Chapter 1

Data and Methods

In our work, we draw from Facebook advertising data to produce estimates of the number of users with particular demographics and interests. In this document, 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.

1.1 Facebook's Advertisement Ecosystem

As the world's largest social network[1], Facebook's targeted advertising system allows marketers and businesses to target ads to its users by their demographic attributes, location, as well as interests and internet behaviors that it has inferred about them. 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 – or at least what Facebook *thinks* they are.

Using the social network's ad system, we can issue 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.

[1] https://www.weforum.org/agenda/2017/03/most-popular-social-networks-mapped/

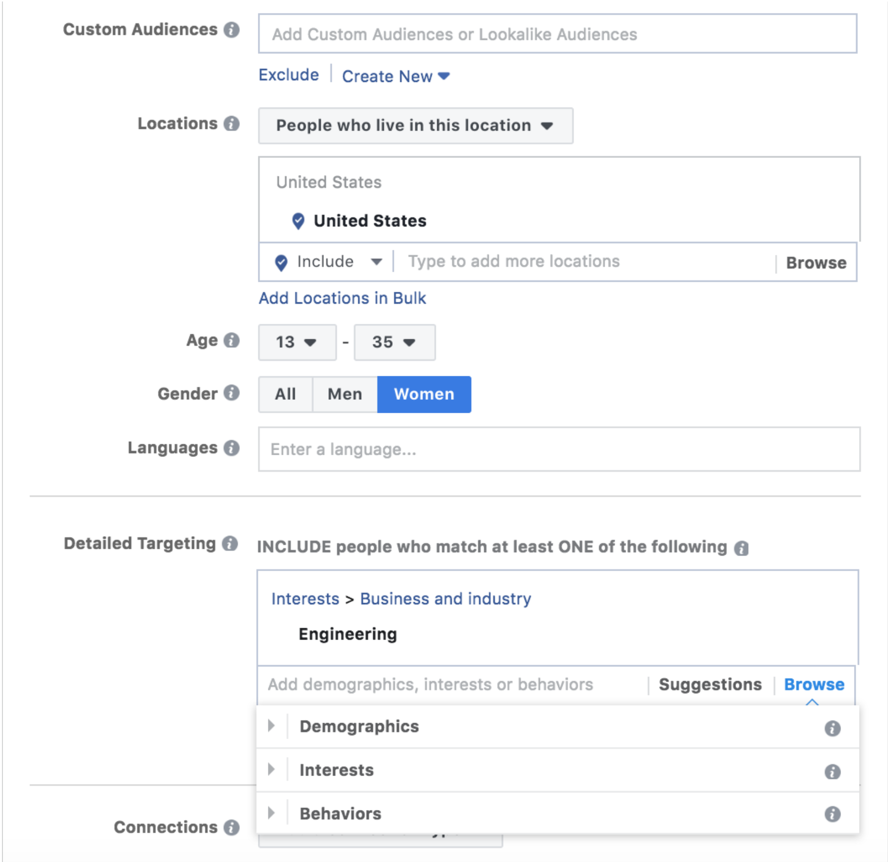

Figure 1.1: Screenshot of Facebook's Ads Manager interface showing some of the major functionalities available. Geographical location, age and gender have been specified. An elaborate list of targeting options are available through the cascading drop-down menus under "Detailed Targeting". Here, the category "Engineering" has been selected.

Facebook provides multiple tools to target audiences on the social network, such as uploading personally identifiable information (PII) or finding people with similar interests to a prior audience; but for our study, we primarily focus on a feature the platform refers to as *Core Audiences*[2]. Here, Facebook allows marketers to explicitly choose the features or characteristics 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.

1.1.1 Available Targeting Options

Figure 1.1 shows the structure of Facebook's *Ads Manager* page, where marketers can construct core audiences for advertisement campaigns. We see that the interface provides the option to target by location, age, gender, language and a wide

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

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 dropping a pin at a particular address and specifying a target radius as small as 1 kilometer around it. For all geographic options, advertisers are able to target people who either live in the area, have recently traveled there, are currently traveling to the area, 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 interests 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. Even 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. Similarly, we find the "Higher education" attribute to approximate an interest in education very well, but its classification under the "Business and industry" category could be misleading.

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 [1], and removing problematic anti-semitic ad categories that were automatically indexed by the ad system [2]. These regular changes to the features warrant caution while collecting data from the platform.

