

Discrimination through optimization: How Facebook’s ad delivery can lead to skewed outcomes

Muhammad Ali*
Northeastern University
mali@ccs.neu.edu

Piotr Sapiezynski*
Northeastern University
sapiezynski@gmail.com

Miranda Bogen
Upturn
miranda@upturn.org

Aleksandra Korolova
University of Southern California
korolova@usc.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Aaron Rieke
Upturn
aaron@upturn.org

ABSTRACT

The enormous financial success of online advertising platforms is partially due to the precise targeting features they offer. Although researchers and journalists have found many ways that advertisers can target—or exclude—particular groups of users seeing their ads, comparatively little attention has been paid to the implications of the platform’s *ad delivery* process, comprised of the platform’s choices about who should see an ad.

It has been hypothesized that this process can “skew” ad delivery in ways that the advertisers do not intend, making some users less likely than others to see particular ads based on their demographic characteristics. In this paper, we demonstrate that such skewed delivery occurs on Facebook, due to market and financial optimization effects as well as the platform’s own predictions about the “relevance” of ads to different groups of users. We find that both the advertiser’s budget and the content of the ad each significantly contribute to the skew of Facebook’s ad delivery. Critically, we observe significant skew in delivery along gender and racial lines for “real” ads for employment and housing opportunities despite neutral targeting parameters.

Our results demonstrate previously unknown mechanisms that can lead to potentially discriminatory ad delivery, even when advertisers set their targeting parameters to be highly inclusive. This underscores the need for policymakers and platforms to carefully consider the role of the ad delivery optimization run by ad platforms themselves—and not just the targeting choices of advertisers—in preventing discrimination in digital advertising¹.

1 INTRODUCTION

Powerful digital advertising platforms fund most popular online services today, serving ads to billions of users daily. At a high level, the functionality of these advertising platforms can be divided into two phases: *ad creation*, where advertisers submit the text and images that comprise the content of their ad and choose targeting parameters, and *ad delivery*, where the platform delivers ads to specific users based on a number of factors, including advertisers’ budgets, their ads’ performance, and the predicted relevance of their ads to users.

As advertising platforms have added features and grown in popularity, many have raised concerns over how they could be misused.

For example, one of the underlying reasons for the popularity of these services with advertisers is the rich suite of *targeting* features they offer during ad creation, which allow advertisers to precisely specify which users (called the *audience*) are eligible to see the advertiser’s ad. The particular features that advertisers can use for targeting vary across platforms, but often include demographic attributes, behavioral information, users’ personally identifiable information (PII), mobile device IDs, and web tracking pixels [11, 67].

Due to the wide variety of targeting features—as well as the availability of sensitive targeting features such as user demographics and interests—researchers have raised concerns about discrimination in advertising, where groups of users may be excluded from receiving certain ads based on advertisers’ targeting choices [63]. This concern is particularly acute in the areas of credit, housing, and employment, where there are legal protections in the U.S. that prohibit discrimination against certain protected classes in advertising [1–3]. As ProPublica demonstrated in 2016 [27], this risk is not merely theoretical: ProPublica investigators were able to run housing ads that explicitly excluded users with specific “ethnic affinities” from receiving them.² Recently, the U.S. Department of Housing and Urban Development (HUD) sued Facebook over these concerns and others, accusing Facebook’s advertising platform of “encouraging, enabling, and causing” violations of the Fair Housing Act [26].

The role of ad delivery in discrimination Although researchers and investigative journalists have devoted considerable effort to understanding the potential discriminatory outcomes of ad targeting, comparatively little effort has focused on ad delivery, due to the difficulty of studying its impacts without internal access to ad platforms’ data and mechanisms. However, there are several potential reasons why the ad delivery algorithms used by a platform may open the door to discrimination.

First, consider that most platforms claim their aim is to show users “relevant” ads: for example, Facebook states “we try to show people the ads that are most pertinent to them” [62]. Intuitively, the goal is to show ads that particular users are likely to engage with, even in cases where the advertiser does not know a priori which users are most receptive to their message. To accomplish this, the platforms build extensive user interest profiles and track

¹The devlivery statistics of ad campaigns described in this work can be accessed at <https://facebook-targeting.ccs.neu.edu/>

* These two authors contributed equally

²In response, Facebook recently banned the use of certain attributes for housing ads, but many other, un-banned, mechanisms exist for advertisers that achieve the same outcome [63]. Facebook agreed as part of a lawsuit settlement stemming from these issues to go further by banning age, gender, and certain kinds of location targeting—as well as some related attributes—for housing, employment, or credit ads [21].

ad performance to understand how different users interact with different ads. This historical data is then used to steer future ads towards those users who are most likely to be interested in them, and to users like them. However, in doing so, the platforms may inadvertently cause ads to deliver primarily to a skewed subgroup of the advertiser’s selected audience, an outcome that the advertiser may not have intended or be aware of. As noted above, this is particularly concerning in the case of credit, housing, and employment, where such skewed delivery might violate antidiscrimination laws.

Second, market effects and financial optimization can play a role in ad delivery, where different desirability of user populations and unequal availability of users may lead to skewed ad delivery [23]. For example, it is well-known that certain users on advertising platforms are more valuable than others [44, 51, 59]. Thus, advertisers who choose low budgets when placing their ads may be more likely to lose auctions for such “valuable” users than advertisers who choose higher budgets. However, if these “valuable” user demographics are strongly correlated with protected classes, it could lead to discriminatory ad delivery due to the advertiser’s budget alone. Even though a low budget advertiser may not have intended to exclude such users, the ad delivery system may do just that because of the higher demand for that subgroup.

Prior to this work, although hypothesized [23, 48, 66], it was not known whether the above factors resulted in skewed ad delivery in real-world advertising platforms. In fact, in response to the HUD lawsuit [26] mentioned above, Facebook claimed that the agency had “no evidence” of their ad delivery systems’ role in creating discrimination [38].

Contributions In this paper, we aim to understand whether ads could end up being shown in a skewed manner—i.e., where some users are less likely than others to see ads based on their demographic characteristics—due to the ad delivery phase alone. In other words, we determine whether the ad delivery could cause skewed delivery *that an advertiser did not cause by their targeting choices and may not even be aware of*. We focus on Facebook—as it is the most mature platform offering advanced targeting features—and run dozens of ad campaigns, hundreds of ads with millions of impressions, spending over \$8,500 as part of our study.

Answering this question—especially without internal access to the ad delivery algorithm, user data, and advertiser targeting data or delivery statistics—involves overcoming a number of challenges. These include separating market effects from optimization effects, distinguishing ad delivery adjustments based on the ad’s performance measured through user feedback from initial ad classification, and developing techniques to determine the racial breakdown of the delivery audience (which Facebook does not provide). The difficulty of solving these without the ad platform’s cooperation in a rigorous manner may at least partially explain the lack of knowledge about the potential discriminatory effects due to ad delivery to date. After addressing these challenges, we find the following:

First, we find that *skewed delivery can occur due to market effects alone*. Recall the hypothesis above concerning what may happen if advertisers in general value users differently across protected classes. Indeed, we find this is the case on Facebook: when we run identical ads targeting the same audience but with varying budgets, the resulting audience of users who end up actually seeing our ad

can range from over 55% men (for ads with very low budgets) to under 45% men (for ads with high budgets).

Second, we find that *skewed delivery can occur due to the content of the ad itself* (i.e., the ad headline, text, and image, collectively called the *ad creative*). For example, ads targeting the same audience but that include a creative that would stereotypically be of the most interest to men (e.g., bodybuilding) can deliver to over 80% men, and those that include a creative that would stereotypically be of the most interest to women (e.g., cosmetics) can deliver to over 90% women. Similarly, ads referring to cultural content stereotypically of most interest to black users (e.g., hip-hop) can deliver to over 85% black users, and those referring to content stereotypically of interest to white users (e.g., country music) can deliver to over 80% white users, even when targeted identically by the advertiser. Thus, despite placing the same bid on the same audience, the advertiser’s ad delivery can be heavily skewed based on the ad creative alone.

Third, we find that *the ad image itself has a significant impact on ad delivery*. By running experiments where we swap different ad headlines, text, and images, we demonstrate that the differences in ad delivery can be significantly affected by the image alone. For example, an ad whose headline and text would stereotypically be of the most interest to men with the image that would stereotypically be of the most interest to women delivers primarily to women at the same rate as when all three ad creative components are stereotypically of the most interest to women.

Fourth, we find that *the ad image is likely automatically classified by Facebook*, and that this classification can skew delivery from the beginning of the ad’s run. We create a series of ads where we add an alpha channel to stereotypically male and female images with over 98% transparency; the result is an image with all of the image data present, but that looks like a blank white square to humans. We find that there are statistically significant differences in how these ads are delivered depending on the image used, despite the ads being visually indistinguishable to a human. This indicates that the image classification—and, therefore, relevance determination—is likely an automated process, and that the skew in ad delivery can be due in large part to skew in Facebook’s automated estimate of relevance, rather than performance of the ad itself.

Fifth, we show that *real-world employment and housing ads can experience significantly skewed delivery*. We create and run ads for employment and housing opportunities, and use our methodology to measure their delivery to users of different races and genders. When optimizing for clicks, we find that ads with the same targeting options can deliver to vastly different racial and gender audiences depending on the ad creative alone. In the most extreme cases, our ads for jobs in the lumber industry reach an audience that is 72% white and 90% male, our ads for cashier positions in supermarkets reach an 85% female audience, and our ads for positions in taxi companies reach a 75% Black audience, even though the targeted audience specified by us as an advertiser is identical for all three. We run a similar suite of ads for housing opportunities, and find skew there as well: despite the same targeting and budget, some of our ads deliver to an audience of over 85% white users, while others deliver to over 65% Black users. While our results only speak to how our particular ads are delivered (i.e., we cannot say how housing or employment ads *in general* are delivered), the significant

skew we observe even on a small set of ads suggests that real-world housing and employment ads are likely to experience the same fate.

