Potential for Discrimination in Online Targeted Advertising

Till Speicher TSPEICHER@MPI-SWS.ORG *MPI-SWS*
Muhammad Ali MUALI@MPI-SWS.ORG *MPI-SWS*
Giridhari Venkatadri VENKATADRI.G@HUSKY.NEU.EDU *Northeastern University*
Filipe Nunes Ribeiro FILIPERIBEIRO@DCC.UFMG.BR *UFOP, UFMG*
George Arvanitakis GEORGE.ARVANITAKIS@EURECOM.FR *MPI-SWS*
Fabrício Benevenuto FABRICIO@DCC.UFMG.BR *UFMG*
Krishna P. Gummadi GUMMADI@MPI-SWS.ORG *MPI-SWS*
Patrick Loiseau PATRICK.LOISEAU@UNIV-GRENOBLE-ALPES.FR *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG* and *MPI-SWS*
Alan Mislove AMISLOVE@CCS.NEU.EDU *Northeastern University*

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

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

Keywords: Discrimination, advertising, Facebook

1. Introduction

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

The targeted advertising ecosystem comprises of (i) *advertisers*, who decide which users an ad should (not) be shown to; (ii) *ad platforms*, such as Google and Facebook, that aggregate data about their users and make it available to advertisers for targeting; and (iii) *users* of ad platforms that are consumers of the ads. The potential for discrimination in targeted advertising arises from the ability of an advertiser to use the extensive personal (demographic, behavioral,

and interests) data that ad platforms gather about their users to target their ads. An intentionally malicious—or unintentionally ignorant—advertiser could leverage such data to preferentially target (i.e., include or exclude from targeting) users belonging to certain sensitive social groups (e.g., minority race, religion, or sexual orientation).

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

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

1. Unfortunately, Facebook was found half a year later to still accept discriminatory ads, despite the fixes it claims were put in place (Angwin et al., 2017a).
2. https://www.facebook.com/business/help/170456843145568
3. https://www.facebook.com/business/help/164749007013531
4. https://business.twitter.com/en/targeting/tailored-audiences.html
5. https://business.pinterest.com/en/blog/new-targeting-tools-make-pinterest-ads-even-more-effective
6. https://business.linkedin.com/marketing-solutions/ad-targeting/matched-audiences
7. https://support.google.com/youtube/answer/2454017

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

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

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

2. Quantifying Ad Discrimination

Next, we begin by outlining different ad targeting methods offered by Facebook. We then discuss the current approach to determining whether an ad is discriminatory and argue why it is inadequate. Finally, we propose a new and intuitive approach to quantify discrimination.

2.1. Methods for targeted ads

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

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

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

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

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

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

2.2. Quantification approaches

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

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

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

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

2.3. Outcome-based discrimination

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

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

Our key observation is that ad targeting, like recruiting, involves selecting the *target audience* (**TA**) from a much larger pool of *relevant audience* (**RA**). The relevant audience of an ad is the set of *all* users in the database **D** who would find the ad useful and interesting and thus might interact with it. Intuitively, the discrimination measure should capture the extent to which the target audience selection is *biased* based on sensitive group membership of relevant users.

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

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

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

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

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

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

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

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

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

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

3. PII-based Targeting

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

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

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

3.1. Potential for discrimination

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

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

3.2. Public data sources

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

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

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

3.3. Discriminatory audience creation

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

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

We then examine how many of these records match to Facebook accounts that can be targeted with advertisements, and then evaluate whether the created audiences are indeed discriminatory.

9. https://www.catalist.us
10. http://dl.ncsbe.gov/index.html?prefix=data/

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

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

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

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

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

Results The results of this experiment are shown in Table 1, and we make a number of interesting observations. *First*, the fraction of voter records that are *Targetable* (*i.e.*, online on a daily basis) is both significantly high (over 65% for most audiences we create) and fairly consistent across custom audiences. The only notable outliers are the Age (55+) audience, with only 57% matching.

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

11. Facebook previously defined the potential reach as "the number of daily active people on Facebook that match the audience you defined through your audience targeting selections."

significantly higher than the fraction of the voter records with the sensitive attribute. Taken together, our results show that advertisers can exploit public records to easily target discriminatory advertisements to a large number of people.

