Auditing Offline Data Brokers via Facebook's Advertising Platform

Giridhari Venkatadri, Piotr Sapiezynski, Elissa M Redmiles, Alan Mislove, Oana Goga, Michelle L Mazurek, Krishna P Gummadi

Auditing Offline Data Brokers via Facebook's Advertising Platform

Giridhari Venkatadri
Northeastern University

Piotr Sapiezynski
Northeastern University

Elissa M. Redmiles
University of Maryland

Alan Mislove
Northeastern University

Oana Goga
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG

Michelle L. Mazurek
University of Maryland

Krishna P. Gummadi
MPI-SWS

ABSTRACT

Data brokers such as Acxiom and Experian are in the business of collecting and selling data on people; the data they sell is commonly used to feed marketing as well as political campaigns. Despite the ongoing privacy debate, there is still very limited visibility into data collection by data brokers. Recently, however, online advertising services such as Facebook have begun to partner with data brokers—to add additional targeting features to their platform— providing avenues to gain insight into data broker information.

In this paper, we leverage the Facebook advertising system—and their partnership with six data brokers across seven countries—in order to gain insight into the extent and accuracy of data collection by data brokers today. We find that a surprisingly large percentage of Facebook accounts (e.g., above 90% in the U.S.) are successfully linked to data broker information. Moreover, by running controlled ads to 183 crowdsourced U.S.-based volunteers, we find that at least 40% of data broker sourced user attributes are not at all accurate, that users can have widely varying fractions of inaccurate attributes, and that even important information such as financial information can have a high degree of inaccuracy. Overall, this paper provides the first fine-grained look into the extent and accuracy of data collection by offline data brokers, helping to inform the ongoing privacy debate.

ACM Reference Format:
Giridhari Venkatadri, Piotr Sapiezynski, Elissa M. Redmiles, Alan Mislove, Oana Goga, Michelle L. Mazurek, and Krishna P. Gummadi. 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313666

1 INTRODUCTION

Data brokers such as Acxiom [5] and Experian [23] have traditionally collected, aggregated, and linked information about people's activities, based on a variety of sources (e.g., voter records, vehicle registries, loyalty cards, and so forth). Their business model is then selling this information to third-parties such as banks, insurance companies, political campaigns, and marketers. The collection practices of data brokers have been the subject of an ongoing privacy debate. For example, data brokers typically only make data available to clients who purchase it, and not to the users who it is actually about [20]. Even worse, public-facing web sites run by data brokers [10] that purport to reveal the data only report a fraction of what they actually have [19, 47]. In fact, outside of a few niche areas (e.g., credit reports), people in the U.S. have limited if any rights to determine the provenance of, correct, or even view the data these companies have on them [38]. As a result, researchers and society at large still have a very limited understanding of the extent, accuracy, or provenance of this data collection ecosystem.

In parallel, *online services* such as Facebook and Google have been collecting information about people's online activities. Their business model is to build advertising platforms that use this data to provide advertisers with fine-grained targeting features [29, 31]. Recently, data brokers and online services have begun partnering together, allowing for the data collected about users online to be linked against data collected offline. This enables online services to provide advertisers with targeting features that concern users' offline information (e.g., advertisers can then target users based on their net worth, purchase behavior, and so forth [4, 34]). While this linking of offline and online data may seem to only further stoke the privacy debate, it does have one significant benefit: it offers a unique opportunity to gain visibility into the data broker ecosystem. Specifically, advertising platforms often provide advertisers with statistics about any audience an advertiser can target, *including those created using data broker-provided attributes.* This provides the first opportunity to gain insight into the extent of data collection by offline data brokers.

In this paper, we use the Facebook targeted advertising platform to study the *coverage* and *accuracy* of data collected by four offline data brokers (Acxiom, Epsilon, Experian, and Oracle Data Cloud - formerly Datalogix) across seven countries (U.S., U.K, Australia, Germany, France, Japan, and Brazil). [1] To examine the coverage of data brokers (i.e., the fraction of the population they have data on), we collect statistics from the Facebook advertising interface on how many Facebook users possess each of more than 600 data broker-provided attributes. Our results demonstrate that a large

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313666

[1] While Facebook reported having partnerships with six different brokers [4], we were able to find targeting attributes corresponding to only four of these brokers on Facebook's advertising platform.

percentage of Facebook accounts (e.g., ranging from over 48% of users in Japan to over 90% of users in the U.S.) are indeed linked to some data broker information, thus demonstrating for the first time the extent of offline-online data linkage, in addition to demonstrating the coverage of offline data brokers. Furthermore, we devise a methodology to study the coverage at a fine granularity based on the targeting mechanisms and statistics commonly provided by online advertising platforms in Section 3. We find that data broker coverage generally increases with the age of users, and can be significantly lower in U.S. counties with higher poverty levels. We also find that data brokers have high levels of coverage for potentially sensitive user attributes such as income (76.9% of all U.S. users), net worth (67.6%), and purchase behavior (86.7%).

Facebook provides users with an Ads Preferences Page [28] where they can see and correct what data Facebook has inferred about them; however, this page does not reveal to users *any* of the attributes sourced from data-brokers [2]. To study the accuracy of these attributes we use a recently-proposed transparency mechanism [53], which consists of running ads that reveal information about their targeting to the targeted users. Such ads, called *Transparency-enhancing advertisements* (Treads), work because advertising platforms by their very functionality only show targeted ads to users who satisfy the targeting of that ad. Thus, by running one distinct Tread targeting each individual data-broker-sourced attribute offered by the ad platform, we can reveal to users which data broker attributes they have.

To study the accuracy of attributes sourced from data brokers, we run Treads to a set of crowdworkers who installed a browser extension that collects the Treads seen by the worker, and surveys them about the accuracy of the inferred broker information. We use the results from 183 workers—corresponding to 1,432 individual inferred data broker attributes—to analyze the overall accuracy of broker and Facebook information, the per-user accuracy of broker and Facebook information, and the accuracy of broker information across various information categories. We find that the wide coverage of data broker information comes at the cost of mediocre accuracy, with around 40% of attributes being reported as "Not at all accurate" by workers. We also find that inaccuracy does not occur in an all-or-none fashion (i.e., users can have widely varying percentages of inaccurate data broker attributes). Finally, we find that even potentially sensitive attributes like financial information have mediocre accuracy, with for example 56% of attributes about investments reported as "Not at all accurate".

Taken together, our results present the first detailed look into the coverage and accuracy data broker ecosystem; our methodology could also be used to help trace the provenance of collected data, as well as to study how data collection practices change over time.

2 BACKGROUND

We begin by providing some background about offline data brokers, before explaining relevant parts of Facebook's advertising interface.

2.1 Offline data brokers

While some data brokers such as Acxiom [5], Epsilon [21], and Datalogix (now part of Oracle Data Cloud [39]) are geared towards aggregating consumer data to sell to marketers and advertisers, other companies such as Experian [23] have a greater focus on creating credit reports for businesses.

Kinds of data collected The data collected by data brokers is diverse and often very sensitive. It can include demographic information such as ethnicity and occupation, household characteristics such as age and number of children, financial information such as income and net worth, life events such as divorces, and credit-related information about property, mortgages, and investments [9, 26]. Other information collected can include sensitive health-related information such as ailments, medications, visual impairments, and health indicators [8, 26]. Publicly available detailed catalogs of data offered by data brokers are hard to find; we were only able to find the following data catalogs corresponding to Experian [26], Oracle Data Cloud [40], and Acxiom [6]. These catalogs show the diversity, sensitivity, and fine-grained nature of the information collected by these data brokers, highlighting the importance of understanding their data collection better.