Taken together, our results paint a distressing picture of heretofore unmeasured and unaddressed skew that can occur in online advertising systems, which have significant implications for discrimination in targeted advertising. Specifically, due to platforms' optimization in the ad delivery stage together with market effects, ads can unexpectedly be delivered to skewed subsets of the advertiser's specified audience. For certain types of ads, such skewed delivery might implicate legal protections against discriminatory advertising; we leave a full exploration of these implications to the legal community. However, our results indicate that regulators, lawmakers, and the platforms themselves need to think carefully when balancing the optimization of ad platforms against desired societal outcomes, and remember that ensuring that individual advertisers do not discriminate in their targeting is insufficient to achieve non-discrimination goals sought by regulators and the public.

Ethics All of our experiments were conducted with careful consideration of ethics. We obtained Institutional Review Board review of our study at Northeastern University (application #18-11-13), with our protocol being marked as "Exempt". We minimized harm to Facebook users when we were running our ads by always running "real" ads (in the sense that if people clicked on our ads, they were brought to real-world sites relevant to the topic of the ad). While running our ads, we never intentionally chose to target ads in a discriminatory manner (e.g., we never used discriminatory targeting parameters). To further minimize the potential for discrimination, we ran most of our experimental ads in categories with no legal salience (such as entertainment and lifestyle); we only ran ad campaigns on jobs and housing to verify whether the effects we observed persist in these domains. We minimized harm to the Facebook advertising platform by paying for ads and using the ad reporting tools in the same manner as any other advertiser. The particular sites we advertised were unaffiliated with the study, and our ads were not defamatory, discriminatory, or suggestive of discrimination.

2 BACKGROUND

Before introducing our methodology and analyses, we provide background on online display advertising, describe Facebook's advertising platform's features, and detail related work.

2.1 Online display advertising

Online display advertising is now an ecosystem with aggregate yearly revenues close to \$100 billion [20]. The web advertising ecosystem is a complex set of interactions between ad publishers, ad networks, and ad exchanges, with an ever-growing set of entities involved at each step allowing advertisers to reach much of the web. In contrast, online services such as Facebook and Twitter run advertising platforms that generally serve a single site (namely, Facebook and Twitter themselves).

In this paper, we focus primarily on single-site advertising platforms, but our results may also be applicable to more general display advertising on the web; we leave a full investigation of the extent to which this is the case to future work. The operation of platforms

such as Facebook and Twitter can be divided into two phases: *ad creation* and *ad delivery*. We provide more details on each below.

Ad creation Ad creation refers to the process by which the advertiser submits their ad to the advertising platform. At a high level, the advertiser has to select three things when doing so:

- (1) *Ad contents*: Advertisers will typically provide the ad headline, text, and any images/videos. Together, these are called the *ad creative*. They will also provide the link where the platform should send users who click.
- (2) *Audience Selection/Targeting*: Advertisers need to select which platform users they would like to see the ad (called the *audience*).
- (3) *Bidding strategy*: Advertisers need to specify how much they are willing to pay to have their ads shown. This can come in the form of a per-impression or per-click bid, or the advertiser can simply place an overall *bid cap* and allow the platform to bid on their behalf.

Once the advertiser has entered all of the above information, they submit the ad for review³; once it is approved, the ad will move to the ad delivery phase.

Ad delivery Ad delivery refers to the process by which the advertising platform shows ads to users. For every opportunity to show a user an ad (e.g., an *ad slot* is available as the user is browsing the service), the ad platform will run an *ad auction* to determine, from among all of the ads that include the current user in the audience, which ad should be shown.

In practice, however, the ad delivery process is somewhat more complicated. *First*, the platforms try to avoid showing ads from the same advertiser repeatedly in quick succession to the same user; thus, the platforms will sometimes disregard bids for recent winners of the same user. *Second*, the platforms often wish to show users relevant ads; thus, rather than relying solely on the bid to determine the winner of the auction, the platform may incorporate a relevance score into consideration, occasionally allowing ads with cheaper bids but more relevance to win over those with higher bids. *Third*, the platforms may wish to evenly spread the advertiser budget over their specified time period, rather than use it all at once, which introduces additional complexities as to which ads should be considered for particular auctions. The exact mechanisms by which these issues are addressed is not well-described or documented by the platforms.

Once ads enter the ad delivery phase, the advertising platforms give advertisers information on how their ads are performing. Such information may include detailed breakdowns (e.g., along demographic and geographic lines) of the characteristics of users to whom their ad is being shown along with some characteristics of users who are clicking on the ad.

2.2 Facebook's advertising platform

In this paper, we focus on Facebook's advertising platform as it is one of the most powerful and feature-rich advertising platforms in use today. As such, we provide a bit more background here

³Most platforms have a review process to prevent abuse or violations of their platforms' advertising policies [7, 71].

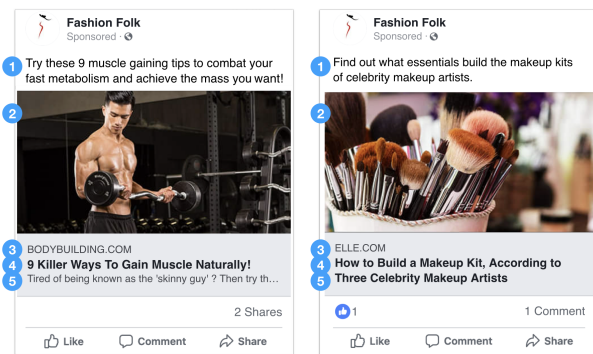


Figure 1: Each ad has five elements that the advertiser can control: (1) the ad headline and text, entered manually by the advertiser, (2) the images and/or videos, (3) the domain, pulled automatically from the HTML meta property `og:site_name` of the destination URL, (4) the title, pulled automatically from the HTML meta property `og:title` of the destination URL, and (5) the description from meta property `og:description` of the destination URL.

about the specific features and options that Facebook provides to advertisers.

Ad contents When placing an ad on Facebook, it must be linked to a *Page*; advertisers are allowed to have multiple Pages and run ads for any of them. Ads can come in multiple forms, such as promoting particular posts on the page. However, for typical ads, the advertiser must provide (a) the headline and text to accompany the ad, and (b) one or more images or videos to show to the user. Optionally, the advertiser can provide a *traffic destination* to send the user to if they click (e.g., a Facebook Page or an external URL); if the advertiser provides a traffic destination, the ad will include a brief description (auto-generated from the HTML meta data) about this destination. Examples showing all of these elements are presented in Figure 1.

Audience selection Facebook provides a wide variety of audience selection (or *targeting*) options [10, 11, 34, 63]. In general, these options fall into a small number of classes:

- *Demographics and attributes:* Similar to other advertising platforms [35, 65], Facebook allows advertisers to select audiences based on demographic information (e.g., age, gender, and location), as well as profile information, activity on the site, and data from third-parties. Recent work has shown that Facebook offers over 1,000 well-defined attributes and hundreds of thousands of free-form attributes [63].
- *Personal information:* Alternatively, Facebook allows advertisers to specify *the exact users* who they wish to target by either (a) uploading the users’ personally identifiable information including names, addresses, and dates of birth [28, 67, 68], or (b) deploying web tracking pixels

on third-party sites [30]. On Facebook, audiences created using either mechanism are called *Custom Audiences*.⁴

- *Similar users:* Advertisers may wish to find “similar” users to those who they have previously selected. To do so, Facebook allows advertisers to create *Lookalike Audiences*⁵ by starting with a source Custom Audience they had previously uploaded; Facebook then “identif[ies] the common qualities of the people in it” and creates a new audience with other people who share those qualities [31].

Advertisers can often combine many of these features together, for example, by uploading a list of users’ personal information and then using attribute-based targeting to further narrow the audience.

Objective and bidding Facebook provides advertisers with a number of *objectives* to choose from when placing an ad [8], where each tries to maximize a different *optimization event* the advertiser wishes to occur. These include “Awareness” (simply optimizing for the most *impressions*, a.k.a. views), “Consideration” (optimizing for clicks, engagement, etc.), and “Conversion” (optimizing for sales generated by clicking the ad). For each objective, the advertiser bids on the objective itself (e.g., for “Awareness”, the advertiser would bid on ad impressions). The bid can take multiple forms, and includes the start and end time of the ad campaign and either a lifetime or a daily budget cap. With these budget caps, Facebook places bids in ad auctions on the advertisers’ behalf. Advertisers can optionally specify a per-bid cap as well, which will limit the amount Facebook would bid on their behalf for a single optimization event.

Facebook’s ad auction When Facebook has ad slots available, it runs an ad auction among the active advertisements bidding for that user. However, the auction does not just use the bids placed by the advertisers; Facebook says [32]:

The ad that wins an auction and gets shown is the one with the highest *total value* [emphasis added]. Total value isn’t how much an advertiser is willing to pay us to show their ad. It’s combination of 3 major factors: (1) Bid, (2) Estimated action rates, and (3) Ad quality and relevance.

Facebook defines “Estimated action rates” as “how well an ad performs”, meaning whether or not *users in general* are engaging with the ad [5]. They define “Ad quality and relevance” as “how interesting or useful we think a given user is going to find a given ad”, meaning how much a *particular user* is likely to be interested in the ad [5].

