Saarland University
Max Planck Institute for Software Systems

Measuring Bias in Facebook's Ad-Targeting Attributes

Master's Thesis in Computer Science

by

Muhammad Ali

supervised and advised by

Prof. Dr. Krishna P. Gummadi

reviewers

Prof. Dr. Krishna P. Gummadi

Prof. Dr. Christian Rossow

August 3, 2018

UNIVERSITÄT DES SAARLANDES

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Eidesstattliche Erklärung

Ich bestätige, dass diese These meine eigene Arbeit ist, und ich habe kein anderes Material verwendet als die, auf die ich mich beziehe.

Non-plagiarism Statement

I confirm that this thesis is my own work, and I have not used material other than the ones to which I refer.

Einverständniserklärung

Ich stimme zu, meine Arbeit der Öffentlichkeit zugänglich zu machen, indem ich sie der Bibliothek der Informatikabteilung hinzufügen lasse.

Declaration of Consent

I agree to make my thesis accessible to the public by having it added to the library of the Computer Science Department.

Saarbrücken, August 3, 2018.

(Muhammad Ali)

Abstract

Targeted advertisements on the internet today have become indispensable for numerous industries, ranging from corporate marketing to election campaigns. Their rise in popularity has partly been due to their unprecedented precision in reaching the right people, as well as a growing online population. However, despite all their strengths, previous work has identified instances where targeted ads can be biased against certain groups of users – for instance, by failing to present them with opportunities (e.g., job ads) that it might be presenting for a competing group. While identifying instances of bias from an end-user's perspective is important, there is also an increasing need to understand where these biases might be originating from in an advertising system.

In this thesis, we look at the targeting advertising system of Facebook, the world's largest social network. Instead of observing ads, our study goes a step deeper and investigates the *targeting attributes/interests* that the advertising platform provides to marketers. We conduct a systematic measurement study to understand whether the attributes used to reach users are themselves biased against certain socially salient groups (e.g., gender or race). We leverage Facebook's public developer APIs to measure user interests on the social network, as well as a survey (N=300) to understand people's interests in the offline world.

Our study finds Facebook to be less biased against women for several job-related attributes when compared to the survey. We also find evidence of Facebook attributing lower-income users to risky financial interests such as gambling. Thus, we find the differences to be both possibly beneficial and detrimental in different cases. Our study helps better understand the advertising ecosystem, and has implications for policy-makers, researchers and software engineers, who might want to design interventions or build more robust ad systems.

Acknowledgements

I would like to thank my supervisor Krishna Gummadi for both giving me the opportunity to work with him as a student, and for mentoring me as a person. I am grateful that he chose to spend his time and energy on teaching me the importance of discipline, and of focusing on the foundations of every problem before rushing to solve it. Hoping to some day be as clear in my thought process and communication as Krishna is a long way to go, but I'd like to try. I am also deeply indebted to Elissa Redmiles – who has been incredibly kind and helpful in the last stages of my thesis, helping me organize and communicate my results. I am fortunate to have worked with Elissa and to have learnt that it is possible to be driven and kind at the same time.

I am thankful to my friends in Saarbrücken for making my time in the city enjoyable. I am glad I got to meet so many intelligent and helpful people like Prabal, Faraz, Mohammad, Bhaskar, Mihai, Alina, Verica, Bilal, and many others that cannot fit into this one sentence. I am also thankful to Fariha for sharing the student lab with me as I wrote my thesis, and for participating in long rants about academia, Pakistani politics, and the futility of anything and everything.

I am incredibly fortunate to have parents who have always supported me in whatever I have done. Perhaps even more importantly, they have always had my back and let me know that I have the space to fail. I would like to thank my siblings for always looking up to me, and always forcing me to lead by example.

Finally, I would like to thank Tazeen from the bottom my heart – for being there in the most difficult of days, and for being a Bridge Over Troubled Water.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern internet applications often rely on user data – either revealed voluntarily by the users, or inferred through their behavior on the application – to personalize the user experience. Such personalization is increasingly commonplace due to the rise in popularity of machine learning algorithms and the availability of large-scale user data. Social networking sites (Facebook, Twitter), search engines (Google, Bing), content sharing platforms (YouTube, SoundCloud), and even modern news sites (NY Times), are prominent examples.

While this personalization often comes in the form of recommending similar items, applications might aim to infer fine-grained *attributes* about the user as well – such as an interest in politics, video games, or a particular genre of music. In addition to personalization, having precise user attributes also allows these applications to power advertisements on their platform. For instance, Twitter, a popular news-based social network, infers as many as 350 such attributes for each of its users – ranging from interest in certain professions to sports, and many more[1]. Given that Twitter never asks its users to explicitly describe their interests, all of the attributes are essentially *inferences* the social network has made about its users based on the content they post and their network on the website.

The primary use case of these user attributes/features is online advertising. Advertisers can leverage the platform's knowledge of its users to show ads only to people who are already interested in their products, a phenomenon referred to as *online targeted advertising*. Thus, the online advertising ecosystem consists of three parties [27]: (i) *advertisers*, who specify the audience they want to reach;

[1] https://business.twitter.com/en/targeting/interest.html

(ii) *ad platforms*, who aggregate user data and infer their interests; (iii) *users*, who are the consumers of the advertisements.

Due to their complexity and potential impact on users, advertising systems have attracted a considerable amount of attention in research. Previously, work has been done to understand the privacy implications of these systems [7, 18, 19, 30, 31]. There has also been work to understand how advertising systems might themselves be discriminatory [12, 30], or might assist malicious advertisers in launching such campaigns [27]; and even how some of these systems might fail to be fully transparent to the users about why they were shown particular ads [1, 12].

Given that the online ads industry is now bigger than the market for television and print advertising in the U.S. [14], it is important to have a deeper understand of existing advertising systems. Having a better understanding of these systems would also be helpful in designing the next generation of targeted advertising platforms, which are beneficial to users but are less prone to the current transparency and discrimination problems.

1.1 Research Questions

While prior work has demonstrated the existence of demographic biases from an end user's perspective i.e. in the *ad delivery*, there is still room for a systematic understanding of how the attribute inferences inside the system might be biased. In our study, we focus our attention on the less studied problem of how the differences in an advertising system's inferences might lead to a bias in advertising. In particular, we ask:

How do the attribute inferences made by a targeted advertising platform change for different demographic groups?

While differences in ad delivery are important to understand how online ads might be biased, our motivation is to look one step deeper, and inside the advertising system itself. An advertising platform whose targeting attribute inferences might be disproportionate for different subgroups of the population could unknowingly discriminate amongst them. For instance, a platform that infers much more men as interested in engineering jobs than women could disadvantage women for STEM job postings; or if the system does not associate high paying jobs with certain ethnic groups, it might be limiting their opportunities. We aim to operationalize and identify such instances of disproportionate inferences in our study.

We focus on Facebook, the world's largest social network [17], because of its high adoption and the popularity of its ads. Facebook has a matured and powerful advertising platform that allows sponsoring content on the social network itself and across its family of applications such as Instagram and Messenger. Figure 1.1 shows an example ad on Facebook that advertises multiple music albums on Amazon.

We choose to do our analyses by demographic groups and not individual users as our concentration is on systematic instance of biases for a group of the population. Moreover, Facebook's ad *reach estimates* form the core measurements we take from the platform. These are aggregated estimates provided by the social network about how many people an ad campaign can reach. These aggregate estimates provide a natural analog for studying the ad platform's inferences for various demographic groups.

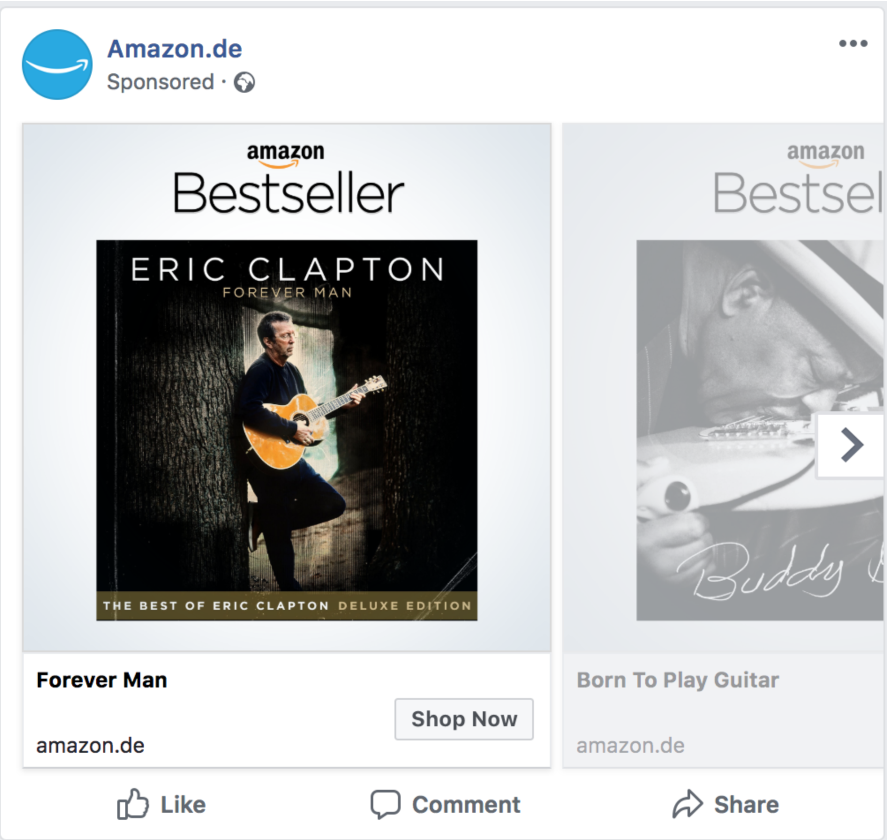


Figure 1.1: Example of an advertisement on Facebook.

To be succinct, we try to answer the following questions in our study:

1. How do Facebook's ad attribute inferences vary across different demographic factors such as gender, ethnicity, and income?
2. Are these differences relatively more or less disparate than what we observe in the real world?
3. How could these differences impact the ads on Facebook, and eventually users on the social network?

1.2 Contributions and Outline

To understand Facebook's attribute inferences, we take measurements of several of its targeting attributes across a variety of demographics. We also later choose a subset of important targeting attributes related to jobs and finances to do fine-grained analyses. To compare the ad platform's inferences with people's actual interests, we conduct a user survey and gather our respondents' interests in different jobs and financial categories.

Our study finds that many ad targeting attributes have noticeable demographic associations, not all of which are troubling. We discuss how these disparities in inferences might give us insight into the users' internet-use patterns as well as how advertisers use the targeting options. We also find that for professional targeting attributes, Facebook surprisingly has less disparity between men and women, and White and Black ethnicities than what our survey suggests. For the financial targeting attributes, we find patterns that might be concerning: debt-related and risky financial products such as gambling and credit cards are more popular with lower income users, while beneficial categories like investment have no such biases.

Our study helps form a better understanding of one the internet's biggest advertising platforms and its internal workings. Characterizing the demographic associations of ad targeting attributes is important for policy makers and software engineers who might want to design interventions in problematic instances, as well as for end-users, who are most affected by these ads. We also believe our study could be beneficial for the research community at large who are interested in using demographic and interest estimates from Facebook ads for various tasks.

The rest of this thesis document is organized as follows. Chapter 2 discusses the background concepts needed for our discussion and reviews important literature in two areas: (i) understanding the potential of bias and discrimination in online advertising, and (ii) using targeted advertising data for computational and social tasks. Chapter 3 discusses in detail the structure of Facebook's advertising platform, the scale of its penetration in a developed country like the United States, the data we collect from the platform, and the techniques we use to collect and analyze our data. Chapter 3 also describes the structure and the purpose of our user survey. In Chapter 4, we present our findings related to possible biases in Facebook's inferences, and their comparison with responses from our user survey. We begin our discussion by characterizing the differences on a high level, and then focus on specific attributes related to jobs and finances. Chapter 5 summarizes our discussion, discusses the implications of our study and sets an outlook for future exploration of the problem space.

