



Case Studies in Public Interest Technology

Citation

Zang, Jinyan. 2021. Case Studies in Public Interest Technology. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368422>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of **Government**
have examined a dissertation entitled

“Case Studies in Public Interest Technology”

presented by **Jinyan Zang**

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature *Latanya Sweeney*
Latanya Sweeney (May 7, 2021 13:02 EDT)

Typed name: Prof. Latanya Sweeney (Chair)

Signature *Jim Waldo*

Typed name: Prof. James H. Waldo

Signature *Nicol Turner-Lee*
Nicol Turner-Lee (May 7, 2021 12:58 EDT)

Typed name: Dr. Nicol Turner-Lee (Brookings Institution)

Date: **May 7, 2021**

Case Studies in Public Interest Technology

A dissertation presented

by

Jinyan Zang

to

The Department of Government

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Political Science

Harvard University

Cambridge, Massachusetts

May 2021

© 2021 Jinyan Zang

All rights reserved.

Case Studies in Public Interest Technology

Abstract

Today, there are multiple ways where digital technologies adversely impacts the public interest, whether that's the spread of misinformation online, the loss of privacy, the threat of algorithmic discrimination, and more. Public interest technology is an emerging field that seeks to use cross-disciplinary techniques to research and address these issues in order to advance the public interest.

For this dissertation, I present three different case studies of public interest tech research projects, each of which focuses on a different technology and relevant public interest. In Chapter 2, I research how Facebook's advertising algorithms can discriminate by race and ethnicity. In Chapter 3, I test how the predictability of Social Security Number (SSN) assignment based on easily accessible data about Americans presents a risk of identity theft. In Chapter 4, I demonstrate how TraceFi, a Wi-Fi based collocation detection technology, can be deployed for COVID-19 contact tracing.

In this dissertation, I propose how we can adapt Lawrence Lessig's pathetic dot model as the "Three Forces Model of Public Interest Tech" to understand the current dysfunctional state of relationships between technology, society, and the public interest, where the public interest is often affected as an output of technology but not fully considered as an input. The three forces of the law, norms, and market can affect a given technology or vice versa which in turn affects the public interest. For different combinations of technologies and public interests, the amount of force exerted by the

law, norms, or market could also differ and so could the degree of feedback between the technology and each of the forces.

Since the normative goal of public interest tech as a field is to ultimately advance the public interest, the goal state of the Three Forces Model demonstrates how the public interest can be an input for the law, norms, and market in how they affect a technology's design and usage, which would in turn affect the public interest. Stakeholders relevant to each of the forces can consider the public interest as a priority in how they interact with a technology and its designer.

In Chapter 5, I present how we can apply the Three Forces Model for Public Interest Tech to each case study to describe the current state and the ideal goal state.

In order to effectively respond to the multiple ways of how digital technologies have adversely impacted the public interest, we need a “whole-society” strategy that coordinates our laws, norms, and markets in how they interact with our technologies to prioritize the public interest. As public interest technologists, we need to work across disciplines to advance the public interest.

Let's get started.

Table of Contents

Abstract	iii
Acknowledgments.....	vi
Chapter 1. Introduction	1
Chapter 2. How Facebook’s Advertising Algorithms Can Discriminate By Race and Ethnicity.....	42
Chapter 3. How Were Social Security Numbers Assigned?.....	121
Chapter 4. Building A Collocation Detection System Using A Wi-Fi Sensor Array for COVID-19 Contact Tracing in A University Setting.....	168
Chapter 5. Conclusion	215
Appendix.....	246

Acknowledgements

I am incredibly grateful to my advisor, Latanya Sweeney, for her mentorship and support over the last 9 years from introducing these topics to me along with Jim Waldo in CS 105 back when I was an undergrad at Harvard, to bringing me to DC as a Research Fellow in Technology and Data Governance at the Federal Trade Commission, to finally supporting my research as my PhD advisor and working together as colleagues at the Data Privacy Lab at Harvard. I am also extremely grateful to Nicol Turner-Lee and Jim Waldo for being a part of my dissertation committee and providing me with expert guidance and advice.

I am incredibly fortunate to have the love and support of my parents, Mengwei Zang and Qinggong Ping, throughout my life and especially for their encouragement throughout my PhD. Coming back to Boston for my PhD also meant that I was able to spend the last several years happily hanging out with and taking care of my grandmother and grandfather, Yiling Han and Songtang Zang. This has been one of the happiest unforeseen consequences of my PhD. We got to take ferries out to the Boston Harbor Islands, visit a medieval castle in Gloucester, and check out an original house moved from Anhui province, where my family is from, to the Peabody Essex Museum. Over the last year, I got to quarantine with my grandmother and enjoy lots of delicious food. My grandfather passed last November from prostate cancer, and I know more than anything that he was incredibly proud of me and wanted to attend a second Harvard graduation ceremony.

Finally, I wanted to thank my amazing boyfriend, Steven Hong. You were by my side through my highs and lows. Last year, while you were busy treating COVID-19 patients at the hospital, you then came home and supported me through my research projects. I am incredibly lucky to have you by my side.

Chapter 1

Introduction

In 1998, Lawrence Lessig proposed his pathetic dot theory to describe the four constraints that shape the behavior of an individual [1]. They are law, norms, market, and architecture (Figure 1.1). Laws can force an individual to obey it or else face legal consequences ex post for disobedience, whether that's a fine or jail time. Norms and social pressure can also influence an individual to do what's expected of them as a member of society. Market forces shape which products an individual can purchase, what do they do to earn a living, and other decisions related to how an individual obtains and spends scarce resources. Finally, Lessig describes "architecture", the fourth constraint, as features of the world, either natural or man-made, that restrict or enable certain behaviors. For example, if there's a wall blocking an individual, they can't see through it. Nor can they easily access an abortion if the nearest abortion clinic is located very far away.

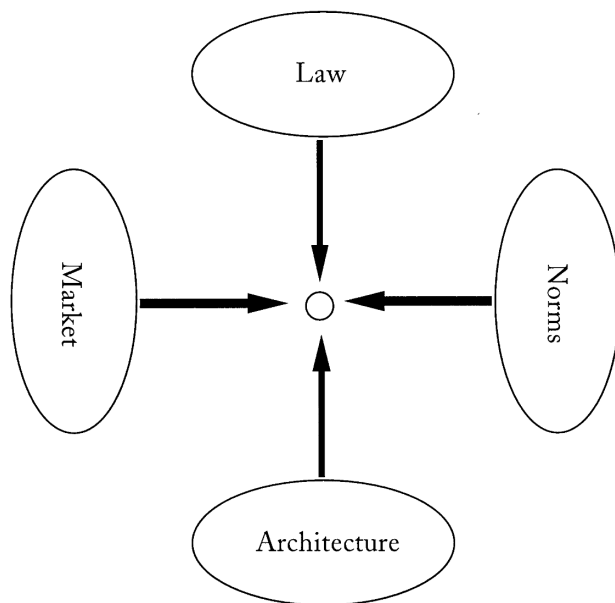


Figure 1.1. Four constraints that shape the behavior of an individual in Lessig's pathetic dot theory [1].

Using this framework, Lessig describes how digital technologies or “cyberspace” can constrain individuals in similar ways as architecture does in the real world [1]. Whether one is a user of a piece of technology or a data subject within a technological system, the design of the technology creates virtual walls or locked doors that can be just as constraining as any physical wall. If you don't have the right authorization as a developer or the ability to hack the code, then the computer program will dictate what functions you can use and what happens to your data.

Lessig expanded on this idea of cyberspace as architecture and how to regulate it in his *Code and Other Laws of Cyberspace* in 1999 [2]. At the time, the internet and other digital technologies were still developing and faced relatively light regulations. Lessig argued in favor of more regulations of the internet whether that's through the law or code in order to preserve important democratic values such as free speech or privacy.

Fast forward two decades, and we see in America today many of the negative repercussions of how these technologies, which still generally face light regulations, have developed to serve corporate interests and profits while harming the public interest. Issues contributing to the “techlash” or “tech backlash” include the spread of misinformation online distorting our politics, the loss of privacy in order to use vital digital tools, the emergence of new ways to discriminate through biased algorithms, and more [3]. Many Americans feel like the hypothetical individuals in Lessig's pathetic dot model, constantly constrained by digital architecture they have no control over [4]. Those who have the power are the designers and often corporations behind each technology, whether that's Facebook creating a content moderation system that doesn't stop the misinformation problem, Google monitoring the searches and emails of its users which sacrifices privacy to target ads, or Northpointe

creating COMPAS, a recidivism risk score algorithm that's potentially biased against African-American defendants. The power that these technology designers have can be just as influential on the lives of Americans as the nation's legislators, judges, and President, but they don't follow the same democratically accountable processes of elections, appointments, and confirmation hearings to gain that power.

One approach is to see these issues not as purely due to technology, but rather as a result of "sociotechnical" systems where the interaction between social and technical systems have complex "cause and effect" relationships [5, 6, 7, 8, 9, 10, 11]. The decisions made by an algorithm can reflect and reinforce the ways that society and people in society already behave. Cathy O'Neil, in her book *Weapons of Math Destruction*, notes that often these tools "punish the poor and oppressed in our society, while making the rich richer" [11]. Thus, using a sociotechnical approach to research algorithmic bias means understanding the complex social, cultural, and organizational contexts where an algorithm is deployed [9]. This includes asking questions about the identity, incentives, preferences, assumptions, biases, and other social attributes of the technology's designers, users, data subjects, customers, and other stakeholders [10, 12]. The sociotechnical approach also has many ties to the Science, Technology, and Society (STS) literature on treating the characteristics, power, and effect of technologies as based on the "network of relations within which a technology is positioned", according to Neyland [10]. Thus, an ethnography-based approach can uncover the principles and expectations of the different parties who are involved with or affected by the technology [10].

As the scale of tech's impact expands beyond the individual to affect large groups of people and society as a whole, the field of "*public interest technology*" has emerged over the last several years as a response by leaders in philanthropy, academia, and governments [13, 14, 15, 16]. According to

Eaves, Felten, McGuiness, Mulligan, and Weinstein, “Public interest technology refers to the study and application of technology expertise to advance the public interest/generate public benefits/promote the public good” [17]. Public interest technology is a cross-disciplinary field using approaches from computer science, data science, social sciences, public policy, and law with a normative goal of applying that expertise for “public interest or common good, as distinguished from the design of technology or technology policy to advance commercial or individual goals and interests” [17].

The normative goal of public interest tech to “advance the public interest” is important in light of how technology today is often developed without considering the public interest as an input. Unforeseen consequences often occur when the technology clashes with society and results in adversely impacting the public interest [18].

Three Forces Model of Public Interest Tech (Current State)

I propose how we can adapt Lessig’s pathetic dot model to understand the current state of relationships between technology, society, and the public interest, where the public interest is only an output rather than an input.

The underlying ambition of Lessig’s pathetic dot theory, which is based on earlier work written by many scholars at the University of Chicago that Lessig described as the “Old Chicago School” [1], is that the four constraints – law, norms, market, and architecture – “constitute a sum of forces that guide an individual to behave” [1]. As previously discussed, we can map digital technology as “architecture”, since it serves as the role of architecture for the digital world. Unfortunately, our current state is a dysfunctional one, where the public interest is often affected as an output of technology but not fully considered as an input, and we can see how that has motivated the techlash and the establishment of public interest tech as a response. The dyadic relationship between technology and the public interest doesn’t exist in a vacuum. But rather it is built by and operates

within the complicated nature of society itself. Thus, Figure 1.2 shows how, in our current state, the three forces of the law, norms, and market can affect a given technology or vice versa which in turn affects the public interest. There are dotted lines from the public interest to the law, norms, and market forces, which represents the mixed presence of the public interest as an input. It may be partially absent or not a significant priority that needs to be addressed. Even if one force considers the public interest as being very important, it may not be the most powerful force that is impacting the technology and technology designer.

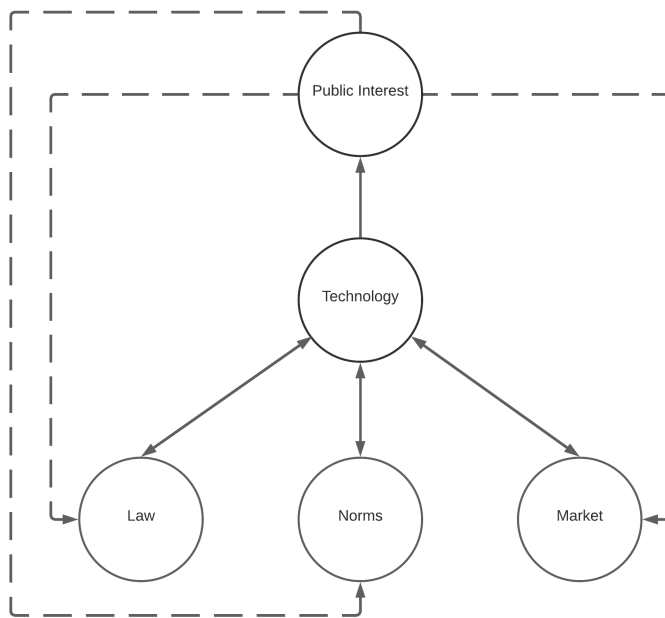


Figure 1.2. The Three Forces Model of Public Interest Tech (Current State). The three forces of the law, norms, and market affect a given technology or vice versa which in turn affects the public interest. There are dotted lines from the public interest to the law, norms, and market forces, which represents the mixed presence of the public interest as an input. It may be partially absent or not a significant priority that needs to be addressed.

In the Three Forces Model, I see the technology designer and the technology being influenced by and interacting with the law, norms, and market to shape our digital world in similar ways to how

the architect and architecture balance the forces of the law, norms, and market to shape our physical world. In architecture, we find a myriad of ways for each of the forces to consider the public interest as an input. For example, zoning laws reflect acceptable development uses by a community while building codes ensure the construction of safe buildings. Architecture schools, professional associations, and professional licenses ensure architects are aware of standardized knowledge on how to design safe and effective buildings that serve residential, commercial, industrial, or other needs. While the market forces, such as the private developer hiring the architect, may be self-interested in maximizing their profits, in the US, we see mandatory affordable housing quotas or Inclusionary Zoning in many urban areas such as Boston, New York, San Francisco, and more as way for public policy to ensure that the market will provide new housing for residents of all incomes [19]. Even though an architect may often be paid by a private developer as a client, they still have to balance between the three forces. An architect may need to push back against a client wanting to build a bigger building than zoning laws would allow. Similarly, an architect may need to push back against a client hoping to cut costs by using untested building materials that may go against their training and industry norms.

Of course, architecture itself can still fall short of serving the public interest, with a prominent example being a lack of affordable housing in many cities across the US that is only partially solved by Inclusionary Zoning [19]. But we see many established ways for the law, norms, and market to consider the public interest when influencing architects and shaping the architecture of the future. With digital technologies, we're only beginning to see similar robust developments in the law, norms, and market to consider the public interest as an important input for technology design. One potential reason is that while humankind has been constructing buildings since time immemorial, the Internet and related digital technologies are fairly new and have only become prominent over the last 30 years.

Koseff describes how in the early days of the Internet, the 1990s, there was a common belief in “Internet Exceptionalism”, which meant that the Internet or “cyberspace” is uniquely different from all previous media and that the rules of property, expression, identity, movement, and context that apply to the physical world, should not apply to the digital one [20]. As a result, we often find that in many of the areas where the greatest clash between current technologies and the public interest are found today, a given technology was not designed with the specific public interest fully incorporated as an input.

We can examine the issues of misinformation, data privacy, and algorithmic bias through the Three Forces Model as examples of how the law, norms, and market interact with a given technology to create an adverse impact on the public interest.

Section 230 of the Communications Decency Act of 1996 gives broad liability protection to digital platforms for publishing, removing, or restricting access to user content [20, 21]. This law means that social media sites such as Facebook, Twitter, or YouTube are not liable for hosting and spreading misinformation on their platforms since the content is posted by their users. In the most egregious case, as was seen in 2016, Russian-affiliated groups may spread misinformation on these platforms through bot accounts, advertisements, and user groups to influence a U.S. Presidential election. While their executives had to testify in multiple Congressional hearings, Section 230 meant that Facebook and Twitter were not liable for hosting Russian-affiliated misinformation on their sites [22]. The issue became even more complicated in 2020, since misinformation may not be entirely from foreign sources but from domestic groups and even average citizens as well. Thus, there’s also a dueling public interest in protecting the free speech rights of American users. While there was a more active effort by Facebook and Twitter to try to restrict misinformation on their platforms for the 2020 elections, misinformation was still widely disseminated, and the January 6 Capitol Insurrection,

motivated by misinformation regarding Donald Trump winning the 2020 election, was organized in part through Facebook Groups [23, 24]. Thus, we see how social media sites can adversely affect the public interest of limiting misinformation (Figure 1.3). While this harm may be justified given the competing public interest of free speech, the broad liability protections of Section 230 means that instead of having this debate through legal processes like the courts, it's happening internally within the management of private tech companies. However, since the status quo benefits the social media companies, there's also feedback between the technology and the law with tech companies lobbying Congress to maintain their liability protections if and when Section 230 is reformed [25].

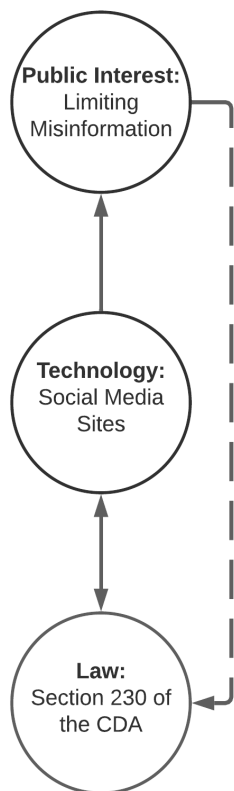


Figure 1.3. Three Forces Model (Current State) for Limiting Misinformation. Since Section 230 of the CDA limits the liability of social media sites for having user-posted misinformation on their platforms, though the law may try to exert some force on social media sites through Congressional Hearings, which is represented by the dotted line. There's also feedback between

the technology and the law to maintain the status quo even though the technology can have an adverse impact on the public interest of limiting misinformation.

For a different example where the role of norms is most relevant (Figure 1.4), we see how the growth of social media sites such as Facebook and Instagram over the last decade has normalized the sharing of personal photos online of our everyday lives. In fact, the social media sites even gamify the process by showing users how many “likes” each photo receives and promoting content from “influencers”, thus, creating a feedback loop between the technology and the norm. On the surface, the new normal of posting photos of ourselves eating, working, studying, hanging out, or in 2020, just living at home, seems relatively innocuous. However, it also facilitates facial recognition companies such as Clearview AI being able to collect 3 billion images from social networks like Facebook and Instagram [26]. Before the advent of these social media sites, it wouldn’t have been possible for Clearview AI to collect so many photos on so many Americans in one place. Thus, we see in the current state how the technology of social media sites adversely impacts the public interest of facial data privacy. Some users may react to learning about Clearview AI by being more concerned about their privacy and stop posting personal photos online or use more strict privacy settings. However, the social media sites themselves may be motivated in maintaining the norm since they profit from user content and engagement, and therefore may want to make privacy settings hard to use or reduce public attention on how others can scrape their sites [27].

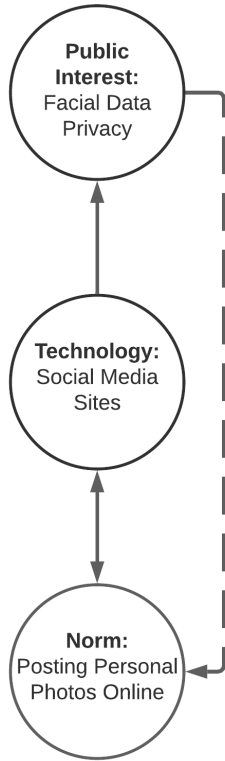


Figure 1.4. Three Forces Model (Current State) for Facial Data Privacy. There’s a feedback loop between social media sites normalizing posting personal photos online and more photos thus being posted on social media sites. However, the dramatic increase in the number of publicly accessible personal photos on these sites adversely impacts the public interest of facial data privacy by creating a rich data source for facial recognition companies and services. The dotted line from the public interest to the norm represents how some users may stop posting personal photos online or use more strict privacy settings due to privacy concerns.

The role of the market is often most relevant in cases of algorithmic bias. For example, in 2013, Sweeney found that Google’s ad platform was statistically significantly more likely to show the word “arrest” in ads for Black-identifying first names than in ads for White-identifying first names. The adverse impact ratio was 40% for Google.com search results and 77% for Reuters.com, which had ads served by Google [28]. One possible cause for this bias is active discriminatory behavior by the

advertiser, Instant Checkmate, using different ad text templates for Black-identifying vs. White-identifying names. Instant Checkmate denies this [28]. Another cause, which often occurs in cases of algorithmic bias, is a feedback loop between an existing bias in the market for background searches to favor checking for arrest records for Blacks vs. Whites being reinforced by Google's advertising algorithm showing ads that are most likely to get clicked. However, this example of algorithmic bias may adversely impact the public interest of anti-discrimination, especially since even individuals without arrest records may show up in ads implying they have one (Figure 1.5). That was the case for Sweeney herself. In this case, due to Sweeney's research, Google stopped showing the arrest language ads from Instant Checkmate, which is represented by the dashed line in Figure 1.5, but what about other markets where algorithmic bias may occur but haven't yet been studied? Other instances of algorithmic bias have been found in facial recognition systems [29], online shopping [30], search engines [31, 32, 33, 34], job sites and hiring software [35, 36, 37], translation services [38], healthcare [39], and other systems.

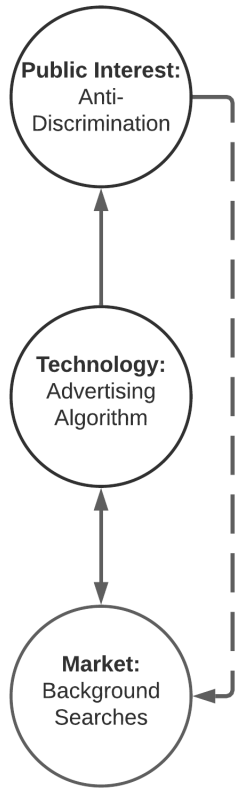


Figure 1.5. Three Forces Model (Current State) for Anti-Discrimination. The market for background searches may be biased in favor of checking for arrest records for Blacks vs. Whites. The advertising algorithm may learn that bias as more ads for Black-identifying names with “arrest” in the text are clicked vs. similarly worded ads with White-identifying names. Over time, this relationship between the ad algorithm and the market may become a feedback loop. However, this example of algorithmic bias may adversely impact the public interest of anti-discrimination, especially since even individuals without arrest records may show up in ads implying they have one. The dotted line from public interest to the market represents how Google eventually stopped showing the arrest language ads as a result of Sweeney’s study.

For different combinations of technologies and public interests, the amount of force exerted by the law, norms, or market could also differ and so could the degree of feedback between the

technology and each of the forces. For this dissertation, I present three different case studies of public interest tech research where there's a different combination of primary forces for each case study of a technology and a relevant public interest. In Chapter 2, I research how Facebook's advertising algorithms can promote racial discrimination. In Chapter 3, I test how the predictability of Social Security Number (SSN) assignment based on easily accessible data about Americans present a risk of identity theft. In Chapter 4, I demonstrate how TraceFi, a Wi-Fi based collocation detection technology can be deployed for COVID-19 contact tracing. In Chapter 5, I present how each research project fits within the Three Forces Model for Public Interest both in terms of the current state and the ideal goal state discussed later in this chapter.

The research projects in Chapters 2 and 3 are looking at issues with existing technologies, Facebook's advertising algorithms and the SSN assignment protocol, respectively. In Chapter 4, I set out to develop a new technology, effective Wi-Fi based collocation detection for contact tracing.

I briefly summarize below how we can apply the Three Force Model to understand the current state of each research project and the findings of each project.

Chapter 2 – How Facebook's Advertising Algorithms Can Discriminate By Race and Ethnicity

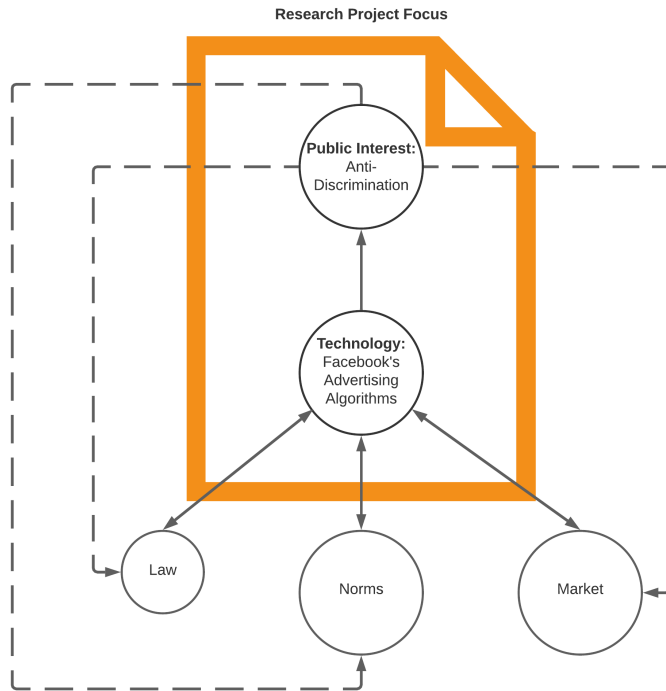


Figure 1.6. The Three Forces Model (Current State) for Chapter 2’s Research Project and Project Focus. The technology is Facebook’s advertising algorithms and the public interest is anti-discrimination. The law is likely exerting a weaker force on Facebook’s advertising algorithms than norms and the market. The research project focused on how Facebook’s advertising algorithms impacted anti-discrimination.

