Comments on
Competition and Consumer Protection in the 21st Century Hearings, Project Number P181201

Alan Mislove

August 4, 2018

These comments are in response to the FTC's request for comments[1] *on competition and consumer protection in the 21st century, and are most closely aligned with topic 4: The intersection between privacy, big data, and competition.*

Advertising now funds most of the popular web sites and internet services: companies including Facebook, Twitter, and Google all provide their services for free, in exchange for collecting data from their users as they interact with the service. One of the primary explanations for the success of these online advertising platforms is their ability to allow advertisers to carefully control *targeting*, or choosing who sees their ads. In contrast with classical advertising media—such as radio, television, and print advertising—online advertising platforms offer a wide variety of mechanisms that advertisers can use to carefully select to whom their ads are shown.

Recently, many advertising services have introduced advanced mechanisms that enable significantly more fine-grained targeting for advertisers. For example, advertisers can now directly target individual users using their personal information on many advertising platforms [2, 11, 9, 7, 3]: commonly called *personally identifiable information (PII)-based targeting*, many services allow the advertiser to upload users' PII or other uniquely-identifying information. Once they do so, the platform then matches the uploaded data against their database and allows the advertiser to advertise to just those users who match. On Facebook, advertisers create custom audiences by literally uploading a comma-separated-values (CSV) file containing up to 15 different types of PII [4] (names, dates of birth, addresses, emails, phone numbers); on Google, they upload a CSV file with up to 7 different types of PII [6]. Other advanced targeting features include pixel-based targeting (using web tracking pixels on third-party sites), "lookalike" targeting (where the advertising platform finds other users similar to a specified group), and physical location targeting (using data collected from mobile apps).

As a technologist, I am concerned about how these advanced targeted advertising features may be misused in order to adversely impact consumers. My research group at Northeastern University has carefully studied multiple such targeting features and our results have demonstrated some of the risks that these services represent, including leaking users' PII and enabling hard-to-detect forms of discrimination. Making this situation worse is the fact that most of these advertising services are open to *any* internet user, meaning most targeting features are available to anyone who signs up as an advertiser (an action that is typically free).

In particular, there are three axes along which our research group has found that targeted advertising services may be used to consumers' detriment, outlined in more detail below:

Limited transparency Users are often unaware of the extent of targeted advertising, a fact that is compounded by the services' general lack of transparency. For example, our recent study ([1], also attached as Attachment #1) published at the NDSS'18 conference carefully examined two of the transparency features offered by Facebook: (1) ad preferences, and (2) ad explanations. We found both of these to be limited, and not offer users a complete view of the data that the services collect about them, or how that data is used by advertisers. In particular, Facebook's ad preferences purports to show users all of the targeting attributes that Facebook has attached to their account; in fact, this service does not show any of the more than 500 targeting attributes that are provided by third-party data brokers (e.g., Acxiom, Oracle Data Cloud, etc) and made available to advertisers. Additionally, Facebook's ad explanations purports to tell users why they were served a given ad, but in fact, the service is incomplete (showing at most one targeting attribute) and sometimes misleading (saying that advertisers "may have" used particular targeting attributes when they did not).

[1] https://www.ftc.gov/news-events/press-releases/2018/06/ftc-announces-hearings-competition-consumer-protection-21st

Unknown security and privacy implications Features like Facebook's custom audiences effectively represent powerful *linking* mechanisms, allowing advertisers to leverage any PII (e.g., from customer data, public records, etc.) to target users. We have carefully studied Facebook's custom audience implementation and demonstrated weaknesses ([10], also attached as Attachment #2) that allow an adversary to exploit the interface to infer users' PII as well as to infer their activity. Specifically, we have shown how an adversary can infer users' full phone numbers knowing just their email address, determine whether a particular user visited a website, and de-anonymize all the visitors to a website by inferring their phone numbers en masse. These attacks can be conducted without any interaction with the victim(s), cannot be detected by the victim(s), and do not require the adversary to spend money or actually place an ad. We proposed a simple and effective fix to the attacks based on reworking the way Facebook de-duplicates uploaded information. After disclosing these attacks to Facebook, they removed certain features of the interface that enabled them; however, our results indicate that PII-based targeting mechanisms may present unique security and privacy vulnerabilities that can lead to direct harms to consumers.

Disadvantages consumers There are special legal protections around how advertisements can be placed that concern housing (the Fair Housing Act), employment (the Age Discrimination in Employment Act), and credit (the Fair Credit Reporting Act). Recently, groups including ProPublica [5] have shown that advertisers on Facebook are able discriminate against users belonging to sensitive groups (i.e., to exclude users belonging to a certain race or gender from receiving their ads). Such criticisms have led, for instance, Facebook to disallow the use of attributes such as ethnic affinity from being used by advertisers when targeting ads related to housing, employment, or financial services. We have examined other targeting features of Facebook's interface and shown ([8], also attached as Attachment #3) that such measures are far from sufficient and that the problem of discrimination in targeted advertising is much more pernicious. In particular, custom audiences could trivially be used by advertisers implement discriminatory filtering; with the advertiser uploading the PII of the resulting discriminatory set to Facebook. Instead of relying on banning particular processes, we argue that discrimination measures should be based on the ultimate targeted population and not on the attributes used for targeting. Overall, our findings call for exploring fundamentally new methods for mitigating discrimination in online targeted advertising.

Taken together, our findings indicate that these advanced targeted advertising features carry substantial risks to consumers, and that the FTC should carefully consider how to adequately protect consumers as the popularity and ubiquity of these services grows.

About the author

Alan Mislove is an Associate Professor and Associate Dean and Director of Undergraduate Programs at the College of Computer and Information Science at Northeastern University, which he joined in 2009. He received his B.A., M.S., and Ph.D. in computer science from Rice University in 2002, 2005, and 2009, respectively. Prof. Mislove's research concerns the security, privacy, and efficiency of newly emerging systems. He work comprises over 50 peer-reviewed papers, has received over 11,000 citations, and has been supported by over $5M in grants from government agencies and industrial partners. He is a recipient of an NSF CAREER Award (2011), a Google Faculty Award (2012), the ACM SIGCOMM Test of Time Award (2017), the USENIX Security Distinguished Paper Award (2017), the NDSS Distinguished Paper Award (2018), the IEEE Cybersecurity Award for Innovation (2017), and his work has been covered by the Wall Street Journal, the New York Times, and the CBS Evening News.

References

[1] 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.

[2] About Customer Match. https://support.google.com/adwords/answer/6379332?hl=en.

[3] Audience Targeting. https://help.pinterest.com/en/articles/targeting.

[4] Custom Audiences from your Customer List. https://www.facebook.com/business/help/606443329504150.

[5] Facebook Lets Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race/.

[6] Format your customer data file. https://support.google.com/adwords/answer/7659867.

[7] LinkedIn Contact Targeting. https://business.linkedin.com/marketing-solutions/ad-targeting/contact-targeting.

[8] 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.

[9] Target Custom Groups of Twitter Users. https://business.twitter.com/en/targeting/tailored-audiences.html.

[10] 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.

[11] What's a Custom Audience from a Customer List? https://www.facebook.com/business/help/341425252616329/.

ATTACHMENT #1

Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations

Athanasios Andreou[§], Giridhari Venkatadri[†], Oana Goga[*], Krishna P. Gummadi[‡], Patrick Loiseau[*‡], Alan Mislove[†]
[§]EURECOM, France

[†]Northeastern University, USA

[*]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, France

[‡]Max Planck Institute for Software Systems (MPI-SWS), Germany

***Abstract*—Targeted advertising has been subject to many privacy complaints from both users and policy makers. Despite this attention, users still have little understanding of what data the advertising platforms have about them and why they are shown particular ads. To address such concerns, Facebook recently introduced two transparency mechanisms: a "Why am I seeing this?" button that provides users with an explanation of why they were shown a particular ad (ad explanations), and an Ad Preferences Page that provides users with a list of attributes Facebook has inferred about them and how (data explanations).**

In this paper, we investigate the level of transparency provided by these two mechanisms. We first define a number of key properties of explanations and then evaluate empirically whether Facebook's explanations satisfy them. For our experiments, we develop a browser extension that collects the ads users receive every time they browse Facebook, their respective explanations, and the attributes listed on the Ad Preferences Page; we then use controlled experiments where we create our own ad campaigns and target the users that installed our extension. Our results show that ad explanations are often *incomplete* and sometimes *misleading* while data explanations are often *incomplete* and *vague*. Taken together, our findings have significant implications for users, policy makers, and regulators as social media advertising services mature.

I. Introduction

In recent years, targeted advertising has become the source of a growing number of privacy concerns for internet users. At the heart of the problem lies the opacity of the targeted advertising mechanisms: users do not understand what data advertising platforms have about them and how this data is being used for ad targeting (i.e., to select the ads that they are shown). This resulting lack of transparency has begun to catch the attention of policy makers and government regulators, who are increasingly introducing laws requiring transparency. For example, the General Data Protection Regulation (GDPR) of the EU establishes a "right to explanations" [9], [26], and the Loi pour une République Numérique of France strengthens the transparency requirements for digital platforms [10].

While many prior studies have focused on bringing transparency to targeted advertising on the web [17], [34], [30], [31], [20], [36], [43], few studies (if any) have focused on social media advertising. Targeting ads on social media differs from traditional ad targeting in multiple important ways: *First*, social media platforms such as Facebook have access to much richer data sources than traditional advertising companies such as DoubleClick (e.g., Facebook has information about the content people are posting, their self-reported demographics, the identities of their friends, web browsing traces, etc). *Second*, social media platforms know detailed personally-identifiable information (PII) of users, and they often allow advertisers to target users based on this information. In comparison, traditional advertisers often only track user browsing behaviors via opaque cookies. As social media sites are now the de-facto portal to the web for many users, bringing transparency to social media advertising is a significant concern.

In response to users' and regulators' concerns, social media platforms recently started offering transparency mechanisms. In particular, Facebook was the first to do so by introducing two features. *First*, Facebook introduced a "Why am I seeing this?" button that provides users with an explanation for why they have been targeted with a particular ad. *Second*, Facebook added an Ad Preferences Page that provides users with an explanation for what information Facebook has inferred about them, how Facebook inferred it, and what information is used for targeting them with advertisements. However, to the best of our knowledge, there has been little examination of these two transparency mechanisms; such a study is all the more important because other social media sites such as Twitter have recently begun introducing similar transparency mechanisms.

In this paper, we take a first step towards exploring the transparency mechanisms provided by social media sites, focusing on the explanations that Facebook provides. However, constructing explanations for social media advertising is a challenging problem as ad impressions are the result of a number of complex processes within Facebook, as well as

Network and Distributed Systems Security (NDSS) Symposium 2018
18-21 February 2018, San Diego, CA, USA
ISBN 1-1891562-49-5
http://dx.doi.org/10.14722/ndss.2018.23191
www.ndss-symposium.org

of interactions between multiple advertisers and Facebook's advertising platform. Here, we narrow our study to the two processes for which Facebook provides transparency mechanisms: the process of how Facebook infers data about users, and the process of how advertisers use this data to target users. We call explanations about those two processes *data explanations* and *ad explanations*, respectively.

Constructing an explanation involves a number of design choices, ranging from the phrasing, to the length, and to the amount of detail provided. As a consequence, what would constitute a *good* explanation is an ill-defined question, as it depends heavily on what the purpose of the explanation is. For instance, explanations can serve to improve the trust placed by users in the site, or simply to satisfy their curiosity in order to enhance the service's utility. Explanations can also be seen as a tool to allow users to control the outcome of the ad targeting system (e.g., the ads they receive), or as a tool for regulators[1] to verify compliance with certain rules (e.g., non-discrimination), or even as a tool for users to detect malicious or deceptive targeting behavior. Different purposes might impose different design choices: for instance, verifying non-discrimination might necessitate an exhaustive list of all targeting attributes used, while such a list may be overwhelming for end users who are simply curious.

We do not attempt to arbitrate on what would be a good explanation. Instead, we identify a number of *properties* that are key for different types of explanations aimed at bringing transparency to social media advertising. We then evaluate empirically how well Facebook's explanations satisfy these properties and discuss the implications of our findings in view of the possible purposes of explanations. Specifically, after providing a detailed account of the different processes involved in Facebook's advertising and the data about users they make available to advertisers (Section II), this paper makes the following contributions:

(*i*) We investigate *ad explanations* (Section III), i.e., explanations of the ad targeting process. We define five key properties of the explanations: *personalization*, *completeness*, *correctness* (and the companion property of *misleadingness*), *consistency*, and *determinism*. To analyze the explanations Facebook provides, we build a Chrome browser extension that collects all the ads users receive, along with the explanations provided for the ads, every time the users browse Facebook. We deploy this extension and collect 26,173 ads and corresponding explanations from 35 users. To study how well Facebook's ad explanations satisfy our five properties, we conduct controlled ad campaigns targeting users who installed the browser extension, and compare the explanation to the actual targeting parameters we defined in the campaign.[2]

Our experiments show that Facebook's ad explanations are often *incomplete* and sometimes *misleading*. We observe that *at most one* (out of the several attributes we targeted users with) is provided in the explanation. The choice of the attribute shown depends deterministically on the type of the attribute (e.g., demographic-, behavior-, or interest-based) and its rarity (i.e., how many Facebook users have a particular attribute). The way Facebook's ad explanations appear to be built—showing only the most prevalent attribute—may allow malicious advertisers to easily obfuscate ad explanations from ad campaigns that are discriminatory or that target privacy-sensitive attributes. Our experiments also show that Facebook's ad explanations sometimes suggest that attributes that were never specified by the advertiser "may" have been selected, which makes these explanations potentially misleading to end users about what the advertiser's targeting parameters were.

(*ii*) We investigate *data explanations* (Section IV), i.e., explanations of the data inferred about a user. We define four key properties of the explanations: *specificity*, *snapshot completeness*, *temporal completeness*, and *correctness*. To evaluate Facebook's explanations, we crawl the Facebook Ad Preferences Page for each user daily using the browser extension, and we conduct controlled ad campaigns that target attributes that are not present in the Ad Preferences Page. Our analysis shows that the data provided on the Ad Preferences Page is *incomplete* and often *vague*. For example, the Ad Preferences Page provides *no information* about data obtained from data brokers, and often does not specify which action a user took that lead to an attribute being inferred. Consequently, users have little insight over how to avoid potentially sensitive attributes from being inferred.

Overall, our study is a first step towards better understanding and improving transparency in social media advertising. While we do not claim that the properties that we have identified form an exhaustive list, we hope that our work will spur further interest from researchers and social media sites to investigate how to improve transparency mechanisms.

II. Advertising on Facebook

Before evaluating the explanations provided by Facebook, we first explore the different processes that are involved when a user is shown an ad, as well as the ad targeting parameters Facebook makes available to advertisers. This information is useful as a reference for evaluating the explanations provided by Facebook and understanding their impact, and for understanding what are the different components we ideally would like to make transparent.

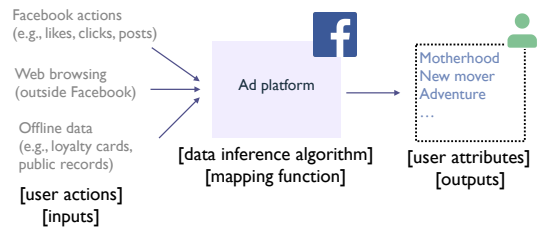
We first separate out the different processes that are responsible for a user receiving an ad, then briefly describe how advertisers can place ads using Facebook's advertising interface, and finally analyze the various targeting methods available to advertisers by studying what data about users is used by each.

A. The processes responsible for receiving an ad

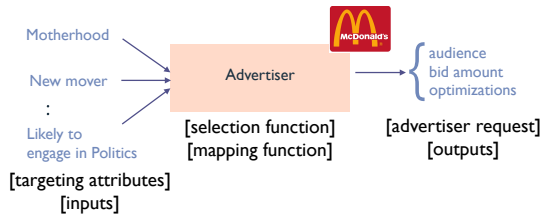
The central goal of our paper is to analyze Facebook's answers to the question *Why am I being shown this ad?* The reason why a user received a particular ad is, however, the result of a complex process that depends on many inputs. To enumerate just a few, it depends on: what the platform thinks the user is interested in, the characteristics of users the advertiser wants to reach, the set of advertisers and the parameters of their campaigns, the bid prices of all advertisers, the active users on the platform at a particular time, and

[1] This is one of the main intended goal of bringing transparency in laws such as the French "loi pour une République Numérique".

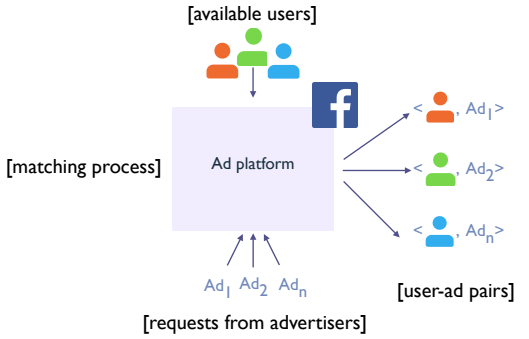
[2] Our study was reviewed and approved by our respective institutions' Institutional Review Boards.



(a) The data inference process



(b) The audience selection process



(c) The user-ad matching process

Fig. 1: The processes responsible for receiving an ad.

the algorithm used to match ads to users. Theoretically, an explanation could provide information about all these complex processes, however, it would be very challenging to do so without overwhelming users.

In this section, we attempt to simplify the task by separating the different processes that are responsible for a user receiving an ad. In social media advertising we can distinguish three responsible components:

(1) *The data inference process* is the process that allows the advertising platform to learn the users' attributes. We can model this process as having three parts (see Figure 1a): (a) the raw user data (the inputs), containing the information the advertising platform collects about a user either online (e.g., pages liked, web browsing activity, uploaded profile information, etc) or offline (e.g., data obtained from data brokers); (b) the data inference algorithm (the mapping function between inputs and outputs), covering the algorithm the advertising platform uses to translate input user data to targeting attributes; and (c) the resulting targeting attributes (the outputs) of each user that advertisers can specify to select different groups of users.

(2) *The audience selection process* is the interface that allows advertisers to express who should receive their ads. Advertisers create *audiences* by specifying the set of targeting attributes the audience needs to satisfy (see Figure 1b; more details in Section II-C). Later, to launch an ad campaign, advertisers also need to specify a bid price and an optimization criterion (e.g., "Reach" or "Conversions", that specify to Facebook what the advertiser's goal is).

(3) *The user-ad matching process* takes place whenever someone is eligible to see an ad [2]. It examines all the ad campaigns placed by different advertisers in a particular time interval, their bids, and runs an auction to determine which ads are selected (see Figure 1c).

An explanation for the data inference process or the audience selection process can provide information about any of the the three components: the *inputs*, the *outputs*, or the *mapping function*. Explanations of the advertising platform matching process are, however, much more complex as the outcome not only depends on the advertising platform and its complex matching algorithm, but also on all the competing advertisers and their corresponding requests as well as all the available users on the platform. In this paper, we focus on explanations of the first two processes, and we refer to them as *data explanations* and *ad explanations* respectively. We leave explanations of the advertising platform matching process for future work. Nevertheless, only explaining the data inference and advertising selection process simplifies the design of explanations while keeping the explanation informative for the user. Note that while data explanations provide information about the decisions of the advertising platform, ad explanations provide information about the decisions of the advertiser. Thus, the set of properties and concerns is different for the two.

B. Placing ads on Facebook

Facebook's advertiser interface allows advertisers to create targeting *audiences*—predefined sets of users that match various criteria (i.e., that have certain *attributes*)—and then place ads that will only be seen by users in a particular audience (see Figure 2). The interface allows advertisers to choose the *location*, *age range*, *gender*, and the *language* of users they wish to target. Additionally, advertisers can browse through a list of predefined *targeting attributes* that can be *demographic*-, *interest*-, or *behavior-based* to further refine their audiences.

In addition to this traditional form of audience selection based on targeting attributes, Facebook introduced a new feature called *custom audiences* in 2012 [18]. In brief, custom audiences allow advertisers to upload a list of PII—including email addresses, or phone numbers, or names along with ZIP codes—of users who they wish to reach on Facebook.[3] Facebook then creates an audience containing only the users who match the uploaded PII.

C. Targeting methods and available data

While there are many ways to target users as described above, we choose to analyze targeting on Facebook through

[3]Other social media sites such as Twitter, Google, Pinterest or LinkedIn also provide similar features.

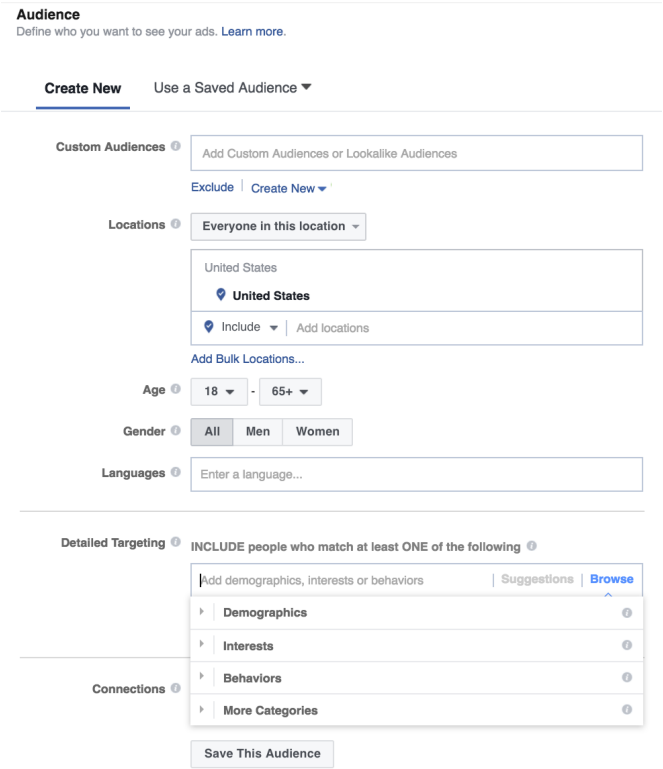


Fig. 2: Facebook's audience creation interface.

the lens of where the targeting data comes from (i.e., the data provenance). Advertisers can target users in three different ways: (1) based on attributes computed by Facebook—we call this approach *traditional Facebook targeting*; (2) based on attributes that are externally sourced from data brokers such as Acxiom and Experian (called partner categories by Facebook)—we call this approach *data broker targeting*; and (3) by directly providing a list of PII corresponding to users —we call this approach *advertiser PII targeting*.

1) Traditional Facebook targeting: This type of targeting is essentially the traditional way to target people, where advertisers can define their audiences by choosing from a predefined list of targeting attributes. This targeting exploits information about users' demographic-, interest-, and behavior-based features that Facebook gathers.

To aggregate information about its users, Facebook has many potential sources of data: information about the activities users perform on Facebook (e.g., the information they provide in their profiles, the pages they like, etc), as well as information Facebook collects about users' activities outside Facebook (e.g., which sites users browse,[4] which Facebook applications they install on their mobile devices, etc).

To more closely examine how advertisers are able to target their ads, we collect the full list of predefined targeting attributes, which is hierarchically organized as a tree with similar attributes grouped under common sub-categories. We

[4]Facebook can use cookies to track visits by users to any webpage that has either a Facebook tracking pixel [24], or a Facebook like button [35], or uses the Facebook login [39] feature.

TABLE I: List of U.S. targeting categories provided by different data sources with the number of attributes in each category. The categories are divided by type: Behavior- (B), Demographic- (D), and Interest-based (I).

Category	FB	Acxiom	Experian	DLX	Epsilon	Other
(B) Anniversary	1	-	-	-	-	-
(B) Consumer Classif.	2	-	-	-	-	-
(B) Digital activities	39	-	-	-	-	-
(B) Expats	74	-	-	-	-	-
(B) Mobile device user	81	-	-	-	-	-
(B) Multicultural affinity	6	-	-	-	-	-
(B) Seasonal and events	2	-	-	-	-	-
(B) Travel	5	-	-	11	-	-
(B) Automotive	-	1	-	151	-	-
(B) Charitable donations	-	5	-	-	4	-
(B) Financial	-	25	-	-	1	-
(B) Job role	-	2	-	1	-	-
(B) Media	-	35	-	-	-	-
(B) Purchase behavior	-	23	3	144	5	-
(B) Residential profiles	-	2	1	-	2	-
(B) B2B	-	-	-	29	-	-
(D) Education	13	-	-	-	-	-
(D) Generation	3	-	-	-	-	-
(D) Home	2	19	1	2	-	-
(D) Life Events	36	-	-	-	-	-
(D) Parents	9	-	-	11	-	-
(D) Politics (US)	8	-	-	-	2	-
(D) Relationship	16	-	-	-	-	-
(D) Work	26	-	-	1	-	-
(D) Financial	-	16	-	-	-	10
(I) Business and industry	39	-	-	-	-	-
(I) Entertainment	70	-	-	-	-	-
(I) Family and relationships	8	-	-	-	-	-
(I) Fitness and wellness	11	-	-	-	-	-
(I) Food and drink	37	-	-	-	-	-
(I) Hobbies and activities	60	-	-	-	-	-
(I) Shopping and fashion	21	-	-	-	-	-
(I) Sports and outdoors	22	-	-	-	-	-
(I) Technology	21	-	-	-	-	-
Other	2	-	-	-	-	-
Total attributes	614	128	5	350	14	10
Audience reach	196M	152M	131M	147M	71M	145M

find that the list varies based on the country of the advertiser's Facebook account.[5] Therefore, we collect the list of targeting attributes across 10 different countries (U.S., U.K., France, Germany, Australia, South Korea, Brazil, Japan, Canada, and India) by creating test accounts in each of these countries. We direct our traffic through proxies in order to create advertising accounts in each of these countries. In total, we collect 1,420 unique targeting attributes across the 10 countries.

In addition, we collect the metadata that Facebook's advertiser interface provides for each predefined attribute: a short *description* of the attribute (e.g., for "Multicultural Affinity" we get the description "People who live in the United States whose activity on Facebook aligns with Hispanic multicultural affinity"); and the *data provenance* of the attribute (i.e., whether the data comes from Facebook or one of its partners such as Acxiom). For each attribute, we create an audience of users with that attribute, and obtain the corresponding *audience reach* estimate (of the number of users in that

[5]Note that the list of predefined targeting attributes varies based on the country where the advertiser creates his Facebook account, and not on the location of users that are targeted.

TABLE II: Sample of targeting attributes offered by Facebook and four data broker partners: Acxiom, DLX, Experian, and Epsilon. Also shown is the category and corresponding audience reach (number of users).

