AUDITING RACE AND GENDER DISCRIMINATION IN ONLINE HOUSING MARKETS

BY

JOSHUA T ASPLUND

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Professor Karrie Karahalios

ABSTRACT

While researchers have developed rigorous practices for offline housing audits to enforce the Fair Housing Act, the online world lacks similar practices. In this work we lay out principles for developing an online fairness audit and demonstrate two examples; gender- and race-based discrimination in online housing advertisements, and personalized recommendation ordering. We employ a controlled sock-puppet audit technique to build online profiles associated with a specific demographic profile or intersection of profiles, and describe the requirements to train and verify profiles of other demographics. We also describe the process used to collect data for the two audits using these sock-puppet profiles. In the first we collect ads served on several sites in order to determine whether the number of housing-related ads served is dependent on the perceived race or gender of the profile. The second compares the ordering of personalized recommendations on major housing and real-estate sites. Using statistical tests, we examine whether the results seen in these areas exhibit indirect discrimination: whether there is correlation between the content served and users' protected features, even if the system does not know or use these features explicitly. We believe this framework provides a compelling foundation for further exploration of housing fairness online.

For the victims of bad algorithms and systematic oppression.

ACKNOWLEDGMENTS

To my beloved wife and best friend Anna, thank you for the sacrifices you have made and your constant support.

My sincere gratitude to my advisor, Professor Karrie Karahalios. Thank you for pushing me to accomplish more, learn new skills, and for supporting my mental health.

To Motahhare Eslami, Rick Barber, Professor Hari Sundaram, and the rest of our research group, for their assistance on my research, and for contributing a wide range of experiences and ideas that made these experiments possible.

To Ryan Cunningham, for mentoring me throughout my degree, and lighting a passion for ethics and policy that has changed my life.

And finally, to my family who have supported me throughout my entire life.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 RELATED WORK 3
2.1 Fairness Definitions 3
2.2 Housing Bias 4
2.3 Computational Audits 5

CHAPTER 3 EXPERIMENTAL DESIGN 8
3.1 Scope 8
3.2 Technical Details 10
3.3 Glossary of Terms 10

CHAPTER 4 PROFILE TRAINING 11
4.1 Site Selection 11
4.2 Profile Validation 12
4.3 Limitations 14

CHAPTER 5 ADVERTISING AUDIT 15
5.1 Data Collection 15
5.2 Ad Categorization 15
5.3 Results 17

CHAPTER 6 RANKING AUDIT 22
6.1 Data Collection 22
6.2 Pricing Results 23
6.3 Geographical Discrimination Results 27

CHAPTER 7 DISCUSSION 29
7.1 Alternate Definitions of Fairness 29
7.2 Concluding an Absence of Bias 29
7.3 Reproducibility 30
7.4 Consequences of Targeting 30
7.5 Detecting and Resolving Bias 31

CHAPTER 8 CONCLUSION 32

APPENDIX A SITE PLAYLISTS 33

REFERENCES 35

CHAPTER 1: INTRODUCTION

Online tracking and algorithmic personalization have opened up complicated new legal, ethical, and policy questions regarding bias and discrimination. As more and more information from users becomes available online, many platforms employ personalization algorithms to target information to users based on their usage behavior. While this might improve the efficiency of information consumption by users, it can also introduce bias along many dimensions including discrimination against a specific group of users (e.g. based on race and gender) [1]. The effects of such biases can perpetuate inequality, and even constitute illegal conduct in some areas [2].

Housing is one area where discrimination based on a protected class (including gender, race, color, disability, religion, familial status, or national origin) is subject to fairness regulations. In the United States of America, the Fair Housing Act (42 U.S.C. §3601-3619), often abbreviated FHA, makes it illegal for a landlord or seller to refuse to sell, rent to, or negotiate with a person due to their inclusion in a protected class.

In order to enforce these housing laws, the U.S. Department of Housing and Urban Development (HUD) performs regular audits to determine whether potential buyers or renters received differential treatment based on one of these protected classes. These studies are generally structured as a paired test [3, 4]. In this style of audit, two testers pose as prospective buyer or renters who differ only in protected classes. The participants reach out to real-estate agents and landlords asking to be shown houses on the market that fit specific criteria, and record which properties they are shown and how they are treated. Because all other attributes such as income are controlled, any difference in treatment is noted as possible discrimination.

The internet, however, has changed the the way people search for housing, making the traditional audits inadequate. Users now have access to listing information that was previously only available to real-estate agents or landlords. Significant portions of the housing search can now take place on property listing sites, housing search engines, and targeted advertising, removing traditional audit subjects from the process. Each of these can personalize their results to users, potentially introducing bias. This bias can be experienced by users in a variety of ways. For example, a specific listing or ad may not be shown to anyone that falls within specific demographic constraints. Members of one group may be given different information about price or availability. One group may receive significantly more advertisements for a property than another. Each of these actions could be in violation of the Fair Housing Act if performed based on a protected class.

The many forms of potential bias in online housing, along with the inadequacy of traditional housing audits, have resulted in calls for new audit techniques [2]. However, the complexity and opacity of the online algorithmic platforms that provide housing information to users makes auditing these platforms much more challenging than traditional housing providers. One of the main challenges to this style of audit is building accurate and realistic profiles. In traditional offline audits, researchers are able to control for protected attributes such as race and gender by selecting names and testers who match the desired characteristics. This type of user-provided profile information is not generally provided to housing websites. In many of these systems, such attributes are inferred indirectly via statistical models which are not available to outside researchers and auditors. In addition to controlling for protected attributes, it is also a challenge to control for a user's behavior in an online housing audit. Such challenges have made transferring offline housing audits to the online world quite difficult.

In this work we take steps towards building online housing audits to investigate gender-based and race-based discrimination in the context of online housing advertisements and search-result ranking. We demonstrate a sock-puppet browsing technique for building agents associated with a specific demographic profile, and discuss methods of validating that these profiles are classified correctly by advertising networks. We train profiles using this system to perform two audits of online housing systems. In the first audit we examine whether there is correlation between number of housing ads served to users and their protected features, regardless of the system's knowledge or use of these features. Our results show significant differences in the number of housing advertisements shown to agents of different races. In the second audit, we analyze differences in the way the listings are ordered, and show that a major property listing site's orders recommended listings differently depending on a visitor's gender. Finally, we discuss how the framework outlined in this paper can provide a basis for further exploration and audits.

CHAPTER 2: RELATED WORK

This work focuses on audits, defined here as systematic evaluations of organizations, networks, or marketplaces. These are generally performed by independent researchers, often treating the test subject as a black-box. These evaluations aim to prove, or disprove, that the system being tested meets formal standards.

In this paper we are concerned with whether the online housing markets provide fair outcomes for users regardless of gender or race.

2.1 FAIRNESS DEFINITIONS

The first step in this process is defining fairness and bias. Clearly and precisely defining "fairness" can be quite complicated as it varies between disciplines and contexts. A recent survey [5] identified more than 20 definitions of fairness. Many of these fit into larger categorizations by the method, target, or purpose of the discrimination. Previous auditing studies have particularly looked at two categorizations of fairness: *direct* vs. *indirect* and *individual* vs. *group* fairness [6, 7, 8, 9].

In the *direct/indirect* categorization of fairness, the focus is on the method of discrimination [10]. Direct discrimination occurs when rules or procedures explicitly impose "disproportionate burdens" on minority or disadvantaged groups. Indirect (or systematic) discrimination is caused by rules or procedures that impose disproportionate burdens without explicitly targeting the specific groups. This includes tactics such as redlining which are designed to discriminate based on racially homogeneous neighborhoods. In algorithmic systems, this categorization of fairness means whether a system takes a protected attribute as direct input or whether the output of the system correlates with some protected attributes, even if these attributes are not used as direct inputs of the system.

Individual/group fairness is based on how subjects are selected from the groups being studied [11]. Individual fairness requires that two similar individuals receive similar treatment. The counterpart, group fairness, requires that demographic subsets of the population are treated equally. These two measurements are not always equivalent, as deciding features may not be evenly spread throughout a population. Within the context of housing, individual fairness states that two prospective customers who differ only in protected traits should be shown the same properties and offered the same rates. Group fairness requires that, on average, a random member of one group should receive the same treatment as a random member of the other. These outcomes can differ based on the distribution of income, net

worth, or education within these groups.

Individual fairness provides the metric for many of the fairness laws in the United States of America: housing (42 U.S.C. §3601-3619), employment (42 U.S.C. §2000e), and public accommodations (42 U.S.C. §2000a). As such, audits focusing on these areas often use it as well. A recent study by Thebault-Spieker et al. of online reputations systems compared the ratings given to gig workers [12]. Participants were given a name, photo, and piece of work, and asked to rate the quality of the work. Because the same piece of work could be paired with any number of profiles, each of the workers is effectively equal, and any differences in the rating of the work can be attributed to bias based on the profile.