Current state:

- The technology is Facebook’s different advertising targeting options: its prepackaged “Detailed Targeting” options, its Lookalike Audiences, and its Special Ad Audiences tools.
- The public interest is anti-discrimination by race and ethnicity.
- The law likely exerts the weakest force despite Facebook having faced multiple lawsuits over discrimination and its advertising platform over the last 5 years, because Facebook hasn’t faced significant financial or criminal risks from the law over this issue and Facebook has argued that it’s ultimately immune from liability due to Section 230.

- Normative differences in how users of different race and ethnicity behave online is likely to significantly impact the discriminatory potential of Facebook’s advertising algorithms. At the same time, many minority users may not expect to see discriminatory targeting online even when legal, which likely contributed to support for the July 2020 boycott.
- Market forces are possibly also strong, since in July 2020, more than 1,000 major corporate advertisers joined a “Stop Hate for Profit” boycott organized by civil rights groups to stop advertising on Facebook for the month. At the same time, Facebook wants to maximize profits by serving the needs of advertisers as best as possible, which may reinforce discriminatory ad targeting.

The research project focused on the dyadic relationship between Facebook’s advertising algorithms and anti-discrimination. The project found that while Facebook’s retirement of multicultural affinity groups in August 2020 has removed one way to target minorities on the platform, its other targeting options, as well as Lookalike and Special Ad Audience tools, can still discriminate by race and ethnicity. While some discriminatory ad targeting may be legal or even desirable, this project demonstrates how there’s a lack of transparency on the discriminatory potential of Facebook’s ad platform which may help cover up the behavior of discriminatory advertisers and undermine the intent of non-discriminatory ones.

- In 2021, Facebook’s “African-American Culture” ad targeting option contained 75% fewer White users than the old “African American (US)” option they removed in the previous year.
- Facebook’s tools to help advertisers find similar users to their existing customers exhibited bias towards including more African-Americans or Whites depending on which racial group was dominant in an advertiser’s customer list, and this was true for the Lookalike Audience

tool, as well as the Special Ad Audience tool that Facebook designed to explicitly not use sensitive demographic attributes when finding similar users.

- Lookalike or Special Ad audiences based on customer lists with either stereotypically African-American or White names or ZIP codes would be even more biased towards including more users of that demographic group.
- Similarly, Lookalike audiences based on Asian customer lists can also become biased towards Asians, reaching up to 100% Asian in one case, and Lookalike audiences based on Hispanics over-represented Hispanics versus Non-Hispanics.

Chapter 3 – How Were Social Security Numbers Assigned?

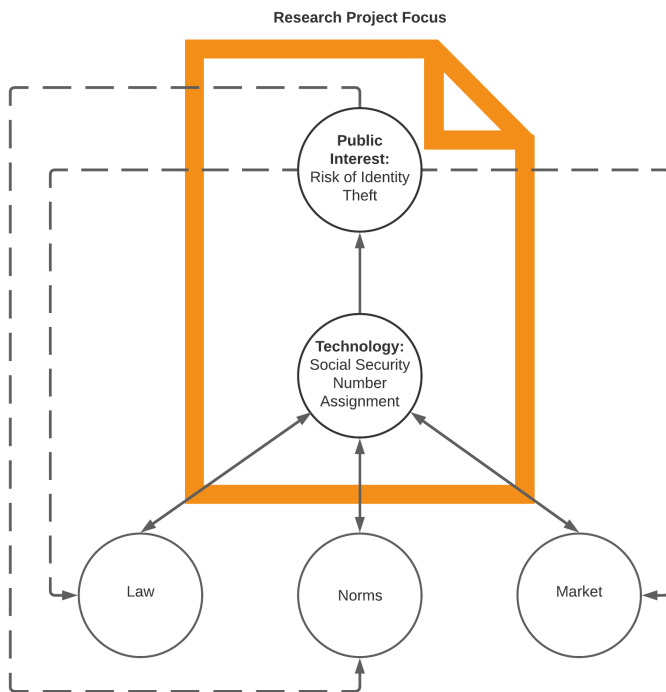


Figure 1.7. The Three Forces Model (Current State) for Chapter 3’s Research Project and Project Focus. The technology is Social Security Number assignment and the public interest is the risk of identity theft. The law, norms, and market are all powerful forces in this case. The research project focused on how Social Security Number assignment impacted the risk of identity theft.

Current state:

- The technology is Social Security Number (SSN) assignment.
- The public interest is the risk of identity theft.
- The law is a significant force since the Enumeration At Birth program by the Social Security Administration starting in 1989 gave SSNs to newborns at birth and SSA also didn't start randomizing SSNs until 2011.
- Since SSNs are used as both identifiers and authenticators in the US today, it's normal for Americans to be protective of their SSN while openly sharing relevant information to predict their SSN such as their date of birth and state of birth.
- Market forces helped proliferate the usage of SSNs as identifiers and authenticators, which created attractive opportunities for identity thieves to exploit. In addition, identity thieves can also turn to dark web markets selling SSNs often found in data breaches.

The research project focused on the dyadic relationship between Social Security Number assignment and the risk of identity theft. The project found strong evidence that SSNs were assigned in a nested loop protocol based on sets of Area Numbers, Group Numbers, Area Numbers, and then Serial Numbers in all 50 states and DC between 1989 and 2011. This means that Americans born between 1989 – 2011 face an additional SSN-based identity theft vulnerability due to how SSA assigned their SSNs at birth.

- I build upon earlier research to propose my own hypothesis about SSN assignment as following a nested loop protocol
- For Americans born between 1989 and 2011, they have SSNs most vulnerable to prediction based on their state of birth and date of birth, due to the Social Security Administration's Enumeration At Birth program

- For SSNs in the Death Master File, I am able to accurately predict the first 5 digits 48% of the time and the first 6 digits 11% of the time
- States with smaller populations were the most vulnerable: I am able to accurately predict the first 5 digits of the SSN in 19 states including DC more than 80% of the time, and for 5 states – Delaware, Idaho, North Dakota, South Dakota, and Wyoming – more than 90% of the time
- It’s time for public policy to focus on solutions that can replace SSNs with alternatives that are designed to be strong authenticators from the start

Chapter 4 – Building A Collocation Detection System Using A Wi-Fi Sensor Array for COVID-19 Contact

Tracing in A University Setting

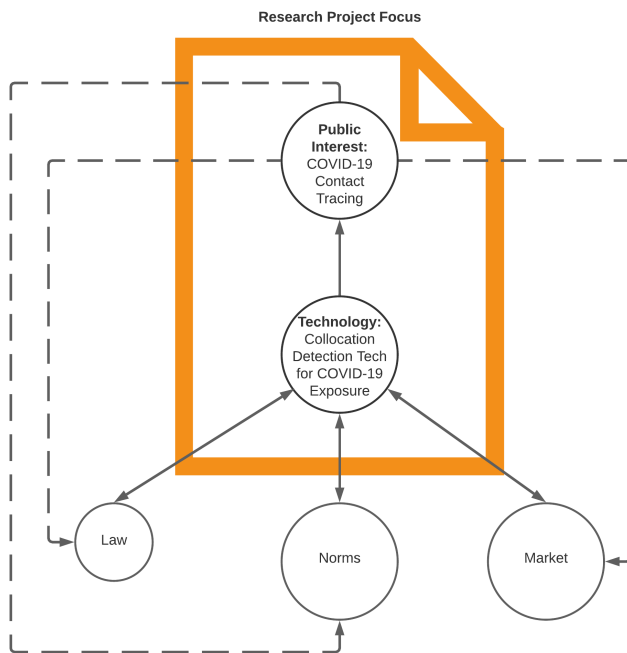


Figure 1.8. The Three Forces Model (Current State) for Chapter 4’s Research Project and Project Focus. The technology is collocation detection tech for COVID-19 exposure and the public interest is COVID-19 contact tracing. The law is the weakest force primarily due to the lack of coordination by the US federal government on contact tracing in 2020. The research project

focused on how collocation detection tech for COVID-19 exposure impacted COVID-19 contact tracing.

Current state:

- The technology is collocation detection tech for COVID-19 exposure.
- The public interest is COVID-19 contact tracing.
- The law is the weakest force primarily due to the lack of coordination by the US federal government on contact tracing in 2020.
- Effective contact tracing technology would need to be easy for users to adopt but also address their privacy concerns.
- Google and Apple were able to leverage their duopoly in the market on mobile operating systems and the corresponding app stores to restrict the types of contact tracing apps allowed on their devices.

The research project focused on the dyadic relationship between collocation detection tech for COVID-19 exposure and COVID-19 contact tracing. The project built and tested TraceFi, a Wi-Fi based collocation detection system that uses a sensor array to accurately detect mobile devices within 6 feet of each other, for possible use in contact tracing without the burden of requiring a user to install an app in order to participate.

- TraceFi is a Wi-Fi based collocation detection system that uses a sensor array to accurately detect mobile devices within 6 feet of each other for possible use in contact tracing without the burden of requiring a user to install an app in order to participate
- We tested multiple machine learning models in a TraceFi pilot across 12 different spaces in 3 different buildings under regular use conditions and found XGBoost models had a peak

sensitivity of 91% and a peak specificity of 86%, with a high median sensitivity of 77% and a high median specificity of 81%

- TraceFi can be used for accurate real-world collocation detection for contact tracing to determine whether 2 devices were within 6 feet for 15 minutes or more and is the first Wi-Fi technology to do so
- We engaged with stakeholders around the university to incorporate their concerns around the ease of adoption, accuracy, cost, and privacy into how we propose including TraceFi data into the contact tracing data flow
- We designed a system for using TraceFi data for contact tracing according to Fair Information Practices that seek to preserve the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive

How can the law, norms, and market apply the public interest on technology?

Since the normative goal of public interest tech is to ultimately “advance the public interest” [17], how can the law, norms, and market apply the public interest on technology?

As discussed above, the technology designer, just like the architect before them, is constrained by the forces of the law, norms, and market. But unlike the architect, the potentially naïve optimism of Internet Exceptionalism from the 1990s meant that the technology designer doesn’t necessarily have to consider the public interest as defined by historical laws, norms, and markets as an input to their design to the same degree [20]. We see the repercussions of that dynamic in the issues that instigated the techlash and the case studies of this dissertation. The goal of public interest tech as a field is to find ways to fix this dynamic by re-incorporating the public interest as an input to the laws, norms, and market forces that affect a technology and its designer. Each of these forces have

a long history of considering what is the public interest, and in recent years, we've seen examples of how these interpretations of the public interest have been applied to the issues relevant to public interest tech.

Public interest and the law

We can see how three common interpretations of public interest in law relate to public interest tech issues.

First, public interest in law can be about representing the interests of the “little guy”, the historically disadvantaged, or the diffuse interest of the masses against the concentrated interests of those with power and wealth [40]. In 1905, Supreme Court Justice Louis Brandeis in an address to the Harvard Ethical Society stated, “Instead of holding a position of independence between the wealthy and the people, prepared to curb the excesses of either, able lawyers have to a large extent allowed themselves to become adjuncts of great corporations and have neglected their obligation to use their powers for the protection of the people... The great opportunity of the American bar is, and will be, to stand again as it did in the past, ready to protect also the interests of the people” [41]. We see Brandeis' call for a nobler obligation for lawyers than simply representing corporate interests in the rise of Legal Aid groups to represent the poor [42] and class action lawsuits to represent the many [43]. For a tech-related example, in 2009, Netflix was sued in a class action lawsuit on behalf of 480,000 customers whose movie watching data was voluntarily disclosed by Netflix for a competition to create a better movie recommendation algorithm [44]. The plaintiffs argued that this violated the Video Protection Privacy Act. Netflix settled with the plaintiffs and cancelled the second iteration of its competition in 2010 [45].

Second, public interest in law can be about representing some substantive interest that is guaranteed by the Constitution or important to the public [40, 46]. Civil rights groups such as the

ACLU and the NAACP Leadership Defense Fund have pursued a strategy of filing successive lawsuits to reduce the spaces and circumstances for racial discrimination [46]. Environmental law groups have pursued similar strategies to advance environmental protection. Conservative groups that support pro-life or pro-gun positions have also argued that they are protecting the public interest in reducing abortion or gun control limits [47]. Beyond the courts, many advocacy groups work to enshrine their substantive interpretation of the public interest in legislation such as with the Civil Rights Act, the Clean Air Act, or “stand-your-ground” laws. For tech-related examples, in Chapter 2, we see how Facebook has been sued by the National Fair Housing Alliance [48], the ACLU [49], the Communications Workers of America [50], and others [51] over issues of discrimination on its advertising platform violating civil rights laws such as the Fair Housing Act and the Civil Rights Act of 1964.

Third, public interest in law can be the decisions made by the elected officials, regulators, and judges acting as agents of the government [40]. For example, in Barron’s Law Dictionary, “public interest” is defined as “a subjective determination by an individual such as a judge or governor, or a group such as a township committee or state legislature of what is for the general good of all people” [52]. Some statutes may explicitly reference the public interest in how public officials should make decisions. For example, the Freedom of Information Act (FOIA) requires officials to waive or reduce fees for duplicating documents if disclosure “is in the public interest because it is likely to contribute significantly to public understanding of the operations or activities of the government and is not primarily in the commercial interest of the requester” [53]. Even if there is not explicit statutory guidance, public officials should be acting based on the public interest in the enforcement of laws and regulations. For example, in August 2012, the Federal Trade Commission (FTC) approved Facebook’s acquisition of Instagram with a 5-0 vote. In its decision, the FTC wrote, “The commission reserves the

right to take such further action as the public interest may require” [54]. In December 2020, the FTC filed an antitrust lawsuit against Facebook seeking to have Facebook’s acquisition of WhatsApp and Instagram reversed [55]. Under this third interpretation of the public interest in law, it’s possible that the FTC was acting on behalf of the public interest both in 2012 when it approved the Instagram acquisition and in 2020 when it sought to reverse it. Of course, it’s also important to consider that political science researchers have found that government bureaucrats can also be self-seeking and prioritize their job security, status, and power rather than the public interest [56].

Public interest and norms

In social contract theory, political philosophers argue that the social contract is the source of political obligations and social norms, including principles of justice [57]. Different political philosophers have then argued for different ways for individuals in society to agree on a shared theory of justice. I present a brief overview of three broad camps of theories of justice, examples of how they apply to issues related to public interest tech, and possible limits of each theory. In these examples, I argue that the public interest is related to what the relevant parties consider as justice.

The first camp, exemplified by Hobbes and Hume, argues for justice as mutual advantage [57, 58]. Individuals agree to compromise for the social contract if it allows them to pursue their separate aims more harmoniously and successfully [57]. According to Barry, justice is “the constraint on themselves that rational self-interested people would agree to as the minimum price that has to be paid in order to obtain the cooperation of others” [58]. Thus, the gains made possible by cooperation have to be enough to justify constraining self-interested actions [59]. As an example, in 2014 and 2015, students and professors at Harvard found in separate research studies disparate impact effects for Asian and African-American users on Airbnb relative to White users [60, 61]. In May 2016, #AirbnbWhileBlack became a trending hashtag on Twitter as users documented the discriminatory

behavior they experienced on the platform [62]. In response, Airbnb hired former Attorney General Eric Holder in July 2016 to improve their anti-discrimination policy [63], required hosts to accept a guest's booking before requesting their photo in October 2018 [64], and launched a study of the racial experience gap on the platform in November 2020, which I helped advise [65]. In Airbnb's case, it's mutually advantageous for both the company and its users to conceive of anti-discrimination as justice, since reducing discrimination would result in more successful bookings for minority users which would also help Airbnb's bottom line. Thus, the public interest of anti-discrimination can be a norm for the platform. However, critics would point out that under justice as mutual advantage, rational individuals would never act against their own self-interest or adhere to agreements that do so [57]. For example, in the case of Sweeney's finding of racial discrimination in online ads with "arrest" in the text [28], it was mutually advantageous for both the advertiser, Instant Checkmate, and the ad platform, Google, to pursue a discriminatory ad delivery strategy rather than a non-discriminatory one, until Sweeney's study drew negative media attention.

The second camp, exemplified by Kant and Rousseau, argues for justice as impartiality [57, 58]. Individuals agree to principles of justice that are mutually justifiable to rational and reasonable people that do not reflect their self-interests [58]. Thus, impartial agents would agree to promote the common good [57]. Relevant tech-related examples for this camp include the many ethical technology design and ethics principles created by different impartial experts or organizations such as the Fair Information Practices created by the OECD in 1980 [66], the Privacy by Design principles created by Ann Cavoukian in the 1990s initially for the Canadian government [67], or the Association for Computing Machinery (ACM) Code of Ethics and Professional Conduct created by the ACM Code 2018 Task Force [68]. If every technology designer in society agrees to follow these principles, then they can serve as a way of asserting the public interest as norms. However, critics like Rawls point out

that there often isn't an unconditional commitment to the common good [57]. If there are no penalties to ignoring these principles, or if it's possible to only pay lip service without making substantive changes, then some technology designers may choose to prioritize their private interests instead.

The third camp, exemplified by Rawls, argues for justice as reciprocity [57]. According to Rawls, "The idea of reciprocity lies between the idea of impartiality...and the idea of mutual advantage" [69]. Reciprocity is the willingness to do one's part to cooperate in society provided that others also do theirs [57]. This may involve acting against one's self-interest but only if others are willing to do the same. Rawls argues that individuals desire to live in a society in which their actions and those of others can be judged as fair and just [69, 70]. We saw how quickly justice as reciprocity played out in the suspensions of President Trump's online accounts across different services after the January 6, 2021 Capitol Insurrection. Within a week of January 6, Twitter, Facebook, YouTube, Snapchat, even Shopify and other companies either temporarily or permanently suspended President Trump's accounts with them [71]. Each tech company individually risked the wrath of users who supported Trump and may see a suspension as censorship, but as the suspension movement gained steam, it became progressively easier for new companies to justify their actions as simply following industry norms. For Twitter, which permanently suspended @realDonaldTrump on January 8, they cited how his tweets violated their public interest framework for world leaders which prohibits the "Glorification of Violence" [72]. Facebook referred its indefinite suspension of President Trump's account to its Oversight Board of 40 experts in law, ethics, human rights, and tech policy [73]. This example also highlights the incredible power that tech companies have over the political arena, such as speech by the President of the United States, and even when they try to act in favor of the public

interest to establish the norms of acceptable speech on their platforms, it still raises fundamental questions about whether that power should be in the hands of tech executives in the first place.

Public interest and the market

Neoclassical economics considers efficient and competitive markets as the “invisible hand” described by Adam Smith that aggregates individual decisions motivated by one’s preferences into socially optimal outcomes [74, 75]. However, not all markets are economically efficient. Market failure can occur when there are externalities either negative or positive that are not being priced in or if there’s a lack of competition due to monopolies or oligopolies. Economics can provide us with a lens to understand how the public interest can be to address these market failures, which also exist in the tech industry today.

In Mankiw’s economics textbook, an externality is defined as “when a person engages in an activity that influences the well-being of a bystander and yet neither pays nor receives any compensation for that effect” [74]. On a societal level, negative externalities can decrease social welfare such as pollution, while positive externalities can increase it such as clean air. Because the cost of the good generating the externality doesn’t include the social cost or benefit, society ends up with too many goods that generate negative externalities, while not enough of the ones that generate positive externalities [74]. Pigou in the 1930s proposes how levying taxes on negative externalities and subsidies on positive externalities can internalize the social cost or benefit of the good generating the externality for the firm or consumer in order to maximize social welfare [76, 77]. Through the lens of externalities, we can see how Section 230’s liability protection for online platforms removes the cost of legal damages for hosting harmful content such as misinformation, which is a negative externality for society, from influencing the behaviors of these firms [78, 79]. On the other hand, there are positive externalities from universal broadband access, including for rural and low income households, as

being proposed in President Biden's American Jobs Plan released on March 31, 2021 [80]. Benefits include greater employment opportunities, especially for remote work, access to tele-health or online learning, and even more efficient agriculture [81]. Without subsidies, these benefits may not outweigh the significant cost for internet service providers to install broadband in rural areas or the low revenues for servicing low-income households, which argues for it being the public interest to subsidize universal broadband access for \$100 billion in President Biden's proposal [80].

Unregulated monopolies or oligopolies can result in another form of market failure as a lack of competition benefits the incumbent firms. Before the 1960s, economic structuralism was the dominant antitrust perspective, which saw monopolies and oligopolies as harmful to the public interest by creating market structures that (1) enable collusion, price-fixing, or market division, (2) block new entrants, or (3) harm consumer, supplier, or worker interests [82]. By the 1970s, the Chicago School of viewing antitrust through the effect of monopolies on increasing prices gained in popularity [83]. This was the basis for Robert Bork's argument that antitrust policy should seek to maximize the "consumer welfare", which the courts have interpreted as being measured through prices [82]. The shift from economic structuralism to the consumer welfare standard for antitrust enforcement creates a problem when tech companies may be able to offer low prices or even "free" products through monopolizing the market and keeping out competitors. Khan argues that since investors rewarded Amazon for pursuing growth above profits, it was able to engage in predatory pricing to dominate e-commerce [82]. With Facebook, we're starting to see a shift back to economic structuralism with antitrust enforcement, since even though its products such as its social network platform is ostensibly "free" to consumers, the FTC still sued Facebook for antitrust violations in December 2020 for acquiring competitors and threatening to turn its technologies against them [84]. For example, in 2008, CEO Mark Zuckerberg's wrote in an email, "It is better to buy than compete" [55].

Three Forces Model of Public Interest Tech (Goal State)

Since the normative goal of public interest tech is to ultimately “advance the public interest” [17], the goal state of the Three Forces Model shown in Figure 1.9 demonstrates how the public interest can be an input for the law, norms, and market in how they affect a technology’s design and usage, which would in turn affect the public interest. At the goal state, the public interest is not some arbitrary, ambiguous concept based on the whims of the technology designer, instead it’s an input in how the law, norms, and market create the rules and incentives that the designer has to follow.

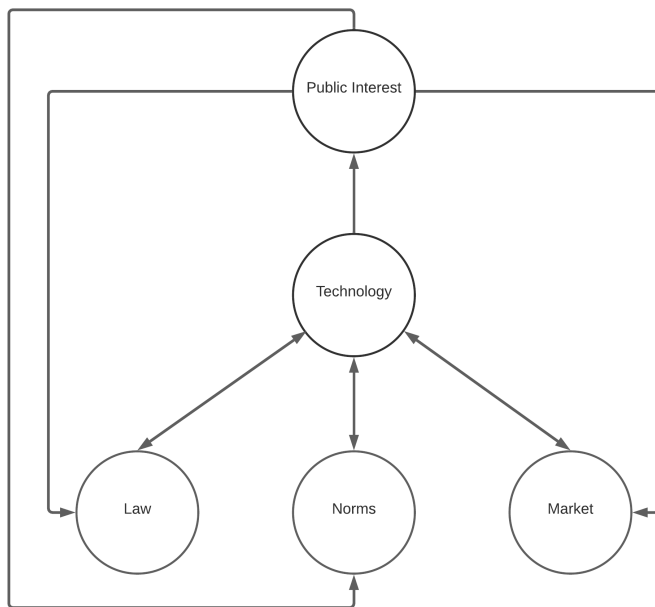


Figure 1.9. The Three Forces Model of Public Interest Tech (Goal State). Public interest is an input for the law, norms, and market. The three forces then affect a given technology or vice versa which in turn affects the public interest.

Stakeholders relevant to each of the forces can consider the public interest as a priority in how they interact with a technology and its designer. In order to uphold the pressure of public interest tech on technology designers who may otherwise prioritize their own self-interests, you need stakeholders enforcing it through the law, advocating for it through norms, or incentivizing it through the market.

For example, the stakeholders for the law may be elected officials, regulators, judges, class action attorneys, legal advocacy groups, and others. Elected officials or regulators may pass a new law or regulation influenced by the public interest. The courts may also see new litigation related to a public interest tech issue and rule in favor of protecting the public interest.

The stakeholders for norms include social activists, advocacy groups, subject matter experts, the media, and others. Activists and advocacy groups may organize public support for protecting the public interest in a given technology. They may also partner with subject matter experts to describe best practices or technology design principles that serve the public interest. The media can garner public attention and pressure.

The stakeholders for the market include competitors, employees, shareholders, customers, and others. Competitors may build competing products that promotes the public interest. Employees, shareholders, and customers can all pressure a company designing a technology to better serve the public interest or else face losses of labor, capital, or revenue. If there are issues of market failure either due to externalities or monopolies, then fixing those larger market failure issues can create an efficient and competitive market environment to serve the public interest.

In the goal state, for different combinations of technologies and public interests, the amount of force exerted by the law, norms, or market – with the public interest as an input – on the technology and its designer could also differ. In Chapter 5, I present how each of the research projects in chapter 2-4 can inform how the law, norms, and market can advance the public interest in the relevant technologies. I briefly summarize this discussion below.