Thus, it is clear that Facebook attempts to identify the users within an advertiser’s selected audience who they believe would find the ad most useful (i.e., those who are most likely to result in an optimization event) and shows the ad preferentially to those users. Facebook says exactly as such in their documentation [4]:

During ad set creation, you chose a target audience ... and an optimization event ... We show your ad to people in that target audience who are likely to get you that optimization event

⁴Google, Twitter, and Pinterest all provide similar features; these are called *Customer Match* [6], *Tailored Audiences*, and *Customer Lists* [55], respectively

⁵Google and Pinterest offer similar features: on Google it is called *Similar Audiences* [36], and on Pinterest it is called *Actalike Audiences* [57].

Facebook goes on to say [62]:

Put simply, the higher an ad’s relevance score, the less it will cost to be delivered.

Statistics and reporting Facebook provides advertisers with a feature-rich interface [24] as well as a dedicated API [52] for both launching ads and monitoring those ads as they are in ad delivery. Both the interface and the API give semi-live updates on delivery, showing the number of impressions and optimization events as the ad is running. Advertisers can also request this data be broken down along a number of different dimensions, including age, gender, and location (Designated Market Area [53], or DMA, region). Notably, the interface and API *do not* provide a breakdown of ad delivery along racial lines; thus, analyzing delivery along racial lines necessitates development of a separate methodology that we describe in the next section.

Anti-discrimination rules In response to issues of potential discrimination in online advertising reported by researchers and journalists [27], Facebook currently has several policies in place to avoid discrimination for certain types of ads. Facebook also recently built tools to automatically detect ads offering housing, employment, and credit, and pledged to prevent the use of certain targeting categories with those ads. [42]. Additionally, Facebook relies on advertisers to self-certify [14] that they are not in violation of Facebook’s advertising policy prohibitions against discriminatory practices [25]. More recently, in order to settle multiple lawsuits stemming from these reports, Facebook stated that they will soon no longer allow age, gender, or ZIP code-based targeting for housing, employment or credit ads, and that they would also block other detailed targeting attributes that are “describing or appearing to relate to protected classes” [21].

2.3 Related work

Next, we detail related work on algorithm auditing, transparency, and discriminatory ad targeting.

Auditing algorithms for fairness Following the growing ubiquity of algorithms in daily life, a community formed around investigating their societal impacts [60]. Typically, the algorithms under study are not available to outside auditors for direct examination; thus, most researchers treat them as “black boxes” and observe their reactions to different inputs. Among most notable results, researchers have shown price discrimination in online retail sites [40], gender discrimination in job sites [15, 41], stereotypical gender roles re-enforced by online translation services [12] and image search [43], disparate performance on gender classification for Black women [13], and political partisanship in search [19, 47, 58]. Although most of the work focused exclusively on the algorithms themselves, recently researchers began to point out that auditors should consider the entire socio-technical systems that include the users of those algorithms, an approach referred to as “algorithm-in-the-loop” [37, 61]. Furthermore, recent work has demonstrated that fairness is not necessarily composable, i.e., for several notions of fairness such as individual fairness [22], a collection of classifiers that are fair in isolation do not necessarily result in a fair outcome when they are used as part of a larger system [23].

Advertising transparency In parallel to the developments in detecting and correcting unfairness, researchers have conducted studies and introduced tools with the aim of increasing transparency and explainability of algorithms and their outcomes. For example, much attention has been dedicated to shedding light on the factors that influence the targeting of a particular ad on the web [49, 50, 56, 73] and on specific services [18, 72].

Focusing on Facebook, Andreou et al. investigated the transparency initiative from Facebook that purportedly tells users why they see particular targeted ads [11]. They found that the provided explanations are incomplete and, at times, misleading. Venkatadri et al. introduced the tool called “TREADS” that attempts to close this gap by providing Facebook users with detailed descriptions of their inferred attributes using the ads themselves as a vehicle [69]. Further, they investigated how data from third-party data brokers is used in Facebook’s targeting features and—for the first time—revealed those third-party attributes to the users themselves using TREADS [70]. Similar to other recent work [54], Venkatadri et al. found that the data from third-party data brokers had varying accuracy [70].

Discrimination in advertising As described above, Facebook has some policies and tools in place to prevent discriminatory ad targeting. However, advertisers can still exclude users based on a variety of interests that are highly correlated with race by using custom audiences [63], or by using location [33, 46]. Separately, Sweeney [64] and Datta et al. [18] have studied discrimination in Google’s advertising system, and have examined the potential parties responsible and how their actions may be interpreted under the law [17].

The work just described deals with identifying possibilities for the advertisers to run discriminatory ads using the platform’s features. In contrast, other researchers, as well as and HUD’s recent complaint, have suggested that discrimination may be introduced by the ad platform itself, rather than by a malicious advertiser [18, 38, 48, 66]. For example, Lambrecht et al. ran a series of ads for STEM education and found they were consistently delivered more to men than to women, even though there are more female users on Facebook, and they are known to be more likely to click on ads and generate conversions. Our work explores this initial finding in depth, separating market effects from optimization effects and exploring the mechanisms by which ads are delivered in a skewed manner.

3 METHODOLOGY

We now describe our methodology for measuring the delivery of Facebook ads. At a high level, our goal is to run groups of ads where we vary a particular feature, with the goal of then measuring how changing that feature skews the set of users the Facebook platform delivers the ad to. To do so, we need to carefully control which users are in our target audience. We also need to develop a methodology to measure the ad delivery skew along racial lines, which, unlike gender, is not provided by Facebook’s existing reporting tools. We detail how we achieve that in the following sections.

3.1 Audience selection

When running ads, we often wish to control exactly which ad auctions we are participating in. For example, if we are running multiple instances of the same ad (e.g., to establish statistical confidence), we do not want the ads to be competing against each other. To avoid this, we use random PII-based custom audiences, where we randomly select U.S. Facebook users to be included in mutually-exclusive audiences. By doing so, we can ensure that our ads are only competing against each other in the cases where we wish them to.

Generating custom audiences We create each custom audience by randomly generating 20 lists of 1,000,000 distinct, valid North American phone numbers (+1 XXX XXX XXXX, using known-valid area codes). Facebook reported that they were able to match approximately 220,000 users on each of the 20 lists we uploaded.

Initially, we used these custom audiences directly to run ads, but while conducting the experiments we noticed that—even though we specifically target only North American phone numbers—many ads were delivered to users outside of North America. This could be caused by users traveling abroad, users registering with fake phone numbers or with online phone number services, or for other reasons, whose investigation is outside the scope of this paper. Therefore, for all the experiments in this paper, we use our custom audiences and additionally specify that we only want to target people located in the U.S.

3.2 Data collection

Once one of our ad campaigns is run, we use the Facebook Marketing API to obtain the delivery performance statistics of the ad every two minutes. When we make this request, we ask Facebook to break down the ad delivery performance according to the attribute of study (age, gender, or location). Facebook’s response to each query features the following fields, among others, for each of the demographic attributes that we requested:

- `impressions`: The number of times the ad was shown
- `reach`: The number of unique users the ad was shown to
- `clicks`: The number of clicks the ad has received
- `unique_clicks`: The number of unique users who clicked

Throughout the rest of the paper, we use the reach value when examining delivery; thus, when we report “Fraction of men in the audience” we calculate this as the reach of men divided by the sum of the reach of men and the reach of women (see Section 3.5 for discussion on using binary values for gender).

3.3 Measuring racial ad delivery

The Facebook Marketing API allows advertisers to request a breakdown of ad delivery performance along a number of axes but it does not provide a breakdown based on race. However, for the purposes of this work, we are able to measure the ad delivery breakdown along racial lines by using location (Designated Market Area, or DMA⁶) as a proxy.

⁶Designated Market Areas [53] are groups of U.S. counties that Neilson defines as “market areas”; they were originally used to signify a region where users receive similar broadcast television and radio stations. Facebook reports ad delivery by location using DMA regions, so we use them here as well.

	DMA Region(s) [53]	Aud.	Records
1	Wilmington	White	450,000
	Raleigh–Durham	White	450,002
2	Greenville–Spartanburg	Black	446,047
	Greenville–New Bern		
	Charlotte Greensboro	Black	446,050

Table 1: Overview of the four North Carolina custom audiences used to measure racial delivery. We divide the most populated DMAs in the state into two groups, and in each group, create two audiences with ~450,000 users of the same race. We then use the statistics Facebook reports about delivery by DMAs to infer delivery by race.

Similar to prior work [63], we obtain voter records from North Carolina; these are publicly available records that have the name, address, race, and often phone number of each registered voter in the state. To maintain fairly equal audience sizes, we partition the most populated North Carolina DMA regions into two groupings that have roughly the same number of users from each racial group that we consider: White and Black. After sampling approximately 900,000 users for each race from their corresponding DMAs, we split these audiences into two, in order to support running multiple ads in parallel on non-overlapping audiences. We upload the voter data from each of these four lists as separate Custom Audiences to Facebook.⁷ The details of the resulting four audiences are shown in Table 1.

When we run ads where we want to examine the ad delivery along racial lines, we run the ads to one audience from the first grouping and the other race’s audience from the second grouping. We then request that Facebook’s Marketing API deliver us results broken down by DMA region. Because we selected DMA regions to be a proxy for race, we can use the results to infer which custom audience they were originally in, allowing us to determine the racial makeup of the audience who saw (and clicked on) the ad.

3.4 Ad campaigns

We use the Facebook Ad API described in Section 2.2 to create all ads for our experiments and to collect data on their delivery. We carefully control for any time-of-day effects that might be present due to different user demographics using Facebook at different times of the day: for any given experiment, we run all ads at the same time to ensure that any such effects are experienced equally by all ads. Unless otherwise noted, we used the following settings:

- *Objective*: Consideration→Traffic⁸
- *Optimization Goal*: Link Clicks
- *Traffic destination*: An external website (that depends on the ads run)
- *Creative*: All of our ads had a single image and text relevant to the ad.