3.4. Summary

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

4. Attribute-based Targeting

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

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

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

4.1. Potential for discrimination

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

These findings raise several questions about the potential for discriminatory targeting using Facebook's curated and free-form attributes. First, given that ethnic affinity-based targeting was disallowed for its potential correlation with ethnicity (race) of users, are there other demographic, behavioral, or interest attributes that are similarly correlated, if not more? Second, given that there exist hundreds of thousands of free-form attributes, can malicious advertisers find *facially neutral* free-form attributes that disproportionately target or exclude users of a sensitive group. For example, an advertiser seeking to create an audience excluding certain ethnic groups may choose to select her target audience from users interested in particular news media sites or magazines.

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

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

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

4.2. Discriminatory audience creation

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

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

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

Free-form attributes: We begin our investigation by gathering an extensive (though not exhaustive) list of free-form attributes that are supported by the Facebook marketing API.[13] The API provides two useful calls that we exploit: i) given a piece of text, the API provides a list of free-form attributes that *match* the given text; and ii) given an attribute, the API provides a list of other *related* attribute suggestions. For instance, the list of related attributes for 'The New York Times' includes 'The Washington Post', 'The Wall Street Journal', and 'The Economist'.

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

We begin by trying to find attributes from the above set of 240,000 attributes that can be

13. https://developers.facebook.com/docs/marketing-api
14. http://www.listofnewspapers.com/
15. https://www.alexa.com/topsites/category/Top/News/Newspapers

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

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

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

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

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

More worryingly, some free-form attributes allow a malicious advertiser to target people based on their beliefs. Table 4 presents a few sensitive free-form attributes from our dataset along with their potential audience in the U.S. These attributes correspond to a large audience with specific religious beliefs, including 'islam' (5.7M), 'catholic church' (6.5M), and 'evangelicalism' (5.6M). Thus, although it is not possible to target religion using curated attributes, one can use free-form attribute targeting to narrow the audience to people who are interested in a specific religion. Finally, we note that it is possible to target or exclude gay and LGBT users (or people sympathetic to their causes) via attributes like 'LGBT community' (21M), 'Gay pride' (13M), 'Same-sex marriage' (4.2M), as well as groups of vulnerable people, including 'Addicted' (100K), 'REHAB' (450K), 'AA' (50K). While this last set of free-form attributes might be useful, for example, for an advertiser to prevent addicted people to receive ads about alcoholic beverages, a discriminatory advertiser could explicitly exclude them.

Using Facebook's attribute suggestions: We first investigate the free-form attributes suggested by Facebook to better understand the criteria used to select these suggestions. Table 5 shows the suggestions returned by the Facebook Marketing API (right column) given a free-form attribute (left column). We selected attributes associated with news outlets biased towards conservative audience to check whether their respective suggestions are also similarly biased. For instance, almost 80% of the audience of

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

Very Conservative - U.S. Facebook Population (13.9%)	
Input Attribute	**Attribute Suggestions**
Townhall.com (79.5%)	The Daily Caller (67.1%), RedState (84.3%), TheBlaze (59.6%), Hot Air (news site) (79.4%)
The American Spectator (70.7%)	The Daily Caller (67.1%), Townhall.com (79.4%), The American Conservative (85.2%), National Review (78.6%), Weekly Standard (72.2%), Human Events (53.3%), Commentary (34.5%), RedState (84.3%), Harper's Magazine (11.7%), U.S. News & World Report (18.6%)
The Patriot Post (70.7%)	American Patriot (68.4%), Patriot Nation (54.3%), Patriot Update (84.2%), NewsBusters.org (78.7%), Guns & Patriots (61.2%), RedEye (9.1%), America's Conservative Voice (74.4%)
American Thinker (67.5%)	National Review (78.6%), Fox Nation (75.3%)
The Cullman Times (63%)	Montgomery Advertiser (40.4%), The Huntsville Times (40.8%), The Tuscaloosa News (44.8%), al.com (42.5%)

Townhall.com[16] are "very conservative" Facebook users, whereas the average amount of very conservative U.S. users of Facebook is about 13.89%. We note that the audiences corresponding to most of the suggested attributes also exhibit a strong bias towards conservative audience. From all suggestions presented, only Harper's Magazine[17] and RedEye[18] have a less conservative audience in comparison with the U.S. distribution.

