ON THE VAGUENESS OF ONLINE PROFILING

The internet is one of the most productive fields of experimentation with profiling and prediction techniques. Especially in online marketing it has become inevitable to track internet users and collect as much information as possible to profile (soon to be) customers for the purpose of predicting and influencing their behavior. A constantly growing number online tracking services pretend to help publishers to "understand" their audience and offer highly differentiated means for segmentation of web users while those affected by the profiles have no knowledge about their *data doubles*. In the context of this paper we understand profiling as the practice of the automatic collection of information about internet users from various sources by a third party that are then enriched by additional assumptions, mostly based on statistics. It has been argued that this conception of profiling has severe impact on conceptions of individual autonomy and privacy as it is understood in legislative regulation (Gutwirth and Hildebrandt 2010; McStay 2010).

Besides theoretical discussions we want to shed light on the daily practice of online profiling. We describe the findings of our study on the extent and form of interest profiles created by Google and other online tracking services. The results support the descriptions of profiling as part of the surveillant assemblage (Haggerty and Ericson 2000) since the profiles created represent internet users by multiple, diverse and unstable profiles. This leads us to the conclusion that we ne need to think about ways to interact with and use profiles to support the autonomy of users.

But while surveillance, especially one that is carried out by state agencies, has lately been in the focus of public discourse resulting in questioning these practices, the possibilities for marketing services to profile and predict consumers is ever extending and as strong as state surveillance. As (Morris 2012) has put it: "[T]he increased use of data analysis has been embraced by two parallel forces: marketing and surveillance". Especially online profiling turns out to be one of the major fields of experimentation about what is possible with regards to how profiling can be performed and how they can be used to nudge users and influence their behavior.

1. INTRODUCTION TO ONLINE PROFILING AND ITS CRITIQUE

From the beginning of the internet age advertising companies have been fascinated by the possibilities to track how readers react to advertisings on websites (Turow 2011). Previously, in offline advertising, marketers had to rely on the data provided by publishers about the size and socio-demographic averages of the audience. With the introduction of banner ads and online advertising marketplaces the marketing departments where abled to track how internet users interact with websites and ads - which ones they click when and what happens afterwards. Since then the technology, together with the number of internet users, has developed rapidly. Today cookie based tracking is enhanced by zombie cookies and beginning to be replaced by browser- (Boda et al. 2012) and canvas fingerprinting (Mowery and Shacham 2012). Tracking services can collect every mouse move and keystroke made on a website without notice and especially without consent. These interaction data is enriched with socio-demographic information and used to characterize individuals on various levels. Although these techniques are sometimes critically discussed in a broader public[1], only 15-30% of web users make use of ad blockers (Pagefair 2015).

The need to profile users and predict their behavior is related to the way online advertising is organized nowadays. The space for ads on websites is sold and purchased in a dynamic market referred to as *programmatic advertisement* combined with a strategy called *real-time bidding (RTB) (Yuan, Wang, and Zhao 2013)*. This inner-advertisement-economy is used to purchase advertisement space on multiple websites dependent on the (assumed) viewer, leading to websites not knowing which ad publisher will show an ad on space they offer. When a website is loaded the main advertisement partner of the website is offering the space for each specific request on an automated platform, he represents the *supplier side*. Space is offered together with information about the browser that is loading a website like which operating system it is running on, which screen resolution the monitor provides but also where the IP-address originates. This initial profiles are correlated with socio-demographic data, for example estimations about the average income of that area of the IP-address or the likeliness of the users of specific devices of having one or the other gender[2]. On the mentioned RTB platform those that want

1 For example when a study on self-censorship was release in cooperation with facebook (Das and Kramer 2013) were data was collected in users' browsers before they pressed send.

2 In the marketing world only two genders exist.

to bring ads of their clients (*demand side*) automatically bid on space/profile offers. To explain the idea behind this strategy (Singer 2012) cited from an investor presentation of Acxiom, an us-based marketing company:

> [Mr. Hughes] logs on to Facebook and sees that his friend Ella has just become a fan of Bryce Computers [...] When Mr. Hughes follows a link to Bryce's retail site, [...] the system recognizes him from his Facebook activity and shows him a printer to match his interest. He registers on the site, but doesn't buy the printer right away, so the system tracks him online. [...] while he scans baseball news on ESPN.com, an ad for the printer pops up again.
> [When returning] to the Bryce site [he is] then offer[d] a sweeter deal: a $10 rebate and free shipping.
> Correctly typecast, Mr. Hughes buys the printer.

In an explanation it is described that Acxiom has build "70 detailed socioeconomic clusters" and Mr. Hughes is characterized in those as a "'savvy single' — meaning he's in a cluster of mobile, upper-middle-class people who do their banking online, attend pro sports events, are sensitive to prices — and respond to free-shipping offers." (Singer 2012)

When it comes to possible impacts of these practices those most commonly addressed are *price discrimination* (Danna and Gandy 2002) and the possibility of a *filter bubble (Pariser 2011)*. The market-liberal idea of perfect price discrimination, sometimes also called *dynamic pricing*, is to be able to 'negotiate' the perfect price for each product in each transaction leading for the best price for the seller as well as the buyer on an open market. The term filter bubble refers to drawbacks of personalization especially when reading news, resulting in less serendipity (Meckel 2012) and closed communities. While debates about the existence of these forms and the extent of personalization are ongoing (Vissers et al. 2014) there is no doubt that there are limits to how "personal" websites and advertising should become (Malheiros et al. 2012). Nevertheless, other industries are beginning to adapt the practice of analyzing internet browsing behavior for other purposes like credit scoring[3] or employee satisfaction[4]. The profiles generated from online behavior are becoming a universal measurement of the individual resulting in a market were they are used as currency.