⁷Unfortunately, Facebook does not report the number of these users who match as we use multiple PII fields in the upload file [67].

⁸This target is defined as: Send more people to a destination on or off Facebook such as a website, app, or Messenger conversation.

- *Audience selection:* We use custom audiences for most of our ads, as described in Section 3.1, and further restrict them to adult (18+) users of all genders residing in the United States.
- *Budget:* We ran most ads with a budget of \$20 per day, and stopped them typically after six hours.

3.5 Measuring and comparing audiences

We now describe the measurements we make during our experiments and how we compute their confidence intervals.

Binary values of gender and race Facebook’s marketing API reports “female”, “male”, and “unknown” as the possible values for gender. Across our experiments, we observe that up to 1% of the audiences are of “unknown” gender. Further, when running our experiments measuring race (and targeting specific DMAs), we observe that a fraction (~10%) of our ads are delivered to audiences outside of our predefined DMAs, thus making it impossible for us to infer their race. This fraction remains fairly consistent across our experiments regardless of what we advertise, thus introducing the same amount of noise across our measurements. This is not entirely unexpected, as we are targeting users directly, and those users may be traveling, may have moved, may have outdated information in the voter file, etc. Regardless, we treat both of these attributes as binary (ignoring the “unknown” gender and any delivery outside of our target DMAs) and leave more complete investigation to future work. Because of this simplifying decision, we recognize that delivery can be skewed with respect to gender non-binary users and/or users of other races in a way that remains unreported in this work.

Measuring statistical significance Using the binary race and gender features, throughout this work, we describe the audiences by the fraction of male users and the fraction of white users. We calculate the lower and upper limits of the 99% confidence interval around this fraction using the method recommended by Agresti and Coull [9], defined in Equation 1:

$$\begin{aligned}
 L.L. &= \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n}, \\
 U.L. &= \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n},
 \end{aligned} \tag{1}$$

where $L.L.$ is the lower confidence limit, $U.L.$ is the upper confidence limit, \hat{p} is the observed fraction of the audience with the attribute (here: male), n is the size of the audience reached by the ad. To obtain the 99% interval we set $z_{\alpha/2} = 2.576$. The advantage of using this calculation instead of the more frequently used normal approximation $p \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is that the resulting intervals fall in the $(0, 1)$ range. Whenever the confidence intervals around these fractions for two audiences are non-overlapping, we can make a claim that the gender or racial makeups of two audiences are significantly different [16]. However, the converse is not true: overlapping confidence intervals do not necessarily mean that the

means are not different (see Figure 4 in [16] for explanation). In this work we report all the results of our experiments but for easier interpretation emphasize those where the confidence intervals are non-overlapping. We further confirm that the non-overlapping confidence intervals represent statistically significant differences, using the difference of proportion test as shown in Equation 2:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{2}$$

where \hat{p}_1 and \hat{p}_2 are the fractions of men (white users) in the two audiences that we compare, n_1 and n_2 are sizes of these audiences, and \hat{p} is the fraction of men (white users) in the two delivery audiences combined. All the results we refer to as statistically significant are significant in this test with a Z -score of at least 2.576.

4 EXPERIMENTS

In this section, we explore how an advertiser’s choice of ad creative (headline, text, and image) and ad campaign settings (bidding strategy, targeted audience) can affect the demographics (gender and race) of the users to whom the ad is ultimately delivered.

4.1 Budget effects on ad delivery

We begin by examining the impact that market effects can have on delivery, aiming to test the hypothesis put forth by Lambrecht et al. [48]. In particular, they observed that their ads were predominantly shown to men even though women had consistently higher click through rates (CTRs). They then hypothesized that the higher CTRs led to women being more expensive to advertise to, meaning they were more likely to lose auctions for women when compared to auctions for men.

We test this hypothesis by running the same ad campaign with different budgets; our goal is to measure the effect that the daily budget alone has on the makeup of users who see the ads. When running these experiments, we keep the ad creative and targeted audience constant, only changing the bidding strategy to give Facebook different daily limits (thus, any ad delivery differences can be attributed to the budget alone). We run an ad with daily budget limits of \$1, \$2, \$5, \$10, \$20, and \$50, and run multiple instances at each budget limit for statistical confidence. Finally, we run the experiment twice, once targeting our random phone number custom audiences, and once targeting all users located in U.S.; we do so to verify that any effect we see is not a function of our particular target audience.

Figure 2 presents the results, plotting the daily budget we specify versus the resulting fraction of men in the audience. The left graph shows the results when we target all users located in the U.S., and the right graph shows the results when we target the random phone number custom audiences. In both cases, we observe that changes in ad delivery due to differences in budget are indeed happening: the higher the daily budget, the smaller the fraction of men in the audience, with the Pearson’s correlation of $\rho = -0.88$, $p_{val} < 10^{-5}$ for all U.S. users and $\rho = -0.73$, $p_{val} < 10^{-3}$ for the custom audiences. The stronger effect we see when targeting all U.S. users may be due to the additional freedom that the ad delivery system has when choosing who to deliver to, as this is a significantly larger audience.

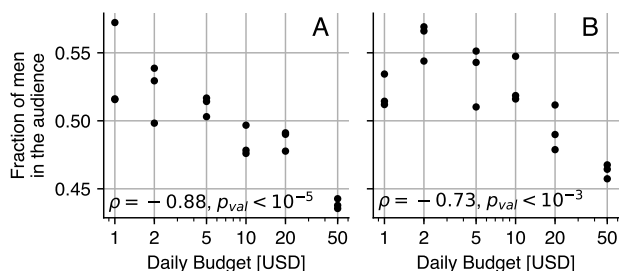


Figure 2: Gender distributions of the audience depend on the daily budget of an ad, with higher budgets leading to a higher fraction of women. The left graph shows an experiment where we target all users located in the U.S.; the right graph shows an experiment where we target our random phone number custom audiences.

To eliminate the impact that market effects can have on delivery in our following experiments, we ensure that all runs of a given experiment use the same bidding strategy and budget limit. Typically we use a daily budget of \$20 per campaign.

4.2 Ad creative effects on ad delivery

Now we examine the effect that the ad creative (headline, text, and image) can have on ad delivery. To do so, we create two stereotypical ads that we believe would appeal primarily to men and women, respectively: one ad focusing on *bodybuilding* and another on *cosmetics*. The actual ads themselves are shown in Figure 1. We run each of the ads at the same time and with the same bidding strategy and budget. We run five instances of each pair of ads in parallel, targeting different custom audiences, to ensure each of our ads is competing against exactly one other of our ads. *Note that we do not explicitly target either ad based on gender; the only targeting restrictions we stipulate are 18+ year old users in the U.S.*

We observe dramatic differences in ad delivery, even though the bidding strategy is the same for all ads, and each pair of ads target the same gender-agnostic audience. In particular, the bodybuilding ad ended up being delivered to over 75% men on average, while the cosmetics ad ended up being delivered to over 90% women on average. Again, this skewed delivery is despite the fact that we—the advertiser—did not specify difference in budget or target audience.

Individual components’ impact on ad delivery With the knowledge that the ad creative can skew delivery, we dig deeper to determine *which* of the components of the ad creative (headline, text, and image) have the greatest effect on ad delivery. To do so, we stick with the bodybuilding and cosmetics ads, and “turn off” various features of the ad creative by replacing them with empty strings or blank images. For example, the bodybuilding experiment listed as “base” includes an empty headline, empty ad text, and a blank white image; it does however link to the domain `bodybuilding.com`. Similarly, the cosmetics experiment listed as “base” includes no headline, text, or image, but does link to the domain `e11e.com`. We then add back various parts of the ad creative, as shown in Figure 1.

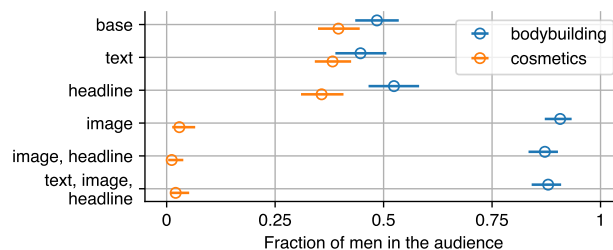


Figure 3: “Base” ad contains a link to a page about either bodybuilding or cosmetics, a blank image, no text, or headline. There is a small difference in the fraction of male users for the base ads, and adding setting the “text” only decreases it. Setting the “headline” sets the two ads apart but the audience of each is still not significantly different than that of the base version. Finally, setting the ad “image” causes drastic changes: the bodybuilding ad is shown to a 91% male audience, the cosmetics ad is shown to very few men, despite the same target audience.

The results of this experiment are presented in Figure 3. Error bars in the figure correspond to 99% confidence intervals as defined in Equation 1. All results are shown relative to that experiment’s “base” ad containing only the destination URL. We make a number of observations. First, we can observe an ad delivery difference due to the destination URL itself; the base bodybuilding ad delivers to 48% men, while the base cosmetics ad delivers to 40% men. Second, as we add back the title and the headline, the ad delivery does not appreciably change from the baseline. However, once we introduce the image into the ad, the delivery changes dramatically, returning to the level of skewed delivery discussed above (over 75% male for bodybuilding, and over 90% female for cosmetics). When we add the text and/or the headline back alongside the image, the skew of delivery does not change significantly compared to the presence of image only. Overall, our results demonstrate that the choice of ad image can have a dramatic effect on which users in the audience ultimately are shown the ad.

Swapping images To further explore how the choice of image impacts ad delivery, we continue using the bodybuilding and cosmetics ads, and test how ads with incongruent images and text are delivered. Specifically, we swap the images between the two ads, running an ad with the bodybuilding headline, text, and destination link, but with the image from cosmetics (and vice versa). We also run the original ads (with congruent images and text) for comparison.