Extent of data collection Data brokers reveal only high-level information about the extent of the data they collect. For example, Acxiom reveals that its data collection "now encompasses more than 62 countries, 2.5 billion addressable consumers and more than 10,000 attributes", for a "comprehensive representation of 68 percent of the world's online population" [7]. Similarly Epsilon reveals that its data covers "virtually every U.S. household" [22], and Experian reveals that it "maintains credit information on approximately 220 million U.S. consumers and 25 million active U.S. businesses", "demographic information on approximately 235 million consumers in 117 million living units across the U.S.", and "information on more than 650 million vehicles in the U.S. and Canada" [24]. However, except to their clients, these data brokers usually do not reveal detailed lists of what kinds of information they collect, and how many users are covered by different kinds of information.

Some data brokers do allow individuals to gain some insight into the data collected about them, such as Acxiom's "About the Data" site [10] and Experian's free credit reports [25]. However, these are only sanitized versions of their data and only show a fraction of the actual data collected [19, 47].

2.2 Facebook's advertising platform

Facebook's advertising platform is one of the largest and most mature online advertising platforms. It leverages Facebook's user data to allow advertisers to target *audiences* of users (i.e., sets of users with particular attributes) with ads.

Targeting parameters When specifying an audience, advertisers must specify the location (e.g., countries, states, or postal codes), age range (between 13 and 65+), and gender of users to target. In addition, advertisers can choose to include or exclude users who have certain attributes, choosing from a list containing 1,121 attributes; of these, 614 attributes come from Facebook and are present in all countries, while the others come from data brokers and can vary across countries [2]. In addition, advertisers are able to create targeting formulas of boolean expressions over attributes [11].

Size estimates Once the advertiser specifies the particular attributes of users to target, Facebook then gives the advertiser a size estimate (called the *potential reach*) that represents the number of

monthly active Facebook users who meet the targeting criteria.[2] Prior work found that the potential reach was calculated by rounding the number of matching users to two significant digits [52].

Partner categories Facebook refers to the offline information obtained by linking its profiles with data brokers (also called Facebook Marketing Partners) as *partner categories* [4]. Facebook reports that partner categories are "available to people targeting audiences located in the United States, Brazil, France, Germany, the United Kingdom, Australia, and Japan"; this includes data from Acxiom, Acxiom Japan, CCC Marketing (for Japan), Epsilon, Experian, Oracle Data Cloud, and Quantium [4]. Facebook recently announced that it would soon be terminating these partnerships and removing these categories [49] in the wake of the Cambridge Analytica scandal, and ultimately did so in October 2018.

Linking partner data One key concern with partner data is the *linking* process: essentially, identities in the partner's database need to be correctly linked with Facebook's user database to enable accurate targeting. While the exact linking methodology is not public, Facebook has stated that it is based on users' unique identifiers [34]:

> For each Facebook user, the company computes a hash code of the person's phone number, email address and other major identifiers and transmit those to Acxiom, Oracle, Epsilon and the other data brokers it works with and requests that they return all available marketing segments they offer for that user.

2.3 Related work

We next review closely related work on data brokers, linkage between offline and online identities, and online advertising platforms.

Coverage of data brokers To the best of our knowledge, there have been no studies analyzing the coverage of data brokers in a fine-grained manner; as previously mentioned, this has been due to the difficulty of conducting such a study owing to the opaque practices of data brokers. While the U.S. Federal Trade Commission (FTC) published a report in 2014 describing data brokers, their sources of data, and their clients [20], they were unable to report on the coverage of data by these data brokers at a fine-grained level.

Accuracy of data brokers There have been a number of anecdotal reports that show that data collected by data brokers can be inaccurate. For example, a journalist obtained a copy of their information held by Oracle Data Cloud [34], and examined their data as provided by Acxiom's "About the Data" site, finding that more than 70% of their attributes from either source were inaccurate. Similarly another journalist [13] found their information from Acxiom highly inaccurate, while yet another [37] found nearly 50% of their personal information purchased for a $50 fee from an undisclosed company inaccurate. While these results are intriguing, they each represent data for only a single individual.

A recent white paper [42] presented a small-scale study of 107 Deloitte U.S. employees, asking them to review their data revealed by a "leading consumer data broker" with "a publicly available, web-based portal that presents users with a variety of personal and household data." These authors restrict themselves to a small, arbitrary subset of 30 attributes, finding that nearly half of the attributes are incorrect for more than 50% of users who have them. They do not report how many of the 107 participants actually had data about them and responded to the survey. Another small-scale study [45] surveyed 8 graduate students about their information collected by Oracle BlueKai as revealed by its transparency feature, with all participants finding some inaccuracies in their profiles. These two studies focus on a non-representative sample of users, on one data broker, and (in the case of the former study) on a small set of attributes provided by the broker; in comparison, our methodology allows us to study a more general sample of users, covering multiple data brokers, avoiding the use of a sanitized version of the data, and focusing on all 507 data broker attributes that Facebook makes available through its advertising platform.

Online-offline linkage Prior work [36] has explored the potential for data brokers to link online data (Facebook profiles) and offline data (voter registration records) in target cities in the U.S., presenting one method to link such profiles. Since data brokers might potentially use more sophisticated methods to accomplish such linkage, our work makes the complementary contribution of empirically demonstrating the extent of such online-offline linkage.

Online advertising platforms A number of recent studies have looked at the privacy implications of online advertising platforms. For example, multiple projects [27, 33, 52] have discovered serious privacy leaks on Facebook's ad platform, and others [48] demonstrated the potential for abusing Facebook's ad platform to launch hard-to-detect discriminatory advertising on Facebook's ad platform. Other researchers [17, 46] have demonstrated that these weaknesses are not confined to Facebook, and are present in Google's ad platform as well. The recent inclusion of data broker-derived data—often consisting of sensitive information such as financial information—could exacerbate such ill effects.

On the other hand, recent work [2] studied Facebook's transparency mechanism that purports to show users why they were targeted with a particular ad. They found that Facebook's transparency feature does reveal information about Facebook's attributes, but *does not* reveal any information about data broker attributes that have been collected about users. Our work is therefore the first to shed light on the usage of data broker information in this ecosystem.

A number of recent studies have used Facebook's advertising platform for various demographic studies, either estimating the coverage of sensitive attributes computed internally by Facebook [16], studying the distribution of health conditions using various indirect markers [35], studying the gender divide on Facebook [30], or studying the migration of populations [55]. Our paper, on the other hand, estimates the extent of linkage between online and offline identities, and the coverage achieved by offline data brokers, using Facebook's advertising platform.

Finally, a number of recent studies have examined the accuracy of information revealed by the transparency mechanisms of online advertising platforms. While one study [50] found that participants' age and gender as revealed by Google's transparency mechanism were accurate for 65% - 74% of participants, and had missing values for 18%-29% of participants; another study [14] found that over 52% of participants reported less than half of their information as listed by the transparency mechanisms of Google, Facebook, and a

[2]Facebook previously defined potential reach as measuring the number of *daily* active users [52]; however, the definition of potential reach has always captured the notion of active users.

small broker Nielsen eXelate as "relevant". Finally, 27% of users in a recent report [1] found their information revealed by Facebook's transparency mechanism inaccurate. Our work instead chiefly focuses on the coverage and accuracy of information aggregated by offline data brokers.

3 COVERAGE

We next describe our methodology for examining the coverage of data collected by data brokers, before analyzing the results.