The drawbacks of online profiling are mostly discussed as an issue of privacy with regard to the effects on individuals or groups leading to voices that argue for

---

3 See (for example): http://www.kreditech.com/what-we-do/ (last visit 29.04.2015)

4 See http://uk.businessinsider.com/workday-talent-insights-can-predict-when-employees-will-leave-2015-4 (last visit: 29.04.2015)

stronger regulation (Gutwirth and Hert 2008; Hildebrandt 2012). But before we discuss other possibilities to react we would like to take a step back and discuss profiling as a power technique and a subtle type of control.

> “Demographics and user statistics are more important than real names and real identities. On the Internet there is no reason to know the name of a particular user, only to know what that user likes, where they shop, where they live, and so on. The clustering of descriptive information around a specific user becomes sufficient to explain the identity of that user” (Galloway 2004, 69)

Galloway describes profiling as a phenomenon referring to Foucault's concept of *biopower* that affects people at the level of information.[5] It abstracts from the individual and it's singularities to control a society and steer towards higher goals like a perfect capitalist market, as it is envisioned with price discrimination. According to Galloway and others like (Tiqqun 2012) power is performed in cybernetic systems. This theory implies that separate systems, that are seen as black boxes from the outside, are influencing each other through feedback loops. There are only a few fixed rules stating what to do or - in a consumer context – what to buy. Instead, each action is observed and results in a change in feedback to reach a status that fulfills the aim of both systems. That is how, from a marketing perspective, the events around Mr. Hughes shopping tour can be described. He needs a printer, Bryce Inc wants to sell printers. By observing each other and reacting to one another the win-win situation of a cheaper printer of Mr Hughes and a Sell for Bryce is reached.

But this story can also be interpreted in other ways. As (McStay 2011) points out this conception turns advertisement into autopoetic systems that include the profile as an abstraction of the user or users and need their continuous input to keep going. The description of behavioral advertisement as a feedback based system allows us to critique the way they expropriate any action someone does online for their purposes to be used as a source of information. Still, this often accepts behavioral advertisement as a system that works correctly, at least for its own purpose. Instead, as McStay points out, there is no natural overlap between the profile generated by online tracking systems and the individual.

> The notion of a virtual data-double should not be confused with that of a doppelganger, in that behavioural profiling systems are predicated on aggregating systems that reveal little, if anything, of a user’s real world self. (McStay 2010)

---

5 One could also refer to (Deleuze 1992) who first described the transformation from disciplinary power to control, but Galloway goes more into detail about the technical aspects of this transformation.

The profiles are modeled to work perfectly within the system of online advertising, but not only can the concept of modeling be criticized. In the next section of this paper we show, based on our analysis of real world profiling for behavioral advertisement, that it can be doubted that these models function as expected.

2. LEVELS OF PROFILES

To discuss online profiling it is necessary to get into some details about the technological background. While others have presented definitions of profiling along the targeted individual or group (Hildebrandt 2006) or based on the process of profiling (Ferraris et al. 2013) in our case it is beneficial to distinguish levels of profiles with regard to the amount of information they can contain. We define different levels of profiles based on capabilities of the underlying techniques to describe an individual. Alongside the definition of pseudonyms (Pfitzmann and Hansen 2010) we define three types of profiles: transaction profiles, role profiles and person profiles. These types of profiles also reflect the development of profiling as the techniques evolved over time.

First, on the network layer *transaction profiles* can be created. Those are limited to the information that is required for a specific transaction. Looking at the internet structure these profiles can be created at the level of HTTP requests. A website owner can combine the information about a request from a single Browser/IP address combination into a profile about the user that is assumed to sit in front of a screen. These profiles can be created exclusively at a server a client interacts with, based on the IP-Address and the user-agent string the browser discloses on each request. This string contains information about the type of web browser, and the operating system. In the authors case this is "*Mozilla/5.0 (X11; Linux x86_64; rv:38.0) Gecko/20100101 Firefox/38.0 Iceweasel/38.2.0.*" These profiles are mostly used for immediate adjustments to a service, for example users can be automatically redirected to the mobile version of a website if the user agent string holds information that shows someone is surfing using a mobile operating system.

In marketing this data is also often used in an aggregated way by website owners to learn which sites are visited how often. In today's rapid software development cycles it is also common to test new versions of a website with a small user group to decide which layout elements are more efficient. A/B Testing is used in combination with transaction profiles to measure the reaction of audience groups to, mostly small, changes. Although transaction profiles are considered anonymous, releasing service providers from data protection principles, additional analysis

can be done on these transaction profiles to combine data from different sessions. E-Commerce business have used this data for simple price discrimination algorithms for example to alter prices based on the position associated with an IP-Address (Valentino-DeVries, Singer-Vine, and Soltani 2012) or based on the device someone is using (Mattioli 2012). These profiles can be disrupted fairly easy. When using a different IP Address (e.g. after changing the WIFI) or after a browser or system update that changes the user agent string, the identifier changes and additional steps are needed to keep track of the user.

With *role profiles* some of these gaps are closed. They can be described as service specific profiles that consist of all encounters of a user visiting and using a website over time. They may be bound to a specific account e.g. in an online social network or an online shop but can also be achieved by tracking user interaction pseudonymously. This is most commonly done with various kinds of cookies (Soltani et al. 2009) that are stored on the users' device. This kind of profiles are used in audience analytics software that is described below and are often enriched with statistical information.

Although these profiles are regarded anonymous, there are often not, simply because the amount of data entries per profile often makes them unique. For example in 2006 AOL released a data set that was thought to be anonymous and that contained search terms and visited websites related to pseudonymous IDs. A journalist of the New York Times randomly identified one user within a few days and visited her at home.[6]

It is still easy to escape these forms of profiling. Using ad blockers prevents cookies from being stored in the first place or they can be deleted afterwards.[7] Cookie tracking also has the disadvantage of being bound to one browser (and therefore device). In times of multiple internet devices per user like laptops, smartphones or even smart TVs, it requires additional tracking mechanisms to perform what is called *cross-device-tracking*.