The results of this experiment are presented in Figure 4, showing the skew in delivery of the ads over time. The color of the lines indicates the image that is shown in the ad; solid lines represent the delivery of ads with images consistent with the description, while dotted lines show the delivery for ads where image was replaced. We make a number of observations. First, when using congruent ad text and image (solid lines), we observe the skew we observed before. However, we can now see clearly that this delivery skew appears to exist from the very beginning of the ad delivery, i.e., before users

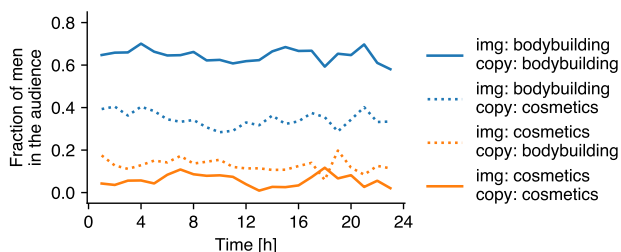


Figure 4: Ad delivery of original bodybuilding and cosmetics ads, as well as the same ads with incongruent images. Skew in delivery is observed from the beginning, and using incongruent images skews the delivery to a lesser degree.

begin viewing and interacting with our ads. We will explore this further in the following section. Second, we see that when we switch the images—resulting in incongruent ads (dotted lines)—the skew still exists but to a lesser degree. Notably, we observe that the ad with an image of bodybuilding but cosmetics text delivers closest to 50:50 across genders, but the ad with the image of cosmetics but bodybuilding text does not. The exact mechanism by which Facebook decides to use the ad text and images in influencing ad delivery is unknown, and we leave a full exploration to future work.

Swapping images mid-experiment Facebook allows advertisers to change their ad while it is running, for example, to update the image or text. As a final point of analysis, we examine how changing the ad creative mid-experiment—after it has started running—affects ad delivery. To do so, we begin the experiment with the original congruent bodybuilding and cosmetics ads; we let these run for over six hours. We then swap the images on the running ads, thereby making the ads incongruent, and examine how ad delivery changes.

Figure 5 presents the results of this experiment. In the top graph, we show the instantaneous ad delivery skew: as expected, the congruent ads start to deliver in a skewed manner as we have previously seen. After the image swap at six hours, we notice a very rapid change in delivery with the ads almost completely flipping in ad delivery skew in a short period of time. Interestingly, we do not observe a significant change in users’ behavior to explain this swap: the bottom graph plots the click through rates (CTRs) for both ads by men and women over time. Thus, our results suggest that the change in ad delivery skew is unlikely to be due to the users’ responses to the ads.

4.3 Source of ad delivery skew

We just observed that ads see a significant skew in ad delivery due to the contents of the ad, despite the bidding strategy and targeting parameters being held constant. However, we observed that the ad delivery skew was present from the very beginning of ad delivery, and that swapping the image in the middle of a run resulted in a very rapid change in ad delivery. We now turn to explore the mechanism that may be leading to this ad delivery skew.

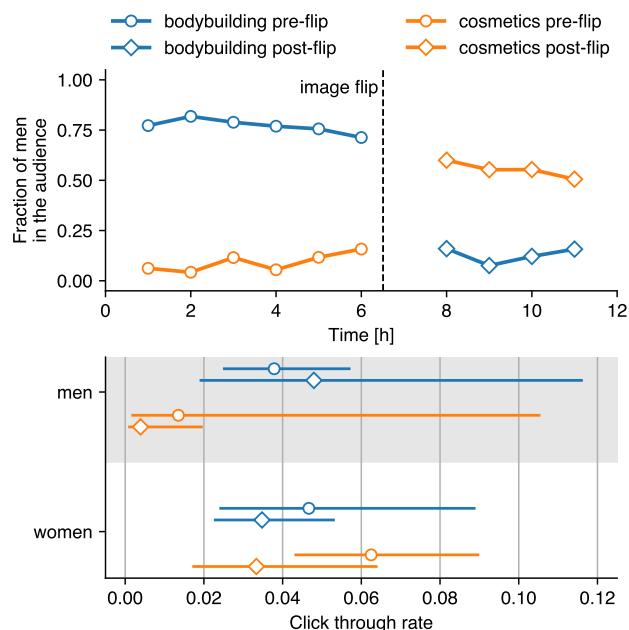


Figure 5: When we flip the image in the middle of the campaign, the ad is reclassified and shown to an updated audience. Here, we start bodybuilding and cosmetics ads with corresponding descriptions and after 6 hours and 32 minutes we flip the images. Within an hour of the change, the gender proportions are reversed, while there is no significant difference between the click through rates per gender pre and post flipping of the images.

Almost-transparent images We begin with the hypothesis that Facebook itself is automatically classifying the ad creative (including the image), and using the output of this classification to calculate a predicted relevance score to users. In other words, we hypothesize that Facebook is running automatic text and image classification, rather than (say) relying on the ad’s initial performance, which would explain (a) the delivery skew being present from the beginning of ad delivery, and (b) how the delivery changes rapidly despite no significant observable change in user behavior. However, validating this hypothesis is tricky, as we are not privy to all of Facebook’s ad performance data.

To test this hypothesis, we take an alternate approach. We use the *alpha channel* that is present in many modern image formats; this is an additional channel that allows the image to encode the *transparency* of each pixel. Thus, if we take an image and add an alpha channel with (say) 99% opacity, all of the image data will still be present in the image, but any human who views the image would not be able to see it (as the image would show almost completely transparent). However, if an automatic classifier exists, and if that classifier is not properly programmed to handle the alpha channel, it may continue to classify the image.

Test images To test our hypothesis, we select five images that would stereotypically be of interest to men and five images that





















No.	Male		Female	
	Visible	Invisible	Visible	Invisible
1				
2				
3				
4				
5				

Table 2: Diagram of the images used in the transparency experiments. Shown are the five stereotypical male and female images, along with the same images with a 98% alpha channel, denoted as invisible. The images with the alpha channel are almost invisible to humans, but are still delivered in a skewed manner.

would stereotypically be of interest to women; these are shown in the second and fourth columns of Table 2.^{9,10} We convert them to PNG format add an alpha channel with 98% opacity¹¹ to each of these images; these are shown in the third and fifth columns of Table 2. Because we cannot render a transparent image without a background, the versions in the paper are rendered on top of a white background. As the reader can see, these images are not discernible to the human eye.

We first ran a series of tests to observe how Facebook’s ad creation phase handled us uploading such transparent images. If we used Reach as our ad objective, we found that Facebook “flattened” these images onto a white background in the ad preview.¹² By targeting ourselves with these Reach ads, we verified that when they were shown to users on the Facebook mobile app or in the

⁹All of these images were cropped from images posted to pexels.com, which allow free non-commercial use.

¹⁰We cropped these images to the Facebook-recommended resolution of 1,080×1,080 pixels to reduce the probability Facebook would resample the image.

¹¹We were unable to use 100% transparency as we found that Facebook would run an image hash over the uploaded images and would detect different images with 100% opacity to be the same (and would refuse to upload it again). By using 98% transparency, we ensure that the images were still almost invisible to humans but that Facebook would not detect they were the same image.

¹²Interestingly, we found that if we instead used Traffic as our ad objective, Facebook would both “flatten” these images onto a white background and then normalize the contrast. This caused the ads to be visible to humans—simply with less detail than the original ads—thus defeating the experiment. We are unsure of why Facebook did not choose to normalize images with the objective for Reach.

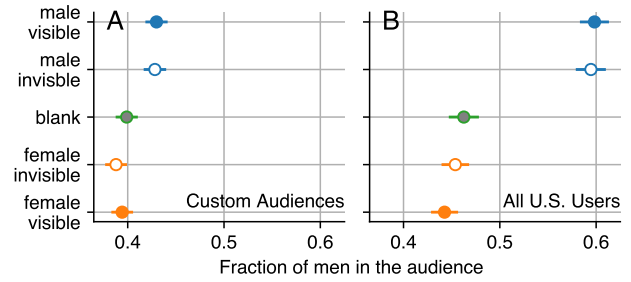


Figure 6: Ad delivery to ads with the images from Table 2, targeting general US audience as well as the random phone number custom audience. The solid markers are visible images, and the hollow markers are the same images with 98% opacity. Also shown is the delivery to truly white images (“blank”). We can observe that a difference in ad delivery exists, and that that difference is statistically significant between the male and female invisible images. This suggests that automated image classification is taking place.

desktop Facebook web feed, the images did indeed show up as white squares. Thus, we can use this methodology to test whether there is an automatic image classifier present by examining whether running different transparent white ads results in different delivery.

Results We run ads with all twenty of the images in Table 2, alongside ads with five truly blank white images for comparison. For all 25 of these ads, we hold the ad headline, text, and destination link constant, run them all at the same time, and use the same bidding strategy and target custom audience. We then record the differences in ad delivery of these 25 images along gender lines. The results are presented in the left graph of Figure 6A, with all five images in each of the five groups aggregated together. Error bars in the plot correspond to the 99% confidence interval calculated using Equation 1. We can observe that ad delivery is, in fact, skewed, with the ads with stereotypically male images delivering to over 42% men and the ads with female delivering to 39% men.

Interestingly, we also observe that the male invisible ads appear to be indistinguishable in performance from the male visible ads, and the female invisible ads appear to be indistinguishable in performance from the female visible ads.