Chapter 2

Background and Related Work

2.1 Background

2.1.1 Online Social Networks

Online social networks refer to websites or internet application where users are able to share content such as text posts, images and videos with a network of connected users [21]. Connections in the network could either be uni- or bi-directional depending on the kind of community the network wants to craft. Twitter, a popular news based social network, functions on uni-directional connections where users are able to *follow* other users on the network and keep up with the information they post. Facebook, the world's largest social network, has a primarily bi-directional structure, where users of the website can establish *friend* relationships and gain access to each others' posted content.

Social networks form an interesting component of the modern internet because of the amount of content shared by their users on these platforms. With volumes of data about public likes and dislikes being created every second, social networks have emerged as powerful data sources of both individual and public opinions.

The scale of social networks such as Facebook and Twitter, and often the ease of accessing data from them makes them a useful tool for research in problems that involve understanding social processes.

2.1.2 Online Targeted Advertising

Targeted advertising refers to choosing an advertisement's audience based on their demographics, behaviors or prior buying history. By online targeted advertising, we refer to targeted advertising campaigns on the internet.

Due to the volumes of user data collected and aggregated by online social networks, they are able to deeply understand the characteristics of their users. This data can, in turn, be used to accurately target ads to users of the social network. Facebook[1], Twitter[2] and LinkedIn[3], all provide elaborate targeted advertising tools to marketers.

It is important to note that the practice of targeted advertising isn't simply limited to the internet. Leaflet marketing is also a form of targeted advertising where the users are targeted based on their presence at an event, or the neighborhood they live in. The internet simply allows these campaigns to run on a much bigger scale.

Similarly, internet applications that might not necessarily be social networks, but collect data about their users also might offer advertising options. For example, Google, the popular search engine, has an extensive advertisement network across its family of applications. Conventional e-mail marketing is also one of the most common forms of online targeted advertising.

However, our study limits its scope to targeted advertising on social networks – and in particular, to the world's largest social network, Facebook [17].

2.2 Related Work

Here, we review prior literature that documents instances of targeted advertising systems being biased against a particular demographic. Further, the Facebook marketing data that we employ in our study has gained considerable traction in the research community as a useful data source. We also review literature that has leveraged Facebook's advertising data for a variety of tasks, from demography to computational social science.

[1] https://www.facebook.com/business/products/ads
[2] https://business.twitter.com/en/solutions/twitter-ads.html
[3] https://business.linkedin.com/marketing-solutions/ads

2.2.1 Bias and Opacity in Targeted Advertising

It has previously been shown that searching for conventionally African-American sounding names such as DeShawn, Darnell and Jermaine results in a higher likelihood of Google's AdSense platform showing advertisements related to public arrest records and background-checking websites [30]. After conducting thousands of web searches with full names of American people, Sweeney [30] was able to show statistically significant differences in both the number of background check ads, as well as ads with the word "arrest" in them when searched with racially associated Black names.

Datta et al. found that women were served fewer ads related to career coaching than men, on Google's ad network [12]. They simulate ad delivery using automated browser-based agents, and observe how changing user profile attributes for the agents result in differentiated ad delivery.

More recently, Till et al. studied how malicious advertisers could use the ad targeting tools provided by Facebook to launch discriminatory ads that aim to exclude certain socially salient groups and/or overrepresent others [27]. The study highlights multiple attack vectors that an advertiser could exploit to build increasingly exclusionary audiences, even when Facebook might put interventions to stop such behavior.

In addition to biases and discrimination in ad delivery, researchers have also worked to highlight the lack of transparency in modern advertising systems as well, where explanation mechanisms for these ads fail to be entirely truthful. Owing to people's concerns and possible unease about marketing on the web [31], social networking and content platforms have started providing tools that let users understand the "inferences" the ad platform has made about them. For example, Google provides the Ad Settings page [29], while Facebook has build features such as Ad Preferences[4] and "Why am I seeing this?" [10].

Datta et al. show in their study how despite significant changes in ads shown after visiting webpages related to substance abuse and disability, Google's Ad Settings page did not reflect that the platform had learnt new attributes about the user [12]. This opacity in Google's ad settings has been replicated in other studies [33] as well.

[4] https://www.facebook.com/ads/preferences/

Similar results have also been found in Facebook's explanation of ads [1], where it has been shown that Facebook has a tendency to provide incomplete explanations. In a majority of cases, the ad explanations have been shown to unspecific – for instance, only revealing that an ad was shown because of a liked page but never mentioning which page.

Such issues with hidden biases and the lack of transparency in these ad platforms motivate us to look under the hood and identify the origins of such biases. Rather than characterizing bias from an end-user perspective with possibly incomplete explanations, we look into the targeting attributes used by advertisers themselves.

2.2.2 Uses of Targeted Advertising Data

In addition to studying possible problems with ad targeting systems, prior work has also found ways to put this novel data source to use in a variety of tasks. Targeted advertising data from platforms such as Google and Facebook has been used to solve problems in areas as diverse as demography [24, 34] and public health [4, 20, 26].

Advertisement estimates from Facebook have been employed to monitor stocks of migrants in the United States [34]. Since the social network allows targeting ads to expatriates from different countries, Zagheni et al. leveraged the estimates for the number of expats provided by the ad platform to study migrants from different countries to the United States. Estimates from Facebook's ads platform have also been used to estimate male fertility rates for multiple countries [24]. By asking the ad platform for the number of parents who have had a child in the last 12 months (another one of Facebook's allowed ad targeting options), the authors were able to approximate metrics for male fertility that correspond to ground-truth data.

Ad targeting data from the popular professional networking site LinkedIn has also been shown to be useful in identifying gender gaps in the workforce [15]. Similar to Facebook, LinkedIn also provides targeted advertising options based on several features, including gender, location, age and field of work. Haranko et al. [15] were able to use estimates from the ad platform to characterize how gender gaps in different professions vary across cities in the United States – establishing LinkedIn's targeted advertising system as a useful data source for such problems.

Advertisement data has also been shown to be useful in problems related to public health. Saha et al. [26] have shown how estimates from Facebook's ad targeting system can be used to characterize mental health awareness for a disorder like schizophrenia. By establishing an index for schizophrenia awareness, the authors were able to use Facebook's ad platform to measure the index for a variety of demographic groups such as different ages or education levels. Other work has also made the case for using advertisement estimates to monitor lifestyle diseases and disorders, such as diabetes or alcoholism, in real-time [4, 20] – highlighting the potential as well as the limitations of the data source.

We take inspiration from, and build on top of the work that has been done before us in terms of tapping into ad targeting data. We believe our study helps build a better understanding of Facebook's ad targeting system, and has implications for the research community at large that employs such data in their studies.

Chapter 3

Data and Methods

In this chapter, we broadly explain the structure of Facebook's advertisement ecosystem, the data we gather from the system, and the techniques we use to analyze it.

3.1 Facebook's Advertisement Ecosystem

As the world's largest social network [17], Facebook's targeted advertising system allows marketers and businesses to target ads to its users by their demographic attributes, location, interests and internet behaviors. Prior to launching an ad, Facebook is also able to report how many people would fit the specified criteria and view the campaign – which it refers to as the *reach estimate*.

The ability of the platform to a) simultaneously target users by both their demographic and behavioral attributes; and b) to provide reach estimates prior to actually launching the ads, makes it a useful tool for understanding the characteristics of a population. It also makes it a powerful tool for understanding what Facebook *thinks* the characteristics of a population are.

We are able to use the system for very specific queries such as "the number of women between the ages of 25 and 29 interested in online shopping" or "the number of African-American men managing small business pages on Facebook", and then compare the given reach estimates across demographic groups – all without ever launching an ad.

On a high level, Facebook categorizes its ad targeting options into three major categories[1]:

1. **Core Audiences:** Here, Facebook allows marketers to choose the features of the population they want to target. These features range from demographic variables such as gender and relationship status to interests in things such as anime movies or hip hop music. Section 3.1.2 discusses the available options in detail.
2. **Custom Audiences:** This feature allows marketers to upload a list of personally identifiable information (PII), which is then matched to accounts on Facebook for targeted advertising.
3. **Lookalike Audiences:** In lookalike targeting, Facebook is able to expand on an initial list of people provided by the marketer (for instance, through a PII list or through a prior ad campaign). Facebook builds the target audience by looking for people with similar interests and features to the ones provided.

We primarily focus on core audiences, as our goal is to study how the targeting attributes provided for such camapigns are inferred differently for different demographic groups.

3.1.1 Scale of Facebook's Penetration

As of the time of this writing, Facebook's advertiser interface reports that 240 million people who live in the United States are targetable with Facebook ads. To understand the scale of this penetration, we stratify this estimate by each county in the U.S. and observe what fraction of the population in each county is targetable. Figure 3.1 shows the fraction of population in each county that is targetable with Facebook's ads. Estimates for each county's total population have been obtained from the U.S. Census Bureau. Since Facebook doesn't provide location targeting by a particular county, we get the population of each county from Facebook by specifying all ZIP codes in the county.

We find that the median penetration over all counties is 59.10%. We also observe that there are no noticeable gaps in access; coastal states (California, Washington, Northeastern states) have higher percentages of their population on Facebook,

[1] https://www.facebook.com/business/products/ads/ad-targeting

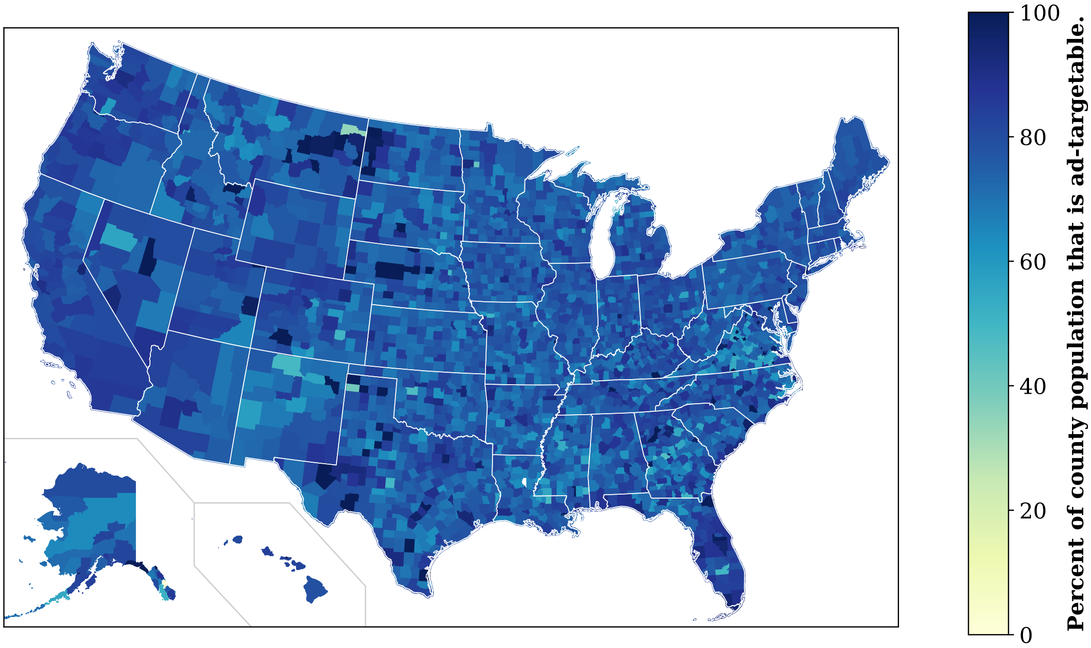


FIGURE 3.1: County-level penetration of Facebook advertisements in the United States.

while states in the South (Virginia, North Carolina, Georgia) frequently have counties with lesser online population. Overall, however, the penetration of Facebook's advertising system looks promising, with no clusters regions that are offline.