The highest level of profiling online users is to monitor all their internet traffic and behavior to build *person profiles*. This is thought to produce a profile that reflects all actions taken online. One example is Facebook, who started by asking all users to identify themselves by their real names. Now they are trying to bind

6 The database is still online at HTTP://SEARCH-ID.COM/ (last seen 19.08.2015)

7 To prevent this other forms of tracking via zombie cookies or browser-fingerprints as well as exclusion of AdBlock users have been developed. In turn privacy enhancing technologies like the TOR Browser have been developed that try to block all kinds of tracking.

users to its platform by including more and more services and offering additional access modalities on multiple devices. This is thought to convince people to spend more and more time within the closed Facebook-Platform. Google, as another example, enables cross-device-tracking with user identification based on the Google-Account that is needed on in the browser as well as on mobile Android-devices. And Verizon recently used its power as an internet service provider to add an identifier at the networking level to be able to follow users online (Mayer 2014; Mayer 2015).

The Google and Facebook approaches require a large user base that is willing to participate in the strategy offered by the service. But since there is still a large fraction of the web and web users that don't participate as well as a growing number of services that want to track users but do not have a large user base, the majority of profiling is based on technologies of transaction and role profiles.

3.EXAMPLES OF ONLINE PROFILING

There are numerous providers of online tracking and profiling services, new ones open up and others are bought and merged regularly. According to Ghostery Inc. a company that has specialized in AdBlocking and measuring of online advertisement over 2100 ad and tracking services are competing on that market.[8] While some gather profiles for the sole purpose of targeted advertisements others also engage in *audience analytics*. These services are targeted at the publishers and offer insights into the audience of a website. Although they mostly focus on aggregated and averaged information about pages visited on a website, services like *google analytics* or *etracker* show publishers the clickstream of a single user as well as demographic information that generated by user visits to other websites.

Not all service providers disclose which attributes and categories are part of their profile model, in the following we will therefore explain only a few, that are accessible for end-users. In the following we describe three examples that offer at least basic insights into how their profiles are constructed. We will then continue on the analysis of one of them to find out what they look like for a number of virtual users.

8 See https://apps.ghostery.com/en/apps/ for details (last seen 18.09.2015)

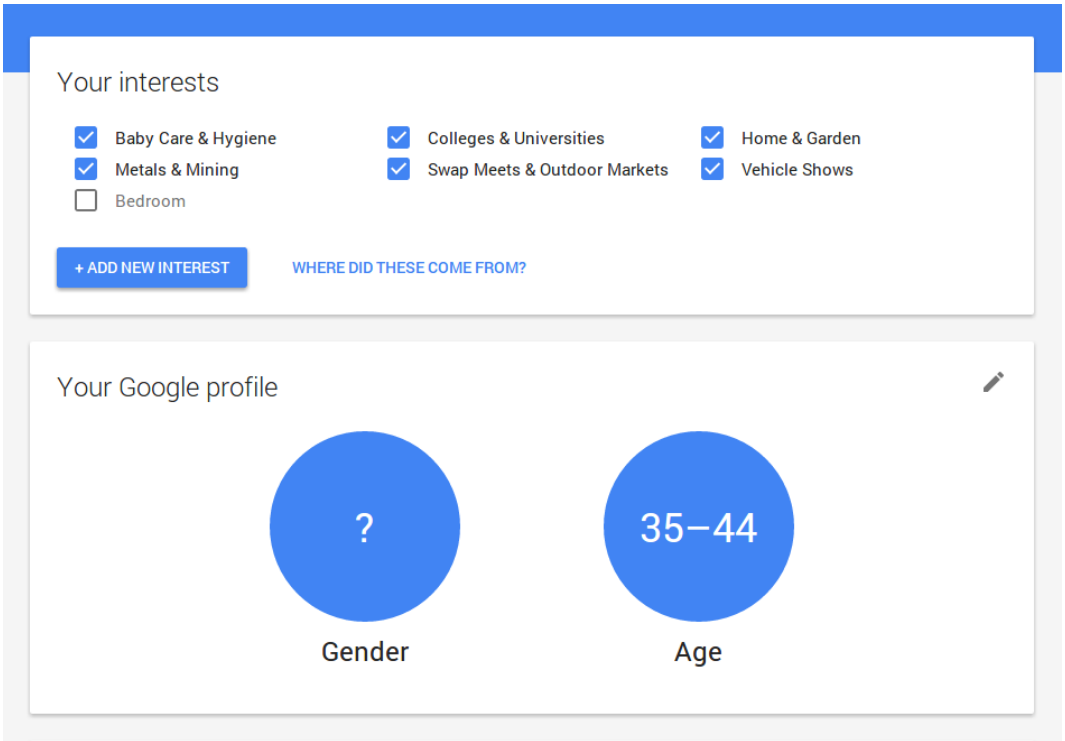

Figure 1: An interest and demographic profile as presented on googles privacy dashboard

1.Google

Google curates a list of 2042 "Ad Interests Categories"[9] of which 946 are related to "localities" while the other 1095 are organized in subcategories related to products. These "interests" are organized in a tree structure based on 24 basic interest categories (Google Interest Categories/GIC) which are further subdivided on up to seven levels like "Sports > Sporting Goods > Combat Sports Equipment > Martial Arts Equipment". The list of categories and the number of interests listed under that GIC are listed in table 1. This list remained unchanged for at least the last two years. While the list covers a broad number of interests it is obvious that it was created for the purpose of targeting advertisements. Therefore, the number off sub-categories for goods and services that are offered, especially on the internet, is significantly larger than those covering activities or interests where there is less competition. For example the number of sub interests for Arts & Entertainment, which contains 53 sub-interests in 'Music and Audio' – related to a still evolving market for selling mp3s as well as streaming services – is remarkably larger than the number of sub-interests in "books & literature" which has only on more level with not content related sub-categories.