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

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

Fox (37%) → Fox News Channel (67%) → Sean Hannity (88%) → Mark Levin (93%) → Rush Limbaugh (93%) → The Rush Limbaugh Show (94%) → The Sean Hannity Show (95%).

4.3. Summary

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

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

5. Look-Alike Audience Targeting

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

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

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

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

5.1. Potential for discrimination

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

5.2. Bias in look-alike audience selection

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

Similar to what we did in Sections 3 and 4, we use the North Carolina voter database to construct several groups of 10,000 randomly selected people based on their ethnicity (Asian, Black, White, Hispanic), gender, political affiliation (registered Democrat or Republican) and age (18-24, 24-35, 35-54, 55+). We construct a source audience corresponding to each group and for each source audience, we ask Facebook to construct look-alike audiences from the US in five percentile ranges: closest matching 2%, 2-4%, 4-6%, 6-8%, and 8-10% of the US population. Note that the audiences in the different percentile ranges do not overlap with one another and each subsequent percentile range becomes less similar (i.e., less closely matching) to the source audience.

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

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

tion reflects the biases in source audience selection.

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

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

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

Knowing the representation bias of each attribute allows us to construct a ranking of attributes from most to least biased; we refer to attributes at the top of the ranking as *over-represented* and those at the bottom to be *under-represented* in a target audience. Table 6 shows the top 5 over-represented and under-represented attributes for the source audience of African Americans and its 2% and 2–4% (the most similar two) look-alike audiences. The table shows that a majority of attributes that were found to be overrepresented in the source audience remain so for the look-alike audiences; similar behavior can be observed for the underrepresented attributes. These results—particularly the presence of multicultural affinity attributes—suggest that Facebook is using its extensive set of attributes to likely infer the biases that we introduced into our source audience. Moreover, it is propagating these biases to the selection of look-alike audiences, constantly over-representing African Americans and under-representing Hispanics and Asian Americans compared to their proportions in the national population.

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

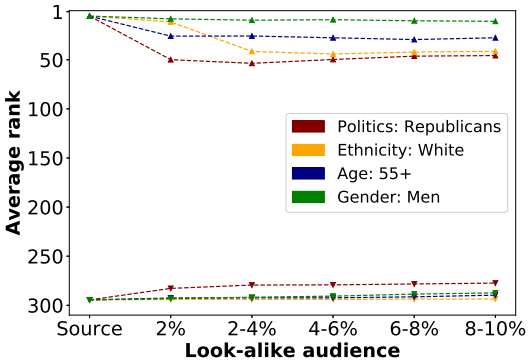

Figure 1: Comparison of the average ranks of top 10 over-represented and under-represented attributes in look-alike audiences built from different types of biased source audiences. Average ranks for over-represented attributes are indicated by upward triangles, downward triangles are used for the average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

5.3. Discriminatory audience creation

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

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

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

sitive attribute was severely exaggerated, we expect the disparity measure to reflect the disparity in favor of the attribute.

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

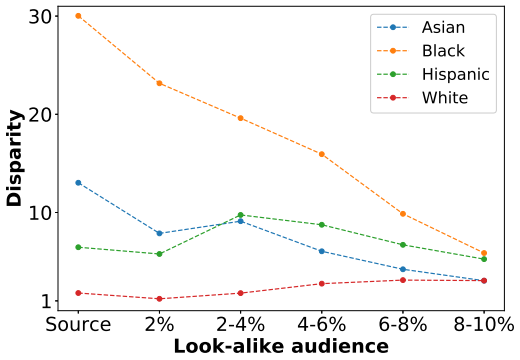

Figure 2: Disparity (in favor) for each ethnicity when the look-alike audiences are created from an audience biased on that ethnicity.

5.4. Summary

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

6. Concluding Discussion

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

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

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

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

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

Acknowledgements

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

References

Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS*, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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