This helps illustrate the scale of the platform in a country like the U.S. where a large fraction of the population is reachable and can be studied with advertisement data. Observing penetration at the level of each county also helps us understand on a fine-grained level which people constitute the sample of our study.

3.1.2 Available Targeting Options

Figure 3.2 shows the structure of Facebook's *Ads Manager* page, where marketers can construct audiences for advertisement campaigns. We see that the interface provides the option to target by location, age, gender, language and a wide array of options under the "Detailed Targeting" section.

For location based targeting, Facebook provides the option to target countries, regions comprising multiple countries (e.g. Asia or the European Economic Area), regions within countries such as provinces or states, cities, ZIP codes, as well as congressional districts in the United States. It also allows radius targeting by

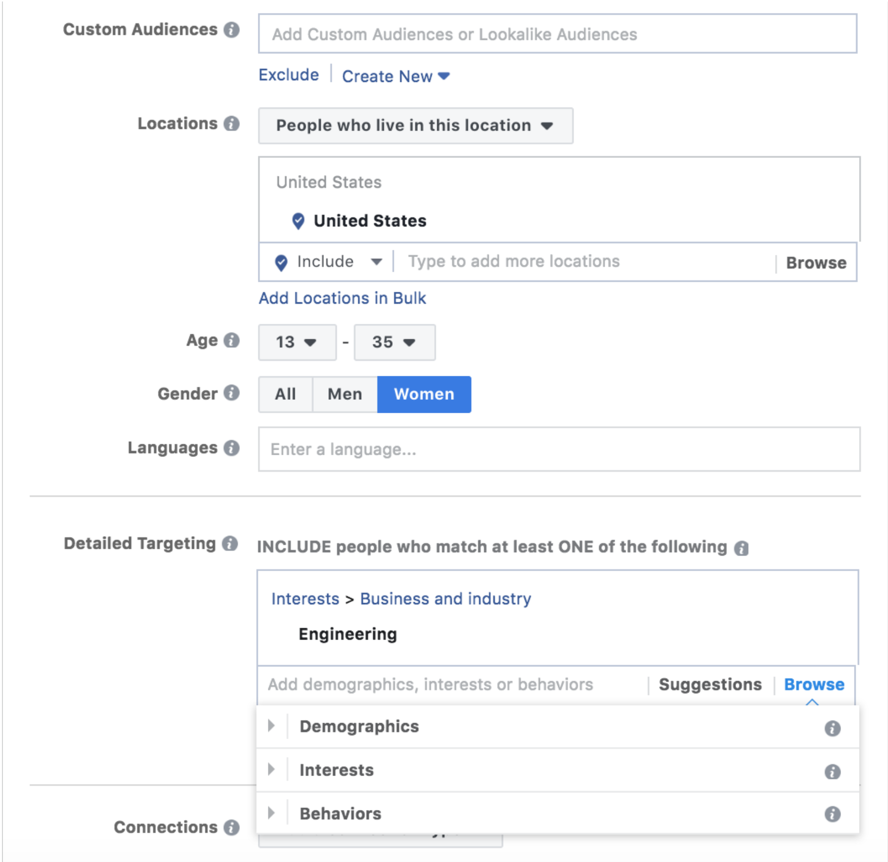


FIGURE 3.2: Screenshot of Facebook's Ads Manager interface showing some of the major functionalities available. Geographical location, age and gender have been specified. An elaborate list of targeting options are available through the cascading drop-down menus under "Detailed Targeting". Here, the category "Engineering" has been selected.

dropping a pin at a particular address and specifying a target radius as small as 1 kilometer around it. For all these options, advertisers are able to target people who either live in the area, have recently traveled there, are currently traveling, or all of the above.

In selecting age, the interface permits arbitrary age ranges with the minimum and maximum age options. Since the social network's minimum allowed joining age is 13, the platform doesn't allow targeting pre-teens.

Under detailed targeting, the platform classifies attributes into three broad categories: demographics, interests and behaviors. Targeting features such as education level, relationship status and employment are mostly grouped under demographics, while interests and behaviors often contain lifestyle choices and internet use behaviors respectively. Within each category, many targeting features are grouped together by themes such as "Fitness and wellness", "Hobbies and activities" etc. However, the demarcation isn't very strict and it is sometimes unclear why a particular feature was grouped under a category. For example, expat and ethnicity targeting attributes are put under behaviors even though they describe demographic traits.

We thus find it more useful to peruse the targeting attributes and identify the ones that are important for a particular study, instead of being guided by how the interface groups them.

It is also important to note that the list of these targetable features is not static. It is subject to moderation if the users report a feature in the list as inappropriate. As an example, this public moderation practice has previously led to Facebook renaming the ethnic affinity feature [2] and removing problematic anti-semitic ad categories that were automatically indexed by the ad system [3]. Similarly, many of the attributes grouped under behaviors are "Partner Categories" obtained through external data broker firms in different countries, such as Acxiom, Experian etc. Facebook plans to discontinue use of these categories in the interest of its users' privacy [22]. These regular changes to the features warrant caution while collecting data from the platform.

In addition to the features under these three categories, Facebook also automatically indexes other interests across the platform which advertisers can search by inputting free text. Many of these features have been observed to be directly related to pages on the social network and are explained in the Ads Manager as targeting "people who have expressed an interest in" or "like pages related to" the particular feature.

Further, the interface permits splitting the detailed targeting option into three sub-fields, allowing for different mechanisms of adding targeting features:

1. **Include**: The default field as shown in Figure 3.2 shows the include option. Features added to this field are combined together with a logical OR operation to match the audience. Each subsequent feature added here acts to expand the audience.

2. **Narrow**: Unlike the include option, features are combined in this field with a logical AND operation. Features added to this field serve to refine and narrow the audience.

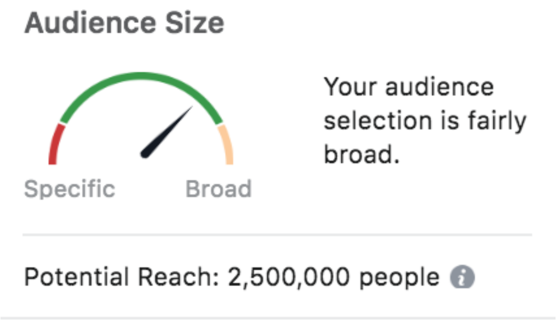


FIGURE 3.3: Size of targetable population (reach estimate) with the targeting options selected in Figure 3.2 i.e. women living in the United States, between the ages of 13 and 35, who are interesting in engineering.

3. **Exclude**: People who match features listed under the exclude section are explicitly excluded from the ad's audience.

Using the three functions, marketers are able to construct arbitrarily complex and niche audiences. For instance, a targeting combination such as:

$$(\text{electrical engineering} \lor \text{software engineering}) \land \text{mathematics} \land \neg\text{design}$$

is perfectly possible and permitted with the current targeting mechanism.

Throughout the process of tweaking the audience by selecting different targeting options, Facebook reports the number of people accessible with the current selection. This is shown under *Potential Reach* in the Ads Manager, and is shown in Figure 3.3. These reach estimates form the foundation of our study and are the primary measurements we take from the online ads network.

It is also important to note that in order to preserve user privacy, the reach estimates are rounded. The exact number of people matching a targeting criteria is never revealed to the advertiser. We note in our study that the platform does not provide estimates less than 1,000 – any estimate lesser than the threshold is rounded up to a thousand. For estimates greater than the threshold, the potential reach is rounded by different step sizes depending on the reach's magnitude [32]. This might not affect large scale studies but could limit our methodology's applications to geographical areas with smaller populations.

3.1.3 The Marketing API

To encourage developers to programmatically create and manage ad campaigns, Facebook has provided public Application Programming Interfaces (APIs).

By making a developer account on the platform and setting up an application with appropriate permissions, we are also able to leverage these APIs for research purposes. In particular, we utilize the Marketing API to specify combinations of targeting features, demographic attributes, and geographical locations – in return for the reach estimates for these combinations.

We also take advantage of the Targeting Search API – the same API that is used to search Facebook's database of features when an advertiser inputs free text. By sending an empty string query, we are able to retrieve all features grouped under interests and behaviors from the platform.

Facebook also provides convenient Software Developer Kits (SDKs) for different programming languages for abstractions in access to the APIs. We make use of the Facebook Business SDK for Python[2] to make queries to the Marketing and Targeting Search APIs.

3.1.4 Data Collected

We collect both demographic and advertisement related estimates from the Marketing API. Demographically, we choose to focus on gender, ethnicity and annual income; while for the advertisement estimates, we query the API for the interest each demographic group shows in the ad-targeting features. We limit our estimates to the United States as it is a high internet penetration country, and also the only region where Facebook explicitly infers ethnicity – or "Multicultural Affinity" as the Ads Manager calls it.

Demographics: For gender, we ask Facebook for the number of men and women in the U.S. For ethnicity, we query for the number of White, Asian, Hispanic and African Americans in the country. Since Facebook doesn't have an explicit White ethnicity attribute, we obtain this estimate by excluding all other ethnic groups recorded by the platform. Annual household income is a data-broker attribute

[2] https://github.com/facebook/facebook-python-business-sdk

with pre-defined income ranges on the platform – we also collect estimates for the number of people in each of these income bins.

Targeting Features: Once we have the list of targeting attributes from the Targeting Search API – 323 interests and 264 behaviors, we are able to iterate over them and ask the Marketing API for a) the number of people interested in the attribute; and b) the number of people from each demographic group (gender, ethnicity and income) interested in the attribute. The latter estimate is made with a logical AND (the narrow option) in the API.

3.2 Methods for Analysis

Once we have collected data from the marketing API, we begin evaluating which targeting features have the highest affiliations with which subgroups of the population.

3.2.1 Quantifying Differences in Interest

To understand how strongly the ad platform associates a demographic with a particular targeting feature, we ask three fundamental questions:

1. What proportion of people in a given demographic group have a particular interest or behavior?
2. Are members of a particular demographic group more likely to be inferred as having this interest or behavior than the general population?
3. Are members of a particular demographic group more likely to be inferred than those in another demographic group?

The first question is straightforward to answer. We refer to the fraction of a group's population that is interested in an attribute as the *penetration* of the attribute in the group. For a demographic d and feature f on the platform, we define penetration as

$$\text{penetration}_g(d, f) = \frac{n_g(d, f)}{n_g(d)}, \tag{3.1}$$

where $n_g(d)$ is the reach estimate (i.e. number of people targetable) by specifying demographic d alone on the Ads Manager. $n_g(d, f)$ refers to the intersection of populations targetable with demographic d and feature f. Because of the nature of the ad ecosystem, a geographical location of the audience must always be specified, which is shown here as g.

Equation 3.1 gives us a simple statistic for the association of each demographic in our dataset to the attributes on the platform. To answer the second question of whether a group is more or less likely to be inferred than the general population, we build on top of penetration and define the notion of *affinity*. We define the affinity for a demographic group d towards a targetable feature f as

$$\text{affinity}_g(d, f) = \frac{n_g(d, f)}{n_g(d)} - \frac{n_g(f)}{N_g}. \tag{3.2}$$

Following similar notation, here $n_g(f)$ refers to the reach estimate for feature f without specifying any demographic; N_g denotes total population in region g according to Facebook. All estimates involved in these computations are obtained with the data collection process described in Section 3.1.4.