9 See HTTPS://SUPPORT.GOOGLE.COM/ADS/ANSWER/2842480 (last seen 18.08.2015)

Interest Category (No. of Subcategories)	
Arts & Entertainment (147)	Travel (27)
News (21)	Autos & Vehicles (95)
Games (42)	Food & Drink (73)
Law & Government (36)	Beauty & Fitness (21)
Finance (50)	Jobs & Education (36)
Computers & Electronics (128)	Reference (30)
Internet & Telecom (34)	Online Communities (18)
Sports (69)	Pets & Animals (15)
Business & Industrial (121)	Books & Literature (9)
People & Society (40)	Home & Garden (48)
Science (25)	Hobbies & Leisure (30)
Shopping (71)	Real Estate (9)

Table 1. Distribution of Google Interest Categories (GIC)

Advertisers can use this lists to specify the target group for an advertisement they want to place on either the google search services or within the *ad sense* network That network covers a large number of websites which offer space on their pages where advertisements are placed by google without the publishers knowing which ones that are. Advertisements placed, e.g. on a news website that participates in ad sense, can refer to either the content of the article a user is reading, or to interests listed in the profile which results in ads that have no relation to the content of the website at all.

Google creates profiles based on cookie tracking in combination with the accounts, if available. Users are tracked on Googles websites and services like Search or GMail as well as on third party websites that either participate in its large advertising network or make use of its analytics service *google analytics*. Personalized tracking across different devices is also possible if users stay logged in to their Google Account in the browser and their Android device(s).

Google also offers insights into the profile they create on their privacy dashboard[10] (see Figure 1). Users are encouraged to keep their profiles up to date and correct them, if necessary to increase the accuracy of the profile (and to help create a person profile). The privacy dashboard includes some basic demographic in-

10 Available at https://www.google.com/dashboard/ (last visited 30.08.2015)

formation like age and gender as well as a number of interests from the list described above. The profile presented on the privacy dashboard gives no hints about the certainty it has about any attribute assigned. Neither does it differentiate between the level of the interest in the interest category tree. Google might assign a general interest in “sports” as well as a very differentiated interest in a specific aspect of “fan fiction”. (Datta, Tschantz, and Datta 2015) have also shown that the privacy dashboard does not necessarily show the actual profile that is used for selecting advertisements to display. It does neither reveal the whole profile that is used, nor do the changes made manually have a measurable effect that can be observed directly.

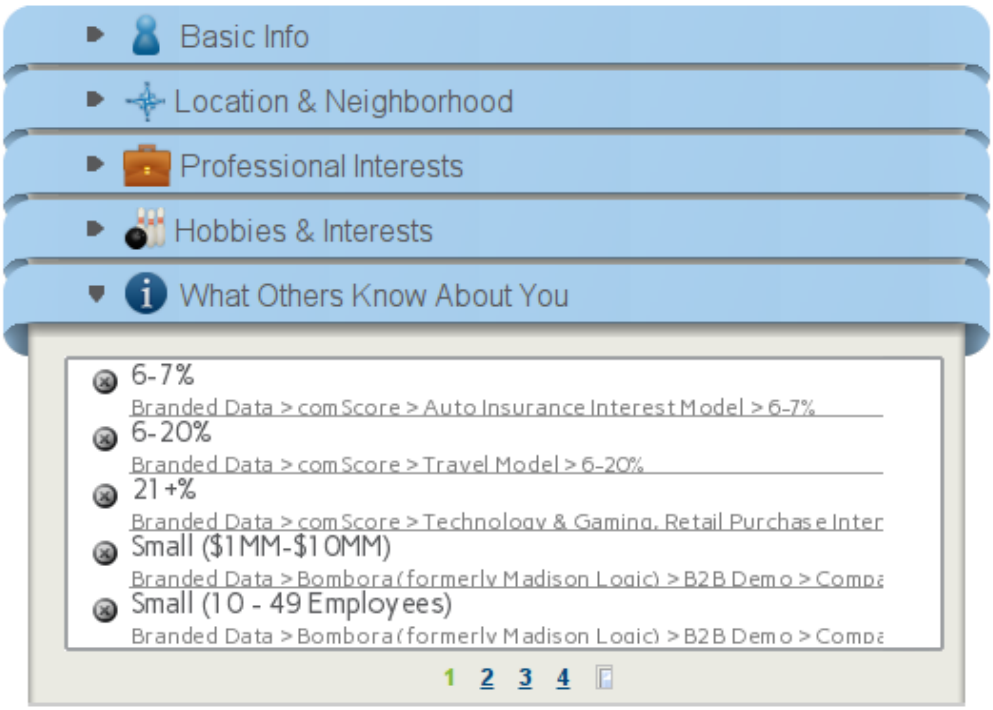

Figure 2: A profile created by bluekai as presented on bluekai.com/registry

2.Bluekai

Bluekai is a marketing service provided and owned by Oracle. It is not as large as Google but has grown over the past years, especially through acquisitions of other companies. In its reports Oracle states that it “ensure[s] that all users who are tagged as 'in-market' have indeed taken actions online to declare themselves as such.” (Oracle 2015). Although this sounds like the accuracy for the attributes being connected to a profile is assured by an “action” taken by a user, normally it is just based on a visit to a website that is defined as being used by people with that attribute. Comparable to Googles privacy dashboard Bluekai offers its “registry”

for users to review and optimize their profile (see figure 2). In difference to Google there is no overview about all possible attributes available, only the categories give some brief insights into the modeling of the profiles.

Bluekai estimates socio-demographic information ('Basic Info') like age and gender as well as data that can be extracted from transaction profiles like the geo-location of an IP address ('Location & Neighborhood'). They also make assumptions about the interest of a user based on the websites visited that were trackable by the service. Those are subsumed under the categories 'Professional Interests' and 'Hobbies & Interests'. The category 'What Others Know About You' include a list of attributes Bluekai has purchased from other services that are specialized in more specific industries. As shown in figure 2 this may include estimations about interests in products as well as additional socio-demographic data like the size of the company someone is working at.