Source	Category	Reach	Targeting attributes
Facebook	(D) Politics (U.S.)	179M	Likely To Engage in Politics (Conservative), Likely To Engage in Politics (Liberal), Likely To Engage in Politics (Moderate), U.S. Politics (Conservative), U.S. Politics (Liberal), U.S. Politics (Moderate), U.S. Politics (Very Conservative), U.S. Politics (Very Liberal)
Facebook	(I) Family and relationships	138M	Dating, Family, Fatherhood, Friendship, Marriage, Motherhood, Parenting, Weddings
Facebook	(B) Consumer classification/India	3100	(A) Affinity for High Value Goods/India, (A+B) Affinity for Mid-High Value Goods/India
Facebook	(D) Parents/All Parents	59M	(0-12 months) New Parents, (01-02 Years) Parents with Toddlers, (03-05 Years) Parents with Preschoolers, (06-08 Years) Parents with Early School Age Children, (08-12 Years) Parents with Preteens, (13-18 Years) Parents with Teenagers, (18-26 Years) Parents with Adult Children, Expectant parents, Parents (All)
Acxiom	(B) Charitable donations	75M	Animal welfare, Arts and cultural, Environmental and wildlife, Health, Political
Acxiom	(B) Financial/Spending methods	140M	1 Line of Credit, 2 Lines of Credit, 3, Active credit card user, Any card type, Bank cards, Gas, department and retail store cards, High-end department store cards, Premium credit cards, Primarily cash, Primarily credit cards, Travel and entertainment cards
Acxiom	(B) Purchase behavior/Store types	34M	High-end retail, Low-end department store
Acxiom	(B) Residential profiles	5M	Recent homebuyer, Recent mortgage borrower
Acxiom	(D) Financial/Net Worth/Liquid assets	74M	$1-$24,999, $25,000-$49,999, $50,000-$99,999, $500K-$1M, $100K-$249K, $250K-$499K, $1M-$2M, $2M-$3M, $3M+ ,
DLX	(B) Automotive/New vehicle buyers (Near market)/Style	102M	Crossover, Economy/compact, Full-size SUV, Full-size sedan, Hybrid/alternative fuel, Luxury SUV, Luxury sedan, Midsize car, Minivan, Pickup truck, Small/midsize SUV, Sports car/convertible
DLX	(B) Purchase behavior/Health and beauty	90M	Allergy relief, Antiperspirants and deodorants, Cosmetics, Cough and cold relief, Fragrance, Hair care, Health and wellness buyers, Men's grooming, Oral care, Over-the-counter medication, Pain relief, Skin care, Sun care, Vitamins
DLX	(B) Automotive/Owners/Vehicle age	95M	0/1 year old, 2 years old, 3 years old, 4/5 years old, 6/10 years old, 11/15 years old, 16/20 years old, Over 20 years old
Experian	(D) Home/Home Ownership	26M	First time homebuyer
Experian	(B) Residential profiles	5M	New mover
Epsilon	(B) Residential profiles	3M	Likely to move
Epsilon	(B) Charitable donations	34M	All charitable donations, Cancer Causes, Children's Causes, Veterans

audience) provided by Facebook (Facebook calls this estimate the "potential reach" [1]).

Table I summarizes these results, with the first column showing the categories present for each type of attribute (behavior-, demographic-, or interest-based), and the second column showing the corresponding number of targeting attributes under each category. While some of these categories such as "Hobbies and activities" may seem quite benign, others such as "Family and relationships" may raise privacy issues in the context of advertising. To help better understand how fine-grained the targeting attributes can be, we present a sample of these in the first group of rows in Table II; the second column of the table contains the parent categories from Table I while the fourth column contains the targeting attributes that fall under that category. For each category, we create an audience of users that have *at least one* of the targeting attributes that fall under that category and obtain the corresponding *audience reach* estimates; these are presented in the third column of Table II. From the table, we observe that Facebook allows advertisers to target people that are "new parents", have an "affinity for high value goods", are "likely to engage in politics (conservative)", are in an "open relationship", etc.

In addition to the list of predefined targeting attributes described above, Facebook also computes other targeting attributes that advertisers can search for by inputting free text, and use to target users. These attributes are predominantly interest-based attributes which correspond to "People who have expressed an interest in or like pages" related to those particular attributes, according to the description found in the advertiser interface. We did not attempt to collect such attributes as there are likely a large number of them, given that there are millions of such pages [19].

2) Data broker targeting: This type of targeting is similar to the traditional-Facebook targeting described above, except for the fact that the targeting attributes are sourced from data brokers (called Facebook Marketing Partners) instead of being mined by Facebook; this data is obtained by Facebook by linking their user data with data from data brokers.

The provenance information present in the metadata of each attribute allowed us to observe that some of the predefined attributes Facebook provides come from various data brokers. In the U.S., Facebook currently works with four data brokers: Epsilon, DLX, Experian, and Acxiom. Table I presents the number of targeting attributes that come from different data brokers in the U.S. We observe from the penultimate row that a large fraction (45%) of targeting attributes come from these data brokers. These targeting attributes capture information such as financial information (e.g., income level, net worth, purchase behaviors, charity, and use of credit cards) that is presumably more difficult for Facebook to determine from its data alone. Each of the last four groups of rows in Table II presents a sample of attributes sourced from a particular data broker; many of the attributes sourced from data brokers may also raise privacy concerns among users.

While Facebook relies mostly on online data, data brokers aggregate information about people both from online sources [23] as well as offline sources such as voter records, criminal records, data from surveys and other data providers such as automotive companies, grocery, drug stores or supermarkets [12], [40], [3], [11].

To study how many Facebook users data brokers have data about, for each data broker (in the U.S.), we create an audience of users who are located in the U.S. and who have *at least one* of the attributes provided by that data broker (in the U.S.); we then obtain the corresponding *audience reach* estimates provided by Facebook's advertiser interface. The last row of Table I presents the audience reach estimates. We were surprised to see that almost all the data brokers have data about

the majority of Facebook users (i.e., their audience reach is generally more than 100M while the audience reach using all attributes provided by Facebook is 196M).

3) Advertiser PII targeting: Besides the traditional forms of targeting through attribute selection, advertisers can directly upload their own list of users they want to reach on Facebook using the custom audience feature. Using this mechanism, Facebook allows advertisers that have collected information about their customer's names and addresses (information typically asked when creating fidelity cards), phone numbers, or email addresses to target them with ads on Facebook. Using this mechanism, advertisers can simply upload a list of phone numbers and target people in the list. Likewise, advertisers can target people that visited their website, installed their mobile application, or interacted with content on their Facebook page.

To implement these features, the Facebook platform effectively links advertiser-provided PII with users on Facebook.[6] Note that Facebook does not reveal the corresponding Facebook accounts to advertisers, it only gives an estimate on the number of people in the custom audience that have an account on Facebook.

D. Summary

Facebook has aggregated a large number of attributes about its users, as seen from the audience reach numbers, both from the activities of users in Facebook, and from data brokers. Through its advertiser interface, Facebook allows advertisers to use very fine-grained and potentially sensitive attributes to target users with ads. Thus, it is important that explanations provide a clear view of how users are targeted and what data Facebook has about them.

III. Audience selection explanations

We begin by examining explanations that concern the audience selection process (see Section II-A). In other words, *what actions did the advertiser take that led to a user being shown an ad?* We call these answers *ad explanations.* This question can be answered in multiple ways and with various degrees of information. For example, an explanation such as, "*you are being shown this ad because the advertiser targets people with accounts on Facebook*" might be a potential explanation, although not a particularly useful one. Therefore, it is critical to analyze such explanations, as their design choices have significant implications on how well users understand how their data is being used by the advertising platform. We first discuss possible properties of ad explanations in general, and then investigate the explanations provided by Facebook and their properties.

A. What is an ad explanation?

As mentioned in Section II-A, ad explanations could provide information about the inputs (the users' information, actions, etc), the outputs (the inferred targeting attributes), or the mapping function between them. The explanations could also provide information about the advertising campaign, such as bid amount or the optimization criteria chosen.

Facebook recently introduced a feature where users can click on a button labeled "Why am I seeing this?" next to each ad they are shown. Facebook then provides explanations to the user such as

> One reason you're seeing this ad is that [advertiser] wants to reach people interested in Facebook, based on activity such as liking pages or clicking on ads. There may be other reasons you're seeing this ad, including that [advertiser] wants to reach people ages [age] and older who live in [location]. This is information based on your Facebook profile and where you've connected to the internet.

Thus, the ad explanations that Facebook provides give some information about the targeting attributes used by the advertiser.

The ad explanation above can be separated into two parts. In the first part—before "There may be other reasons you're seeing this ad"—Facebook provides attributes asserting that they have been used by the advertiser for the audience selection. We simply call these *attributes*. In the second part, Facebook provides additional attributes with the caveat that they *may* have been used by advertiser—we call these *potential attributes*. Most explanations that we observed (76%) can be separated in this way (i.e., include both attributes and potential attributes), while the remainder do not include the second part (i.e., they have no potential attributes).[7]

B. Properties of ad explanations

We now examine the properties that ad explanations could have. Let us suppose that an advertiser targeted users by creating an audience with the following attributes:

$$A = (a_1 \text{ AND } a_2) \text{ OR } a_3 \text{ OR } \neg a_4$$

and that we have four users with the following attributes $U_1 = \{a_1, a_2, a_{991}, a_{992}\}$, $U_2 = \{a_3, a_{993}, a_{994}\}$, $U_3 = \{\neg a_4, a_{995}\}$, $U_4 = \{a_1, a_2, a_{996}\}$. There are a number of properties that the platform's ad explanations could satisfy:

a) Correctness: We say that an explanation is *correct* if every attribute and potential attribute listed has been used by the advertiser. In our example, only a_1, a_2, a_3, or $\neg a_4$ should appear in the explanation if it is to be correct. However, because of potential attributes, not all explanations that do not meet this definition are incorrect. Specifically, we say that an explanation is *incorrect* if there exists an attribute listed that was actually not used by the advertiser. We say that an explanation is *misleading* if all of its attributes listed were used by the advertiser, but there exists a potential attribute listed that was not used by the advertiser. Thus, we note that a misleading explanation is neither correct nor incorrect.

In our example, an explanation with attributes a_1 and a_2 and potential attribute a_{997} is misleading, as a_{997} was

[6] Investigating the accuracy of such matching is important—but beyond the scope of this paper—as previous work showed that matching at large scale is often inaccurate [25].

[7] While placing our own ads, we found that the explanations *without* the second part only occurred when we selected *no* targeting attributes beyond age, gender, and location.

not specified by the advertiser. However, if the explanation included a_{997} as an attribute (rather than a potential attribute), we would then call the explanation incorrect. Fortunately, for the remaining properties, we do not need to make the distinction between attributes and potential attributes; the attributes mentioned next can be of either type.

b) Personalization: Ad explanations can either be *non-personalized* (i.e., the explanation is the same for all users that received the ad) or *personalized* (i.e., the explanation differs from user to user). Using our example above, one non-personalized ad explanation would be to report all of the attributes specified by the advertiser. In contrast, personalized ad explanations might only show the attributes that are specified by the advertiser that also match the user. For example, U_1's explanation might be $\{a_1, a_2\}$, U_2's might be $\{a_3\}$, etc. Personalized ad explanations may be more useful for users who want to only know why *they* were shown the ad, but non-personalized explanations might be more useful for users who want to know more about the set of all users who the advertiser was targeting.

c) Completeness: A *complete* ad explanation should list all the attributes $a_1, a_2, a_3, \neg a_4$ for non-personalized ad explanations, while for personalized ad explanations, it should list the entire subset of $a_1, a_2, a_3, \neg a_4$ attributes for which Facebook has information about the user.

A *succinct* (incomplete, yet useful) ad explanation would limit the number of listed attributes to the most important ones, for some useful notion of "importance." We will see later in the section that Facebook currently shows only one attribute in each ad explanation, regardless of the number of attributes used by the advertiser. Succinct ad explanations might be preferred over complete ad explanations if users are overwhelmed by a large number of attributes that appear in the explanation. However, constructing succinct ad explanations requires ranking the importance of attributes. Among other criteria, such a ranking could be based on:

(1) an attribute's *rarity* in the entire Facebook user population (i.e., based on the fraction of Facebook users that have that attribute); intuitively, if 90% of users on Facebook have a_1 and only 1% have a_2, including attribute a_2 in the ad explanation would be more informative than including a_1.

(2) an attribute's perceived *sensitivity*; having a particular political leaning may be a more prevalent feature than playing tennis, but the former might be more privacy sensitive than the latter. Moreover, the perceived sensitivity of an attribute varies from user to user, so a personalized explanation may be able to capture different users' rankings.

d) Consistency: In the case of personalized ad explanations, the platform could ensure *consistent* explanations across users who match the same subset of attributes. In our example above, the ad explanations given to users U_1 and U_4 would need to be the same if the platform provided consistent ad explanations.

e) Determinism: Finally, *deterministic* ad explanations would give the same ad explanation to a user for all ads that were placed with the same targeting attributes. On the contrary, non-deterministic ad explanations may cycle through multiple explanations at different times. Note that non-deterministic ad explanations might be necessary if ad explanations are personalized and the input data Facebook has about a user changes over time.

In the rest of the section, we analyze Facebook's ad explanations based on the properties defined above.

C. Measurement methodology

To study the ad explanations that Facebook provides, we wrote a browser extension that gathers ad explanations for all the ads received by users on Facebook. To check the properties of ad explanations, we conduct controlled ad campaigns that target volunteers who installed the browser extension, gather the ad explanations provided, and check which attributes are represented in the ad explanations.

1) Browser extension to collect ad explanations: We develop a browser extension for Chrome that records the ads the users receive whenever they browse Facebook, as well as the respective explanations that Facebook provides. Once an ad appears, it is captured by the extension and forwarded to a server we control. We detect the ads based on specific unique characteristics, such as the "Sponsored" tag, that make them different from other posts. Ads can either appear as posts in the user's feed—called *front ads*—or can appear on the right of the screen—called *side ads*.

We also capture the ad explanation URL that is linked to by a "Why am I seeing this?" button on each ad. Facebook imposes a rate limit for the requests to these URLs. Specifically, usually after 10 requests/hour, the service stops delivering explanations for some time. Thus, we send the explanation URL requests to a scheduler that does not make more than 10 requests/hour. Moreover, to avoid unnecessary requests (while allowing us to study consistency), once we collect an explanation for a particular ad for a given user, we do not collect the explanation for the same ad if shown again to the same user for a period of two days. The process does not interfere with the browsing experience of the user. Moreover, the number of requests we make to Facebook is trivial when compared to the number of requests that take place when a user browses Facebook.

We collect ads and explanations from 35 users for a total of 5 months (accumulated across all the users). We recruit users by advertising our browser extension on a personal basis to our co-workers and families. In total, we collect 26,173 unique ads and their corresponding ad explanations; we refer to this dataset as the AD-DATASET.

2) Design of controlled experiments: To test the properties of ad explanations, we launch ad campaigns where we control the targeting attributes and collect the explanations Facebook provides. Our goal is to investigate how the targeting attributes that we select are represented in the explanations users receive.

The primary challenge in designing these controlled experiments is to collect the explanations corresponding to *our* ad campaigns. Therefore we launch ad campaigns that try to target the people that installed our browser extension. Since the number of users that installed our browser extension (called *monitored users*) is limited, we employ several strategies to

increase the likelihood that the monitored users receive the ads so that we can collect the ad explanations:

Selection of targeting attributes: For the monitored users we gather the targeting attributes that appear in their Facebook Ad Preferences Page [5]. Depending on the type of the experiment, we either use the most common attributes across our monitored users to target ads, or unique attributes that can single out a user.

High bid: To ensure that our ads would be delivered effectively, we placed bids that were higher than the value suggested by Facebook. For most of the experiments, our bid was 25€ per 1,000 impressions, while the suggested bid by Facebook was typically 7–8€ per 1,000 impressions.

Campaign objective: We created campaigns that optimized for "Reach." According to Facebook, this particular campaign objective, when selected, shows the ads to the maximum number of people (rather than showing the ad to people that are the most likely to click on the ad).

Location: Since most of the users using our browser extension live in the same city (of about 150K inhabitants), we targeted this city in our ad campaigns to narrow the audience and have a higher chance to collect the ad explanation.

Custom list: In some of our experiments, to narrow our audience even more, we used three custom lists: one comprising of 900 public U.S. voter records, one comprising of 9,350 public U.S. voter records from North Carolina [8], and one comprising of 10,000 public French mobile phone numbers. To each of these lists, we also added our monitored users. We used each custom list for the appropriate experiments in order to maximize the probability that the ads would reach the monitored users; we observed that if the audience reach is less than 20, the campaign often fails. Thus, we always tried to achieve an audience reach that was larger than 20 for every possible combination of targeting attributes that we attempted.

Finally, to ensure that we can identify explanations corresponding to different ad campaigns, each ad had unique text, which in combination with the advertiser identity, made them uniquely identifiable. Our ads were generic with neutral content. They made use of stock photos provided by Facebook, and the accompanying text was suggesting users to spend their vacation in Saarbrücken, Germany, or Nice, France (e.g., "This spring, the number one destination is Saarbrücken!"). We did not include any links or track conversions for any ad.

In total, we performed 135 different ad campaigns. Out of the 135 experiments, 96 reached at least one monitored user and 65 reached more than one user. In total, we gathered 254 ad explanations for our own ads from 14 unique monitored users that were targeted for these experiments. In the remainder of the section, whenever we refer to controlled experiments, we only consider the 96 successful experiments.

3) Impact of the small/biased dataset: The goal of our controlled experiments is to test whether Facebook explanations satisfy the properties we defined, such as completeness or correctness. The key to design such experiments is to be able to both target an account and collect the respective explanation. The number of users we monitor only affects the probability that we can observe the corresponding explanation. Even with a small number of users, we were able to observe the corresponding explanations of most of our ad campaigns.

While our users are not representative of the Facebook population as a whole, they are spread across 3 countries in Europe as well as the U.S. While proving that explanations *always* satisfy certain properties is likely impossible even with a much larger user base, proving that explanations fail to satisfy certain properties only requires one example.

4) Ethics: All experiments and data collection presented in this paper were reviewed by the Ethical Review Board of the University of Saarland and approved; they were also reviewed and approved by the Institutional Review Board of Northeastern University. We limited our data collection to just what was necessary to measure the ad explanations and did not record other user behavior (e.g., the browser extension was only active when the users were browsing Facebook, and only uploaded information about the ads they were shown). Moreover, our extension did not fetch any additional ads that the user would not have otherwise been shown or click on any ads; thus, we did not affect advertisers in any way.

Our data collection is compliant with Facebook's Terms of Service (https://www.facebook.com/legal/terms). Under Protecting People's rights (5th section, 7th point) "If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it." We did all of the above.

D. Evaluation of Facebook's ad explanations

Using the data described above, we now study the properties of the explanations provided by Facebook.

1) Overview: Recall that Facebook's ad explanations typically have two parts: the first part starts with "One reason you're seeing this ad ..." or "You're seeing this ad because ...", and the second part starts with "There may be other reasons you're seeing this ad ...".

The first part of the ad explanations varies greatly across all of the ad explanations we observed. If we focus only on the first part of the ad explanations for the ad explanations that have both parts, we can group (the first part of) explanations based on their underlying pattern and attribute type. Table III shows the different explanation types we identified together with typical examples for each type; overall, we observed 10 different structures for the first part of the explanations.

In contrast, the second part of the explanations always contains age, location, and gender information, and has the format:

> There may be other reasons why you're seeing this ad, including that [advertiser] wants to reach [gender] aged [age range] who live or have recently been in [location]. This is information based on your Facebook profile and where you've connected to the Internet.

TABLE III: Examples of the first part of ad explanations provided by Facebook (we underlined the sources of data Facebook mentions as well as emphasizing the variable text that changes from explanation to explanation depending on the ad).

Explanation type	Example of explanations	Count
LANGUAGE	One reason why you're seeing this ad is that BOREDOM THERAPY wants to reach people who SPEAK "ENGLISH (US)". This is based on information from sources such as your Facebook profile.	404
DEMOGRAPHICS	One of the reasons why you're seeing this ad is because we think that you may be in the "MILLENNIALS" audience. This is based on what you do on Facebook.	149
BEHAVIORS	One of the reasons why you're seeing this ad is because we think that you may be in the "GMAIL USERS" audience. This is based on what you do on Facebook.	239
INTERESTS	One reason why you're seeing this ad is that ACER wants to reach people interested in ELECTRONIC MUSIC, based on activity such as liking pages or clicking on ads.	4,621
DATA BROKERS	One reason you're seeing this ad is that CANAL FRANCE wants to reach people who are part of an audience created based on data provided by ACXIOM. Facebook works with data providers to help businesses find the right audiences for their ads.	78
PII-BASED TARGETING	One reason you're seeing this ad is that AAAS - THE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE wants to reach people who have visited their website or used one of their apps. This is based on customer information provided by AAAS - THE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE. One reason you're seeing this ad is that ACTIMEL added you to a list of people they want to reach on Facebook. They were able to reach you because you're on their customer list or you've provided them with your contact information off of Facebook. One reason you're seeing this ad is that ABOUT YOU added you to an audience of people they want to reach on Facebook. This is based on activity such as watching their Facebook videos, sharing links to their website on Facebook and interacting with their Facebook content. One reason you're seeing this ad is that SHAUN T wants to reach people who like their page.	696
PROFILE DATA	One reason you're seeing this ad is that AEGEAN AIRLINES wants to reach people with RELATIONSHIP STATUS "ENGAGED" on their Facebook profiles. One reason why you're seeing this ad is that EY CAREERS wants to reach people with THE SCHOOL/UNIVERSITY UNIVERSITÄT DES SAARLANDES - SAARLAND UNIVERSITY listed on their Facebook profiles. One reason you're seeing this ad is that ATENAO - TRANSLATION agency wants to reach people with THE EDUCATION LEVEL "DOCTORATE DEGREE" listed on their Facebook profiles.	144
LOOKALIKE AUDIENCE	One reason why you're seeing this ad is that AUTODESK STUDENTS wants to reach people who may be similar to their customers.	1,314
MOBILE DATA	One reason why you're seeing this ad is that CDU SAARBRÜCKEN-SCHEIDT wants to reach people WHO WERE RECENTLY NEAR THEIR BUSINESS. This is based on information from your Facebook profile and your mobile device.	142
SOCIAL NEIGHBORHOOD	One reason why you're seeing this ad is that CARTIER wants to reach people whose friends like their Page.	188

Note that the value of the gender field can be either "men", "women", or "people", as Facebook allows advertisers to target "All" genders as shown in Figure 2.

Looking closely at the examples in Table III, we can see that the ad explanations often provide information about who the advertiser is, what targeting attributes they used, and what the underlying source for these targeting attributes is. The underlying data sources mentioned are very diverse, including "your Facebook profile", "where you've connected to the internet", "liking pages", "clicking on ads", and "what you do on Facebook", among others.

We now turn to examine whether the explanations match the properties described in Section III-B.

2) Traditional Facebook targeting: We first examine ads placed using only targeting attributes that are provided by Facebook. After examining these explanations, we then look at explanations for data broker targeting and finally advertiser PII targeting.

a) Personalization: In the AD-DATASET, there exist 10,936 unique ads that provide different explanations for at least two users. This suggests that explanations are personalized. In order to verify this, we performed controlled experiments where we created a targeting audience $A = (a_1$ OR $a_2)$ where a_1 and a_2 were interest-based attributes.[8] We picked the interests so that there are two users that installed our browser extension, where one had a_1 but not a_2 and one had a_2 but not a_1. We performed two such ad campaigns. In all campaigns the ad reached both users, and the ad explanation for each user was different, showing in each case only the interest attribute that each user had. Thus, ad explanations on Facebook are personalized.

b) Completeness: In all ad explanations collected in the AD-DATASET, there is *at most one* attribute that appears in the (first part of the) ad explanation. This raises questions about the completeness of the ad explanations given the fact that the Facebook advertiser interface allows advertisers to use multiple attributes, and it is unlikely that *all* advertisers in our dataset only used one targeting attribute.

To verify that only one attribute is shown even if multiple attributes are specified by the advertiser, we conducted 28 controlled experiments that target three attributes $A = (a_1$ AND a_2 AND $a_3)$ and 51 that target two attributes $A = (a_1$ AND $a_2)$. We varied the precise attributes targeted in each ad campaign. In all explanations provided by Facebook across all monitored users, only one attribute was ever shown, while all users had all attributes. Thus, we observe that Facebook's ad explanations are incomplete.

This incompleteness of explanations raises several questions regarding whether there is a strategy behind *which* attribute appears in the explanation. Due to practical limitations on the number of monitored users and controlled experiments we could perform, we cannot provide definite answers as to which attribute is selected; however, we test the impact of several parameters on the explanations:

[8]For clarity, we omit from A the location or custom list, however, all our experiments in this section use these targeting options to narrow the audience, see Section III-C2.

(1) *Does the order of selected attributes affect the shown attribute?* We performed four experiments with two pairs of interest-based attributes where, for each pair, we tried both orderings of attributes A_1 = (a_1 AND a_2) and A_2 = (a_2 AND a_1). The order did not affect the ad explanation shown.

(2) *Does the rarity of the attributes affect the shown attribute?* We conducted 23 controlled experiments where A_i = (a_1 AND a_2) and where both a_1 and a_2 are of the same type (behavior-, demographic- or interest-based), and where a_1 was more common than a_2. In all 52 ad explanations we collected from all users, the attribute that was the most common always appeared in the ad explanation. For example, for targeting *"Video games (915M users)* AND *Time (823M)"* and *"Video games (915M)* AND *Photography (659M)"*, "Video Games" would be chosen. This result suggests (but does not conclusively prove) that Facebook chooses the *most common attribute* to include in the ad explanation. If this is in fact the case, this choice opens the door for malicious advertisers to obfuscate their true targeting attributes by always including a very popular attribute (e.g., "Facebook access (mobile): all mobile devices (2B)") in their targeting attributes.

(3) *Does the type of the attributes affect the shown attribute?* While our experiments suggest that for attributes of the same type (behavior-, demographic- or interest-based), rarity is the factor that decides which attribute will be shown in the explanation, this does not apply when the attributes are of *different* types. We performed 37 controlled experiments A_i = (a_1 AND a_2) where a_1 and a_2 are of different types (e.g., a_1 is demographic- and a_2 is behavior-based) as well as 24 experiments A_i = (a_1 AND a_2 AND a_3), where a_1, a_2, a_3 are of at least two different types. We tested demographic-, behavior-, interest-, and PII-based targeting attributes. Table IV shows all the pairs of attributes that were used in our experiments, the type of the attribute that appears in the ad explanation, and the number of experiments for each pair.

As we can observe in the table, the order appears to be deterministic. We observe that: DEMOGRAPHIC > INTEREST > PII-BASED > BEHAVIOR. That is, our results suggest that whenever the advertiser uses one demographic-based attribute in addition to other attributes in its targeting, the demographic-based attribute will be the one in the explanation. If this is in fact the case, this choice is potentially impactful to users as previous research shows that users often consider behavior attributes more sensitive than the demographic ones [37].

(4) *Do logical operators affect the shown attribute?* Despite the fact that advertisers can include negation when selecting attributes, we observe no ad explanation in the AD-DATASET that contains a negation. To validate that negated attributes do not appear in ad explanations, we conducted three controlled experiments using the NOT operator with interest-, behavior- and demographic-based attributes. In none of the experiments did we see the respective attribute in the explanation. Instead, the explanations included a custom list explanation, which was our non-negated attribute in the experiments.

c) Consistency: In our controlled experiments, for the 65 ads that reached more than one of the monitored users, the explanations were the same for 61 users. The rest of four correspond to explanations that are personalized (i.e., the users that received the ad do not have the same attributes). Thus, we have no evidence that Facebook ad explanations are not consistent.

TABLE IV: Dominance of attribute types.

Attribute types selected	Shown in explanation	Experiments
Demographic AND Behavior	Demographic	3
Demographic AND Behavior AND PII-Based	Demographic	4
Demographic AND PII-Based	Demographic	1
Demographic AND Demographic AND PII-Based	Demographic	3
Interest AND Demographic	Demographic	3
Interest AND Demographic AND PII-Based	Demographic	2
Interest AND Behavior	Interest	3
Interest AND Behavior AND PII-Based	Interest	2
Interest AND PII-Based	Interest	26
Interest AND Interest AND PII-Based	Interest	10
Behavior AND Behavior AND PII-Based	PII-Based	3
Behavior AND PII-Based	PII-Based	1

d) Correctness: We observed that in some of our controlled experiments the ad explanations provided by Facebook contain, in the second part of the explanations, potential attributes that we never specified in our targeting, namely location-related attributes.

