing Information Flows Between Ad Exchanges Using Retargeted Ads," *25th USENIX Security Symposium*, 2016.
[6] BlueKai Taxonomy Report (2016) [online].
[7] K. Brookman, P. Rouge, A. Alva, C. Yeung, "Cross-Device Tracking: Measurement and Disclosures," *Proceedings on Privacy Enhancing Technologies*, 2017.
[8] F. Brunton. H. Nissenbaum, "Vernacular resistance to data collection and analysis: A political theory of obfuscation," *First Monday* vol. 16, 2011.
[9] F. Brunton, H. Nissenbaum. "Obfuscation: A UserâĂŹs Guide for Privacy and Protest," *MIT Press*, 2015.
[10] W. Christl, S. Spiekermann, "Networks of Control," 2016 [online].
[11] Cliqz/WhoTracks.Me. "June Update - Do You Consent?" 2018 [online] .
[12] F. Dankar, K. El Emam, "A Theoretical Model for Obfuscating Web Navigation Trails," *EDBT/ICDT 2013 Workshops*, doi:10.1145/2457317.2457341, 2013.
[13] A. Datta, M.C. Tschantz, A. Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination." *Privacy Enhancing Technologies*, doi:10.1515/popets-2015-0007, 2015.
[14] M. Degeling, T. Herrmann, "Your Interests According to Google - A Profile-Centered Analysis for Obfuscation of Online Tracking Profiles," [online], arXiv:1601.06371 [cs], 2016.
[15] M. Degeling, "On The Vagueness Of Online Profiling," *Profile, Predict, Prevent. Blockchain Workshops*, 2015.
[16] S. Englehardt, D. Reisman, C. Eusbank, P. Zimmermann, J. Mayer, A. Narayanan, E. Felten, "Cookies That Give You Away: The Surveillance Implications of Web Tracking," *24th International Conference on World Wide Web*, doi:10.1145/2736277.2741679, 2015.
[17] S. Englehardt, A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," *ACM CCS*, 2016.
[18] H. Elmeleegy, L. Yinan, Q. Yan, P. Wilmot, and W. Mingxi, "Overview of Turn Data Management Platform for Digital Advertising," *VLDB Endow*, 2013.
[19] D. Evans, "The Online Advertising Industry: Economics, Evolution, and Privacy," *Journal of Economic Perspectives*, vol. 23, no. 3, pp. 37âĂŞ60, 2009.
[20] O. Gandy, "The Panoptic Sort - A Political Economy of Personal Information," *Westview Press*, 1993.
[21] A. Danna, O. Gandy, "All That Glitters is Not Gold: Digging Beneath the Surface of Data Mining," *Journal of Business Ethics*, vol. 40, p. 373âĂŞ386, 2002.
[22] J. Gomez,T. Pinnick, A. Soltani, "KnowPrivacy", *UC Berkeley, School of Information*, 2009.
[23] E. Gilbert, "Widespread Underprovision on Reddit," *Proc. of the Conference on Computer Supported Cooperative Work*, doi:10.1145/2441776.2441866, 2013.
[24] G. Sharad, J. Hofman, and M. Sirer, "Who Does What on the Web: A Large-Scale Study of Browsing Behavior," *ICWSM*, 2012.
[25] A. Hannak, G. Soeller, D. Lazer, A. Mislove, C. Wilson. "Measuring Price Discrimination and Steering on E-commerce Web Sites," *Proc. of the IMCâĂŹ14*, doi:10.1145/2663716.2663744, 2014.
[26] S. Gutwirth, M. Hildebrandt, "Some Caveats on Profiling," *Data Protection in a Profiled World*, p. 31âĂŞ41, Springer, 2010.
[27] D.C. Howe, H. Nissenbaum. "TrackMeNot: Resisting Surveillance in Web Search," *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society*, 2009.
[28] D.C. Howe, "Surveillance Countermeasures: Expressive Privacy via Obfuscation," *Datafied Research*, vol. 4, no. 1, 2015.
[29] A. Karaj, S. Macbeth, R. Berson, and J. M. Pujol. "WhoTracks.Me: Monitoring the Online Tracking Landscape at Scale." ArXiv:1804.08959 [Cs], 2018.
[30] D. Kamerer, "Estimating Online Audiences: Understanding the Limitations of Competitive Intelligence Services," *First Monday*, vol. 18, no. 5, 2013.
[31] A. McStay, "The Mood of Information: A Critique of Online Behavioural Advertising,' *The Continuum International Publishing Group*, 2011.
[32] J. R. Mayer, J. C. Mitchell, "Third-Party Web Tracking: Policy and Technology," *IEEE Symposium on Security and Privacy*, p 413âĂŞ27, doi:10.1109/SP.2012.47, 2012.

[33] W. Melicher, M. Sharif, J. Tan, L. Bauer, M. Christodorescu, and P. Leon, "(Do Not) Track Me Sometimes: UsersâĂŹ Contextual Preferences for Web Tracking," *Privacy Enhancing Technologies*, doi:10.1515/popets-2016-0009, 2016.