From the source code of the service it is also reproducible that Bluekai offers services for re-targeting of ads, as described in the example of Mr. Hughes. A profile may also include a shopping history ('Things You May Want To Buy') as well as predictions of the likeliness for buying goods based on items were the shopping process was not completed ('Things You May Have Bought').

3.Quantcast

Quantcast is one of the largest providers in the audience measurement market and also offers services in online advertisement. Similar to Google analytics website owners can use a code provided by Quantcast to measure the number of visitors to their site. Using their services is being rewarded with analytic information that is enriched with additional data about the audience (see Figure 3). This includes distribution of users in groups like gender, age, education and income but also rather sensitive data like ethnicity or political views. Table 2 lists detailed categories measured by Quantcast. The data shown for each website is always in relation to the average american internet users. Websites owner can choose to publish data related to their site on the Quantcast homepage since the website is also frequently used by marketers to decide which websites fit their target group. Besides traditional cookie tracking Quantcast uses questionnaires to collect additional information about the audience of sites. Users are then asked to provide details about their socio-demographic data in exchange for rewards like coupons.

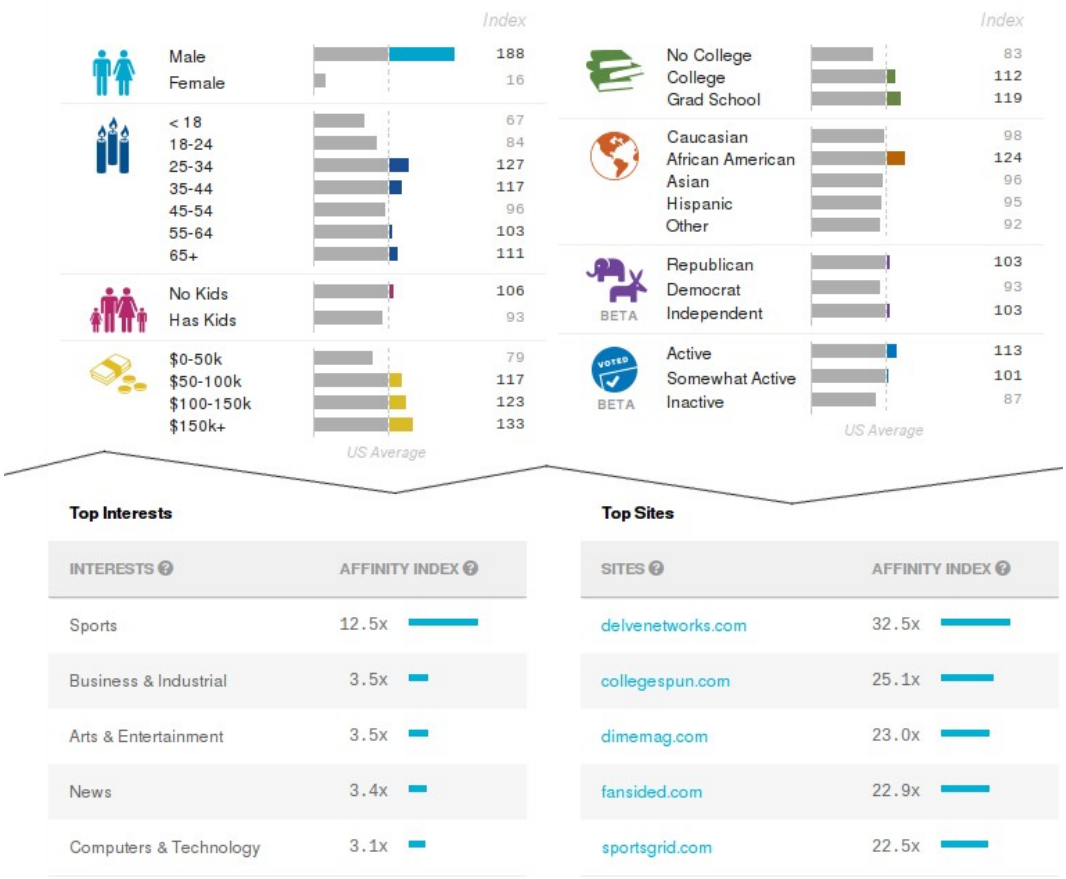

Figure 3: Site specific audience profile as presented at quantcast.com

(Kamerer 2013) has shown that the Quantcast data is highly inaccurate. Quantcast and similar data brokers collect data from various sources combining data from tracking scripts, panel users and third party data without disclosing their methodology in detail. While Kamerer only could prove that the number of visitors estimated by Quantcast, Alexa and Compete differ from first-hand measurements it is reasonable to assume that similar inaccuracies embedded in the other data attributes.

Gender	Age	Children	Income	Education	Ethnicity
Male: 49% Female: 51%	<18: 18% 18-24: 12% 25-34: 17% 35-44: 17% 45-55: 17% 55-64: 10% 65+: 2%	No: 51% Yes: 49%	0-50k: 51% 50-100: 29% 100-150k: 12% >150k: 8%	No College: 45% College: 41% Grad School: 45%	Caucasian: 75% African America: 9% Asian: 4% Hispanic: 9% Other: 1%

Table 2: Detailed categories and attributes of Quantcast data showing the average american distribution Quantcast is comparing each side with.

4. A STUDY OF THE EXTENT OF ONLINE PROFILING BY GOOGLE

The examples help to understand how and which data is used for modeling profiles. In the next step we want to analyze the actual extent of these profiles. To do so we simulated 500 internet users that surfed to 100 pages each and had a look at the profiles presented at googles privacy dashboard. Our findings show that most services that offer profile only have access to a small fraction of the users internet behavior. Google, the company that owns the largest tracking network, is able to track up to 60% of a users' website visits. Nevertheless, they are constructing profiles and pretend to know a user very well.