To explore this, we performed 65 controlled experiments where we did not specify any location and the audiences were created using custom lists: A_i = (*Custom List* AND a_i), or A_i = (*Custom List* AND a_i AND a_j), where a_i, a_j are various attributes. Despite the fact that we selected *no* location, all of the corresponding ad explanations contained the following text in the second part:

> There may be other reasons why you're seeing this ad, including that [advertiser] wants to reach people ages 18 and older who live [in/near] [location].

where [location] included "Germany", "Saarbrücken, Saarland", "Paris, Île-de-France", "Nice, Provence-Alpes-Côte d'Azur", "Ayía Paraskeví, Attiki, Attica (region)", depending on the user. This shows that Facebook adds potential attributes to ad explanations that advertisers never specified in their targeting, which makes them misleading. In all of our experiments, the location listed in the ad explanation corresponded to the current location of the user receiving the ad. Our intuition is that when the location is not specified by the advertiser, Facebook is automatically adding the current location of the user receiving the ad as a potential attribute to the ad explanation (and not the location of the advertiser). We do not believe that Facebook is intentionally constructing misleading ad explanations, but our finding underscores the importance of ensuring that ad explanations accurately capture the reasons why a user was targeted.

e) Determinism: In the AD-DATASET, we observed that 12,144 ads were seen multiple times by the same user. Of these, we found that 3% of the ads had at least two different explanations given to the same user. For 55% of these cases the change is in the second part of the explanation, and corresponds to the explanation having different targeting

locations in each ad (potentially because the user was in a different places when he received the ad). Thus, Facebook's ad explanations do not appear to always be deterministic.

3) Data-broker targeting: In the AD-DATASET, we collected 78 ad explanations that mentioned data brokers. In these cases, the actual targeted attribute is not given; instead, the user is told they were part of an audience based on data provided by a specific data broker (see Table III). This is in contrast with the fine-grained attributes that *advertisers* can choose from in the Facebook advertiser interface (e.g., income level, see Table II). To verify this, we conducted three controlled experiments where $A = (a_i)$, with a_i being an attribute provided by Acxiom. As before, we observed that the explanation did not mention the actual attribute, but instead simply said it was "based on data provided by Acxiom." This indicates that when advertisers use data-broker-provided targeting attributes, Facebook provides incomplete explanations to users.

4) Advertiser-PII targeting: Finally, we examine how Facebook's explanations change when advertisers use PII-based targeting (e.g., uploading the user's PII to add them to an audience, using a custom list). Across all explanations we found when using PII-based targeting, Facebook provides explanations like "you're on their customer list" or "you've provided them with your contact information off of Facebook." Unfortunately, Facebook does not reveal to the user *which* PII the advertiser provided (e.g., their email address, phone number, etc). Yet again, we find that the explanations provided by Facebook are incomplete; this issue is especially acute when the advertisers are targeting users directly with their PII.

E. Summary

Across all of our experiments, we consistently found that Facebook's explanations are *incomplete* and sometime *misleading*, often omitting key details that would allow users to understand and potentially control the way they are targeted. Many times, the ways in which the explanations are incomplete make it difficult for users to understand whether sensitive information was used: by appearing to pick the most common attribute to show, by not providing the actual attribute when advertisers use data-broker-provided attributes, and by not revealing the PII that advertisers provided when using PII-based targeting.

IV. DATA INFERENCE EXPLANATIONS

We now turn to examine the data inference process, and Facebook's explanations that attempt to answer the question *what data about me is Facebook inferring and making available to advertisers to target me with ads?* We call these answers *data explanations*. Similar to the previous section, we first discuss key properties of data explanations and then test whether the explanations provided by Facebook satisfy these properties.

A. What is a data explanation?

As mentioned in Section II-A, data explanations can provide information about the inputs, the outputs, or the mapping function of the data inference process. For example, an explanation for *outputs* could simply list all the attributes the advertising platform has inferred about the user or it could provide additional information such as the platform's confidence that the user actually has the given attribute, or whether the attribute has an expiration date. An explanation for the *mapping function* could simply say "We inferred that you like Pizza from your activity on Facebook" or could give a more fine grained answer such as "We inferred that you like Pizza because you checked in to Joe's Pizza on 27 June 2017". An explanation for the mapping function could additionally say *how* it is inferring an attribute such as "We use DBpedia to infer attributes from your Facebook likes", or even specify *when* the platform usually updates the profile of a user.

The amount of information that can be presented in an explanation is therefore large. However, the advertising platform might not wish for their "formula" to be revealed to the users, as it might be considered intellectual property by the platform.

Facebook provides an Ad Preferences Page [5] that shows users the advertising attributes it has inferred about them (i.e., the outputs). Facebook also gives explanations about the actions that led to the inference of a particular attribute (i.e., Facebook provides information about the mapping function of the data inference system), see Figure 3. We next discuss what are some key properties for such explanations.

B. Properties of data explanations

Let us suppose that a user U performed a set of actions i_n on Facebook (i.e., the inputs), and that Facebook inferred a set of attributes o_n about the user from these activities (i.e., the outputs). And let us suppose the mapping function for inputs to outputs had the rule

$$(i_1 \text{ AND } i_2) \text{ OR } i_3 \quad \Longrightarrow \quad o_1,\, o_2,\, o_3$$

We next describe the types of data explanations a platform could provide.

a) Specificity: A data explanation is *precise* if it shows the *precise* activities that were used to infer an attribute about a user. A precise explanation for o_1 might be "we inferred o_1 because you took the actions i_1 and i_2", while a vague explanation might be "we inferred o_1 because of what you do on Facebook." We say that an explanation is precise enough when it is *reproducible*. Precise explanations are preferable over vague explanations as they provide actionable information that users can use to control what the advertising platform is inferring about them.

b) Snapshot completeness: A data explanation is snapshot complete if the explanation shows *all the inferred attributes* about the user that Facebook makes available. A complete data explanation for a user who took action i_3 would be $\{o_1, o_2, o_3\}$, while an incomplete data explanation would be $\{o_1\}$.

The number of attributes the advertising platform has inferred about a user can sometimes be large. Thus, it might be desirable to list the attributes by their importance, for some measure of importance (e.g., how rare/uniquely identifying is the attribute, how many ads received by the user were shown because of the particular attribute, etc). We leave a more in-depth exploration of the best design choices to future work.

c) Temporal completeness: In our experimental results, we observe that the attributes inferred about users change quite often. Hence, for a system that is highly dynamic, snapshot completeness is not enough and it is important for the explanation to be temporally complete and show *all* the attributes inferred about a user over a period of time. Moreover, it may be equally important to learn that the platform *removed* an attribute as it is to learn that it inferred it in the first place. Thus, a temporally complete explanation is one where the platform shows all inferred attributes over a specified period of time.

d) Correctness: A correct explanation is one that only shows the activities that actually lead to the inference of the attributes. Correct explanations for o_1 would include $\{i_1 \text{ AND } i_2\}$, or $\{i_3\}$. An incorrect explanation would be $\{i_4 \text{ AND } i_2\}$. It is important, when analyzing the properties of a data explanation, not to confuse the properties of the explanations with the properties of the inference algorithm. For example, an explanation might be correct, even if the attributes inferred are incorrect (i.e., the user is not interested in a particular attribute).

Note that, while specificity and correctness are properties of explanations of the mapping function, snapshot and temporal completeness are properties of explanations of the outputs.

C. Measurement methodology

To study what data explanations Facebook provides, we crawl the information on the Ad Preferences Page daily over a 5 month period for the 35 monitored users.

The Ad Preferences Page provides insights on three aspects: *interests*: the interests Facebook inferred about the user from his activity on Facebook such as the pages he liked; *advertisers*: the advertisers connected to the user (advertisers whose ads the user clicked on, advertisers whose webpages he visited, and advertisers who have the user's contact information); and *categories*: the demographic and behavioral information Facebook has collected or has inferred about the user based on data inside or outside of Facebook (see Figure 3). To analyze this information, our Chrome extension collects all the attributes present on the Ad Preferences Page on a daily basis. For interests alone, Facebook provides explanations of why they inferred the particular interest; we collect these explanations as well.

D. Evaluation of Facebook's data explanations

We now examine the data we collected from our 35 users to better understand the properties of Facebook's data explanations.

1) Overview: We first examine the number of attributes that Facebook reports to each user. We find that the number of reported attributes varies widely by user, ranging from 4 to 893 attributes, with an average of 247 and a median of 153. Across all users, we find that most reported attributes were interest-based (93%), followed by behavior-based (5%) and demographic-based (2%).

We also examine how often these reported attributes change (recall that we collect the reported attributes daily for each user). We measure changes using *divergence*, which is simply

$$|Set_{day1} \oplus Set_{day2}|$$

where $\oplus$ denotes the disjunctive union of the sets. Thus, the divergence is simply the number of attributes added or removed. Across all users, we find that the average daily divergence ranges from 0 to 82, with an average of 10.7. Thus, we see that the inferred attributes change somewhat rapidly (on average, 4.3% of attributes change per day).

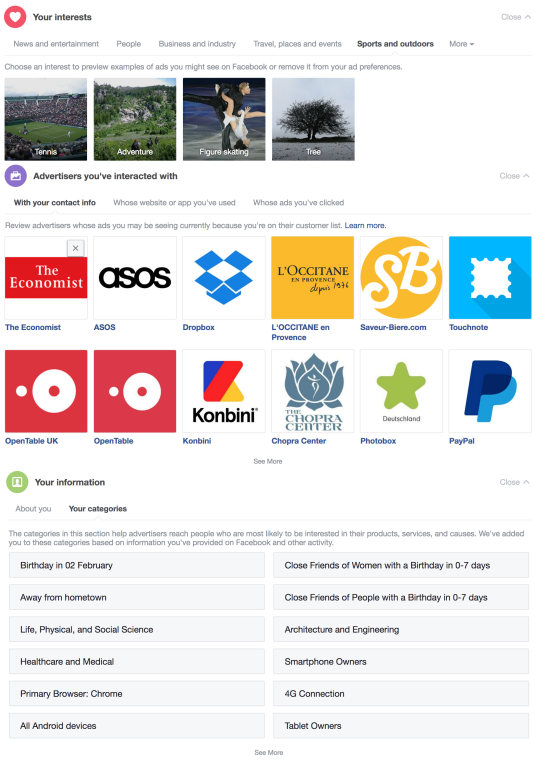


Fig. 3: Example of information provided in the Ad Preferences Page.

Next, we turn to examine whether the explanations meet the properties we outlined in Section IV-B. Recall that Facebook only provides data explanations for interest attributes; thus, these are the explanations we examine for the remainder of this section.

2) Specificity: Out of the 9,929 different data explanations we collected, we extracted five distinct patterns; these are shown in Table V. The explanations are usually short, generic, and they mostly refer to ad clicks, page likes or app installations. While explanations that refer to app installs, as well as explanations that refer to preferences that the users added themselves, are *precise*, the majority (97%) of data explanations are not. For example, the vast majority of interest explanations are due to liked pages and ad clicks, but Facebook does not specify *which* page or ad led to the interest attribute.

3) Snapshot completeness: To evaluate the snapshot completeness, we test whether Facebook allows advertisers to target users based on attributes that do not appear in their Ad Preferences Page. Thus, for each user, we check whether there are attributes that appear in their ad explanations but

TABLE V: Overview of data explanations we observed.

Pattern	Explanations
You have this preference because you liked a page related to [interest]	4,518
You have this preference because you clicked on an ad related to [interest]	4,352
You have this preference because we think it may be relevant to you based on what you do on Facebook, such as pages you've liked or ads you've clicked	785
You have this preference because you installed the app [app]	249
This is a preference you added	25

which never appeared in their Ad Preferences Page, we call them these *hidden attributes*. In our dataset, we found a total of 205 hidden attributes for 24 distinct users, 55 of these are profile attributes such as schools, languages, or relationship status, and the rest are interest-, behavior-, or demographic-based attributes. It is important to note that this does not mean explanations are definitely incomplete, as we may have missed some attributes that only appeared briefly in the Ad Preferences Page (i.e., for less than one day).

To verify whether we can target people with attributes that do not appear in their Ad Preferences Page, we launched several controlled experiments targeting an audience with different attributes that are not present in a user's Ad Preferences Page. If the monitored user receives an ad from one of these campaigns with an ad explanation containing the attribute, it means that Facebook allows advertisers to target him with attributes that are not shown in the Ad Preferences Page.[9]

We tested six data broker attributes, out of which two resulted in successful campaigns with a data broker explanation for a monitored user; we also tested four profile data and language attributes, out of which two were observed in a data explanation for at least one monitored user. While we observed that most of the profile data attributes appear in some form in the "About Page", or "Facebook Settings" of a user, we observed that *no* data broker attributes appear in the Ad Preferences Page (or other places) of any of our monitored users. According to a statement by a Facebook representative [15], the absence of data broker attributes from the Ad Preferences Page is a deliberate choice, motivated by the fact that the data was not collected by Facebook. Due to this decision, Facebook's data explanations are not complete, as no data broker attributes are ever shown to users.

4) Temporal completeness: Despite the rapid changes in inferred attributes that we observe above, Facebook does not provide any historical information about the attributes it had inferred about a user. Thus Facebook's data explanations do not exhibit temporal completeness.

5) Correctness: Testing correctness precisely is challenging, as the provided data explanations are vague and do not reveal the exact page the user liked, or the ad the user clicked.

[9]In the Self-Serve Ads Terms Facebook says "In instances where we believe doing so will enhance the effectiveness of your advertising campaign, we may broaden the targeting criteria you specify." Thus, to be sure that the user received the ad because Facebook thinks he in interested in the attribute, it is not enough for the user to receive the ad of our ad campaign, but the attribute also needs to appear in the explanation.

In order to briefly test correctness, we created a fake Facebook account, and liked 7 Facebook pages related to U.S. Politics and 15 pages related to TV Shows. We run the experiment in a controlled environment, in a browser with no history, and we did not perform any other actions on Facebook besides liking the mentioned pages. From these 22 likes, Facebook inferred 27 interests; all of these interests had data explanations like "You have this preference because you liked a page related to [interest]." Thus, we did not find any indication that explanations were incorrect. While a more comprehensive set of experiments is required for more complete results, we leave such an exploration to future work.

E. Summary

While the Ad Preferences Page does bring some transparency to the different attributes users can be targeted with, the provided explanations are *incomplete* and often *vague*. Facebook does not provide information about data-broker-provided attributes in its data explanations or in its ad explanations. This means that currently users have no way of knowing what data broker attributes advertisers can use to target them. This is despite the fact that close to half of the targeting attributes come from data brokers and they have an audience reach similar to Facebook's own targeting attributes.

V. Related work

A. Bringing transparency to targeted advertising

While there have been many studies on online advertising, ad auctions, tracking, and ad blocking in general, we focus next only on the studies that are the closest to our proposal; we refer the reader to [44] for a more general overview of the work in the space. We split the related works according to the kind of transparency they aim to provide.

Ad-level transparency: Two studies [17], [34] proposed techniques to detect whether an ad is contextual, re-targeted or behavioral. A few other studies took the next step and proposed methods to detect *why* the ads are being targeted, that is, what particular user action triggered the targeting of a particular ad [30], [31], [20], [36]. At a high level, these approaches monitor the actions of users (e.g., the emails users receive and send, the videos users see on YouTube) and they propose methods to estimate the likelihood that a given ad was shown due to a given input by performing controlled experiments. In contrast, we investigate how explanations provided by Facebook reveal information about why an ad has been targeted.

User-level transparency: Closest to our work are three tools: Floodwatch [6] and EyeWnder [4] collect the ads people receive while browsing the internet and provide aggregate statistics about them; and MyAdChoices [36] detects whether an ad is interest-based, generic, or retargeted, and allows users to selectively block certain types of ads. None of the tools focus on social media advertising and they do not analyze ad explanations. Two other studies analyzed the Google Ad Settings [7] (which is the equivalent of the Facebook Ad Preferences Page). Datta et al. [20] checked whether users receive different ads if they change their *categories* in the Google Ad Settings in order to detect discrimination. Willis et al. [43] investigated whether the Google Ad Setting pages

reveal *all* the categories Google inferred about a user and found that some behavioral ads were not explained by the revealed inferred categories. In contrast, we provide definite proof that Facebook makes available more targeting attributes to advertisers than it reveals to users.

Platform-level transparency: A few measurement studies bring insights into various aspects of the ad ecosystem. Barford et al. [16] focus mainly on presenting aggregated statistics by crawling ads at large scale. Using experiments based on artificial personas, they also study the relation between personas and advertiser categories and test whether an ad is behavioral. This study, however, does not focus on social media ad targeting but rather on the traditional ad ecosystem that targets users when they browse the Internet.

B. Analyzing Facebook's advertiser interface

A number of studies have investigated Facebook's advertiser interface and its pitfalls. For instance, ProPublica, an investigative journalism organization, showed that advertisers can create ads related to housing, while excluding users based on race, an act which is illegal [13]. More recently, ProPublica, as part of their "Breaking the Black Box" series [14], investigated whether Facebook informs users sufficiently about the use of data brokers in advertising [15] and found that while advertisers can target users with attributes provided by data brokers, they do not mention it in the Ad Preferences Page. Our work confirms this finding but also goes beyond in investigating other types of transparency.

Finally, Korolova et al. [28] proposed an attack that exploits Facebook's advertiser interface to infer private attributes of Facebook users. Later work by Venkatadri et al. [41] demonstrated that more advanced attacks are possible through the custom audience advertiser interface. However, the focus of these studies is not transparency, but on pinpointing vulnerabilities in the advertising interface.

C. Interpretability of decision making systems

Transparency and interpretability have been the focus of many recent studies in the context of automated decision making systems, with many previous works acknowledging the importance of having more interpretable models [22], [42], [33]. A first line of work focuses on providing explanations to existing algorithms/decision making systems, by studying for example, what are the inputs that have the biggest impact on the outputs [21], or by uncovering how the model behaves locally around specific predictions [38]. A second line of work aims at building algorithms that are interpretable by design by integrating interpretability constraints in their optimization functions [27], [32]. The main use-case for interpretable models is for the domain experts to understand whether the algorithm is behaving appropriately or not. In our work, we study explanations that are provided to users with the goal of making sure that users get satisfactory and useful explanations. Our work offers a different perspective on how to build good explanations and, to our knowledge, is the first empirical study of real-world explanations in social media advertising.

While many studies emphasize that explanations and transparency mechanisms bring trust to a platform [33], [38], Weller [42] warns that platforms can manipulate users to trust their system, with explanations that are not useful to them. The "Copy Machine" study [29] shows that useless explanations that did not provide any actual information were almost equally successful in gaining trust as meaningful explanations. Our study shows the different ways in which explanations offered by Facebook fail to provide adequate information to end users or worse, provide them with misleading information.

VI. Conclusion

In this paper, we investigated transparency mechanisms for social media advertising by analyzing Facebook's ad explanations and data explanations. We devised a set of key properties that such explanations could satisfy, such as correctness, completeness and specificity; we then performed a series of controlled ad campaigns to analyze whether Facebook's explanations satisfy such properties.

Our experiments demonstrated that Facebook's ad explanations are often incomplete and sometimes misleading, and that Facebook's data explanations are incomplete and often vague. These findings have important implications for users, as they may lead them to incorrectly conclude how they were targeted with ads. Moreover, these findings also suggest that malicious advertisers may be able to obfuscate their true targeting attributes by hiding rare (and potentially sensitive) attributes by also selecting very common ones. To make matters worse, Twitter recently introduced explanations that are similar to Facebook's explanations. This underscores the urgent need to provide properly designed explanations as social media advertising services mature. We hope that our study will provide a basis to guide such a design.

To complement our work, it would be interesting to perform a study on how users react to different possible explanations that can be provided. This would explore another dimension that could further inform the explanations' design choices. Yet we believe that it is important first to understand explanations at a technical level in order to understand their vulnerabilities. Hence, we leave such a study for future work.

Facebook's explanations only provide a partial view of its advertising mechanisms. To move towards greater transparency we built a tool, AdAnalyst, that works on top of Facebook and provides explanations with some of the missing properties. AdAnalyst keeps track of historical data about ads and explanations to provide users with a temporal view; and it provides a wider perspective by aggregating data across users. The tool can be downloaded and installed from http://adanalyst.mpi-sws.org/. We hope that AdAnalyst will help increase the transparency of Facebook advertising and that it will allow users to detect malicious and deceptive advertising.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF through grants CNS-1563320 and CNS-1616234, by ANR through grants ANR-17-CE23-0014 and ANR-16-TERC-0012-01, by Institut Mines Telecom through the "Future & Ruptures" program and by a Data Transparency Lab grant. We acknowledge funding from the Alexander von Humboldt Foundation as well.

REFERENCES

[1] "About potential reach," https://www.facebook.com/business/help/1665333080167380, accessed: 2017-11-30.

[2] "About the delivery system: Ad auctions," https://www.facebook.com/business/help/430291176997542, accessed: 2017-11-30.

[3] "Datalogix segments," http://bit.ly/2qzt5oI, accessed: 2017-11-30.

[4] "Eyewnder," http://www.eyewnder.com/, accessed: 2017-11-30.

[5] "Facebook ad preferences," https://www.facebook.com/ads/preferences/, accessed: 2017-11-30.

[6] "Floodwatch," https://beta.floodwatch.me/, accessed: 2017-08-11.

[7] "Google ad settings," https://myaccount.google.com/, accessed: 2017-11-30.

[8] "US voter list information ," http://voterlist.electproject.org/, accessed: 2017-11-30.

[9] "EU General data protection regulation," Apr. 2016, accessible from https://www.eugdpr.org/.

[10] "LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique," Journal Officiel de la République Française n° 0235 du 8 octobre 2016, Oct. 2016, accessible at https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte.

[11] Acxiom, "Consumer data products catalog," http://bit.ly/2rjzWFT, accessed: 2017-11-30.

[12] ——, "Privacy faq," http://bit.ly/2qupYAo, accessed: 2017-11-30.

[13] J. Angwin and T. Parris Jr., "Facebook lets advertisers exclude users by race," http://bit.ly/2eXf7ap, October 28, 2016, accessed: 2017-11-30.

[14] J. Angwin, T. Parris Jr., and S. Mattu, "Breaking the black box: What Facebook knows about you," http://bit.ly/2driPIj, September 28, 2016, accessed: 2017-11-30.

[15] ——, "Facebook doesn't tell users everything it really knows about them," http://bit.ly/2ieiNsq, December 27, 2016, accessed: 2017-11-30.

[16] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *WWW*, 2014.

[17] J. M. Carrascosa, J. Mikians, R. Cuevas, V. Erramilli, and N. Laoutaris, "I always feel like somebody's watching me: measuring online behavioural advertising," in *ACM CoNEXT*, 2015.

[18] J. Constine, "Facebook lets businesses plug in CRM email addresses to target customers with hyper-relevant ads," http://tcrn.ch/2q0JdxP, September 20, 2012, accessed: 2017-11-30.

[19] B. Darwell, "Facebook platform supports more than 42 million pages and 9 million apps," http://bit.ly/28YXb1H, April 27, 2012, accessed: 2017-11-30.

[20] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," in *PETS*, 2015.

[21] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *IEEE S&P*, 2016.

[22] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint 1702.08608*, 2017.

[23] Experian, "Product and service privacy policies," http://www.experian.com/privacy/prod_serv_policy.html, accessed: 2017-11-30.

[24] Facebook, "How does the conversion pixel track conversions?" http://bit.ly/2peqORu, accessed: 2017-11-30.

[25] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, "On the reliability of profile matching across large online social networks," in *ACM KDD*, 2015.

[26] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," in *WHI*, 2016.

[27] B. Kim, J. A. Shah, and F. Doshi-Velez, "Mind the gap: A generative approach to interpretable feature selection and extraction," in *NIPS*, 2015.

[28] A. Korolova, "Privacy violations using microtargeted ads: A case study," in *IEEE ICDMW*, 2010.

[29] E. J. Langer, A. Blank, and B. Chanowitz, "The mindlessness of ostensibly thoughtful action: The role of 'placebic' information in interpersonal interaction." *Journal of personality and social psychology*, 1978.

[30] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation." in *USENIX Security*, 2014.

[31] M. Lecuyer, R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu, "Sunlight: Fine-grained targeting detection at scale with statistical confidence," in *ACM CCS*, 2015.

[32] B. Letham, C. Rudin, T. H. McCormick, D. Madigan *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, 2015.

[33] Z. C. Lipton, "The mythos of model interpretability," in *WHI*, 2016.

[34] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "AdReveal: Improving transparency into online targeted advertising," in *ACM HotNets*, 2013.

[35] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *IEEE S&P*, 2012.

[36] J. Parra-Arnau, J. P. Achara, and C. Castelluccia, "MyAdChoices: Bringing Transparency and Control to Online Advertising," *ACM Trans. Web*, 2017.

[37] A. C. Plane, E. M. Redmiles, M. L. Mazurek, and M. C. Tschantz, "Exploring user perceptions of discrimination in online targeted advertising," in *USENIX Security*, 2017.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *ACM KDD*, 2016.

[39] N. Stokes, "Should you use Facebook or Google to log in to other sites?" http://bit.ly/1kxEP3X, May 6, 2017, accessed: 2017-11-30.

[40] D. Tynan, "Acxiom exposed: A peek inside one of the world's largest data brokers," http://bit.ly/2qvQQjy, May 15, 2013, accessed: 2017-11-30.

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

[42] A. Weller, "Challenges for transparency," in *WHI*, 2017.

[43] C. E. Wills and C. Tatar, "Understanding what they do with what they know," in *ACM WPES*, 2012.

[44] S. Yuan, A. Z. Abidin, M. Sloan, and J. Wang, "Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users," *arXiv preprint arXiv:1206.1754*, 2012.

ATTACHMENT #2

Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface

Giridhari Venkatadri[†], Athanasios Andreou[§], Yabing Liu[†],
Alan Mislove[†], Krishna P. Gummadi[‡], Patrick Loiseau[*‡], Oana Goga[*]
[†]Northeastern University [§]EURECOM [‡]MPI-SWS [*]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG

***Abstract*—Sites like Facebook and Google now serve as de facto data brokers, aggregating data on users for the purpose of implementing powerful advertising platforms. Historically, these services allowed advertisers to select which users see their ads via *targeting attributes*. Recently, most advertising platforms have begun allowing advertisers to target users *directly* by uploading the personal information of the users who they wish to advertise to (e.g., their names, email addresses, phone numbers, etc.); these services are often known as *custom audiences*. Custom audiences effectively represent powerful *linking* mechanisms, allowing advertisers to leverage any PII (e.g., from customer data, public records, etc.) to target users.**

In this paper, we focus on Facebook's custom audience implementation and demonstrate attacks that allow an adversary to exploit the interface to infer users' PII as well as to infer their activity. Specifically, we show how the adversary can infer users' full phone numbers knowing just their email address, determine whether a particular user visited a website, and de-anonymize all the visitors to a website by inferring their phone numbers en masse. These attacks can be conducted without any interaction with the victim(s), cannot be detected by the victim(s), and do not require the adversary to spend money or actually place an ad. We propose a simple and effective fix to the attacks based on reworking the way Facebook de-duplicates uploaded information. Facebook's security team acknowledged the vulnerability and has put into place a fix that is a variant of the fix we propose. Overall, our results indicate that advertising platforms need to carefully consider the privacy implications of their interfaces.