Group unfairness, and specifically indirect group unfairness, is often studied in situations where there are no existing fairness policies. A study in 2015 analyzed the effects of geography on digital crowdsourcing platforms [13]. It showed that the time it took to travel to a potential job and the relative wealth of the task area play a large part in the worker's decision to take a job. This rewards workers from areas with a higher socio-economic status since they do not have to travel the same distance, and therefore systematically discourages poorer and minority workers from the gig economy.

Throughout this paper we will define bias using the HUD definition of *disparate impact*. This states that "(the) illustrations of unlawful housing discrimination in this part may be established by a practice's discriminatory effect, even if not motivated by discriminatory intent" (78 FR 11459). Therefore, when we conclude that a system is biased we do not attribute any malice to that fact; it is a simple statement that the outcomes or effects are different across demographics.

In this paper, we investigate *indirect individual* fairness in online housing by evaluating whether two profiles that differ only in gender or race receive the same housing information.

2.2 HOUSING BIAS

Housing discrimination has been a topic of active research for more than half a century [3], but the introduction of the internet has substantially enlarged the area. Here we discuss both online and offline housing audits.

2.2.1 Traditional Housing Audits

Housing discrimination in non-internet contexts has been a topic of interest and research for decades. The first nation-wide HUD Housing Discrimination Survey (HDS) was performed in 1977, and additionally national studies have been performed in 1989, 2000, 2012

[3]. The 2012 audit showed continued decreases in overall discrimination, but still significant differences in the treatment of prospective renters and buyers along demographic lines.

The Fair Housing Act reaches beyond the search process as well. The places and methods of advertising housing must be fair. The initial interpretation of the law was that a seller cannot post an advertisement that directly excludes members of a protected class. A 1991 decision in *Ragin v. New York Times* (923 F.2d 995 (1991)) ruled that implicit discrimination fell within the law, and therefore that showing people of color primarily in service roles constituted illegal discrimination. Additionally, §109.25.a of the HUD regulations make it clear that limiting the reach of advertisements to demographically homogeneous areas falls afoul of the FHA as well.

2.2.2 Online Housing Bias

Recent studies of bias on the online rental site Airbnb [14, 15] show significant racial bias against prospective renters with distinctively African American names. The most recent [15] demonstrated that African American users were turned down for rentals 58 percent of the time, as opposed to Caucasian renters who were turned down only 50 percent of the time.

These results reaffirm previous studies on racial bias in online rentals both in the United States [16, 17, 18] and in Europe [19]. In each of these experiments, researchers emailed landlords to inquire about vacant apartments advertised online and signed the emails with names that implied the senders' race. In all of the studies, customers with traditionally non-English/American names received fewer responses overall and more negative responses. The prevalence of such biases in housing, along with the opacity of online algorithmic housing platforms, have raised requests for algorithmic housing audits during recent years [20, 21].

2.3 COMPUTATIONAL AUDITS

As discussed in the previous section, the principles of audits have begun to be applied to computational systems. These have not only been confined to online housing, but have expanded to diverse topics from health information to social networks to search engines.

Information security has presented several unique audit contexts due to the variety and value of the information involved. Previous papers have covered system designs to describe audit methods for structures from the underlying network topology [22] all the way up to a complex heterogeneous web service [23].

Privacy has also been a vital area for computational audits. [24] describe privacy audits for social network sites that would enable users to more easily determine how their data is being

shared [24]. Another paper by Bates et al. describes how government-sponsored wiretapping programs can be amended to allow open and transparent auditing from tamper-resistant logs [25]. Finally, [26] describe a system for auditing medical record, a system which is heavily protected by privacy law, but requires substantial oversight as medical records change hands [26].

2.3.1 Sock-puppet Audits

The methods demonstrated in this paper fall into a style known as "sock-puppet" audits. The term sock-puppet refers to automated systems that mimic realistic, but controlled, user behavior. This enables repeatable tests while still offering results generalizable to human users.

The AdFisher paper by Datta et al. [6] described a method of building user profiles to mimic profiles of differing demographics using automated browsing as well as explicit advertisement settings. Their experiments used these sock-puppets to explore a wide range of different topics including transparency and equality. We will discuss some of their finding in greater detail in a later sub-section.

Online marketplaces present a situation ripe for bias and discrimination since users have no ability to see the prices available to their peers. Hannk et al. described an audit strategy for detecting differential pricing online [27] using a combination of voluntarily submitted browsing profiles and algorithmically generated browsing behavior. Their experiments showed that more than half of the e-commerce sites audited performed some level of personalization.

The many types of potential bias in online systems, including price discrimination, have resulted in calls for auditing online algorithmic systems [28]. The two types of bias on which we have chosen to focus are *advertising bias* and *ranking bias*. Below, we briefly describe these biases and discuss existing audits in each space.

2.3.2 Auditing Advertising Bias

As was discussed in section 2.2.1, the contents and presentation of an advertisement can have discriminatory effects. Marketing dangerous or predatory topics to minority audiences can have profound effects on economic, physical, psychological, and societal health [29]. Empowering this with personalized online tracking and targeting may improve the advertisement's efficiency, but it can reinforce existing biases and inequalities by training algorithmic systems with biased data, or for biased outcomes.

For example, Google search results for stereotypically African American names return ads

containing the word "arrest" significantly more than for white names [30]. This creates both individual harms in circumstances such as job searches where employers routinely search for information about the applicant, but it also creates societal harms by reinforcing racist stereotypes that African Americans are more likely to commit crime.

In another example, Datta et al. found that women in their study received fewer ads encouraging them to take high paying jobs than men [6]. They attribute this to a few unique ads that were shown thousands of times over the course of the audit that were highly correlated with the gender of the user.

As discussed in Chapter 1, Facebook received a fair housing complaint against them in August 2018 (Assistant Secretary for Fair Housing & Equal Opportunity v. Facebook) for allowing advertisers to explicitly target customers that belong to demographic groups.

A very recent study by Ali et al. showed that modern advertising platforms (specifically Facebook) can cause discriminatory effects without the advertiser's input or awareness [31]. This presents a dangerous possibility of unwittingly reinforcing negative stereotypes that have been "learned" by a complex system.

2.3.3 Auditing Ranking Bias

The influence of search engines has greatly accelerated the study of ranking effects on user interaction. Users overwhelmingly favor the top results in search pages[32, 33]. Keane et al. demonstrated that users continued to click on the top links in Google search results even if the order of the results was reversed [34].

Because ordering can effect decision making so drastically, it is important that the ranking algorithm is fair. A 2018 study analyzed bias in resume search results [7]. It measured whether there were significant differences between the ranking of male and female job seekers based on inferred gender. The study concluded that the systems were generally fair on an individual basis, but that there was significant group bias against feminine candidates.

In this work, we investigate advertising and ranking bias in online housing by providing online advertising and housing listing platforms different inputs and evaluating the difference in outputs. This method involves many challenges including profile building, and controlling for protected attributes while imitating real users' behavior that we describe below.

CHAPTER 3: EXPERIMENTAL DESIGN

Our experiments aimed to answer the following questions:

- Does the perceived race and gender of a user searching for housing online affect the number and/or type of housing advertisements served to that individual? If so, why?
- Does the ranking algorithm for recommended properties on housing listing sites Trulia and Realtor.com weight property attributes differently based on the perceived race or gender of the visitor? If so, how do those weights change?
- Do the profile training techniques presented in this research produce advertising profiles that are significantly differentiated and reflect the desired attributes?

Our experimental design for the first two questions drew inspiration from previous audits in the space, especially the nation-wide HUD Housing Discrimination Survey [3, 2]. These follow the pattern of paired-testing discussed in chapters 1 and 2, where each test is performed by at least two individuals who differ only in protected classes.

Each audit or experiment consisted of hundreds of agents making dozens of measurements. An agent's life cycle consisted of two main steps: 1) training a browser profile and 2) collecting experimental data. Details on the methods of profile training and testing are covered in greater detail in chapter 4, and data collection in sections 5.1 and 6.1

In this section we will discuss the scope, technical details, and terminology of these audits.

3.1 SCOPE

There are a nearly limitless number of audits that could be performed in this area, considering the number of legally protected attributes, size of the United States housing market, and breadth of online services involved in housing. Even performing the simplest audits required limiting several of these variables.

3.1.1 Demographics

Each agent in the experiment was assigned an identity at the intersection of two demographic categories: race and gender. All other protected classes were controlled by removing any related topics so that additional variables would not confound the results.

The first set of categories we considered was gender. Agents were assigned to either the male or female profile training. We recognize that this does not account for transgender or

non-binary individuals. However, a lack of data on browsing behavior for these individuals made it prohibitively difficult to define a profile training strategy to include these individuals.

The other categorization was by racial or ethnic background, labeling agents as either Caucasian, African American, Hispanic, or Asian. These were chosen as they are the four largest racial and ethnic groups in the United States. This category is referred to as "race" throughout this paper, although there is active debate as to whether a Hispanic background is racial or ethnic in nature. We chose to use the term race because two-thirds of Hispanic Americans consider their background as part of their racial identity [35]. Additionally, this decision enabled simpler terminology across the audits.

3.1.2 Locations

This audit focused on two areas: the Chicago, Illinois metropolitan area and the Champaign/Urbana, Illinois area.