3.1 Methodology for studying coverage

Obtaining data broker attributes Similar to prior work [2], we use the "web inspector" feature of our web browser to identify the API call made by Facebook's ad interface to retrieve the list of targeting attributes shown to an advertiser.[3] These attributes are organized in a hierarchy, with the highest level categories being "Interests", "Demographics", and "Behaviors". These are then subdivided into sub-categories; the targeting attributes themselves are the leaf nodes of this hierarchy. For all attributes, the result of the API call contains the name of each targeting attribute, its parents in the hierarchy, and the total number of users with that attribute. For partner attributes, the result also contains the name of the partner (data broker) that the attribute was sourced from, and a brief description about the attribute.

We notice that the list of targeting attributes shown depends on the country in which the advertising account is created (and not, for example, on the country of the audience being targeted). Thus, to obtain the list of targeting attributes across different countries, we create Facebook accounts while logged in to an Amazon Web Services (AWS) instance based in that country, and then obtain the corresponding list of attributes shown. We collected these lists of targeting attributes in April 2017 for all seven countries (U.S., U.K, Australia, Germany, France, Japan, and Brazil) where Facebook reports it offers partner categories [4].

We found partner categories in all seven countries, with a varying number of data brokers in these different countries as can be seen from the first two columns of Table 1. These included all but two brokers—Quantium (for Australia) and CCC Marketing (for Japan)—reported by Facebook [4]; this may be because Facebook allows advertisers to request additional categories on a case-by-case basis from certain partners (as opposed to offering them to all advertisers via the ad interface). In addition, we collected these lists of targeting attributes for all three other countries (Canada, South Korea, and India) where we could create AWS instances; as expected, we found no partner categories corresponding to these countries.

Measuring extent of data broker coverage While it would be ideal to measure the percentage of *all* Facebook identities that have been linked to data broker information, the platform offers no way of estimating this. Instead we focus on the set of Facebook identities that are *targetable* by advertisements, since Facebook provides us estimates of the size of this set (via the *potential reach* estimates shown when creating an ad).[4]

[3]One only needs a Facebook account in order to be an advertiser and use Facebook's ad interface.

[4]These estimates could vary over longer intervals of time as they count active users [52]; thus, we ensure that all measurements corresponding to that population are made within short intervals of time. Also, these estimates are rounded to have two significant digits [52]; we verified that the magnitude of this measurement error does not affect our findings.

Country	Partner	Attribute count	Targetable		Percent
			Overall	Partner	
U.S.	**All**	507	210M	190M	90.5%
	Acxiom	128	210M	160M	76.2%
	Datalogix	350	210M	160M	76.2%
	Others[5]	10	210M	150M	71.4%
	Experian	5	210M	140M	66.7%
	Epsilon	14	210M	130M	61.9%
Australia	**All**	58	16M	13M	81.3%
	Experian	34	16M	12M	75.0%
	Acxiom	24	16M	9.1M	56.9%
U.K.	**All**	139	39M	29M	74.4%
	Acxiom	103	39M	22M	56.4%
	Datalogix	19	39M	17M	43.6%
	Experian	17	39M	15M	38.5%
Germany	Acxiom	60	31M	20M	64.5%
France	Acxiom	21	32M	18M	56.3%
Brazil	Experian	20	120M	61M	50.8%
Japan	Acxiom	17	25M	12M	48.0%

Table 1: Coverage of different data brokers across countries with partner categories. We show the total number of broker attributes, the number of Facebook identities that are targetable (Overall), the number of these identities that have at least one attribute from that broker (Partner), and the resulting coverage. Countries with more than one broker have a row indicating the coverage of all the brokers together (All).

We then estimate the coverage achieved by a broker within a particular population: *First*, we obtain the number of targetable identities corresponding to all users in that population (e.g., by obtaining the potential reach targeting that population). *Second*, we obtain the number of targetable identities within that population that have at least one attribute from that particular broker (we can do this since Facebook's ad interface allows us to target an OR of attributes, as mentioned in Section 2). While each of these numbers represents subsets of the overall population that are targetable, their ratio provides an estimate of the coverage in the overall population.

We are then able to study the coverage across different subpopulations (e.g., across genders, ages, locations, etc.), as Facebook allows us to additionally filter by those attributes (see Section 2).

Limitations Our methodology has a few limitations worth discussing. *First*, the population of targetable identities could be a biased sample of both the overall population of Facebook users, and of the overall population of that targeted country as a whole. We briefly explore this bias by focusing on the U.S., comparing the distribution of targetable identities with those of the overall population of the country (sourced from the U.S. Census Bureau's 2017 American Community Survey [ACS] 1-Year Estimates [3]). From Table 2, we see that the distribution across gender is similar for both targetable identities and for the U.S. population. However, we can see that the population of targetable identities is biased towards lower ages, with a larger percentage of people aged 44 or below than in the overall U.S. population. Nevertheless, this sample can still provide us a unique opportunity to study the coverage of data broker information across a large sample of users.

	Value	U.S. Census	Targetable	Surveyed (§4)
Age	15-24	13.3%	17.9%	22.9%
	25-34	13.8%	26.2%	37.1%
	35-44	12.6%	17.9%	22.4%
	45-54	13.0%	14.6%	8.7%
	55-64	12.9%	11.7%	7.1%
	65+	15.6%	10%	0.0%
Gender	Male	48.7%	47.8%	40.9%
	Female	51.3%	52.2%	56.8%

Table 2: Demographics of the entire U.S. population (U.S. Census), all targetable U.S. Facebook users (Targetable), and the crowdsourced users surveyed in Section 4 (Surveyed).

Second, data brokers might provide only a subset of their information to Facebook. However, Facebook has revealed that it requests "all available marketing segments they offer" from its partners [34], suggesting this is not the case. *Third*, the measured coverage of data brokers depends on the extent of linkage between Facebook profiles and offline data broker information, and thus can only provide a lower bound on the actual coverage of the data brokers. In the next section, we first demonstrate that there is a high extent of linkage between Facebook profiles and offline data broker information, thus suggesting our lower bounds are tight.

3.2 Analysis of coverage

We now examine the results on data broker coverage obtained using the methodology just described.

Extent of online-offline identity linkage We begin by examining the fraction of Facebook accounts that are linked with *any* data broker information. To do so, for each country, we measure the percentage of targetable Facebook identities that have *at least one* attribute from any of the brokers offering attributes within that country. This percentage reveals the extent of linkage of Facebook identities with offline information and serves as a lower bound.

Table 1 shows the percentage of targetable identities in different countries that have at least one data broker attribute; for countries with multiple data brokers, we include an **All** line that matches users from any of the data brokers. We make two observations: *First*, the extent of linkage is surprisingly high for the U.S., with over 90% of targetable Facebook identities having at least one data broker-provided attribute; similar high percentages of linkage are observed for Australia (81.3%), and for the U.K. (74.4%). *Second*, the extent of linkage is higher for countries where Facebook links to a larger number of data brokers (and obtains a larger number of attributes), indicating that even for the other countries, Facebook might have the ability to link a larger percentage of identities if it simply partnered with more data brokers. Taken together, these results suggest that offline and online data are being linked at massive scale. Having observed that our methodology can yield good lower bounds, we move on to further characterize the coverage.

Coverage of individual data brokers We measure the overall coverage achieved by various individual data brokers[5] in each of the seven countries we study; the results are presented in Table 1.

[5]We observed that 10 particular targeting attributes for the U.S. were labeled as coming from "data providers", indicating they are sourced from multiple brokers; in a previous crawl earlier in 2017, we observed that these were exclusively sourced from Acxiom. We mark these as "Others" when reporting results.