Chapter 2 – How Facebook’s Advertising Algorithms Can Discriminate By Race and Ethnicity

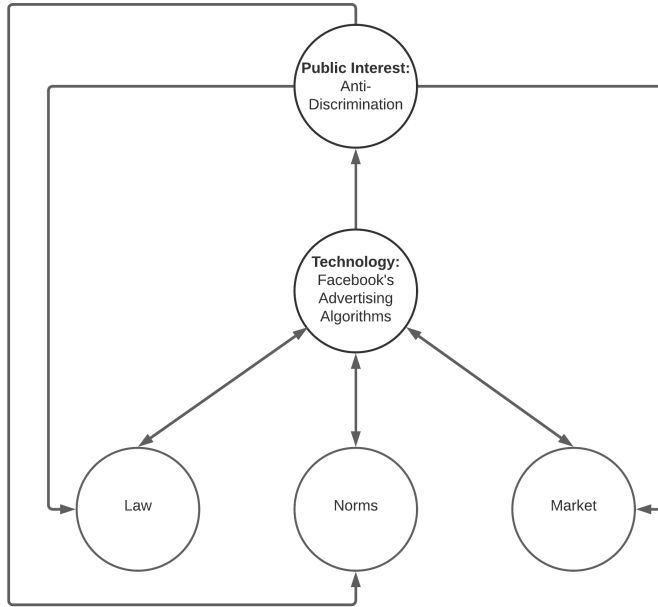


Figure 1.10. The Three Forces Model (Goal State) for Chapter 2’s Research Project. The public interest of anti-discrimination can be an input to how the law, norms, and market affect Facebook’s advertising algorithms.

Goal state:

- **Law:** Existing laws on civil rights can be enforced to require greater transparency of advertisers’ targeting strategies on Facebook, and new legislation may avoid implementing a “fairness through unawareness” standard as public policy for discrimination liability protection. Section 230 reform may also be considered to clarify Facebook’s liability for discriminatory advertising.
- **Norms:** Future civil rights audits at Facebook can study the potential for discrimination on its platform and establish a new trend of inviting public scrutiny similar to the annual publication of workforce diversity reports at many large tech companies.

- **Market:** Advertisers and users of Facebook and its competitors can pressure them to improve the disclosure of the demographics of audiences targeted by ads and other digital experiences.

Chapter 3 – How Were Social Security Numbers Assigned?

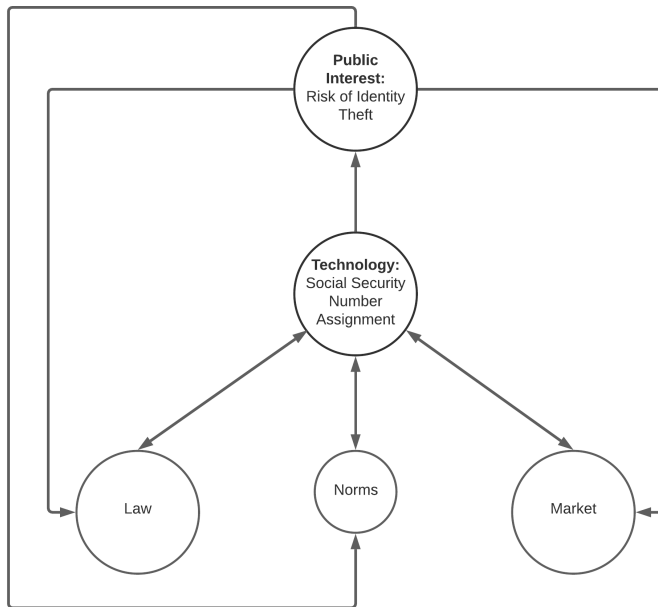


Figure 1.11. The Three Forces Model (Goal State) for Chapter 3’s Research Project. The public interest of reducing the risk of identity theft can be an input to how the law, norms, and market affect the role of Social Security Numbers in society, with a change in norms likely being the weakest force since many Americans are already protective of giving out their SSNs.

Goal state:

- **Law:** US public policy can invest in creating better technology design solutions to replace Social Security Numbers as de facto national identifiers and authenticators.
- **Norms:** Changing the current common practice, endorsed by the IRS, of obscuring the first 5 digits of an SSN on a form or paycheck while revealing the last 4 digits, may have a moderate benefit on preventing the revealing of SSNs. But ultimately changing norms is likely the

weakest force since many Americans are already protective of giving out their SSNs, but the research project demonstrates how there's underlying correlations in their SSN and their date and state of birth, which individuals can't fix on their own.

- **Market:** Marketplace alternatives to SSNs as authenticators have emerged in the private and public sectors and can replace SSNs with competitive alternatives that avoid creating a single point of failure.

Chapter 4 – Building A Collocation Detection System Using A Wi-Fi Sensor Array for COVID-19 Contact Tracing in A University Setting

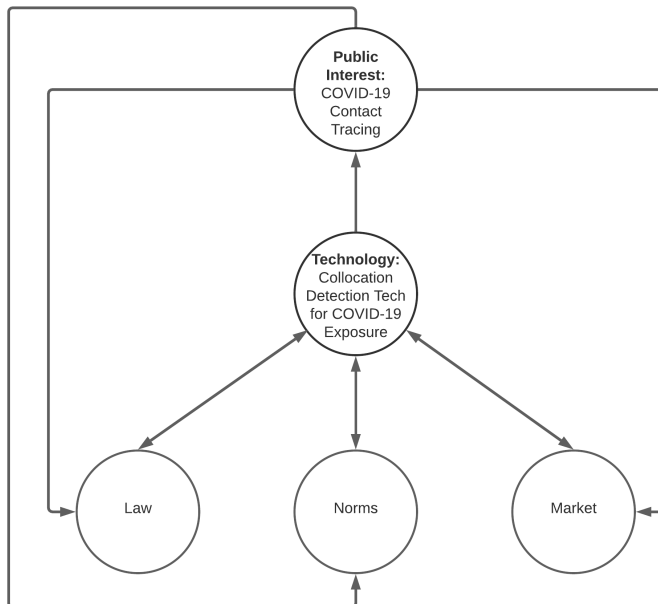


Figure 1.12. The Three Forces Model (Goal State) for Chapter 4's Research Project. The public interest was an input from the outset to how we considered the law, norms, and market forces in the design of the data flow system for TraceFi's collocation predictions to support the work of contact tracers at Harvard University Health Services while protecting the privacy of individual device owners.

Goal state:

- **Law:** Ideally, the law should be updated to clarify how HIPAA protections should apply to digital contact tracing technologies. In the case of TraceFi, we propose a privacy wall between the TraceFi System maintained by the Data Privacy Lab and Harvard University Health Services (HUHS) and Harvard University Information Technology (HUIT).
- **Norms:** We designed the data flow system to follow best practices described in the Fair Information Practices and GDPR principles. We also proposed making TraceFi opt out in order to ensure ease of adoption. Finally, we sought to approach the ideal privacy model for contact tracing of preserving the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive.
- **Market:** We sought out stakeholder input throughout the pilot study with 13 digital town halls and regular discussions with partners throughout the university. We also compared TraceFi to GAEN apps based on stakeholder concerns of ease of adoption, accuracy, cost, and privacy.

Building on the Three Forces Model of Public Interest Tech

In this dissertation, I present the Three Forces Model of Public Interest Tech and how it can be applied to the current and goal states of three case studies of different technologies and public interests that serve as the focus of the research projects. The Three Forces Model can illustrate why the current state often finds the public interest adversely impacted by technology when it is not a significant input but rather primarily an output of the sociotechnical process. The Three Forces Model can also illustrate how to achieve the normative goal of the field by having the public interest become an input to the law, norms, and market through actions taken by the relevant stakeholders.

Future work in public interest tech can examine how to address multiple public interests at once, especially if they appear to be countervailing public interests such as the free speech rights of the President of the United States versus the goal of reducing misinformation. Another major issue to

be explored is how to integrate the public interests of different nations and societies for technologies that span across countries.

The disruptions of digital technology on the public interest are occurring in multiple ways from the spread of misinformation, the loss of privacy, the rise of algorithmic discrimination, the threats of election interference, and more. In order to effectively respond, we need a “whole-society” strategy that coordinates our laws, norms, and markets to prioritize the public interest in how they impact our technologies. As public interest technologists, we need to work across disciplines to “advance the public interest” [17].

For Latanya Sweeney, she goes one step further in telling her students that she wants them to “save the world”.

Let’s get started.

References

1. Lessig L. The New Chicago School. *The Journal of Legal Studies*. Vol 27. No S2. 661–691. 1998. <https://doi.org/10.1086/468039>.
2. Lessig L. *Code and Other Laws of Cyberspace*. Basic Books, Inc. USA. 1999.
3. Atkinson R D et al. A Policymaker’s Guide to the “Techlash”—What It Is and Why It’s a Threat to Growth and Progress. Information Technology & Innovation Foundation. 2019. <https://itif.org/publications/2019/10/28/policymakers-guide-techlash>.
4. Knight Foundation. *Techlash? America’s Growing Concern With Major Technology Companies*. Knight Foundation. 2020. <https://knightfoundation.org/reports/techlash-americas-growing-concern-with-major-technology-companies/>.
5. Bostrom R P and Heinen J S. MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*. Vol 1. No 3. 17–32. February 3, 1977. <https://doi.org/10.2307/248710>.
6. Trist E. *The Evolution of Socio-Technical Systems: A Conceptual Framework and an Action Research Program*. 1981.
7. Baxter G and Sommerville I. Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*. Vol 23. No 1. 4–17. January 1, 2011. <https://doi.org/10.1016/j.intcom.2010.07.003>.
8. Selbst A D, Boyd D, Friedler S A, Venkatasubramanian S, and Vertesi J. Fairness and Abstraction in Sociotechnical Systems. in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019. 59–68. <https://doi.org/10.1145/3287560.3287598>.
9. Rosenbaum H and Fichman P. Algorithmic accountability and digital justice: A critical assessment of technical and sociotechnical approaches. *Proceedings of the Association for Information Science and Technology*. Vol 56. No 1. 237–244. January 1, 2019. <https://doi.org/https://doi.org/10.1002/pr2.19>.
10. Neyland D. *Accountability and the Algorithm BT - The Everyday Life of an Algorithm*. Neyland D, Editor. Springer International Publishing. Cham. 2019. 45–71.
11. O’Neil C. *Weapons of Math Destruction*. Crown Random House. 2016.
12. Musiani F. Governance by algorithms. *Internet Policy Review*. 2013. <https://doi.org/10.14763/2013.3.188>.
13. Schneier B. *Public-Interest Technology Resources*. Public Interest Tech. 2020. <https://public-interest-tech.com/>. Accessed February 3, 2021.

14. Freedman Consulting. A Pivotal Moment: Developing A New Generation of Technologists for the Public Interest. 216AD.
15. Ford Foundation. 5 reasons you might be a Public Interest Technologist. Equals Change Blog. 2018. <https://www.fordfoundation.org/just-matters/equals-change-blog/posts/5-reasons-you-might-be-a-public-interest-technologist/>.
16. New America Foundation. Public Interest Technology. New America Foundation. 2020. <https://www.newamerica.org/pit/about/>.
17. Eaves D, Felten E, McGuinness T, Mulligan D K, and Weinstein J. Defining Public Interest Technology. New America Foundation. 2020. <https://www.newamerica.org/pit/blog/defining-public-interest-technology/>.
18. Sweeney L. In Technology We Trust. .
19. Ramakrishnan K, Treskon M, and Greene S. Inclusionary Zoning: What Does the Research Tell Us about the Effectiveness of Local Action? 2019.
20. Kosseff J. The Twenty-Six Words That Created the Internet. Cornell University Press. 2019.
21. Congressional Research Service. Social Media: Misinformation and ContentModerationIssues for Congress. 2021.
22. Parlapiano A and Lee J. The Propaganda Tools Used by Russians to Influence the 2016 Election - The New York Times. The New York Times. February 16, 2018. <https://www.nytimes.com/interactive/2018/02/16/us/politics/russia-propaganda-election-2016.html>.
23. Horwitz J. Facebook Knew Calls for Violence Plagued “Groups,” Now Plans Overhaul. The Wall Street Journal. January 31, 2021. <https://www.wsj.com/articles/facebook-knew-calls-for-violence-plagued-groups-now-plans-overhaul-11612131374>.
24. Mack D, Mac R, and Bensinger K. “If They Won’t Hear Us, They Will Fear Us”: How The Capitol Assault Was Planned On Facebook. BuzzFeed News. January 19, 2021. <https://www.buzzfeednews.com/article/davidmack/how-us-capitol-insurrection-organized-facebook>.
25. McCabe D. Tech Companies Shift Their Posture on a Legal Shield, Wary of Being Left Behind. The New York Times. December 15, 2020. <https://www.nytimes.com/2020/12/15/technology/tech-section-230-congress.html>.
26. Heilweil R. The world’s scariest facial recognition company, explained. Vox. May 8, 2020. <https://www.vox.com/recode/2020/2/11/21131991/clearview-ai-facial-recognition-database-law-enforcement>.

27. Collins K. Facebook hopes to “normalize” idea of data scraping leaks, says leaked internal memo. CNET2. April 20, 21AD. <https://www.cnet.com/news/facebook-hopes-to-normalize-idea-of-data-scraping-leaks-says-leaked-internal-memo/>.
28. Sweeney L. Discrimination in online Ad delivery. *Communications of the ACM*. Vol 56. No 5. 44–54. 2013. <https://doi.org/10.1145/2447976.2447990>.
29. Buolamwini J and Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. Vol 81. 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>.
30. Hannak A, Soeller G, Lazer D, Mislove A, and Wilson C. Measuring Price Discrimination and Steering on E-Commerce Web Sites. in *Proceedings of the 2014 Conference on Internet Measurement Conference*. 2014. 305–318. <https://doi.org/10.1145/2663716.2663744>.
31. Kay M, Matuszek C, and Munson S A. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015. 3819–3828. <https://doi.org/10.1145/2702123.2702520>.
32. Kulshrestha J et al. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017. 417–432. <https://doi.org/10.1145/2998181.2998321>.
33. Robertson R E, Jiang S, Joseph K, Friedland L, Lazer D, and Wilson C. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact.* Vol 2. No CSCW. November 2018. <https://doi.org/10.1145/3274417>.
34. Datta A, Datta A, Makagon J, Mulligan D K, and Tschantz M C. Discrimination in Online Advertising: A Multidisciplinary Inquiry. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. Vol 81. 20–34. <http://proceedings.mlr.press/v81/datta18a.html>.
35. Chen L, Ma R, Hannák A, and Wilson C. Investigating the Impact of Gender on Rank in Resume Search Engines. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018. 1–14. <https://doi.org/10.1145/3173574.3174225>.
36. Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, and Wilson C. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017. 1914–1933. <https://doi.org/10.1145/2998181.2998327>.
37. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuter*. October 10, 2018. <https://www.reuters.com/article/us-amazon-com-jobs-automation->

insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

38. Bolukbasi T, Chang K-W, Zou J, Saligrama V, and Kalai A T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. in NIPS. 2016. NIPS. <https://www.microsoft.com/en-us/research/publication/quantifying-reducing-stereotypes-word-embeddings/>.
39. Obermeyer Z, Powers B, Vogeli C, and Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Vol 366. No 6464. 447 LP – 453. October 25, 2019. <https://doi.org/10.1126/science.aax2342>.
40. Edwin Rekosh. Who defines the public interest? *SUR*. 2004. <https://sur.conectas.org/en/defines-public-interest/>.
41. Brandeis L. The Opportunity in the Law. *American Law Review*. Vol 39. 55–63. 1905.
42. Cummings S L. The Internationalization of Public Interest Law . *Duke Law Journal* . Vol 57. No 4. 891–1036. <https://heinonline.org/HOL/P?h=hein.journals/duklr57&i=919>.
43. Marcus D. The Public Interest Class Action . *Georgetown Law Journal*. Vol 104. No 4. 777–834. <https://heinonline.org/HOL/P?h=hein.journals/glj104&i=787>.
44. Singel R. Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims. *Wired*. December 17, 2009. <https://www.wired.com/2009/12/netflix-privacy-lawsuit/>.
45. Buley T and The Firewall. Netflix Settles Privacy Lawsuit, Cancels Prize Sequel. *Forbes*. March 12, 2010. <https://www.forbes.com/sites/firewall/2010/03/12/netflix-settles-privacy-suit-cancels-netflix-prize-two-sequel/?sh=32e7ccf5951e>.
46. Rabin R L. Lawyers for Social Change: Perspectives on Public Interest Law . *Stanford Law Review*. Vol 28. No 2. 207–262. <https://heinonline.org/HOL/P?h=hein.journals/stflr28&i=228>.
47. Southworth A. Conservative Lawyers and the Contest over the Meaning of Public Interest Law . *UCLA Law Review*. Vol 52. No 4. 1223–1278. <https://heinonline.org/HOL/P?h=hein.journals/uclalr52&i=1237>.
48. NATIONAL FAIR HOUSING ALLIANCE; FAIR HOUSING JUSTICE CENTER, INC.; HOUSING OPPORTUNITIES PROJECT FOR EXCELLENCE, INC.; FAIR HOUSING COUNCIL OF GREATER SAN ANTONIO v. Facebook. United States District Court, Southern District of New York. <https://nationalfairhousing.org/wp-content/uploads/2019/03/2018-06-25-NFHA-v.-Facebook.-First-Amended-Complaint.pdf>.
49. Sherwin G. How Facebook Is Giving Sex Discrimination in Employment Ads a New Life. *ACLU*. 2018. <https://www.aclu.org/blog/womens-rights/womens-rights-workplace/how-facebook-giving-sex-discrimination-employment-ads-new>.

50. Communications Workers of America vs. T-Mobile US, Inc. Amazon.com, Inc., Cox Communications, Inc., Cox Media Group, LLC, and similarly situated employers and employment agencies, DOES 1 through 1,000. United States District Court, Northern District of California.
<https://doi.org/https://www.onlineagediscrimination.com/sites/default/files/documents/og-cwa-complaint.pdf>.
51. Statt N. Facebook signs agreement saying it won't let housing advertisers exclude users by race. The Verge. July 24, 2018. <https://www.theverge.com/2018/7/24/17609178/facebook-racial-discrimination-ad-targeting-washington-state-attorney-general-agreement>.
52. Gifis S. Barron's Law Dictionary. Barron's Educational Series. 2010.
53. Markman S. FOIA Update: New Fee Waiver Policy Guidance. 1987.
54. Tabor A. Proposed Acquisition of Instagram, Inc. by Facebook, Inc. File No. 121-0121. Federal Trade Commission. 2012.
https://www.ftc.gov/sites/default/files/documents/closing_letters/facebook-inc./instagram-inc./120822zeiglerinstagramcltr.pdf.
55. Tracy R. The U.S. vs. Facebook: Why Is the FTC Filing an Antitrust Lawsuit Now? The Wall Street Journal. December 10, 2020. <https://www.wsj.com/articles/the-u-s-vs-facebook-why-is-the-ftc-filing-an-antitrust-lawsuit-now-11607607238>.
56. Niskanen W A. Bureaucracy and representative government. 1971.
57. Hausman D M and McPherson M S. Economic Analysis, Moral Philosophy and Public Policy. Cambridge University Press. Cambridge, UNITED KINGDOM. 2006.
58. Barry B. Theories of Justice: A Treatise on Social Justice. University of California Press. 1989.
59. Hartley C. Two Conceptions of Justice as Reciprocity. Social Theory and Practice. Vol 40. No 3. 409–432. April 8, 2014. <http://www.jstor.org.ezp-prod1.hul.harvard.edu/stable/24332305>.
60. Gilheany J, Wang D, and Xi S. The Model Minority? Not on Airbnb.com: A Hedonic Pricing Model to Quantify Racial Bias against Asian Americans. Technology Science. 2015.
61. Edelman B G and Luca M. Digital Discrimination: The Case of Airbnb.com. SSRN Electronic Journal. 2014. <https://doi.org/10.2139/ssrn.2377353>.
62. Parkinson H J. #AirBnBWhileBlack hashtag highlights potential racial bias on rental app. The Guardian. May 5, 2016.
<https://www.theguardian.com/technology/2016/may/05/airbnbwhileblack-hashtag-highlights-potential-racial-bias-rental-app>.

63. Wong J. Airbnb hires former attorney general Eric Holder to fight discrimination. The Guardian. July 20, 2016. <https://www.theguardian.com/technology/2016/jul/20/airbnb-hires-eric-holder-racial-discrimination-bias>.
64. Fulwood S. Airbnb announces booking policy change to head off outcry over persistent racial discrimination. ThinkProgress. October 24, 2018. <https://archive.thinkprogress.org/airbnb-changes-photo-policy-combat-racial-discrimination-4f71c375553a/>.
65. Airbnb. A new way we're fighting discrimination on Airbnb. Airbnb Resource Center. 2020. <https://www.airbnb.com/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201>.
66. Dixon P. A Brief Introduction to Fair Information Practices. World Privacy Forum. 2008. <https://www.worldprivacyforum.org/2008/01/report-a-brief-introduction-to-fair-information-practices/>.
67. Cavoukian A. Privacy by Design...take the challenge. 2009.
68. ACM. ACM Code of Ethics and Professional Conduct. ACM. 2018. .
69. Rawls J. Political Liberalism. Columbia University Press. 1993.
70. Knight J. JUSTICE AND FAIRNESS. Annual Review of Political Science. Vol 1. No 1. 425–449. June 1, 1998. <https://doi.org/10.1146/annurev.polisci.1.1.425>.
71. Soo Z. YouTube suspends Trump's channel for at least a week. AP. January 13, 2021. <https://apnews.com/article/youtube-suspend-trump-channel-1-week-0f6166b8a3d3452709968aa1ba934cdc>.
72. Twitter. Permanent suspension of @realDonaldTrump. Twitter Blog. 2021. https://blog.twitter.com/en_us/topics/company/2020/suspension.html.
73. Dwoskin E. Facebook outsources its decision to ban Trump to oversight board. The Washington Post. January 21, 2021. <https://www.washingtonpost.com/technology/2021/01/21/facebook-oversight-board-trump-ban/>.
74. Mankiw G. Principles of Economics. South-Western College Pub. 2013.
75. Finlayson A C, Lyson T A, Pleasant A, Schafft K A, and Torres R J. The "Invisible Hand": Neoclassical Economics and the Ordering of Society1. Critical Sociology. Vol 31. No 4. 515–536. July 1, 2005. <https://doi.org/10.1163/156916305774482183>.
76. McClure J and Watts T. The Greatest Externality Story (N)ever Told. The American Economist. Vol 61. No 2. 157–177. April 9, 2016. <https://www-jstor-org.ezp-prod1.hul.harvard.edu/stable/26725777>.

77. Boudreaux D J and Meiners R. Externality: Origins and classifications. *Natural Resources Journal*. 2019.
78. Verveer P. Countering Negative Externalities in Digital Platforms. 2019.
79. Edelman B and Stemler A. From the digital to the physical: Federal limitations on regulating online marketplaces. *Harvard Journal on Legislation*. 2019. <https://doi.org/10.2139/ssrn.3106383>.
80. Goovaerts D. Biden targets \$100B for universal broadband access in \$2T plan. *Fierce Telecom*. March 31, 2021. <https://www.fiercetelecom.com/telecom/biden-targets-universal-broadband-access-2t-plan>.
81. Brake D and Bruer A. How to Bridge the Rural Broadband Gap Once and For All. 2021.
82. Khan L M. Amazon's antitrust paradox. *Yale Law Journal*. 2017.
83. Posner R A. The Chicago School of Antitrust Analysis. *University of Pennsylvania Law Review*. 1979. <https://doi.org/10.2307/3311787>.
84. Horwitz J. Zuckerberg's Deal Making for Facebook Is Central to Antitrust Cases. *The Wall Street Journal*. December 10, 2020. https://www.wsj.com/articles/zuckerbergs-deal-making-for-facebook-is-central-to-antitrust-cases-11607596201?mod=article_inline.

Chapter 2

How Facebook's Advertising Algorithms Can Discriminate By Race and Ethnicity

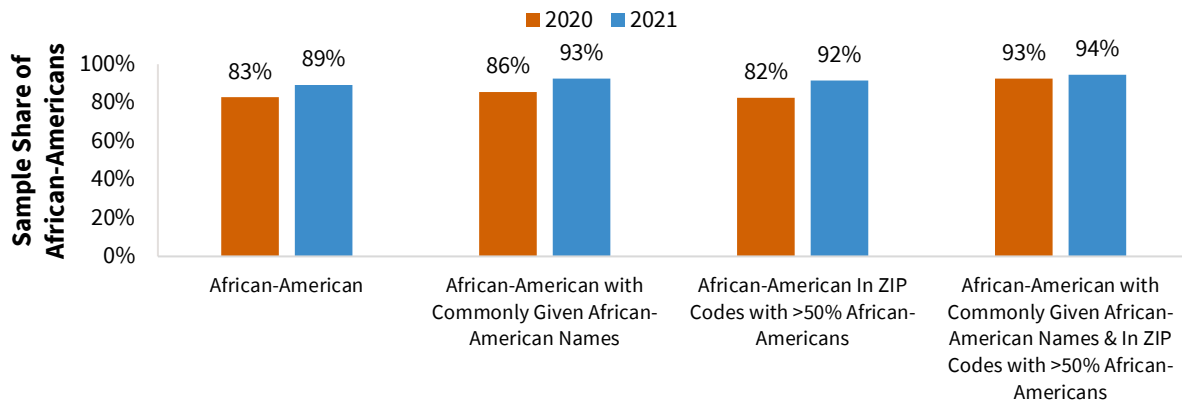
Jinyan Zang

Highlights

- This study examines the racial and ethnic biases of Facebook's advertising platform in January 2020 and January 2021, before and after a major July 2020 boycott of Facebook by advertisers over issues of misinformation and civil rights.
- In 2021, Facebook's "African-American Culture" ad targeting option contained 75% fewer White users than the old "African American (US)" option they removed in the previous year.
- Facebook's tools to help advertisers find similar users to their existing customers exhibited bias towards including more African-Americans or Whites depending on which racial group was dominant in an advertiser's customer list, and this was true for the Lookalike Audience tool, as well as the Special Ad Audience tool that Facebook designed to explicitly not use sensitive demographic attributes when finding similar users.
- The degree of bias towards including more African-Americans or Whites in a Lookalike or Special Ad audience was larger when using customer lists of individuals with racially stereotypical names or ZIP codes as the basis for each tool.