As shown in Figure 6A, we verify that the fraction of men in the delivery of the male ads is significantly higher than in female-centered and neutral ads, as well as higher in neutral ads than in female-centered ads. We also show that we cannot reject the null hypothesis that the fraction of men in the two versions of each ad (one visible, one invisible) are the same. Thus, we can conclude that the difference in ad delivery of our invisible male and female images is statistically significant, despite the fact that humans would not be able to perceive any differences in these ads. This strongly suggests that our hypothesis is correct: that Facebook has an automated image classification mechanism in place that is

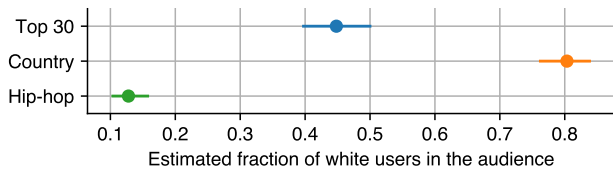


Figure 7: We run three campaigns about the best selling albums. *Top 30* is neutral, targeting all. *Country* implicitly targets white users, and *Hip-hop* implicitly targets Black users. Facebook classification picks up on the implicit targeting and shows it to the audience we would expect.

used to steer different ads towards different subsets of the user population.¹³

To confirm this finding, we re-run the same experiment except that we change the target audience from our random phone number custom audiences (hundreds of thousands of users) to all U.S. users (over 320 million users). Our theory is that if we give Facebook’s algorithm a larger set of auctions to compete in, any effect of skewed delivery would be amplified as they may be able to find more users for whom the ad is highly “relevant”. In Figure 6B we observe that the ad delivery differences are, indeed, even greater: the male visible and invisible images deliver to approximately 60% men, while the female visible and invisible images deliver to approximately 45% men. Moreover, the statistical significance of this experiment is even stronger, with a Z value over 10 for the ad delivery difference between the male invisible and female invisible ads.

4.4 Impact on real ads

We have observed that differences in the ad headline, text, and image can lead to dramatic difference in ad delivery, despite the bidding strategy and target audience of the advertiser remaining the same. However, all of our experiments thus far were on test ads where we typically changed only a single variable. We now turn to examine the impact that ad delivery can have on realistic ads, where all properties of the ad creative can vary.

Entertainment ads We begin by constructing a series of benign entertainment ads that, while holding targeting parameters fixed, implicitly target users of different races. Namely, we run three ads leading to lists of best albums in the previous year: general top 30 (neutral), top country music (stereotypically of interest mostly to white users), and top hip-hop albums (stereotypically of interest mostly to Black users). We find that Facebook ad delivery follows the stereotypical distribution, despite all ads being targeted in the same manner and using the same bidding strategy. Figure 7 shows the fraction of white users in the audience in the three different ads, treating race as a binary (Black users constitute the remaining fraction). Error bars represent 99% confidence intervals calculated using Equation 1.

¹³It is important to note we not know exactly how the classification works. For example, the classifier may also be programmed to take in the “flattened” images that appear almost white, but there may sufficient data present in the images for the classification to work. We leave a full exploration of how exactly the classifier is implemented to future work.

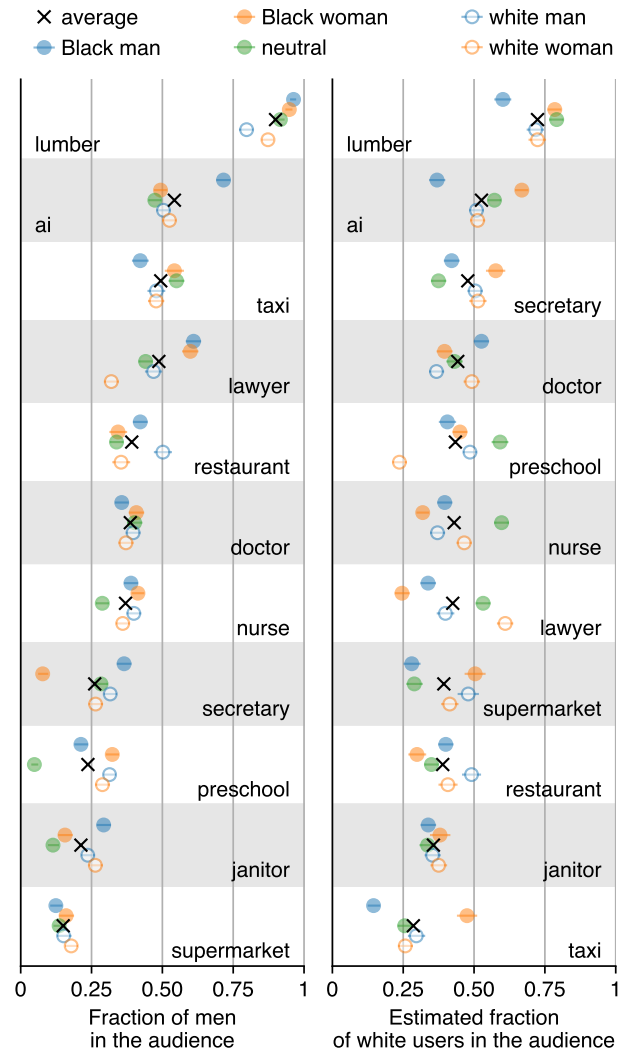


Figure 8: Results for employment ads, showing a breakdown of ad delivery by gender (left figure) and race (right figure) in the ultimate delivery audience. The labels refer to the race/gender of the person in the ad image (if any). The jobs themselves are ordered by the average fraction of men or white users in the audience. Despite the same bidding strategy, the same target audience, and being run at the same time, we observe significant skew along on both racial and gender lines due to the content of the ad alone.

Neutral ads are seen by a relatively balanced, 45% white audience, while the audiences receiving the country and hip-hop ads are 80% and 13% white, respectively. Assuming significant population level differences of preferences, it can be argued that this experiment highlights the “relevance” measures embedded in ad delivery working as intended. Next, we investigate cases where such differences may not be desired.

Employment ads Next, we advertise eleven different generic job types: artificial intelligence developer, doctor, janitor, lawyer, lumberjack, nurse, preschool teacher, restaurant cashier, secretary, supermarket clerk, and taxi driver. For each ad, we customize the text, headline, and image as a real employment ad would. For example, we advertise for taxi drivers with the text “Begin your career as a taxi driver or a chauffeur and get people to places on time.” For each ad, we link users to the appropriate category of job listings on a real-world job site.

When selecting the ad image for each job type, we select five different stock photo images: one that has a white male, one that has a white female, one that has a black male, one that has a black female, and one that is appropriate for the job type but has no people in it. We run each of these five independently to test a representative set of ads for each job type, looking to see how they are delivered along gender and racial lines. Thus, the target audiences that we use for these experiments are the North Carolina audiences described in Section 3.3. We run these ads for 24 hours, using the objective of Traffic, all targeting the same audience with the same bidding strategy.

The results of this experiment are presented in Figure 8, plotting the distribution of each of our ads along gender (left graph) and racial (right graph) lines. As before, the error bars represent the 99% confidence interval calculated using Eq. 1. We can immediately observe drastic differences in ad delivery across our ads along both racial and gender lines: our five ads for positions in the lumber industry deliver to over 90% men and to over 70% white users in aggregate, while our five ads for janitors deliver to over 65% women and over 75% black users in aggregate. Recall that the only difference between these ads are the ad creative and destination link; we (the advertiser) used the same bidding strategy and target audience, and ran all ads at the same time.

It is important to note that we cannot make conclusions about how ads for different jobs are delivered *in general*, as we only have studied how our particular set of ads were delivered (i.e., we do not claim that Facebook delivers *all* employment ads for lumberjacks primarily to white users). However, the fact that we observe significantly skewed delivery suggests that employment ads run by real-world employers are likely subject to skewed delivery as well.

Housing ads Finally, we create a suite of ads that advertise a variety of housing opportunities, as discrimination in online housing ads has recently been a source of concern [26]. We vary the type of property advertised (rental vs. purchase), the implied cost (fixer-upper vs. luxury), and the presence of a family in the ad image (just the house vs. a Black family vs. a white family). In each ad, the cost is implied through wording of the ad as well as the accompanying image. Ads with either of the families present also mention the word ‘family’ in the description. Each ad leads to a listing of houses for sale or rental apartments in North Carolina on a real-world housing site. Simultaneously, we ran a baseline ad with generic (non-housing) text that simply links to google.com. All of the ads ran for 12 hours, using the objective of Traffic, all targeting the same North Carolina audiences and using the same bidding strategy.

We present the results in Figure 9 (interestingly, we found little skew for the housing ads along gender lines, and we omit those

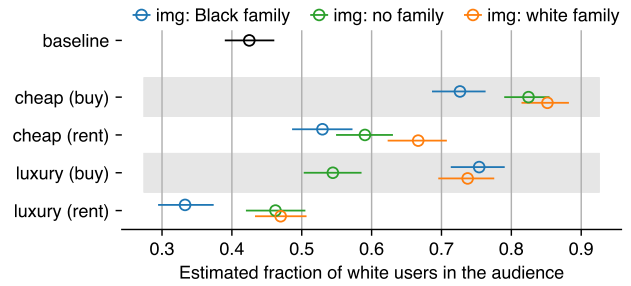


Figure 9: Results for housing ads, showing a breakdown in the ad delivery audience by race. Despite being targeted in the same manner, using the same bidding strategy, and being run at the same time, we observe significant skew in the makeup of the audience to whom the ad is delivered (ranging from over 85% white users to over 65% Black users).

results). We observe significant ad delivery skew along racial lines in the delivery of our ads, with certain ads delivering to an audience of over 85% white users while others delivering to an audience of as little as 35% white users. As with the employment ads, we cannot make claims about what particular properties of our ads lead to this skew, or about how housing ads in general are delivered. However, given the significant skew we observe with our suite of ads, it indicates the further study is needed to understand how real-world housing ads are delivered.

5 CONCLUDING DISCUSSION

To date, the public debate about discrimination in digital advertising has focused heavily on the targeting features offered by advertising platforms, and the ways that advertisers can misuse those features.