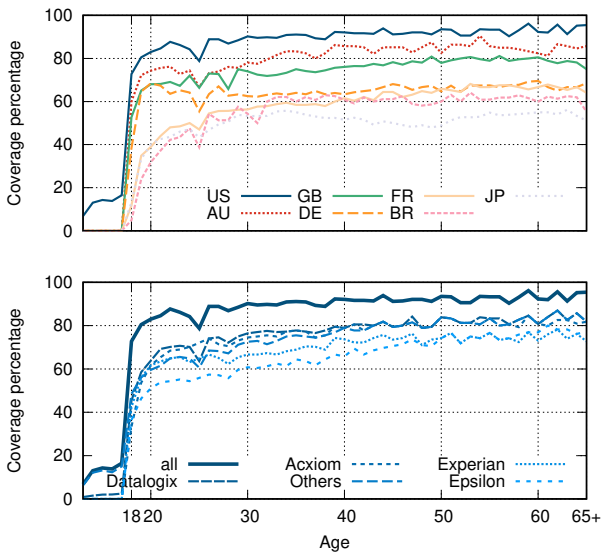

Figure 1: Breakdown of data broker coverage by age across multiple countries (top) and only within the U.S. (bottom).

We make three observations: *First*, we see that almost all the brokers achieve a coverage of at least half the targetable population in the country, irrespective of the country, showing that data brokers are successful in aggregating information about users at a global level. *Second*, we also see that the coverage can be as high as 76.2% (for Acxiom and Datalogix (now part of Oracle Data Cloud [39]) in the U.S.) and 75% (for Experian in Australia). *Third*, we see that the coverage achieved by the same broker can vary significantly (e.g., Experian covers over 66%, 75%, 38%, and 50% in the U.S, Australia, U.K., and Brazil, respectively).

Variation in coverage across ages One powerful feature of Facebook's advertising service is the ability to sub-divide any audience by additional features. As an example, we briefly study the variation of overall coverage of all data brokers in a country with users' age at the top of Figure 1. Similarly, we study the variation in coverage of individual data brokers from the U.S. at the bottom of Figure 1.

In either case, we study variation of coverage across ages taken from the set {13, 14, 15, ..., 65+}; these correspond to the ages that the interface allows us to target. We firstly observe that across countries, or across brokers in the U.S, the coverage rapidly increases between the ages of 18 and 20; this might be because these ages roughly correspond to the ages when people start acting with financial independence, and indicate that data brokers are rapidly able to cover people once they start acting as independent adults. We secondly observe that the coverage generally increases up to the age of 30, after which the rate of increase with age slows down; one exception is Epsilon in the U.S, whose coverage increases from 60.8% at age 30 to over 77.2% at age 65+. Finally, we observe that while the coverage for people below 18 years of age is zero for countries other than the U.S.; the corresponding coverage for the U.S. is non-zero, not dropping below 6.8% for the minimum studied age of 13. We observe that the data brokers in the U.S. with non-zero coverage for these users are Oracle Data Cloud (dropping from a

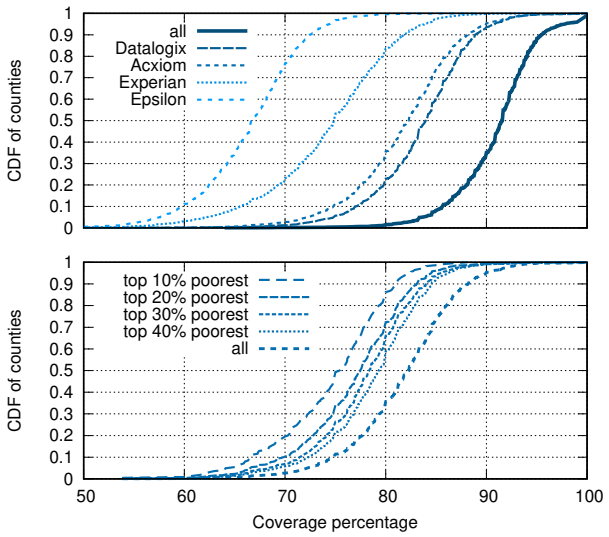

Figure 2: Cumulative distribution function (CDF) of data broker coverage across different U.S. counties (top), and of Acxiom for counties ranked by population below poverty level, as rated by the U.S. Census (bottom).

Category	Coverage
Behaviors → Purchase behavior	86.7%
Demographics → Financial → Income	76.9%
Demographics → Home → Home Ownership	72.3%
Behaviors → Automotive	68.7%
Demographics → Financial → Net Worth	67.6%
Behaviors → Financial → Spending methods	65.6%
Demographics → Home → Home Type	45.6%
Behaviors → B2B → Company size	44.1%
Demographics → Home → Home Type → Home value	43.1%
Behaviors → B2B → Industry	40.5%
Behaviors → Charitable donations	40.5%
Behaviors → Financial → Investments	40.0%
Demographics → Financial → Net Worth → Liquid assets	38.5%
Behaviors → Media → Television → Show Genre	28.7%
Behaviors → Travel	26.1%

Table 3: List of all 15 categories of data broker attributes provided on Facebook, along with the coverage (i.e., the fraction of targetable U.S. Facebook users who have at least one attribute in that category). We observe impressive coverage, with a significant variance across categories.

2.4% coverage for age 17 to a 0.8% coverage for age 13), and the brokers labeled by Facebook as "data providers", marked "Others" in the plot (dropping from 15% for age 17 to 6.3% for age 13).

Variation in coverage across U.S. counties We move on to study the variation of data broker categories across fine-grained geographic areas, plotting the CDF of the variation across different U.S. counties in Figure 2 (top). Since we cannot directly target counties on Facebook's ad platform—but we can target ZIP codes—we obtain a mapping from ZIP codes to counties [51]; we filter out military ZIP codes, P.O. box ZIP codes, and unique ZIP codes (dedicated for use by large organizations), giving us 29,916 ZIP codes in total. We further only consider ZIP codes from counties from the 50 states of the U.S. (3,110 counties in total). In the figure, we observe that the broker coverage varies widely across counties, roughly varying between 65% and 95% for Acxiom and Oracle Data Cloud, 50% and 80% for Epsilon, and 50% and 90% for Experian.

While on one hand, increased data broker coverage can have negative privacy implications, on the other hand there are a number of benefits to consumers from data broker coverage, such as better fraud prevention, improved product offerings, and more relevant (tailored) advertisements [20]. Thus, we study whether the extent of data broker coverage is correlated with the socioeconomic development of counties, specifically measured by the percentage of population in a county that is below the poverty level as per the 2012-2016 American Community Survey 5-Year Estimates [12]. In Figure 2 (bottom), we plot the CDF of coverage of one data broker (Acxiom) across the top percentiles of counties (in terms of having the highest fraction of population below poverty line), in addition to plotting the CDF of coverage across all counties. While we only show results for one data broker (owing to space constraints), we find similar results with other data brokers, and with the overall coverage of all data brokers taken together. We see that the coverage by Acxiom is indeed lower for counties with higher fractions of population below poverty line; for example, while 64.3% of all counties have at least 80% of their population covered by Acxiom, only 13.8% of those in the top 10th percentile have a similar coverage by Acxiom. This shows that data brokers indeed achieve lower coverage in counties with higher poverty rates.

Variation in coverage across attributes To understand the variation in coverage across various kinds of information, we group the 507 broker attributes available for the U.S., according to the second-to-last level of the hierarchy presented by Facebook, which we refer to as a *category*. Doing so gives us 15 categories in total where the "leaf" attributes are provided by data brokers. For example, the category Demographics → Financial → Net Worth → Liquid assets represents peoples' total liquid assets; all of the leaf attributes in this category together cover 100M people, and there are nine different leaf attributes including $1-$25K (covering 23M people), $25K-$50K (14M people), ..., up to >$3M (2.7M people). Similarly, the category Behaviors → Purchase behavior represents the kinds of purchases people make; it covers 169M people and has 175 different leaf attributes including Coupon users (100M people), Childrens' cereals (24M people), and Over-the-counter medication (43M people).