A large positive affinity would indicate that members of demographic d are considered much more likely by the ad platform to be interested in feature f, as compared to the general population in region g. Analogously, a large negative value would mean Facebook does not associate group d with the feature f and some other subgroup of the population in g is more highly interested.

For answering the third question of whether one demographic is more likely than another to be inferred for an attribute, we compare the affinity of the attribute for both groups. We refer to the difference in affinity (or alternatively, penetration) for two groups d_1 and d_2 as their *disparity* on targeting feature f,

$$\begin{aligned} \text{disparity}_g(d_1, d_2, f) &= \text{penetration}_g(d_1, f) - \text{penetration}_g(d_2, f) \\ &= \frac{n_g(d_1, f)}{n_g(d_1)} - \frac{n_g(d_2, f)}{n_g(d_2)}. \end{aligned} \tag{3.3}$$

Naturally, for two groups d_1 and d_2 that have large disparity for a feature f, Facebook has a very different understanding of their interest in f. As a result, a higher fraction of the group with the larger affinity might end up seeing content related to f. Moreover, targeted ads also present the opportunity to expand these differences. By disproportionately showing ads for attributes that might be relevant for both demographics, disparities might reinforce themselves over time. Therefore, having a notion of disparity between two demographics allows us to observe the differences Facebook believes these groups have.

Using this framework, we are able to answer questions like whether there are significant differences between men and women in Facebook's inference of professional features such as engineering; or what kind of industries the ad platform associates most with African Americans. Perhaps most importantly, we are able to do this without launching any malicious or harmful ads and only through the reach estimates provided in the Ads Manager.

3.2.2 Statistical Testing

One of the questions we frequently ask in our study is whether any difference in two proportions – for instance, the penetrations of an attribute for two groups – is statistically significant. We make use of the 2-sample χ^2 test for equality of proportions to answer these questions. The test has a null hypothesis that the two proportions p_1 and p_2 being compared are equal i.e.

$$H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2,$$

and is conveniently available as the `prop.test()` function in R, or as `chi2_contingency()` in Python's `scipy` package.

Because we conduct our analyses on the level of a country's population, we frequently deal with large sample sizes – situations where a χ^2 proportion test might be too sensitive. In such situations, we use the proportion test in conjunction with a measure of effect size. We use Cohen's h, which is a measure of distance between two proportions or probabilities. We use the nondirectional variant of the statistic, which measures the difference between two proportions p_1 and p_2 as

$$h = \left|2\left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right)\right|. \quad (3.4)$$

As a rule of thumb [11], $h = 0.20$ is seen as a small effect size, while $h = 0.50$ and $h = 0.80$ are seen as medium and large effect sizes respectively. While the proportion test is able to tell us whether differences in our proportions are statistically significant, it does not describe the size of these differences. Having a notion of the effect size allows us to put the result of the proportion test in perspective, and determine which results are meaningful, or *practically significant* [13].

3.3 Survey for User Interests

In addition to collecting data from Facebook's APIs, we also conduct an online survey ($N = 300$) as an alternate source for opinion data. We design the survey on the Qualtrics platform[3], and do our census-representative fielding through a survey sampling firm.

Having survey data in addition to Facebook's estimates allows us to have ground-truth data for how each demographic is interested in some of Facebook's targeting attributes. Since Facebook has an overwhelming amount of targeting features, we only pick a subset of them related to jobs and finances to ask our survey respondents (motivation explained further in Section 4.3).

We administer our survey to a census-representative sample of 300 respondents in the United States. On a high level, we ask respondents about their interest in pursuing a job in multiple areas. We also question them about their interest in different financial products such as credit cards, mortgage loans, as well as their interest in gambling. Each respondent was also asked for their gender, age, ethnicity and race, as well as annual income. Demographics for the respondents were recorded at the end of the survey to avoid bias in the opinion questions.

Questions from the survey, and a demographic breakdown of our survey respondents is given in Appendix A.

[3] https://www.qualtrics.com/

Chapter 4

Results and Insights

In this section, we discuss the results we find in our analysis of Facebook's targeting attributes. We begin by observing the quality of Facebook's demographic estimates in Section 4.1. Section 4.2 characterizes on a high level how the ad platform's attribute inferences differ for different demographic groups. Later in Section 4.3, we specifically focus on job- and finance-related targeting attributes, and compare Facebook's inferences with our conducted user survey.

4.1 Quality of Facebook's Demographic Estimates

Since our study focuses on Facebook's ad-targeting features and their associations along different demographic dimensions, it is imperative that Facebook's estimates for these demographic variables are reliable. The demographic variables we look at, are gender, ethnicity, age, completed education, and annual household income.

Gender, age and education are self-reported on the social network, while ethnicity (available only in the United States) is an inferred feature. Income estimates for U.S. residents on the platform are obtained through partnerships with data brokers [1]. Facebook plans on discontinuing these partnerships soon following increased scrutiny from government bodies over possible misuses of public data [22, 28].

Errors in demographic estimates therefore, might arise due to faulty self-reporting, problems in Facebook's inference mechanism or even because of problems in data acquired from external sources.

To investigate the quality of demographic estimates from Facebook's API, we compare them with ground truth data from the U.S. Census Bureau. For comparing gender, ethnicity, age and education level, we use the American Community Survey's (ACS) 2016 5-year estimates; for comparing income, we use household income estimates from the 2017 Current Population Survey (CPS). Both ACS and CPS are programs sponsored by the U.S. Census Bureau and thus provide federal and official data about the population.

Table 4.1 shows the comparison of demographic variables across both datasets. Because of the large sample sizes, it is important to account for the effect size while using χ^2 proportion tests across the datasets. We use Cohen's h (defined in Section 3.2) to quantify the effect size for the difference of proportions. We note that the fluctuations in the estimates for gender and ethnicity are practically insignificant. Age estimates are also fairly reliable with only one younger age group of 25-29 year olds being overrepresented. Of the demographic variables we consider, we find the estimates for education to be least reliable, where the platform has a significant propensity to overestimate the number of college graduates. Perhaps more surprisingly, we find Facebook's estimates for annual income to be highly representative of ground-truth data, given that data brokers have previously been criticized for having often inaccurate data [16]

We believe these results demonstrate that even though not everyone is on Facebook or targetable with its online ads, there are no significant skews in the predictions that the ad platform makes for gender, ethnicity, age and income. Overall, this practice gives us confidence that analyses built on top of these estimates are reliable and not spurious due to poor inferences on the social network.

4.2 Differences Across Demographics

Once we have made sure that the demographic estimates provided by the advertising platform are accurate, we observe how inferences for interest- or behavior-based attributes vary for said demographics. We pay attention to this question as it might have direct impact on the kind of content that users consume on the social network.

Many of these differences could possibly be benign. Targeting attributes like cosmetics or motorcycles might have gender connotations that would not be seen

Variable	Value	Facebook	% Census Bureau	Δ
Gender	Male	45.83	49.21	-3.38
	Female	54.16	50.78	+3.38
Ethnicity	White	64.58	61.95	+2.63
	Black	14.16	12.63	+1.53
	Asian	3.08	5.21	-2.13
	Hispanic	13.75	17.32	-3.57
Completed education	Less than high school	3.16	10.09	-6.93*
	High school	17.08	21.41	-4.33
	College	31.25	13.61	+17.64**
	Graduate school	4.11	7.71	-3.6
Age	15-19	5.83	6.69	-0.86
	20-24	12.91	7.09	+5.81
	25-29	13.75	6.89	+6.85*
	30-34	10.83	6.69	+4.13
	35-39	10.41	6.29	+4.11
	40-44	8.33	6.39	+1.93
	45-49	8.33	6.59	+1.73
	50-54	7.08	6.99	+0.08
	55-59	6.25	6.69	-0.44
	60+	14.16	20.39	-6.23
Income	$40,000 - $49,999	6.25	8.37	+0.08
	$50,000 - $74,999	15	16.95	-1.95
	$75,000 - $99,999	13.75	12.25	+1.5
	$100,000 - $124,999	7.08	8.71	-1.63
	$125,000 - $149,999	6.67	5.41	+1.26
	$150,000+	12.86	13.56	-0.7

TABLE 4.1: Percentage demographic makeup of Facebook's population in the United States, compared with ground truth data from the U.S. Census Bureau. The difference in Facebook is shown as Δ. χ^2 test of proportions yields $p < 0.001$ for all measurements, with small effect size (Cohen's $h < 0.2$) for unmarked differences. *$h = 0.2$; **$h = 0.4$.

as problematic, but perhaps even helpful in an advertisement context. However, if attributes related to jobs, education, or finances are disproportionately associated with different demographic groups, it might eventually affect the opportunities presented to the social network's users. To characterize such patterns hidden in the ad platform, we see how the targeting attributes are distributed across gender, ethnicity and income levels.

Figure 4.1 shows how the ad platform's inferences for men and women vary over all available targeting attributes listed under "Interests" and "Behaviors" in the Ads

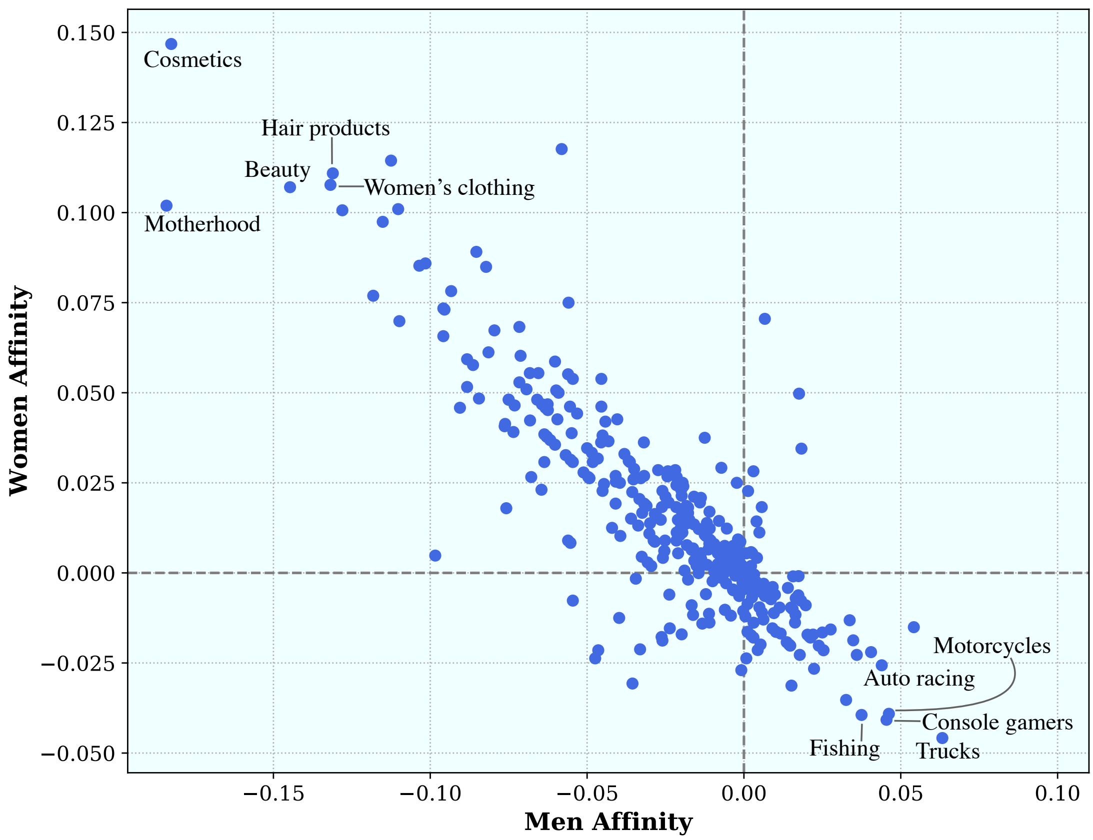


FIGURE 4.1: Differences in Facebook's inferences of ad targeting attributes for men and women in the United States. Each point corresponds to a targeting attribute available for advertisers in the Ads Manager; the top 5 most disparate points for each gender have been labeled. Affinty has the same definition as in Section 3.2; the grey dotted lines correspond to the affinity value if each group was as interested in an attribute as the entire U.S. population.