We chose to include Chicago in this audit due to both size and demographics. The city is one of the largest metropolitan areas in the United States, and therefore will have a very large number of properties available at any time. A smaller town or city may be unable to show the same level of personalization simply because there are fewer available properties for the site to display. Additionally, the Chicago area is both one of the most diverse and most segregated cities in the United States[36]. This increases verisimilitude because it is realistic that a person of that gender or race would be searching for a property in that area. It also presents a situation ripe for racial redlining, where landlords or sellers sort buyers into neighborhoods by race.

We included the Champaign/Urbana area in order to compare a smaller community to Chicago. It is still somewhat racially diverse, with Hispanic, African American, and Asian populations each comprising at least 10% of the population [37]. Furthermore, it is one of the largest metropolitan areas in Illinois after Chicago, and therefore has a housing market large enough for a variety of listings in different areas and at different price-points.

3.1.3 Timeline

Agents were run in blocks, where a mixture of differing profiles were trained and executed simultaneously. This was done to control for unintended effects that might be caused by changes to the site or network during the course of the experiment.

Blocks of training and collection were scheduled throughout the week and at varying times in order to minimize any data that might be inferred due to the day or time an agent visited

a page.

We started collecting results for the advertising audit in early August 2018, and started collecting ranking data in early December 2018. We continued collecting both through mid-March, however we limited the data we analyzed in the Advertising audit to ads collected before February 2019. We chose to limit the data due to the release of several large stories focusing of Facebook's handling of housing ads [38, 39]. Although these stories did not directly affect our data sources, there was still a possibility that the increased scrutiny in this area would cause the sites we were to testing to change their behavior and invalidate the study.

3.2 TECHNICAL DETAILS

At the beginning of the experiment each agent is represented as a fresh browser profile with no cookies or history. Next, the agent is assigned both a gender and race. It browses the internet following specific patterns in order to build a representative browser profile. Finally, the trained profile visits target sites and collects experimental data for analysis.

Each agent was represented by a browser profile in a discrete instance of Firefox 63 running on an Ubuntu 16.04 server. The browser instances were controlled by the Python programming language and the Selenium automation framework. All browsing traffic was routed through a proxy so that it would appear to originate in the correct geographic location.

3.3 GLOSSARY OF TERMS

agent An agent is an individual participant in the experiment. In a traditional audit this term would correspond to an experimenter performing the audit. In this framework, an agent refers to a unique browser profile.

profile When used within the context of profile training, the term profile refers to the set of data collected by an advertising network about a single user.

protected class A demographic category upon which it is illegal to discriminate. Under United States housing law these include gender, race, color, disability, religion, familial status, or national origin.

CHAPTER 4: PROFILE TRAINING

Because our audit follows the pattern of paired testing, we required a method of providing data about participants (such as age, gender, or race) to the subjects of the audit, online advertising networks and housing sites. This meant building a profile on these sites.

These profiles represent prospective home buyers who differ in gender and race, but are similar in location, income, and other important categories. This was done by browsing sites that were disproportionately visited by members of a specific demographic group. This method of inferred profile building was used for two reasons: necessity and verisimilitude.

First, there was no acceptable option for self-selection in the agent's advertising profiles. In earlier papers the experimenters had direct control over interests and visibility into inferred characteristics through tools such as Google Ad Settings [6] [40]. However, using these tools for audits has gotten significantly more difficult in the intervening years. Many sites, including Google, now require an account on the service to access these tools and explicitly forbid creating or using automated accounts in their Terms of Service. Additionally, in an effort to reduce possible bias, many advertisers have removed the ability to specify or view the demographic information we need. The combination of these factors made direct profile building infeasible.

Second, this method matches the way that advertisers would determine the interests of a user [41]. Most networks do not have detailed profile information supplied by the user and must rely on statistical inference. Using browsing to build profiles may reveal latent bias due to inferred data; an advertiser or network may make a decision based on a proxy attribute of a demographic group rather than the demographic itself. This kind of bias might be missed in an explicit profile, but would be caught with behavioral profiles.

4.1 SITE SELECTION

We gathered the representative sites from Quantcast, an internet analytics company that provides demographic breakdowns of site visitors in the United States. The site provides statistics on gender, age, income, ethnicity, and education [42, 43].

Relative popularity in Quantcast's metrics is measured through the demographic index [42]: the probability that a site visitor is a member of that demographic divided by the likelihood that a visitor to a random site is a member of that demographic. Therefore, a demographic index 100 for Hispanic visitors would say that there is no difference in the number of Hispanic visitors from what would be expected on any random site, where a 200

would demonstrate that twice as many visitors are Hispanic than average.

When selecting sites, we chose sites that had a demographic index above 140 for gender and ethnic minorities, or above 120 for Caucasian sites. We chose a second threshold for Caucasian users because a small increase in the demographic index for a significant majority group implies a sizable decrease in the number of minority visitors. More specifically, a demographic index for Caucasian visitors of 120 suggests that 90% of the site's visitors are Caucasian, significantly more than the 75% baseline.

In order to create the list of sites, or "playlist", for each demographic, researchers browsed the top 500 sites on Quantcast's rankings and selected all of those marked as directly measured by the network. We then discarded any that did not publish audience metrics. Finally, we sorted through and collected any sites that matched the demographic index threshold and put them on their corresponding lists.

We removed any sites that had a significantly large demographic index in other areas, specifically age and income. Income of the agents must be controlled since it is legal to discriminate in housing based on income, and therefore a strong correlation between a profile and high or low income could introduce legally acceptable bias into the results, invalidating any finding of illegal bias. Although age is a protected class, removing it from the profiles helps ensure that additional bias is not being introduced by an unmeasured variable.

A set of control sites were also selected from the top 100 most visited sites. These sites had minimal demographic skew, and therefore should not correlate with any specific race or gender. Separate control agents were run using these sites for training alongside the experimental agents to ensure that personalization occurred.

Finally, we verified that each site selected uses Google tracking on their homepages. In order to accomplish this, we tested whether the Google Ads script was loaded while rendering the homepage. This guaranteed that each page visit would be recorded by the advertising network under test.

Each agent visited between twelve and fourteen sites over three sessions during training. The agent stayed each site for roughly one minute and scrolled around the page in order to mimic human behavior. The exact time and behavior were randomized to avoid bot detection. In all, training a profile took roughly one hour.

4.2 PROFILE VALIDATION

One of the main challenges we had in training profiles was validating them; i.e. whether the assigned gender or race to a profile is correct. While there was no direct way to ensure

if a profile is completely accurate, we used a proxy method to measure the accuracy of the profile building process.

The key to this method is finding a feature that is independent from those used to train the profile, but correlated with the desired demographic. For example, if the desired demographic were college-aged students, measuring a significant increase in the number of ads for apartments advertised as "Close to Campus" would strongly indicate that the correct demographic profile had been learned.

This method has several strengths that make it desirable for these audits. For one, it is entirely indirect: it does not assume that the advertising system has explicitly labeled the desired demographic value, only that it knows correlations between interest groups, and that the profile fits within the desired set of connected interests.

Additionally, this method can be extended to new advertising networks with relative ease. Direct measurements through tools such as settings pages [6] require modifying the audit system for every new network. This proxy method can be updated and extended by verifying that the training pages visited and the ads or pages collected are covered by the new network.

4.2.1 Gender Validation

To validate whether the assigned gender to a profile was accurate, we investigated the ad category of apparel (clothing and jewelry) that the majority of its ads were clearly gendered, either by the style or model featured. We categorized these advertisements manually, looking up items if the intended gender was unclear. Any unclear items that were not explicitly gendered on the seller's website were left uncategorized.

Once all of the apparel related ads were categorized, we counted the numbers of male and female ads for each agent. We then performed an ANOVA test to determine whether there was a statistically significant correlation between the number of female ads and the gender of the agent. The test showed a statistically significant difference ($p = 0.048$), which supports our hypothesis that the advertising network learned the intended gender of the agents.

4.2.2 Race Validation

While we could not validate all categories of races, we used Spanish vs. English language to distinguish between Hispanic and other races. In order to do so, we categorized the Spanish ads and compared the frequency of these ads between Hispanic and non-Hispanic profiles. We found that while Hispanic profiles received some Spanish ads, none of the non-Hispanic

profiles ever received a Spanish ad. This, while not comprehensive, suggests that our profile training method seems to have distinguished between Hispanic and non-Hispanic races.

As discussed earlier, this method is a starting point to validate the profile building process in algorithmic audit techniques.

4.3 LIMITATIONS

While we believe that our methods are robust and accurate, the structure and ecosystem of online advertising present new difficulties when compared to previous audits.

The primary limitation of this experimental method is validating whether advertisers have inferred the desired data. In traditional audits the agents either met with real-estate agents in person or sent correspondence with personal information. This guaranteed that the person or group under audit knew the protected information regardless of whether they used it. Using direct methods such as Google Ad Settings allows a researcher to verify that the desired information was present in the listed interests. However, with inferred profiles it is difficult to verify that the correct attributes have been learned.