- Similarly, Lookalike audiences can also become biased towards Asians, reaching up to 100% Asian in one case when using a customer list of Asians with stereotypical names and ZIP codes, and Lookalike audiences based on Hispanics over-represented Hispanics versus Non-Hispanics.

Sample share of African-American voters in Lookalike Audiences based on lists of NC voters with different traits



Traits of NC Voters on Facebook Used to Create Lookalike Audiences

As the customer list used to create the Lookalike audience contain more stereotypically African-American traits in terms of names and ZIP codes, the sample shares of African-American voters in the corresponding Lookalike audiences became more biased, increasing up to 93% in 2020 and 94% in 2021.

Abstract

Over the last 5 years, Facebook has faced repeated criticism and lawsuits over the potential for discrimination on its ad platform. In July 2020, advocacy groups organized a high-profile boycott of Facebook's advertising platform over issues of misinformation and civil rights that successfully pressured over a thousand major corporations to stop advertising on Facebook for the month. Facebook responded by releasing its civil rights audit and announcing the removal of its much-criticized multicultural affinity groups such as "African American (US)", "Asian American (US)", and "Hispanic (US – All)" as ad targeting options. This study examines if Facebook has gone far enough to prevent discrimination on its advertising platform. I collected data on Facebook's ad platform in January 2020 and in January 2021. I compared the racial and ethnic breakdown of the old multicultural affinity groups against similar sounding cultural interest groups such as "African-American Culture", "Asian American Culture", and "Hispanic American Culture" that were still usable by advertisers in 2021. I also used a set theory approach to study racial and ethnic biases in Facebook's Lookalike Audience and Special Ad Audience algorithms in both time periods.

Results summary: I found that in 2021 Facebook's "African-American Culture" ad targeting option contained 75% fewer White users than the old "African American (US)" option they removed in the previous year. Facebook's tools to help advertisers find similar users to their existing customers exhibited bias towards including more African-Americans or Whites depending on which racial group was dominant in an advertiser's customer list, and this was true for the Lookalike Audience tool, as well as the Special Ad Audience tool that Facebook created to explicitly not use sensitive demographic attributes when finding similar users. The degree of bias towards including more African-Americans or Whites in a Lookalike or Special Ad audience was larger when using customer lists of individuals with racially stereotypical names or ZIP codes as the basis for each tool. There was also significant bias

towards Asians in Lookalike audiences based on Asians, reaching up to 100% Asian in one case. Finally, Lookalike audiences based on Hispanics over-represented Hispanics versus Non-Hispanics. This study shows that Facebook's ad platform can be used to discriminate by race and ethnicity through using different cultural interest groups as targeting options or through using the Lookalike and Special Ad Audience tools in 2021. It also provides evidence that "fairness through unawareness", the idea that discrimination is prevented by eliminating the use of protected class variables or close proxies in a model, does not reduce the potential for algorithmic bias.

Introduction

Facebook is the second largest digital advertising platform, behind Google, in the US today [1]. It offers advertisers multiple tools to target their ads at different users on the platform.

- An advertiser can use Facebook’s own “Detailed Targeting” options which are categories of users that share the same demographic, interest, or behavior based on Facebook’s analysis of its user data.
- Facebook also allows advertisers to create a “Custom Audience” or a list of customers or individuals that the advertiser already has data on for Facebook to target directly.
- Facebook also offers to create a “Lookalike Audience”, which is based on an advertiser’s existing Custom audience by finding the users whom Facebook has identified as the most similar to the ones currently in the Custom audience.
- Finally, for ads related to housing, employment, and credit, Facebook can create a “Special Ad Audience” which is like a Lookalike audience except Facebook does not use sensitive attributes such as “age, gender or ZIP code” in considering which users are similar enough to include [2].

Over the last 5 years, Facebook has faced repeated criticism, lawsuits, and controversies over the potential for discrimination on its ad platform. Journalists have demonstrated how easy it is to exclude users whom Facebook has classified as being in racial or ethnic affinity groups from being targeted by housing or employment ads [3, 4]. Researchers have demonstrated racial and ethnic biases in Facebook’s Lookalike Audience and Special Ad Audience algorithms [5]. Facebook has been sued by the National Fair Housing Alliance [6], the ACLU [7], the Communications Workers of America [8], the U.S. Department of Housing and Urban Development [9], and others [10] over issues of discrimination on its advertising platform violating civil rights laws such as the Fair Housing Act and

the Civil Rights Act of 1964. There is also ongoing controversy over how Facebook’s platform can be used by political actors, both foreign and domestic, to spread misinformation and especially target racial and ethnic minorities in the 2016 and 2020 election cycles [11, 12, 13, 14, 15].

In July 2020, a high-profile boycott of Facebook’s advertising platform over issues of misinformation and civil rights was organized by advocacy groups including the NAACP, the Anti-Defamation League, Color of Change, and others to call on major corporations to stop advertising on Facebook for the month [16]. More than 1,000 large companies including Microsoft, Starbucks, Target, and others participated in the boycott [17].

On July 8, 2020, Facebook released its own civil rights audit conducted by Laura Murphy, former Director of the ACLU Legislative Office, and attorneys at the law firm Relman Colfax [18]. The audit criticized Facebook for having “placed greater emphasis on free expression” instead of balancing that with the “value of non-discrimination” [18]. Regarding the audit, Facebook Chief Operating Officer Sheryl Sandberg stated, “it is the beginning of the journey, not the end” [19].

As a concrete step of that journey, on August 11, 2020, Facebook announced that it will retire its controversial “multicultural affinity” groups that allowed advertisers to target users whom Facebook has categorized as “African American (US)”, “Asian American (US)” or “Hispanic (US – All)” [20].

However, has Facebook gone far enough to prevent discrimination on its advertising platform?

While not every case of advertising discrimination is illegal or even potentially undesirable, such as the case of marketing textbooks to students, this study seeks to reveal to what degree can Facebook’s ad platform carry out racial and ethnic discrimination through its different tools, which may be the intended or unintended goal of different advertisers on Facebook.

I studied this question by testing Facebook’s advertising platform in two waves, first in January 2020 and again in January 2021. While in 2021 Facebook no longer offers multicultural affinity groups as targeting options for advertisers, I tested the similar sound cultural interest groups that Facebook still offered as targeting options, such as “African-American Culture”, “Asian American Culture”, and “Hispanic American Culture”. I also tested if Facebook’s other advertising tools, such as Lookalike Audiences and Special Ad Audiences, can discriminate by race and ethnicity. Thus, this study examined the following questions about Facebook’s advertising platform:

- Are the cultural interest groups as racially and ethnically homogenous as the old multicultural affinity groups?
- Do Lookalike and Special Ad audiences reflect racial and ethnic biases depending on the lists of individuals used to generate them?
- Is the degree of racial and ethnic bias in Lookalike and Special Ad audiences affected by well-established racial factors from the offline world such as the name or ZIP code of individuals used to create the audience?
- What are the differences in the type and degree of bias observed in Facebook’s advertising tools in 2021 versus 2020?

Background

The Rise of Digital Advertising and Microtargeting

In 2019, digital advertising spending (\$129 billion) eclipsed traditional advertising (\$109 billion) for the first time in the US , and Facebook itself accounted for 22% of all digital ad spending, second only behind Google’s 37% market share [1]. Thus, law enforcement agencies, advocacy groups, the media, and researchers are increasingly focused on the need to prevent advertising discrimination on the country’s second most popular digital advertising platform.

Facebook’s advertising platform provides multiple targeting tools to help an advertiser “microtarget” only the users they want which are the focus of this study.

- “Detailed Targeting” options allow an advertiser to target a prepackaged group of Facebook users who share common attributes based on Facebook’s data analysis of the ads they click, the pages they engage with, the activities they do on its websites, and other data. “Detailed Targeting” options are organized into three categories: demographics, interests, and behaviors (Figure 2.1).

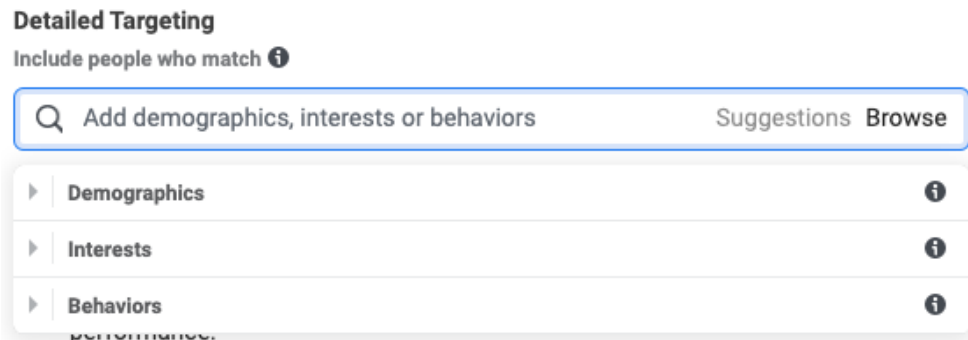


Figure 2.1. Adding Detailed Targeting Option to a Facebook Advertising Campaign.

- “Custom Audiences” allow an advertiser to upload their own contact list of customers or individuals for Facebook to target or to create an audience by integrating Facebook’s trackers on their websites or apps. When an advertiser uploads a customer list, Facebook then matches data fields such as email, phone number, first name, last name, city, state, country, ZIP code, date of birth, gender, and more with the data it has on its users. Choosing a Custom audience means targeting ads only to the Facebook users that were matched to the advertiser’s customer list (Figure 2.2). Facebook can also create Custom audiences based on which Facebook users have interacted with an advertiser’s content on the Facebook platform

such as commenting on a video, liking an Instagram post, attending a Facebook event, and more (Figure 2.3).

Create an Audience From a Customer List

Prepare Your Customer List

Your customer list is a CSV or TXT file that contains information used to build your audience. Identifiers in your customer list are used to match with Facebook users. The more identifiers you provide, the better the match rate.

Include at least one main identifier ⓘ

Email Phone Number Mobile Advertiser ID Facebook App User ID

Facebook Page User ID First Name Last Name

Include more identifiers ⓘ

City State/Province Country ZIP/Postal Code Date of Birth

Year of Birth Gender Age

Add value information to create a value-based lookalike ⓘ

Customer Value

[Download List Template](#)

[See Formatting Guidelines](#)

[Import From Mailchimp](#)

Your Customer List Information Is Hashed

Before the list is sent to Facebook for your audience to be created, we use a cryptographic security method known as hashing, which turns the identifiers into randomized code and cannot be reversed.

[Learn More](#)

Back Next

Figure 2.2. Identifiers Used to Match an Advertiser’s Customer List with Users on Facebook.

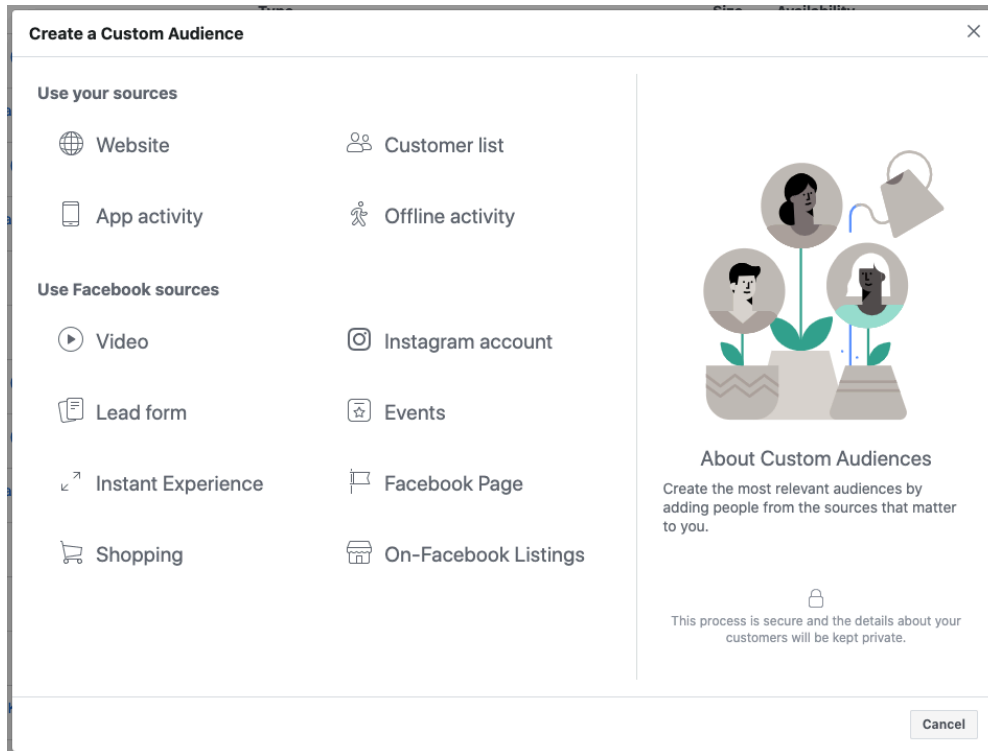


Figure 2.3. Sources for Custom Audience on Facebook.

- “Lookalike Audiences” allow an advertiser to reach people similar to a designated source audience created through the Custom Audiences tool. According to Facebook, “You choose a source audience, and we identify the common qualities of the people in it. Then we find people like them, using your selected location and desired audience size.” The advertiser can choose one or multiple countries as the “Audience Location”, and Facebook will find the 1-10% of the population of Facebook users in those countries that are most similar to the advertiser’s source audience (Figure 2.4).

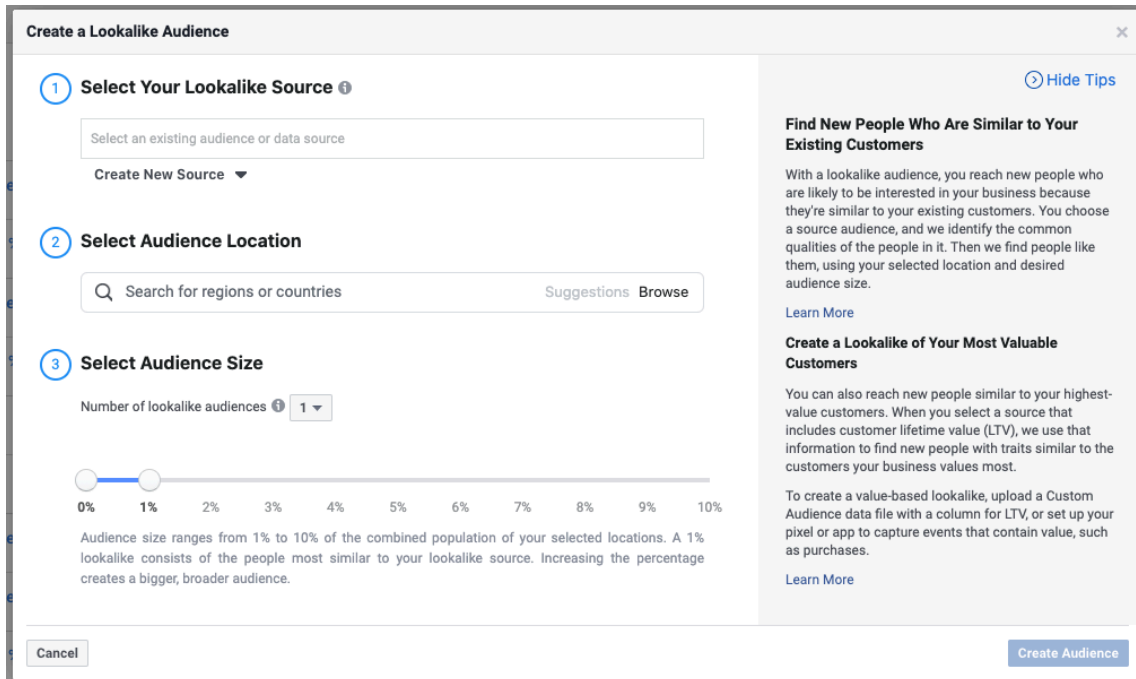


Figure 2.4. Creating a Lookalike Audience.

- “Special Ad Audiences” allow an advertiser to create a Lookalike Audience that finds similar people to their source audience “in online behavior without considering things like age, gender or ZIP code” [2] specifically for ads in the categories of housing, employment, and credit which are regulated by anti-discrimination laws (Figure 2.5). The settings for creating a Special Ad Audience are very similar to the Lookalike Audience settings. The advertiser chooses a source audience, then selects one or multiple countries as the “Audience Location”, and finally asks Facebook to find the 1-10% of the population of Facebook users in those countries that are most similar to the advertiser’s source audience (Figure 2.6).

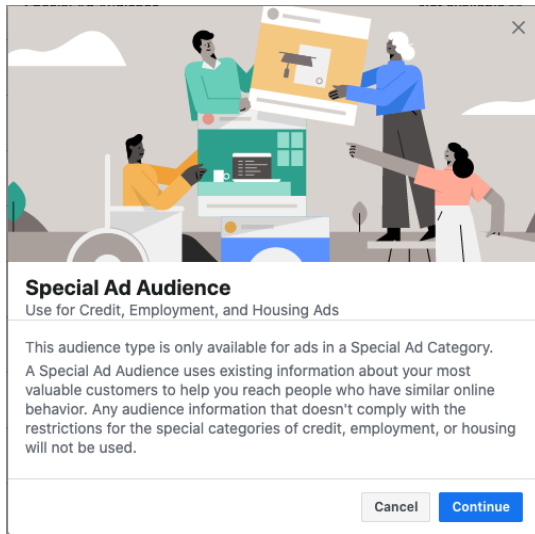


Figure 2.5. Special Ad Audience Tool.

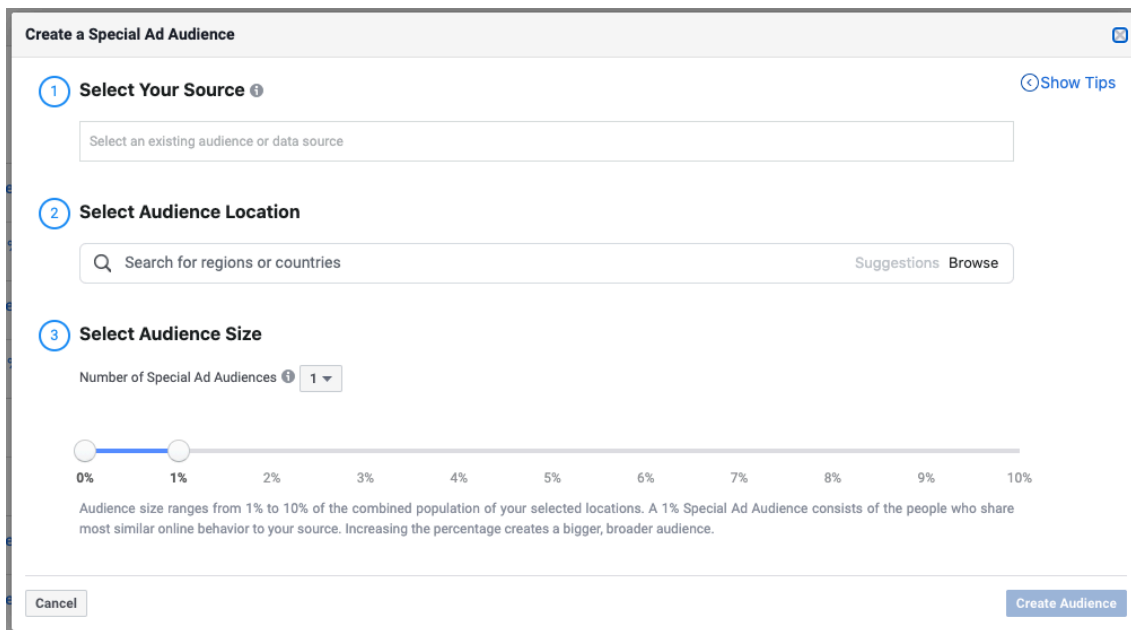


Figure 2.6. Creating a Special Ad Audience.

An advertiser can use the different tools of Facebook's ad platform to define the desired intersection of attributes in whom they want to see an ad by using Facebook's Audience selection tool (Figure 2.7). First, the advertiser is able to choose one or multiple Custom, Lookalike, or Special Ad

audiences to “include” as potential targets. The advertiser can also choose one or multiple Custom or Lookalike audience to “exclude” as potential targets, though they are not able to choose to exclude a Special Ad audience. Next, in terms of geographical targeting options, the advertiser can choose to only target individuals currently living in, recently located in, or traveling to a country, state, city, ZIP code, media market, or Congressional District. Another option is to drop a pin on the map and target all individuals within a setting of 1 to 50 miles of the pin’s location. The advertiser also has options to target by age, gender, and language. The advertiser can choose to add one or more of the predefined “Detailed Targeting” options based on demographics, interests, and behaviors created by Facebook to “include” or “exclude” as criteria for their target audience. There are also additional options for an advertiser to include or exclude users who have connections with the advertiser’s Facebook pages, apps, or events.

Audience
Define who you want to see your ads. [Learn More](#)

Create New Audience Use Saved Audience ▾

Custom Audiences Create New ▾

INCLUDE people who are in at least ONE of the following

EXCLUDE people who are in at least ONE of the following

Locations
Location - Living In:

- United States: North Carolina

Age
18 - 65+

Gender
All genders

Detailed Targeting
All demographics, interests and behaviors
 Detailed Targeting Expansion:

- Off

Languages
All languages

[Hide Options ▸](#)

Connections

Figure 2.7. Facebook’s Audience selection tool.

Digital advertising platforms like Facebook give advertisers new abilities to target individuals which were never in one place before when compared to traditional advertising.

First, with print or broadcast advertising, the advertiser will reach anyone who reads the newspaper or watches the TV show where ad is placed. On Facebook, the advertiser can target only specific individuals that they want to reach through Facebook’s Custom Audiences or match a very

specific intersection of different targeting criteria in terms of demographics, interests, behaviors, and more (Figure 2.7).

Second, with physical mail or email advertising, the advertiser can only reach those whose physical or email address they already have. On Facebook, the advertiser can potentially target millions of new individuals who are very similar to their existing customers without needing to know the contact information for this new audience through Facebook's Lookalike Audiences or Special Ad Audiences tools.

The increased targeting capabilities of Facebook's advertising platforms also present new risks for discriminatory advertising to have a widespread impact.

Facebook and Race

Given this potential for discrimination, it's important to note that unlike gender or age, Facebook doesn't ask their users to provide their race directly. However, until August 11, 2020, the Facebook Ads platform did infer a user's "Multicultural Affinity" based on their behavior on Facebook and how similar it is to others of the same affinity group. Thus, Facebook offered advertisers the ability to target by Multicultural Affinity classifications for "African American", "Asian American", and "Hispanic", with the following Hispanic sub-categories: "Hispanic - Bilingual", "Hispanic - Spanish Dominant", and "Hispanic - English Dominant".

According to Facebook:

The word "affinity" can generally be defined as a relationship like a marriage, as a natural liking, and as a similarity of characteristics. We are using the term "Multicultural Affinity" to describe the quality of people who are interested in and likely to respond well to multicultural content. What we are referring to in these affinity groups is not their genetic makeup, but their affinity to the cultures they are interested in...The Facebook multicultural targeting solution is based on affinity, not ethnicity. This provides advertisers with an opportunity to serve highly relevant ad content to affinity-based audiences [21].

While Facebook argued that these multicultural affinity classifications do not facilitate racial and ethnic discrimination by advertisers, it has faced repeated criticism from journalists, researchers, civil rights groups, and law enforcement agencies over this issue.

Media Attention

In October 2016, ProPublica journalists Julia Angwin and Terry Parris Jr. found that “Facebook lets advertisers exclude users by race” [3]. Advertisers could include and exclude users from being targeted based on Facebook’s “Ethnic Affinity” categories, which included African American, Asian American, and Hispanic. In order to demonstrate the possibility of discrimination, ProPublica ran their own housing -related ads on Facebook using these categories to exclude users of a given ethnic affinity (Figure 2.8). Facebook responded that ethnic affinity categories existed as part of their “multicultural advertising” effort and that ethnic affinity was not the same as race but rather a membership category that Facebook created based on the pages and posts a user liked or engaged with on its website [3].

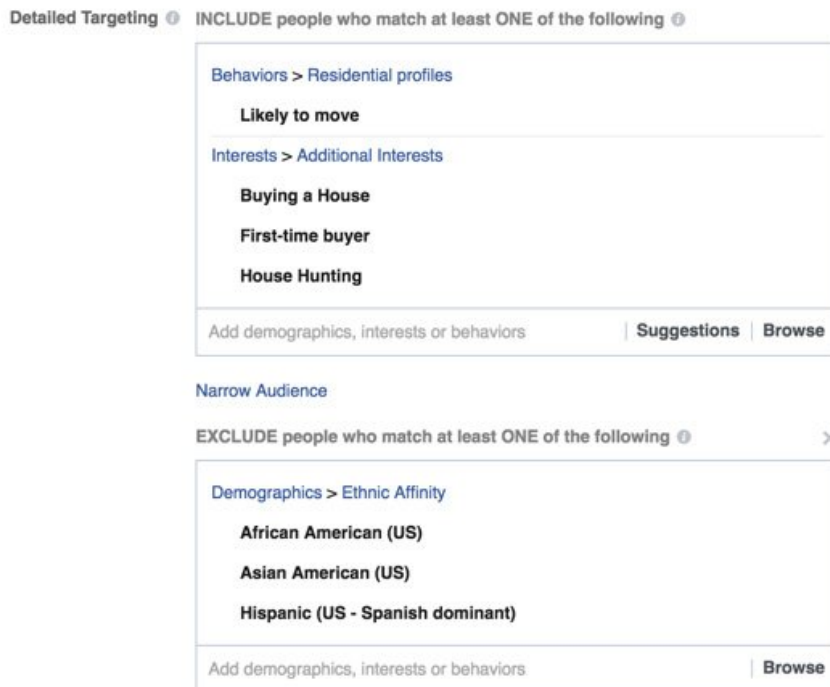


Figure 2.8. ProPublica’s example in October 2016 of using the “Ethnic Affinity” categories within Facebook’s Detailed Targeting options to potentially exclude minority users from seeing a housing-related ad in a discriminatory manner [3].