In addition to the features characterized 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 1.1 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.
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:

$$(\text{electrical engineering} \vee \text{software engineering}) \wedge \text{mathematics} \wedge \neg\text{design}$$

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 1.2. These reach estimates form the foundation of our study and are the primary measurements we take from the online ads network.

1.1.2 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

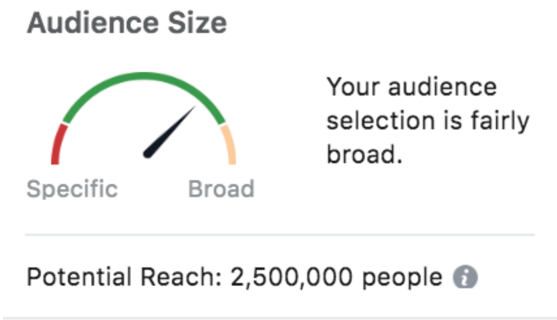

FIGURE 1.2: Size of targetable population (reach estimate) with the targeting options selected in Figure 1.1 i.e. women living in the United States, between the ages of 13 and 35, who are interesting in engineering.

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[3] to make queries to the Marketing and Targeting Search APIs.

1.1.3 Data Collected

We collect both demographic and advertisement related estimates from the Marketing API.

Demographically, we choose to focus on age, gender, ethnicity and education; we discontinue using education for our later analyses because of poor quality estimates from the ad platform (more detail in Section 1.2).

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

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

United States as it is a high internet penetration country, and also the only region where Facebook infers user 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. We query for age by breaking down the ages into five year bins starting from 15 until "60+", which is the maximum targetable age on the platform.

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 interested in the attribute. The latter estimate is made with a logical AND operation i.e. with the "narrow" option in the API.

1.2 Quality of Facebook's 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. Gender, age and education are self-reported on the social network, while ethnicity (available only in the United States) is an inferred feature. Errors might arise either due to faulty self-reporting or due to problems in Facebook's inference mechanism.

To investigate the quality of demographic estimates from Facebook's API, we compare the demographic estimates to ground truth data from the American Community Survey's 2016 5-year estimates. The American Community Survey (ACS) is a survey administered by the United States Census Bureau in addition to the decennial census to annually update housing and demographic statistics [3]. We use the developer API[4] provided by ACS to query for ground truth data.

As of the time of this writing, the Ads Manager reports that 240 million people who live in the United States are targetable with Facebook ads. To understand the

[4] https://www.census.gov/data/developers/data-sets.html

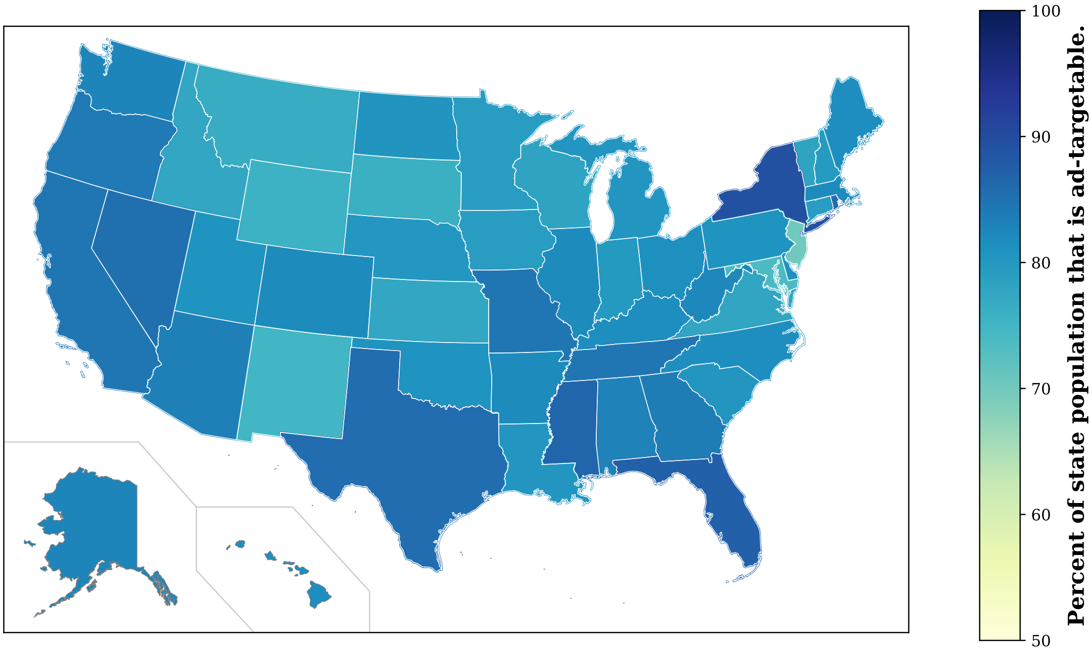

FIGURE 1.3: State level penetration of Facebook advertisements in the United States. In every state, at least half of the population is reachable with targeted advertising.