Another difficulty is undesired correlation during profile building. Individual fairness is not sensitive population-wide statistics, but profile systems are designed to make statistical deductions from known data. Therefore, if presented two profiles, one is Caucasian and the other African American but equal in all other known aspects, the advertiser may reasonably deduce that the African American user has a lower income. Although this may be true on average [44], this is illegal discrimination according to the FHA.

CHAPTER 5: ADVERTISING AUDIT

The first audit concerns housing-related ads served to users while browsing the internet. Our goal was to determine whether the number of housing related ads served to an agent was correlated with their gender or race.

5.1 DATA COLLECTION

Trained profiles searched for housing and collected the ads they were served. Three main categories of sites were chosen for ad collection: Google search results, local newspapers' online housing listings, and national news homepages. Each site serves ads through Google's AdSense network, which ensured that an agent's profile could be correlated across all the sites visited during training and data collection. Search results were chosen as a source of data due to the high proportion of relevant ads. A collection of different search terms were used by each agent, focusing on houses for sale or rent in the surrounding area. The terms included "houses for sale in [location]", "[location] rentals", and "houses in [location]". The local newspapers were chosen due to the relevant page topics and high number of display advertising slots served. The national news pages were chosen as they target a neutral audience, and because they provides a good baseline for the ads an agent would see on non-housing sites.

In addition to the text, image, or frame-grab of an ad, we collected relevant metadata including information about the search term used, time of day, target URL, and the agent that collected the ad. We did not collect the position of each ad because the framework code that retrieves the location of an element returned erroneous results.

5.2 AD CATEGORIZATION

The next step after visiting the aforementioned websites was finding and categorizing the housing ads.

Although the question of how many overall housing-related ads were seen is important, it gives very little information on how the users' treatment differed. It was important to have more detailed descriptions of the content of advertisement in order to better understand the mechanics of the advertising networks and investigate possible explanations for behavior.

To this end we defined categories of ads, focusing on of sellers and their relationship to the user. This was done so that we could determine if any particular class of advertiser

displayed bias, rather than measuring the entire ecosystem at once.

Ads were placed into one of the following categories:

- **Listing:**
  - **Sale Listing:** An ad to purchase a specific property or community. The ad must state either the address or name of the property, and it must be currently on the market.
  - **Rental Listing:** An ad to rent a specific property or community. Like listing ads it must give the property information in the ad. This includes short-term rentals.
- **Seller:**
  - **Realtor:** An advertisement for a real-estate agent, group, or property manager. These ads must be directly associated with a professional who can broker a property sale or rental.
  - **Rental Group:** An ad for a company or group that facilitates rentals, but does not own or manage the properties themselves. Companies like AirBnB fall in this category.
  - **Foreclosure:** Ads for search engines and sellers specializing in foreclosed properties. These ads use the term "foreclosed" or "bank-owned" property in their text.
  - **Rent-to-Own:** An ad for a rent-to-own seller. Any ad that used the term "rent-to-own" was included in this category, unless it contained a sale or rental listing that was currently for sale.
- **Listing Aggregator:** A search engine or site that collects other listings.
- **Property Search Engine:** A service that collects and searches property information, but is not primarily intended for buying/selling
- **Loans:** Ads that directly reference housing related loans such as mortgage services. Does not include credit card ads, savings accounts, or other general banking ads.
- **Unrelated:** Any ad that does not fall into one of the above categories.

Many papers have introduced novel techniques for determining which ads are relevant. One method described by Liu et al. is to leverage machine learning to classify advertisements [45].

Race	Gender	Agents	Ads
African American	Female	94	2540
	Male	92	2665
Asian	Female	94	2720
	Male	98	2729
Caucasian	Female	94	2630
	Male	92	2755
Hispanic	Female	95	2596
	Male	95	2668
	Control	12	437

Table 5.1: The number of agents in each demographic group, and the total number of ads seen by them.

They did this by detecting which ads differ in topic from the page content, and are therefore likely targeted. Another method was described in Adscape [41]. This leverages Google's AdSettings page with manual classification in order to estimate the proportion of ads served that were targeted to the user.

The number of important categories and inclusion of image and video ads made manual classification the most sensible method. Although it was more labor intensive than many alternatives, it allowed greater accuracy when dealing with subtle distinctions in the type of ads, as well as collecting information about the contents of non-text advertisements. Categorization was performed in batches by a single researcher, working from a list of unique advertisements collected.

5.3 RESULTS

We collected ads from 766 different trained profile agents each assigned to one of nine distinct treatments - roughly 95 of each of the eight experimental treatments and 12 of the control. The experiment collected $21,740$ ads in total, of which $2,262$ were unique. A detailed breakdown by demographic is included in table 5.1.

First, we filtered out any ads irrelevant to the specific topic of interest. Next we grouped the agents by gender and race and counted the number of ads seen of the specific category.

Once the ads had been counted, we compared the number of ads observed by each group using a two-way ANOVA test. This is a multi-level statistical test that can test the significance of several variables simultaneously. This made it a very good match for our data as we could test both demographic variables in a single experiment.

The ANOVA test requires the data to be normal and homogeneous in order to make

	Female	Male
African American	15.170	16.739
Asian	16.383	16.133
Caucasian	16.149	17.967
Hispanic	15.653	15.916

Table 5.2: The average number of housing-related ads seen per agent. Caucasian men see significantly more housing-related ads than any other group, and African American women see significantly fewer.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	4	1018.18	254.54	2.30	0.0569
Gender	1	273.89	273.89	2.48	0.1159
Race & Gender	3	388.93	129.64	1.17	0.3192
Residuals	1106	122331.01	110.61		

Table 5.3: The results of a two-way ANOVA test with the number of ads seen as dependent on the agent's gender and race. The p-value for race is not quite lower than the standard 0.05 level, but suggests that there may be some differential treatment, even if it is not statistically significant.

accurate conclusions, so this was verified before the statistical analysis. These assumptions were verified using the Fligner-Killeen test for homogeneity of variance, the Shapiro-Wilk normality test, and inspection of both the Residual vs. Fitted and Normal Q-Q plots.

If the data was satisfactorily well-behaved, we performed the ANOVA test and recorded the results. Since the data for several categories of ads did not match the underlying assumptions of the statistical test, we only make conclusions in categories where they were valid.

5.3.1 Overall Housing Ads

The first comparison was whether there was any bias in the overall housing-related ads. This combined all of the coded categories above, except for unrelated ads. We saw $12,418$ housing-related ads, with the average number seen per-demographic in table 5.2.

In general the values are clustered between 15.50 and 16.50, except for Caucasian men who saw substantially more housing ads, and African American women who saw substantially fewer. Although these numbers suggest there may be some difference in treatment, statistical analysis is needed to make a certain claim.

In table 5.3 we can see that the ANOVA test does not detect significant bias in any area at the $p = 0.05$ level. However, the p-score for race is just barely above at $p = 0.0569$. This, combined with the visual inspection of table 5.2 suggests that there may be some bias that

	Female	Male
African-American	0.617	0.620
Asian	0.585	0.510
Caucasian	0.532	0.652
Hispanic	0.558	0.642

Table 5.4: The average number of listing advertisements seen by each agent over the course of the audit. The male agents see more listings on average, except for Asian agents.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	3	5.29	1.76	0.79	0.4982
Gender	1	0.75	0.75	0.34	0.5610
Race & Gender	3	0.71	0.24	0.11	0.9559
Residuals	193	428.46	2.22		

Table 5.5: The results of a two-way ANOVA test with the agent's gender and race as independent variables and the number of listing advertisements as the dependent variable.

is being obscured by the number of degrees of freedom.

In order to investigate this further, we analyzed the data pair-wise: selecting one subset of the dataset and comparing it to its complement. Running the ANOVA test only comparing two racial-groups (Caucasian and non-Caucasian) shows a significant difference ($p = 0.033$) in treatment, while the other three permutations show a significance value close to the original. This demonstrates that, while there may not be significant differences when considering all variables at once, the treatment of Caucasian agents was significantly different than the treatment of others. Since the data for Caucasian visitors never made up more than one third of the data under consideration, this bias was not as clear to the original method of analysis.

5.3.2 Listing Ads

We see a different story with listing ads (sale listing and rental listing categories). We were able to collect 444 listing ads, with the breakdown of counts in table 5.4. As in table 5.2, Caucasian men saw the most advertisements of this type, however in this category Asian men saw the fewest. In general, male visitors saw more listing ads than their female counterparts, except for Asian visitors where the pattern reversed.

We do not see any significant bias in the number of listings using the multi-factor ANOVA test (table 5.5). Unlike the previous category, there is no evidence of bias found by pairwise tests either. While this is an encouraging result, we cannot conclude that the system is unbiased in this area without further analysis. This topic will be discussed at greater length

in section 7.2.

5.3.3 Rent-to-Own Ads

One area of housing loans that is particularly interesting is that of Rent-to-Own (RtO) properties. Also known as "contract-for-deed" arrangement, they promise the deed to a property (or the option to purchase the deed) upon completion of contract terms, usually including a set period of on-time payments and paying both taxes and upkeep for the property.