I. Introduction

Data brokers are businesses whose revenue model revolves around aggregating information about individuals from a variety of public and private sources. Traditional data brokers include Acxiom, Datalogix, and Equifax, who sell access to the collected data to third parties, including advertisers, marketers, and political campaigns. Recently, online services like Google and Facebook have become *de facto* data brokers: while they do not typically sell access to the collected data directly, they instead use the collected data to build powerful advertising services that have data on billions of users worldwide. Consequently, any breaches or hacks potentially threaten the privacy of large numbers of users, making them vulnerable to potential fraud, harassment, and identity theft attacks [6].

These new data brokers provide interfaces to their advertising services that enable advertisers to target their ads to users with specific attributes (e.g., all 35-year-old males living in Detroit); these groups are referred to as audiences. The interfaces provide advertisers with basic statistics about their selected audience, including an estimate of its size. In ground-breaking work in 2011, Korolova [18] demonstrated that malicious Facebook advertisers could select attributes that are "microtargeted", or chosen so that they are satisfied only by a single user. As a result, Korolova was able to use the audience size statistics to infer users' demographic information that was set to be private. In response, Facebook disallowed microtargeting, placing a minimum size on audiences.

Recently, data brokers such as Facebook [42] and Google [1] have introduced a new feature on their advertising interfaces: *custom audiences*. Instead of creating audiences based on user attributes, advertisers can now upload personally identifying information (PII) about specific users; the platform then locates matching accounts and creates an audience consisting of only these users. The advertiser can then use this audience when placing ads, thereby showing their ads only to the specific users whose information they uploaded. For example, a small business may know the names and addresses of its customers; using custom audiences, the business can upload this information to Facebook, and then target these users with advertising directly. The custom audience feature has proven popular with advertisers: it allows them to directly select *the users* to whom their ad is shown, as opposed to only selecting *the attributes of the users* [42], [28].

At its core, the custom audience feature is a *linking* mechanism, enabling advertisers to link various forms of user PII that they have collected to the information collected by the advertising platform. Advertisers come with various pieces of user PII—they may have the email addresses of some people, the names and addresses of others, and the phone numbers of a few more—and the platform links all of these disparate pieces of PII to the users' accounts. Unfortunately, if not implemented carefully, the custom audience feature can inadvertently leak bits of information about users to advertisers.

In this paper, we focus on auditing Facebook's advertising interface, as Facebook's advertising service is one of the most mature and well-used advertising platforms. Indeed, we show that Facebook's advertising interface leaks sensitive personal information about users by reporting coarse-grained information about the size of a custom audience. For example, malicious Facebook advertisers can infer the phone number of a user given only their email address, or they can infer whether a given user has visited a webpage the advertiser controls. None of these attacks require any interaction with the victim, none can be detected by the victim, and all can be performed without actually placing any ads or paying any

money. These attacks can be particularly devastating for user privacy: for example, they enable malicious users to infer the phone numbers of celebrities or politicians [31], [25], allow oppressive governments to identify and intimidate citizens who dissent online [30], enable adversaries to easily identify users' mobile numbers for purposes of "phone porting" attacks [27], and allow website operators to de-anonymize users who visit websites that may contain embarrassing or censored content.

To develop our attacks, we carefully study the features of the Facebook advertising interface. The attacks we present are enabled by two characteristics of this interface that we discover: *First*, we show that the size statistics reported by the interface are obfuscated using *rounding*; this enables us to create audiences that negate the effect of rounding by having a size that falls exactly on the rounding threshold. *Second*, we demonstrate that the interface *de-duplicates* multiple PII records that refer to the same user when reporting the size statistics; combined with the previous observation, we are able to determine whether two pieces of PII refer to the same user.

Overall, our paper makes four contributions:

- We characterize the Facebook custom advertising interface, revealing the two characteristics that enable our attacks.
- We show how an adversary can abuse the custom audience linking mechanism to infer other PII of a victim, knowing only basic information about the victim. This attack, for example, allows an adversary to infer the phone number or name/location of a victim, given their email address.
- We demonstrate how the adversary can use Facebook's third-party tracking Javascript to de-anonymize visitors to their website (e.g., by inferring visitors' phone numbers).
- We propose a mitigation that modifies how de-duplication is implemented. In brief, instead of de-duplicating at the *user* level, Facebook should de-duplicate at the *PII* level.

Ethics Throughout this study, we have ensured that all of our experiments meet community ethical standards. *First*, all of the input voter records that we used as a basis for creating threshold audiences are available as public records. *Second*, we did not collect or obtain the personal information of any Facebook users; our small set of users only provided us with their email addresses and phone numbers for validation. Our attacks did not interact with their accounts in any way, reveal any additional information about them that we did not already know, or provide to Facebook any information about them that Facebook did not already have. *Third*, we have responsibly disclosed all of the attacks to Facebook, providing them with early drafts of this submission. Their security team has acknowledged the vulnerability and has put into place a fix that is a variant of the fix we propose in Section VI.

II. Background

Most of the popular web and mobile application-based services today are funded by advertising, where the end users are provided the service for free in exchange for being shown ads.

Site	Name	Email	Phone number	City or ZIP	State or Province	Birthday, Gender	Employer	Site user ID	Mobile advertiser ID	Min. Size
Facebook	✓	✓	✓	✓	✓	✓	✗	✓	✓	20
Instagram	✓	✓	✓	✓	✓	✓	✗	✓	✓	20
Twitter	✗	✓	✓	✗	✗	✗	✗	✓	✓	500
Google	✓	✓	✓	✓	✗	✗	✗	✓	✓	1,000
Pinterest	✗	✓	✗	✗	✗	✗	✗	✗	✓	100
LinkedIn	✗	✓	✗	✗	✗	✗	✓	✗	✓	100

TABLE I: User attributes that advertisers can upload to create custom audiences in various advertising platforms. Also shown is the minimum custom audience size that the sites allow.

Most online advertising platforms are implemented using auctions, where advertisers bid on keywords or search terms in the traditional search-based advertising platforms (e.g., Google, Yahoo, Bing); or user demographics, interests, behaviors, and other user information in online social network advertising platforms (e.g., Facebook, Twitter, Pinterest). Whenever a user views a page, the platform runs an auction and displays the ads requested by the winning advertiser(s).

On most online social media advertising platforms, advertisers are allowed to create *audiences* (simply, groups of users) for convenience. For example, an advertiser may wish to advertise to a certain group of users multiple times; the advertiser is allowed to first define this audience (e.g., all users living in a certain region), and then can submit ads (and corresponding bids) to be shown to these users. Today, there are two primary ways an advertiser can create an audience:

1. Attribute-based audiences Advertisers can create audiences by specifying the properties of the users whom they wish to target. For example, advertisers can create audiences using any combination of targeting attributes, such as location (city or ZIP), age, gender, languages, likes, interests, etc.

2. PII-based audiences Recently, a new method for creating audiences has been developed: audiences created based on information that uniquely identifies specific users (i.e., using personally identifiable information, PII). These PII-based audiences come in two flavors. *First*, the advertiser can upload the PII of the users they wish to target (e.g., their email addresses, their names and cities of residence, etc.). The advertising platform then creates an audience of platform users matching the PII.[1] We refer to these audiences as *custom audiences*, with the nomenclature referring to the fact that the audiences are created in a custom fashion by the advertiser.

We provide an overview of the PII types that advertisers can upload when creating custom audiences on various platforms

[1]Most platforms provide advertisers with guarantees that the platforms themselves are not capturing customer information uploaded for custom audience creation. This is typically implemented using *hashing*, where all advertiser-provided data is hashed before being uploaded to the platform (in fact, some platforms allow advertisers to upload already-hashed data records) [41], [34], [16], [4].

in Table I. As can be observed, Facebook (and Instagram, owned by Facebook) allows the widest variety of personal information to be uploaded. However, all platforms allow advertisers to create custom audiences based on email addresses, and most allow creation based on mobile advertising IDs (essentially per-mobile-device identifiers) [40].[2] We note that the custom audience feature is given different names on different platforms: Facebook's Custom Audiences [42], [35], Twitter's Tailored Audiences [36], Google's Customer Match [1], Pinterest's Audiences [26], and LinkedIn's Audience Match [3].

Second, advertisers can create audiences based on users who interact with the advertiser's Facebook applications or (external) website. To facilitate external website tracking, the platform gives the advertiser some code (often referred to as a tracking pixel, as it was historically implemented as a one-pixel image) to include on their website; when users visit the advertiser's website, the code makes requests to the social media platform, thereby adding the user to an audience. We refer to these audiences as *tracking pixel audiences*.

It is important to underscore the distinction between PII-based audiences and traditional attribute-based audiences: with attribute-based audiences, advertisers could only specify the *attributes* of the users they wanted to target (e.g., an advertiser could create an audience of female users in Seattle). With PII-based audiences, advertisers instead specify the particular *users* they want to target, either by uploading known email addresses, names, or other personal information, or by selecting users who visited an external website the advertiser controls.

III. Facebook's PII-Based Audiences

For the remainder of the paper, we focus on Facebook's advertising platform, as it is the largest and most successful social media advertising platform. An advertiser can use a number of different features of Facebook's advertising interface when placing an ad. Only the first few of these steps are relevant to our attacks; the adversary does not need to actually place an ad. In this section, we describe the important features of Facebook's advertising interface relevant to PII-based audiences that we use throughout the paper.

A. Creating custom audiences

Advertisers create custom audiences using a web interface where they are allowed to upload 15 different types of user information; these include external unique identifiers such as Email, Phone number, and Mobile advertiser ID,[3] as well as fields related to name (First name, Last name), age (Date of birth, Year of birth, Age), sex (Gender), and location (ZIP/Postal code, City, State/Province, Country). In addition, advertisers can specify Facebook app user ID and Facebook page user ID, which are obfuscated identifiers that are generated when a user installs an advertiser's Facebook application or likes their page. Behind the scenes, Facebook first hashes advertiser-provided data before uploading it for matching, ostensibly to assure advertisers that the matching process will not reveal the advertisers' customer data [41].

We experimented with this interface to determine what set(s) of fields are *required* in order to initiate the matching process (e.g., would it be enough to simply upload a First name of "John" and match all users named John on Facebook?). Among the 15 fields, we found that only five of them can be used *alone* to create a custom audience: Email address, Phone number, Mobile advertiser ID, Facebook app user ID, and Facebook page user ID. If the advertiser wishes to use only the remaining 10 fields, they must provide First name, Last name, and one of the following sets of fields: (City, State/Province), (ZIP), or (Date of birth); they can provide additional information if desired. Finally, we note that advertisers are allowed to upload files with different information for different users (e.g., a file with the email addresses of some users, and the phone numbers of others).

It typically takes up to a few hours for Facebook to create the custom audience after the advertiser uploads the PII. As far as we can tell, there is no limit on the size of lists that can be uploaded; the audience creation process takes proportionally longer for longer lists, e.g., a few hours for 10 million records.

To prevent advertisers from targeting individual users (likely as a response to Korolova's work [18]), most platforms have policies about the minimum number of matched members for each PII-based audience to be usable. We show this information in Table I. While Facebook requires at least 20 individuals when creating a custom audience [24], most other sites, such as Google AdWords, Twitter, Pinterest, only allow advertisers to use a custom audience if it contains of hundreds of users or more. In the case of Facebook, it will actually create audiences with fewer than 20 users, but will not allow an advertiser to advertise to *only* that audience (more on this later).

Finally, once the audience is created, Facebook reports success and tells the advertiser the number of matched records (note that Facebook may not be able to match all of the PII records that were uploaded to Facebook accounts). We refer to the number of matched records as the *audience size*. The custom audience size is obfuscated, and we explore the mechanism by which it is obfuscated in the next section.

B. Creating tracking pixel audiences

Advertisers create tracking pixel audiences using a separate web interface by simply providing a name for the audience, and then including the Javascript code provided by Facebook in their external website. As with custom audiences, Facebook provides an obfuscated audience size and requires at least 20 individuals for the audience to be usable.

[2] LinkedIn is a business-centric social network and offers an entirely different feature, Employer, as their account-targeting option. LinkedIn allows advertisers to upload lists of up to 30,000 companies, and can target ads to those companies' employees. However, this feature is not self-service and involves working with the LinkedIn account team.

[3] Mobile advertiser ID represents a mobile-OS-provided identifier that is unique for each device (but can be reset by the user). This is useful for advertisers who wish to target mobile users who have already installed the advertiser's app.

C. Obtaining potential reach

The second advertising feature that we use is the potential reach. To place an ad, the advertiser can choose an existing PII-based audience to advertise to, or *combine* that audience with any other PII-based audiences that the advertiser has previously created. This avoids requiring the advertiser to submit multiple ads (one for each audience) or to re-upload the same data multiple times.

When combining audiences, Facebook allows advertisers to *include* users who appear in any number of existing PII-based audiences, and then *exclude* users who appear in any number of such audiences. When convenient, we denote these operations as union ($\cup$) and set-minus ($\setminus$). In testing the interface, we found that exclude trumps include (e.g., if a user is part of any excluded audience, the user will not be in the final combined set regardless of how many included audiences they appear on). For simplicity, we refer to the resulting set of users as the *combined audience*. It must be noted that Facebook allows the inclusion or exclusion even of audiences that have fewer than 20 users.

Once the advertiser has selected the combined audience they wish to advertise to, they then proceed to place a bid and upload the actual advertisement. As these steps are not necessary for our attacks, we do not discuss them in detail.

However, there is one feature of Facebook's interface that we do use: the number of *active* users that are in the combined audience. Facebook provides advertisers with this number, called the *potential reach*. Facebook [43] defines it as[4]

> *... the number of daily active people on Facebook that match the audience you defined through your audience targeting selections*

We refer to users who are "daily active" as *targetable*.

It is important to note how potential reach and audience size differ: the audience size only applies to PII-based audiences and includes *all* matched Facebook accounts, while the potential reach can be applied to both PII-based audiences and combinations of such audiences, but only includes "daily active" Facebook accounts. Thus, for a single audience, the potential reach is always less than or equal to the audience size. Additionally, we can obtain the potential reach for combinations of audiences via includes and excludes.

Similar to the audience size, the potential reach that Facebook reports is obfuscated; we explore this in the next section.

D. Determining audience intersection size

The final advertising feature that we utilize is the audience comparison page. Advertisers are likely to have created multiple PII-based audiences; in order to help them understand these audiences and decide which to use, Facebook's advertising interface allows advertisers to measure the overlap between different pairs of PII-based audiences that they have created. When convenient, we denote this operation as intersection ($\cap$).

[4]While this definition was provided at the time we conducted the experiments, Facebook has since changed the definition of potential reach [2]. However, the new definition still captures the notion of active users.

We interacted with this feature using some of the PII-based audiences we created, and observed it had two important characteristics. *First*, the interface only supports PII-based audiences with an audience size of at least 1,000; smaller audiences are not available when using this interface. *Second*, the intersection size shown is based on audience size, and *not* on potential reach. *Third*, this intersection size is obfuscated; we explore the obfuscation mechanism in the next section.

IV. Facebook Implementation Characteristics

Having described the key features of Facebook's interface, we now begin our analysis. We examined how the key features are implemented, and found two implementation characteristics that enable our attacks; we first describe the datasets we used for testing before describing the two characteristics. The experiments described in this section were performed between January and March 2017; Facebook's fix in response to our disclosure has since changed how some of these features work.

A. Datasets

To study how the PII-based audience interface is implemented, we need PII-based audiences to test with. To create custom audiences, we use two sources of data: *First*, we use the phone numbers and email addresses of 100 recruited friends and family members. For all of these users, we only use the data that Facebook already had listed in their accounts and was visible to us. *Second*, for experiments where we needed large numbers of records, we used public voter records from North Carolina [37]. In brief, records of registered voters in North Carolina are publicly available and contain voters' names and ZIP codes, among other fields. Obviously, not all registered voters will have Facebook accounts and both the Facebook and voter data could be out-of-date or inaccurate; thus, when uploading sets of voter records to create custom audiences, we expect that not all records would match a Facebook account.

To create a tracking pixel audience, we set up a test website with a Facebook tracking pixel created under our advertising account. We recruited 20 friends and family members with active Facebook accounts to visit the page.

B. Calculating size statistics via rounding

As noted in the previous section, there are three different mechanisms within Facebook's advertising interface that report the "size" of a PII-based audience: (1) the *audience size* representing the total number of matched accounts of a custom audience or accounts in a tracking pixel audience, (2) the *audience intersection size* representing the total number of accounts two PII-based audiences have in common, and (3) the *potential reach* representing the number of "daily active" users in a PII-based audience (or combination of such audiences). Recall that all three of these numbers were obfuscated in some way; we now determine how they are actually calculated.

Potential reach We first examine potential reach, described in Section III-C. We use our browser's developer tools to examine the network requests made by Facebook's advertising interface when we choose audiences to include or exclude. We

find that the interface makes `GET` requests to an API endpoint, which returns JSON objects that include the corresponding potential reach. We created a script to automate the process of making requests; we then used our script to collect the potential reach of over 58K different combinations of custom audiences (based on voter data) and targeting attributes, and found the output to have the following properties:

Granularity Looking at the distribution of potential reach values, we never observed any values smaller than 20, or that were not a multiple of 10; besides, we observed that larger values had few significant digits. Looking closely, we found that the values all come from the set

$$\{20, 30, ... \ 990, 1000, 1100, ... \ 10000, 11000, ..., 100000, ...\}$$

meaning they are multiples of 10 up to 1,000, multiples of 100 up to 10,000, multiples of 1,000 up to 100,000, and multiples of 10,000 beyond 100,000.

Consistency Next, we examine how consistent the potential reach values are over time. We first measure the consistency over short time scales by running 1,000 API queries back-to-back (roughly one query every second) for the same audience and targeting attributes. We repeat this experiment five different times with two different audiences, and find all the potential reach values within each run of 1,000 were exactly the same. This result shows that the audience size statistics are consistent over short intervals.

To measure consistency over longer time scales, we queried the API every five minutes for the potential reach of four different audiences, issuing 150 queries for each audience over a period of over 12 hours. The potential reach of three of the audiences remained constant, while the potential reach of the fourth audience changed slightly at one point during the 12 hour experiment and stayed at the new value. This result shows that across longer time scales, the potential reach values are largely consistent; the small changes in potential reach are consistent with Facebook's definition of potential reach as the number of *active* users in the audience.

Finally, we measure the consistency of custom audience creation by re-uploading 15 lists of PII multiple times within the same day to create custom audiences. We find that the potential reach remained consistent for 14 of the lists, with the final list changing by 10 users at one point. Again, this small change is consistent with the definition of potential reach.

Monotonicity Next, we examine whether the potential reach monotonically increases when we add more users to the list we upload. Specifically, we upload a series of lists such that each list has one record added beyond the previous one. Thus, in the corresponding series of custom audiences created, each audience would contain *at most* one additional Facebook account compared to the previous one (if Facebook could not find a match for the added record, or if the matched user was not active, we would expect the potential reach of the audience to be the same as that of the previous one).

We then study whether the corresponding potential reach values increase monotonically. For example, starting with a random sample of 300 voter records, we sequentially add 16 records to the list and upload all 17 lists. The potential reach for the initial sample was 30. This value increased to 40 as the first record was added, and then increased to 50 after the 14th record was added, showing that the potential reach statistics increase monotonically. We repeated this experiment with multiple audiences; in all cases, the same pattern held: the potential reach values increased monotonically, using the minimum granularity steps described above.

Summary Overall, our results strongly suggest that Facebook is *rounding* the raw potential reach value in steps of 10, 100, 1,000, or 10,000, depending on its magnitude.

Audience size Next, we examine the audience size, described in Sections III-A and III-B. We repeated all of the experiments that we conducted above for potential reach; for brevity, we omit the details and simply describe the inferred behavior. We observe that the audience size has values in $\{20, 30, ... \ 90, 100, 200, ...\}$, meaning they are multiples of 10 up to 100 and multiples of 100 thereafter. We also observe that audience size has greater consistency than potential reach, as we did not observe any cases where the number of matched accounts changed over short or long time periods; this makes sense, as audience size is only the number of accounts, and does not consider user activity. Finally, we observe that audience size has similar monotonicity to potential reach.

Summary We again observe that Facebook appears to be rounding the actual size of the audience in steps of 10 or 100, depending on its magnitude.

Audience intersection size Finally, we examine the characteristics of the audience intersection size, as described in Section III-D. We repeated all of the experiments that we performed on potential reach in Section IV-B on the intersection size calculation. We found the calculation to be consistent and monotonic, but with different granularity: the interface rounds the intersection size in steps of 5% of the *smaller* audience size. For example, if we intersected a audience with size 1,000 and another with size 4,500, the answer would be a multiple of 50.

Summary We observe that Facebook appears to be rounding the actual size of the intersection in steps of 5% of the smaller of the two audience sizes.

C. De-duplicating PII from the same user

Next, we describe the second implementation characteristic that enables our attacks. We found that when Facebook has an audience—or combination of audiences—that was created with different PII that both refer to the same user, Facebook only "counts" that user once when reporting the potential reach, audience size, or audience intersection size. For example, suppose that an advertiser uploaded two custom audiences: one containing the phone number of one of their customers (list A), and another containing the email address of the same customer (list B). If the advertiser then asked for the potential

	Audience Size	Potential Reach			
		Alone	$\cup\, C_{1\mathrm{p}}$	$\cup\, C_{1\mathrm{e},2\mathrm{p}}$	$\cup\, C_{1\mathrm{e},1\mathrm{p}}$
$C_{1\mathrm{e}}$	50	40	40	90	40
$C_{1\mathrm{p}}$	50	40	–	90	40
$C_{1\mathrm{e},2\mathrm{p}}$	100	90	–	–	90
$C_{1\mathrm{e},1\mathrm{p}}$	50	40	–	–	–

TABLE II: Audience size and potential reach returned by Facebook for different audiences, and combinations of audiences (described in Section IV-C). In all cases, Facebook appears to be de-duplicating records both *within* and *across* audiences before calculating the statistics.

reach of targeting users in the combination $A \cup B$, this user would only be counted once.

To demonstrate this behavior, we describe experiments we conducted on each of the three "size" mechanisms using data from our 100 friends and family members. We randomly divided these into two distinct groups: G_1 containing 50 users, and G_2 containing the remaining 50 users.

Audience size To test how Facebook de-duplicates audiences, we created four custom audiences:

- $C_{1\mathrm{e}}$, with the email addresses of all users in G_1 (50 records)
- $C_{1\mathrm{p}}$, with the phone numbers of all users in G_1 (50 records)
- $C_{1\mathrm{e},2\mathrm{p}}$, with the email addresses of all users in G_1 (50 records), and the phone numbers all users in G_2 (50 more records)
- $C_{1\mathrm{e},1\mathrm{p}}$, with the email addresses of all users in G_1 (50 records) and the phone numbers of all users in G_1 (50 more records)

We then record the audience size that Facebook reports. As shown in the second column of Table II, we found that $C_{1\mathrm{e}}$ and $C_{1\mathrm{p}}$ have an audience size of 50, and $C_{1\mathrm{e},2\mathrm{p}}$ has an audience size of 100, as expected. However, we found that $C_{1\mathrm{e},1\mathrm{p}}$ has an audience size of 50 users. We repeated this experiment with different sets of users of varying sizes, and found the same behavior. Thus, this experiment shows that when creating a custom audience, PII referring to the same user—even if present in separate records in the uploaded list—is de-duplicated when creating the custom audience.

Potential reach To test how Facebook de-duplicates combinations of audiences when calculating potential reach, we request the potential reach for different combinations of the audiences we created. For example, we ask for the potential reach if we "include" audiences $C_{1\mathrm{e}}$ and $C_{1\mathrm{p}}$. The results of this experiment are shown in Table II, and we make a number of observations. *First*, the third column of the table shows the potential reach value; this is typically lower than the audience size, as the potential reach only includes active users. *Second*, we observe that Facebook appears to be de-duplicating records, even *across* audiences. For example, when we "include" audiences $C_{1\mathrm{e}}$ and $C_{1\mathrm{p}}$, we observe that the total potential reach is 40, even though the potential reach of each audience individually is also 40. We ran similar experiments by combining custom audiences and tracking pixel audiences, and found the same behavior.

Audience intersection size Finally, we explore whether Facebook de-duplicates records when computing the audience intersection size. To do so, we need to modify our custom audiences, as the minimum custom audience size for using the audience intersection interface is 1,000. We create "extended" versions of $C_{1\mathrm{e}}$, $C_{1\mathrm{p}}$, $C_{1\mathrm{e},2\mathrm{p}}$, and $C_{1\mathrm{e},1\mathrm{p}}$ (denoted by $EC_{1\mathrm{e}}$, $EC_{1\mathrm{p}}$, etc.) by adding "padding" records from the North Carolina voter list; for each audience, we repeatedly add random records and re-upload until the audience size reaches 1,000 users.

We then measure the audience intersection size between the EC_i lists; we find that Facebook again appears to be de-duplicating *across* custom audiences. For example, the intersection size between $EC_{1\mathrm{e}}$ and $EC_{1\mathrm{p}}$ is 40, even though those two custom audiences were created using different PII. We repeated this experiment with a variety of different audience sizes and found the same behavior.

In summary, Facebook appears to be de-duplicating users when calculating the audience size, potential reach, and audience intersection size, both within a single audience and across combinations or intersections of audiences.

V. Attacks

In this section, we show how Facebook's advertising interface can be abused by an adversary to link different identities belonging to the same user. We describe three distinct attacks in this section, each with different threat models.

A. De-anonymizing web visitors

We begin by demonstrating the first of our three attacks, which allows an adversary running a website to determine if a particular Facebook user has visited their site.

Threat model We assume that the adversary runs a website where they have installed a Facebook tracking pixel, that the adversary wishes to determine whether or not a particular victim Facebook user has visited this website, and that the adversary knows enough PII of their victim that they can include them in a Facebook custom audience (e.g., their email address; see Section III-A). We assume that the adversary has access to Facebook's advertising system (it is important to note that any user can sign up for Facebook advertising, and no authorization or approval process is necessary). We also assume that the victim is a "daily active" Facebook user.