1.Simulating users

We used Reddit.com as a source for link lists that are bound to a specific user comparable to browsing histories. Reddit.com is an online community platform where users publish links to content on the web to discuss and rate it. While a large fraction of those links is posted to curate subreddits where users chat about content related to a specific interest, there remains a group of users that use reddit as a bookmarking platform. They mainly post links to websites they have visited to a personal profile page. These pages are publicly available and Reddit offers an interface (API) for external programmers to access this content in a structured form. We used a script to collect 500 of these link-profiles that matched our demands being:

- Link lists should consist of at least 60 links, where 100 is the maximum number of links provided per user.
- These 60 or more links should point to at least 60 external sites. Since many users post a lot of reddit internal links. And they
- should point to at least 20 different domains.

We then used these link lists together with an automated web browsing script[11] to simulate these users. The automated browser surfed to each of the 100 sites within one browsing session and collected any data that was exchanged between the browser and the servers.

11 We used phantomjs (HTTPS://WWW.PHANTOMJS.ORG) with Adobe Flash support to automate the browsing process.

2. Measuring the extent of tracking

The whole dataset contained 45829 links to 7123 domains distributed to 506 users. Each users' link list contained 96 URLs to 44 different domains in average. The ten most visited websites of each user made up 59,6% of all of her links. The most linked websites were related to news websites like theguardian.com, nytimes.com or reuters.com as well as entertainment websites like imgur.com, youtube.com or reddit.com itself. This top list differs from more generalized lists of the most visited website[12] which rank google.com, faceobok.com and amazon.com the highest. While our dataset therefore does not represent average browsing behavior it is still a good source to study online tracking since pages like facebook.com oder amazon.com are closed platforms that do not contain tracking scripts of third parties which we wanted to study.

In total we measured that 20% of all the web traffic (HTTP and HTTPS requests) produced was directed to third parties. While these requests to servers do not serve the content of the pages, they are not solely providing tracking services. Other third party sources include scripts from third party websites like Social-Plugins or externally hosted images that do not necessarily serve tracking purposes. However, about 50% of all web traffic was directed to only 50 domains of which 31 are directly related to advertisement networks. Table 3 shows the amount of sites from the link lists that connected to one or more of these trackers. It shows that only Google is able to track more than 50% of the users' website visits.

% of link profils	Services/Domains
50-60%	Google (google-analytics.com; doubleclick.com/net)
40-50%	scorecardresearch.com (audience analytis), facebook.com, twitter.com
30-40%	quantserver.com (audience analytics; source of Quantcast.com)
10-20%	adnxs.com, taboola.com, outbrain.com, bluekai.com, disqus.com, rubiconproject.com, addthis.com (advertisment networks)
5-10%	chartbeat.com, optimizely.com, amazon-adsystems.com, krxd.net

Table 3: Percentage of the link profile that was tracked by different tracking services

12 The toplist referred to in related literature is http://www.alexa.com/toplist

If we add those requests to google services that are not explicitly designed for tracking (e.g. googleapis and fonts) 81,63% of all websites in our test set connect to google.

These numbers are in line with simlar studies like (Gomez, Pinnick, and Soltani 2009) who found google to be able to track 80% of 766,000 domains they analyzed. Recently (Acar et al. 2014) found that a technology they called *cookie syncing* is being used for cooperation between tracking services to extend the reach of each service. This allowed multiple tracking services to be aware of visits of 50% of the top 3000 most visited web sites.

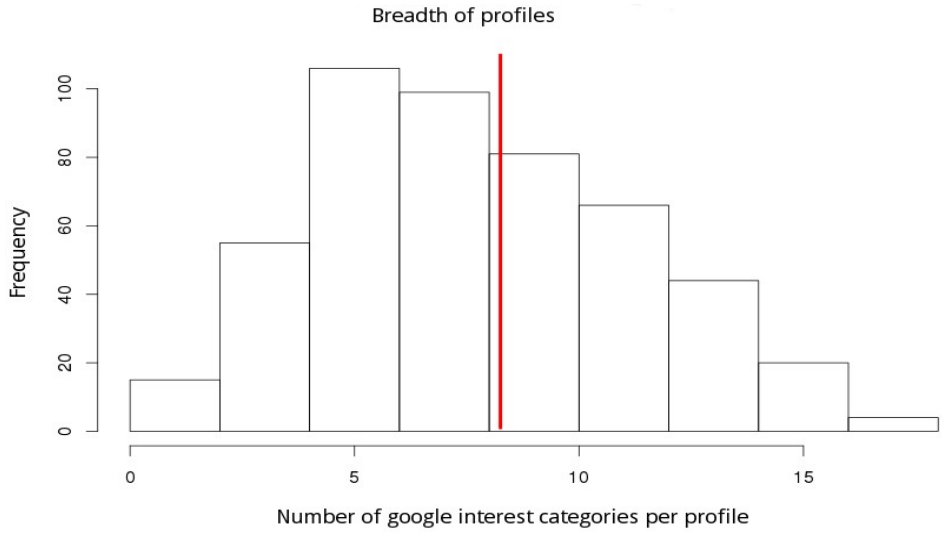

Figure 4: The breadth of a profile is the number of google interest categories a user is assigend after visiting 100 web sites. The red line marks the average. Standard deviation 3.53.

3.Profiles created by google

After all websites of a link list were visited we turned the script to the google privacy dashboard that shows the profile created about a user, as described above. The analysis of the privacy dashboard reveals that more than 17 interests are assigned to a user after visiting 100 pages ranging from 1 to 37 out of the 1095 possible.

As explained above, the privacy dashboard does not differentiate between the level of an interest with regard to the taxonomy created by Google. To make the profiles comparable we therefore defined the breadth of a profile as the number of interest categories (GIC) that a profile contained. This number was derived by only counting the base interest for each interest so that multiple sub-interests (like “Hip Hop Music” and “Rock Music”) only resulted in one GIC point (for “Mu-

sic). Figure 4 shows the distribution of this breadth measurement. In average the breadth of the profiles was 8.25.