For each category, Table 3 shows the percentage of the U.S. population that is covered by at least one attribute corresponding to that category. We find that the coverage can vary significantly with the kind of targeting information, and can be as high as 86.7% for purchase behavior, and higher than 65% for sensitive categories like net worth, spending methods, home ownership, and income, but much lower 38.5% for net worth in the form of liquid assets.

4 ACCURACY

We move on to study the accuracy of data broker attributes, focusing on U.S.-based users. We first discuss the methodology we use before diving into the results.

4.1 Methodology for studying accuracy

Revealing user attributes to users While Facebook reveals Facebook-derived attributes, it has been shown to not reveal users' data broker attributes [2]. Thus, we leverage a recently-proposed mechanism that uses Facebook's advertising platform to enforce transparency on itself. Called *Transparency-enhancing advertisements* (Treads), the mechanism allows an advertiser to reveal platform-collected information about users by targeting them with ads that contain information about their targeting [53]. Specifically, we reveal to users (study participants) whether they have a particular data broker attribute by running one ad targeting all study participants, and another ad targeting only those participants who have that particular data broker attribute. If a user receives the first ad, we know they are targetable; if they receive the second as well, we know they have the data broker attribute.

To be able to target participants, we ask them to 'Like' our study's Facebook page. We then target Treads to the set of users who liked our Facebook page, with one Tread for each of the 507 data broker categories in the U.S (from Section 3.1). Since explicitly including the categories in the ad may be against Facebook's advertising policies [53], we encode each attribute by a unique numeric identifier, which we then translate into innocuous-looking text. For example, we might assign a particular attribute the identifier 852, which is then translated into "We have 8 and 52 ideas. Stay tuned."

Data collection In order to decode these Treads and to infer a user's data broker attributes, we built a browser extension for Google Chrome. The extension scans for our ads on Facebook as the user normally browses, and locally stores only the numeric identifiers in the Treads that are seen (for privacy reasons).

We then use our extension to survey users about the accuracy of their data broker attributes. As a baseline, we also survey users about the accuracy of information about them collected internally by Facebook. To do so, we gather the user's Facebook ad preferences page [28] just prior to the survey; we call these *Facebook attributes*.

When surveying participants, we randomly shuffle the data broker attributes and Facebook attributes, without revealing which is which to users. In order to avoid user fatigue, we survey them on a random sample of at most 25 data broker attributes and at most 30 attributes overall. For each user attribute, we show the user the attribute, and ask them questions about its accuracy. Specifically, we ask users whether the inferred attribute is "Not at all accurate", "Somewhat accurate", "Mostly accurate", or "Completely accurate." For data broker attributes, we also show Facebook's description of the attribute to the user.[6] For Facebook attributes, this description is not available; hence, we only show the attribute. Finally, we ask about their demographic information. Users are asked to answer only those questions they feel comfortable answering.

Changed targeting attributes Early-on during our experiments in this section (in August 2018), we found Facebook stopped supporting 120 out of the 507 original targeting attributes. [7] By freshly obtaining the set of targeting attributes from the API (in the same way as in Section 3.1), we found that each of the above 120 attributes was replaced by a similar attribute. However, the partner for all these attributes was now described as a generic "data license", perhaps indicating they were being sourced from multiple data brokers. In any case, we ran an additional 120 Treads for these attributes, and report accuracy for these under the column of "Others."

[6] For example, the attribute Behaviors → Charitable donations → Veterans, has the description, "People who are interested in donating to veteran causes."

[7] Of these 120 attributes, 96 were sourced from Acxiom (corresponding to information such as net worth, investments, spending methods, home type, charitable donations etc.), 21 were from Oracle (corresponding to information such as travel), two from Experian (corresponding to information such as home ownership) and one from Epsilon (corresponding to the length of residence).

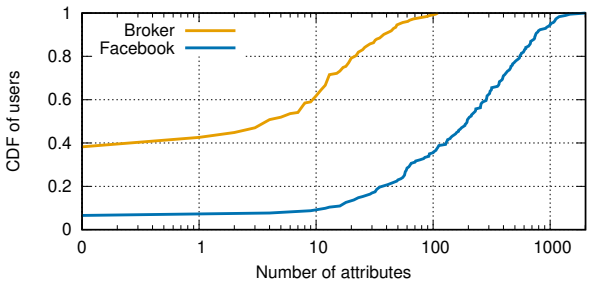

Figure 3: Distribution of inferred broker and Facebook attributes across the 183 participants who took the survey.

4.2 User recruitment

Our user study was reviewed and approved by Northeastern University's Institutional Review Board (IRB). We recruited users from Prolific [44], a crowdsourcing platform geared towards researchers; the quality of data obtained from Prolific users has been found to be comparable to that of the dominant crowdsourcing platform, Amazon Mechanical Turk [41]. We used Prolific's filters to selectively target our study at users aged 18 and above, residing in the U.S., and who use Facebook on a regular basis (at least once a month). We also followed Prolific's guidelines to selectively obtain high-quality participants [18], only considering users who had at least 10 prior submissions on Prolific, and an approval rating of at least 95 out of 100. This gave us a pool of over 4,000 potential participants, from which we recruited 300 users.

Participants were invited to participate in a research study about their reactions to their Facebook information (we did not mention data brokers until the end of the study so as not to bias participants). Participants were first asked to install our extension, for which they were paid $2. Participants were also asked to browse Facebook as they normally would in the subsequent days, and were informed how our extension works. After a period of a month, we followed up with the 231 users that still had our extension correctly installed, and requested them to take the survey (via the extension). We compensated participants $8 for completing the survey. 183 users completed the survey; this attrition rate of 39% (compared to the initial 300 users recruited) is well within the range of attrition rates observed in longitudinal studies on Prolific [18].[8]

Figure 3 shows the distribution of the number of data broker and Facebook attributes for these users who took the survey.[9] While around 61% of these users had at least one data broker attribute, 41% of the users had 10 or more data broker attributes, with the median user having four attributes. The median participant had 196

[8] The attrition was partially due to delays (of over a week) on our part in setting up the survey due to technical issues, and due to some users mistakenly assuming that the study had ended.

[9] Due to a bug in our extension in the early stages of our extension (where it failed to scan some users' Facebook home pages for our Treads, we potentially missed some users' data broker attributes. Thus, the plot potentially shows a lower bound on the actually revealed users' broker attributes.

Facebook attributes; given our limits on the number of attributes surveyed, the median number that we surveyed them on was 22.

Participant demographics To quantify the bias in the set of surveyed participants, we compare their demographics with those of the overall U.S. population [3]. For age/gender, we compare against the population of targetable Facebook identities (see Section 3).

Table 2 compares the age and gender distribution of participants. We find 56.8% of participants are female, as opposed to 52.2% of targetable Facebook identities and 51.3% of the entire U.S. population. We also find 82.4% of participants are younger than 45, as opposed to 62% of the targetable Facebook identities, and 39.7% of the offline population. Overall, this indicates a slight bias towards women, and a larger bias towards younger people, that is qualitatively similar to the bias of targetable Facebook identities.