Threshold audiences The insight behind the method is to create a *threshold audience*: an audience or combination of audiences where the size statistic of interest—potential reach, audience size, or audience intersection size—lies right *before* or right *after* Facebook's rounding threshold. We call an audience with a size right before the rounding threshold a *lower threshold audience*, and an audience with a size right after the rounding threshold an *upper threshold audience*. For example, if Facebook rounds to the nearest 10 using the "5-or-higher" rule, an audience with a size of 84 would be a lower threshold audience (with size reported as 80), an audience with a size of 85 would be an upper threshold audience (with size

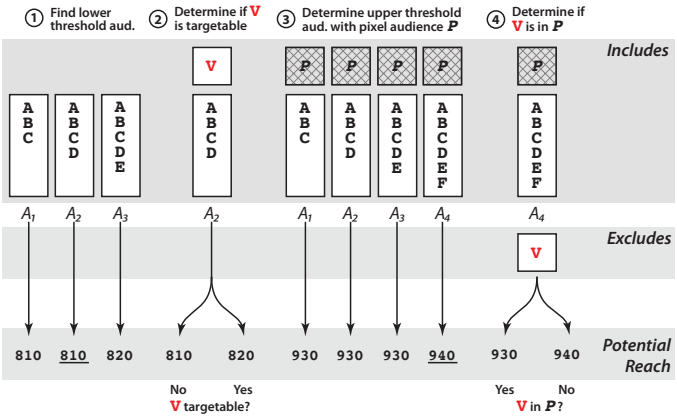


Fig. 1: Diagram of algorithm for determining if victim **V** visited the adversary's website. ① We first create custom audiences $A_1 ... A_n$ with increasing numbers of records and identify a lower threshold audience (e.g., A_2). ② We then determine if **V** is targetable by obtaining the potential reach of this audience union a custom audience V consisting only of **V** (i.e., $A_2 \cup V$). To determine whether **V** is in the tracking pixel audience $\boldsymbol{P}$, ③ we first identify an upper threshold audience containing $\boldsymbol{P}$ by successively combining $\boldsymbol{P}$ with each A_i (i.e., $A_4 \cup \boldsymbol{P}$).; ④ we then determine whether **V** is in $\boldsymbol{P}$ by excluding V from this combination (i.e., $A_4 \cup \boldsymbol{P} \setminus V$) and seeing if the potential reach drops.

reported as 90), and an audience with a size of 86 would *not* be a threshold audience. We use threshold audiences throughout the paper to enable our attacks.

Is the victim targetable? De-anonymizing web visitors boils down to determining whether a given user is part of a tracking pixel audience. To do so, we use the potential reach statistic, which only counts targetable (daily active) users. Thus, we first need to determine whether the victim is targetable; to do so, we upload a series $\{L_1, L_2, ..., L_n\}$ of lists to Facebook, each containing one additional record beyond the previous one. To create these lists, the adversary can use any source of PII (voter records, random phone numbers, etc.). Facebook creates custom audiences $\{A_1, A_2, ..., A_n\}$ corresponding to each of these lists. We also create a custom audience V consisting only of our victim user.

For each A_i, we obtain the potential reach of that audience. We then identify a lower threshold audience as the last audience before the potential reach changes (if multiple such audiences exist, we can simply choose one of them). For example, if the potential reach of A_1 and A_2 is 810 and the potential reach of A_3 is 820, then A_2 is our lower threshold audience (as A_3 is simply A_2 with one additional user).

Once we have identified the lower threshold audience, we then ask for the potential reach of the lower threshold audience *union* V; if the potential reach changes, we know that we can target the victim, otherwise, the victim is not a "daily active" user or our external information was not sufficient to target them. Continuing the example above, we examine the potential reach of $A_2 \cup V$: if the potential reach is 820, the victim has an active Facebook account; if it is 810, they do not. This process is shown in the first two steps of Figure 1.

Is the victim in the tracking pixel audience? If the victim is targetable, we can then proceed to determine whether they are in the tracking pixel audience P (i.e., have visited the adversary's webpage). We first find an upper threshold audience for potential reach by successively combining P with the A_i lists we uploaded. For example, if the potential reach of $A_3 \cup P$ is 930 and $A_4 \cup P$ is 940, then $A_4 \cup P$ is an upper threshold audience. Finally, we take this upper threshold audience and exclude V, containing only our victim (e.g., we ask for the $A_4 \cup P \setminus V$). If the potential reach drops, we know the victim is in P; if not, we know the victim is not in P.

Evaluation To evaluate the effectiveness of this attack, we recruited 40 volunteers from our friends and family; each provided us with the email address with which they log into Facebook. Using our Facebook advertiser account, we created a Facebook tracking pixel and installed it on a webpage we control. We divided the group in half, and had 20 volunteers visit this webpage from the browser they normally use; we had the remaining 20 volunteers not visit the website.

We then conducted the attack using the 40 email addresses to determine whether they had visited our webpage. We successfully inferred that all 20 of the volunteers who did not visit our webpage were not part of the tracking pixel audience. We also successfully inferred that 18 of the volunteers who did visit the webpage were part of the tracking pixel audience. We closely investigated the two users where the attack did not succeed: we found that both had an anti-tracking browser extension installed that prevented the Facebook tracking pixel from communicating with Facebook. Thus, our attack succeeded for all users who visited our webpage and became members of the tracking pixel audience.

B. *Inferring a victim's PII*

We now turn to demonstrate a more powerful attack: an adversary with knowledge of only a victim's email address (or other PII) can infer a victim's other PII (e.g., their phone number, or their name and city/state). This attack also demonstrates how an attacker can use audience size statistics instead of potential reach, meaning all Facebook users are vulnerable to this attack (i.e., not just "daily active" users).

Threat model Our threat model for this attack is similar to the threat model in Section V-A, with two modifications: *First*, the adversary is not running a website, but has a piece of PII for a victim user (e.g., their email address) and wishes to infer other PII. *Second*, the victim no longer needs to be a "daily active" user; they only need to have a Facebook account.

Is the victim in a custom audience? Our attack relies on the ability to determine whether the victim is a member of a custom audience the adversary uploaded; let us call this audience A, created from PII list L. In contrast to the previous

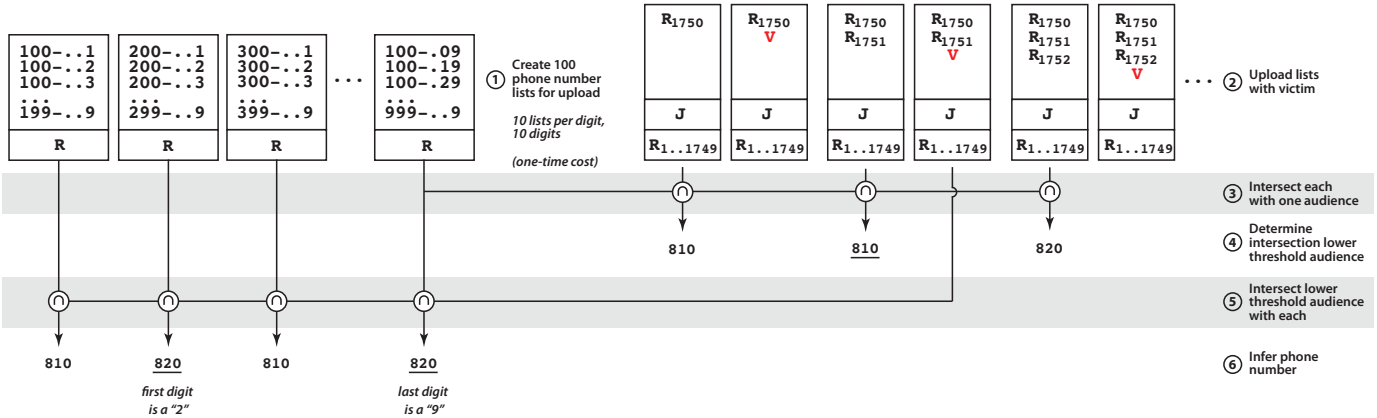


Fig. 2: Algorithm for determining phone number of the victim **V**, using two sets of extra records: R, a set of 1,949 records, and J, a set of 200 records. ① We upload lists of all phone numbers sharing a certain digit along with R. ② We then upload lists of 1,749 members of R with increasing additional members from R, both with and without **V**. ③ Next, we identify the threshold audience by intersecting each of the audiences without **V** with one of the phone number audiences; ④ the threshold audience is right before the intersection size changes. ⑤ We then intersect the threshold audience with **V** with each of the phone number audiences. ⑥ We infer each digit of **V**'s phone number by looking for where the intersection size is higher.

attack, we aim to use the audience size statistic to do so; thus, we need a different approach than before.

To conduct the attack, we rely on the custom audience intersection size feature described in Section III-D. The naïve approach would be to upload a custom audience containing only the victim and then ask Facebook for the size of the intersection between A and this audience; if there is any intersection, we know the victim is in A (and if there is not, we know the victim is not in A). Unfortunately, there are two properties of the intersection feature that make this approach infeasible: Facebook only allows intersections of custom audiences of at least 1,000 users each, and they only show the intersection at the granularity of 5% of the smaller audience size. We therefore must go through extra steps in order to determine whether the victim is in A.

We use a set R of 1,949 records of other users' PII (e.g., voter records) and a separate set J consisting of 50 additional records (R, J, and L should not share any users). We will describe the attack first making a simplifying assumption for clarity; we will show how to remove the assumption shortly. Let us assume for the moment that all records in R and J can be matched to Facebook accounts. We then upload two custom audiences to Facebook:

- C_1: The union of R and L
- C_2: The union of R, J, and the victim's PII

We can observe that both C_1 and C_2 have at least 1,000 users (due to the size of R), and can therefore be used in the custom audience intersection feature. We can further observe that the intersection between the two audiences is determined solely by R and whether the victim is a member of A (recall that R, L, and J share no users, and that list L matches audience A). Finally, we can observe that the smaller of the two custom audiences (C_2) has exactly 2,000 users, and that, by construction, the intersection is a lower threshold audience: we know the intersection size is either $|R|$ (if the victim is not in A) or $|R|+1$ (if the victim is). Since the interface will round to the nearest 5% of 2,000 (i.e., 100) and $|R|$ was selected with 1,949 users, and since Facebook de-duplicates users when computing intersections, the custom audience intersection size that Facebook returns for C_1 and C_2 will either be 1,900 (if the victim is not in A) or 2,000 (if the victim is). Thus, we can determine whether the victim is in A, and therefore in PII list L.

Dealing with unmatched users We now show how to remove the assumption that all records in R and J will match Facebook accounts. To do so, we need to address the fact that removing this assumption will prevent us from creating a lower threshold audience by design. Instead, we must find a lower threshold audience in a similar manner to how we did before. To do so, we "hold back" 10% of R (i.e., 200 records, twice the rounding threshold[5]) and upload multiple versions of C_2; we also increase the size of J to account for the fact that not all uploaded records will match (e.g., we can create J from 200 records). Specifically, we upload a series of custom audiences C_2^i and C_3^i, for i =1749...1949, where

- C_2^i: The union of $R_{1..i}$, J, and the victim's PII
- C_3^i: The union of $R_{1..i}$ and J

where $R_{1..i}$ denotes PII records 1 through i in R. We can then use the C_3^i to first locate a lower threshold intersection audience in a similar manner as before: we intersect each C_3^i with C_1, and choose the C_3^i before the intersection size changes. Let us call this lower threshold intersection audience C_3^k. We then intersect C_1 with C_2^k (the latter being simply C_3^k with the victim's PII added in); if the intersection size is the same as before, we know that the victim is not in A. If the intersection size increases, we know the victim is in A.

[5]We choose to hold back *twice* the rounding threshold worth of users as we need to ensure that we "cross" a rounding threshold when creating the C_3^i audiences, enabling us to find a lower threshold audience

Applying the attack to phone numbers We now show how to apply this attack to infer a victim's phone number. To do so, we partition the space of all possible phone numbers into sets that all share a common digit. Specifically, we create subsets L_{ij} of all possible phone numbers, where L_{ij} is a set of phone numbers whose i-th digit has a value of j. For example, if phone numbers are eight digits, L_{12} will contain `20 00 00 00`, `20 00 00 01`, up through `29 99 99 99`. This gives us $10d$ sets, each of size 10^{d-1}, where d is the number of digits in the number.

With each of these lists, we are now ready to conduct the attack. We simply repeat the methodology proposed above, determining whether the victim is a member of each L_{ij}. Note that because J and R should never share any users with L_{ij}, the adversary should choose users for R and J from a different country than the one the victim is from. For each i, we should find exactly one j such that the victim is in L_{ij}. Figure 2 gives a diagram of this attack.

Evaluation To evaluate the effectiveness of this attack, we recruit 22 friends and family members who have Facebook accounts, from two regions: 14 from the Boston area and 8 from France. In order to infer these users' phone numbers, we create the L_{ij} lists for both regions.

For Boston, we create a total of 140 lists: two area codes (617 and 857), where phones of each area code have 7 digits we need to infer (recall that US phone numbers have the structure `XXX YYY ZZZZ` where `XXX` is the area code). Each list contains 1M phone numbers, all with a single digit in common. Each of these lists took around 30 minutes to upload.

For France, we create a total of 82 lists: French mobile phone numbers have nine digits where the first digit is either 6 or 7. We generate all 200M possible numbers starting with 6 or 7 and use this to construct our sets (and hence each L_{ij} contains 10M phone numbers); it took over four hours to upload each list. We uploaded all 82 lists of phone numbers (two to determine the first digit, and 10 for each of the remaining eight digits) over a period of a week.

It is important to note that uploading these initial lists is, by far, the most expensive part of the attack. However, the resulting audiences can be re-used to infer the phone number of any user (i.e., the L_{ij} audiences are not victim-specific).

We then conduct our attack to infer the phone numbers of all 22 users. We find that we are successfully able to infer the numbers of 11 of the 14 users in Boston, and of all 8 of the users in France. In the cases where we succeeded, we were able to infer each users' phone number in under 20 minutes. We carefully examined the three users on whom the attack failed, and we found that one user had never provided their phone number to Facebook, while the other two users had actually provided *multiple* phone numbers to Facebook. As a result, we inferred that these two users were members of multiple L_{ij} for a given digit i (e.g., we inferred that a user's second digit is both a 2 and a 5); in the Appendix, we describe and evaluate a modified approach for determining both of the underlying numbers. For the one user that had never provided their phone number, no L_{ij} matched for any digit.

C. De-anonymizing users en masse

Recall that in Section V-A, we demonstrated an attack that allowed the adversary to determine if a specific user was a member of a tracking pixel audience. Utilizing the techniques developed in the previous attack, we now present a much more powerful version of that attack: the ability to infer en-masse the PII of the members of any audience. For example, we show how this can be used to de-anonymize all visitors to a webpage: by placing Facebook's tracking pixel on the webpage, the adversary can determine the phone numbers of all members of this audience.

Threat model As in Section V-A, we assume that the adversary runs a website where they have installed a Facebook tracking pixel, and that the adversary wishes to de-anonymize all visitors to this website. We assume that the visitors are "daily active" Facebook users (those visitors who are not "daily active" are not subject to the attack). In addition, as before, we assume that the adversary has access to PII of some Facebook users—e.g., from voter records—in order to be able to create threshold audiences.

Inferring the first digit(s) Let us denote our tracking pixel audience as P. We first find an upper threshold audience containing P using the same technique as in Section V-A: we upload a series of increasing A_i audiences and identify an A_t such that $A_t \cup P$ falls just over the rounding threshold.

We then use the L_{ij} from Section V-B, where each L_{ij} contains all possible phone numbers whose i-th digit has a value of j. Specifically, we check each of the L_{1j} to determine if any of the users in P have each possible $j \in 0..9$ as the first digit. To do so, we simply exclude the L_{1j} from $A_t \cup P$, denoted $A_t \cup P \setminus L_{1j}$; if the potential reach of this combination is lower than the potential reach of $A_t \cup P$, we know at least one member of P has a phone number with a first digit of j. At this point, we have determined all of the first phone number digits (e.g., we might determine that all users in P have phone numbers starting with 2 or 8).

Inferring further digits Now, we have a set of prefixes (e.g., 2 and 8 from above), and we wish to determine the next digit(s) for each prefix (e.g., 21, 26, or 87). The natural way to do so would be to create lists L for every possible prefix (e.g., all phone numbers starting with 20, ...), but this quickly becomes unmanageable as there are 10^d possible prefixes of length d. Instead, we effectively construct these lists using the exclude feature and the L_{ij} we already uploaded. Specifically, let us denote L^c_{ij} as the *complement* of L_{ij}—all phone numbers that do *not* have j as the ith digit—we can express it as $L^c_{ij} = \cup_{k \neq j} L_{ik}$. We can then exclude L^c_{ij} from our audience to focus only on the prefix of interest.

Let us suppose we have the prefix 8 and we wish to determine all following digits for phone numbers that start with 8. We can then determine an upper threshold set containing $P \setminus L^c_{18} = P \setminus (L_{10} \cup L_{11} ... L_{17} \cup L_{19})$; let us denote this $P \cup A_t \setminus L^c_{18}$. We can then exclude each L_{2j} from this as

well and see if the potential reach drops; if so, it indicates there is at least one user with the prefix $8j$. For example, if the potential reach of $P \cup A_t \setminus (L_{18}^c \cup L_{24})$ is lower than the potential reach of $P \cup A_t \setminus L_{18}^c$, then we know there is at least one user whose phone number starts with 84.

This approach naturally generalizes to longer prefixes by excluding the complement of each prefix digit. Suppose we have a prefix of 379 and we wish to determine the set of next digit(s). We would construct an upper threshold list containing $P \setminus (L_{13}^c \cup L_{27}^c \cup L_{39}^c)$, denoted $P \cup A_t \setminus (L_{13}^c \cup L_{27}^c \cup L_{39}^c)$. We can then additionally exclude each L_{4j} to determine if any of the phone numbers starting with 379 have j as the next digit.

Recursively inferring all phone numbers Given the methodology above, it is straightforward to infer all of the phone numbers in an audience. We first infer the first digits of phone numbers corresponding to members of P. Subsequently, in iteration k, we take each phone prefix inferred from iteration $k-1$ (which would be of length $k-1$), append to it a digit $d \in \{0, 1, \cdots, 9\}$ and check whether the resulting phone prefix of length k matches at least one member of P. If so, we add the prefix of length k to the list of prefixes inferred in iteration k. This terminates by outputting all the phone numbers of members of P for which Facebook has a phone number.

Evaluation We place a Facebook pixel to track visitors to a webpage that we created specifically for this experiment, and recruit 14 friends and family from the Boston area that have Facebook accounts, and have a phone beginning with area code either 617 or 857. The volunteers were asked to visit the webpage while logged in to their Facebook account. Using the methods from section V-A with the email addresses provided by the volunteers, we found that two volunteers had installed ad-blockers and were not in the pixel audience, that another volunteer was not a "daily active" user, and a final volunteer did not have a phone number associated with their Facebook account. Thus, we expect that we should be able to infer 10 phone numbers from our pixel audience.

Our method output nine phone numbers that correctly matched the phone numbers of nine out of the 10 remaining volunteers. For the 10th user, we found that user had provided multiple phone numbers to Facebook; we describe and evaluate a modified approach for how such users can be handled in the Appendix. Our method took around one hour to do the entire search and output the phone numbers by sending queries to Facebook in a serial fashion; this time could be cut down significantly by parallelizing the queries sent to Facebook.

D. Discussion

We now discuss a few issues our attacks bring up.

Cost We first briefly quantify the cost of launching our attacks, and discuss potential optimizations to reduce this cost. The costs are summarized in Table III.

Number and size of phone number lists: If we represented the enumerated phone numbers from section V-B using a base b different from 10, we would need $b\lceil \log_b N \rceil$ lists of size $\frac{N}{b}$ each, where N is the number of possible phone numbers. For example, to reduce the size of each list (e.g., to avoid detection), the attacker could use a base of 100, which would require 400 lists of 100K numbers each for a particular US area code (compared to 70 lists of 1M numbers each, as before). Though these numbers seem large and potentially easy to detect, the attacker could partition these lists across multiple accounts to avoid detection or rate limiting (see Section VI-A).

Number and size of padding lists: In addition to the lists of enumerated phone numbers, our attacks require the uploading of lists of padding records. For our email–phone linkage attack (Section V-B), we require up to 200 lists of padding records (containing fewer than 2,000 records each) to find a threshold list, and 200 more with the victim added to each of the previous 200 lists. Alternately, by first finding the threshold list in the first round of uploads, and then adding the victim only to the threshold list, we only need one list (as opposed to 200) with the victim added. Furthermore, by searching for a threshold list using two rounds of uploads, the number of lists can be reduced to 30, where in the first round we upload lists whose sizes increase in increments of 10 to find a range of 10 consecutive sizes within which the threshold list lies. In the second round, we find the actual threshold list from within the appropriate range.

For our attacks de-anonymizing particular web visitors (Section V-A) and de-anonymizing users en-masse (Section V-C), we need r lists containing $1 \cdots r$ users respectively (to find threshold lists), where r is the granularity of rounding applied to the website's victim audience size. For audiences of size greater than 1,000, the value of r can be in the hundreds or greater (see section IV-B). However, by working with subsets of the audience of less than 1,000 users, the value of r can be reduced to 10 and the lists can be re-used with different subsets. For example, to infer the phone numbers of a website audience with 8,000 members, it would be necessary to use around 100 padding lists to find a threshold list. However, if at a time we only attempt to learn any matching phone numbers that end with a particular digit (by excluding those that have phone numbers that end with other digits), the size of the reduced audience would be roughly a tenth of the original size ($\sim$800), which would be rounded to the nearest 10 and would only require 10 padding lists.

Time taken: For our email–phone linkage attack (Section V-B), there is a setup time of a few days to upload the phone number lists (only done once) and of a few hours to find a threshold audience (done once for every batch of victims); parallelization can reduce this time. Then, it typically only takes an additional hour to find the phone number of any victim.

For our attack de-anonymizing particular web visitors (Section V-A), once the phone number and padding lists are all uploaded (similar cost as above), we only need to upload a record with the victim's information (roughly 20 minutes). Then, we only need to run around r queries, taking a few minutes per victim.

Attack step	Num of lists uploaded	Size of each list	Queries	Time taken	Times performed
Upload phone lists	82 (France)	20M (France)	0	$<$ 1 week (France)	Once
	70 (US)	1M (US)	0	$<$ One day (US)	Once
Upload padding lists for V-A	r	Up to r	0	1-2 hours if $r = 100$	Once
Upload padding lists for V-C	r	Up to r	0	1-2 hours if $r = 100$	Once
Find threshold list for V-B	200 (one upload round)	$<$ 2,000	200	3–4 hours	Once per batch
	30 (two upload rounds)	$<$ 2,000	30	4–6 hours	Once per batch
V-A inference step	1 (victim's record)	1	~r	Up to 20 min	Once per victim
V-B inference step	1 (threshold list plus victim)	$<$ 2,000	One per phone list	Up to 1 hour	Once per victim
V-C inference step	0	–	$O(md(r+10))$	$O(md(r+10))$ seconds	Once per audience

TABLE III: The cost of performing various steps for the three attacks described in the paper.

For our en-masse de-anonymization attack (Section V-C), once the lists are all uploaded (similar cost as above), the worst-case requires $O(md(r+10))$ queries (where d is the number of digits in the phone number, m is the total number of matching phone numbers, and r is the rounding applied to the audience size as previously described); e.g., these would require around two hours when $m = 10$, $d = 7$, and $r = 100$ (at a query each second). However, this worst-case estimate assumes that phone numbers do not share any prefixes; when phone numbers share prefixes, the time taken for inferring a particular prefix is amortized across all the phone numbers that share the prefix. Hence, for larger audiences, the attack can potentially scale significantly sub-linearly. Also, as pointed out in Section V-C, by uploading the lists to multiple advertiser accounts and inferring the matching phone numbers in parallel, the time taken can go down proportionally.

Generality Any similar advertising platform would be incentivized to provide the functionality that Facebook's advertising platform provides (custom audiences, de-duplication, size statistics, etc.), as many of these are requested by the advertisers. Indeed, as described in Section II, other advertising platforms provide PII-based targeting, and provide similar size statistics. While Instagram, Google [38], and Twitter allow the use of both email addresses and phone numbers, Pinterest and LinkedIn allow the use of both emails and mobile advertiser IDs (identifiers for the user's mobile device). This suggests the potential for similar attacks on other platforms, but due to space constraints, we leave exploring them to future work. However, it is worth noting that most other platforms have larger minimum sizes for audiences; this alone does not significantly raise the difficulty of our attacks, as the attacker need only pad any uploaded audiences with additional records (e.g., voter records).

In order to check if other platforms are vulnerable, a similar process as presented in this paper can be followed: (i) determine the properties of size statistics (e.g., consistency, monotonicity, presence of noise), and whether the platform de-duplicates different kinds of PII; if so, (ii) collect padding lists with appropriate PII and size (based on the granularity of size statistics) and evaluate the attacks presented in Section V.

Other PII linkages When inferring a victim's PII (Section V-B), the adversary need not be limited to phone numbers; any PII that can be enumerated can be used by assigning a unique d-digit number as an index to each enumerated PII element, and applying the same attack to infer the matching d-digit index. For example, since Facebook's advertising platform allows audiences to be created using combinations of (First name, Last name, City, State) or (First name, Last name, Date of birth), the same attack could be used to infer the phone number of a victim knowing only their name and city, or their name and date of birth. Worse, the adversary could enumerate combinations of common names and potential dates of birth in the U.S., assign them indices, and then infer the name and birthdate of a victim given only their email address.

VI. Defenses

We now address how Facebook can fix the attacks.

A. Defense approaches and non-fixes

Recall that all of our attacks were enabled by two implementation characteristics from Section IV: (1) the rounding of audience size statistics, and (2) the de-duplication of users. Simply removing all size estimates is likely to be impractical, as this feature is likely useful to advertisers. Thus, Facebook must find a tradeoff between providing utility to advertisers and ensuring user privacy; using coarse-grained audience size estimates alone is insufficient, as we showed an adversary can construct audiences that fall on the rounding threshold. We first discuss potential defense approaches that raise the cost for the attacker, but do not prevent the attack.

Anomaly detection A platform could use anomaly detection (e.g., on the rate of queries, the size of lists, the contents of lists, etc.) to identify malicious advertisers. The main challenges with this approach are that the attacker could spread their queries out across multiple accounts, across time, and mix in other legitimate queries to evade detection. Moreover, the attacker need not upload easily-detectable lists (e.g., lists with all phone numbers that begin with a single digit); the attacker can easily make it more difficult to detect a pattern. Thus, as observed in other contexts (such as fake account detection), approaches based on anomaly detection struggle to provide guarantees of robustness.

Rate-limit API queries Alternatively, the platform could rate-limit API queries; while this will increase the time taken to conduct the attack, the attacker can again overcome the constraints by using multiple accounts. Thus, this raises the bar for attackers, but is insufficient for determined or powerful attackers, or for attackers who are only interested in targeting certain users (e.g., celebrities or dissidents).

Financial disincentives Alternatively, an advertising platform could employ financial disincentives by charging advertisers for every audience created or size estimate obtained. While such disincentives would increase the cost to the attacker, they may not be well-received by advertisers, who often create a large number of audiences and run many campaigns. Moreover, advertisers who wish to target particular victims would likely not be dissuaded by such costs.

Noisy size estimates While the addition of random noise would make the proposed attacks harder, an adversary could still circumvent that by making many repeated queries and performing statistical analysis on the results. For example, if Facebook adds uniform noise between 0 and $n-1$ (without rounding), then $\frac{log(\epsilon)}{2log(1-\frac{1}{n})}$ samples of each of two audiences are necessary to say whether the sizes come from the same distribution with probability $1-\epsilon$. If we set ϵ = 0.01 and $n = 20$, then we require only 45 samples of each audience size.

Applying differential privacy [7] might be robust against the described attacks. However, as pointed out by Korolova [18], deploying differential privacy is challenging in advertising platforms as the adversary can easily create multiple advertising accounts and issue many queries to the platform.

B. Secure mechanism that restricts de-duplication

Overall, there does not appear to be a simple and robust solution based on modifying how audience size statistics are calculated that would prevent our attacks while maintaining utility. Instead, we propose a solution based on changing the way Facebook de-duplicates users.

Model of the platform For the purpose of the defense we propose, we assume the advertising platform provides advertisers the following functionalities: (1) creating PII-based audiences by either uploading records or using tracking pixels, (2) obtaining the size of these PII-based audiences, and (3) combining these PII-based audiences using $\cup$, $\cap$, and $\setminus$, and obtaining the size of the resulting audience.