There has been significant discussion about the conditions of these agreements, with the various state investigators and the US Department of Defense describing them as a form of predatory lending [46, 47], but then concluding that this type of sale was not a credit sale, but a lease. They are generally targeted at low-income buyers, and previous lawsuits have alleged that these companies specifically target African American buyers and neighborhoods.

In our analysis we saw that African-American men received more ads for this type of loan than any other group. This conclusion is supported by the pairwise tests in tables 5.6 to 5.9 that show the respective ANOVA results for each test.

Looking at table 5.6, we can see that there is a significant correlation ($p = 0.007$) between the gender and race of the user and the number of RtO ads seen. This confirms that African-American men see more of these ads compared to other groups, and supports the conclusion that these dangerous lending practices are often targeted towards minorities.

5.3.4 Other Categories of Ads

We did not see any significant differences in treatment for any other categories of housing ads. This was due to a few limiting factors. First of all, several categories had very few ads, which made error bounds too large for reasonable conclusions. Secondly, several categories had bimodal distributions of views. In these, a few agents saw ads many times, but most agents never saw any. This violates many of the assumptions of the ANOVA test, and also compounded the first limitation. Finally, a few of the categories were sufficiently large, but failed one or more of the assumptions.

The first and third limitations may be addressed in the future with significantly larger sample sizes. The second, however, is likely to be a problem regardless of the sample size due to the behavior of online advertisements; ads tend to "follow" a user to increase effectiveness. Thus, alternative analysis may be necessary for situations where this is common.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	1	40.12	40.12	2.43	0.1226
Gender	1	4.72	4.72	0.29	0.5940
Race & Gender	1	125.30	125.30	7.59	0.0071
Residuals	88	1452.85	16.51		

Table 5.6: Pairwise comparison of the number of RtO ads seen. Comparing African-American users to all others.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	1	1.41	1.41	0.08	0.7824
Gender	1	2.61	2.61	0.14	0.7073
Race & Gender	1	2.81	2.81	0.15	0.6967
Residuals	88	1616.18	18.37		

Table 5.7: Pairwise comparison of the number of RtO ads seen. Comparing Asian users to all others.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	1	15.29	15.29	0.84	0.3609
Gender	1	2.28	2.28	0.13	0.7236
Race & Gender	1	10.01	10.01	0.55	0.4594
Residuals	88	1595.41	18.13		

Table 5.8: Pairwise comparison of the number of RtO ads seen. Comparing Caucasian users to all others.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	1	1.11	1.11	0.06	0.8050
Gender	1	2.68	2.68	0.15	0.7013
Race & Gender	1	23.41	23.41	1.29	0.2590
Residuals	88	1595.79	18.13		

Table 5.9: Pairwise comparison of the number of RtO ads seen. Comparing Hispanic users to all others.

CHAPTER 6: RANKING AUDIT

The other audit we performed was an analysis of whether the gender or race of a user is correlated with the ranking of recommended property search results. This analysis focused specifically on online listing services.

In order to select sites for this audit we performed an initial survey of several of the most popular online listing services, including Zillow, Trulia, Redfin, Realtor.com, Homesnap, and Forsalebyowner.com. A site could be considered only if it met a few criteria: First, it must have an algorthmically generated "recomended" sort. This immediately ruled out a few sites that only sorted based on concrete characteristics. Second, that sort must be personalized. Zillow and Redfin were removed as they did not exhibit any personalization. Both display pre-generated recommendations that were identical for every agent that visited the site. Finally, we chose sites that allowed scraping the pages of interest in their robots.txt files. This was to done in order to be "good neighbors" to these sites by following best practices. All of the remaining sites fulfilled this requirement.

This survey left us with two popular listing sites to audit: Trulia.com and Realtor.com. Both republish MLS and allow searching properties for rental or sale. Additionally, both default to a personalized sort and include Schema.org metadata in their search pages. This simplified data collection and reduced the number of requests to the respective sites.

6.1 DATA COLLECTION

In the data collection process, the trained browser profile visited the homepage and submitted the desired location using the main search bar. When the results were returned, the program verified that they were sorted using the personalized sort. On Trulia this is called "Just for you", and "Relevant Listings" on Realtor.com. If they were not sorted correctly, the desired sort was selected from the drop-down box. The program then scrolled down the page while extracting the metadata for each listing from the page's HTML. We collected the unique ids of the listings along with their addresses, latitude and longitude, number of bedrooms and bathrooms, and price.

We discarded any commercial properties or listings above $1.5 million dollars after verifying that there was no correlation between the number of properties of this kind and the gender or race of the user.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	3	7.1×10^{10}	2.4×10^{10}	0.211	0.8886
Gender	1	3.3×10^{11}	3.3×10^{11}	2.929	0.0870
Position	1	2.0×10^{13}	2.0×10^{13}	177.7	< 0.001
Gender, Race	3	4.0×10^{11}	1.3×10^{11}	1.181	0.3153
Race, Position	3	2.8×10^{11}	9.2×10^{10}	0.824	0.4805
Gender, Position	1	5.8×10^{11}	5.8×10^{11}	5.240	0.0221
Gender, Race, Position	3	1.8×10^{11}	5.9×10^{10}	0.532	0.6602
Residuals	6545	7.3×10^{14}	1.1×10^{11}		

Table 6.1: A three-way ANOVA test on the dataset collected from Realtor.com with an agent's gender and race and a listing's position as the independent variables, and the listing's price as the dependent variable. We can see a significant interaction between gender and position, implying that the rankings are ordered differently by price depending on the gender of the user.

6.2 PRICING RESULTS

We used a similar technique to analyze the rankings as we did the advertisements. For each listing service we performed a multi-factor ANOVA test with the profile variables (gender and race) and the ranking position as independent variables, and the listing attributes (price, number of bedrooms, distance) as the dependent variable.

We expect to see no direct correlation between either of the profile variables and the index, as the count and frequency of the indexes is identical between all measurements. The applicable tests for this audit are those that measure the correlation between one of the profile variables and one of the listing attributes. Statistically significant correlation of this type implies that the listing attribute is weighted differently depending on the user's demographics.

We also verified that personalization did not effect other attributes of the listings. Most importantly, we did not see differential pricing on any site we surveyed. This is a welcome departure from the 2012 HUD audit that found substantial price discrimination in one-on-one interactions [2].

6.2.1 Realtor.com

We collected 436 measurements from Realtor.com over the course of three months. Each of these records contains the metadata for each of the top 30 listings as well as the time collected, unique identifier of the browser profile, and the trained gender and race of the profile. In total we collected information on 789 distinct properties.

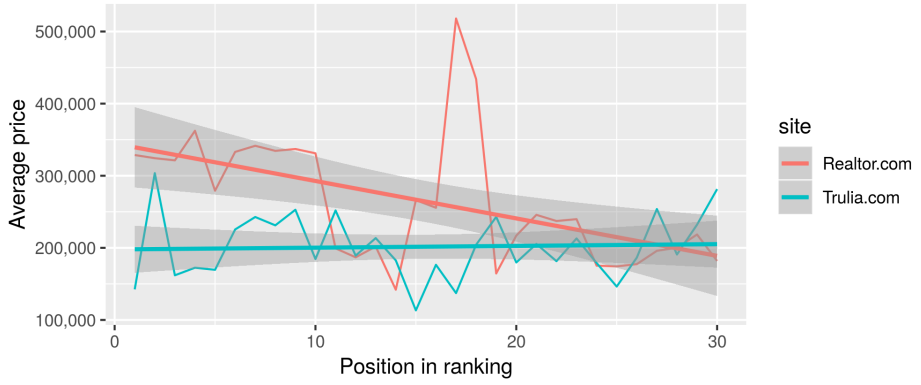

Figure 6.1: Average price of listings at a given index, comparing Realtor.com and Trulia.com. Notice that Realtor.com starts with more expensive properties and tends towards cheaper ones as you go down the list, where Trulia maintains a fairly stable average price.

	Female	Male
African American	378,123.30	343,618.80
Asian	338,620.70	364,870.80
Caucasian	395,230.90	317,609.40
Hispanic	359,624.90	344,917.60

Table 6.2: The average price (in USD) of listings seen on Realtor.com for housing searches in Chicago IL.

Next we performed a multi-factor ANOVA test on the provided data. The results of the test are contained in section 6.2.1. The most immediately obvious result is that the position of a listing is highly correlated with its price. We can see this correlation plotted in fig. 6.1. The effect of this is that the average price of a listing on Realtor.com decreases as you scroll down the suggestions.

We can also see that there is a statistically significant ($p = 0.022$) correlation between gender, position, and price in the results. This follows the strong correlation between position and price, but implies that men and women receive different orderings than men. While the results of this statistical test give us evidence of bias, they do not describe the ways in which the system treats users differently.

Looking at the correlation between gender, position, and ranking in fig. 6.2, we see that women are presented more expensive properties than men at the top of the list, but less expensive properties as they move down the list. The trend lines cross around position 15, half-way down the ranking.