Manager. Here, we use the notion of affinity defined in Section 3.2 to analyze the inferences. Each point in the plot corresponds to a targeting attribute provided to advertisers. Positive (and negative) values of affinity reflect that a higher (or lower) fraction of the gender's population is inferred interested than the general U.S. population; the grey dotted line would represent the affinity if penetration for the gender were the same as the global population. The top 5 most disparate attributes for both genders have been annotated.

We can observe in Figure 4.1 that the most disparate attributes seem to be conventional interests associated with the genders. It isn't particularly concerning that Facebook thinks women are more interested in cosmetics than men – in fact, it might be illustrative of good personalization on the social network. However, it is interesting to note that many more targeting features are female- than male-leaning. Moreover, the magnitude of affinity for the most female-dominant attributes is much more than the male-dominant ones. These patterns could either exist because in the U.S., a higher fraction of women use Facebook than men (Table 4.1), or because women are more likely to interact with content and click on ads. Prior literature does offer some evidence for the latter explanation, where women were observed to be more likely to click on spam on Facebook [25]. Since Facebook's inferences rely on ad clicks and page likes [9], Figure 4.1 might suggest that women perform those activities more than men. We observe that points in the top-right and bottom-left quadrants are a result of imperfect stratification into the two gender groups – most likely either because some users do not classify themselves as male of female on the platform, or because of the reach estimate rounding process.

Figure 4.2 shows how the platform's inferences vary for two major ethnic groups in the U.S.: White and Black Americans. Unlike the results for gender, here many of the attributes are crowded around the axis, implying that the ethnic connotations of the targeting attributes are not too strong. Nevertheless, the annotated points suggest that the platform does have associations for certain attributes – for instance, country music is associated with Whites, while hip hop music has a bigger Black population.

Perhaps surprisingly, we also notice some attributes related to mobile web use, such as access over 4G networks, and from tablets and smartphones, are more highly associated with Black Americans. This is verified by prior work which has shown that in the United States, Blacks are more likely than Whites to be "smartphone only" users [23]. We believe this shows promise for using Facebook ad estimates for studies of web use as well.

For a more detailed picture, Table 4.2 shows the top 10 most disparate targeting features for men over women, Blacks over Whites, and high income users over low income users. For each feature, we show its position in the Ads Manager, the penetration for both demographic groups, and the disparity between the two.

We pick these three examples in particular as we believe they summarize the kinds of disparities we find within the ads system. The targeting attributes with the highest male penetration compared to women are reflective of personalization for men. In this situation, Facebook is simply capturing that men are much more interested in console games, outdoor activities, vehicles etc. than women. None of the results strike as problematic or somehow disadvantageous to women.

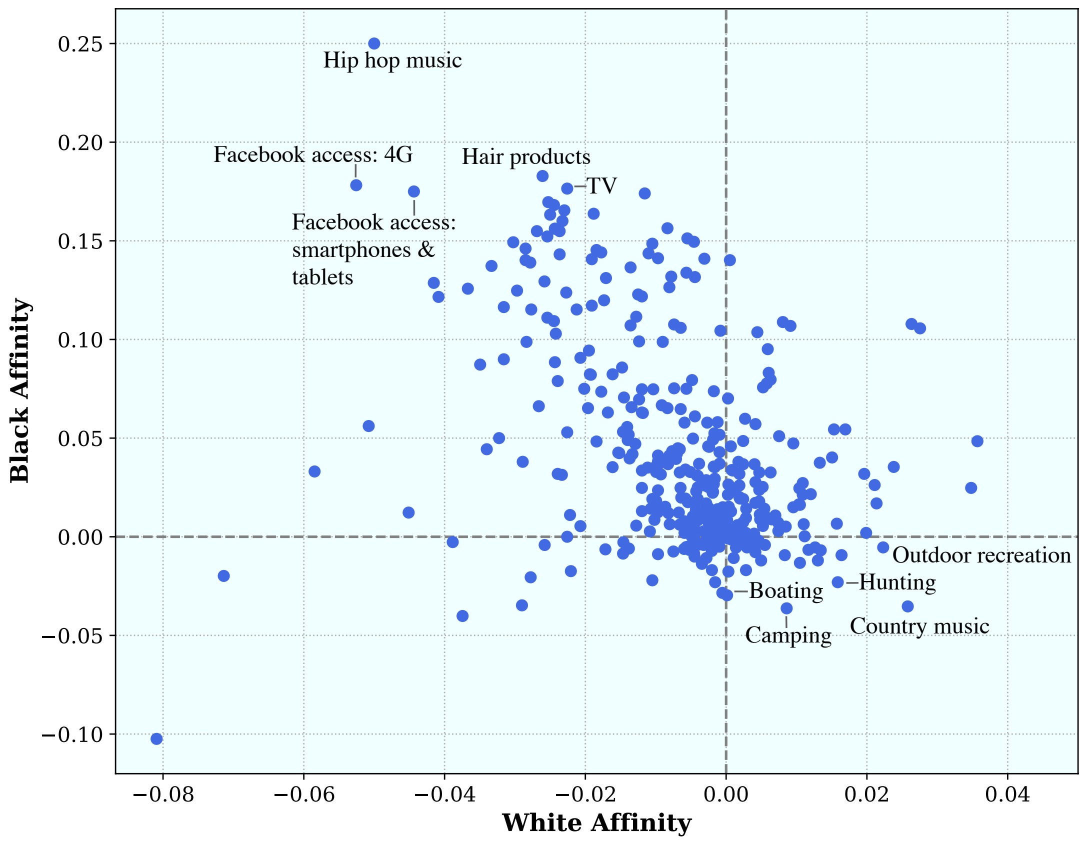


Figure 4.2: Differences in Facebook's inferences of ad targeting attributes for Black and White Americans.

For the attributes with highest disparity between Blacks and Whites, we also observe personalization, but in addition, it also includes results that are perhaps a bit surprising. For instance, Blacks being more interested in sales jobs or beauty products has no known normative explanation. Attributes like these and the ones related to web-use as mentioned earlier are most likely results of differences in how the demographic uses Facebook. Such results might not be indicative of biases hidden within the system but they present the opportunity to use Facebook's targeting APIs as a tool to understand people's internet usage patterns.

Finally, comparing high income users to low income users has similar results. There are multiple luxury features such as golf, skiing or Middle Eastern cuisine that have been personalized for high end users; but there are also results that might reveal bias. In particular, the advertising platform presents a possibility for exclusionary advertising by associating architecture and higher education with high-income users. These differences could perhaps translate to lower-income users receiving less content about higher education, and being presented less opportunity on the social network than richer users.

Our results from these high level analyses show that, depending on the demographic groups being observed, the differences in Facebook's inferences can range all the way from benign to insightful to concerning.

Table 4.2: Top 10 most disparate targeting attributes for three demographic groups: men, Blacks and high income users. For each group, the penetration of the attributes is compared with a competing group in the same demographic category. Penetration refers to the fraction of the demographic group that is inferred interested according to Facebook, disparity is simply the difference in penetration for two groups – as defined in Section 3.2. Targeting attributes are listed in decreasing order of disparity.

Gender: Penetration for men is more than women.

Attribute	Men	Women	Disparity
Interests → Hobbies and activities → Trucks	0.209	0.1	0.109
Behaviors → Digital activities → Console gamers	0.145	0.059	0.086
Interests → Hobbies and activities → Motorcycles	0.155	0.069	0.086
Interests → Sports and outdoors → Fishing	0.2	0.123	0.077
Interests → Sports and outdoors → Auto racing	0.127	0.058	0.069
Interests → Politics and social issues → Military	0.2	0.131	0.069
Interests → Sports and outdoors → Hunting	0.191	0.123	0.068
Interests → Technology → Game consoles	0.136	0.074	0.062
Interests → Sports and outdoors → Basketball	0.282	0.223	0.059
Interests → Politics and social issues → Veterans	0.118	0.065	0.053

Ethnicity: Penetration for Blacks is more than Whites.

Attribute	Black	White	Disparity
Interests → Entertainment → Hip hop music	0.5	0.2	0.3
Behaviors → Mobile Device User → Facebook access (network type): 4G	0.882	0.652	0.23

Behaviors → Mobile Device User → Facebook access (mobile): smartphones and tablets	1.0	0.781	0.219
Interests → Shopping and fashion → Hair products	0.441	0.232	0.209
Interests → Entertainment → TV	0.676	0.477	0.199
Interests → Technology	0.853	0.658	0.195
Interests → Shopping and fashion → Beauty	0.676	0.484	0.192
Interests → Entertainment → Music	0.853	0.665	0.188
Interests → Business and industry → Sales	0.588	0.4	0.188
Interests → Business and industry → Business	0.706	0.523	0.183

Income: Penetration for high income users ($350,000+) is more than lower income users ($40,000 - $75,000).

Attribute	**High Income**	**Low Income**	**Disparity**
Interests → Business and industry → Architecture	0.244	0.161	0.083
Interests → Business and industry → Higher education	0.401	0.327	0.074
Interests → Business and industry → Interior design	0.289	0.22	0.069
Interests → Food and drink → Middle Eastern cuisine	0.093	0.038	0.055
Interests → Fitness and wellness → Yoga	0.223	0.169	0.054
Interests → Sports and outdoors → Surfing	0.137	0.084	0.053
Interests → Sports and outdoors → Skiing	0.122	0.07	0.052
Interests → Sports and outdoors → Golf	0.168	0.118	0.05
Interests → Food and drink → Greek cuisine	0.08	0.03	0.05
Interests → Food and drink → Veganism	0.183	0.133	0.05

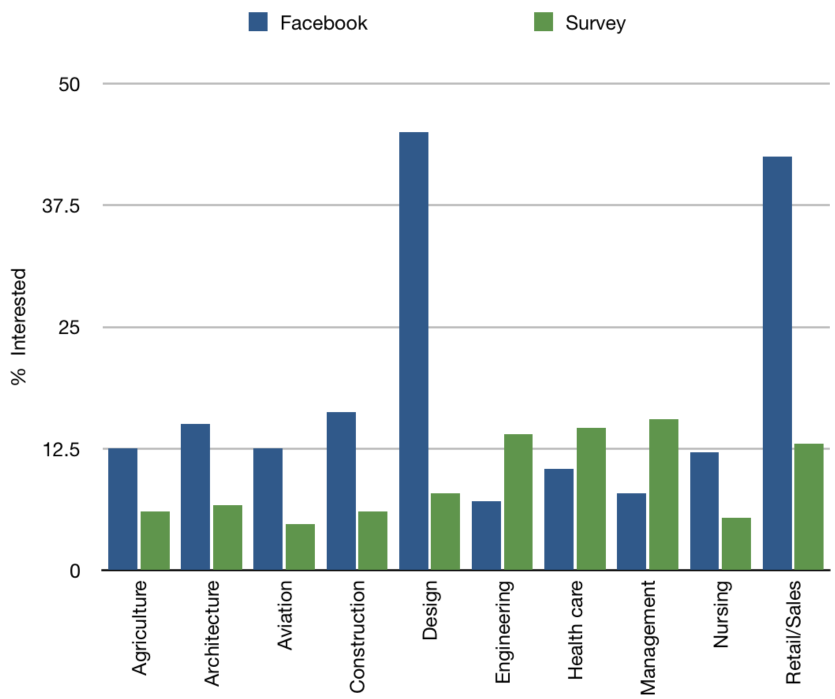
4.3 Fine-grained Analyses

After observing the differences in Facebook's inferences on a macro-level, we focus our attention on financial and professional targeting attributes. We specifically choose these two categories, as unlike entertainment choices or personal interests, ads for jobs or financial products might have immediate socioeconomic consequences for a user. It also presents an interesting avenue where we could investigate whether the ad platform has certain gender or ethnic stereotypes for occupations. Table 4.3 shows the professional and financial targeting attributes we choose for our analyses. All of the professional attributes listed are currently grouped within "Business and industry" under interests in the Ads Manager. Most financial attributes in Table 4.3 are also under the same hierarchy but grouped under "Personal finance" within the business and industry category. The only exception is gambling, which is listed under games in the entertainment category.