We find that 63.4%, 8.1%, and 8.7% of our participants reported a single race of White, Black, and Asian, respectively. Thus our participant pool has a smaller percentage of White and Black participants and a larger percentage of Asian participants compared to the general U.S. population which has 72.3%, 12.7%, and 5.6% respectively of these races. 12.6% of our participants reported themselves as "Hispanic or Latino", as opposed to 18.1% of the overall population. Finally, our partipants are spread across 39 U.S. states.

Limitations Inaccuracy in data broker information revealed via Facebook's advertising platform can arise from two sources: (i) inaccuracy in the information collected by data brokers, and (ii) inaccuracy in linking data broker information to Facebook accounts. Thus, our methodology measures the cumulative effect of these two sources of inaccuracy, providing an upper bound on the inaccuracy of information collected by data brokers. However, as discussed in Section 2, Facebook only uses unique identifiers such as phone numbers and email addresses to match data broker information [34], thus indicating that the inaccuracy introduced due to the linking process is likely limited. Thus, our upper bounds are likely tight.

Additionally, the information exposed by Facebook via its advertising platform may only cover a subset of information held by data brokers. However, recent reports have stated that Facebook receives "all available marketing segments offered for [each] user" from data brokers [34]. Moreover, since our methodology relies on Treads to reveal user attributes, not all data broker attributes of each user might be successfully revealed (e.g., if some Treads that are not delivered). Finally, our methodology relies on the accuracy of attributes as reported by participants; while each participant is the best source for evaluating the accuracy of their own data, they might misunderstand what certain attributes mean, may lie, or might be subject to fatigue (which we try to limit). Besides, participants might not recall past behavior such as purchases, or imperfectly estimate information such as their financial data.

4.3 Analysis of accuracy

We next analyze the responses from participants in order to understand the accuracy of information collected by data brokers.

Overall accuracy of attributes We begin by analyzing the overall accuracy of the collected data broker attributes, shown in Table 4. We find a largely bimodal distribution: 40.6% of the 1,432 surveyed broker attributes are reported "Completely accurate", while 40.5% of them are reported "Not at all accurate." By contrast, a higher percentage (proportion test, $p < 0.001$) of Facebook attributes are reported "Completely accurate" (51.7% of the 3,464 surveyed attributes). Also, a lower percentage ($p < 0.001$) of Facebook attributes are reported "Not at all accurate" (25.3% of the 3,464 surveyed attributes). This may be due to the significantly larger amount of data that Facebook can directly observe in users' activities, when compared to data brokers. The low accuracy of data broker information has significant implications, given the widespread use of data broker information for various purposes, including for background checks.

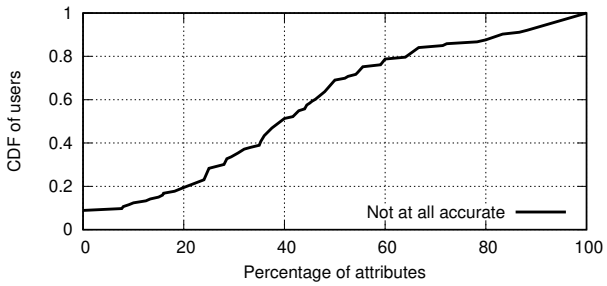

Figure 4: CDF of the fraction of users' data broker attributes rated as "Not at all accurate." We see a wide distribution.

Are inaccurate attributes stale? To study whether data broker attributes are simply inaccurate, or were previously accurate but is now out of date (*stale*), we ask users whether attributes they rated as anything other than "Completely Accurate" were accurate in the past. Specifically, we show the distribution of user responses to the question "Do you think this attribute would have accurately described you at a previous time? (For example, would it have been more accurate last year?)" in Table 5. We see that around 24% of incorrect attributes from data brokers are stale (with users responding with either "Yes" or "Probably yes"). In contrast, the percentage of incorrect attributes from Facebook that are stale is higher (32.3%). Thus, data broker attributes appear to suffer from many inaccuracies, only partially explained by out-of-date data.

User-level accuracy Inaccuracy in a user's attributes might often be because of data brokers' inability to uniquely or correctly identify users from different source databases (e.g., the user might have a very common name). Thus, it might be expected that users' attributes are inaccurate in an all-or-none fashion (i.e., that either almost all a user's attributes are incorrect, or almost all a user's attributes are correct). To study whether this is the case, we show the variation in accuracy by user in Figure 4.

Contrary to our expectation, we see that users typically have a subset of their broker attributes inaccurate, with the percentage of attributes marked "Not at all accurate" varying widely between 0% and 100%, with the median user having around 40% of their broker attributes marked "Not at all accurate." Besides, more than 90% of users who have data broker attributes have a non-zero fraction of their attributes marked "Not at all accurate." This means that inaccuracy of data broker attributes is a problem potentially affecting most people (rather than an unfortunate few).

Variation in accuracy across attributes Finally, to study how the accuracy of broker attributes varies across types or categories

	Data Brokers	Acxiom	Datalogix	Experian	Epsilon	Others	Facebook
Responses	1,432	95	728	55	18	536	3,464
Not at all accurate	40.5% ± 2.5%	27.4% ± 9.0%	42.0% ± 3.6%	21.8% ± 10.9%	55.6% ± 23.0%	42.2% ± 4.2%	25.3% ± 1.4%
Somewhat accurate	13.6% ± 1.8%	14.7% ± 7.1%	14.4% ± 2.6%	20.0% ± 10.6%	0.0% ± 0.0%	12.1% ± 2.8%	15.2% ± 1.2%
Mostly accurate	5.2% ± 1.2%	4.2% ± 4.0%	6.2% ± 1.7%	12.7% ± 8.8%	5.6% ± 10.6%	3.4% ± 1.5%	7.9% ± 0.9%
Completely accurate	40.6% ± 2.5%	53.7% ± 10.0%	37.4% ± 3.5%	45.5% ± 13.2%	38.9% ± 22.5%	42.4% ± 4.2%	51.7% ± 1.7%

Table 4: Aggregated user-reported accuracies for attributes from different data brokers and Facebook. The first column shows the accuracies for attributes from all data brokers taken together. 95% confidence intervals are shown for sample percentages.

	Data Brokers	Acxiom	Datalogix	Experian	Epsilon	Others	Facebook
Responses	833	44	445	28	11	305	1,596
Yes/Probably yes	24.0% ± 2.9%	29.5% ± 13.5%	23.8% ± 4.0%	32.1% ± 17.3%	36.4% ± 28.4%	22.3% ± 4.7%	32.3% ± 2.3%
No/Probably not	76.0% ± 2.9%	70.5% ± 13.5%	76.2% ± 4.0%	67.9% ± 17.3%	63.6% ± 28.4%	77.7% ± 4.7%	67.7% ± 2.3%

Table 5: Aggregated user-responses about the past accuracies for attributes from different data brokers and from Facebook. 95% confidence intervals are shown for sample percentages.

of information, we consider the variation of accuracy across the 15 higher-level "categories" in Facebook's hierarchy (from Section 3.2). Since we have only a relatively small number of responses for some categories, we group together categories that have fewer than 30 responses and which have a common ancestor in the hierarchy. For example, we group together Income, Net worth, and Liquid assets, which have a common ancestor (Demographics → Financial).

We show the distribution of responses in Table 6. We make multiple observations: *First*, we observe mediocre accuracy across all categories. Indeed, even categories involving important financial information such as income/net worth/liquid assets, or home ownership/home type/home value can have a high degree of inaccuracy, with 47.2% and 42.3% of respective attributes marked "Not at all accurate." This has important implications as this information might be used in important situations such as credit or background checks.