	Female	Male
African American	224,921.70	224,172.60
Asian	226,058.70	220,708.50
Caucasian	220,961.40	227,589.70
Hispanic	217,756.30	230,006.20

Table 6.3: The average price (in USD) of listings seen on Realtor.com for housing searches in Champaign and Urbana IL.

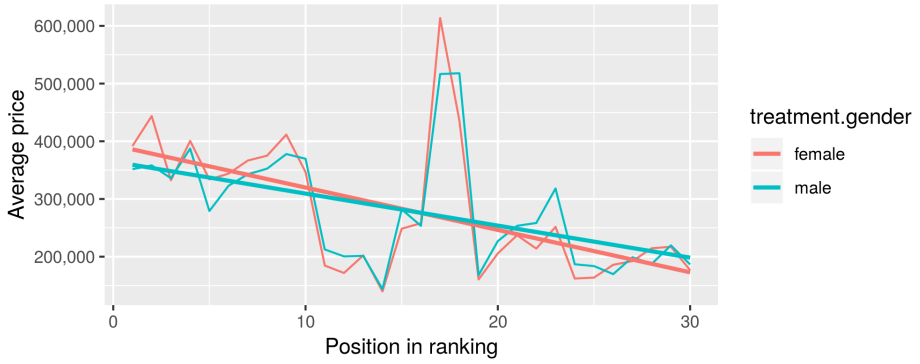

Figure 6.2: Comparison of average listing price at a given index by gender. While the trend-line of the average price for women starts above the one for men, the two cross at index 16.

We also see a large spike in average price at index 17. Upon inspection we discovered that this spot is often used for larger, more expensive properties such as penthouse apartments.

We can also look at the houses served to the agents. Table 6.2 lists the average listing price seen by each race for Chicago, and table 6.3 gives the averages for the Champaign/Urbana area. In Chicago we see that the trend of women being served more expensive properties holds, except for Asian users. Asian women were served significantly cheaper properties than their male counterparts.

This pattern reverses in the smaller town, however. In the Champaign/Urbana area men receive recommendations for slightly more expensive properties. Once again, the trend is swapped for Asian agents, with Asian women receiving more expensive recommendations.

6.2.2 Trulia.com

On Trulia we collected 118 measurements over the course of one and a half months. Like Realtor.com, we collected the listing information as well as metadata about the scraping

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Race	3	6.3×10^9	2.1×10^9	0.14	0.9372
Gender	1	9.7×10^8	9.7×10^8	0.06	0.8013
Position	1	1.9×10^{11}	1.9×10^{11}	12.63	0.0004
Gender, Race	3	1.3×10^{10}	4.3×10^9	0.28	0.8375
Race, Position	3	3.0×10^9	1.0×10^9	0.07	0.9780
Gender, Position	1	1.4×10^9	1.4×10^9	0.09	0.7605
Gender, Race, Position	3	3.4×10^9	1.1×10^9	0.07	0.9742
Residuals	9274	1.4×10^{14}	1.5×10^{10}		

Table 6.4: A three-way ANOVA test on the dataset collected from Trulia.com with an agent's gender and race and a listing's position as the independent variables, and the listing's price as the dependent variable.

	Female	Male
African American	211,871.50	214,731.00
Asian	211,399.20	217,588.40
Caucasian	215,972.70	213,385.90
Hispanic	212,214.00	214,387.50

Table 6.5: The average price (in USD) of listings seen on Trulia.com for housing searches in Champaign and Urbana IL.

profile.

The results of the multi-factor ANOVA test are shown in table 6.4. Our results do not show evidence of any statistically significant discrimination on either gender or bias.

Although we do not see significantly different rankings between different races, there are still some differences between the listings seen. For example, tables 6.5 and 6.6 show the average house price by race. In contrast with the Realtor.com results, we see the largest difference between the Asian and Hispanic agents, with the others falling between. The higher average cost for Asian agents is especially interesting as it is far more different than the average than any of the others.

	Female	Male
African American	1,394,098.00	1,320,769.00
Asian	1,215,298.00	1,115,562.00
Caucasian	1,258,380.00	1,316,573.00
Hispanic	1,313,913.00	1,219,628.00

Table 6.6: The average price (in USD) of listings seen on Trulia.com for housing searches in Chicago IL.

	African American	Asian	Caucasian	Hispanic	p-value
Douglas	18	12	16	16	0.747
Calumet Heights	12	10	12	11	0.970
Kenwood	10	10	9	10	0.994
Washington Park	20	21	22	20	0.987
West Lawn	12	11	12	12	0.996

Table 6.7: A comparison between Douglas, the most racially disparate neighborhood in the Trulia results, compared to a random sample of four other neighborhoods from that dataset. We can see that listings in this area were seen less by Asian agents than any other racial group.

	African American	Asian	Caucasian	Hispanic	p-value
Lake View	8	15	19	21	0.099
Loop	8	7	9	8	0.969
Norwood Park	5	4	6	5	0.940
Cicero	8	10	5	5	0.630
Chatham	9	11	11	10	0.966

Table 6.8: Comparison between neighborhoods comparing all 30 listings collected for each page

6.3 GEOGRAPHICAL DISCRIMINATION RESULTS

Another area of interest for this audit was whether the race of a user influenced the neighborhoods they were recommended. This type of bias can be thought of as a companion to redlining, where lenders restrict mortgages in certain neighborhoods on the basis of race [48, 49]. This leads to indirect segregation: members of minority groups are technically allowed to move outside their neighborhoods, but are prevented from doing so because they are unable to find funding. This results in racially segregated communities as well as depriving minority groups from building greater net worth [50, 51].

This portion of the audit uses the same set of data as the other ranking experiments, but a different statistical test. Because we are interested in how many users are recommended each neighborhood, we use a simple χ^2-test rather than the ANOVA. We compare the actual distribution for each neighborhood to the expected distribution, and if the p-value is lower than the threshold $p = 0.05$ we conclude that there is biased treatment.

We saw very little variance in the number of times a neighborhood was seen by each racial group on Trulia. The χ^2 tests summarized in table 6.7 showed that the lowest p-value is the Douglas neighborhood, with $p = 0.7468$. Thus there is no evidence of geographic discrimination for Trulia.com.

We see a very different story from Realtor.com. Looking at all of the listings in table 6.8,

	African American	Asian	Caucasian	Hispanic	p-value
Lake View	7	15	19	21	0.060
Loop	7	5	8	7	0.872
Norwood Park	5	4	6	5	0.940
Cicero	8	10	5	5	0.624
Chatham	9	11	11	10	0.974

Table 6.9: Comparison between neighborhoods comparing the top 15 listings collected for each page

we see that there is one neighborhood, Lake View, that that is significantly different at the $p = 0.1$ level. Further restricting to the top 15 listings in the ranking (table 6.9, this value decreases further, implying that the neighborhood is not only shown less to African American users, but it is shown further down as well.

Although this is not a conclusive demonstration of geographical discrimination, it does open the door for further exploration in this area.

CHAPTER 7: DISCUSSION

In this paper we described two audits on online housing markets; the first studying online ad targeting and the second measuring the effects of personalized search result ranking.

In the first audit we found differences in the number of housing ads shown on the basis of race, and concluded that Caucasian agents saw significantly more housing-related advertisements than other agents who searched for housing. This implies that the advertiser's target audiences were more closely aligned with the Caucasian group, leading to indirect bias. The first audit also found that advertisements for predatory Rent-to-Own programs were seen by African American users significantly more than any other group.

The second audit uncovered evidence of gender bias in the ordering of suggested properties on the online housing site Realtor.com. Women were recommended more expensive properties towards the top of the list, and less expensive properties towards the bottom. We also concluded that African American users were significantly less likely to be recommended houses in the Lake View neighborhood of Chicago. While this may be unintentional, techniques like this have been used in the past to enforce indirect segregation by pricing minorities out of desirable neighborhoods [52].

While our results show some bias in these systems, we do not claim that any laws have been broken. Further audits would be required to show a pattern of systematic bias in one of these platforms.

7.1 ALTERNATE DEFINITIONS OF FAIRNESS

The Fair Housing Act and existing housing audits focus specifically on individual fairness. However, this definition of fairness does not consider the ways that systematic bias based on a protected class can create inequalities in other areas. The 2012 Housing Discrimination Study notes that paired-testing studies require choosing non-representative testers from their respective demographics, and that this may mask group unfairness [2]. Indirect audits such as the one described here can be modified to measure this by favoring representative profiles over paired profiles.

7.2 CONCLUDING AN ABSENCE OF BIAS

While we can compare differences in the two results, our current statistical methods cannot conclude that there is no bias in this system. This follows the historical focus on the presence

of bias rather than its absence. A 2017 paper by Thebault-Spieker et al. demonstrated methods of proving the absence of bias in online 5-star ranking systems [12]. While their methods are not directly applicable to this type of audit, the integrating similar ideas could greatly strengthen future housing audits.

We did not find any significant evidence of bias in the ranking of properties on Trulia.com. Once again, we cannot conclude that the system is not biased, but the lack of obvious discrimination is an important result. The goal of any fairness audit is to detect bias so that it can be resolved, and negative results are an important part of that process.