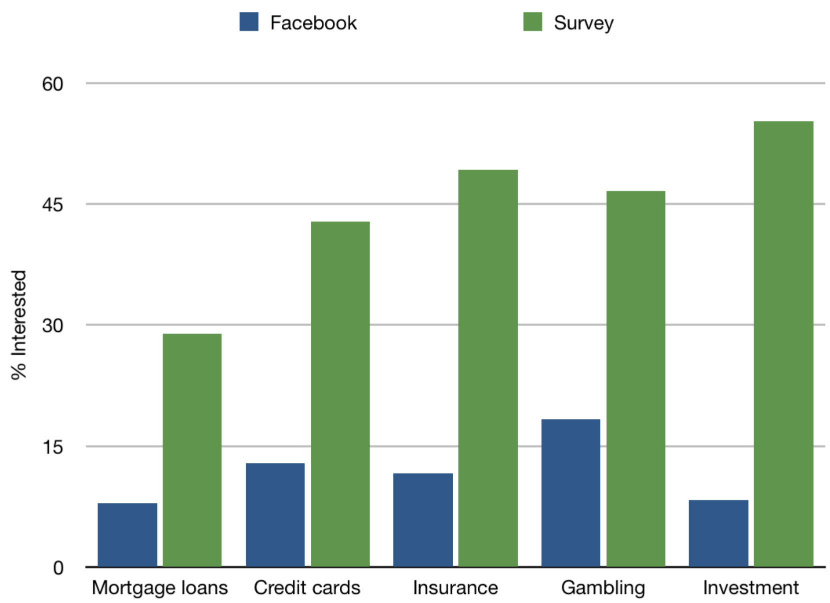
Category	Targeting Attribute
Finances	Credit cards Mortgage loans Insurance Investment Gambling
Professions	Agriculture Architecture Aviation Construction Engineering Design Health care Management Nursing Sales

TABLE 4.3: Financial and professional targeting attributes used in our study.

We obtain estimates for interest in these categories from the online ad system as well as from our user survey. Figure 4.3 shows the comparison of public interest in professional and financial topics between Facebook ads and our survey.



(A) Industries/Professions



(B) Financial interests

FIGURE 4.3: Comparison of professional and financial interests between Facebook and user survey.

It is immediately evident that out of all industries observed (Figure 4.3a), Facebook reports significantly larger proportion of its population as being interested ($p < 0.001; h > 0.50$) in sales and design than the survey. For other industries, the results are mixed: in industries like architecture, aviation or construction, Facebook reports more interest than the survey; while for engineering, health care and management, our survey respondents express slightly more interest than Facebook's estimates. However, many of these differences are much smaller in magnitude than the exaggeration done by Facebook for design and sales.

On the other hand, our survey respondents declare significantly more interest ($p < 0.001$; $h > 0.50$) in the financial categories than what Facebook infers about them (Figure 4.3b). Possible explanations for these differences could either be that Facebook isn't a large enough marketplace for financial product ads, or that users do not actively interact with such content on Facebook. In the former case, the low interest could be a result of advertisers not using the attributes enough, while in the latter, it would be explained by a lack of user engagement.

To interpret these differences, it is also helpful to understand that inferences for ad targeting attributes on Facebook work in a self-reinforcing cycle: ad clicks and page likes by users result in them being inferred as interested [9]; and marketers advertise for said ad clicks or page likes [8]. Thus particularly high interests like the ones observed for sales and design could be a result of this compounded supply and demand phenomenon.

Furthermore, the fact that design and sales are so different from the other industry categories might also suggest that advertisers are using these attributes quite differently/for different purposes as compared to the other industries. Given that all of the professional attributes that we use are grouped under the same hierarchy in the Ads Manager, it shows that advertisers might disregard how the attributes are organized and use what they deem suitable.

4.3.1 Professions

In addition to these high-level observations, we stratify the estimates for the professional attributes by ethnicity and gender to observe the differences for each demographic group. Table 4.5 compares the fraction of men and women who are interested in the selected professions according to Facebook and our survey. We also compare the interest for each profession on Facebook with employment data

Facebook	Bureau of Labor Statistics (BLS)
Agriculture	Farmers, ranchers, and other agricultural managers
	Graders and sorters, agricultural products
	Miscellaneous agricultural workers
Architecture	Architects, except naval
	Drafters
Aviation	Aircraft pilots and flight engineers
	Aircraft mechanics and service technicians
Construction	Construction and extraction occupations
	Construction managers
Design	Designers
Engineering	Aerospace engineers
	Agricultural engineers
	Biomedical engineers
	Chemical engineers
	Civil engineers
	Computer hardware engineers
	Electrical and electronics engineers
	Environmental engineers
	Industrial engineers, including health and safety
	Marine engineers and naval architects
	Materials engineers
	Mechanical engineers
	Mining and geological engineers, including mining safety engineers
	Nuclear engineers
	Petroleum engineers
	Engineers, all other
	Engineering technicians, except drafters
Health care	Healthcare practitioners and technical occupations
Management	Management occupations
Nursing	Registered nurses
Sales	Sales and related occupations

Table 4.4: Mapping from Facebook's professional targeting attributes to professions recorded by the Bureau of Labor Statistics (BLS).

from the U.S. Bureau of Labor Statistics (BLS)[1]. Table 4.4 also shows the mapping we do from targeting categories listed on Facebook to professions recorded by BLS. We obtain BLS estimates for a particular industry by summing over all mapped professions listed in Table 4.4.

We compare against BLS to have a frame-of-reference with respect to actual employment statistics in the U.S., and to see whether Facebook is more or less biased

[1] https://www.bls.gov/cps/cpsaat11.htm

comparatively.

Industry	Facebook Men	Facebook Women	BLS Men	BLS Women	Survey Men	Survey Women
Agriculture	9.0	**15.38**	**0.94**	0.32	**9.66**	3.18
Architecture	11.82	**16.92**	**0.18**	0.06	**11.03**	3.18
Aviation	12.73	**13.08**	**0.16**	0.01	**8.28**	1.91
Construction	**16.36**	15.38	**5.61**	0.2	**11.72**	1.27
Engineering	**8.64**	6.08	**1.6**	0.26	**22.07**	7.64
Design	35.45	**52.31**	0.27	**0.3**	4.14	**12.1**
Health care	6.0	**14.62**	1.46	**4.24**	9.66	**20.38**
Management	6.55	**8.46**	**6.84**	4.38	**18.62**	14.01
Nursing	4.45	**16.15**	0.2	**1.76**	0.69	**10.19**
Sales	38.18	**46.15**	**5.13**	4.8	**15.17**	12.1

TABLE 4.5: Percentage of gender population interested in different professions. Comparison of Facebook's interest for both genders is done with U.S. Bureau of Labor Statistics' (BLS) employment data, as well as interest expressed in survey. Statistically significant ($p < 0.05$) differences in proportions with $0.5 > h \geq 0.2$ are highlighted in yellow, $h \geq 0.5$ are highlighted in orange. **Bold** shows which gender is more interested in the profession in each dataset.

It is apparent from Table 4.5 that the fraction of men and women interested in Facebook, and the fraction actually employed according the BLS are vastly different. So we can be certain that the inferences on Facebook are not predictive of how many people are employed in the profession. Comparing level of interest across Facebook and the survey shows us that often Facebook results for men's interests correspond to what the survey respondents declared. Facebook's estimates for male interest in agriculture, architecture, and aviation etc. are fairly close to the survey responses. However, it seems to not line up with the interest for women. Female survey respondents consistently express less interest than men in professions like agriculture, architecture and aviation. Interestingly, Facebook happens to inflate interest for women beyond men in its ad estimates. It is interesting to note that estimates from the survey of which gender is more interested in a profession (highlighted in bold in Table 4.5) line up perfectly against BLS' employment data. Facebook, on the other hand, reports women as more interested for most professions – except construction and engineering, and even for those professions, it seems to reduce the disparity reflected in the survey or BLS' employment data.

In line with what we observe in Figure 4.3a, Facebook significantly overestimates the interest for sales and design industries, even when stratified by gender. This provides further evidence that perhaps Facebook's inferences for these attributes are not indicative of interest in sales or design jobs, and that advertisers might be using these attributes for different purposes.

These results for professional targeting attributes follow a similar trend to the general distribution observed in Figure 4.1 – where the ad platform infers women to be generally more interested for most attributes. In the case of job ad targeting, this could perhaps be seen as a good thing, in that Facebook could help bridge the gender gap for employment ads.

Similar to the gender differences in Table 4.5, Table 4.6 shows ethnic differences in professional interests across Facebook, BLS, and the survey.

	Facebook		**BLS**		**Survey**	
Industry	White	Black	White	Black	White	Black
Agriculture	**13.55**	11.18	**0.93**	0.12	**7.85**	0.0
Architecture	14.84	**17.94**	**0.17**	0.03	**6.81**	6.25
Aviation	13.55	**14.12**	**0.12**	0.03	5.24	**6.25**
Construction	16.13	**16.76**	**4.07**	1.51	**6.81**	3.12
Engineering	7.1	**7.94**	**1.13**	0.41	**15.18**	12.5
Design	44.52	**52.94**	**0.39**	0.12	7.33	**15.62**
Health care	10.32	**14.71**	**3.54**	2.7	13.61	**21.88**
Management	7.1	**11.76**	**7.58**	3.36	17.28	**25.0**
Nursing	11.61	**13.82**	**1.23**	0.97	4.71	**6.25**
Sales	40.0	**58.82**	**1.28**	0.87	**12.57**	12.5

TABLE 4.6: Percentage of population in ethnic group interested in different professions. Annotation scheme and table organization is the same as in Table 4.5. Because of small counts while comparing with ethnicities in the survey, we use Fisher's exact test to compare differences in proportions.

Here too, we notice many of the same patterns. Facebook's estimates are not reflective of employment data from BLS. Consistent with prior observations, estimates for design and sales are hyper-inflated. Facebook consistently inflates the interest for both ethnicities for most professions, except engineering, health care and management. Looking at the Facebook estimates, we can observe that for most professions, there are minute differences in the penetration for Blacks and Whites, even when other data shows otherwise – such as in agriculture, where Black survey respondents show no interest whatsoever; or in management, where Whites are more than twice as employed than Blacks according to BLS.

These results suggest that even though Facebook's inferences for professions along ethnicity might consistently be different from both employment data and our survey results, they are not particularly detrimental to either group. In fact, the low disparity in inferences might even be helpful for bridging employment gaps for high paying jobs in engineering or management – which are evident from BLS' data.

4.3.2 Finances

We also examine in our study how Facebook's inferences for financial attributes change for different income profiles. Figure 4.4 shows how interest in gambling, investment and credit cards on Facebook changes with increasing income. We see a steady decline in the interest for gambling and credit cards. We also observe that the interest in investment only drops slightly for the higher income users, but the difference is not as pronounced as the other two attributes.