One year after their initial reporting, ProPublica journalists again found in November 2017 that “Facebook (Still) Letting Housing Advertisers Exclude Users by Race” [4]. ProPublica was able to purchase housing-related ads that excluded on the basis of multicultural affinity groups, as well as mothers of high school kids, people interested in wheelchair ramps, Jews, expats from Argentina, and Spanish speakers, which are other categories that appear related to “protected classes” in anti-discrimination law. Every ad was approved within three minutes. Facebook acknowledged that “This was a failure in our enforcement and we’re disappointed that we fell short of our commitments” [4]. The main difference in advertiser experience that ProPublica noted was that “Ethnic Affinity” was now renamed “Multicultural Affinity” and moved from the Demographics section of targeting options to the Behaviors section. Since Facebook also let advertisers select which ZIP codes to target, ProPublica was also able to target their ads at only majority non-Hispanic White ZIP codes in Brooklyn, as a demonstration of a practice they described as similar to “redlining” [4], a historical discriminatory practice by landlords, brokers, and lenders to oftentimes exclude African-Americans from moving to predominantly White neighborhoods, which is now prohibited by the Fair Housing Act.

Academic Research

Academic researchers have also found discriminatory potential in Facebook’s other targeting tools such as Lookalike Audiences and Special Ad Audiences. A December 2019 study found that Lookalike and Special Ad audiences could be significantly biased depending on the demographics of the source audience [5]. For example, when the source audience was all women, the Lookalike audience was 96.1% women, and the Special Ad audience was 91.2% women. Similar relationships

were also observed for different racial groups. For example, a source audience that was 100% Black created a Lookalike audience which had a 61% overlap with a given list of 900,000 other Black voters and only a 16% overlap with a second list of 900,000 White voters in the same state. The corresponding Special Ad audience had a 62% overlap with the Black voter list and a 12% overlap with the White voter list. On the other hand, a Lookalike Audience based on a source audience that was 100% White had a much smaller 17% overlap with the Black voter list and a much larger 42% overlap with the White voter list. Similarly, the corresponding Special Ad Audience had a 10% overlap with the Black voter list and a higher 36% overlap with the White voter list.

Legal Actions Taken Against Facebook and Its Advertisers

There have also been multiple litigation filed on the issue of advertising discrimination filed against Facebook. On November 3, 2016, a class action lawsuit, *Mobley v. Facebook*, was filed in U.S. District Court arguing that Facebook violated federal anti-discrimination laws for housing (Fair Housing Act) and employment (Civil Rights Act) by citing the ProPublica reporting [22, 23]. On March 27, 2018, a coalition of housing advocacy groups, led by the National Fair Housing Alliance (NHFA), filed a lawsuit against Facebook for violating the Fair Housing Act by allowing advertisers to discriminate against legally protected groups such as mothers, the disabled, and Spanish-language speakers [24]. The NHFA conducted their own investigation of Facebook's advertising platform, and they also were able to exclude individuals that Facebook has classified as "Disabled American Veterans", "moms of preschool kids", and interested in "English as a second language" [6]. On September 18, 2018, the ACLU, the Communications Workers of America (CWA), and the employment law firm Outten & Golden LLP filed charges with the Equal Employment Opportunity Commission (EEOC) against Facebook and ten major corporations that targeted ads for jobs to younger male Facebook users only, excluding all women and older users [7]. In a separate 2018 lawsuit filed by the

Communications Workers of America vs. T-Mobile, Amazon, and 1,000 other large employers using Facebook ads, the plaintiffs allege that not only did the employers use Facebook's prepackaged targeting options in a discriminatory manner, but that they also used Facebook's Lookalike Audience tool to target candidates who are demographically similar to their existing workforce in ways that marginalized older workers [8, 25].

Regulatory agencies and other law enforcement offices have also investigated and filed suit against Facebook for advertising discrimination. In 2016, Washington State's Attorney General's Office started a 20-month investigation of Facebook's advertising platform. The investigators were able to purchase ads that excluded protected categories of people from being targeted and found real-world examples of ads that did just that [10, 26, 27]. According to the Attorney General Bob Ferguson, "Facebook's advertising platform allowed unlawful discrimination on the basis of race, sexual orientation, disability and religion...That's wrong, illegal, and unfair" [26].

On March 28, 2019, the U.S. Department of Housing and Urban Development (HUD) sued Facebook, because it "unlawfully discriminates based on race, color, national origin, religion, familial status, sex, and disability by restricting who can view housing-related ads on Facebook's platforms and across the internet" and it "mines extensive data about its users and then uses those data to determine which of its users view housing-related ads based, in part, on these protected characteristics" [9].

Political Controversy and Criticism from Civil Rights Advocacy Groups

Finally, while it is not unlawful in the United States to target political or news ads in a racially biased manner, the ability for Facebook's advertising platform to be used to spread misinformation or inflame racial tensions among different demographic groups has been a high-profile political controversy in the 2016 and 2020 General Elections.

Researchers studying the 3,519 ads that Facebook shared with Congress as part of the investigations into Russian interference with the 2016 elections found that many of the ads focused on black identity issues, such as police shootings, BlackLivesMatter, and discrimination [28]. The most popular Facebook targeting options used by the Russians included targeting users interested in Martin Luther King, African-American Civil Rights Movement, African-American history, Black Power, and other related categories [15]. In fact, 17 ads even used Facebook’s “African-American (US)” multicultural affinity group [15]. Another study found that across all the Russian-linked ads disclosed by Facebook, 52% had more than double the proportion of African-Americans in their target audience compared to Facebook’s US baseline [29]. Besides the Russians, other political actors may have also used Facebook’s advertising tools to target minority voters in controversial ways in 2016. For example, in October 2016, Trump campaign manager, Brad Parscale, announced that they will target an ad, which featured Clinton disparaging young African-Americans as “superpredators”, using Facebook’s advertising platform to “only the people we want to see it”, in order to suppress Clinton’s votes from African-Americans [30].

In 2020, there have been similar controversies about misinformation campaigns targeting minority voters on Facebook. For example, racial appeals and a focus on American social unrest during the summer of 2020 have been part of misinformation campaigns by groups linked to Russia, Iran, and China on Facebook and Twitter [11, 12]. Domestic political actors have also used these platforms to promote misleading or false ads [11, 13, 14]. For example, in August 2020, *The Washington Post* reported that FreedomWorks, a conservative political advocacy group established by David and Charles Koch, spent \$1,500 on a Facebook ad by the Protect My Vote page using LeBron James’ picture and a misquoted line to criticize mail-in ballots [14]. This ad targeted voters in swing states with high concentrations of minority voters [14]. Other political groups were able to exploit

Facebook’s fact-checking system to re-post and distribute nearly identical copies of previously taken down ads that Facebook’s fact-checking partners helped remove [13]. A significant amount of misinformation also targeted Hispanics in 2020, particularly Hispanic voters in Florida [31, 32, 33, 34]. According to Longoria, one tactic was to draw on “anti-blackness” bias to promote the idea that Black individuals were “harassing” Latinos under the guise of activism, such as one viral video that was shared 180,000 times on Facebook of two Black women harassing a Latino family celebrating at a party, which falsely labelled the women as members of Black Lives Matter [31]. Other misleading ads for example labelled Biden as a “comunista” or stated that Harris supported abortion up to the minutes before birth [31, 32]. A common problem was that fact-checking was not as robust or enforced for Spanish-language ads and posts on Facebook [33].

Due to the widespread misinformation, often targeting minority users, on Facebook, on June 17, 2020, civil rights groups such as the NAACP, the Anti-Defamation League, Color of Change, and others launched the “Stop Hate for Profit” campaign to pressure major corporate advertisers to stop advertising on Facebook for the month of July 2020 [35]. More than 1,000 major advertisers joined the boycott including Microsoft, Starbucks, Unilever, Target, and more [17]. The organizers of the boycott outlined 10 recommendations for Facebook to adopt such as hiring a C-suite executive to review the company’s products for discrimination, hate, and bias, participating in a regular third-party audit on identity-based misinformation and hate, stopping the amplification of content with ties to hate, misinformation, or conspiracies, ending the exemption of politicians from fact-checking, and other changes [35]. The organizers of the boycott argued that the racism on Facebook reflects the issues of systemic racism in America today which are worsened by Facebook’s technologies and its historically laid back approach to moderation [36].

Racial Discrimination by Humans

Outside of Facebook, researchers have found that racial discrimination and racial bias can occur in many different situations in everyday life, whether that's applying for a job or a loan or trying to get a doctor's appointment. In these cases, it's individual humans or a group of people working within a larger organization who are making the biased decisions. Oftentimes, similar to how Facebook doesn't collect race directly as a variable from its users, discrimination by others in society may also not rely on an explicit racial variable but rather using more implicit proxies for race.

One common proxy variable is to use the name of the individual to discriminate by race. Different racial groups tend to have distinct sounding first and/or last names [37, 38]. Field experiments have shown that discrimination may occur for individuals with different racially-affiliated names [39, 40]. For example, in 2003, Bertrand and Mullainathan found that resumes with White-sounding first names received 50 percent more callbacks for interviews than resumes with Black-sounding first names [39]. In 2016, Kang, DeCelles, et al. found a similar interview callback gap of 40% for James (White-sounding first name) versus Lamar (Black-sounding first name) when both resumes appear "White", i.e. the resumes don't describe participation in ethnically affiliated groups like an African-American fraternity, but no statistical difference in the lower callback rates for both names when the resumes appear "Black". For resumes that appear "Asian", Luke Zhang (White-sounding first name, Asian-sounding last name) received a statistically significant 83% more interview callbacks than Lei Zhang (Asian-sounding first and last name) [40]. A meta-analysis of these field experiments in 2017 found no change in African-American callback rates since 1989 while there was a decline in discrimination against Latinos [41]. Audit studies have also looked at the response rate gap to a housing inquiry, a mortgage application, a request for a doctor's appointment, a request for help from a public official, and more [42, 43, 44].

Another method for racial discrimination is based on the home of the individual since historical factors have resulted in high levels of racial segregation in many US cities [45]. Before the passage of the Fair Housing Act in 1968, lenders and the real estate industry commonly practiced “redlining”, where certain minority neighborhoods were declined access or offered inadequate access to affordable mortgages and minority renters and borrowers were often prevented from moving to White-dominant neighborhoods [45]. While the Fair Housing Act made redlining and housing discrimination illegal, disparities in access to housing and borrowing still exist between racial groups [45]. For example, a 2015 study in Baltimore found that race is the most statistically significant factor in predicting who gets a mortgage, and the disparity ratio of loans to the population is 210% for Whites and 27% for African-Americans [46]. Similar patterns exist for 61 metro areas around the country [47]. Hanson and Hawley found in their audit study across the 10 largest US cities that African-Americans faced lower response rates from landlords than Whites in racially-mixed neighborhoods, areas with rent above the median rent for the city, and neighborhoods close to the city center or first ring suburbs [44]. According to research by Raj Chetty, Nathaniel Hendren, et al. these disparities contribute to the “substantially lower rates of upward mobility and higher rates of downward mobility” for African-Americans relative to Whites, “leading to large income disparities that persist across generations” [48]. Chetty et al. found that less than 5% of Black children grew up in rich areas with a poverty rate below 10% while more than 63% of White children did [48]. Amongst researchers of algorithmic bias, these findings have contributed to concerns over the inclusion of geography-related variables such as ZIP codes as data inputs that may serve as proxy variables for race and contribute to algorithmic bias [49, 50].

Due to this research on how discrimination often occurs “offline”, this study examines if the algorithms used by Facebook to create Lookalike and Special Ad audiences exhibit higher degrees of

racial and ethnic bias when using individuals with racially and ethnically stereotypical names and ZIP codes in the source audience.

Discrimination by Algorithms and Anti-Discrimination Laws

The Civil Rights Movement and its legislative successes have resulted in three major federal laws that forbid racial and other types of discrimination in housing (Fair Housing Act), employment (Civil Rights Act of 1964) and credit (Equal Credit Opportunity Act), including in advertisements for those sectors [49]. These laws provide a legal recourse for justice for victims of discrimination and a potential threat of legal repercussions to discriminatory actors. Thus, it's these laws that the ACLU, HUD, and others have argued that Facebook and advertisers on its platform have violated. However, as discrimination moves from decisions made by humans to those by machines and algorithms, how to detect discrimination and enforce existing anti-discrimination laws has become more complicated.

Latanya Sweeney's 2013 study on Google's ad platform was an early, high-profile example of a digital audit study to detect discrimination. She found that statistically significantly more ads using Black-identifying first names had the word "arrest" in the ad's text than ads using White-identifying first names, with an adverse impact ratio of 77% for Reuters.com, which had ads served by Google, and an adverse impact ratio of 40% for Google.com search results [51]. Other instances of algorithmic bias have been found in facial recognition systems [52], online shopping [53], search engines [54, 55, 56, 57], job sites and hiring software [58, 59, 60], translation services [61], healthcare [62], and other systems.

Since much of the work on algorithmic discrimination has focused on the products of private sector companies, the legal concepts of "disparate treatment" and "disparate impact", which were established by Title VII of the 1964 Civil Rights Act are the primary jurisprudence that is applicable [63].

Disparate treatment focuses on intentional discrimination, when individuals are treated differently because of their protected class attribute, such as their race, color, national origin, religion, sex, disability, or familial status [63]. There can be overt evidence of disparate treatment such as when a lender has a policy of a higher credit limit for older borrowers versus younger borrowers [64]. Or there can be comparative evidence of disparate treatment such as when two borrowers who are otherwise similar get treated differently by a lender on the basis of their protected class attribute [64].

Disparate impact occurs when there is disproportionate burden in outcomes for a specific group, where showing intentionality is not required [63, 64]. There is a burden-shifting framework to decide whether a company is liable under disparate impact. First, the plaintiff needs to show evidence of disproportionate outcomes across demographic groups. The Equal Employment Opportunity Commission's (EEOC) 80-20 Rule is often cited by algorithmic bias literature though it primarily applies to labor issues [65, 66]. Then, the defendant corporation can respond to try to show if it had a "business necessity" to justify its decision-making process [63, 64]. For example, a job that required lifting heavy supplies may argue that it's a business necessity to include a candidate's ability to lift weights in its hiring process, even if it leads to hiring more male versus female candidates. Finally, the plaintiff must show that there is a less discriminatory alternative that could meet the business necessity [63].

Applying these concepts to the different Facebook ad targeting tools, if an advertiser uses Facebook's targeting options in an intentionally discriminatory way such as creating age limits, gender criteria, or choosing the multicultural affinity targeting options like "African-American (US)" then that may be considered to be disparate treatment. If an advertiser has a predominantly White customer base, they then use either Facebook's Lookalike Audience or Special Ad Audience tools to create a new target audience of other users whom Facebook has identified as being similar to their

existing customers. If the resulting target audience is racially biased by being predominantly White, in following the disparate impact framework, it doesn't matter whether the advertiser or Facebook intentionally meant to discriminate. First, the plaintiffs would need to show that Black and other minority users were impacted disproportionately by not being targeted by the advertiser relative to White peers. Then, the advertiser may argue it was a business necessity to use Facebook's tools to find new ad audiences similar to their existing customers, and Facebook may argue that its business necessity meant creating the best possible Lookalike or Special Ad audience to serve their advertising clients. Finally, the plaintiffs would need to show that there are less discriminatory alternatives that could still achieve the same business necessities without predominantly targeting more White users.

This study examines to what degree would Facebook's racially-affiliated targeting options lead to disparate treatment by race and ethnicity and also to what degree would Facebook's Lookalike Audience and Special Ad Audience tools lead to disparate impact by race and ethnicity depending on the demographics of the source audience.

Facebook's Response to Charges of Discrimination

Facebook has become more responsive over time in making changes to its advertising platform after being repeatedly accused of discrimination by journalists, advocacy groups, and law enforcement agencies.

In response to the original ProPublica report from 2016 [3], Facebook announced efforts in February 2017 to improve its anti-discrimination efforts such as creating enforcement tools to disapprove ads that use multicultural affinity groups – the new name “Ethnic Affinity” – as targeting options when offering housing, employment, or credit opportunities. Facebook also required advertisers to self-certify that they are not discriminating [67].

On July 24, 2018, Facebook signed a legally binding agreement with the state of Washington to make changes to its advertising tools as a result of the 20-month investigation conducted by Washington Attorney General Bob Ferguson's office [10]. Facebook agreed to pay \$90,000 in costs and fees to the Attorney General's Office and removed targeting options that may allow advertisers to exclude on the basis of race, religion, sexual orientation, veteran status, and other protected classes for housing, employment, credit, and insurance ads [10]. On August 21, 2018, Facebook announced it will eliminate 5,000 targeting options related to ethnicity or religion, such as "Native American culture", "Passover", "Evangelicalism", and "Buddhism" from being used by advertisers [68, 69].

On March 19, 2019, Facebook settled with the ACLU, the Communications Workers of America, the National Fair Housing Alliance, and others on their multiple lawsuits against Facebook over advertising discrimination [70, 71]. Facebook agreed to create a separate portal for ads regarding housing, employment, and credit [72]. On this portal, Facebook will offer "a much more limited set of targeting options so that advertisers cannot target ads based on Facebook users' age, gender, race, or categories that are associated with membership in protected groups, or based on zip code or a geographic area that is less than a 15-mile radius, and cannot consider users' age, gender, or zip code when creating 'Lookalike' audiences for advertisers" [72].

On August 26, 2019, Facebook started implementing this agreement by requiring advertisers purchasing employment, housing, and credit ads to use the tools of "Special Ad Categories" instead of the regular Facebook advertising tools [2]. The "Special Ad Categories" tools restrict options that "allow targeting by age, gender, ZIP code, multicultural affinity, or any detailed options describing or appearing to relate to protected characteristics" and advertisers have to create a Special Ad Audience that finds similar users to their existing customers "in online behavior without considering things like age, gender or ZIP code" rather than a normal Lookalike Audience [2].

Before settling these lawsuits, Facebook has repeatedly argued that it is not liable for discrimination on its platform due to the actions of its advertisers, because it is protected by Section 230 of the Communications Decency Act [26, 68]. According to Facebook, since it is a platform rather than a publisher, then Section 230 means that it should not be liable for the content that it hosts [26]. In 2018, the U.S. Department of Justice has filed Statements of Interest in two different discrimination-related lawsuits against Facebook arguing that it does not believe Section 230 liability protections apply to Facebook’s advertising platform [73, 74].

Finally, in response to the July 2020 Stop Hate for Profit boycott organized by the NAACP, the Anti-Defamation League, Color of Change, and others, Facebook released its civil rights audit conducted by Laura Murphy, former Director of the ACLU Legislative Office, and attorneys at the law firm Relman Colfax on July 8, 2020 [18]. Similar to the motivations of the boycott’s organizers, the civil rights audit primarily focused on issues of misinformation and hate speech on Facebook and how it has “placed greater emphasis on free expression” instead of balancing that with the “value of non-discrimination”, which the report noted did not need to be mutually exclusive [18]. For example, the audit criticized Facebook for deciding that President Trump’s post stating “when the looting starts the shooting starts” – in regards to the summer’s Black Lives Matter protests – did not violate its content policies about the incitement of violence and thus left it up without any warning labels [18]. Regarding advertising discrimination, the audit acknowledged that HUD has filed charges against Facebook for violating fair housing laws with its ad targeting options [18]. It also acknowledges that research from Northeastern University and Upturn [5] has shown it’s possible for the Special Ad Audiences tool to be biased despite not using protected class information [18]. Regarding this audit and Facebook’s civil rights policies, Facebook’s Chief Operating Officer, Sheryl Sandberg, acknowledged that “it is the beginning of the journey, not the end” [19]. On August 11, 2020, Facebook announced that it will

finally retire its controversial multicultural affinity advertising targeting options for racial and ethnic groups [20].

The boycott by most major advertisers ended by August 2020, and Facebook does not appear to have suffered significant financial damage as a result of the boycott, though the boycott also coincided with the COVID-19 pandemic, which saw many small and medium businesses increase their web presence and online sales [75]. Besides the civil rights audit, Facebook has also responded to some of the demands of the boycott's organizers such as banning Holocaust denial and blackface posts [75] and agreeing to hire a civil rights vice president, though not a C-suite executive as demanded by the boycott's organizers [76].

At the end of 2020, Facebook has also begun to update some of its "race-blind" policies that were having disproportionate impact on marginalized groups. In December 2020, Facebook started re-engineering its automated moderation systems, which previously did not distinguish between groups who have historically been targets of hate speech versus groups who were not [77]. Thus, comments such as "White people are stupid" were treated the same way as anti-Semitic or racist slurs [77]. Black users complained that the old system removed posts such as "Thank a Black woman for saving our country", and civil rights experts argued that "you can't have the conversation if it is being filtered out, bizarrely, by overly blunt hate speech algorithms" [77]. Facebook is de-prioritizing its moderation of negative comments about "Whites", "men", and "Americans" as less likely to be harmful, while acknowledging that underrepresented groups need more protection [77].

This study examines changes in the discriminatory potential of Facebook's advertising platform from January 2020 to January 2021. It was a tumultuous year, which saw organized action by its users, advocacy groups, and its largest advertisers to apply pressure for greater civil rights protections on the platform. Facebook has also begun to publicly take some steps towards reform,

such as removing multicultural affinity targeting options and re-engineering its content moderation system.

Has Facebook gone far enough to prevent discrimination on its advertising platform?

Methods

Overview

This study is the first to examine the racial and ethnic breakdowns of Facebook's multicultural affinity groups and the similar sounding cultural interest groups. It is also the first to determine the racial and ethnic biases of Facebook's Lookalike Audience and Special Ad Audience tools across multiple time periods and whether the degree of bias is affected by using individuals with racially stereotypical names and ZIP codes in the source audience.

I conducted the study over 2 waves, with Wave 1 occurring in January 2020 and Wave 2 in January 2021. During each wave, I conducted two types of tests of Facebook's ad platform to answer my research questions about potential racial and ethnic biases.

- Test 1 – Studying the Racial and Ethnic Breakdowns of Targeting Options By Facebook
 - Detailed Targeting options
 - 2020 – Multicultural Affinity groups
 - African American (US)
 - Asian American (US)
 - Hispanic (US – All)
 - 2021 – Cultural Interest groups
 - African-American Culture
 - Asian American Culture
 - Hispanic American Culture

- Research question: Are the cultural interest groups as racially and ethnically homogenous as the old multicultural affinity groups?
- Test 2 – Studying the Bias in Facebook’s Lookalike Audience and Special Ad Audience Tools
 - Using A Set Theory Approach
 - Lookalike Audiences (2020 and 2021)
 - Special Ad Audiences (2020 and 2021)
 - Research questions:
 - Do Lookalike and Special Ad audiences reflect racial and ethnic biases depending on the lists of individuals used to generate them?
 - Is the degree of racial and ethnic bias in Lookalike and Special Ad audiences affected by well-established racial factors from the offline world such as the name or ZIP code of individuals used to create the audience?
 - What are the differences in the type and degree of bias observed in Facebook’s advertising tools in 2021 versus 2020?

In each wave, I used that month’s North Carolina voter list of ~8 million voters to test Facebook’s advertising tools. This study tests how Facebook’s ad platform relates to different samples of African-American, White, Asian, and Hispanic voters, which are racial and ethnic groups that exist in the voter data and also as Facebook targeting options. For each test, I first create lists of voters with known race and ethnic data. I then input those lists to different Facebook ad tools to see if the estimated reach according to Facebook changes in a biased way based on the combination of the demographics of the list that is used and the Facebook tool being tested.

Preparing the North Carolina Voter Data for Testing on Facebook

The North Carolina voter data is useful for this study, because it includes important personal information that Facebook can use to match a voter to a user profile, and it also contains self-reported race and ethnicity data that I use to create subsets of voters for each test.

The North Carolina voter dataset is publicly available and contains the following variables that overlap with what Facebook requests in order to match an uploaded list of voters with their corresponding Facebook profiles for ad targeting through the Custom Audience tool:

- First name
- Last name
- City
- State
- ZIP code
- Country
- Gender
- Age
- Year of birth
- Phone number

I filtered the voter data for “Active” and “Verified” voters to ensure that I am using voters with the most up-to-date city and ZIP code fields to match against Facebook profiles.

Voter registration forms in North Carolina ask voters to fill out their race and ethnicity which is captured in the state voter data (Figure 2.9). I then used the African American, Asian, and White race categories and the Hispanic and Not Hispanic ethnicity categories to create subsets of voters to test the degree of bias in Facebook’s advertising platform.