In this paper, we set out to investigate a different question: *to what degree and by what means may advertising platforms themselves play a role in creating discriminatory outcomes?*

Our study offers an improved understanding of the mechanisms behind and impact of ad delivery, a process distinct from ad creation and targeting. While ad targeting is facilitated by an advertising platform—but nominally controlled by advertisers—ad delivery is conducted and controlled by the advertising platform itself. We demonstrate that, during the ad delivery phase, advertising platforms can play an independent, central role in creating skewed, and potentially discriminatory, outcomes. More concretely, we have:

- Replicated and affirmed prior research suggesting that market and pricing dynamics can create conditions that lead to differential outcomes, by showing that the lower the daily budget for an ad, the fewer women it is delivered to;
- Shown that Facebook’s ad delivery process can significantly alter the audience the ad is delivered to compared to the one intended by the advertiser based on the content of the ad itself. We used public voter record data to demonstrate that broadly and inclusively targeted ads can end up being differentially delivered to specific audience segments, even when we hold the budget and target audience constant.

- Demonstrated that skewed ad delivery can start at the beginning of an ad’s run. We also showed that this process is likely automated on Facebook’s side, and is not a reflection of the early feedback received from users in response to the ad, by using transparent images in ads that appear the same to humans but are distinguishable by automatic image classification tools, and showing they result in skewed delivery.
- Confirmed that skewed delivery can take place on real-world ads for housing and employment opportunities by running a series of employment ads and housing ads with the same targeting parameters and bidding strategy. Despite differing only in the ad creative and destination link, we observed skewed delivery along racial and gender lines.

We briefly touch on the broader implications of our findings.

Limitations It is important to note that while we have revealed certain aspects of how ad delivery is accomplished, and the effects it had on our experimental ad campaigns, we cannot make broad conclusions about how it impacts ads more generally. For example, we observe that all of *our ads* for lumberjacks deliver to an audience of primarily white and male users, but that may not hold true of *all ads* for lumberjacks. However, the significant ad delivery skew that we observe for our employment and housing ads strongly suggests that such skew is present for such ads run by real-world advertisers.

Policy implications Our findings underscore the need for policymakers and platforms to carefully consider the role of the optimizations run by the platforms themselves—and not just the targeting choices of advertisers—in seeking to prevent discrimination in digital advertising.

First, because discrimination can arise in ad delivery independently from ad targeting, limitations on ad targeting—such as those currently deployed by Facebook to limit the targeting features that can be used—will not address discrimination arising from ad delivery. On the contrary, to the extent limiting ad targeting features prompts advertisers to rely on larger target audiences, the mechanisms of ad delivery will have an even greater practical impact on the ads that users see.

Second, regulators, lawmakers, and platforms themselves will need to more deeply consider whether and how longstanding civil rights laws apply to modern advertising platforms in light of ad delivery dynamics. At a high level, federal law prohibits discrimination in the marketing of housing, employment and credit opportunities. A detailed consideration of these legal regimes is beyond the scope of this paper. However, our findings show that ad platforms themselves can shape access to information about important life opportunities in ways that might present a challenge to equal opportunity goals.

Third, in the U.S., Section 230 of the Communications Decency Act (CDA) provides broad legal immunity for internet platforms acting as publishers of third-party content. This immunity was a central issue in recently-settled litigation against Facebook, who argued its ad platform should be protected by CDA Section 230 in part because its advertisers are “wholly responsible for deciding where, how, and when to publish their ads.” [29] Our research shows that this claim is misleading, particularly in light of Facebook’s role

in determining the ad delivery outcomes. Even absent unlawful behavior by advertisers, our research demonstrates that Facebook’s own, independent actions during the delivery phase are crucial to determining how, when, and to whom ads are shown, and might produce unlawful outcomes. These effects can be invisible to, and might even create liability for, Facebook’s advertisers.

Thus, the effects we observed could introduce new liability for Facebook. In determining whether Section 230 protections apply, courts consider whether an internet platform “materially contributes” to the alleged illegal conduct. Courts have yet to squarely consider how the delivery mechanisms described in this paper might affect an ad platform’s immunity under Section 230.

Fourth, our results emphasize the need for increased transparency into advertising platforms, particularly around ad delivery algorithms and statistics for real-world housing, credit, or employment ads. Facebook’s existing ad transparency efforts are not yet sufficient to allow researchers to analyze the impact of ad delivery in the real world.

Potential mitigations Given the potential impact that discriminatory ad delivery can have on exposure to opportunities available to different populations, a natural question is how ad platforms such as Facebook may mitigate these effects. This is not straightforward, and is likely to require increased commitment and transparency from ad platforms as well as development of new algorithmic and machine learning techniques. For instance, as we have demonstrated empirically in Section 4.1 (and as [23] have shown theoretically), skewed or unfair ad delivery can occur even if the ad platform refrains from refining the audience supplied by the advertisers according to the predicted relevance of the ad to individual users. This happens because different users are valued differently by advertisers, which, in a setting of limited user attention, leads to a tension between providing a useful service for users and advertisers, making the opportunities those advertisers are sharing available to different user populations in a fair way, and the platform’s own revenue goals.¹⁴

Thus, more advanced and nuanced approaches to addressing the potential issues of discrimination in digital advertising are necessary. One promising direction for such work may be for the platforms to develop and deploy ad delivery algorithms that consider the fairness of the entire outcome, rather than merely of individual ad campaigns, as part of their optimization objective. For example, the recently introduced notion of preference-informed fairness [45], which generalizes the fairness notions of envy-freeness [39] (widely adopted in the economics community) and individual fairness [22] (widely adopted in the theoretical computer science community), may be applicable to the digital advertising setting, and, under certain assumptions, permits efficient optimization.

Digital advertising increasingly influences how people are exposed to the world and its opportunities, and helps keep online services free of monetary cost. At the same time, its potential for negative impacts, through optimization due to ad delivery, is growing. Lawmakers, regulators, and the ad platforms themselves need to address these issues head-on.

¹⁴A formal statement of this claim for the theoretical notions of individual fairness [22] and its generalization, preference-informed fairness, can be found in [45].

ACKNOWLEDGEMENTS

The authors thank NaLette Brodnax and Christo Wilson for their invaluable feedback on the manuscript and Martin Goodson for pointing out erroneous confidence intervals. The authors also acknowledge Hannah Masuga, a graduate fellow at Upturn, whose initial experiments during her fellowship inspired this research. This work was funded in part by a grant from the Data Transparency Lab.

ERRATA

v2: In the version of the paper published on April 3rd, 2019 we wrongly stated in Section 3.5 that ~40-50% of ads were delivered outside of our predefined DMAs. In version v2 we corrected this figure to ~10%. Further, in response to a request from Facebook, we changed the axis labels from “Fraction of white users in the audience” to “Estimated fraction of white users in the audience”.

v3: We changed the method of calculating confidence intervals from normal approximation to the method described by Agresti and Coull [9]. All confidence intervals presented in the figures throughout the paper use this method. The change does not affect any of the conclusions. Notably, after the change the confidence intervals in Figure 4 no longer cross 0.