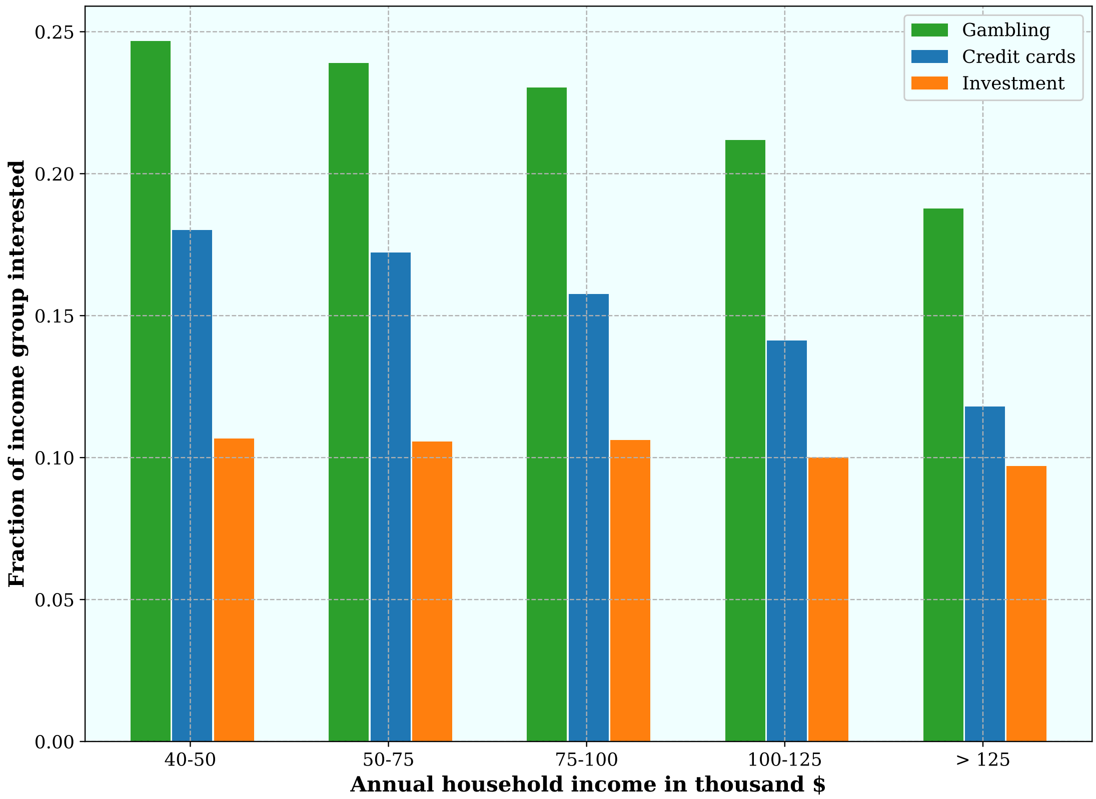


FIGURE 4.4: Change in interest for financial targeting attributes on Facebook with increasing annual income.

We find this pattern concerning as gambling is the riskiest of the three financial interests and has been shown to be addictive [5]. Despite its possible harms to low income users, Facebook believes more of them are interested than richer users.

Similarly, the ad platform associates lower income users with credit cards, a debt-related product, more often than high income users.

These trends, though concerning, could simply be a reflection of interests in the real world. To understand whether Facebook is simply reflecting a real-world phenomenon, we compare these results to responses in our survey. Figure 4.5 shows interest for the same financial attributes from our survey, over different annual incomes. Each income bin in Figure 4.5 has at least 25 survey respondents.

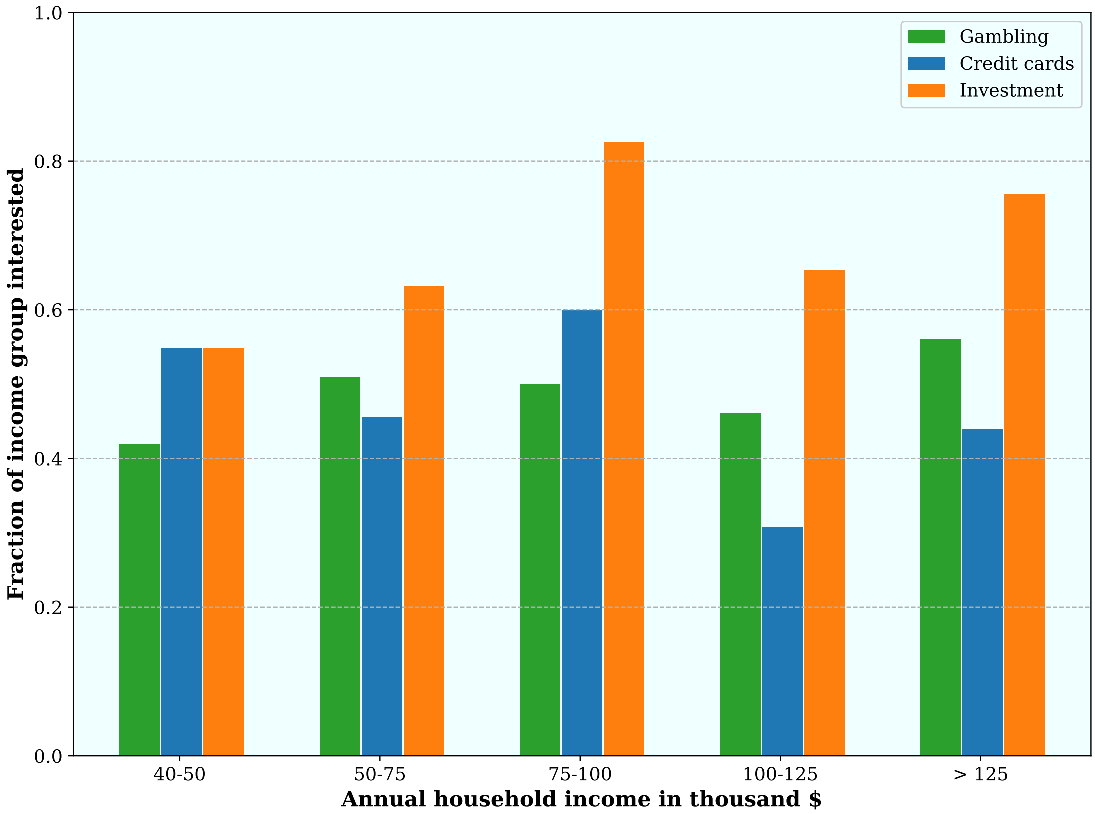


FIGURE 4.5: Change in interest for financial targeting attributes in survey with increasing annual income.

Contrary to the trend in ad targeting data, our survey suggests a slight increase in interest for gambling as income increases. Interest for credit cards also slightly increases before declining for the higher income brackets (> $100,000), but is more random than the trend for gambling. We also observe a noticeable increase in interest for investment as income increases. All of these trends are in contrast with Facebook's estimates in Figure 4.4.

Such disparities between Facebook and real-world data could be troubling because the ad platform, with its skewed world-view, essentially presents an opportunity to advertisers to target financially vulnerable users. Given that it has been documented that gambling advertisers particularly seek lower-income users [6], such biases in the ad platform might assist malicious advertisers.

As mentioned earlier, inference for these targeting attributes work in a self-reinforcing cycle where clicking on ads and interacting with advertised content leads to the user being considered interested [9]. Even though Facebook considers much less people as interested in finances when compared to the survey (Figure 4.3b), what we find concerning is that by harboring such biases, the platform might allow for the disparities to exacerbate over time. If the core audience for credit card and gambling ads are believed to be lower-income users, this would lead to increased ad delivery for them, which could further lead to stronger inferences, and so on.

However, not all disparities that we observe here are detrimental. For instance, Facebook also infers low income users as slightly more interested in investment options as well. Assuming that advertisers who target for investment products aren't predatory, this would provide a leveled playing field for different income groups, giving everyone the opportunity to participate. However, as is evident in Figure 4.4, the potential benefit that the platform might present to lower income users in terms of advertisement ads might be overshadowed by the risk it could expose them to, with ads for gambling or credit cards.

Chapter 5

Conclusions and Outlook

5.1 Summary

The motivation of our study was to take a modern, large-scale targeted advertising system and observe if the user attributes inferred by the system have any hidden biases. We argue on the importance of understanding attribute inferences as they are one of the major tools used by advertisers. Understanding an advertising platform's inferences gives us insight into what the system thinks about its users, and whether it is capable of unknowingly discriminating between certain socially salient groups in the user-base.

We choose Facebook's advertising system, and analyze how the ad-targeting attributes on the platform are inferred differently based on gender, ethnicity and annual income, for users in the United States. We take all demographic and interest estimates from the advertising platform itself.

We conduct a systematic measurement study where we first ensure that Facebook's demographic estimates are accurate by comparing them with census data. We then characterize on a high-level how available attributes in the interface vary along gender, ethnicity and income. Later, we focus on a subset of the attributes related to jobs and finances, and present fine-grained analyses by comparing with a user survey that we conduct for the same interests.

We would like to summarize our results keeping in light the research questions we pose in Section 1.1. The first question we ask of the advertising system is:

1. How do Facebook's ad attribute inferences vary across different demographic factors such as gender, ethnicity, and income?

We find noticeable differences in Facebook's attribute inferences for these demographic groups. Our results in Section 4.2 show that the ad platform considers women and Blacks more interested than men or Whites respectively, for a majority of attributes. We argue that the differences could be attributed to how these demographic groups use Facebook differently. Even though most of the differences we find are reflective of benign personalization, we do identify that high income users are more often inferred interested in architecture jobs and higher education – and the platform thus holds potential to exclude lower-income users from important opportunities. The second question that we asked was:

2. Are these differences relatively more or less disparate than what we observe in the real world?

To compare against "real-world" interests, we conduct a survey with 300 respondents where we ask them their interest in various jobs and financial products. Due to the ad platform's tendency to consider women and Blacks more interested, we find that surprisingly, the disparity in inferences for professional attributes on Facebook is lesser than our survey. We discuss these results in Section 4.3 and believe that these differences might help alleviate gender and ethnic disparity in the workforce. Our third question was:

3. How could these differences impact the ads on Facebook, and eventually users on the social network?

In Section 4.3, while comparing Facebook's targeting attributes with survey responses, we also find that Facebook associates gambling more often with low- than high-income users. On one hand, the ad platform's inferences for jobs are less disparate than our survey and might help alleviate disparity in the workforce – but on the other hand, its ability to associate gambling with low-income users and higher education with high-income users shows how it could affect ad delivery and perhaps limit opportunities for some demographics. We also discuss in Section 4.3.2 how the ad platform's internal mechanics for inferring interests might lead to these differences being exacerbated over time.

5.2 Outlook

This thesis assists in better understanding a small but important part of the large, complex ecosystem of targeted advertising. Figure 5.1 shows three major components of the advertising ecosystem [27]: advertisers, the advertising platform, and the users. We illustrate here that the advertising system itself is made up of several components, the user attributes which we study just being one of them. While this thesis works to measure the latent biases in the ad platform's user attributes, we do not tackle the question of how these biases concretely translate to differences in ad delivery.

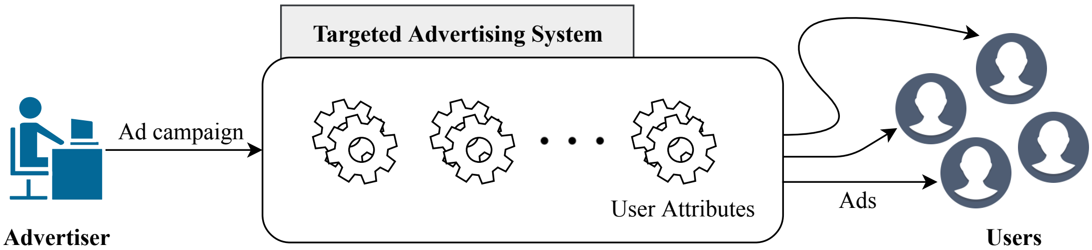


FIGURE 5.1: An illustration of the targeted advertising ecosystem with its major constituents: the advertiser, the targeting system, and the users.