NORTH CAROLINA VOTER REGISTRATION APPLICATION (fields in red text are required)

2020.02

06w

<p>1 Indicate whether you are qualified to vote or preregister to vote based on U.S. citizenship and age.</p> <p style="text-align: center;">Are you a citizen of the United States of America? IF YOU CHECKED "NO" IN RESPONSE TO THIS CITIZENSHIP QUESTION, DO NOT SUBMIT THIS FORM. YOU ARE NOT QUALIFIED TO VOTE.</p> <p style="text-align: center;">Will you be at least 18 years of age on or before election day?</p> <p style="text-align: center;">Are you at least 16 years of age and understand that you must be 18 years of age on or before election day to vote? IF YOU CHECKED "NO" IN RESPONSE TO BOTH OF THESE AGE QUESTIONS, DO NOT SUBMIT THIS FORM. YOU ARE NOT QUALIFIED TO REGISTER OR PREREGISTER TO VOTE.</p>			
<p>2 Provide your full legal name.</p> <p>Last Name <input type="text"/> Suffix <input type="text"/></p> <p>First Name <input type="text"/></p> <p>Middle Name <input type="text"/></p>		<p>3 Provide your date of birth and identification information.</p> <p>Date of Birth (MM/DD/YYYY) <input type="text"/> / <input type="text"/> / <input type="text"/></p> <p>State or Country of Birth <input type="text"/></p> <p>NC Driver License or NC DMV ID Number <input type="text"/></p> <p>Last 4 Digits of Social Security Number <input type="text"/></p> <p><input type="checkbox"/> Check if you do not have a driver license or Social Security number. State Voter Registration Number (Optional: To locate, check "Voter Lookup" at www.NCSBE.gov.) <input type="text"/></p>	
<p>4 Provide your residential address - where you physically live. Do not enter a P.O. Box or a mail drop location.</p> <p>Address Number <input type="text"/> Street Name and Type <input type="text"/></p> <p>Address Line 2 (e.g., apartment, lot or unit number) <input type="text"/></p> <p>City <input type="text"/> State <input type="text"/> Zip Code <input type="text"/></p> <p>County <input type="text"/> Have you lived at this address for 30 or more days? <input type="checkbox"/> Yes <input type="checkbox"/> No If "No", date moved? <input type="text"/></p>		<p>5 Provide a mailing address.</p> <p>Do you receive mail at your residential <input type="checkbox"/> Yes <input type="checkbox"/> No. If "No", you are required to provide a mailing address.</p> <p>Mailing Address Line 1 <input type="text"/></p> <p>Mailing Address Line 2 <input type="text"/></p> <p>Mailing Address Line 3 <input type="text"/></p> <p>City <input type="text"/> State <input type="text"/> Zip Code <input type="text"/></p>	
<p>No Physical Address? If you do not have an address, use the space to the right to illustrate where you normally live or sleep. Write in the names of the nearest crossroads (or streets). Draw an X on the map to show where you live or usually sleep.</p> <p>IMPORTANT: You should also provide a valid mailing address above to permit the board of elections to send you a voter card.</p>			
<p>6 Provide your demographic information (optional).</p> <p>Gender <input type="checkbox"/> Male <input type="checkbox"/> Female</p> <p>Ethnicity <input type="checkbox"/> Not Hispanic/Latino <input type="checkbox"/> Hispanic/Latino</p> <p>Race <input type="checkbox"/> African American/Black <input type="checkbox"/> American Indian/Alaska Native <input type="checkbox"/> Asian <input type="checkbox"/> Multiracial <input type="checkbox"/> Native Hawaiian/Pacific Islander <input type="checkbox"/> White <input type="checkbox"/> Other</p>		<p>7 Provide your choice for political party affiliation.</p> <p><input type="checkbox"/> Constitution Party <input type="checkbox"/> Libertarian Party <input type="checkbox"/> Other</p> <p><input type="checkbox"/> Democratic Party <input type="checkbox"/> Republican Party</p> <p><input type="checkbox"/> Green Party <input type="checkbox"/> Unaffiliated</p> <p>If you select a party that is not recognized in North Carolina, you will be registered as <i>Unaffiliated</i>.</p>	
<p>8 Complete if you are currently registered to vote in another NC county or in another state. (This information will be used to cancel your previous voter registration in the other county or state.)</p> <p>First Name Used in Last Registration <input type="text"/> Middle Name Used in Last Registration <input type="text"/> Last Name Used in Last Registration <input type="text"/> Suffix <input type="text"/></p> <p>Address Where You Were Last Registered <input type="text"/> City/State/Zip Code of Last Registration <input type="text"/> County of Last Registration <input type="text"/></p>			
<p>9 Provide your contact information (optional). (This information is helpful if we need to contact you concerning your voter registration. Your contact information may be disclosed as a public record.)</p> <p>Area Code <input type="text"/> Phone Number <input type="text"/> Email Address <input type="text"/> Would you like to be contacted to be a poll worker? <input type="checkbox"/> Yes <input type="checkbox"/> No</p>			
<p>10 Sign below to attest to your qualifications to vote.</p> <p style="text-align: center;">FRAUDULENTLY OR FALSELY COMPLETING THIS FORM IS A CLASS I FELONY UNDER CHAPTER 163 OF THE NC GENERAL STATUTES.</p> <p>I attest, under penalty of perjury, that in addition to having read and understood the contents of this form, that: (1) I am a United States citizen, as indicated above; (2) I am at least 18 years of age, or will be by the date of the general election; or I am at least 16 years old and understand that I must be at least 18 years old on the day of the general election to vote; I shall have been a resident of North Carolina, this county, and precinct for 30 days before the date of the election in which I intend to vote; (3) I will not vote in any other county or state after submission of this form and if I am registered elsewhere, I am canceling that registration at this time; and (4) I am not currently serving a felony sentence, including any probation, post-release supervision, or parole OR I am serving an extended term of probation, post-release supervision, or parole, I have outstanding fines, fees, or restitution, and I do not know of another reason that my probation, post-release supervision, or parole was extended.</p> <p>X _____ Signature Required _____ Date</p>			

Figure 2.9. Example North Carolina Voter Registration Form in 2021. This form requests the voter to provide their demographic information in terms of ethnicity and race highlighted in the red box.

Since Test 2 involved seeing if the degree of bias in Lookalike and Special Ad audiences changed when using lists of individuals with commonly given names of one race or ethnicity for the source audience, I started with the ethnicolr library in Python to predict a voter's race or ethnicity based on their first and last name [78]. I then created subsets of voters of each demographic group who have commonly given names of that race or ethnicity according to ethnicolr predictions. Research has shown that there are trends in popular first and last names being given to babies of different races and ethnicities [37, 38, 51]. In fact, these trends are often used in political science and social science research to predict the race and ethnicity of individuals in a dataset that doesn't contain explicit race and ethnicity fields [38]. A common approach is to use the U.S. Census Bureau's Frequently Occurring Surnames dataset which contains all last names occurring more than 100 times in the Decennial Census and the racial and ethnic breakdown of each last name [79]. However, using only last names may incorrectly predict the race of some individuals, especially African-Americans, who have last names that are also frequently used by White families but may have more racially distinct first names. Thus, Sood and Laohaprapanon created the ethnicolr library which uses a Long Short Term Memory neural network trained on Florida's voter registration data to predict an individual's race or ethnicity using both their first and last name [78]. They found that their model using both names, which had a precision of 83% and a recall of 84%, performed better than a model using only the last name in out of sample testing. The ethnicolr library was released in 2018 and has since been used in multiple research studies in medicine, political science, economics, education, and other subjects [80, 81, 82, 83, 84]. Figure 2.10 shows the results from the ethnicolr library using 4 names. The model outputs probabilities that a given first and last name belongs to one of 4 classes: White, African-American, Asian, and Hispanic. It then predicts a label based on which class had the highest probability. The first two rows compare "Jinyan Zang" versus "Jinyan Zane", where the

change in the last name resulted in the predicted label changing from Asian to White. The last two rows compare “Latanya Sweeney” versus “Tanya Sweeney” where the change in the first name resulted in the predicted label changing from African-American to White.

First name	Last name	Predicted Label	Probabilities			
			White	African-American	Asian	Hispanic
Jinyan	Zang	Asian	19%	2%	77%	2%
Jinyan	Zane	White	49%	31%	14%	5%
Latanya	Sweeney	African-American	10%	88%	1%	1%
Tanya	Sweeney	White	89%	9%	1%	2%

Figure 2.10. Example Results from the ethnicolr Library in Python Using 4 Names. Changing from “Jinyan Zang” to “Jinyan Zane” resulted in the predicted label switching from Asian to White. Changing from “Latanya Sweeney” to “Tanya Sweeney” also resulted in the predicted label switching from African-American to White.

Since the Test 2 also involved seeing if the degree of bias in Lookalike and Special Ad audiences changed based on using voters living in racial or ethnic enclaves, I started with the U.S. Census Bureau’s American Community Survey to find the racial and ethnic breakdown of every ZIP code in North Carolina [85]. According to the Census, North Carolina is 71% White, 22% African-American, and 3% Asian, and it is also 10% Hispanic [86]. Therefore, I considered African-Americans living in ZIP codes with >50% African-Americans, Whites living in ZIP codes with >90% Whites, Asians living in ZIP codes with >20% Asians, and Hispanics living in ZIP codes with >20% Hispanics as having a racially stereotypical ZIP code.

With ethnicolr predictions based on names and Census data identifying racial enclave ZIP codes, I created 4 versions of 10K voter samples for each race / ethnicity to study the demographic biases of Lookalike and Special Ad audiences in Test 2 (Figure 2.11).

Race / Ethnicity	10K Voter Samples Used to Create Lookalike and Special Ad Audiences
African-American	African American
	African-American with Commonly Given African-American Names
	African-American In ZIP Codes with >50% African-Americans
	African-American with Commonly Given African-American Names & In ZIP Codes with >50% African-Americans
White	White
	White with Commonly Given White Names
	White In ZIP Codes with >90% Whites
	White with Commonly Given White Names & In ZIP Codes with >90% White
Asian	Asian
	Asian with Commonly Given Asian Names
	Asian In ZIP Codes with >20% Asians
	Asian with Commonly Given Asian Names & In ZIP Codes with >20% Asians
Hispanic	Hispanic
	Hispanic with Commonly Given Hispanic Names
	Hispanic In ZIP Codes with >20% Hispanics
	Hispanic with Commonly Given Hispanic Names & In ZIP Codes with >20% Hispanics

Figure 2.11. 10K Voter Samples with Different Traits Used to Create Lookalike and Special Ad Audiences for Each Race / Ethnicity for Test 2 Analysis.

Test 1 – Studying the Racial and Ethnic Breakdowns of Targeting Options By Facebook

I started by recording the target size for the relevant Facebook targeting options in 2020 and 2021 using its ad planning tool shown in Figure 2.12. For 2020, the options were “African American (US)”, “Asian American (US)”, and “Hispanic (US – All)”. In 2021, the options were “African-American

Culture”, “Asian American Culture”, and “Hispanic American Culture”. I then compared the target size of each Facebook option to the corresponding estimate from the US Census for the 18+ population of that demographic group [87].

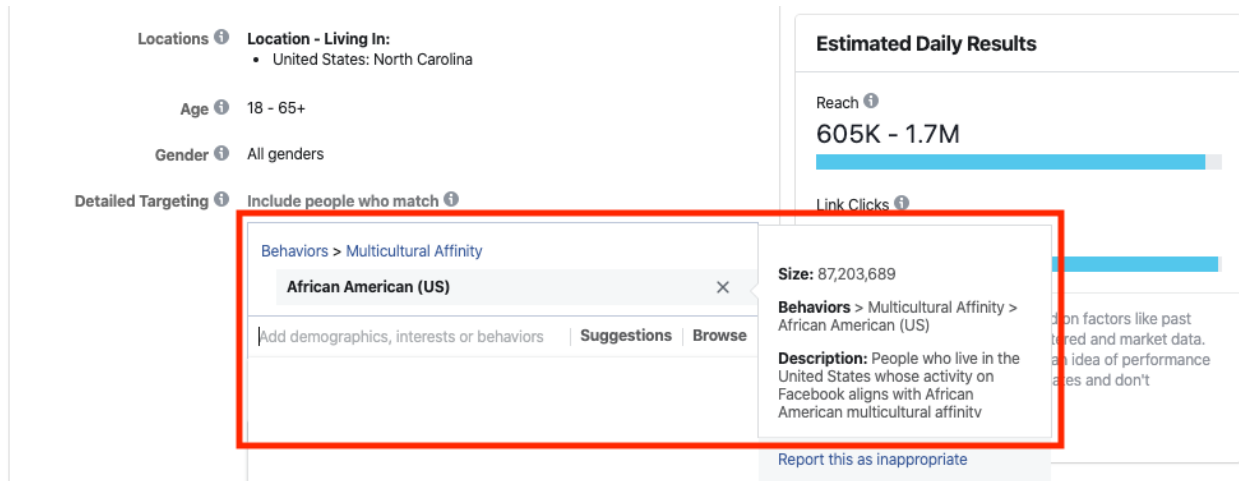


Figure 2.12. Target Size of the “African American (US)” Option in 2020. The size in 2020 was 87,203,689 Facebook users highlighted in the red box.

For each wave, I created different Custom audiences on Facebook by segmenting the North Carolina voter list by different racial groups – African-American, Asian, and White – and by different ethnic groups – Hispanic and Non-Hispanic. I then randomly sampled 10,000 voters from each of the segments to use as the basis to create a Lookalike audience of the 1% most similar users in the United States on Facebook.

I iterated and set each Custom or Lookalike audience as the target on Facebook’s ad planning tool (Figure 2.13). I then added the targeting option being tested under the “Detailed Targeting” setting of Facebook’s ad planning tool and recorded the updated daily reach estimate. In the example shown in Figure 2.14, an ad that targeted the African-American voters Custom audience and also matched the “African-American Culture” interest option would only reach 142,000 users daily, which

represents 37% of the reach of an ad that targeted the same African-American voters without the additional “African-American Culture” criteria (Figure 2.13).

The screenshot displays the Facebook Ad Planning Tool interface. At the top, there are 'Edit' and 'Review' buttons. The main area is divided into two columns. The left column is titled 'Audience' and contains several sections: 'Create New Audience' and 'Use Saved Audience'; 'Custom Audiences' with a 'Create New' button and a search bar containing '2_black.csv'; 'Locations' with 'United States: North Carolina'; 'Age' with '18 - 65+'; 'Gender' with 'All genders'; 'Detailed Targeting' with a search bar containing 'Add demographics, interests or behaviors'; 'Detailed Targeting Expansion' with an unchecked checkbox; 'Languages' with 'All languages'; and a 'Save This Audience' button. The right column is titled 'Audience Definition' and shows a gauge chart with 'Specific' and 'Broad' markers, and a note that 'Audience definition is unavailable'. Below this is the 'Estimated Daily Results' section, which is highlighted with a red box. It shows 'Reach' as '131K - 379K' and 'Link Clicks' as '1.5K - 4.4K'. A disclaimer at the bottom of the results section states: 'The accuracy of estimates is based on factors like past campaign data, the budget you entered, market data, targeting criteria and ad placements. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results. Were these estimates helpful?'.

Figure 2.13. Example of the Estimated Daily Reach of African-American Voters Custom Audience on Facebook’s Ad Planning Tool in 2021. The estimated daily reach was 379K highlighted in the red box.

[Edit](#) [Review](#)

Audience

Define who you want to see your ads. [Learn More](#)

[Create New Audience](#) Use Saved Audience ▾

Custom Audiences Create New ▾

Customer List

2_black.csv

Exclude

Locations

Location - Living In:

- United States: North Carolina

Age

18 - 65+

Gender

All genders

Detailed Targeting

Include people who match ⓘ

Interests > Additional Interests > African-American culture

African-American culture

 Suggestions Browse

Exclude Narrow Audience

Detailed Targeting Expansion ⓘ

Reach people beyond your detailed targeting selections when it's likely to improve performance.


Languages

All languages

[Show More Options ▾](#)

Save This Audience

Audience Definition



Audience definition is unavailable.

Potential Reach: Unavailable ⓘ

Estimated Daily Results

Reach ⓘ

49K - 142K

Link Clicks ⓘ

563 - 1.6K

The accuracy of estimates is based on factors like past campaign data, the budget you entered, market data, targeting criteria and ad placements. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

[Were these estimates helpful?](#)

Figure 2.14. Example of the Estimated Daily Reach of Targeting an African-American Voters Custom Audience Who Match the “African-American Culture” Interest Option on Facebook’s Ad Planning Tool in 2021. The estimated daily reach was 142K, which is 37% of the reach in Figure 2.13.

When the estimated reach is below 1 million, Facebook’s ad planning tool in general rounds to the nearest thousand by using “K”, so that may have introduced some small rounding errors into the results shown in this study. I also maximized the ad budget to \$1 million in most cases to ensure the maximum reach estimate is used.

Test 2 – Studying the Bias in Facebook’s Lookalike Audience and Special Ad Audience

Tools Using A Set Theory Approach

In order to study the racial and ethnic breakdowns of a Lookalike or Special Ad audience, I used the set theory approach described in Sapiezynski et al. [5].

For example, to study the African-American versus White bias of a Lookalike or Special Ad audience, I first created a 2 million voter sample made up of 1 million randomly sampled African-American voters and 1 million randomly sampled White voters. Then from the remaining voters in North Carolina not in the sample, I created lists of 10,000 randomly sampled voters of different racial- or ethnic-related traits to create corresponding Lookalike and Special Ad audiences of the 1% most similar Facebook users in the United States (Figure 2.11). Thus, to see if a Lookalike audience based on African-Americans was biased towards including more African-American than White voters, I first measured the estimated reach of the 1 million African-American voters Custom audience (Figure 2.15). I then measured the estimated reach of the 1 million African-American audience if it excluded the Lookalike Audience based on African-Americans (Figure 2.16). Finally, I repeated the process for the 1 million White voters Custom audience.

Audience

Define who you want to see your ads. [Learn More](#)

Create New Audience Use Saved Audience ▾

Custom Audiences Create New ▾

Customer List

2_black_1M.csv

🔍 Search existing audiences

Exclude

Locations

Location - Living In:

- United States: North Carolina

Age

18 - 65+

Gender

All genders


Detailed Targeting

All demographics, interests and behaviors

Detailed Targeting Expansion:

- Off

Audience Definition



Audience definition is unavailable.

Potential Reach: Unavailable ⓘ

Estimated Daily Results

Reach ⓘ

103K - 299K

Link Clicks ⓘ

1.2K - 3.6K

The accuracy of estimates is based on factors like past campaign data, the budget you entered, market data, targeting criteria and ad placements. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

[Were these estimates helpful?](#)

Figure 2.15. Example of the Estimated Daily Reach of Targeting a 1 Million African-American Voters Custom Audience in 2021. The estimated reach was 299K highlighted the red box.

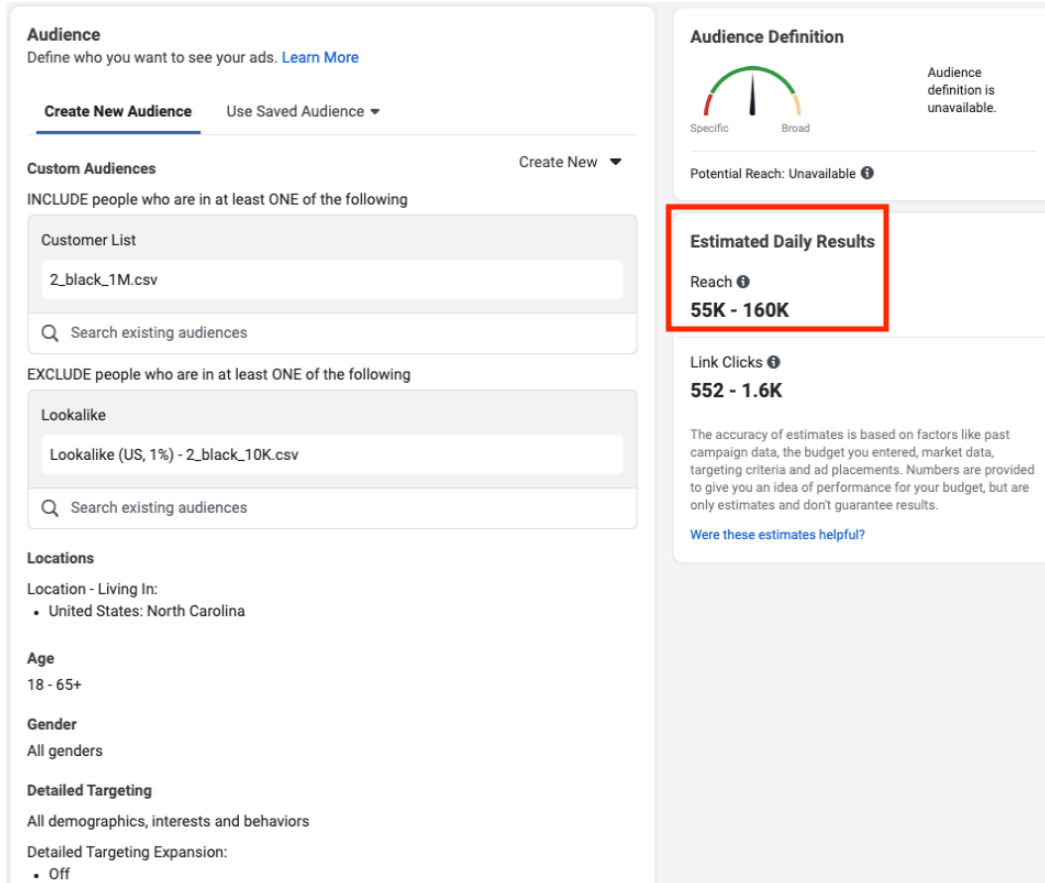


Figure 2.16. Example of the Estimated Daily Reach of Targeting a 1 Million African-American Voters Custom Audience While Excluding a Lookalike Audience Based on African-American Voters in 2021. The estimated reach was 160K which is 139K fewer users than Figure 2.15.

In this case, the estimated reach of the 1 million African-American voters decreased by 139,000 when excluding the Lookalike audience based on African-Americans, but the estimated reach of the 1 million White voters decreased by only 17,000 under the same circumstances. Thus, within the intersection of the Lookalike audience and the 2 million voter sample, 89% of the overlap were African-American voters while 11% were White voters as shown in Figure 2.17. This means that African-Americans were over-represented by being far above the 50% baseline in the 2 million voter sample. In this paper, I refer to statistics like the 89% as the sample share of African-American voters

in the Lookalike audience and statistics like the 11% as the sample share of White voters in the Lookalike audience. I also describe a demographic group being above the expected baseline level as “over-represented” and being under the baseline as “under-represented”. If the set theory approach found that a Lookalike or Special Ad audience over-represented a demographic group, then I describe that audience in this paper as being “biased” towards that race or ethnicity.

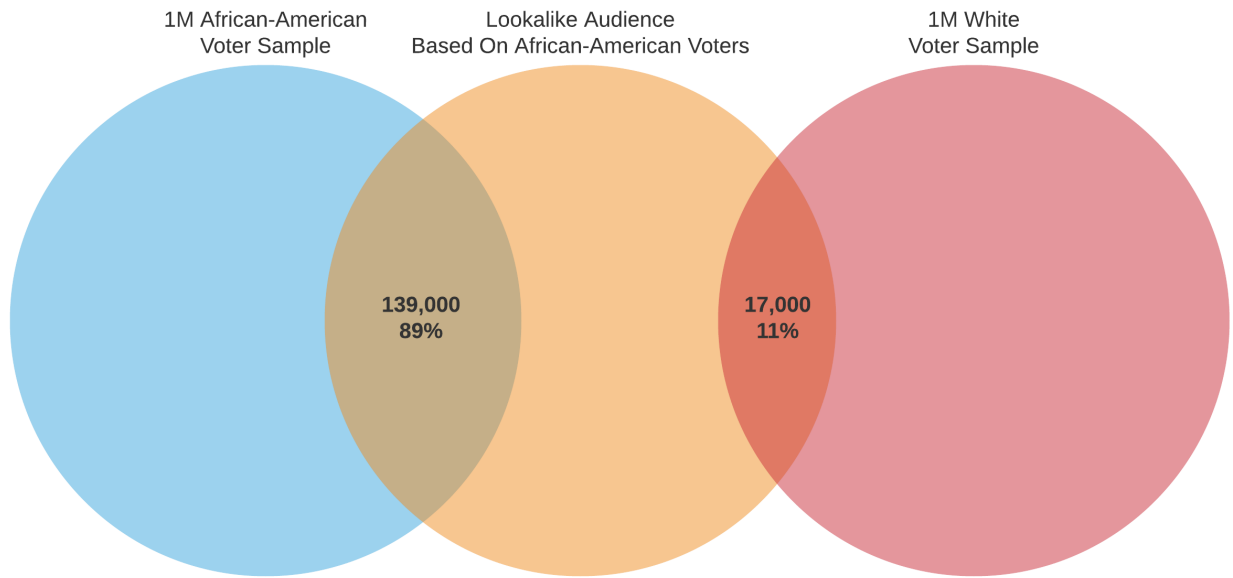


Figure 2.17. Example Breakdown of the Intersection of the Lookalike Audience Based on African-American Voters with the 1 Million African-American Voter Sample and the 1 Million White Voter Sample in 2021. Because 89% or 139,000 of the intersection were African-American voters versus just 11% being White voters, the Lookalike audience over-represented African-American voters and appears to be biased towards African-Americans.

Since there are far fewer Asian voters in North Carolina compared to African-Americans and Whites, Figure 2.18 shows that the set theory approach used a 150,000 voter sample as the comparison set with 50,000 Asian, African-American, and White voter samples each to study the degree of bias in a Lookalike or Special Ad audience towards Asians. Similarly, Figure 2.19 shows that I

used a 200,000 voter sample as the comparison set with 100,000 Hispanic and Non-Hispanic voter samples to study the degree of bias in a Lookalike or Special Ad audience towards Hispanics. Finally, since Facebook doesn't allow Special Ad audiences to be in the "Exclusion" position, I simply flipped the settings described earlier by first measuring the reach of the Special Ad audience on its own and then measuring the reach of the Special Ad audience while excluding a given voter sample, in order to implement the set theory approach for measuring biases of Special Ad audiences.