Second, we observe that some categories have higher accuracies than others. For example, the Purchase behavior category has a higher percentage of attributes marked "Completely accurate" than the Automotive category (proportion test, $p < 0.001$), and even the Income/Net worth/Liquid assets category ($p < 0.002$). While it is hard to explain why this might be the case, one potential reason is that purchase behavior is often tracked via loyalty cards provided by the data broker themselves [15], thus tracking users more directly.

Third, we find a less bimodal distribution for some categories than others. For example, the Purchase behavior, Income/Net worth/Liquid assets, and Television → Show Genre categories all have a larger percentage of intermediate accuracy levels ("Somewhat accurate" or "Mostly accurate") than the Automotive, and Spending methods categories (with $p < 0.006$ for all six pairwise comparisons). We explain why this might be the case in the following section.

Qualitative analysis of inaccuracy To better characterize how certain attributes are inaccurate, we briefly analyze participants' (free-text) responses to the question "What is incorrect about it?", asked for attributes that they did not mark "Completely accurate." Participants provided responses for 822 out of the 833 inaccurate broker attributes; however, a majority of these responses simply stated that the attribute did not correspond to the participant. We make a number of qualitative observations from the remaining responses (which are thus limited to a subset of all responses).

First, participants often reported errors of degree for categories like Purchase behavior, Income/Net worth/Liquid assets, and Television → Show Genre. For example, for attributes such as "food enthusiast" or "gadget enthusiast" that fall under the Purchase behavior category, participants responded that they might like food or occasionally purchase gadgets, but not to the extent of an enthusiast. Similarly, participants indicated that they only occasionally watched shows of a particular television show genre. For attributes under the Income/Net worth/Liquid assets category, which typically specify a range of income or net worth, participants indicated that the ranges were incorrect by different margins (ranging from "close" to "an order of magnitude below"). This explains why such categories might have less bimodal distributions compared to other categories such as Automotive, and Spending methods, which typically have attributes that are more likely to incur binary errors (e.g., make of car owned).

Second, we observed some anecdotal reasons from participants explaining how particular attributes were stale (i.e., accurate in the past but no longer so). Some responses indicated users changing their purchase behavior in response to changed life circumstances: two participants had purchased particular children's food items when they had children, or cared for children in the past. Similarly, a few participants had purchased pet products when they owned a pet. Finally, one participant had stopped buying a particular kind of frozen food in order to be more healthy. Other responses indicated reasons for changes in finance-related information such as home value (no longer accurate due to appreciation), or particular kinds of investments (participant no longer had that kind of investment).

Third, for categories like Income, Net worth, Liquid assets, or Home value, participants reported having both higher and lower values compared to their data broker attributes (which show a particular range of values for that category). While two participants reported having higher and lower incomes each, six participants reported having lower values for net worth. Nine participants reported having fewer liquid assets (of which two reported that their family might collectively have a matching amount of liquid assets), while three participants reported having more liquid assets. Finally, two participants reported having homes with higher values.

Fourth, we found some egregious errors (either in magnitude or sensitivity). For example, one unemployed student was inferred

Category	Total	Not at all accurate	Somewhat accurate	Mostly accurate	Completely accurate
Purchase behavior	541	26.1% ± 3.7%	18.7% ± 3.3%	7.0% ± 2.2%	48.2% ± 4.2%
Income/Net worth/Liquid assets	53	47.2% ± 13.4%	20.8% ± 10.9%	5.7% ± 6.2%	26.4% ± 11.9%
Home Ownership/Home type/Home value	111	42.3% ± 9.2%	6.3% ± 4.5%	1.8% ± 2.5%	49.5% ± 9.3%
Automotive	252	60.7% ± 6.0%	8.7% ± 3.5%	4.0% ± 2.4%	26.6% ± 5.5%
Spending methods	141	35.5% ± 7.9%	6.4% ± 4.0%	2.1% ± 2.4%	56.0% ± 8.2%
B2B Company size/Industry	48	66.7% ± 13.3%	6.2% ± 6.8%	8.3% ± 7.8%	18.8% ± 11.0%
Charitable donations	27	40.7% ± 18.5%	18.5% ± 14.7%	3.7% ± 7.1%	37.0% ± 18.2%
Financial → Investments	50	56.0% ± 13.8%	14.0% ± 9.6%	4.0% ± 5.4%	26.0% ± 12.2%
Television → Show Genre	119	37.8% ± 8.7%	21.0% ± 7.3%	5.9% ± 4.2%	35.3% ± 8.6%
Behaviors → Travel	38	39.5% ± 15.5%	10.5% ± 9.8%	5.3% ± 7.1%	44.7% ± 15.8%

Table 6: Aggregate user-reported accuracies for attributes corresponding to different "categories" of information. The first column shows a description of the different categories of information, while the remaining columns show the aggregate distribution of user responses. 95% confidence intervals are shown for sample percentages.

to be a C-Suite (i.e., top level) executive, while another two unemployed participants (one of whom was a student) were inferred to be corporate executives. Finally, a few participants who were non-drinkers were inferred to be likely to buy alcoholic beverages.

5 DISCUSSION

Implications of findings Recent policy is moving in the direction of requiring greater transparency and accuracy of data collection; for example, the principles of the General Data Protection Regulation (GDPR) in the European Union require organizations collecting personal data to ensure "lawfulness, fairness and transparency" and to ensure "the personal data is accurate and up-to-date, having regard to the purposes for which it's processed, and correct it if not" [43]. In this light, the findings in this paper about the coverage and lack of accuracy of information collected by data brokers calls for greater efforts from the brokers to (a) allow people to control and correct what data is collected about them, and (b) warn consumers of their data about cases where the data may not be accurate.

Future applicability of methodology Facebook removed partner categories as of October 2018 [4], meaning that it is no longer possible to study the coverage and accuracy of data brokers via Facebook's advertising platform. However, our methodology is still applicable on other platforms such as Twitter, which offers similar categories sourced from Acxiom and Oracle Data Cloud [32].

In addition, our methodology could also be used to study the coverage and accuracy of information internally gathered by online advertising platforms such as Facebook, Google, and Twitter, themselves. The existing transparency mechanisms of these platforms, revealing collected user information, have been shown to reveal an incomplete view to users [2, 17, 54]. Thus, our methodology for studying accuracy using Treads could potentially reveal the accuracy of the data that these platforms have actually collected.

Ethics All our experiments were performed as per community ethical standards. Our experiments measuring the coverage of data brokers only rely on aggregate statistics about users, and thus do not involve collection of data about individual users.

For our experiments measuring the accuracy of data broker information (approved by Northeastern University's IRB), we take multiple precautions to ensure the privacy of participants. *First*, our extension stores inferred attributes about participants locally (using their browser's local storage), and does not collect any personally identifying information from users. *Second*, participants are only required to answer questions they are comfortable answering; any attributes that participants are not comfortable answering questions about are not shared with our server. *Third*, we store the completed surveys in anonymized form. *Fourth*, users who unexpectedly see ads that explicitly state or imply their personal attributes might consider this a privacy violation. We address this concern by targeting the Treads that we run only to participants who have read our consent forms and signed up for our study, and obfuscating the targeting information in these Treads (as opposed to explicitly stating the targeting information directly).

6 CONCLUSION

The extent of data collection by online sites and offline data brokers, and the extent of linkage between online and offline identities of users are both crucial inputs to ongoing debates about user privacy. This paper used Facebook's advertising platform to (a) measure for the first time the actual extent of linkage between online (Facebook) identities and offline identities (as compiled by data brokers), (b) measure for the first time the coverage of data brokers from seven countries in a fine-grained manner, and (c) measure the accuracy of information from data brokers for U.S.-based users. Taken together, our analysis and results of the fine-grained coverage of data brokers provide new insights into the workings of data brokers, that can feed into ongoing privacy debates. We hope this paper serves as a starting point for more such analyses in the future.