In parallel, some of our results (Figure 4.3a) suggest that even when grouped together by the advertising platform, some attributes might be used differently by advertisers. We therefore believe understanding how advertisers perceive the targeting attributes is an important aspect of the system that needs further work.

We believe building a more holistic understanding of bias in the advertising ecosystem requires studying: (i) how advertisers perceive and use targeting attributes; (ii) how the targeting attributes might be biased (this thesis); and (iii) measuring how the disparities in attribute inference impact ad delivery. Having such an end-to-end understanding would help both in designing better systems in the future, and making current systems more robust.

In addition to our study's main objectives, as a byproduct, our study also shows the potential for using advertising data to understand people's web-use behaviors. The rich targeting attributes related to device usage, shopping behaviors and interests can be leveraged as a real-time data source to complement conventional survey-based approaches.

Appendix A

User Survey Details

A.1 Survey Questions

Here, we list the questions from our survey which have been used in the analyses for the results.

Interests. Questions in this section were shuffled for each respondent. The respondents were not shown that they were in the interests section of the survey.

Q1: Have you ever been interested in pursuing a job or career in any of following professions or in learning more about any of the following professions? Multiple fields may apply.

- Agriculture (sorters, farmers, ranchers, agricultural manager, etc.)
- Architecture (drafters, architects, etc.)
- Aviation (pilots, flight engineers, aircraft mechanics, etc.)
- Construction
- Design (interior design, graphic design, UI design, etc.)
- Engineering (mechanical, computer, civil, etc.)
- Health care (administrative and practitioners, not including nursing)
- Management (any management or supervisory level role)
- Nursing (licensed and registered nurses, nurse practitioners)
- Retail (sales, customer service, etc.)
- Other [text entry]
- I don't know
- Prefer not to answer

Q2: Are you interested in applying for a mortgage or buying a home or learning more about mortgage and home purchasing options?

- Yes
- No
- I don't know
- Prefer not to answer

Q3: Are you interested in investing money or learning more about investment options?

- Yes

- No
- I don't know
- Prefer not to answer

Q4: Are you interested in gambling (e.g., buying a lottery ticket, visiting a casino, betting on a sports event, playing video or online poker, playing bingo for money, betting on a horse race, etc.) or learning more about gambling options?

- Yes
- No
- I don't know
- Prefer not to answer

Q5: Are you interested in buying insurance or learning more about insurance options? This may include health insurance, life insurance, car insurance, and etc.

- Yes
- No
- I don't know
- Prefer not to answer

Q6: Are you interested in enrolling in a B.A., B.S., M.S., M.A., Ed.D., Ph.D., or other college or graduate level degree or certificate program or in learning more about college or graduate education options?

- Yes
- No
- I don't know
- Prefer not to answer

Q7: Are you interested in obtaining a credit card or learning more about credit card options?

- Yes
- No
- I don't know

- Prefer not to answer

Q8 (if answered yes to Q7): Which of the following types of credit cards are you interested in obtaining or learning more about, if any?

- Gas or retail store cards
- Premium credit cards
- Travel or entertainment credit cards
- High-end department store (e.g., Neiman Marcus, Saks Fifth Avenue) credit cards
- Other [text entry]
- I don't know
- Prefer not to answer

Demographics. Questions in this section were not shuffled and were presented in the order listed here. Demographic details were always asked at the end of the survey. Respondents were not informed that they were in the demographics part of the survey.

Q9: Please specify the gender with which you most closely identify.

- Female
- Male
- Other
- Prefer not to answer

Q10: How old are you today? [text entry]

Q11: What is the highest level of school you have completed or the highest degree you have received?

- High school incomplete or less
- High school graduate or GED (includes technical/vocational training that doesn't count towards college credit)
- Some college (some community college, associate's degree)
- Four year college degree/bachelor's degree
- Some postgraduate or professional schooling, no postgraduate degree

- Postgraduate or professional degree, including master's, doctorate, medical or law degree
- Other
- Prefer not to answer

Q12: Are you of Hispanic, Latino, or Spanish origin, such as Mexican, Puerto Rican or Cuban?

- Yes
- No
- Prefer not to answer

Q13: Which of the following describes your race?

- White
- Black of African-American
- Asian or Asian-American
- Native American/American Indian/Alaska Native
- Native Hawaiian/Other Pacific Islanders
- Some other race
- Prefer not to answer

Q14: Last year, that is in 2017, what was your total family annual income from all sources, before taxes?

- Less than $10,000
- $10,000 to less than $20,000
- $20,000 to less than $30,000
- $30,000 to less than $40,000
- $40,000 to less than $50,000
- $50,000 to less than $75,000
- $75,000 to less than $100,000
- $100,000 to less than $125,000
- $125,000 to less than $150,000
- $150,000 to less than $250,000
- $250,000 to less than $350,000

- $350,000 to less than $500,000
- More than $500,000
- I don't know
- Prefer not to answer

A.2 Survey Demographics

Table A.1 gives the demographic breakdown of our survey respondents. We also compare the proportions with U.S. Census Bureau data using 2-sample χ^2 test for equality of proportions.

Variable	**Value**	% **Survey**	**Census Bureau**	Δ
Gender	Male	46.03	49.21	-3.18
	Female	49.84	50.78	-0.94
Ethnicity	White	60.63	61.95	-1.32
	Black	10.16	12.63	-2.47
	Asian	5.08	5.21	-0.13
	Hispanic	17.78	17.32	+0.46
Completed education	Less than high school	2.54	10.09	-7.55*
	High school	13.02	21.41	-8.39*
	College	25.71	13.61	+12.1*
	Graduate school	8.41	7.71	+0.7
Age	20-24	10.15	7.09	+3.06*
	25-29	7.61	6.89	+0.72
	30-34	8.25	6.69	+1.56
	35-39	8.88	6.29	+2.59
	40-44	8.88	6.39	+2.49
	45-49	9.84	6.59	+3.25*
	50-54	9.52	6.99	+2.53
	55-59	7.93	6.69	+1.24
	60+	23.17	20.39	+2.78
Income	$40,000 - $49,999	9.84	8.37	+1.47
	$50,000 - $74,999	18.09	16.95	+1.14
	$75,000 - $99,999	12.69	12.25	+0.44
	$100,000 - $124,999	8.25	8.71	-0.46
	$125,000 - $149,999	5.39	5.41	-0.02
	$150,000+	7.61	13.56	-5.95*

TABLE A.1: Demographic breakdown of survey respondents. $*p < 0.05$.

Table A.1 suggests that the survey is fairly census-representative for gender, age, ethnicity and income. For education, however, it leans towards more educated people than the general U.S. population.

Bibliography

[1] 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 *The Network and Distributed System Security Symposium (NDSS)*, 2018.

[2] Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProPublica*, November 2017.

[3] Julia Angwin, Madeleine Varner, and Ariana Tobin. Facebook Enabled Advertisers to Reach 'Jew Haters'. *ProPublica*, September 2017.

[4] Matheus Araujo, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 253–257. ACM, 2017.

[5] Carlos Blanco, Paulo Moreyra, EV Nunes, Jeronimo Saiz-Ruiz, and Angela Ibanez. Pathological Gambling: Addiction or Compulsion? In *Seminars in clinical neuropsychiatry*, volume 6, pages 167–176, 2001.

[6] Mattha Busby. Revealed: how gambling industry targets poor people and ex-gamblers. https://www.theguardian.com/society/2017/aug/31/gambling-industry-third-party-companies-online-casinos. Accessed: 2018-07-30.

[7] Claude Castelluccia, Mohamed-Ali Kaafar, and Minh-Dung Tran. Betrayed by Your Ads! In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–17. Springer, 2012.

[8] Facebook Help Center. Advertising Objectives. https://www.facebook.com/business/help/197976123664242. Accessed: 2018-07-30.

[9] Facebook Help Center. How does Facebook decide which ads to show me and how can I control the ads I see? https://www.facebook.com/help/562973647153813?helpref=faq_content.
Accessed: 2018-07-29.

[10] Facebook Help Center. Why am I seeing ads from a specific business or advertiser? https://www.facebook.com/help/1674984446161704.
Accessed 2018-07-17.

[11] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ, 1988.

[12] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[13] SM Ellis and HS Steyn. Practical Significance (Effect Sizes) Versus, or in Combination with Statistical Significance (p-values): Research Note. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, 12(4):51–53, 2003.

[14] eMarketer. US Digital Ad Spending to Surpass TV this Year. https://www.emarketer.com/Article/US-Digital-AdSpending-Surpass-TV-this-Year/1014469., 2016.
Accessed: 2018-07-31.

[15] Karri Haranko, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. Professional gender gaps across us cities. *arXiv preprint arXiv:1801.09429*, 2018.

[16] Melanie Hicken. Find out what Big Data knows about you (it may be very wrong). https://money.cnn.com/2013/09/05/pf/acxiom-consumer-data/.
Accessed: 2018-07-29.

[17] Rosamond Hutt. The world's most popular social networks, mapped. https://www.weforum.org/agenda/2017/03/most-popular-social-networks-mapped/.
Accessed: 2018-07-27.

[18] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. What Matters to Users? Factors that Affect Users' Willingness to Share Information with Online Advertisers. In *Proceedings of the ninth symposium on usable privacy and security*, page 7. ACM, 2013.

[19] Miguel Malheiros, Charlene Jennett, Snehalee Patel, Sacha Brostoff, and Martina Angela Sasse. Too Close for Comfort: A Study of the Effectiveness and Acceptability of Rich-Media Personalized Advertising. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 579–588. ACM, 2012.

[20] Yelena Mejova, Ingmar Weber, and Luis Fernandez-Luque. Online Health Monitoring using Facebook Advertisement Audience Estimates in the United States: Evaluation Study. *JMIR public health and surveillance*, 4(1), 2018.

[21] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[22] Facebook Newsroom. Shutting Down Partner Categories. https://newsroom.fb.com/news/h/shutting-down-partner-categories/. Accessed: 2018-07-23.

[23] Andrew Perrin. Smartphones Help Blacks, Hispanics Bridge Some – but not all – Digital Gaps with Whites, 2017.

[24] Francesco Rampazzo, Emilio Zagheni, Ingmar Weber, Maria Rita Testa, and Francesco Billari. Mater Certa Est, Pater N umquam: What can Facebook Advertising Data Tell Us about Male Fertility Rates? *arXiv preprint arXiv:1804.04632*, 2018.

[25] Elissa M Redmiles, Neha Chachra, and Brian Waismeyer. Examining the Demand for Spam: Who Clicks? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 212. ACM, 2018.

[26] Koustuv Saha, Ingmar Weber, Michael L Birnbaum, and Munmun De Choudhury. Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates. *Journal of medical Internet research*, 19(5), 2017.

[27] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Conference on Fairness, Accountability and Transparency*, pages 5–19, 2018.

[28] Nick Statt. Facebook will no longer allow third-party data for targeting ads. https://www.theverge.com/2018/3/28/17174854/facebook-shutting-down-partner-categories-ad-targeting-cambridge-analytica. Accessed: 2018-07-23.

[29] Google Support. Control the ads you see. https://support.google.com/ads/answer/2662856.

[30] Latanya Sweeney. Discrimination in Online Ad Delivery. *Queue*, 11(3):10, 2013.

[31] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *proceedings of the eighth symposium on usable privacy and security*, page 4. ACM, 2012.

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

[33] Craig E Wills and Can Tatar. Understanding What They Do with What They Know. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, pages 13–18. ACM, 2012.

[34] Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, 43(4):721–734, 2017.