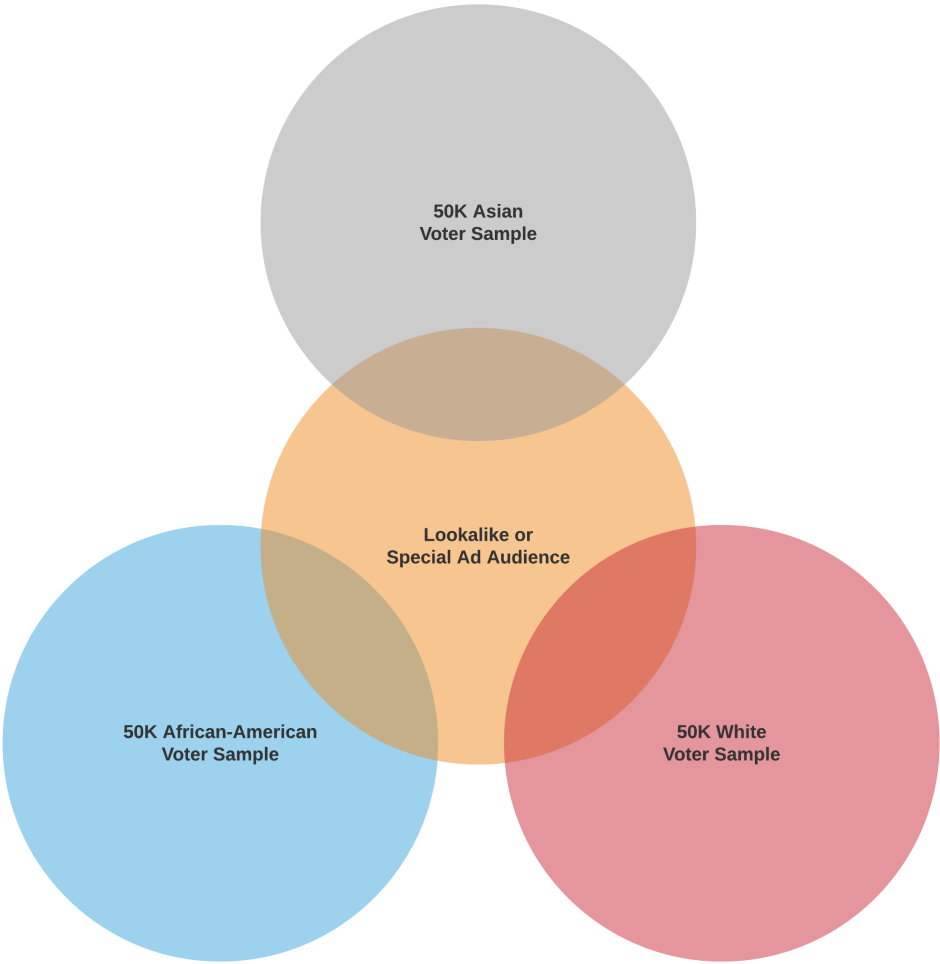


Figure 2.18. Set Theory Approach to Study the Bias Towards Asians in Lookalike or Special Ad Audiences Using a 150,000 Sample of Asian, African-American, and White Voters.

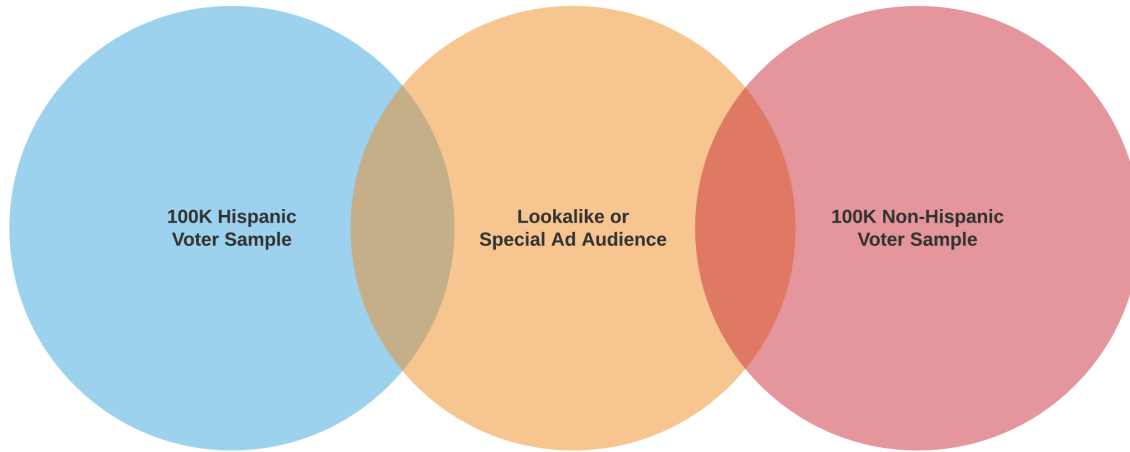


Figure 2.19. Set Theory Approach to Study the Bias Towards Hispanics in Lookalike or Special Ad Audiences Using a 200,000 Sample of Hispanic and Non-Hispanic Voters.

Results

Test 1 – Studying the Racial and Ethnic Breakdowns of Targeting Options By Facebook

When compared to the relevant Census population estimates, Facebook’s ad targeting sizes became much closer to the Census estimates by 2021 for Asians and Hispanics. In addition, in 2021, Facebook’s “African-American Culture” ad targeting option contained 75% fewer White users than the old “African American (US)” option they removed in the previous year, while the number of non-Asians and Non-Hispanics increased significantly for the Asian and Hispanic related targeting options.

In 2020, Facebook’s multicultural affinity group targeting option for “African-American (US)” had 2.49 times more people than the Census population estimate, while “Asian American (US)” was 0.28 times the Census estimate and “Hispanic (US – All)” was 0.50 times the Census estimate (Figure 2.20). In 2021, the number of Facebook users interested in “African-American Culture” was 2.26 times the Census estimate, while the target size of “Asian American Culture” was 0.96 times the Census estimate and that of “Hispanic American Culture” was 1.31 times the Census estimate (Figure 2.20).

Thus, the target size of Facebook’s Asian American Culture and Hispanic American Culture option grew closer to the Census population estimate in 2021 while the target size of the African-American Culture option stayed at twice the population estimate which was similar to the “African American (US)” option in 2020.

	Facebook Advertising Targeting Option	Facebook Target Size	Census (Age 18+ Population)	Facebook-Census Ratio
2020	African American (US)	87,203,689	35,079,870	2.49x
	Asian American (US)	4,972,438	17,502,608	0.28x
	Hispanic (US – All)	21,542,628	43,089,980	0.50x
2021	African-American Culture	79,388,010	35,079,870	2.26x
	Asian American Culture	16,807,470	17,502,608	0.96x
	Hispanic American Culture	56,515,880	43,089,980	1.31x

Figure 2.20. Facebook Advertising Target Size to Census Population Estimate Ratio. The Census Population Estimates use the US Census’ 2019 Population Estimates [87] for each race or ethnicity for individuals 18 and older since that was the age limit for the Facebook Detailed Targeting. Facebook’s ad targeting sizes became much closer to the Census estimates by 2021 for Asians and Hispanics.

In 2020, the share of NC voters on Facebook that could be reached by targeting “African American (US)” was 43% of African-American voters, 23% of Asian voters, and 39% of White voters (Figure 2.21). However, since there are far more White voters overall than African-American voters in North Carolina, this means that the “African-American (US)” targeting option could reach approximately 150,000 African-American voters and 428,000 White voters (Figure 2.21). In 2021, the conditional probability of being interested in “African-American Culture” was significantly higher for African-American voters at 37% versus only 8% for Asian and White voters (Figure 2.21). This resulted in 142,000 African-American voters, 109,000 White voters, and only 2,000 Asian voters being interested

in “African-American Culture” (Figure 2.21). Thus, 75% fewer Whites were interested in “African-American Culture” in 2021 compared to “African American (US)” in 2020.

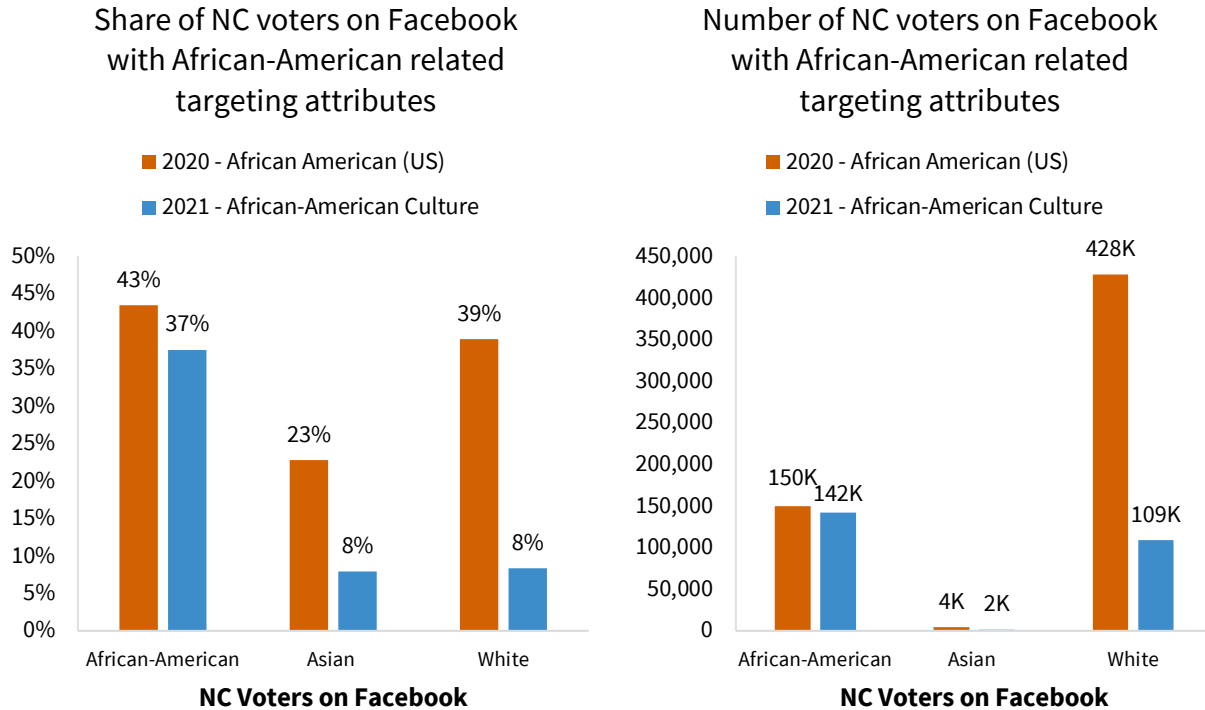


Figure 2.21. Share and Number of NC Voters on Facebook with African-American Related Targeting Attributes. 75% fewer Whites were interested in African-American Culture in 2021 compared to “African American (US)” in 2020.

In 2020, the share of Asian voters on Facebook reached by the “Asian American (US)” option was 8.9% which dwarfed the 0% of African-American voters and the 0.1% of White voters (Figure 2.22). However, in terms of absolute numbers, this means that approximately 1,600 Asian and White voters could be reached with the same multicultural affinity option (Figure 2.22). In 2021, the “Asian American Culture” target option could reach more non-Asians with a 7.4% share of Asian voters but also 2.5% share of African-American voters and 1.4% share of White voters (Figure 2.22). This means in absolute numbers that targeting “Asian American Culture” would reach only 1,400 Asian voters which

is far less than the corresponding 9,400 African-American voters and 18,000 White voters who share the same interest (Figure 2.22).

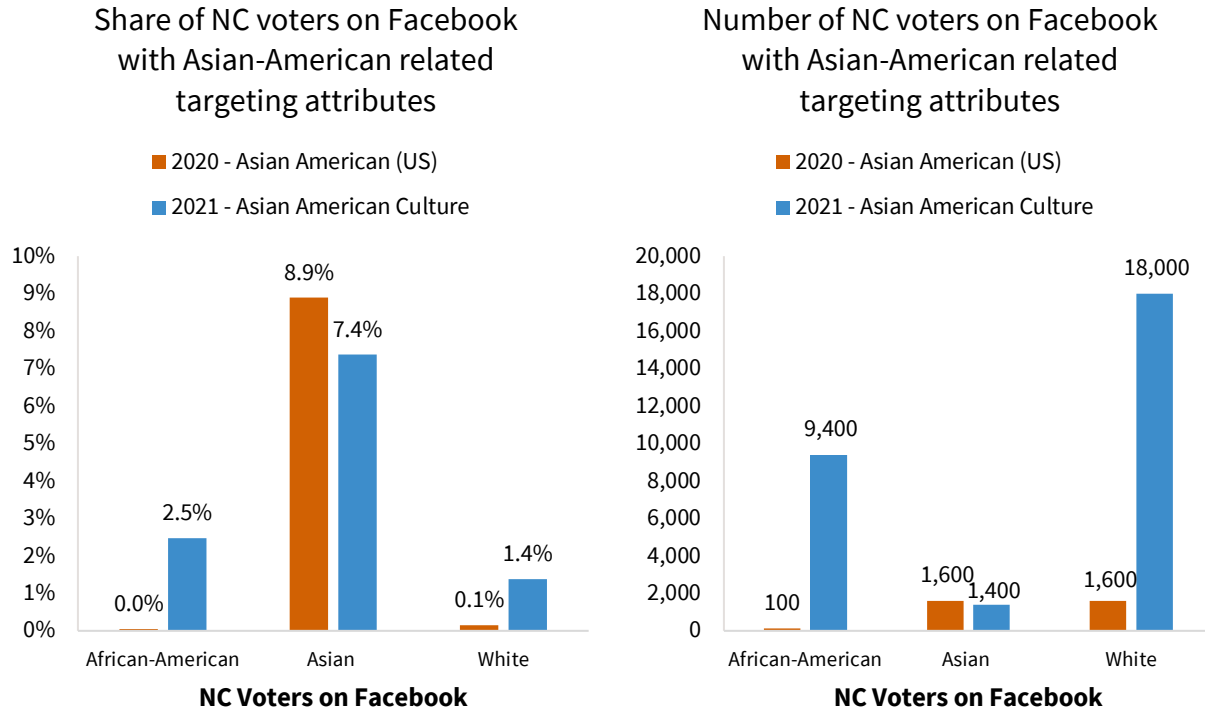


Figure 2.22. Share and Number of NC Voters on Facebook with Asian-American Related Targeting Attributes. In 2021, the “Asian American Culture” target option reached far more non-Asians than Asians, with an estimated reach of 1,400 Asian voters but 9,400 African-American voters and 18,000 White voters.

In 2020, Hispanic voters were more likely than Non-Hispanic voters to be reached by Facebook’s “Hispanic (US -All)” targeting option in both relative and absolute terms. 27.3% of Hispanic voters on Facebook had the “Hispanic (US - All)” attribute compared to only 0.5% of Non-Hispanic voters, which represented approximately 15,000 Hispanic voters and 6,400 Non-Hispanic voters (Figure 2.23). In 2021, only 9.8% of Hispanic voters were classified as interested in “Hispanic American Culture”, and 1.1% of Non-Hispanic voters had the same interest (Figure 2.23). This results

in only 5,800 Hispanic voters but 16,000 Non-Hispanic voters being reached by the targeting option, since there are far more Non-Hispanic than Hispanic voters on Facebook (Figure 2.23).

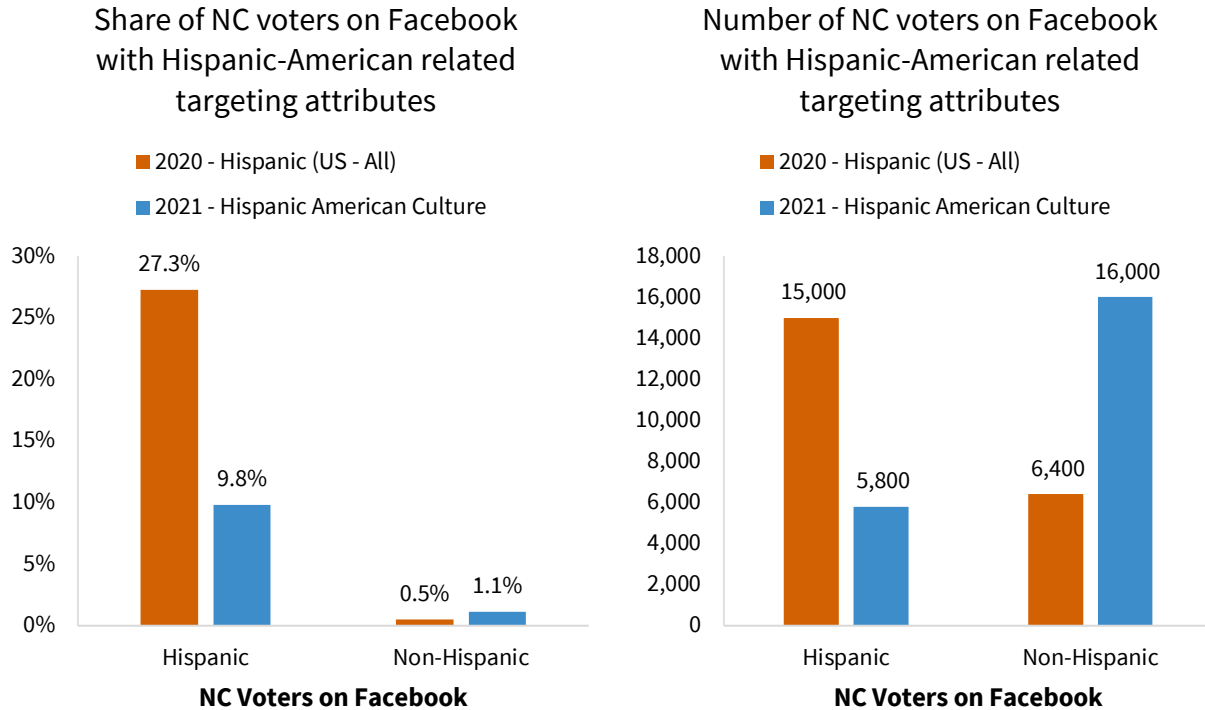


Figure 2.23. Share and Number of NC Voters on Facebook with Hispanic-American Related Targeting Attributes. In 2020, more than twice the number of Hispanic voters than Non-Hispanic voters are targeted by “Hispanic (US – All)”, but that ratio flips in 2021 with more than twice the number of Non-Hispanics than Hispanics reached by the “Hispanic American Culture” targeting option.

I also found that in 2021, Lookalike audiences based on African-American, Asian, White, or Hispanic voters tended to have similar shares interested in “African-American Culture” or “Hispanic American Culture” as the shares of the corresponding voter lists themselves. However, all Lookalike audiences, regardless of which voter list was used to create them, had about 2-3% of their users being

interested in “Asian American Culture”, including the Lookalike audience based on Asian voters. For more details about these results, see Appendix A.

Test 2 - Studying the Bias in Facebook’s Lookalike Audience and Special Ad Audience Tools

Using A Set Theory Approach

In both 2020 and 2021, Facebook’s Lookalike and Special Ad audiences were biased in over-representing the sample share of African-American or White voters depending on which race was dominant in the customer list used as the source audience. This bias increased when using customer lists with stereotypically African-American or White names or ZIP codes. Similar biases were observed for Lookalike audiences based on Asians or Hispanics, with a Lookalike audience in one case having a 100% sample share of Asian voters when using a customer list of Asians with stereotypically Asian names and ZIP codes. In a shift from 2020, I found that Special Ad audiences based on Asians did not significantly over-represent Asian voters in 2021.

When studying the breakdown of the 2 million NC voter sample intersecting with Lookalike audiences, 83% of the overlap between the sample and a Lookalike audience based on African-Americans were African-American voters in 2020 and that increased to 89% in 2021 (Figure 2.24). As the customer list used to create the Lookalike audience takes on more stereotypically African-American traits by having commonly given African-American names, living in a ZIP code with >50% African-Americans, or both, the sample shares of African-American voters in the Lookalike audiences increased up to 93% in 2020 and up to 94% in 2021 (Figure 2.24).

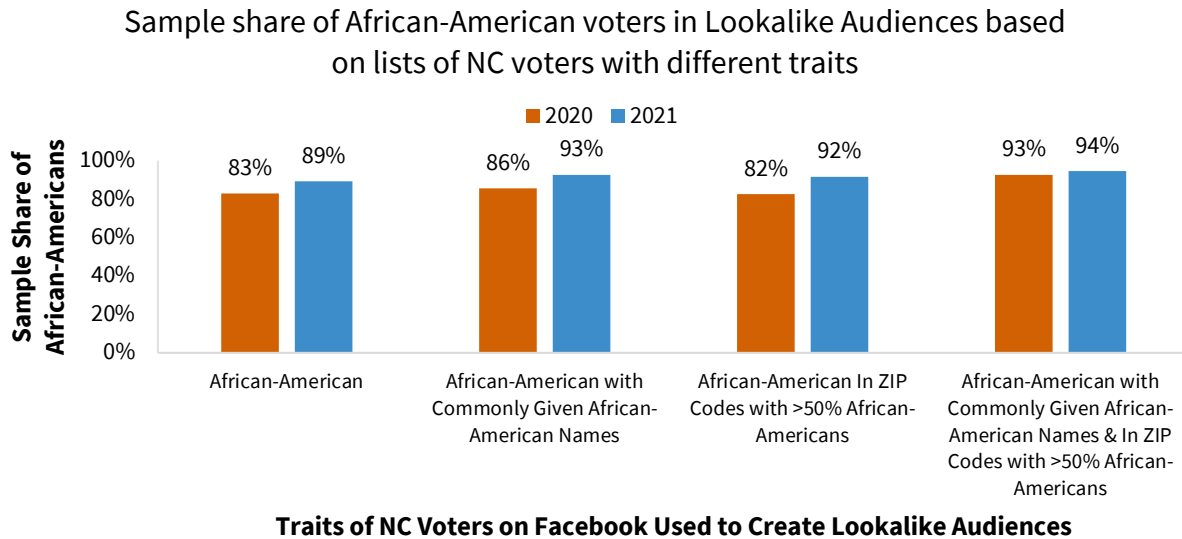


Figure 2.24. Sample Share of African-American Voters in Lookalike Audiences Based on Lists of NC Voters with Different Traits. As the customer list used to create the Lookalike audience contain more stereotypically African-American traits in terms of names and ZIP codes, the sample shares of African-American voters in the corresponding Lookalike audiences became more biased, increasing up to 93% in 2020 and 94% in 2021.

Similarly, I also found that Lookalike audiences based on Whites contained a large majority of White voters in their overlap with the 2 million NC voter sample in both waves. In 2020, 73% of the voters shared between the 2 million voter sample and the Lookalike audience based on White voters were also White, and in 2021 the rate was similar at 71% (Figure 2.25). As the White voters used to create the Lookalike audience took on additional White-affiliated traits with their name, living in a >90% White ZIP code, or both, the sample shares of Whites peaked at 87% in 2020 and 84% in 2021 for the resulting Lookalike audiences (Figure 2.25).

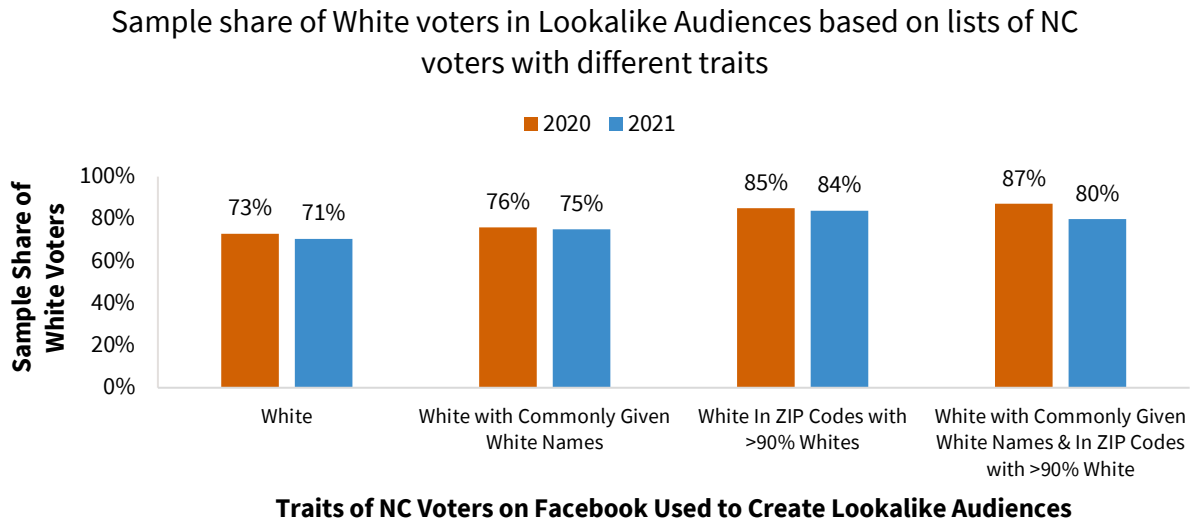


Figure 2.25. Sample Share of White Voters in Lookalike Audiences Based on Lists of NC Voters with Different Traits. Lookalike audiences based on Whites over-represented White voters in both waves, with a sample share of 73% White voters in 2020 and 71% in 2021.