4.Volatility of the profiles

To make sure that our measurement was correct we tried to reproduce the data a few weeks after our first test. We were surprised to find out that, although the same websites were visited in the same order using the same tools, we found that the profiles did not match. The breadth of the profiles dropped from 8.25 to 7.38 and there were also differences in the GICs assigned. Only 51% of the interest categories from the first test were assigned in the second one, too. We found 1.84 (s=1.32) interest categories to be not present in the second run for the same link list while 1.29 (s=1.72) interest categories were shown on the privacy dashboard that were not assigned in the first place.

We can think of several explanations for this high variance. First, due to the effects of *real time bidding,* every time a website is visited, another tracker can be present. While Google might have served the ad on a page load during the first test, it might not do so on the next. Second, it is likely that Google assigns interests not based on the content of the site but rather by using a relation to the interests that are known about other users of the same site. For example, if user *A* reads a news article about topic 1 that was also read by users *B* and *C.* If users *B* and *C* both “expressed” their interest in topic 2 by visiting another website, this might lead to topic 2 (and the corresponding interest) being assigned to user *A* as well. Depending on how Google weights these relations, e.g. by ranking recent visits to websites higher in terms of how they represent an interest, the profiles change frequently dependent not only of which websites A visits, but also how B and C behave. A third factor might be that those websites are more prominent in our dataset that update their content frequently, like those of large news agencies (*theguardian*, *nytimes* etc.). Although the specific article visited during our test could have lead to the same interest based on the articles content, Google might relate other interests to the site and consecutively to the user. This is likely when, during a later visit, the website feature other topics on it's front page or even in sidebars that change from one visit to another. We found an example for this behavior when we tested the interests assigned after the visit of a single website. Multiple parallel issued requests to the front page of *wired.com,* a large IT news site, resulted in the interest “Computer and Electronics”, for all sessions, but also in up to two other, highly varying interests. This is related to the fact that the

front page of wired.com is rendered differently and features different articles on each visit.

As it was shown for Quantcast and other audience analytics services, they make use of data sets bought from third parties. In times of big data it is often assumed that using more data from more sources has advantages above assuring the quality and methodology of the statistics in use. Based on the weight that is given to a secondary data set in the profile model inaccuracies from all sources might add up.

5. CONCLUSION

We have discussed the methods of online profiling and the promises they make towards marketers. It is based on the conception that it is possible to understand a personality by reviewing her browsing history and that this knowledge enables them to interact and influence the user. But with the evolution of the technology to a state were each internet users is thought to be targeted individually based on her person profiles, it has instead started to create data doubles with whom is interacted instead of the user.

It is often argued that computer-based judgments based on profiling will lead to more fairness because of a reduction of human bias (Youyou, Kosinski, and Stillwell 2015). We assume that in the future these judgments will not be based on fixed, user-curated profiles but – in line with big data practices - with unstable and volatile profiles that are a combination of transaction profiles but also third party information from multiple sources. This will result in more intransparency and therefore unfairness since the algorithms work with many correlations and dependencies. Therefore, decisions made are neither reproducible nor comprehensible by humans even if they have access to the profiles created since they won't know for sure if this is the one that a decision is based on. A profile is created at the moment it is used, to make a decision based on the individuals' previous behavior, context information (e.g. the time of day) and also the previous behavior of other internet users.

Although advertisement agencies promise to create models of users and audience that represent each single internet user, so that a feedback loop between ads and users can be established, this is not the case. Instead, the profiles are vague and can be assumed to be incorrect to a large extent. For advertisement agencies inaccuracies are no problem at all since the still allow to place ads, but users are

left with no options to adopt since there is no way of knowing what feedback one's actions will provoke.

Due to the high rate of failure, mismeasurement and not-measurable parts of an individual, online tracking services have begun to create profiles that are vague and indecisive only that the idea of profiling is to help to make decisions (Ferraris et al. 2013). Instead of creating profiles that are closer to person profiles our study shows that profiles are dependent on various context variables.

We would argue, inline with McStay, to leave the notion of profiles behind that are perceived as a (partial) representation and description of an individual. While this might hold true for the facebook profile curated by oneself, the profiles that are used within data processing operations by advertisers and others are much more complex and dependent on the context they are used for. Although some services offer insights into their profiling practice, this information can not be trusted. To understand the complex relations between oneself, the online behavior and reactions of the systems, a new form of *information* and *data mining literacy (Berendt 2012)* is necessary if people want to keep or even re-gain autonomy towards the services and to be enable to interact with their profiles. To do so, technologies that support influencing and obfuscation (Brunton and Nissenbaum 2011) of profiles are needed.

REFERENCES

Acar, G., C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. 2014. “The Web Never Forgets: Persistent Tracking Mechanisms in the Wild.” In *Proceedings of the 21st ACM Conference on Computer and Communications Security*.

Berendt, Bettina. 2012. “Data Mining for Information Literacy.” In *Data Mining: Foundations and Intelligent Paradigms*, edited by Dawn E. Holmes and Lakhmi C. Jain, 265–97. Intelligent Systems Reference Library 25. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-23151-3_12.

Boda, Károly, Ádám Máté Földes, Gábor György Gulyás, and Sándor Imre. 2012. “User Tracking on the Web via Cross-Browser Fingerprinting.” In *Information Security Technology for Applications*, edited by Peeter Laud, 31–46. Lecture Notes in Computer Science 7161. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-29615-4_4.

Brunton, Finn, and Helen Nissenbaum. 2011. “Vernacular Resistance to Data Collection and Analysis: A Political Theory of Obfuscation.” *First Monday* 16 (5). http://firstmonday.org/ojs/index.php/fm/article/view/3493.

Danna, Anthony, and Oscar H. Gandy. 2002. “All That Glitters Is Not Gold: Digging Beneath the Surface of Data Mining.” *Journal of Business Ethics* 40 (4): 373–86. doi:10.1023/A:1020845814009.