scale of this penetration, we stratify this estimate by each state and observe what fraction of the population in each state is targetable. Figure 1.3 shows the fraction of population in each state that is targetable with Facebook's targeted ads. We find that the lowest penetration is in New Jersey with 58.13% of the population being targetable. The average penetration over all states is 69.45%. 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.

Further, Table 1.1 shows the comparison of gender, ethnicity, education and age estimates across both datasets. Because of the large sample sizes, it is important to use χ^2 proportion tests while accounting for the effect size. We use Cohen's h to quantify the effect size for the difference of proportions.

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 and age. 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.

Variable	Value	% Facebook	ACS	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

TABLE 1.1: Percentage demographic makeup of Facebook's population in the United States, compared with ground truth data from the American Community Survey (ACS). 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. *$h = 0.2$; **$h = 0.4$.

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.

1.3 Methods for Analysis

Once we have collected data from the marketing API and ensured that Facebook's demographic estimates corroborate with real-world data, we begin evaluating which targeting features have the highest affiliations with which subgroups of the population.

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

1. How likely it for people from a demographic to be inferred interested in a targeting feature?

2. Are members of the demographic more likely to be inferred than the general population?

3. Are they more likely to be inferred than another particular demographic?

To answer the first question, we look at what fraction of a group's population is interested in an attribute according to Facebook. We refer to this ratio as the *penetration* of the attribute in the group. For a demographic d and feature f on the platform, we define penetration as

$$\text{penetration}_g(d, f) = \frac{n_g(d, f)}{n_g(d)}, \tag{1.1}$$

where $n_g(d)$ is the reach estimate (i.e. number of people targetable) by specifying demographic d alone on the platform. $n_g(d, f)$ 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 1.1 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

$$\text{affinity}_g(d, f) = \frac{n_g(d, f)}{n_g(d)} - \frac{n_g(f)}{N_g}. \tag{1.2}$$

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 1.1.3.

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,

$$\begin{aligned}\text{disparity}_g(d_1, d_2, f) &= \text{penetration}_g(d_1, f) - \text{penetration}_g(d_2, f) \\ &= \frac{n_g(d_1, f)}{n_g(d_1)} - \frac{n_g(d_2, f)}{n_g(d_2)}.\end{aligned} \tag{1.3}$$

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 would 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.

It also allows us to be more rigorous by performing statistical testing on the disparity and see whether the differences are statistically significant. For a set of targeting features F and a feature $f^* \in F$ with positive value of the disparity statistic $T = \text{disparity}_g(d_1, d_2, f^*)$, we are able to compute the p-value with a maximum likelihood estimate as

$$\begin{aligned}p &= \Pr[\ \text{disparity}_g(d_1, d_2, f) \geq T\ ] \\ &= \frac{\left|\{f \in F \mid \text{disparity}_g(d_1, d_2, f) \geq T\}\right|}{|F|}\end{aligned} \tag{1.4}$$

In our situation, F is the set of all features that we have gathered from Facebook's APIs. Similarly, for negative values of the disparity statistic, we look towards more extreme negative values to compute the p-value,

$$
\begin{aligned}
p &= \Pr[\ \text{disparity}_g(d_1, d_2, f) \leq T\ ] \\
&= \frac{\big|\{f \in F \mid \text{disparity}_g(d_1, d_2, f) \leq T\}\big|}{|F|}
\end{aligned} \tag{1.5}
$$

This simple but useful way of significance testing gives us a clear idea of the strength of any difference in feature disparities. 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. Moreover, we are able to do this without launching any malicious or harmful ads and only through the reach estimates provided in the Ads Manager.

Bibliography

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

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

[3] U.S. Department of Commerce. American Community Survey Information Guide. April 2017.