[34] J. Nielson (2011, dec), "How Long Do Users Stay on Web Pages?" [online].

[35] L. Olejnik, T. Minh-Dung, and C. Castelluccia, "Selling Off Privacy at Auction," 2013.

[36] J. Pagliery, "Russian âĂŸMethbotâĂŹ Fraud Steals $180 Million in Online Ads." *CNNMoney*, 2016 [online].

[37] E. Pariser, "The Filter Bubble: What The Internet Is Hiding From You," Penguin , 2011.

[38] G. Papadakis, R. Kawase, E. Herder, and W. Nejdl, "Methods for Web Revisitation Prediction: Survey and Experimentation." *User Modeling and User-Adapted Interaction*, vol. 25, no. 4, p. 331âĂŞ69, 2015. doi:10.1007/s11257-015-9161-7.

[39] M. Duggan, A. Smith "6% of online adults are reddit users," *Pew Internet and American Life Project*, 2013.

[40] E. Rader and R. Gray, "Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed," *33rd ACM Conference on Human Factors in Computing Systems*, 2015. doi:10.1145/2702123.2702174

[41] A. Rao, F. Schaub, and N. Sadeh. "What Do They Know about Me? Contents and Concerns of Online Behavioral Profiles,", 2015. arXiv:1506.01675 [Cs]

[42] Z. Rodgers, "Google Adds Cross-Device Metrics To DoubleClick, Partially Answers FacebookâĂŹs âĂŸPeopleâĂŹ Power," *AdExchanger* [online], 17. June 2015.

[43] F. Roesner, K. Tadayoshi, and D. Wetherall, "Detecting and Defending Against Third-Party Tracking on the Web," *9th USENIX Conference on Networked Systems Design and Implementation*, 2012.

[44] V. Le Pochat, T. Van Goethem, and W. Joosen. "Rigging Research Results by Manipulating Top Websites Rankings." ArXiv:1806.01156 [Cs], 2018.

[45] B. Ur, P. Leon, L. Cranor, R. Shay, and Y. Wang, "Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising," *8th Symposium on Usable Privacy and Security*, 2012.

[46] S. Schelter, and J. Kunegis, "On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl," arXiv:1607.07403 [cs], 2016.

[47] P. Singer, F. FlÃűck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, "Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?" *23rd Int. Conference on World Wide Web*, doi:10.1145/2567948.2576943, 2014.

[48] E. Spyromitros-Xioufis, G. Petkos, S. Papadopoulos, R. Heyman, and Y. Kompatsiaris. "Perceived Versus Actual Predictability of Personal Information in Social Networks." *Internet Science. Lecture Notes in Computer Science 9934*, doi:10.1007/978-3-319-45982-0, 2016.

[49] L. Sweeney, "Discrimination in Online Ad Delivery," *Social Science Research Network*, 2013.

[50] T. Theodoridis, S. Papadopoulos, and Y. Kompatsiaris, "Assessing the Reliability of Facebook User Profiling." *24th Int. Conference on World Wide Web*, 2015, doi:10.1145/2740908.2742728.

[51] J. Turow, "The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth," *Yale University Press*, 2012.

[52] V. Toubiana, L. Subramanian, and H. Nissenbaum, "TrackMeNot: Enhancing the privacy of Web Search," arXiv:1109.4677, 2011.

[53] http://www1.icsi.berkeley.edu/ mct/pubs/TR-16-003.pdfM. Tschantz, S. Egelman, J. Choi, N. Weaver, and G. Friedland. "The Accuracy of the Demographic Inferences Shown on GoogleâĂŹs Ad Settings." International Computer Science Institute, Tech-Report, 2016.

[54] T. Vissers, N. Nikiforakis, N. Bielova, and W. Joosen, "Crying Wolf? On the Price Discrimination of Online Airline Tickets," HotPET Symposium (2014).

[55] M. Ward, "It Is Easy to Expose UsersâĂŹ Secret Web Habits, Say Researchers," *BBC News*, 2017.

[56] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer, "Off the Beaten Tracks: Exploring Three Aspects of Web Navigation." *15th Int. Conference on World Wide Web*, doi:10.1145/1135777.1135802, 2006

[57] Lilian Weng, "User Browsing Behavior in New Tabs of Firefox, " [online] 2013

[58] A. Karaj, S. Macbeth, R. Berson, and J.M. Pujol, "WhoTracks.Me: Monitoring the Online Tracking Landscape at Scale." ArXiv:1804.08959 [Cs], 2018.

[59] Z. Yu, S. Macbeth, K. Modi, and J. Pujol, "Tracking the Trackers," *25th Int. Conference on World Wide Web*, doi:10.1145/2872427.2883028, 2016.

[60] S. Yuan, J. Wang, and X. Zhao, "Real-time Bidding for Online Advertising: Measurement and Analysis," *7th Int. Workshop on Data Mining for Online Advertising*, doi:10.1145/2501040.2501980, 2013.

[61] K. Zickuhr. "Online Activities," *Pew Research Center*, 2010.