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
{andreou}@eurecom.fr
[†]Northeastern University, USA
{venkatadri.g@husky.neu.edu, amislove@ccs.neu.edu}
[*]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, France
{oana.goga, patrick.loiseau}@univ-grenoble-alpes.fr
[‡]Max Planck Institute for Software Systems (MPI-SWS), Germany
{gummadi}@mpi-sws.org

***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 1-891562-49-5
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.

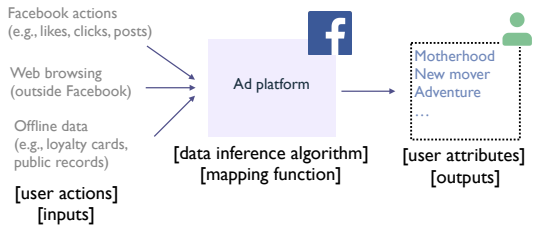
We first separate out the different processes that are responsible for a user receiving an ad, then briefly describe how advertisers can place ads using Facebook's advertising interface, and finally analyze the various targeting methods available to advertisers by studying what data about users is used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

A. The processes responsible for receiving an ad

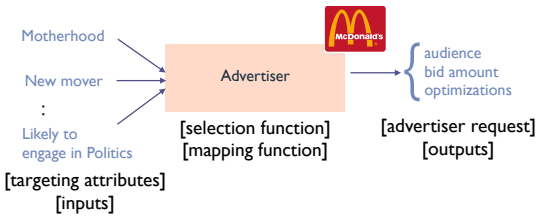
The central goal of our paper is to analyze Facebook's answers to the question *Why am I being shown this ad?* The reason why a user received a particular ad is, however, the result of a complex process that depends on many inputs. To enumerate just a few, it depends on: what the platform thinks the 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".

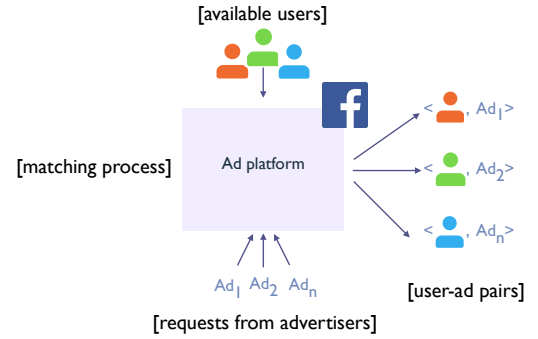
[2] Our study was reviewed and approved by our respective institutions' Institutional Review Boards.



(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.

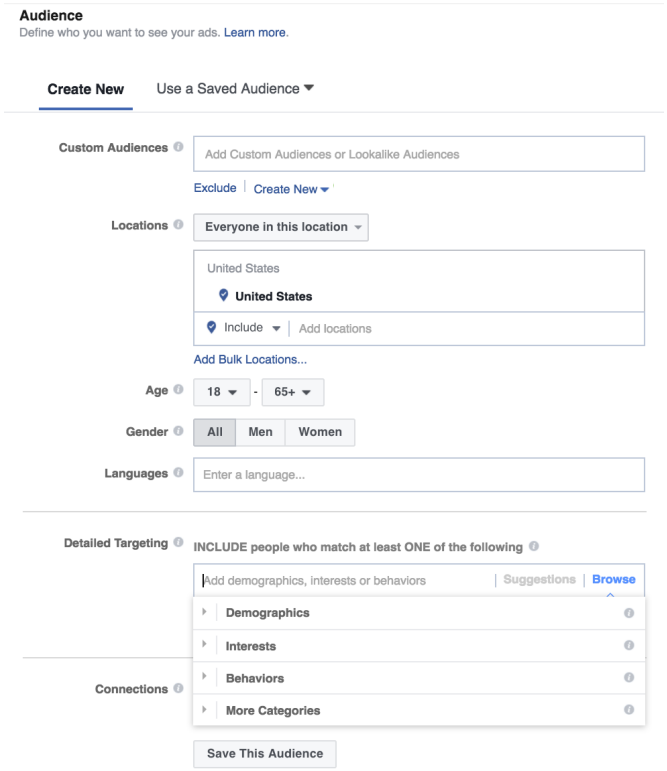


Fig. 2: Facebook's audience creation interface.

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

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

[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.

[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 $\{i_1$ AND $i_2\}$, or $\{i_3\}$. An incorrect explanation would be $\{i_4$ 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).

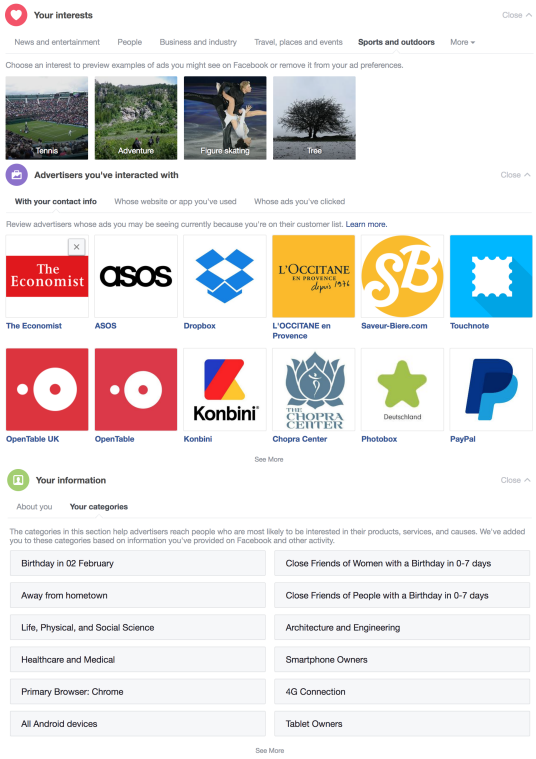


Fig. 3: Example of information provided in the Ad Preferences Page.

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.