Proposed mechanism As is done on Facebook today (see Section III-A), the platform should define a set of *PII types* P_i, where each PII type is defined by a set of attributes that uniquely identify users. The PII types can contain a *single attribute* (e.g., {Email}), or *multiple attributes* (e.g., {Name, Date of Birth}). However, all PII types should be *minimal*, meaning if any attribute is removed, it is no longer a PII type (e.g., if {Name, Date of Birth} is a PII type, then {Name} cannot be a PII type).

Since tracking pixel audiences are not created by explicitly uploading records, we need to treat them differently: each tracking pixel audience is viewed as a unique PII type P_i.

We require that the platform have a *priority order* of PII types $\{P_1, P_2, \cdots\}$; it can be arbitrary, as long as it is static.

Advertisers can create custom audiences by uploading a list of records L, with each record containing an arbitrary number of attribute values. Note that we impose no constraint on what advertisers can upload; the security guarantees will come from how this upload is treated by the platform. Hence, a record might contain attributes that comprise zero, one, or several PII types and can potentially contain some attributes of a multi-attribute PII (without containing the entire multi-attribute PII).

The platform then turns the uploaded records into a custom audience, internally represented as a *set* of *reduced records*. For each uploaded record, the platform finds the smallest i such that all the attributes comprising P_i are present in the uploaded record; the platform then discards all other attributes in the record. This reduced record is added to the set representing the custom audience (with de-duplication). If no complete PII type from the ordered list is contained in the record (e.g., the record only contains a value for attribute Name which alone does not constitute a PII type), then the record is dropped.

Finally, the platform provides the size estimate of the custom audience created from L, denoted $s(L)$, by *matching* each of the reduced records against the user database. For each reduced record, if the attributes actually match an account, $s(L)$ is increased by 1. Thus, a user may be counted twice if they appear in the custom audience as two different reduced records (e.g., if the user's {Email} and {Phone} are in two separate reduced records).

The precise security guarantees provided by our solution are given in the following result.

Theorem 1. No size estimate $s(L)$ for any L can give any extra information about a PII type P_k of a user to an attacker who knows another PII type P_j of the user, and knows all existing values of P_k in the database. Formally, for any user u, any PII types P_j and P_k ($k \neq j$), any y such that there exists some user with $P_k = y$, and any list of records L,

$$Pr(P_k^{(u)} = y | P_j^{(u)} = x, s(L)) = Pr(P_k^{(u)} = y | P_j^{(u)} = x). \quad (1)$$

where $P_k^{(u)}$ denotes the value of PII type P_k for user u as per the platform's internal database.

Theorem 1 shows that our proposed method to compute size estimates $s(L)$ is secure against the attacks in this paper. Specifically, we consider an attacker that knows a PII type P_j about a victim u (i.e., knows that $P_j^{(u)} = x$), and we show that the result of $s(L)$ for any L gives no extra information about any other PII type P_k of the same individual. Theorem 1 follows directly, using Bayes rule, from the following lemma.

Lemma 1. Consider any user u, any PII types P_j and P_k ($k \neq j$) and any y such that there exists a user with $P_k = y$. For any list L, given that $P_j^{(u)} = x$, $s(L)$ is independent of whether $P_k^{(u)} = y$.

Proof. We prove that for any list L, the value of $s(L)$ is the same whether or not $P_k^{(u)} = y$.

First note that the result is trivial if L does not contain at least one record with $P_j = x$ and one record with $P_k = y$ (possibly the same record). Then we distinguish two cases.

Case 1: suppose that there is no record in L with both $P_j = x$ and $P_k = y$. Since we know that there exists a user with $P_k =$

y, those records will be counted independently of whether or not $P_k^{(u)} = y$. Then the result follows from the fact that records with different PII types are not de-duplicated, regardless of whether they belong to the same user.

Case 2: suppose that there is one record with $P_j = x$ and $P_k = y$. Then one of the PII types (or both if there are other higher priority PII types in the record) will be disregarded independently of whether they both correspond to the same user or not, and the resulting list falls in Case 1. □

Operations on the lists Until this point, we have only addressed the size statistics of a custom audience resulting from an uploaded list. We now discuss computing the size statistics of a combination of PII-based audiences, using $\cup$, $\cap$, and $\setminus$ operations. The definition of $s(...)$ for combinations follows the same principle as the definition of $s(L)$. Let L_1 and L_2 be two PII-based audiences. Then $s(L_1 \cup L_2)$, $s(L_1 \cap L_2)$, $s(L_1 \setminus L_2)$ are computed based on the corresponding operations on the reduced records that make up L_i (in the case of tracking pixel audiences, all records are single-attribute and are inherently reduced). As with $s(L)$, the final size statistic is then computed by matching each of the resulting reduced records after applying the operation against the user database, counting users multiple times if they are represented by two separate reduced records.

Then, the Theorem 1 holds on $s(L_1 \cup L_2)$, $s(L_1 \cap L_2)$, $s(L_1 \setminus L_2)$; this follows directly from the following lemma.

Lemma 2. Consider any user u, PII types P_j and P_k ($k \neq j$) and y such that there exists a user with $P_k = y$. For any lists L_1 and L_2, given that $P_j^{(u)} = x$, $s(L_1 \cup L_2)$, $s(L_1 \cap L_2)$ and $s(L_1 \setminus L_2)$ are independent of whether $P_k^{(u)} = y$.

Proof. The proof is similar to that of Lemma 1: we show that the size estimate of the audience resulting from $L_1 \cup L_2$, $L_1 \cap L_2$, or $L_1 \setminus L_2$ as defined above is the same whether or not $P_k^{(u)} = y$. First note that the result is trivial unless one of the lists contains at least one record with $P_j = x$ and the other at least one record with $P_k = y$. Then we distinguish two cases.

Case 1: If there is no record with both $P_j = x$ and $P_k = y$ in either list, then the result follows from the fact that records with different PII types are not de-duplicated, regardless of whether they belong to the same user.

Case 2: If there is one record with $P_j = x$ and $P_k = y$, then one of the PII types (or both if there are other higher priority PII types in the record) will be removed independently of whether they both correspond to the same user or not, and we are back to Case 1. □

C. Defense discussion

We now provide a brief discussion of our proposed defense.

Implementation Our proposed defense—replacing the existing matching algorithm by the one outlined above—is straightforward to implement. While we assumed that custom audiences were internally represented as sets of reduced records, this does not need to be the case to implement our solution; the platform could instead represent them as today, but then *tag* the users in the list with the record by which they were matched (to facilitate calculating the size statistics).

Accuracy of size estimates Our proposed method will overestimate the number of users in the audience if there are multiple records with different PII types corresponding to the same user; this is intentional as de-duplicating these users opens the door to our attacks. However, any defense must make a trade-off between utility to advertisers and privacy for users; our defense guarantees that the PII linkage attacks proposed in this paper will not be possible, while not significantly reducing the utility of size estimates to advertisers. In rounding size estimates (as they do today), Facebook already provides inaccurate statistics to advertisers; this shows that Facebook has decided that obfuscating the true value to provide privacy for users is acceptable. Given this, we believe that our defense strikes a better tradeoff than what Facebook already does. Our defense will not affect advertisers who have only one type of PII, or those who upload lists that do not actually contain multiple references to the same user.

Limitations We note that, as stated in the theorem, $s(L)$ gives no extra information *beyond* information about existence of an individual with a given value y of PII type P_k. Indeed, our mechanism does not prevent an attacker from discovering whether or not there *exists* an individual with $P_k = y$. This is important in particular for the multi-attribute PII case. Suppose for instance that the attacker knows that $P_j^{(u)} = x$ and that he knows parts of the attributes that constitute $P_k^{(u)}$ (a multi-attribute PII). Then the attacker could still potentially learn information on $P_k^{(u)}$ by finding out which of the existing PII type P_k are consistent with his background information. For example, if a user's name happened to be unique globally, the attacker could upload combinations of the {Name, Date of Birth}; the one that matched would reveal the user's date of birth. However, this is a separate attack from the ones in this paper, as it is an attack on the creation of custom audiences themselves; we hope to address it in future work.

Facebook's fix Facebook has acknowledged the vulnerabilities reported in this paper and has confirmed to us that it has deployed a stricter version of our defense: no size statistics will be provided for audiences created using multiple PII attributes. Additionally, no size estimates will be provided when combining audiences that were created using different PII attributes. This stricter defense uses our fix's core idea of preventing linking, but does so by providing no size statistics when multiple PII attributes are used; compared to our suggested defense, it may be easier to implement but provides lower utility to advertisers.

VII. Related work

Obtaining user information Due to the large amount of personal data that online social networking sites collect, they have been the vector for many privacy attacks over the

years. Krishnamurthy et al. [19] found that multiple pieces of users' personal information (name, city, zip code, email address, phone numbers, gender, birthday, age, employer, friends, activities, and interests) are either always available or available by default on most of the 12 OSN sites they examined, implying that potential privacy leakages on these sites can pose a significant issue. For example, the "Download Your Information" tool allowed users to inadvertently download the contact data for their friends, including email addresses and phone numbers, which they were not supposed to have access to [11]. Moreover, studies have developed methodologies to infer user attributes such as gender, political views, and religious views [22], [32], [15]; these use features such as homophily to make predictions about users' attributes.

Even worse, some social networking sites have been observed leaking private user information to third parties: for example, Facebook was observed leaking user IDs, usernames, and even personal details (e.g., name, gender, other profile data, friends, networks, wall posts, photos, and likes) to advertisers, typically via Request URIs and HTTP headers (e.g., Host, User-Agent, Referrer, and Cookie) [20], [9]. Even though Facebook has since fixed the issue [29], Facebook Messenger was observed leaking PII to third parties in 2016 [10].

Exploiting advertising services Korolova [18] demonstrated in 2011 that the microtargeting capabilities of Facebook's advertising system may allow advertisers to draw inferences about users who click on a Facebook ad, such as their age or sexual orientation. Using attribute-based targeting, she targeted the user so precisely that she created an audience of exactly one user;[6] she could then create multiple audiences with different values of the unknown attribute and determine which of the advertisements results in a click.

In comparison to our work, there are two main differences: first, Korolova's attack requires that there be a set of targeting attributes that *uniquely* (or almost uniquely) identify the user among all Facebook users (e.g., Male, lives in St. Louis, age 32, ...). Instead, we only need one piece of personal information (often easily available) sufficient to target the user via a custom audience (e.g., their email address). Second, our attacks show how user information beyond targeting attributes can be inferred (i.e., we show how users' phone numbers or names can be inferred from their email address). Third, Facebook mitigated Korolova's attack by imposing a minimum audience size of 20; our work shows that even with this mitigation, attacks are still possible.

Linking user identities The custom audience feature represents a linking mechanism between external data and Facebook's data. Researchers have examined the related problem of linking the accounts of a single user across multiple services [39], [17], [5], [21], [13], [12], [14]. For example, Vosecky et al. proposed an approach to link identities by exploiting user profile attributes, including name, date of birth, and email [39]. Balduzzi et al. [5] used email addresses to discover multiple identities of a user. However, as was recently measured by Goga et al. [14], only a few attributes like usernames and real names are available in OSN sites reliably, which makes such linkage attacks successful only for a small fraction of users. As a result, recent studies have moved beyond users' personal information and examined using textual and image content for linking [21], [13], [12], [14]. In our work, we are not proposing a new method of linking, but are instead examining the privacy implications of a linking mechanism provided by advertising platforms.

Custom audiences There has been surprisingly little academic study of custom audiences. The most recent related study by Minkus et al. [23] empirically examined how offline information (such as voter records) could be matched to public Facebook profiles, thereby enabling the inference of features such as the user's residential address, date and year of birth, and political affiliation. Tucker [33] investigated how users' perception of control over their personal information affects how likely they are to click on online advertising on Facebook, and found that giving users control over their private information can benefit advertising on Facebook. This implies that users want to control their own data used in online advertising; however, the current privacy settings [8] give users very few options. Even worse, users do not have control over their offline data, which can be used in the custom audiences feature. Our work is the first to study custom audiences directly and to point out the privacy leaks that occur via the custom audience interface.

VIII. Conclusion

The vast amounts of user data that social networking services have collected is now utilized by their advertising platforms to allow advertisers to target users via their PII. In this paper, we have shown how the inclusion of PII-based targeting opens up new privacy leaks in advertising platforms. By giving advertisers fine-grained control over the set of users targeted, and by providing them with coarse-grained statistics of audience sizes, the platforms open themselves to powerful attacks that can let an adversary learn private information about users. While we have proposed a solution to the attacks we uncovered, our work shows that platforms need to carefully audit their interfaces when introducing PII-based targeting.

Acknowledgments

We thank the anonymous reviewers and our shepherd, Carmela Troncoso, for their helpful comments. We also thank Paul Francis and Deepak Garg for their valuable feedback, Vincent Toubiana and Franck Baudot at CNIL for helping us bring the problem to Facebook's attention, and Facebook's security team for implementing the fix. This research was supported by NSF through grants CNS-1563320 and CNS-1616234, ANR through grants ANR-17-CE23-0014 and ANR-16-TERC-0012-01, Institut Mines Telecom through the "Future & Ruptures" program, the Alexander von Humboldt Foundation, and a Data Transparency Lab grant.

[6] At the time, Facebook did not have any minimum size requirement on audiences, thus enabling the attack.

ATTACHMENT #2

References

[1] About Customer Match. https://support.google.com/adwords/answer/6379332?hl=en.

[2] About Potential Reach. https://www.facebook.com/business/help/1665333080167380?helpref=faq_content.

[3] Account Targeting. https://business.linkedin.com/marketing-solutions/ad-targeting/account-targeting.

[4] Audience Targeting. https://help.pinterest.com/en/articles/targeting.

[5] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing Social Networks for Automated User Profiling. *RAID*, 2010.

[6] T. S. Bernard, T. Hsu, N. Perlroth, and R. Lieber. Equifax Says Cyberattack May Have Affected 143 Million in the U.S. https://www.nytimes.com/2017/09/07/business/equifax-cyberattack.html.

[7] C. Dwork. Differential Privacy. *ICALP*, 2006.

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

[9] Facebook Leaks Usernames, User IDs, and Personal Details to Advertisers. http://www.benedelman.org/news/052010-1.html.

[10] Facebook Messenger Chatbots Can Leak Your Private Information. https://www.techworm.net/2016/09/facebook-messenger-chatbots-can-leak-private-information.html.

[11] D. Guarini. Experts Say Facebook Leak of 6 Million Users' Data Might Be Bigger Than We Thought. http://www.huffingtonpost.com/2013/06/27/facebook-leak-data_n_3510100.html.

[12] O. Goga. Matching User Accounts Across Online Social Networks: Methods and Applications. Ph.D. Thesis, Pierre and Marie Curie University, 2014.

[13] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting Innocuous Activity for Correlating Users Across Sites. *WWW*, 2013.

[14] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi. On the Reliability of Profile Matching Across Large Online Social Networks. *ACM KDD*, 2015.

[15] N. Z. Gong and B. Liu. You Are Who You Know and How You Behave: Attribute Inference Attacks via Users' Social Friends and Behaviors. *USENIX Security*, 2016.

[16] How Google uses Customer Match data. https://support.google.com/adwords/answer/6334160.

[17] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying Users Across Social Tagging Systems. *AAAI ICWSM*, 2011.

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

[19] B. Krishnamurthy and C. E. Wills. On the Leakage of Personally Identifiable Information via Online Social Networks. *ACM SIGCOMM WOSN*, 2009.

[20] B. Krishnamurthy, K. Naryshkin, and C. E. Wills. Privacy leakage vs. Protection measures: the growing disconnect. *IEEE W2SP*, 2011.

[21] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan. HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling. *ACM SIGMOD*, 2014.

[22] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. *ACM WSDM*, 2010.

[23] 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.

[24] Marketing API: Custom Audience. https://developers.facebook.com/docs/marketing-api/reference/custom-audience.

[25] R. Nazarian. Facebook Loophole Exposes Private Phone Numbers, Here's How to Close It. https://www.digitaltrends.com/social-media/facebook-phone-number-hackers-flaw/.

[26] New Targeting Tools Make Pinterest Ads Even More Effective. https://business.pinterest.com/en/blog/new-targeting-tools-make-pinterest-ads-even-more-effective.

[27] N. Popper. Identity Thieves Hijack Cellphone Accounts to Go After Virtual Currency. https://www.nytimes.com/2017/08/21/business/dealbook/phone-hack-bitcoin-virtual-currency.html.

[28] T. Peterson. How Facebook's Custom Audiences Won Over Adland. http://adage.com/article/digital/facebook-s-custom-audiences-won-adland/297700/.

[29] Protecting Privacy with Referrers. Facebook Engineering's Notes. http://www.facebook.com/notes/facebook-engineering/protecting-privacy-with-referrers/392382738919.

[30] S. E. Rasmussen and J. C. Wong. Facebook Was Where Pakistan Could Debate Religion. Now it's a Tool to Punish 'blasphemers'. https://www.theguardian.com/technology/2017/jul/19/facebook-pakistan-blasphemy-laws-censorship.

[31] D. Storm. Hacker Says He Can Get Phone Numbers on Facebook Which Are Not Supposed to be Public. http://www.computerworld.com/article/3158109/security/hacker-says-he-can-get-phone-numbers-on-facebook-which-are-not-supposed-to-be-public.html.

[32] K. Thomas, C. Grier, and D. M. Nicol. Unfriendly: Multi-party Privacy Risks in Social Networks. *PETS*, Springer-Verlag, 2010.

[33] C. E. Tucker. Social Networks, Personalized Advertising, and Privacy Controls. *Journal of Marketing Research*, 2014.

[34] Tailored Audiences File Data. https://dev.twitter.com/ads/audiences/file-data.

[35] Target Facebook Ads to People on Your Contact List. https://www.facebook.com/business/a/custom-audiences.

[36] Target Custom Groups of Twitter Users. https://business.twitter.com/en/targeting/tailored-audiences.html.

[37] US Voter List Information. http://voterlist.electproject.org.

[38] Upload Data Files and Manage Your Customer Match Audiences. https://support.google.com/adwords/answer/6276125?hl=en.

[39] J. Vosecky, D. Hong, and V. Y. Shen. User Identification Across Multiple Social Networks. *NDT*, 2009.

[40] What Are Mobile Advertising IDs and When Should I Use Them with Custom Audiences? https://www.facebook.com/business/help/570474483033581.

[41] What Happens When I Upload My Customer List to Facebook? https://www.facebook.com/business/help/112061095610075.

[42] What's a Custom Audience from a Customer List? https://www.facebook.com/business/help/341425252616329/.

[43] What's the Difference Between Estimated Daily Reach and Potential Reach? https://www.facebook.com/business/help/1438142206453359.

Appendix A
Handling victims with multiple phone numbers

When describing our attacks in Section V, we found that they did not work as anticipated for users who had uploaded *multiple* phone numbers to Facebook. We now describe and evaluate extensions to our attacks that allow us to handle such users.

A. Extending email-phone inference attack

In Section V-B, we described how our email-phone number linkage attack fails when the victim has provided multiple phone numbers to Facebook. When this happens, we infer *multiple* values for each digit of the phone number, corresponding to the multiple phone numbers.

We first describe a solution that involves the same one-time preparatory steps as the attack described in Section V-B. We then briefly describe how, by conducting additional one-time preparatory steps, the attacker could significantly reduce the effort required per victim to disambiguate and infer the multiple matching phone numbers, thereby scaling up the attack.

Naïve method for disambiguation Let us assume that the attack in Section V-B infers multiple values for some digits of the phone number, and that the maximum number of values inferred for a digit is n (for simplicity, let us assume this occurs only at one digit i). This indicates that Facebook must have at least n phones corresponding to this user. For each value j inferred for the digit i, we perform the following steps:

1) Create the set L_{ij}^{poss} of all possible phone numbers with value j at that digit i, and whose other digits each have one of the values inferred for that respective digit.
2) Represent each phone number in L_{ij}^{poss} with a unique d-digit index (where $d = log_{10}(|L_{ij}^{poss}|)$).
3) Use the attack described in section V-B to link the given email to the d digit index corresponding to the matching phone number; this could be accomplished by dividing L_{ij}^{poss} into $10d$ subsets just as in Section V-B.

This method requires uploading $10d$ custom audiences for each phone to be inferred. However, the size of L_{ij}^{poss} is small compared to the number of all possible phones. For example, when $n = 3$, $|L_{ij}^{poss}| \leq 3^{10}$ for 10 digit phone numbers; hence, it would take less time to create these custom audiences compared to the L_{ij} created in Section V-B. In general, an attacker should be able to use this method to infer all matching phone numbers in a few hours. In case there are multiple matching phone numbers in one of the L_{ij}^{poss}, we can recursively apply the same method to further disambiguate those numbers.

Scaling up the disambiguation We now briefly describe how an attacker could speed up the disambiguation with some additional one-time effort at the beginning. One method to do this would be to upload sets of phone numbers $L_{i_1 j_1 i_2 j_2}$ whose i_1-th digit has a value j_1 and whose i_2-th digit has a value j_2, with one set corresponding to each possible combination of i_1,i_2,j_1, and j_2. Checking whether the victim user exists in $L_{i_1 j_1 i_2 j_2}$ allows the attacker to determine whether at least one of the matching phone numbers has a value j_1 for its i_1-th digit *and* has a value j_2 for its i_2-th digit. It is then trivial to do the disambiguation between the possible phone numbers, by appropriately checking whether the victim user exists in $L_{i_1 j_1 i_2 j_2}$, and to thereby infer all the matching phone numbers.

This method does not require the attacker to create any additional audiences when conducting the attack, allowing the attacker to infer all the victim's matching phone numbers in well under an hour. However, it requires the attacker to upload $100\binom{d}{2}$ sets upfront (one set corresponding to each possible combination of i_1,i_2,j_1, and j_2), where d is the total number of digits in the phone number. For example, a 10 digit phone number would require the attacker to upload 4,500 custom audiences, while a 8 digit phone number would require the attacker to upload 2,800 custom audiences. This would take a few days, and would have to be launched across multiple advertiser accounts (since each account allows the creation of at most 500 PII-based audiences); however, it would not be difficult for a determined attacker.

Evaluation We apply the naïve disambiguation technique to the two volunteers with multiple phones from V-B; we find we are correctly able to infer the matching phones for these two users (who have provided two and four phone numbers to Facebook, respectively).

B. Extending en-masse de-anonymization attack

In Section V-C, we described how our attack to de-anonymize members of a PII-based audience en-masse would fail to infer phone numbers of victims who provided multiple phone numbers to Facebook. We first explain why the en-masse de-anonymization attack fails to infer the matching phone numbers for users who have provided multiple phone numbers to Facebook. We then introduce notation that we use for our modified attack; and then describe the modified attack itself.

Explanation for failure of the attack We explain the failure of the method with an example: assume a user u who has provided two phone numbers to Facebook (`791 3444` and `799 5485`) visits our webpage and is added to the pixel audience. As part of its breadth-first search for matching phone numbers, our attack would first determine that there exists at least one matching phone number in the pixel audience starting with the prefix 79. In the next iteration, the breadth-first search would then construct an upper threshold list A_t such that $P \cup A_t \setminus (L_{17}^c \cup L_{29}^c)$ falls over the rounding threshold, and then additionally exclude L_{3j} to check whether any of the matching phone numbers beginning with 79 have j as the next digit. At the end of the iteration, the search would determine that there exists at least one matching phone number corresponding to each of the prefixes 791 and 799.

The following iteration is where the failure occurs. In order to determine digits following the prefix 791, the breadth-first search would first construct an upper threshold list such that $P \cup A_t \setminus (L_{17}^c \cup L_{29}^c \cup L_{31}^c)$ falls over the rounding threshold, and then additionally exclude L_{4j}. We see immediately that by excluding $(L_{17}^c \cup L_{29}^c \cup L_{31}^c)$ when determining the upper threshold list, we automatically exclude users whose phone numbers start with the prefix 799; since exclude trumps include, and since u has a phone number starting with 799, u is excluded from the list. As a result, we cannot infer any additional digits of this phone number for u; the same problem occurs when trying to infer the digit after the 799 prefix.

Notation As before, we denote the size of a PII-based audience A as $s(A)$. We also assume that indices begin from 1. We also use regular-expression-like notation to refer to a set of phone numbers, with the i-th term of a regular expression r denoting the set of possible values for the i-th digit of the phone numbers in the set (we denote this set of possible values by $r[i]$). For example the set of phone numbers $\{$`791 3444`, `799 5485`$\}$ is denoted by the regular expression $r = 79(1 \vee 9)(3 \vee 5)4(4 \vee 8)(4 \vee 5)$ where $r[3] = \{1, 9\}$ is the set of possible values for the third digits of the phone numbers in the set. Regular expressions are either empty (denoted by $\varnothing$) or of length k, corresponding to the first k digits of the matching phone numbers, where $k \in \{1, ..., d\}$ (d is the total number of digits in the phone number); an empty regular expression is assumed to match all possible phone numbers.

Given a regular expression r of length k that specifies the possible values for the first k digits of matching phone

numbers, and given a set S of possible values for the $(k+1)$-th digit, we can construct a regular expression $r' = r.S$ that is satisfied only by phone numbers whose first k digits satisfy r and whose $(k+1)$-th digit falls in S. Note that as a natural consequence of this definition, if S denotes the possible values for the first digit, $r = \varnothing.S$ is satisfied only by phone numbers that have first digits that fall in S.

In order to select phone numbers that satisfy a regular expression r from a PII-based audience P, we do the following: for each digit index i whose possible values are specified by r, we exclude $(\cup_{j \in r[i]} L_{ij})^c$ from P (this is alternately accomplished by excluding $(\cup_{j \notin r[i]} L_{ij})$ from P). We denote this previously described operation by $P \setminus r$; note that $P \setminus \varnothing = P$. If we want to additionally exclude some list L from $P \setminus r$, we denote this by $P \setminus (r, L)$

We now describe the attack do the en-masse deanonymization and infer all matching phone numbers when users have provided multiple phone numbers to Facebook.

Inferring all matching phone numbers The attack is described in Figure 3, and is conducted by calling the DEANONYMIZEPHONESWRAPPER function with the pixel audience P and the lists of phone numbers L_{ij} as parameters; it returns the set L_{inf} of all phone numbers of all (active) users in P. At a high level the attack has the following steps (the outer loop in the DEANONYMIZEPHONES function):

1) Determine a set of regular expressions, each of which are satisfied by the phone numbers of a minimal number of users in P, and which together cover as many matching phone numbers as possible (line 8 in DEANONYMIZEPHONES).
2) Disambiguate the phone numbers satisfying each regular expression (loop starting at line 9 in DEANONYMIZEPHONES).
3) Exclude the inferred phone numbers from P (by uploading a list with the inferred phone numbers and then excluding the created audience), and repeat the previous steps, until there are no more phone numbers to be inferred from P .

The final step is necessary because in some cases, certain phone numbers can "hide" other phone numbers from being inferred, given the way our algorithm works; by excluding inferred phone numbers from P and performing another round, we can infer these phone numbers that were hidden in the previous round. We first briefly explain how the attack determines a set of regular expressions in step 1, and then briefly explain how it disambiguates the phone numbers satisfying a regular expression in step 2.