Das, Sauvik, and Adam Kramer. 2013. “Self-Censorship on Facebook.” In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6093.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. 2015. “Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination.” *Proceedings on Privacy Enhancing Technologies* 2015 (1): 92–112. doi:10.1515/popets-2015-0007.

Deleuze, Gilles. 1992. “Postscript on the Societies of Control.” *October* 59 (January): 3–7.

Ferraris, Valeria, Francesca Bosco, G. Cafiero, Elena D’Angelo, and Y. Suloyeva. 2013. “Defining Profiling.” SSRN Scholarly Paper ID 2366564. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2366564.

Galloway, Alexander R. 2004. *Protocol: How Control Exists After Decentralization*. Cambridge, MA: MIT Press.

Gomez, Joshua, Travis Pinnick, and Ashkan Soltani. 2009. “KnowPrivacy.” Berkeley, CA: UC Berkeley, School of Information. http://www.knowprivacy.org/.

Gutwirth, Serge, and Paul De Hert. 2008. “Regulating Profiling in a Democratic Constitutional State.” In *Profiling the European Citizen*, edited by Mireille Hildebrandt and Serge Gutwirth, 271–302. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-1-4020-6914-7_14.

Gutwirth, Serge, and Mireille Hildebrandt. 2010. “Some Caveats on Profiling.” In *Data Protection in a Profiled World*, edited by Serge Gutwirth, Yves Poullet, and Paul De Hert, 31–41. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-90-481-8865-9_2.

Haggerty, Kevin D., and Richard V. Ericson. 2000. “The Surveillant Assemblage.” *The British Journal of Sociology* 51 (4): 605–22. doi:10.1080/00071310020015280.

Hildebrandt, Mireille. 2006. “Profiling: From Data to Knowledge.” *Datenschutz Und Datensicherheit - DuD* 30 (9): 548–52. doi:10.1007/s11623-006-0140-3.

———. 2012. “The Dawn of a Critical Transparency Right for the Profiling Era.” *Stand Alone*, 41–56. doi:10.3233/978-1-61499-057-4-41.

Kamerer, David. 2013. “Estimating Online Audiences: Understanding the Limitations of Competitive Intelligence Services.” *First Monday* 18 (5). http://firstmonday.org/ojs/index.php/fm/article/view/3986.

Malheiros, Miguel, Charlene Jennett, Snehalee Patel, Sacha Brostoff, and Martina Angela Sasse. 2012. “Too Close for Comfort: A Study of the Effectiveness and Acceptability of Rich-Media Personalized Advertising.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 579–88. CHI ’12. New York, NY, USA: ACM. doi:10.1145/2207676.2207758.

Mattioli, D. 2012. “On Orbitz, Mac Users Steered to Pricier Hotels.” http://online. wsj.-com/article/SB10001424052702304458604577488822667325882.html.

Mayer, Jonathan. 2014. “How Verizon’s Advertising Header Works.” *Web Policy*. October 24. http://webpolicy.org/2014/10/24/how-verizons-advertising-header-works/.

———. 2015. “The Turn-Verizon Zombie Cookie.” Blog. *Web Policy*. January 14. http://webpolicy.org/2015/01/14/turn-verizon-zombie-cookie/.

McStay, Andrew. 2010. “Profiling Phorm: An Autopoietic Approach to the Audience-as-Commodity.” *Surveillance & Society* 8 (3): 310–22.

———. 2011. *The Mood of Information: A Critique of Online Behavioural Advertising*. A&C Black.

Meckel, Miriam. 2012. “Rettet Den Zufall - Digital Serendipity.” Accessed October 5. http://digitalserendipity.org/.

Morris, Adam. 2012. “Whoever, Whatever: On Anonymity as Resistance to Empire.” *Parallax* 18 (4): 106–20. doi:10.1080/13534645.2012.714560.

Mowery, Keaton, and Hovav Shacham. 2012. “Pixel Perfect: Fingerprinting Canvas in HTML5.” *Proceedings of W2SP*.

Oracle. 2015. “Oracle Data Cloud - Data Directory.” Oracle. http://www.oracle.com/webfolder/assets/cloud-data-directory/index.html.

Pagefair. 2015. “The 2015 Ad Blocking Report.” http://blog.pagefair.com/2015/ad-block-explorer/.

Pariser, Eli. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Pfitzmann, Andreas, and Marit Hansen. 2010. *A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management*. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

Singer, Natasha. 2012. "Acxiom, the Quiet Giant of Consumer Database Marketing." *The New York Times*, June 16, sec. Technology. https://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html.

Soltani, Ashkan, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. 2009. "Flash Cookies and Privacy." In . http://papers.ssrn.com/abstract=1446862.

Tiqqun. 2012. *The Cybernetic Hypothesis*. translation collective. https://translationcollective.files.wordpress.com/2012/06/cybernetic-hypothesis.pdf.

Turow, Joseph. 2011. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven: Yale University Press.

Valentino-DeVries, Jennifer, Jeremy Singer-Vine, and Ashkan Soltani. 2012. "Websites Vary Prices, Deals Based on Users' Information." *Wall Street Journal*, December 24, sec. Tech. http://online.wsj.com/news/articles/SB10001424127887323777204578189391813881534.

Vissers, Thomas, Nick Nikiforakis, Nataliia Bielova, and Wouter Joosen. 2014. "Crying Wolf? On the Price Discrimination of Online Airline Tickets." *HotPET Symposium*. http://www.securitee.org/files/pdiscrimination_hotpets2014.pdf.

Youyou, Wu, Michal Kosinski, and David Stillwell. 2015. "Computer-Based Personality Judgments Are More Accurate than Those Made by Humans." *Proceedings of the National Academy of Sciences*, January, 201418680. doi:10.1073/pnas.1418680112.

Yuan, Shuai, Jun Wang, and Xiaoxue Zhao. 2013. "Real-Time Bidding for Online Advertising: Measurement and Analysis." In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, 3:1–3:8. ADKDD '13. New York, NY, USA: ACM. doi:10.1145/2501040.2501980.