REFERENCES

- [1] 12 CFR § 202.4 (b) – Discouragement. <https://www.law.cornell.edu/cfr/text/12/202.4>.
- [2] 24 CFR § 100.75 – Discriminatory advertisements, statements and notices. <https://www.law.cornell.edu/cfr/text/24/100.75>.
- [3] 29 USC § 623 – Prohibition of age discrimination. <https://www.law.cornell.edu/uscode/text/29/623>.
- [4] About Ad Delivery. <https://www.facebook.com/business/help/1000688343301256>.
- [5] About Ad Principles. <https://www.facebook.com/business/about/ad-principles>.
- [6] About Customer Match. <https://support.google.com/adwords/answer/6379332?hl=en>.
- [7] About Twitter Ads approval. <https://business.twitter.com/en/help/ads-policies/introduction-to-twitter-ads/about-twitter-ads-approval.html>.
- [8] About advertising objectives. <https://www.facebook.com/business/help/517257078367892>.
- [9] ALAN, A. AND A. C. B. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 52, 2 (1998), Taylor & Francis.
- [10] ANDREOU, A., SILVA, M., BENEVENUTO, F., GOGA, O., LOISEAU, P., AND MISLOVE, A. Measuring the Facebook Advertising Ecosystem. In *Network and Distributed System Security Symposium* (San Diego, California, USA, Feb. 2019).
- [11] ANDREOU, A., VENKATADRI, G., GOGA, O., GUMMADI, K. P., LOISEAU, P., AND MISLOVE, A. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook’s Explanations. In *Network and Distributed System Security Symposium* (San Diego, California, USA, Feb. 2018).
- [12] BOLUKBASI, T., CHANG, K.-W., ZOU, J. Y., SALIGRAMA, V., AND KALAI, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural and Information Processing Systems* (Barcelona, Spain, Dec. 2016).
- [13] BUOLAMWINI, J. AND GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency* (New York, New York, USA, Feb. 2018).
- [14] Certify Compliance to Facebook’s Non-Discrimination Policy. <https://www.facebook.com/business/help/136164207100893>.
- [15] CHEN, L., HANNAK, A., MA, R., AND WILSON, C. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction* (Montreal, Canada, April 2018).
- [16] CUMMING, G. AND FINCH, S. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist* 60, 2 (2005), American Psychological Association.
- [17] DATTA, A., DATTA, A., MAKAGON, J., MULLIGAN, D. K., AND TSCHANTZ, M. C. Discrimination in Online Personalization: A Multidisciplinary Inquiry. In *Conference on Fairness, Accountability, and Transparency* (New York, New York, USA, Feb. 2018).
- [18] DATTA, A., TSCHANTZ, M. C., AND DATTA, A. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Privacy Enhancing Technologies Symposium* (Philadelphia, Pennsylvania, USA, June 2015).
- [19] DIAKOPOULOS, N., TRIELLI, D., JENNIFERSTARK, AND MUSSENDEEN, S. I Vote For—How Search Informs Our Choice of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (2018), Oxford University Press, M. Moore and D. Tambini, eds.
- [20] Digital Ad Spend Hits Record-Breaking \$49.5 Billion in First Half of 2018, Marking a Significant 23% YOY Increase. <https://www.iab.com/news/digital-ad-spend-hits-record-breaking-49-5-billion-in-first-half-of-2018/>.
- [21] Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising. <https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/>.
- [22] DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O., AND ZEMEL, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (2012).
- [23] DWORK, C. AND ILVENTO, C. Fairness Under Composition. (2018). <https://arxiv.org/abs/1806.06122>.
- [24] Facebook Ads Manager. <https://www.facebook.com/business/help/200000840044554>.
- [25] Facebook Advertising Policies, Discriminatory Practices. https://www.facebook.com/policies/ads/prohibited_content/discriminatory_practices.
- [26] Facebook Engages in Housing Discrimination with Its Ad Practices, U.S. Says. <https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html>.
- [27] Facebook Lets Advertisers Exclude Users by Race. <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race/>.
- [28] Facebook Marketing API – Custom Audiences. <https://developers.facebook.com/docs/marketing-api/custom-audiences-targeting/v3.1>.
- [29] Facebook Motion to Dismiss in *Onuoha v. Facebook*. <https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.34.0.pdf>.
- [30] Facebook: About Facebook Pixel. <https://www.facebook.com/business/help/742478679120153>.
- [31] Facebook: About Lookalike Audiences. <https://www.facebook.com/business/help/164749007013531>.
- [32] Facebook: About the delivery system: Ad auctions. <https://www.facebook.com/business/help/430291176997542>.
- [33] FAIZULLABHOY, I. AND KOROLOVA, A. Facebook’s Advertising Platform: New Attack Vectors and the Need for Interventions. *Computing Research Repository* (Mar. 2018), <https://arxiv.org/abs/1803.10099>, Workshop on Technology and Consumer Protection (ConPro).
- [34] GHOSH, A., VENKATADRI, G., AND MISLOVE, A. Analyzing Facebook Political Advertisers’ Targeting. In *Workshop on Technology and Consumer Protection* (San Francisco, California, USA, May 2019).
- [35] Google: About audience targeting. <https://support.google.com/google-ads/answer/2497941?hl=en>.
- [36] Google: About similar audiences on the Display Network. <https://support.google.com/google-ads/answer/2676774?hl=en>.
- [37] GREEN, B. AND CHEN, Y. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Conference on Fairness, Accountability, and Transparency* (Atlanta, Georgia, USA, Jan. 2019).
- [38] HUD Sues Facebook Over Housing Discrimination and Says the Company’s Algorithms Have Made the Problem Worse. <https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms>.
- [39] HAL, V. Efficiency, equity and envy. *Journal of Economic Theory* 9 (1974).
- [40] HANNAK, A., SOELLER, G., LAZER, D., MISLOVE, A., AND WILSON, C. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *ACM Internet Measurement Conference* (Vancouver, Canada, Nov. 2014).
- [41] HANNAK, A., WAGNER, C., GARCIA, D., MISLOVE, A., STROHMAIER, M., AND WILSON, C. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *ACM conference on Computer Supported Cooperative Work* (Portland, Oregon, USA, Feb. 2017).
- [42] Improving Enforcement and Promoting Diversity: Updates to Ads Policies and Tools. <http://newsroom.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools/>.
- [43] KAY, M., MATUSZEK, C., AND MUNSON, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction* (2015).
- [44] KIM, L. The Most Expensive Keywords in Google Ad Words. (2011). <http://www.wordstream.com/blog/ws/2011/07/18/most-expensive-google-adwords-keywords/>.
- [45] KIM, M. P., KOROLOVA, A., ROTHBLUM, G. N., AND YONA, G. Preference-Informed Fairness. (2019). <https://arxiv.org/abs/1904.01793>.
- [46] KOROLOVA, A. Facebook’s Illusion of Control over Location-Related Ad Targeting. Medium (Dec. 2018). <https://medium.com/@korolova/facebook-illusion-of-control-over-location-related-ad-targeting-de7f865aee78>.
- [47] KULSHRESTHA, J., ESLAMI, M., MESSIAS, J., ZAFAR, M. B., GHOSH, S., GUMMADI, K., AND KARAHALIOS, K. Quantifying Search Bias: Investigating Sources of Bias for

- Political Searches in Social Media. In *ACM conference on Computer Supported Cooperative Work* (Portland, Oregon, USA, Feb. 2017).
- [48] LAMBRECHT, A. AND TUCKER, C. E. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. (2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260.
- [49] LECUYER, M., DUCOFFE, G., LAN, F., PAPANCEA, A., PETSIOS, T., SPAHN, R., CHAINTREAU, A., AND GEAMBASU, R. XRay: Enhancing the Web's Transparency with Differential Correlation. In *USENIX Security Symposium* (San Diego, California, USA, Aug. 2014).
- [50] LECUYER, M., SPAHN, R., SPILIOPOULOS, Y., CHAINTREAU, A., GEAMBASU, R., AND HSU, D. Sunlight: Fine-grained Targeting Detection at Scale with Statistical Confidence. In *ACM Conference on Computer and Communications Security* (2015).
- [51] LIU, Y., KLIMAN-SILVER, C., KRISHNAMURTHY, B., BELL, R., AND MISLOVE, A. Measurement and Analysis of OSN Ad Auctions. In *ACM Conference on Online Social Networks* (Dublin, Ireland, Oct. 2014).
- [52] Marketing API. <https://developers.facebook.com/docs/marketing-apis/>.
- [53] Neilson DMA® Regions. <https://www.nielsen.com/intl-campaigns/us/dma-maps.html>.
- [54] NEUMANN, N., TUCKER, C. E., AND WHITFIELD, T. How Effective is Black-box Digital Consumer Profiling and Audience Delivery?: Evidence from Field Studies. *Social Science Research Network Working Paper Series* (2018).
- [55] New Targeting Tools Make Pinterest Ads Even More Effective. <https://business.pinterest.com/en/blog/new-targeting-tools-make-pinterest-ads-even-more-effective>.
- [56] PARRA-ARNAU, J., ACHARA, J. P., AND CASTELLUCCIA, C. MyAdChoices: Bringing Transparency and Control to Online Advertising. *ACM Transactions on the Web* 11 (2017).
- [57] Pinterest: Audience targeting. <https://help.pinterest.com/en/business/article/audience-targeting>.
- [58] ROBERTSON, R. E., JIANG, S., JOSEPH, K., FRIEDLAND, L., LAZER, D., AND WILSON, C. Auditing Partisan Audience Bias within Google Search. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction* (2018).
- [59] SAEZ-TRUMPER, D., LIU, Y., BAEZA-YATES, R., KRISHNAMURTHY, B., AND MISLOVE, A. Beyond CPM and CPC: Determining the Value of Users on OSNs. In *ACM Conference on Online Social Networks* (Dublin, Ireland, Oct. 2014).
- [60] SANDVIG, C., HAMILTON, K., KARAHALIOS, K., AND LANGBORT, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *International Communication Association Conference* (2014).
- [61] SAPIEZYNSKI, P., ZENG, W., ROBERTSON, R. E., MISLOVE, A., AND WILSON, C. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web* (San Francisco, California, USA, May 2019).
- [62] Showing Relevance Scores for Ads on Facebook. <https://www.facebook.com/business/news/relevance-score>.
- [63] SPEICHER, T., ALI, M., VENKATADRI, G., RIBEIRO, F. N., ARVANITAKIS, G., BENEVENUTO, F., GUMMADI, K. P., LOISEAU, P., AND MISLOVE, A. On the Potential for Discrimination in Online Targeted Advertising. In *Conference on Fairness, Accountability, and Transparency* (New York, New York, USA, Feb. 2018).
- [64] SWEENEY, L. Discrimination in online ad delivery. *Communications of the ACM* 56, 5 (2013).
- [65] Twitter: Ad targeting. <https://business.twitter.com/en/targeting.html>.
- [66] Upturn Amicus Brief in *Onuoha v. Facebook*. <https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf>.
- [67] VENKATADRI, G., LIU, Y., ANDREOU, A., GOGA, O., LOISEAU, P., MISLOVE, A., AND GUMMADI, K. P. Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface. In *IEEE Symposium on Security and Privacy* (San Francisco, California, USA, May 2018).
- [68] VENKATADRI, G., LUCHERINI, E., SAPIEZYNSKI, P., AND MISLOVE, A. Investigating sources of PII used in Facebook's targeted advertising. In *Privacy Enhancing Technologies Symposium* (Stockholm, Sweden, July 2019).
- [69] VENKATADRI, G., MISLOVE, A., AND GUMMADI, K. P. Treads: Transparency-Enhancing Ads. In *Workshop on Hot Topics in Networks* (Redmond, Washington, USA, Nov. 2018).
- [70] VENKATADRI, G., SAPIEZYNSKI, P., REDMILES, E. M., MISLOVE, A., GOGA, O., MAZUREK, M., AND GUMMADI, K. P. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *International World Wide Web Conference* (San Francisco, California, USA, May 2019).
- [71] What it means when your ad is pending review. <https://www.facebook.com/business/help/204798856225114>.
- [72] WILLS, C. E. AND TATAR, C. Understanding What They Do with What They Know. In *Workshop on Privacy in the Electronic Society* (2012).
- [73] eyeWnder_Experiment. <http://www.eyewnder.com/>.