Determining a set of regular expressions Note that for users with multiple matching phone numbers, it is not possible to directly determine all matching phone numbers just by using the phone number lists L_{ij} that we created. This is because of the constraint imposed by Facebook's platform that we can only exclude a disjunction of custom audiences (and not a conjunction). The best we can do is to learn all possible values of users' phone numbers at each digit for different users (similar to what we get from the email phone linkage attack for users with multiple phone numbers).

Therefore, we aim to determine a set of regular expressions, each of which is satisfied by a minimal number of users' phone numbers. For each such regular expression, inferring which phone numbers satisfying it are present in the pixel audience (i.e., performing step 2) would require us to repeat the de-anonymization attack focused on the set of phone numbers satisfying the regular expression. Since each of these regular expressions are satisfied by a small set of phone numbers, we would be able to perform step 2 and disambiguate the underlying phone numbers for each relatively quickly.

The procedure to compute these regular expressions (described in GETREGEXES function in Figure 4) works by searching the tree of all possible phone numbers, similar to the procedure described in Section V-C; the only difference being that it recursively computes regular expressions of matching phone numbers, rather than individual matching phone numbers. In iteration k, the procedure takes each regular expression r of length $k-1$, and the set of k-th digits of phone numbers that match r (denoted by $nextDigits[r]$), and then finds the minimal subsets of $nextDigits[r]$ for which we do not run into the problem that the original de-anonymization ran into (of one prefix of a user excluding another); these are then appended to r to find regular expressions of size k. In order to find the minimal subsets of $nextDigits[r]$ described above, we check for each subset of $nextDigits[r]$ if we are able to infer additional digits ($nextDigitSet$) at index $(k+1)$; if no, then we discard that subset, if yes, then save $nextDigitSet$ for use in the next iteration.[7]

Disambiguating phone numbers The procedure for disambiguating the phone numbers satisfying a regular expression (described in the DISAMBIGUATE function in Figure 3) is very similar to the naive disambiguation method proposed earlier in the Appendix to disambiguate phone numbers matching an email address. We find the index i_{max} of the digit for which the regular expression allows the maximum number of possible values; for each possible value in $r[i_{max}]$, we fix that value, enumerate all possible phone numbers which have that fixed value at i_{max} and which satisfy the regular expression, and then recursively call DEANONYMIZEPHONES to learn all the phone numbers within this enumerated set of users that are part of the pixel audience.

Discussion We first note that the de-anonymization attack is performed in multiple rounds of inferring regular expressions, and then disambiguating phone numbers. Phone numbers of users with only one phone number are guaranteed to be discovered in the very first round, and will be represented by regular expressions r that match them uniquely (i.e., $r[i]$ is a singleton set for all i). Also, note that while the attack has been

[7]As a consequence of this procedure, the GETREGEXES function produces regular expressions that have the following properties: (i) For any user, they are satisfied by all his phone numbers, or none of his phone numbers, and (ii) Each regular expression of size k covers the minimal number of users possible for a regular expression of size k (for $k \in \{1, ...d-1\}$)

```
function DeAnonymizePhonesWrapper(P, [L_ij])                    ▷ Infer all matching phone numbers in P
    L_inf ← DeAnonymizePhones(P, {}, [L_ij])
    return L_inf
end function
function DeAnonymizePhones(P, L_inf, [L_ij])                    ▷ Infer all matching phone numbers in P \ L_inf
    repeat
        L_new ← {}
        rSet ← GetRegexes(P, L_inf, [L_ij])
        for r ∈ rSet do
            L_disambiguated ← Disambiguate(r, P, L_inf, [L_ij])
            L_new ← L_new ∪ L_disambiguated
        end for
        L_inf ← L_inf ∪ L_new
    until |L_new| = 0
    return L_inf
end function
function Disambiguate(r, P, L_inf, [L_ij])                      ▷ Disambiguate all phone numbers in P \ L_inf that satisfy r
    L ← ∪_{i,j} L_ij
    Find i_max such that |r[i_max]| is maximum
    For v ∈ r[i_max] create L^v (list of phones in L that satisfy r and have value v at index i_max)
    Assign each phone in L^v a unique index
    Partition L^v into L^v_ij (where L^v_ij is the list of phones in L^v whose index has i-th digit j)
    L_disambiguated ← DeAnonymizePhones(P, L_inf, [L^v_ij])
    return L_disambiguated
end function
```

Fig. 3: Algorithm to de-anonymize phone numbers when users have multiple phones.

described assuming that the PII being inferred is the phone number, as mentioned previously in the paper, the same attack can be used for inferring other PII (by enumerating possible values for the PII and by inferring the index of matching users in the enumerated set).

Evaluation We apply the extended version of the en-masse de-anonymization attack on the audience containing 14 volunteers from Boston from Section V-C. Recall that in Section V-C we were already able to infer the phone numbers for nine users with single phone numbers, and found one user in the audience who had provided multiple phone numbers to Facebook. The modified attack outputs 11 phone numbers, nine of which correctly match the nine users with single phone numbers, and the remaining two of which correctly match the user who had provided multiple phones. Our method took around two hours to do the entire search and output the phone numbers by sending queries to Facebook in a serial fashion; as mentioned in Section V-C, this time could be cut likely down significantly by parallelizing the queries sent to Facebook.

```
function GetRegexes(P, L_inf, [L_ij])                ▷ Get regular expressions of user's phone numbers in P \ L_inf
    rSet ← {∅}
    nextDigits[∅] ← {0, 1, ..., 9}
    for i ∈ [1, ..., d − 1] do
        rSet_curr ← {}
        for r ∈ rSet do
            rSet_new, nextDigits_curr ← UpdateRegexes(r, nextDigits[r], i + 1, P, L_inf, [L_ij])
            rSet_curr ← rSet_curr ∪ rSet_new
            for r′ ∈ nextDigits_curr do
                nextDigits[r′] ← nextDigits_curr[r′]
            end for
        end for
        if i < d − 1 then
            rSet ← rSet_curr
        else
            rSet ← {r.nextDigits[r] : r ∈ rSet_curr}
        end if
    end for
    return rSet
end function
function UpdateRegexes(r, digits, i, P, L_inf, [L_ij])
    rSet ← {}, nextDigits ← {}
    Create a list D containing subsets of digits in increasing order of size
    D_selected ← {}
    index ← 1
    while index ≤ len(D) do
        digits_curr ← D[index]
        If digits_curr ⊃ digits_selected for some digits_selected ∈ D_selected, increment index and skip to next iteration
        Find upper threshold audience such that s(P ∪ A_t \ (L_inf, r.digits_curr)) falls just over rounding threshold τ
        nextDigitSet ← {j : s(A_t ∪ P \ (L_inf, r.digits_curr, L_ij)) ≤ τ}
        if |nextDigitSet| > 0 then
            rSet ← rSet ∪ {r.digits_curr}
            nextDigits[r.digits_curr] ← nextDigitSet
            D_selected ← D_selected ∪ digits_curr
        end if
        index ← index + 1
    end while
    return rSet, nextDigits
end function
```

Fig. 4: Helper function that finds regular expressions matching the phone numbers of users in audience.

ATTACHMENT #3

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Potential for Discrimination in Online Targeted Advertising

Till Speicher
MPI-SWS

Muhammad Ali
MPI-SWS

Giridhari Venkatadri
Northeastern University

Filipe Nunes Ribeiro
UFOP, UFMG

George Arvanitakis
MPI-SWS

Fabrício Benevenuto
UFMG

Krishna P. Gummadi
MPI-SWS

Patrick Loiseau
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG and *MPI-SWS*

Alan Mislove
Northeastern University

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recently, online targeted advertising platforms like Facebook have been criticized for allowing advertisers to discriminate against users belonging to sensitive groups, i.e., to exclude users belonging to a certain race or gender from receiving their ads. Such criticisms have led, for instance, Facebook to disallow the use of attributes such as ethnic affinity from being used by advertisers when targeting ads related to housing or employment or financial services. In this paper, we show that such measures are far from sufficient and that the problem of discrimination in targeted advertising is much more pernicious. We argue that discrimination measures should be based on the targeted population and not on the attributes used for targeting. We systematically investigate the different targeting methods offered by Facebook for their ability to enable discriminatory advertising. We show that a malicious advertiser can create highly discriminatory ads without using sensitive attributes. Our findings call for exploring fundamentally new methods for mitigating discrimination in online targeted advertising.

Keywords: Discrimination, advertising, Facebook

1. Introduction

Much recent work has focused on detecting instances of discrimination in online services ranging from discriminatory pricing on e-commerce and travel sites like Staples (Mikians et al., 2012) and Hotels.com (Hannák et al., 2014) to discriminatory prioritization of service requests and offerings from certain users over others in crowdsourcing and social networking sites like TaskRabbit (Hannák et al., 2017). In this paper, we focus on the potential for discrimination in *online advertising*, which underpins much of the Internet's economy. Specifically, we focus on *targeted advertising*, where ads are shown only to a subset of users that have attributes (features) selected

© 2018 T. Speicher *et al.*

by the advertiser. Targeted ads stand in contrast to non-targeted ads, such as banner ads on websites, that are shown to all users of the sites, independent of their attributes.

The targeted advertising ecosystem comprises of (i) *advertisers*, who decide which users an ad should (not) be shown to; (ii) *ad platforms*, such as Google and Facebook, that aggregate data about their users and make it available to advertisers for targeting; and (iii) *users* of ad platforms that are consumers of the ads. The potential for discrimination in targeted advertising arises from the ability of an advertiser to use the extensive personal (demographic, behavioral, and interests) data that ad platforms gather about their users to target their ads. An intentionally malicious—or unintentionally ignorant—advertiser could leverage such data to preferentially target (i.e., include or exclude from targeting) users belonging to certain sensitive social groups (e.g., minority race, religion, or sexual orientation).

Recently, the Facebook ad platform was the target of intense media scrutiny (Angwin and Parris Jr., 2016) and a civil rights lawsuit for allowing advertisers to target ads with an attribute named "ethnic affinity." After clarifying that a user's "ethnic affinity" does not represent the user's ethnicity, but rather represents how interested the user is in content related to different ethnic communities, Facebook agreed to not allow ads related to housing, employment, and financial services be targeted using the attribute (Facebook, 2017) and renamed it to "multicultural affinity."[1]

In this paper, we conduct a systematic study of the potential for discriminatory advertising on the Facebook advertisement platform. We focus on Facebook because it is one of the largest online advertising platforms in terms of number of users reached by ads, the number of advertisers, and the amount of personal data gathered about the users that is made available to advertisers. Furthermore, Facebook is an innovator in introducing new methods for targeting users, such as *custom audience*[2] and *look-alike audience*[3] targeting that are then subsequently adopted by other online social media and social networking platforms like Twitter,[4] Pinterest,[5] LinkedIn,[6] and YouTube.[7] Thus, many of our findings may also be applicable to these other online ad targeting platforms as well.

Our study here is driven by the following high-level question: *What are all the different ways in which a Facebook advertiser, out of malice or ignorance, can target users in a discriminatory manner (i.e., include or exclude users based on their sensitive attributes like race)?*

To answer this question, we begin by proposing an intuitive measure to quantify discrimination in targeted ads. We then systematically investigate three different targeting methods (attribute-based targeting, PII-based targeting, and look-alike audience targeting) offered by Facebook for their ability to enable discriminatory advertising. At a high-level, we find that all three methods enable advertisers to run highly discriminatory ads. Worse, we show that the existing solution approaches of banning the use of certain attributes like "ethnic affinity" in targeting is not only inadequate, but does not even apply in two out of the three ad targeting methods.

While our findings primarily serve to demonstrate the perniciousness of the problem of discriminatory advertising in today's ad platforms, it also lays the foundations for solving (i.e., detecting and mitigating) ad discrimination.

2. Quantifying Ad Discrimination

Next, we begin by outlining different ad targeting methods offered by Facebook. We then discuss the current approach to determining whether an ad is discriminatory and argue why it is inade-

1. Unfortunately, Facebook was found half a year later to still accept discriminatory ads, despite the fixes it claims were put in place (Angwin et al., 2017a).

2. https://www.facebook.com/business/help/170456843145568
3. https://www.facebook.com/business/help/164749007013531
4. https://business.twitter.com/en/targeting/tailored-audiences.html
5. https://business.pinterest.com/en/blog/new-targeting-tools-make-pinterest-ads-even-more-effective
6. https://business.linkedin.com/marketing-solutions/ad-targeting/matched-audiences
7. https://support.google.com/youtube/answer/2454017

quate. Finally, we propose a new and intuitive approach to quantify discrimination.

2.1. Methods for targeted ads

Facebook gathers and infers several hundreds of attributes for all of its users, covering their demographical, behavioral, and interest features[8] (Andreou et al., 2018). Some of those attributes, such as gender or race, are considered *sensitive* meaning targeting (i.e., including or excluding) people based on those attributes is restricted by law for certain types of advertisements (e.g., those announcing access to housing or employment or financial services (Barocas and Selbst, 2016)).

Facebook allows advertisers to select their target audience in three ways:

1. *Attribute-based targeting:* Advertisers can select audiences that have (or do not have) a certain attribute (or a combination of attributes), e.g., select users who are "men", "aged 35", and are interested in "tennis."

2. *PII-based (custom audience) targeting:* Advertisers can directly specify who should be targeted by providing a list of personally identifiable information (PII) such as phone numbers or email addresses.

3. *Look-alike audience targeting:* Advertisers can ask Facebook to target users who are similar to (i.e., "look like") their existing set of customers, specified using their PII.

2.2. Quantification approaches

Next, we discuss three basic approaches to quantifying discrimination and their trade-offs.

1. Based on advertiser's intent: An intuitive (moralized) way to quantify discrimination would be to base it on the advertiser's intent. However, not only is such a measure challenging to operationalize (i.e., to measure from empirical observations), but it also overlooks the harmful effects of unintentionally discriminatory ads that may be placed by a well-meaning but ignorant or careless advertiser. In this paper, we do not consider such approaches.

2. Based on ad targeting process: Another approach to determine whether an ad is discriminatory is based on the *process* used to target the ads. Any ads placed using the right process would be non-discriminatory (by definition), while those using a wrong process would be declared discriminatory (by definition). Existing approaches, such as those that determine whether an ad is discriminatory based on the use of sensitive attributes (e.g., "ethnic affinity") in targeting, fall under this category. As we show in this paper, attempting to quantify discrimination based on the process (means or methods) of targeting is quite difficult when there exist multiple different processes for targeting users. Instead, in this work, we advocate for a third approach.

3. Based on targeted audience (outcomes): We propose to quantify discrimination based on the outcomes of the ad targeting process, i.e., the audience selected for targeting. Put differently, we do not take into account how users are being targeted but only who they are. Outcome-based approaches to quantifying discrimination have the advantage that they can be generally applied to all scenarios, independently of the employed method of targeting. We discuss one such method in the next section.

2.3. Outcome-based discrimination

To formalize our discrimination measure, we will assume that an ad platform like Facebook keeps track of a database $\mathbf{D} = (u_i)_{i=1,\ldots,n}$ of user-records u_i where each user is represented by a vector of boolean attributes, i.e., $u_i \in \mathbb{B}^m$. We denote the sensitive attribute (e.g., race or gender) that we are interested in a particular situation by $s \in \{1, \ldots, m\}$ and its value for a user u by u_s. The corresponding sensitive group $\mathbf{S}$ is the set of all users that have the sensitive attribute, i.e., $\mathbf{S} = \{u \in \mathbf{D} \,|\, u_s = 1\}$.

To measure outcome- (i.e., targeted audience-) based discrimination, we define a metric for how discriminatory an advertiser's targeting is. It is inspired by the *disparate impact* measure that is frequently used to detect discrimination in selecting candidates from a pool of applicants in recruiting and housing allotment scenarios (Barocas and Selbst, 2016).

Our key observation is that ad targeting, like recruiting, involves selecting the *target audience* (**TA**) from a much larger pool of *relevant audience* (**RA**). The relevant audience of an ad is

8. https://www.facebook.com/business/learn/facebook-ads-choose-audience

the set of *all* users in the database $\mathbf{D}$ who would find the ad useful and interesting and thus might interact with it. Intuitively, the discrimination measure should capture the extent to which the target audience selection is *biased* based on sensitive group membership of relevant users.

We define the *representation ratio* measure for sensitive attribute s to capture how much more likely a relevant user u is to be targeted when having the sensitive attribute compared to not having it. More specifically, it is the ratio between the fraction of relevant audience with attribute s that are selected for targeting and the fraction of relevant audience without attribute s that are selected, i.e.,

$$\text{rep_ratio}_s(\mathbf{TA}, \mathbf{RA}) = \frac{|\mathbf{TA} \cap \mathbf{RA}_s| / |\mathbf{RA}_s|}{|\mathbf{TA} \cap \mathbf{RA}_{\neg s}| / |\mathbf{RA}_{\neg s}|}, \quad (1)$$

where $\mathbf{RA}_s = \{u \in \mathbf{RA} \mid u_s = 1\}$ and $\mathbf{RA}_{\neg s} = \{u \in \mathbf{RA} \mid u_s = 0\}$.

Based on the representation ratio we define a measure that we call *disparity in targeting*, defined for a sensitive attribute s as:

$$\text{disparity}_s(\mathbf{TA}, \mathbf{RA}) = \max\left(\text{rep_ratio}_s(\mathbf{TA}, \mathbf{RA}), \frac{1}{\text{rep_ratio}_s(\mathbf{TA}, \mathbf{RA})}\right). \quad (2)$$

Note that it is important to compute disparity based on the relevant audience $\mathbf{RA}$ because $\mathbf{RA}$ may have a very different composition in terms of attribute s than the whole database $\mathbf{D}$. For example, an ad for men's clothes may have a relevant audience $\mathbf{RA}$ with a gender-ratio highly skewed towards men. A random selection of users from $\mathbf{RA}$ would be non-disparate with respect to $\mathbf{RA}$, but might be highly disparate with respect to $\mathbf{D}$. Similarly, for the same targeted audience (including mostly males), some ads could be non-discriminatory (e.g., ads for men's clothes) while others could be highly discriminatory (e.g., ads for high-paying jobs), depending on the corresponding relevant audience. Throughout the paper, we implicitly assume that, for the sensitive attributes considered, the relevant audience has the same distribution as the global population; and we show that the advertiser can include or exclude certains groups based on the sensitive attribute—hence the ad targeting is discriminatory.

We propose to detect discriminatory targeting using our disparity measure as follows: we declare a targeting formula as discriminatory when its disparity for some sensitive attribute value group exceeds a certain threshold (i.e., the group is over- or under-represented). For instance, a reasonable threshold value may be 1.25, mimicking the popular "80%" disparate impact rule (Biddle, 2005), to declare a group over- or under-represented.

In addition to disparity, we would be interested in the recall of an ad, which quantifies how many of the relevant users with the sensitive attribute the discriminatory ad targets or excludes. It can be defined as

$$\text{recall}(\mathbf{TA}, \mathbf{RA}') = \frac{|\mathbf{TA} \cap \mathbf{RA}'|}{|\mathbf{RA}'|}, \quad (3)$$

where $\mathbf{RA}'$ might be the restriction of $\mathbf{RA}$ to $\mathbf{RA}_s$ or $\mathbf{RA}_{\neg s}$, depending on whether the discriminatory advertiser wants to target or exclude users with the sensitive attribute s.

3. PII-based Targeting

In this section, we show how the audience targeting mechanism based on personally identifiable information (PII) recently introduced by Facebook can be exploited by advertisers to covertly implement discriminatory advertising. We first briefly describe the PII-based audience targeting feature of Facebook; we then explain how this feature can be exploited to implement discriminatory advertising. Next, we explain how public data sources have data that advertisers can use to implement discriminatory advertising, and finally demonstrate the feasibility of such an attack by using information from public records to create audiences for advertising that are discriminatory.

PII-based audience selection: While Facebook traditionally allowed advertisers to select audiences to advertise to by specifying attributes of the audience (e.g., age, gender, etc.), Facebook recently introduced *custom audiences*. This feature allows advertisers to specify *exactly* which users they want to target by specifying personally identifying information (PII) that uniquely identifies those users. Facebook allows 15 different types of PII to be used, including phone numbers, email addresses, and combinations of name with other attributes (such as date of birth or ZIP code). The advertiser uploads a file containing a

list of PII; Facebook then matches these PII to Facebook accounts to create a custom audience.

Custom audiences can be viewed as implementing a *linking* function that allows advertisers to link the large amounts of external personal data available today with Facebook's user information. The linking function that custom audiences provide to advertisers is not a one-to-one function (i.e., advertisers cannot determine the exact Facebook account of a given person), but rather, it is an *aggregate function* that maps PII to a group of Facebook users. In the next section, we show that, despite this limitation, custom audiences can be abused to covertly implement discriminatory advertising by exploiting external data to create lists that selectively include only people with the sensitive attribute.

3.1. Potential for discrimination

To implement discriminatory advertising using custom audiences, an advertiser could simply create a list of PII corresponding selectively to people who have the sensitive attribute, uploading this list of PII to create a custom audience, and then advertising to that custom audience. Since the advertiser does not upload the sensitive attribute (instead uploading only a list of PII), and since the advertising platform itself may not have the sensitive user attribute, such targeting becomes difficult to detect.

Most advertisers already possess significant amounts of customer information (e.g., customer data, information from data brokers); however, even if they do not have such data, there are many other sources of data—including public records, data brokers, and web data—that can be accessed for free or at low cost. We next describe public sources of data from which one can get sensitive attributes for large sets of people; we then demonstrate how these data sources can be used in combination with custom audiences to implement discriminatory advertising.

3.2. Public data sources

An increasing amount of information about people is publicly available; we now briefly discuss how advertisers could obtain large amounts of external personal information.

Race, age, and gender Most U.S. states release voter records that contain the personal information of all registered voters (names, phone numbers, addresses, etc) along with other sensitive attributes such as race, age, and gender. For example, date of birth and gender are available in the records released by 38 and 34 states, respectively (Minkus et al., 2015); race information is available in the records of eight states (North Carolina, New Mexico, Louisiana, Tennessee, Alabama, Georgia, Florida, and South Carolina) (Ansolabehere and Hersh, 2013). Even when the race or gender is not available they can often be predicted with reasonable precision from other attributes (Mislove et al., 2011). For example, Tang et al. (Tang et al., 2011) propose a technique to infer the gender of a person from their name with an accuracy of 96.3% while covering more than 95% of users. Other companies such as Catalist[9] aggregate voter records from states and infer missing values of gender (from the first name) and race (from the name and address); the resulting race attributes matched voters' self-reported race 91% of the time (Ansolabehere and Hersh, 2013).

Criminal history People with criminal records—even those who have completed their sentence—are often victims of discrimination. We quickly survey the U.S. and find that more than 40 states in the U.S. make criminal records available online, and that 18 states offer free access to their state-wide criminal record databases; these records often contain significant amounts of personal information such as name, race, gender, and date of birth, along with the specific criminal record. Thus, advertisers can easily create custom audiences consisting only of users in this vulnerable population.

3.3. Discriminatory audience creation

We briefly demonstrate how it is possible to create discriminatory custom audiences on today's advertising platforms. Note that we *did not* actually advertise to these users or affect them in any way. Rather, our goal here is simply to demonstrate that using only public sources of data, advertisers can target protected classes and vulnerable populations with little effort.

9. https://www.catalist.us

Table 1: Results from experiment creating custom audiences using only users with certain attributes from the North Carolina voter records. For each sensitive attribute, we created and uploaded a custom audience of 10K random voters with that attribute. Shown is the total number of records per attribute, the number of Facebook users in the resulting *Targetable* custom audience, and the percentage of *Targetable* users who *match the sensitive attribute* as per Facebook's estimates.

	Voter Records		**Facebook Users**		**Validation of Custom Audience**
Attribute	Number	Percent	Targetable	Targetable %	% matching sensitive attribute
Male	3,438,620	45.5%	6,500	65%	81.5%
Female	3,995,533	52.8%	7,000	70%	91.4%
White	5,303,383	70.1%	6,800	68%	83.8%
Black	1,694,220	22.4%	6,300	63%	82.5%
Asian	79,250	1.0%	6,600	66%	28.8%
Hispanic	163,236	2.2%	5,900	59%	50.8%
Age (18-34)	1,985,117	26.2%	7,100	71%	80.3%
Age (35-54)	2,496,648	33.0%	6,900	69%	79.7%
Age (55+)	3,068,745	40.6%	5,700	57%	61.4%

We downloaded the public voter records from North Carolina,[10] giving us 7.5M records. Using data from the voter records, we then created custom audiences on Facebook for each sensitive attribute, selecting a random subset of 10K users from the voter file with each attribute. For example, we created a custom audience of women by uploading a list of 10K voters listed as female; we created a custom audience of white users by uploading a list of 10K voters listed as white. We created these custom audiences by uploading records containing the following fields: last name, first name, city, state, zip code, phone number, and country.

We then examine how many of these records match to Facebook accounts that can be targeted with advertisements, and then evaluate whether the created audiences are indeed discriminatory. Whenever we target an audience (either based on attributes, or by specifying a custom audience), Facebook provides an estimate of the number of users in the audience who can be targeted with advertisements; this estimate is called the *potential reach*.[11] We first target only the custom audiences created, without any additional targeting attributes specified, and use the potential reach estimate to measure how many records in the audience are *Targetable*.

Finally, in order to validate that advertisements targeted to these custom audiences would indeed be discriminatory, we take each custom audience and then target users with the corresponding sensitive attribute (e.g., for the male voter records audience, we target the Male attribute); we then measure the potential reach, and use the potential reach to measure what percentage of the *Targetable* users in the audience actually have that sensitive attribute (according to Facebook). Ideally, the percentage of *Targetable* users with the sensitive attribute would be 100%; however, Facebook may not know the attributes of some users, may have errors in their matching algorithm, or there may be errors in the user-provided data, making this percentage smaller.

It is important to note that definitions in our data sources (the voter file and census data) do not always line up with the targeting options that Facebook presents. For race, Facebook does not provide race directly but instead provides "ethnic affinity"; this is the same targeting parameter by which Facebook was accused of allowing discriminatory advertising (Angwin and Parris Jr., 2016).

Results The results of this experiment are shown in Table 1, and we make a number of interesting observations. *First*, the fraction of voter records that are *Targetable* (*i.e.*, online on a daily basis) is both significantly high (over 65% for most audiences we create) and fairly consistent across custom audiences. The only notable

10. http://dl.ncsbe.gov/index.html?prefix=data/
11. Facebook previously defined the potential reach as "the number of daily active people on Facebook that match the audience you defined through your audience targeting selections."

outliers are the Age (55+) audience, with only 57% matching.

Second, we observe that the fraction of the *Targetable* audience that matches the sensitive attribute, although it varies fairly widely across the different sensitive attributes, is consistently much higher than the fraction of the general adult population that has those sensitive attributes (assuming the voter records to be representative of the general adult population). In particular, for many sensitive attributes including gender, most races, and all ages, the percentage of *Targetable* audience that matches the sensitive attribute is higher than 80%. We suspect that the reason this fraction is low for the Asian attribute is due to the fact that race is an attribute that users typically do not upload to Facebook directly; however, we leave determining the source of this inconsistency to future work. We also note that even for these cases, the fraction of the *Targetable* audience that matches the sensitive attribute is significantly higher than the fraction of the voter records with the sensitive attribute. Taken together, our results show that advertisers can exploit public records to easily target discriminatory advertisements to a large number of people.