7 ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their helpful comments. This research was supported in part by the Data Transparency Lab, by NSF grant CNS-1616234, by the French National Research Agency (ANR) through the ANR-17-CE23-0014 grant, by the ERC Advanced Grant "Foundations for Fair Social Computing" (No. 789373), and by a UMIACS contract under the partnership between the University of Maryland and DoD. Additionally, Elissa Redmiles acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and a Facebook Fellowship.

REFERENCES

[1] Facebook Algorithms and Personal Data. *PewResearchCenter*, 2019. http://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/.

[2] A. Andreou, G. Venkatadri, O. Goga, K. P. Gummadi, P. Loiseau, and A. Mislove. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. *NDSS*, 2018.

[3] ACS Demographic and Housing Estimates (2017). https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_DP05&prodType=table.

[4] About Partner Categories. https://web.archive.org/web/20180911174110/https://www.facebook.com/business/help/298717656925097.

[5] Acxiom. https://www.acxiom.com/.

[6] Acxiom Data Catalog. http://www.stephenfortune.net/projects/puppetry/data_products_catalog.pdf.

[7] Acxiom Global Coverage. https://www.acxiom.com/news/acxiom-launches-global-data-navigator-tool-offering-marketers-visibility-into-global-audiences/.

[8] Acxiom Healthcare Portfolio. https://www.acxiom.com/news/acxiom-expands-healthcare-solutions-portfolio-new-patients-insights-package/.

[9] Acxiom InfoBase. https://www.acxiom.com/what-we-do/infobase/.

[10] Acxiom transparency website. https://aboutthedata.com/.

[11] Advanced Targeting and Placement. https://developers.facebook.com/docs/marketing-api/targeting-specs.

[12] American Fact Finder. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

[13] T. Bergin. How a data mining giant got me wrong. https://www.reuters.com/article/us-data-privacy-acxiom-insight/how-a-data-mining-giant-got-me-wrong-idUSKBN1H513K.

[14] M. A. Bashir, U. Farooq, M. Shahid, M. F. Zaffar, and C. Wilson. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. *Proceedings of the Network and Distributed System Security Symposium (NDSS 2019)*, 2019.

[15] J. Constine. Datalogix Raises $25M to Pump Juicy Offline Purchase Data Into Google and Facebook. https://techcrunch.com/2013/04/25/datalogix-offline-purchase-data/.

[16] J. G. Cabañas, A. Cuevas, and R. Cuevas. Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. *27th USENIX Security Symposium (USENIX Security 18)*, USENIX Association, 2018.

[17] A. Datta, M. C. Tschantz, and A. Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *PETS*, 2015.

[18] K. Damer. What is your dropout / attrition rate in longitudinal studies? http://help.prolific.ac/getting-started-launching-your-first-study/what-is-your-dropout-attrition-rate-in-longitudinal-studies.

[19] Data Broker Axciom's Flawed Transparency Effort. *MIT Technology Review*. https://www.technologyreview.com/s/519111/data-broker-axcioms-flawed-transparency-effort/.

[20] Data Brokers: A Call for Transparency and Accountability. https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014.

[21] Epsilon. https://www.epsilon.com.

[22] Epsilon U.S. Coverage. https://us.epsilon.com/data-driven-marketing-solutions/people-based-marketing-data-solution.

[23] Experian. https://www.experian.com.

[24] Experian Corporate Factsheet. https://www.experian.com/corporate/experian-corporate-factsheet.html.

[25] Experian Free Credit Report. https://www.experian.com/consumer-products/free-credit-report.html.

[26] Experian Services Catalog. https://www.experian.com/assets/data-university/brochures/ems-list-services-catalog.pdf.

[27] I. Faizullabhoy and A. Korolova. Facebook's Advertising Platform: New Attack Vectors and the Need for Interventions. *CoRR*, https://arxiv.org/abs/1803.10099, Workshop on Technology and Consumer Protection (ConPro), 2018.

[28] Facebook Ads Preferences. https://www.facebook.com/ads/preferences.

[29] Facebook Advertising. https://www.facebook.com/ads/create/.

[30] D. Garcia, Y. M. Kassa, A. Cuevas, M. Cebrián, E. Moro, I. Rahwan, and R. Cuevas. Facebook's gender divide. *CoRR*, https://arxiv.org/abs/1710.03705, 2017.

[31] Google AdWords. https://adwords.google.com/.

[32] Introducing partner audiences (Twitter). https://blog.twitter.com/marketing/en_us/a/2015/introducing-partner-audiences.html.

[33] A. Korolova. Privacy Violations Using Microtargeted Ads: A Case Study. *Journal of Privacy and Confidentiality*, 3(1), 2011.

[34] K. Leetaru. The Data Brokers So Powerful Even Facebook Bought Their Data - But They Got Me Wildly Wrong. Forbes, 2018. https://www.forbes.com/sites/kalevleetaru/2018/04/05/the-data-brokers-so-powerful-even-facebook-bought-their-data-but-they-got-me-wildly-wrong/.

[35] A. Matheus, M. Yelena, W. Ingmar, and B. Fabricio. Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. *ACM WebSci*, 2017.

[36] T. Minkus, Y. Ding, R. Dey, and K. W. Ross. The City Privacy Attack: Combining Social Media and Public Records for Detailed Profiles of Adults and Children. *ACM COSN*, 2015.

[37] C. R. Miller. I Bought a Report on Everything That's Known About Me Online. https://www.theatlantic.com/technology/archive/2017/06/online-data-brokers/529281/.

[38] Mapping, and Sharing, the Consumer Genome. http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html.

[39] Oracle Data Cloud. https://www.oracle.com/applications/customer-experience/data-cloud/.

[40] Oracle Data Cloud Data Directory. http://www.oracle.com/us/solutions/cloud/data-directory-2810741.pdf.

[41] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 2017.

[42] Predictably inaccurate: The prevalence and perils of bad big data. https://www2.deloitte.com/insights/us/en/deloitte-review/issue-21/analytics-bad-data-quality.html.

[43] Principles of the GDPR. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr_en.

[44] Prolific - Home Page. 2018. https://prolific.ac/.

[45] A. Rao, F. Schaub, and N. M. Sadeh. What do they know about me? Contents and Concerns of Online Behavioral Profiles. *CoRR*, abs/1506.01675, 2015.

[46] L. Sweeney. Discrimination in Online Ad Delivery. *SSRN*, 2013.

[47] N. Singer. Acxiom Lets Consumers See Data It Collects. https://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html.

[48] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove. On the Potential for Discrimination in Online Targeted Advertising. *FAT**, 2018.

[49] Shutting Down Partner Categories. https://newsroom.fb.com/news/h/shutting-down-partner-categories/.

[50] M. C. Tschantz, S. Egelman, J. Choi, N. Weaver, and G. Friedland. The accuracy of the demographic inferences shown on Google's Ad Settings. *WPES*, 2018.

[51] United States ZIP Codes. https://www.unitedstateszipcodes.org/.

[52] G. Venkatadri, Y. Liu, A. Andreou, O. Goga, P. Loiseau, A. Mislove, and K. P. Gummadi. Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface. *IEEE S&P*, 2018.

[53] G. Venkatadri, A. Mislove, and K. P. Gummadi. Treads: Transparency-Enhancing Ads. *HotNets*, 2018.

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

[55] E. Zagheni, I. Weber, and K. Gummadi. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Popul. Dev. Rev.*, 43(4), 2017.