7.3 REPRODUCIBILITY

One weakness of these audits is a lack of reproducibility. Practically every factor in these systems is constantly evolving, from the set of ads currently being served, to the targeting and pricing of an advertising campaign, and even the way user profiles are interpreted. This puts researchers in a difficult position: auditors must collect as much data as possible in order to catch any confounding variables, and must carefully validate that the system they are measuring did not change substantially during the course of their audit.

Researchers looking to replicate the findings of an audit like these can use the same tools and methods as the original experiment, but should be very aware that any number of variables may have changed between their measurements and the originals, and therefore lead to very different results.

7.4 CONSEQUENCES OF TARGETING

It is important for online advertising companies to consider whether targeted housing ads are desirable for their business. These ads open up an area of legal and ethical liability that may not be offset by potential benefits. This has been demonstrated recently in two complaints against Facebook, one of the largest online advertising companies; a complaint by HUD (Assistant Secretary for Fair Housing & Equal Opportunity v. Facebook) and a lawsuit by the National Fair Housing Alliance (18 Civ. 2689). These complaints focus on the platform's tools which allow potential advertisers to select which demographics to include or exclude in their audience, and demonstrate that advertisers can discriminate directly by not showing ads based on gender, disability, familial status, religion, and/or national origin.

Additionally, a recent study by Ali et al. [31] showed that explicit discrimination is not necessary to produce prejudiced outcomes. In the study they made several ad purchases

on Facebook in several employment and housing related topics, and then tracked the demographics of the users who saw their ads. Although they did not target these ads to any specific demographic, they saw significant skew in the demographics of the users who received the ads. For example, ads for jobs in the lumber industry reached an audience that was 72% white and 90% male, while ads for cashier positions reached 75% African-American audience. Even more concerning is the fact that the discriminatory appears to be a byproduct of automated classification and image recognition. This creates a system where bias is practically inevitable, whether the advertiser desires it or not.

This type of bias is unlikely to be confined to Facebook. In fact, Facebook may have the best tooling to both find and reduce this type of discrimination due to their detailed user profiles. More opaque advertising systems, such as Google's, require indirect tests such as the ones described in this paper since the platforms do not report on the demographics of the users who received an ad.

7.5 DETECTING AND RESOLVING BIAS

Many sites address the liability caused by personalized recommendations by serving an algorithmically sorted list to all visitors and refreshed regularly. This guarantees that two users with the same query at the same time will receive the same recommendations. For areas where fairness is legally required, such as housing and employment, this can be a prudent alternative. The algorithm may still perpetuate existing bias by under- or over-representing certain neighborhoods and communities, but the system as a whole would be fair under this style of paired-testing audit.

If a service provider chooses to serve targeted housing ads or listings, it is important that they regularly audit their own systems for fairness. This is especially important if the service uses opaque or uninterpretable decision-making systems. These would preferably be performed by an unbiased third-party, but could also be accomplished by an isolated team within the company.

Finally, companies should open up access to their systems for fully independent auditing. As it currently stands, external audits not officially sanctioned by the company run the risk of violating the site's Terms of Service, especially if the researchers access the site using automated tools. As long as these sites are not open to external experiments, the public has no way to verify that they meet the legal standard of fairness.

CHAPTER 8: CONCLUSION

Our two audits investigate important emerging areas of housing fairness: online advertising and search-result ranking. While in some of the explored categories, our results demonstrate that a user's treatment is dependent on a protected class (either race or gender), there are also many categories that we did not find any types of bias. We argue that both of these results are equivalently important as both serve the goal of conducting an audit.

The proposed audit framework can provide a foundation for further audits going forward. The most recent Housing Discrimination Study [2] surveyed 28 metropolitan areas across the United States of America in order to provide a conservative measure of housing discrimination in the county. An online audit of this scale would certainly be a sizeable undertaking, but the lower requirements for participants, travel, and other logistical concerns could make it feasible to conduct on a more regular basis than the roughly ten year pattern of existing audits.

The profile building technique we have used in our auditing platform is also adaptable. An auditor can add new categories or variables simply by changing the sites visited, the training schedule, or the agent's behavior while browsing a page. Even attributes such as the IP address and browser used can be modified in order to build a more representative profile. These behavioral models could be further strengthened by user studies providing truly representative browsing data.

APPENDIX A: SITE PLAYLISTS

Listing A.1: Control Sites

```
amazon.com
google.com
facebook.com
wikipedia.org
twitter.com
youtube.com
ebay.com
yahoo.com
walmart.com
quora.com
craigslist.org
nytimes.com
paypal.com
target.com
bing.com
```

Listing A.2: Male

```
cnbc.com
gizmodo.com
deviantart.com
drudgereport.com
politico.com
deadspin.com
jalopnik.com
theverge.com
bodybuilding.com
gfycat.com
thechive.com
stackoverflow.com
washingtonexaminer.com
kotaku.com
polygon.com
tiebreaker.com
sbnation.com
androidcentral.com
westernjournal.com
flightaware.com
streamable.com
comicbook.com
liveleak.com
```

Listing A.3: Female

```
goodreads.com
thoughtcatalog.com
quizlet.com
thekitchn.com
thepennyhoarder.com
king.com
hometalk.com
yourtango.com
findagrave.com
petfinder.com
sixflags.com
romper.com
patient.info
apartmenttherapy.com
getitfree.us
```

Listing A.4: African American

```
urbandictionary.com
nickiswift.com
bodybuilding.com
quizlet.com
theroot.com
tiebreaker.com
jezebel.com
sbnation.com
yourtango.com
ajc.com
getitfree.us
```

Listing A.5: Asian

```
yelp.com
urbandictionary.com
glassdoor.com
stackexchange.com
cnbc.com
```

```
gizmodo.com
deadspin.com
lifehacker.com
variety.com
jalopnik.com
theverge.com
bodybuilding.com
gfycat.com
quizlet.com
thekitchn.com
stackoverflow.com
kotaku.com
sbnation.com
flightaware.com
androidcentral.com
streamable.com
apartmenttherapy.com
```

Listing A.6: Caucasian

```
whitepages.com
definition.org
topixoffbeat.com
kiwireport.com
worldlifestyle.com
startribune.com
cheatsheet.com
hometalk.com
westernjournal.com
findagrave.com
yourdailydish.com
petfinder.com
gardeningknowhow.com
dailykos.com
```

Listing A.7: Hispanic

```
urbandictionary.com
theverge.com
bodybuilding.com
gfycat.com
quizlet.com
kotaku.com
univision.com
polygon.com
yourtango.com
sixflags.com
androidcentral.com
axs.com
streamable.com
comicbook.com
liveleak.com
```

REFERENCES

[1] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen, "Proxy Non-Discrimination in Data-Driven Systems," *arXiv:1707.08120 [cs]*, Jul. 2017, arXiv: 1707.08120. [Online]. Available: http://arxiv.org/abs/1707.08120

[2] M. Austin Turner, S. Rob, K. Levy Diane, D. Wissoker, C. Aranda, and R. Pitingolo, "Housing discrimination against racial and ethnic minorities 2012," *Washington, DC: US Department of Housing and Urban Development*, 2013.

[3] R. E. Wienk, United States., and National Committee Against Discrimination in Housing., *Measuring racial discrimination in American housing markets: the housing market practices survey.* Washington: Division of Evaluation, Office of Policy Development and Research, U.S. Dept. of Housing and Urban Development, 1979, no. ca. 350 p. [Online]. Available: //catalog.hathitrust.org/Record/000754490

[4] A. F. Schwartz, *Housing Policy in the United States.* Routledge, Aug. 2014. [Online]. Available: https://www.taylorfrancis.com/books/9781135045234

[5] A. Narayanan, "21 fairness definitions and their politics," *New York, NY, USA*, 2018.

[6] A. Datta, M. C. Tschantz, and A. Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," *arXiv:1408.6491 [cs]*, Aug. 2014, arXiv: 1408.6491. [Online]. Available: http://arxiv.org/abs/1408.6491

[7] L. Chen, R. Ma, A. Hannk, and C. Wilson, "Investigating the Impact of Gender on Rank in Resume Search Engines," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018. [Online]. Available: http://doi.acm.org/10.1145/3173574.3174225 pp. 651:1–651:14.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *arXiv:1610.02413 [cs]*, Oct. 2016, arXiv: 1610.02413. [Online]. Available: http://arxiv.org/abs/1610.02413

[9] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *International Conference on Machine Learning*, Feb. 2013. [Online]. Available: http://proceedings.mlr.press/v28/zemel13.html pp. 325–333.

[10] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD*, Y. Li, B. Liu, and S. Sarawagi, Eds. Las Vegas, Nevada, USA: ACM, 2008. [Online]. Available: https://doi.org/10.1145/1401890.1401959 pp. 560–568.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2090236.2090255 pp. 214–226.