3.4. Summary

We explored the inherent risks that custom audiences induce for end users by allowing the linking of external information with Facebook's user data. We demonstrated the ease with which malicious advertisers could leverage the custom audience feature now present on advertising platforms like Facebook to implement discriminatory advertising. In fact, the wide variety of sources of public data available today means that even if an advertiser does not possess customer records of its own, it can easily find data sources to feed into custom audience creation.

4. Attribute-based Targeting

In this section, we examine how Facebook's attribute-based targeting mechanism can be used to launch discriminatory ads. First, we briefly explain how attribute-based targeting works and then examine the potential for abusing it.

Attribute-based audience selection: In brief, attribute-based targeting refers to the process of selecting an ad audience by specifying that recipients need to have a certain attribute or a combination of attributes; this is the traditional way of targeting ads on Facebook. For each user in the US, Facebook tracks a list of over 1,100 binary attributes spanning demographic, behavioral and interest categories that we refer to as *curated attributes*. Additionally, Facebook tracks users' interests in entities such as websites, apps, and services as well as topics ranging from food preferences (e.g., pizza) to niche interests (e.g., space exploration). We refer to these as *free-form attributes*, as they number at least in hundreds of thousands. It is unclear how exactly Facebook infers these attributes, but from their own description[12] this information can be gathered in many different ways such as user activity on Facebook pages, apps and services, check-ins with Facebook, and accesses to external webpages that use Facebook ad technologies. Beyond specifying a target region, language, age and gender for their ad, advertisers can choose that an ad should be shown to people that have some of these curated or free-form attributes turned on or off.

4.1. Potential for discrimination

The potential for discrimination on the Facebook ad platform was first publicly highlighted when researchers discovered the ability to exclude people based on their "ethnic affinity" (a curated attribute) when targeting ads related to housing (Angwin and Parris Jr., 2016). Facebook responded by banning the use of ethnic affinity attribute for certain types of ads (Facebook, 2017). More recently, researchers discovered the ability to target people interested in or holding anti-semitic viewpoints via free-form attributes like "jew haters" (Angwin et al., 2017b).

These findings raise several questions about the potential for discriminatory targeting using Facebook's curated and free-form attributes. First, given that ethnic affinity-based targeting was disallowed for its potential correlation with ethnicity (race) of users, are there other demographic, behavioral, or interest attributes that are similarly correlated, if not more? Second, given that there exist hundreds of thousands of free-form attributes, can malicious advertisers

12. https://www.facebook.com/ads/about/?entry_product=ad_preferences

Table 2: Most inclusive and exclusive curated attributes for each race. In parentheses are the recall and representation ratio for a population from North Carolina. These were obtained by uploading voter records filtered to contain only a single race, and then measuring the size of the subaudience targeted by each attribute. Attributes present in less than 5% of the population are not considered.

Race	Most inclusive	Most exclusive
Asian	US Politics: Liberal (8%, 2.76) Frequent travelers (15%, 2.70) Interest: Vegetarianism (7%, 2.23)	US Politics: Very Conservative (14%, 0.30) African American affinity (17%, 0.41) Interest: Country music (20%, 0.48)
Black	African American affinity (17%, 7.06) US Politics: Very Liberal (12%, 6.44) Interest: Online games (9%, 4.91)	US Politics: Very Conservative (14%, 0.18) US Politics: Conservative (17%, 0.22) Interest: Mountain biking (6%, 0.35)
Indian	Interest: Motorcycles (7%, 2.08) Interest: Online games (9%, 2.04) Interest: Ecotourism (6%, 1.96)	US Politics: Very Conservative (14%, 0.50) Away from hometown (22%, 0.51) Primary OS Mac OS X (7%, 0.56)
White	US Politics: Very Conservative (14%, 5.19) US Politics: Conservative (17%, 3.77) Interest: Hiking (11%, 2.27)	African American affinity (17%, 0.15) US Politics: Very Liberal (12%, 0.16) Interest: Online games (9%, 0.20)

find *facially neutral* free-form attributes that disproportionately target or exclude users of a sensitive group. For example, an advertiser seeking to create an audience excluding certain ethnic groups may choose to select her target audience from users interested in particular news media sites or magazines.

To answer these questions and understand how vulnerable the Facebook ad platform is to these kinds of *indirect discrimination*, we investigate how strongly curated attributes other than "ethnic affinity" correlate with ethnicity and whether free-form attributes that are facially unrelated to sensitive attributes can be used as proxies for sensitive attributes. We executed these experiments by automatically querying the Facebook ad interface for the number of people belonging (or not belonging) to sensitive groups that have a certain curated or free-form attribute.

4.2. Discriminatory audience creation

We now explore how both curated and free-form attributes are correlated with ethnicity.

Curated attributes: We conduct our analysis in the way described in Section 3.3. We use the custom audience mechanism to create groups of people from the North Carolina voter records that only contain particular ethnicities (White, African-American, Asian, and Hispanic). We then create sub-audiences by choosing to only target users matching each curated attribute and observe the size estimates of these sub-audiences. The percentage of users from each audience for whom Facebook inferred a curated attribute reveals how prevalent the attribute is within the audiences of different ethnicities.

The top three inclusive and exclusive attributes per ethnicity are shown in Table 2. The results point out that ethnic affinity is by far not the only and—in many cases—not even the most disparate feature with respect to ethnicity. For example, when targeting Asians on Facebook, it is more effective to do so based on political leaning or eating habits. The tradeoffs between representation ratio and recall for members outside the sensitive group, which an advertiser has to consider when aiming to exclude sensitive group members, can also be gauged from the table. In particular, there are a number of curated attributes with low representation ratio (i.e., high disparity), some of which achieve high recall for members not belonging to the sensitive group.

Free-form attributes: We begin our investigation by gathering an extensive (though not exhaustive) list of free-form attributes that are supported by the Facebook marketing API.[13] The API provides two useful calls that we exploit: i) given a piece of text, the API provides a list of free-form attributes that *match* the given text; and ii) given an attribute, the API provides a

13. https://developers.facebook.com/docs/marketing-api

Table 3: Free-form attributes that may be used for discriminatory targeting. We show the percentage of the attribute audience that are members of the sensitive group as well as the fraction of the U.S. Facebook population that are members of the sensitive group, as a reference.

Free-form Attribute	Potential Target (PT)	PT Audience (%)	US Audience (%)
Marie Claire	Female	90%	54%
myGayTrip.com	Man interested in Man	38.6%	0.38%
BlackNews.com	African American affinity	89%	16%
Hoa hoc Tro Magazine	Asian American affinity	95%	3.4%
Nuestro Diario	Hispanic affinity	98%	16%

Table 4: Examples of free-form attributes that can be targeted by advertisers. In the parenthesis, we show the number of audience that can be targeted or excluded with the attribute.

Topic	Free-form attributes
Religion	Islam (5.7M), Catholic Church (6.5M), Evangelicalism (5.6M)
LGBT	LGBT community(21M), Gay pride (13M), Same-sex marriage(4.2M)
Vulnerable people	Addicted (100K), REHAB (450K), AA (50K), Support group (610K)

list of other *related* attribute suggestions. For instance, the list of related attributes for 'The New York Times' includes 'The Washington Post', 'The Wall Street Journal', and 'The Economist'.

We start with a seed set of names of news outlets extracted from three different sources: Google News (Leskovec et al., 2009), List of Newspapers,[14] and the top 1,000 newspapers from Alexa.[15] We first identify around 3,000 free-form attributes that exactly match with the names of the news outlets. We then execute a snowball sampling on these attributes, using Facebook's related attribute suggestions recursively starting from them. This process resulted in retrieving nearly 240,000 free-form attributes.

We begin by trying to find attributes from the above set of 240,000 attributes that can be used to primarily target or exclude people belonging to sensitive groups. Table 3 shows example free-form attributes that could be exploited for discriminatory targeting. For example, the attribute 'Marie Claire' has an audience with 90% of women, a much larger fraction than the proportion of U.S. women in Facebook (54%). Similarly, the attribute 'myGayTrip.com' has an audience of 38.6% men interested in men, while only 0.38% of the U.S. population in Facebook consists of men interested in men. We also identified a number of attributes with very biased audiences in terms of racial affinities. For example, 'BlackNews.com' has an audience with 89% of the users with African American affinity (in contrast with 16% of African American affinity in the reference population), the audience of 'Hoa hoc Tro Magazine' is composed of 95% users with Asian American affinity, which corresponds to 28 times more in comparison with the reference population. Similarly, 'Nuestro Diario' has an audience with 98% of Hispanic affinity (16% on the reference population). These results suggest that a malicious advertiser could easily find free-form attributes to launch discriminatory ads based on gender, race, and sexual orientation.

More worryingly, some free-form attributes allow a malicious advertiser to target people based on their beliefs. Table 4 presents a few sensitive free-form attributes from our dataset along with their potential audience in the U.S. These attributes correspond to a large audience with specific religious beliefs, including 'islam' (5.7M), 'catholic church' (6.5M), and 'evangelicalism' (5.6M). Thus, although it is not possible to target religion using curated attributes, one can use free-form attribute targeting to narrow the audience to people who are interested in a specific religion. Finally, we note that it is possible to target or exclude gay and LGBT users (or people sympathetic to their causes) via attributes like 'LGBT community' (21M), 'Gay

14. http://www.listofnewspapers.com/
15. https://www.alexa.com/topsites/category/Top/News/Newspapers

Table 5: Suggestions for the most conservative news outlets. The left column shows a set of free-form attributes for conservative news outlets and the right column shows the corresponding free-form attributes suggested by Facebook. The percentage of very conservative users in the audience of each of these free-form attributes is shown in parentheses.

Very Conservative - U.S. Facebook Population (13.9%)

Input Attribute	**Attribute Suggestions**
Townhall.com (79.5%)	The Daily Caller (67.1%), RedState (84.3%), TheBlaze (59.6%), Hot Air (news site) (79.4%)
The American Spectator (70.7%)	The Daily Caller (67.1%), Townhall.com (79.4%), The American Conservative (85.2%), National Review (78.6%), Weekly Standard (72.2%), Human Events (53.3%), Commentary (34.5%), RedState (84.3%), Harper's Magazine (11.7%), U.S. News & World Report (18.6%)
The Patriot Post (70.7%)	American Patriot (68.4%), Patriot Nation (54.3%), Patriot Update (84.2%), NewsBusters.org (78.7%), Guns & Patriots (61.2%), RedEye (9.1%), America's Conservative Voice (74.4%)
American Thinker (67.5%)	National Review (78.6%), Fox Nation (75.3%)
The Cullman Times (63%)	Montgomery Advertiser (40.4%), The Huntsville Times (40.8%), The Tuscaloosa News (44.8%), al.com (42.5%)

pride' (13M), 'Same-sex marriage' (4.2M), as well as groups of vulnerable people, including 'Addicted' (100K), 'REHAB' (450K), 'AA' (50K). While this last set of free-form attributes might be useful, for example, for an advertiser to prevent addicted people to receive ads about alcoholic beverages, a discriminatory advertiser could explicitly exclude them.

Using Facebook's attribute suggestions: We first investigate the free-form attributes suggested by Facebook to better understand the criteria used to select these suggestions. Table 5 shows the suggestions returned by the Facebook Marketing API (right column) given a free-form attribute (left column). We selected attributes associated with news outlets biased towards conservative audience to check whether their respective suggestions are also similarly biased. For instance, almost 80% of the audience of Townhall.com[16] are "very conservative" Facebook users, whereas the average amount of very conservative U.S. users of Facebook is about 13.89%. We note that the audiences corresponding to most of the suggested attributes also exhibit a strong bias towards conservative audience. From all suggestions presented, only Harper's Magazine[17] and RedEye[18] have a less conservative audience in comparison with the U.S. distribution.

A malicious advertiser could exploit free-form attribute suggestions from Facebook in two different ways. First, a malicious advertiser could exploit the Facebook's attribute suggestions to discover attributes that are facially neutral, but are similarly biased as a given free-form attribute. For example, one suggestion from Facebook for 'myGayTrip.com' is the free-form attribute 'Matt Dallas', who is a gay actor.[19] 19.4% of the audience for 'Matt Dallas' are men interested in men, which is 51 times more than the U.S. distribution (0.38%). Thus, a malicious advertiser may use 'Matt Dallas' as a facially neutral proxy for targeting or excluding gay users.

Second, an advertiser can use the suggestion mechanism to search for extremely biased free-form attributes. For example, suppose an advertiser is interested in conservative leaning audiences and the most biased free-form attribute they know is 'Fox', with 37% of conservative audience. The advertiser can start with 'Fox' and keep choosing more and more conservative attribute suggestions until she reaches attributes with extreme conservative audience bias. Below, we show a sequence of suggested attributes, starting from 'Fox', that leads to 'The Sean Hannity Show', a free-form attribute with 95% conservative audience.

Fox (37%) → Fox News Channel (67%) → Sean Hannity (88%) → Mark Levin (93%) → Rush

16. https://www.facebook.com/townhallcom/
17. https://www.facebook.com/HarpersMagazine/
18. https://www.facebook.com/TheRedEye/
19. https://en.wikipedia.org/wiki/Matt_Dallas

Limbaugh (93%) → The Rush Limbaugh Show (94%) → The Sean Hannity Show (95%).

4.3. Summary

In this section, we demonstrated that many curated attributes (beyond ethnic affinity) exhibit correlations with sensitive attributes like race, which makes them potential vectors for discrimination. We also investigated whether the free-form attribute targeting mechanism allows advertisers to target or exclude sensitive groups of users in a discriminatory manner. Specifically, we showed that advertisers can circumvent existing limitations on targeting users based on their interests in sensitive topics like religion and sexual orientation. Furthermore, we show that malicious advertisers can exploit Facebook's suggestions to discover new facially neutral free-form attributes that allow extremely biased targeting.

5. Look-Alike Audience Targeting

In this section, we show how the recently introduced look-alike audience targeting mechanism can be exploited by advertisers to covertly implement discriminatory advertising. We first briefly describe the look-alike audience targeting feature of Facebook; we then explain how this feature can be exploited to implement discriminatory advertising.

Look-alike audience selection: Recently, Facebook introduced the *look-alike audience* targeting feature to help advertisers reach people that are *similar to* (i.e., look like) their existing set of customers.[20] Look-alike audiences are a particularly useful feature for advertisers who have limited data about their customers and want to grow their customer base. Advertisers can use it to outsource the job of marketing (i.e., identifying the attributes of their potential customers and finding them) to Facebook.

To select look-alike audiences, advertisers need to first provide Facebook with information about their existing (initial) set of customers called the *source audience*. An advertiser can choose source audience users in a variety of ways, including by uploading their customers' PII (similar to creating a custom audience) or by specifying them to be the fans of their Facebook page.

After specifying a source audience, Facebook allows advertisers to specify a geographical region (either countries or groups of countries) from which the look-alike audience should be chosen. Facebook orders (ranks) all users in the geographical region based on their similarity to (i.e., how closely they look like) the source audience and allows advertisers to select look-alike audiences by specifying a percentile range (e.g., <2% or 2%-4%) over these ordered users from the geographical region's population. Thus, an advertiser can select X to Y percentile of closest matching users from any country's population to target. In practice, Facebook limits Y to 10%.

5.1. Potential for discrimination

Our concern is that a malicious advertiser seeking to place discriminatory advertisements could exploit look-alike audiences as follows: they could start by creating a highly biased (highly discriminatory) source audience and use the look-alike audience feature to find a larger set of users that is similarly biased, effectively scaling the bias to much larger populations. Put differently, our concern is that when the source audience is discriminatory, its look-alike audience would also be discriminatory. In the following sections, we first investigate whether biases in source audience selection propagate to look-alike audience selection. Later, we show how an advertiser seeking to selectively target people of a particular race could simply create a small (in the order of a few thousands) but highly biased source audience consisting primarily of people of a particular race (as described in Section 3.3) and use it to effectively target a large (in the order of tens of millions) yet similarly—or worse, exaggeratedly—biased look-alike audience.

5.2. Bias in look-alike audience selection

In this section, we construct several highly biased source audiences and check if and how the selection biases in source audience propagate to look-alike audiences.

Similar to what we did in Sections 3 and 4, we use the North Carolina voter database to

20. https://www.facebook.com/business/help/164749007013531

Table 6: Top 5 most over-represented and under-represented attributes in a source audience of African Americans and its two closest look-alike audiences. In parentheses, we show the value of the representation bias of each attribute.

Over-represented Attributes	**Under-represented Attributes**
Source Audience	
African American affinity (5.52)	Asian American affinity (0.09)
US politics: very liberal (3.21)	Hispanic (Spanish dominant) affinity (0.09)
Liberal content engagement (2.98)	Expats: Mexico (0.11)
Interest: Gospel music (2.64)	Hispanic (all) affinity (0.18)
Interest: Dancehalls (2.51)	Expats: all countries (0.22)
2% Look-Alike Audience	
African American affinity (5.24)	Hispanic (Spanish dominant) affinity (0.10)
Liberal content engagement (4.16)	Expats: Mexico (0.13)
US politics: very liberal (3.29)	Asian American affinity (0.13)
Interest: Gospel music (3.07)	Hispanic (all) affinity (0.19)
Interest: Soul music (2.32)	Expats: all countries (0.24)
2–4% Look-Alike Audience	
African American affinity (5.06)	Asian American affinity (0.17)
Liberal content engagement (3.61)	Hispanic (Spanish dominant) affinity (0.18)
US politics: very liberal (3.37)	Expats: Mexico (0.19)
Interest: Gospel music (2.72)	Hispanic (all) affinity (0.29)
Interest: Dancehalls (2.54)	Expats: all countries (0.37)

construct several groups of 10,000 randomly selected people based on their ethnicity (Asian, Black, White, Hispanic), gender, political affiliation (registered Democrat or Republican) and age (18-24, 24-35, 35-54, 55+). We construct a source audience corresponding to each group and for each source audience, we ask Facebook to construct look-alike audiences from the US in five percentile ranges: closest matching 2%, 2-4%, 4-6%, 6-8%, and 8-10% of the US population. Note that the audiences in the different percentile ranges do not overlap with one another and each subsequent percentile range becomes less similar (i.e., less closely matching) to the source audience.

Each of the five look-alike audiences we create (for every source audience of 10,000 people) consist of approximately 4.2 million people, thus totaling to an approximate of 21.1 million unique people in the US. Thus, look-alike audiences allow expansion of the source audience by over three orders of magnitude. The key remaining question is whether the look-alike audience selection reflects the biases in source audience selection.

To capture the biases in our audience selection, we define a measure that we call *representation bias* of a target audience for every user attribute f maintained by Facebook. Simply put, representation bias captures how disproportionately an attribute is observed amongst the target audience ($\mathbf{TA}$) compared to the people in the geographic location from where the look-alike audience is being selected (the geographic location is the US in our scenario and we refer to people in the US as the relevant audience, $\mathbf{RA}$). More formally, the representation bias of an attribute f in an audience is defined as

$$\text{rep_bias}_f(\mathbf{TA}, \mathbf{RA}) = \frac{|\,\mathbf{TA}_f\,|}{|\,\mathbf{TA}\,|}\frac{|\,\mathbf{RA}\,|}{|\,\mathbf{RA}_f\,|}, \quad (4)$$

where similar to the representation ratio (Equation (1)), $\mathbf{TA}_f$ and $\mathbf{RA}_f$ are the subsets of people with attribute f in $\mathbf{TA}$ and $\mathbf{RA}$ respectively. We leave out attributes with very low prevalence in Facebook from our analysis (i.e., attributes for which $|\,\mathbf{RA}_f\,|/|\,\mathbf{RA}\,| < 0.01$).

Knowing the representation bias of each attribute allows us to construct a ranking of attributes from most to least biased; we refer to attributes at the top of the ranking as *over-represented* and those at the bottom to be *under-represented* in a target audience. Table 6 shows the top 5 over-represented and under-represented attributes for the source audience of African Americans and its 2% and 2–4% (the most similar

two) look-alike audiences. The table shows that a majority of attributes that were found to be overrepresented in the source audience remain so for the look-alike audiences; similar behavior can be observed for the underrepresented attributes. These results—particularly the presence of multicultural affinity attributes—suggest that Facebook is using its extensive set of attributes to likely infer the biases that we introduced into our source audience. Moreover, it is propagating these biases to the selection of look-alike audiences, constantly over-representing African Americans and under-representing Hispanics and Asian Americans compared to their proportions in the national population.

To further validate our findings above, we take the top 10 over-represented and under-represented attributes in the source audience and computed their average rank in the look-alike audiences. We performed these computations for differently biased source audiences (selected along the basis of gender, age, ethnicity and political affiliation). Figure 1 shows how the average rank changes across the look-alike audiences given by Facebook. We can see that attributes that were most over- and under-represented in source audience tend to stay, on average, amongst the most over- and under-represented in the look-alike audiences, respectively. These results strengthen our inference that the look-alike audience feature in Facebook is able to both capture the biases in a source audience and propagate the biases to the larger audiences it helps construct.

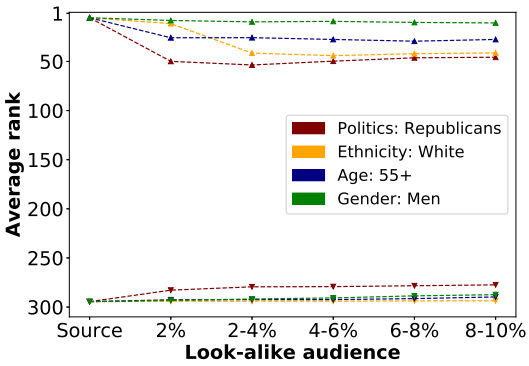


Figure 1: Comparison of the average ranks of top 10 over-represented and under-represented attributes in look-alike audiences built from different types of biased source audiences. Average ranks for over-represented attributes are indicated by upward triangles, downward triangles are used for the average ranks of under-represented attributes.

Figure 2 shows how source audiences that were disparate in favor of an ethnic group tend to produce look-alike audiences also disparate in favor of that ethnicity; although as the audiences become less similar, the disparity tapers off. Only one of these audiences, the 2% look-alike audience for White, has a disparity below 1.25, the threshold obtained from the 80% disparate impact rule (Biddle, 2005). These results show that

5.3. Discriminatory audience creation

Having observed that the look-alike audience selection mimics the biases of the source audience selection, we now check whether the bias propagation is sufficiently strong to lead to discriminatory audience creation. To answer this question, we compute the disparity of the sensitive attribute on which the source audience was biased, and observe how disparate that attribute remains in the look-alike audiences made by Facebook. Note that since we are observing look-alike audiences built from source audiences where the sensitive attribute was severely exaggerated, we expect the disparity measure to reflect the disparity in favor of the attribute.

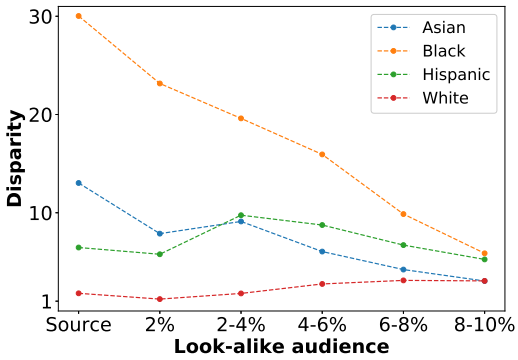


Figure 2: Disparity (in favor) for each ethnicity when the look-alike audiences are created from an audience biased on that ethnicity.

look-alike audiences selected using highly biased source audiences can be highly discriminatory.

5.4. Summary

In this section, we investigated whether it is possible to start with a small discriminatory source audience and then leverage Facebook's look-alike audience feature to construct a considerably larger discriminatory audience. We show that in order to select a look-alike audience, Facebook tries to infer the attributes that distinguish the audience from the general population and propagates these biases in the selection of look-alike audiences. Such bias propagation can amplify the explicit (intentionally created) or implicit (unintentionally overlooked) biases in a source audience of a few thousand to a look-alike audience of tens of millions. As Facebook is actively involved in the selection of the look-alike audience, one might argue that Facebook needs to be more accountable for the selection of such a discriminatory audience.

6. Concluding Discussion

Recently, concerns have been raised about the potential abuse of online advertising platforms to target ads related to housing, employment, and financial services only to users of a particular race, in violation of anti-discrimination laws. In this paper, we set out to investigate the following high-level question: *can a malicious advertiser leverage the different targeting methods offered by platforms like Facebook to target users in a discriminatory manner?* At a high-level, our study makes the following contributions:

(i) We argue that the determination of whether a targeted ad is discriminatory should not be made based on the use or non-use of specific user attributes by advertisers. Rather, inspired by the notion of *disparate impact* (Feldman et al., 2015), we propose a simple outcome-based measure for discriminatory targeting that is computed independently of the user attributes used in targeting.

(ii) Next, using public voter record data in the US, we conduct an empirical study demonstrating that several user attributes in Facebook, beyond the much-criticized "ethnic affinity," show strong positive and negative correlations with users belonging to different races. Worse, Facebook's related attribute suggestions can be exploited by advertisers to discover facially-neutral attributes that can be used for highly discriminatory audience targeting. Thus, simply banning certain attributes is insufficient to solve the problem.

(iii) Finally, we explore the vulnerability of two previously overlooked methods of targeting supported by Facebook namely, *PII-based (custom) audience targeting* and *look-alike audience targeting*. We show that both these methods can be exploited by a malicious advertiser to include or exclude users with certain sensitive features *at scale* (i.e., in the order of tens of millions of users).

Future work – Towards detecting and mitigating ad discrimination: Our study here has largely focussed on *understanding the problem* of discriminatory advertising rather than *proposing solutions* for detecting or mitigating discriminatory targeting. However, in the process, we lay the foundations for the future solutions. First, the discrimination measure proposed here could be used when designing procedures to detect discrimination in the future. Second, we argue that the look-alike audience selection feature also presents a promising solution to the problem of mitigating discrimination in audience selection. Specifically, ad platform providers could expand the targeted audience to include look-alike (most similar) users that belong to under-represented groups (rather than select all look-alike audience). We plan to explore the effectiveness of this approach in mitigating discrimination in future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by the Alexander von Humboldt Foundation, CAPES, Fapemig, and NSF grants CNS-1563320 and CNS-1616234.

References

Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Investigating ad transparency mech-

anisms in social media: A case study of Facebook's explanations. In *NDSS*, 2018.

Julia Angwin and Terry Parris Jr. Facebook lets advertisers exclude users by race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race, 2016.

Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (still) letting housing advertisers exclude users by race. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin, 2017a.

Julia Angwin, Madeleine Varner, and Ariana Tobin. Facebook enabled advertisers to reach 'jew haters'. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters, 2017b.

Stephen Ansolabehere and Eitan Hersh. Gender, race, age and voting: A research note. *Politics and Governance*, 2013.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 2016.

Dan Biddle. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower, 2005.

Facebook. https://newsroom.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools, 2017.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM KDD*, 2015.

Anikó Hannák, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring price discrimination and steering on e-commerce web sites. In *ACM IMC*, 2014.

Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from TaskRabbit and Fiverr. In *ACM CSCW*, 2017.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *ACM KDD*, 2009.

Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting price and search discrimination on the internet. In *ACM HotNets*, 2012.

Tehila Minkus, Yuan Ding, Ratan Dey, and Keith W. Ross. The city privacy attack: Combining social media and public records for detailed profiles of adults and children. In *ACM COSN*, 2015.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of Twitter users. In *AAAI ICWSM*, 2011.

Cong Tang, Keith Ross, Nitesh Saxena, and Ruichuan Chen. What's in a name: A study of names, gender inference, and gender behavior in Facebook. In *DASFAA*, 2011.