[12] J. Thebault-Spieker, D. Kluver, M. A. Klein, A. Halfaker, B. Hecht, L. Terveen, and J. A. Konstan, "Simulation Experiments on (the Absence of) Ratings Bias in Reputation Systems," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–25, Dec. 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3171581.3134736

[13] J. Thebault-Spieker, L. G. Terveen, and B. Hecht, "Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '15. New York, NY, USA: ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2675133.2675278 pp. 265–275.

[14] B. G. Edelman and M. Luca, "Digital Discrimination: The Case of Airbnb.com," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2377353, Jan. 2014. [Online]. Available: https://papers.ssrn.com/abstract=2377353

[15] B. Edelman, M. Luca, and D. Svirsky, "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics*, vol. 9, no. 2, pp. 1–22, Apr. 2017. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/app.20160213

[16] A. G. Carpusor and W. E. Loges, "Rental Discrimination and Ethnicity in Names," *Journal of Applied Social Psychology*, vol. 36, no. 4, pp. 934–952, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0021-9029.2006.00050.x

[17] B. Hogan and B. Berry, "Racial and Ethnic Biases in Rental Housing: An Audit Study of Online Apartment Listings," *City & Community*, vol. 10, no. 4, pp. 351–372, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6040.2011.01376.x

[18] A. Hanson and M. Santas, "Field Experiment Tests for Discrimination against Hispanics in the U.S. Rental Housing Market," *Southern Economic Journal*, vol. 81, no. 1, pp. 135–167, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.4284/0038-4038-2012.231

[19] A. M. Ahmed and M. Hammarstedt, "Discrimination in the rental housing market: A field experiment on the Internet," *Journal of Urban Economics*, vol. 64, no. 2, pp. 362–372, Sep. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094119008000181

[20] A. M. Ahmed, L. Andersson, and M. Hammarstedt, "Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants?" *Land Economics*, vol. 86, no. 1, pp. 79–90, Feb. 2010. [Online]. Available: http://le.uwpress.org/content/86/1/79

[21] A. Chander, "The Racist Algorithm 2017 Survey of Books Related to the Law: Reviews," *Michigan Law Review*, vol. 115, pp. 1023–1046, 2016. [Online]. Available: https://heinonline.org/HOL/P?h=hein.journals/mlr115&i=1084

[22] B. E. Ujcich, A. Miller, A. Bates, and W. H. Sanders, "Towards an accountable software-defined networking architecture," in *2017 IEEE Conference on Network Softwarization (NetSoft)*, Jul. 2017, pp. 1–5.

[23] A. Bates, W. U. Hassan, K. Butler, A. Dobra, B. Reaves, P. Cable, T. Moyer, and N. Schear, "Transparent Web Service Auditing via Network Provenance Functions," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, event-place: Perth, Australia. [Online]. Available: https://doi.org/10.1145/3038912.3052640 pp. 887–895.

[24] A. Gutierrez, A. Godiyal, M. Stockton, M. LeMay, C. A. Gunter, and R. H. Campbell, "Sh@re: Negotiated audit in social networks," in *2009 IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2009, pp. 74–79.

[25] A. Bates, K. R. B. Butler, M. Sherr, C. Shields, P. Traynor, and D. Wallach, "Accountable wiretapping or I know they can hear you now," *Journal of Computer Security*, vol. 23, no. 2, pp. 167–195, Jan. 2015. [Online]. Available: http://content.iospress.com/articles/journal-of-computer-security/jcs515

[26] S. E. Oh, J. Y. Chun, L. Jia, D. Garg, C. A. Gunter, and A. Datta, "Privacy-preserving Audit for Broker-based Health Information Exchange," in *Proceedings of the 4th ACM Conference on Data and Application Security and Privacy*, ser. CODASPY '14. New York, NY, USA: ACM, 2014, event-place: San Antonio, Texas, USA. [Online]. Available: http://doi.acm.org/10.1145/2557547.2557576 pp. 313–320.

[27] A. Hannk, G. Soeller, D. Lazer, A. Mislove, and C. Wilson, "Measuring Price Discrimination and Steering on E-commerce Web Sites," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, ser. IMC '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2663716.2663744 pp. 305–318.

[28] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," *Data and discrimination: converting critical concerns into productive inquiry*, p. 23, 2014.

[29] R. D. Petty, A.-M. G. Harris, and T. Broaddus, "Regulating Target Marketing and Other Race-Based Advertising Practices," *Mich. J. Race & L.*, vol. 8, p. 61, 2002.

[30] L. Sweeney, "Discrimination in Online Ad Delivery," *Queue*, vol. 11, no. 3, pp. 10:10–10:29, Mar. 2013. [Online]. Available: http://doi.acm.org/10.1145/2460276.2460278

[31] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, "Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes," *arXiv:1904.02095 [cs]*, Apr. 2019, arXiv: 1904.02095. [Online]. Available: http://arxiv.org/abs/1904.02095

[32] Z. Guan and E. Cutrell, "An Eye Tracking Study of the Effect of Target Rank on Web Search," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: ACM, 2007. [Online]. Available: http://doi.acm.org/10.1145/1240624.1240691 pp. 417–420.

[33] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An Experimental Comparison of Click Position-bias Models," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1341531.1341545 pp. 87–94.

[34] M. T. Keane, M. O'Brien, and B. Smyth, "Are People Biased in Their Use of Search Engines?" *Commun. ACM*, vol. 51, no. 2, pp. 49–52, Feb. 2008. [Online]. Available: http://doi.acm.org/10.1145/1314215.1314224

[35] P. R. Center, "Multiracial in America: Proud, diverse and growing in numbers," *Pew Research Center Social & Demographic Trends*, 2015.

[36] N. Silver, "The most diverse cities are often the most segregated," 2015. [Online]. Available: https://fivethirtyeight.com/features/the-most-diverse-cities-are-often-the-most-segregated/

[37] "U.S. Census Bureau QuickFacts: Champaign County, Illinois." [Online]. Available: https://www.census.gov/quickfacts/champaigncountyillinois

[38] S. Biddle, "Facebooks Ad Algorithm Is a Race and Gender Stereotyping Machine, New Study Suggests," Apr. 2019. [Online]. Available: https://theintercept.com/2019/04/03/facebook-ad-algorithm-race-gender/

[39] A. Robertson, "What happens next in the housing discrimination case against Facebook?" Apr. 2019. [Online]. Available: https://www.theverge.com/2019/4/2/18286660/facebook-hud-housing-discrimination-case-section-230-legal-defense

[40] C. E. Wills and C. Tatar, "Understanding What They Do with What They Know," in *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, ser. WPES '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2381966.2381969 pp. 13–18.

[41] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and Analyzing Online Display Ads," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2566486.2567992 pp. 597–608.

[42] S. Simpson, "Glossary," Mar. 2019. [Online]. Available: http://help.quantcast.com/hc/en-us/articles/115013851427-Glossary

[43] S. Simpson, "Reading Our Audience Measurement Reports," Sep. 2018. [Online]. Available: http://help.quantcast.com/hc/en-us/articles/115014120368-Reading-Our-Audience-Measurement-Reports

[44] U.S. Census Burea, "Median Household Income by Race," 2018. [Online]. Available: https://www.census.gov/content/dam/Census/library/visualizations/2018/demo/p60-263/figure1.pdf

[45] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "AdReveal: improving transparency into online targeted advertising," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. College Park, MD, USA: ACM, Nov. 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2535771.2535783 p. 12.

[46] M. Goldstein and A. Stevenson, "Contract for Deed Lending Gets Federal Scrutiny - The New York Times," May 2016. [Online]. Available: https://www.nytimes.com/2016/05/11/business/dealbook/contract-for-deed-lending-gets-federal-scrutiny.html

[47] US Department of Defense, "Report on Predatory Lending Practices Directed at Members of the Armed Forces and Their Dependents," 2006.

[48] Y. Zenou and N. Boccard, "Racial Discrimination and Redlining in Cities," *Journal of Urban Economics*, vol. 48, no. 2, pp. 260–285, Sep. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094119099921666

[49] V. J. Roscigno, D. L. Karafin, and G. Tester, "The Complexities and Processes of Racial Housing Discrimination," *Social Problems*, vol. 56, no. 1, pp. 49–69, Feb. 2009. [Online]. Available: https://academic.oup.com/socpro/article/56/1/49/1644423

[50] K. E. Henkel, J. F. Dovidio, and S. L. Gaertner, "Institutional Discrimination, Individual Racism, and Hurricane Katrina," *Analyses of Social Issues and Public Policy*, vol. 6, no. 1, pp. 99–124, Dec. 2006. [Online]. Available: https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-2415.2006.00106.x

[51] F. L. Pincus, "Discrimination Comes in Many Forms: Individual, Institutional, and Structural," *American Behavioral Scientist*, vol. 40, no. 2, pp. 186–194, Nov. 1996. [Online]. Available: https://doi.org/10.1177/0002764296040002009

[52] A. T. King and P. Mieszkowski, "Racial Discrimination, Segregation, and the Price of Housing," *Journal of Political Economy*, vol. 81, no. 3, pp. 590–606, May 1973. [Online]. Available: https://www.journals.uchicago.edu/doi/abs/10.1086/260060