In 2020, Lookalike audiences based on Asian voters appeared to have a moderate tendency to favor Asians, which became a significantly stronger trend in 2021. In 2020, when intersecting the 150,000 voter sample with Lookalike audiences based on Asian voters, 51% of the voters that overlapped were Asian, which steadily increased to 83% for the Lookalike audience based on Asian voters with commonly given Asian names and lived in Asian enclave ZIP codes (Figure 2.26). In 2021, those rates started at 68% for the Lookalike audience based on Asians and increased to 100% for the Lookalike audience based on Asians with stereotypically Asian traits by name and ZIP code (Figure 2.26). In all cases, the sample shares of Asian voters were far above the 33% baseline in the 150,000 voter sample. In addition, it's possible that the actual sample share of Asian voters in the most extreme case was slightly below 100%. A few Black and White voters could have been part of the intersection between the Lookalike audience and the voter sample, but those voters were not

captured due to the way Facebook rounds the advertising reach estimate on its ad planning tool to the nearest 1,000.

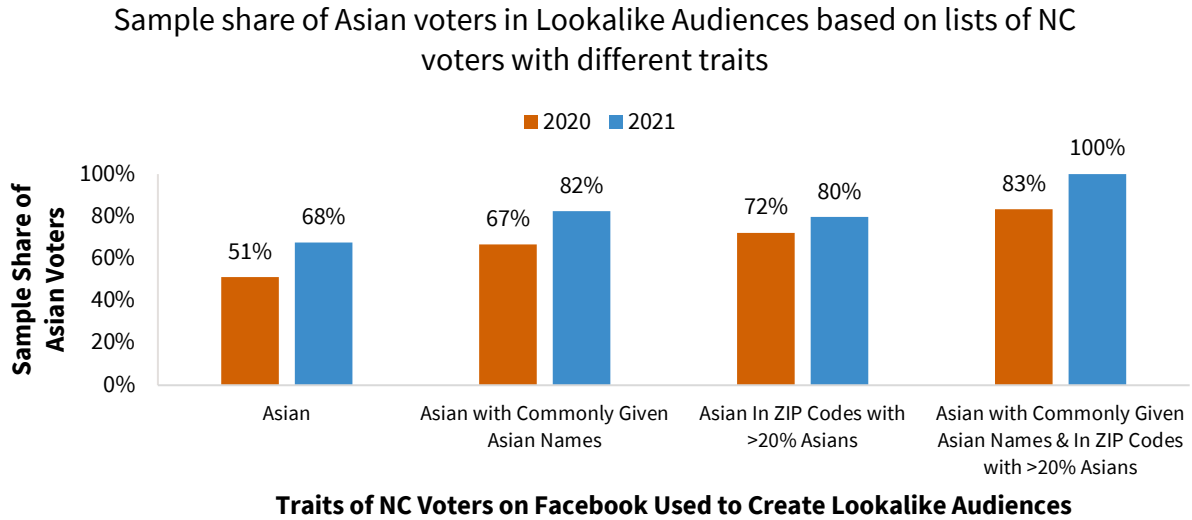


Figure 2.26. Sample Share of Asian Voters in Lookalike Audiences Based on Lists of NC Voters with Different Traits. In 2021, Lookalike audiences based on Asians were significantly biased towards including Asian voters, reaching up to 100% sample share of Asian voters in one case.

In both 2020 and 2021, Lookalike audiences based on Hispanics had a bias for including more Hispanic voters than Non-Hispanic voters. In 2020, Hispanic voters accounted for 71% of the overlap between the 200,000 voter sample and the Lookalike audience based on Hispanics, which was similar to the 69% share seen in 2021 (Figure 2.27). As the customer list used to create the Lookalike audience appeared more stereotypically Hispanic by name or ZIP code, the sample share of Hispanic voters only increased slightly to 75% in 2020, and 79% in 2021 (Figure 2.27).

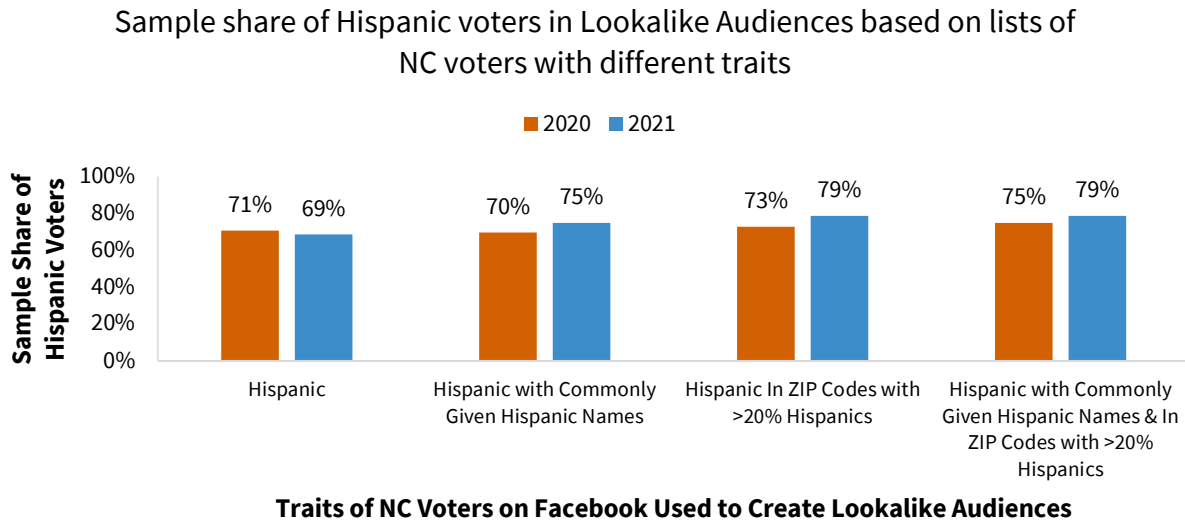


Figure 2.27. Sample Share of Hispanic Voters in Lookalike Audiences Based on Lists of NC Voters with Different Traits. In both 2020 and 2021, Lookalike audiences based on Hispanics had a bias for including more Hispanic voters than Non-Hispanic voters, with a sample share of 71% Hispanic voters in 2020 and 69% in 2021.

Even though Facebook created Special Ad audiences as an anti-discrimination tool for housing, employment, and credit-related ads, I found that Special Ad audiences could still be biased towards including more African-American or White voters depending on the demographics of the source audience.

In 2020 and 2021, the Special Ad audiences based on African-Americans demonstrated significant biases towards including more African-American voters relative to Whites. In 2020, the sample shares of African-American voters in the relevant Special Ad audiences started at 83% and went up to 97% for the Special Ad audience based on African-Americans with stereotypically African-American names and ZIP codes (Figure 2.28). These sample shares are similar to those observed for the Lookalike audiences based on African-Americans shown in Figure 2.24. In 2021, slightly fewer

African-Americans were in the intersection of the 2 million voter sample with the relevant Special Ad audiences, starting at a 76% sample share and increasing up to 89% for the Special Ad audience based on African-Americans with commonly given African-American names and living in ZIP codes with >50% African-Americans (Figure 2.28).

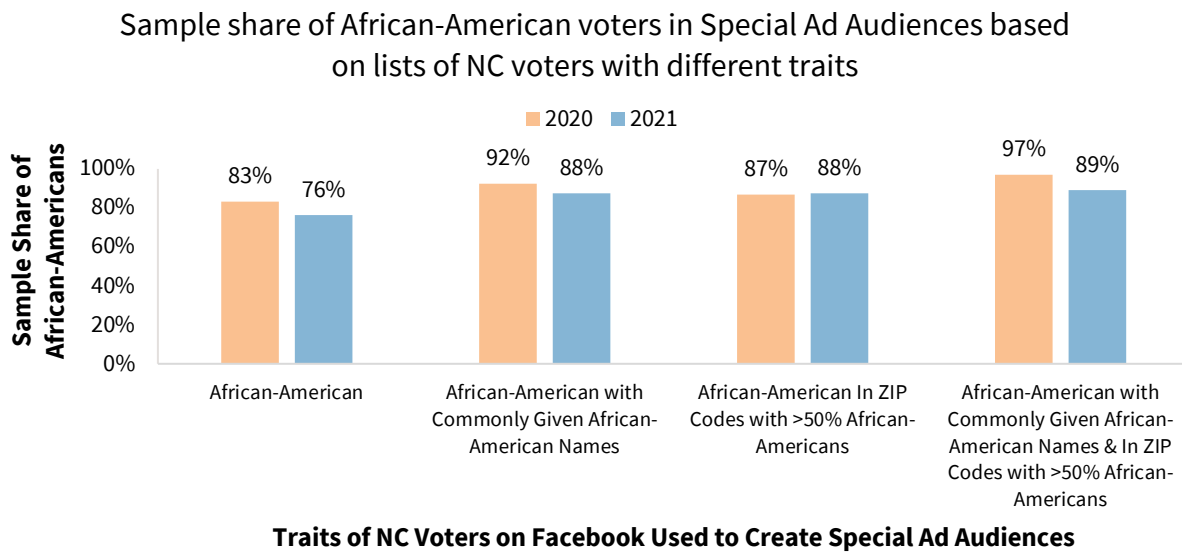


Figure 2.28. Sample Share of African-American Voters in Special Ad Audiences Based on Lists of NC Voters with Different Traits. In both waves, the Special Ad audiences based on African-Americans demonstrated significant biases towards including more African-American voters relative to Whites, with a sample share of 83% African-American voters in 2020 and 76% in 2021.

Special Ad audiences based on Whites also exhibited strong biases towards including more White voters than African-Americans at sample shares similar to the Lookalike audiences shown in Figure 2.25. In 2020, White voters were 83% of the overlap between the 2 million voter sample and the Special Ad audience based on Whites, and in 2021 the sample share was 81% (Figure 2.29). In 2020, the Special Ad audience based on Whites with commonly given White names had the largest bias with a 90% sample share of White voters (Figure 2.29). In 2021, the Special Ad audience based on Whites

living in >90% White ZIP codes had the largest bias with a sample share of 91% of White voters (Figure 2.29).

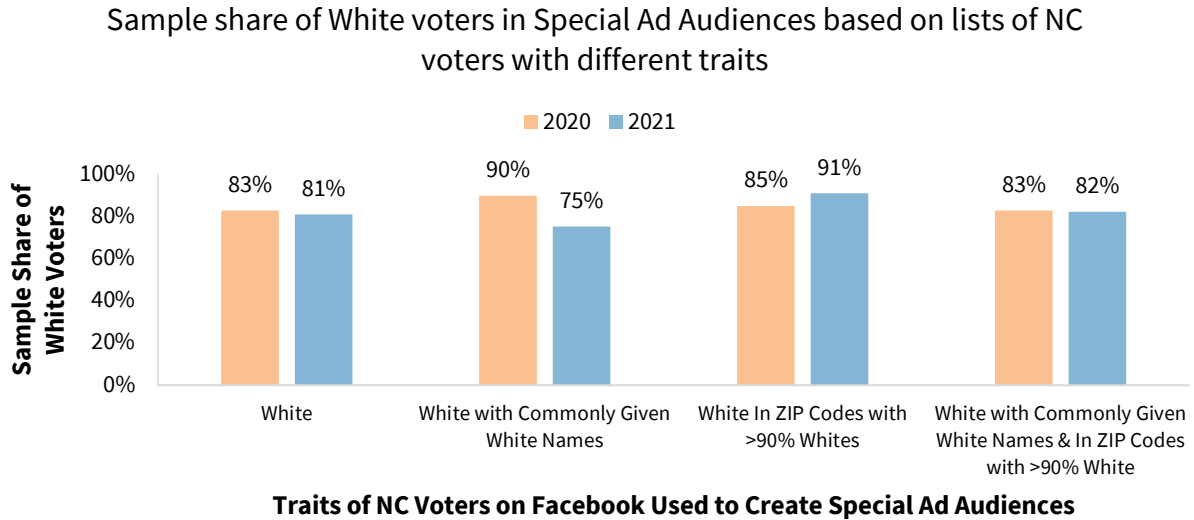


Figure 2.29. Sample Share of White Voters in Special Ad Audiences Based on Lists of NC Voters with Different Traits. In 2020, White voters were over-represented at 83% of the overlap between the 2 million voter sample and the Special Ad audience based on Whites, and same was true for 2021, with the sample share of White voters being 81%.

Interestingly, while in 2020, Special Ad audiences based on Asian voters demonstrated a bias towards Asians, that bias was significantly reduced in 2021. For example, in 2020, the sample share of Asian voters in Special Ad audiences based on different types of Asian voters started at 44% and increased up to 67% when using Asians with stereotypically Asian names and ZIP codes to create the Special Ad audience (Figure 2.30). In 2021, those sample shares go from 36% to 44% for the corresponding Special Ad audiences (Figure 2.30). When considering that the 150,000 voter sample used for each test has a baseline of 33% or 50,000 Asian voters, this means that Asian voters were only slightly over-represented in the intersection of the voter sample and the relevant Special Ad

audiences in 2021. This contrasts significantly with the very strong bias towards Asians, up to 100% sample share in one case, observed in Lookalike audiences based on similar customer lists of Asians shown in Figure 2.26.

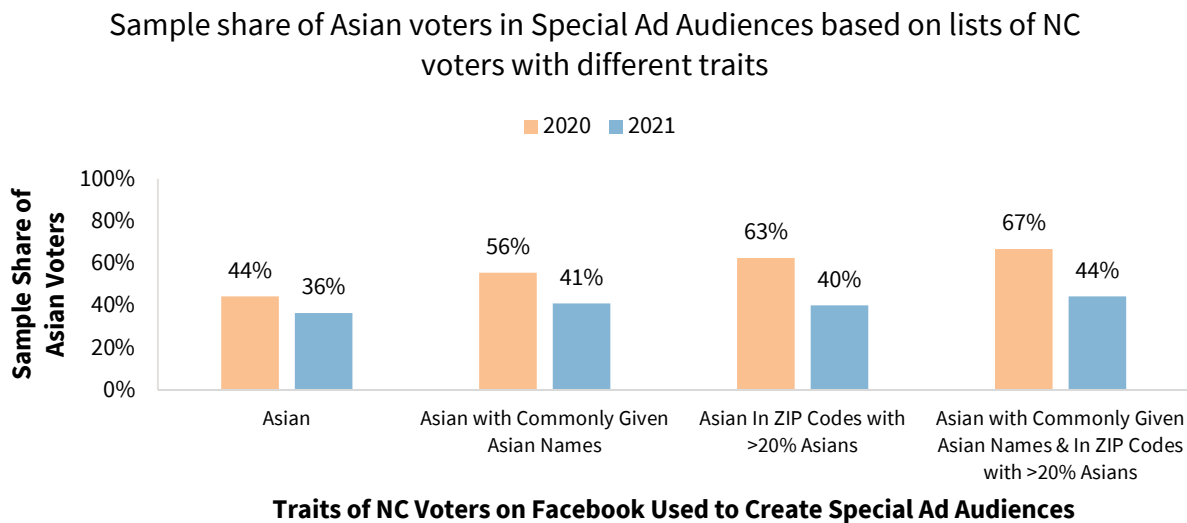


Figure 2.30. Sample Share of Asian Voters in Special Ad Audiences Based on Lists of NC Voters with Different Traits. Interestingly, while in 2020, Special Ad audiences based on Asian voters demonstrated a bias towards Asians, that bias was significantly reduced in 2021, with sample shares of Asian voters within 8 percentage points of the expected baseline of 33% for all four tests of Special Ad audiences.

I found that Special Ad audiences based on different types of Hispanics were not consistently biased towards including more Hispanic than Non-Hispanic voters in each wave. In 2020, only 3 of out of the 4 Special Ad audiences based on Hispanics had a sample share of Hispanic voters more than 10 percentage points above the 50% baseline in the 200,000 voter sample (Figure 2.31). In 2021, it was also 3 out of 4 Special Ad audiences, though a slightly different combination of audiences (Figure 2.31). For example, the Special Ad audience based on Hispanics with commonly given Hispanic names

had a relatively low 47% sample share of Hispanic voters in 2020, while a Special Ad audience based on similar individuals in 2021 had a much higher 82% sample share of Hispanic voters (Figure 2.31). These patterns contrast with the consistent over-representation of Hispanic voters in Lookalike audiences based on Hispanics shown in Figure 2.27.

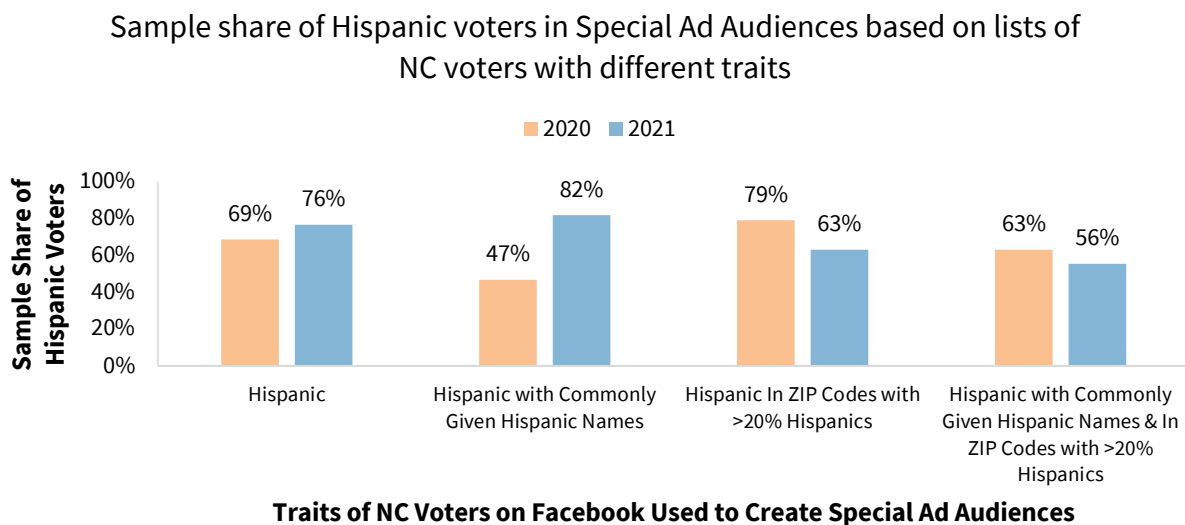


Figure 2.31. Sample Share of Hispanic Voters in Special Ad Audiences Based on Lists of NC Voters with Different Traits. Special Ad audiences based on different types of Hispanics were not consistently biased towards including more Hispanic than Non-Hispanic voters in 2020 and 2021.

Discussion

This study shows that despite the advertising boycott and Facebook’s response to address civil rights issues in 2020, there are multiple ways to use the tools of Facebook’s advertising platform such as its targeting options, Lookalike Audiences, and Special Ad Audiences to discriminate by race and ethnicity in 2021.

Key Findings

Based on the results of this study, I had the following key findings:

- In 2021, Facebook’s “African-American Culture” ad targeting option contained 75% fewer White users than the old “African American (US)” option they removed in the previous year.** On the other hand, targeting options related to Asians and Hispanics included more non-Asians and Non-Hispanics by 2021. In 2020, 43% of African-American voters and 39% of White voters can be reached by the “African American (US)” targeting option and. In 2021, those rates were 37% and 8% respectively for the “African-American Culture” targeting option (Figure 21). In 2021, the number of White voters reached by the “Asian American Culture” targeting option increased by 16,400 compared to the White voters reached by the “Asian American (US)” option in 2020, while the number of Asian voters reached by the two options declined by 200 over the same period (Figure 22). Similarly, the number of Hispanics reached by the “Hispanic American Culture” targeting option in 2021 was 9,200 fewer than the reach of the “Hispanic (US – All)” option in 2020, while the number of Non-Hispanics increased by 9,600 over the same year (Figure 23).
- Facebook’s tools to help advertisers find similar users to their existing customers exhibited bias towards including more African-Americans or Whites depending on which racial group was dominant in an advertiser’s customer list, and this was true for the Lookalike Audience tool, as well as the Special Ad Audience tool that Facebook designed to explicitly not use sensitive demographic attributes when finding similar users.** In 2020, the sample share of African-American voters in Lookalike audiences based on African-Americans was 83%, which increased to 89% in 2021 (Figure 2.24). For Special Ad audiences based on the same list of African-Americans, in 2020, the sample share of African-American voters was 83%, and in 2021, it was 76% (Figure 2.28). Lookalike audiences based on Whites saw a slightly smaller degree of bias with a 73% sample share of White voters in 2020 and a

71% sample share in 2021 (Figure 2.25). Special Ad audiences based on White voters had a 83% sample share of White voters in 2020 and 81% in 2021 (Figure 2.29).

- **The degree of bias towards including more African-Americans or Whites in a Lookalike or Special Ad audience was larger when using customer lists of individuals with racially stereotypical names or ZIP codes as the basis for each tool.** Lookalike audiences using a list of African-Americans with commonly given African-American names or living in ZIP codes with >50% African-Americans had sample shares of African-American voters up to 93% in 2020 and up to 94% in 2021 (Figure 2.24). Special Ad audiences based on similar customer lists had up to 97% sample shares of African-American voters in 2020 and 89% in 2021 (Figure 2.28). Lookalike audiences based on Whites with stereotypically White names or ZIP codes had sample shares of White voters up to 87% in 2020 and 84% in 2021 (Figure 2.25).
- **Similarly, Lookalike audiences can also become biased towards Asians, reaching up to 100% Asian in one case when using a customer list of Asians with stereotypical names and ZIP codes, and Lookalike audiences based on Hispanics over-represented Hispanics versus Non-Hispanics. Finally, in a shift from 2020, Special Ad audiences based on Asians did not appear to over-represent them in 2021.** Figure 2.26 shows that the sample share of Asian voters in a Lookalike audience based on Asians is 51% in 2020 and 68% in 2021, compared to a baseline of 33% Asian in the 150,000 voter sample. The sample share of Asian voters increased up to 83% in 2020 and 100% in 2021 for the Lookalike audience based on Asians with commonly given Asian names and living in ZIP codes with >20% Asians. Special Ad audiences based on Asians in 2020 had a 44% sample share of Asian voters, which increased up to 67% for audiences based on Asians with stereotypically Asian names and ZIP codes (Figure 2.30). On the other hand, in 2021, all of the Special Ad audiences based on Asians had

sample shares of Asian voters within 8 percentage points of the expected baseline of 33%.

Finally, in 2020, the sample shares of Hispanic voters in Lookalike audiences based on different types of Hispanics ranged from 70-75% and in 2021 ranged from 69% - 79%, which are above the 50% baseline share of Hispanics in the 200,000 voter sample (Figure 2.27).

I found that removing “African-American (US)”, “Asian American (US)”, and “Hispanic (US – All)” as targeting options in August 2020 did not mean other similar sounding cultural interest groups, which still existed, could not be used for racial and ethnic targeting in 2021. In fact, 75% fewer Whites were targeted by the “African-American Culture” option in 2021 when compared to the “African-American (US)” option in 2020. I also found that Lookalike and Special Ad audiences can become biased to include more African-Americans or Whites based on which race is more dominant within the customer list used as the source audience, especially if the customers also have racially stereotypical names and ZIP codes. The sample shares of African-American voters reached up to 93%-94% for some Lookalike audiences in 2020 and 2021. Lookalike audiences based on Asians or Hispanics were also biased towards over-representing the corresponding demographic group in both 2020 and 2021, with the degree of bias reaching up to a 100% sample share of Asian voters for a Lookalike audience based on Asians with stereotypically Asian names and ZIP codes. One shift in 2021 was that Special Ad audiences based on Asians did not appear to over-represent them, unlike in 2020.

Limitations of This Study

While the results of this study describe how biased are Facebook’s ad targeting algorithms in 2020 and 2021, they are not able to fully describe why these biases occur. Even if an algorithm wasn’t intentionally designed to be discriminatory, bias can still occur through a series of different mechanisms.

- Underlying biases in the decisions made by humans may be reflected in the training data [60].

- The training data may be unrepresentative or incomplete [52, 88, 89].
- A reinforcement learning algorithm may become biased over time due to the biased behaviors of users [57, 90].
- Trade-offs in balancing an algorithm's performance in different fairness and accuracy metrics may result in biased outcomes for different demographics groups [65, 91, 92, 93, 94, 95, 96, 97, 98].
- A complex algorithmic decision-making system like Facebook's ad platform doesn't have just one algorithm but rather a series of different algorithms interacting with one another, so even if individual algorithms are not biased on their own, their interactions may result in biased outcomes [99].

In this case, it's possible that a combination of multiple causes contributed to the biased outcomes observed in this study. Researchers have found that racial and ethnic groups tend to behave differently from each other online. They visit different websites [100, 101, 102, 103], follow different social media [104, 105], and even browse the web using different devices [106, 107]. In addition, as an online social network, Facebook also has data on the friends of each user. Researchers have found that Americans tend to have very racially homogenous friend networks [108]. For White Americans, on average 91% of their social network are also White [108]. While for Black Americans, on average 83% of their social network are also Black [108]. Similarly, 75% of White Americans and 65% of Black Americans report having a core social network defined as "people with whom they discuss important matters" being entirely of their own race [108]. Future work would need to be done, ideally with a deeper access to Facebook's data and systems than what's possible from this external digital audit approach, in order to explain why Facebook's advertising algorithms are biased.

Finally, this study focused on how Facebook's algorithms treated African-Americans, Whites, and Asians as racial groups and Hispanics and Non-Hispanics as ethnic groups. This was due to how these demographic groups are the focus of Facebook's own ad targeting options and also exist as categories in the North Carolina voter data used for testing. Future studies may also examine other demographic groups such as Native Americans, multiracial individuals, and others. The way that Facebook generally rounds the reach estimate on its ad planning tool to the nearest thousand may require other approaches than the set theory approach used here for smaller demographic groups, since the reach estimate may not change by more than 1,000 users when using different ad targeting settings.

Implications for Facebook

While not every case of advertising discrimination is illegal or even potentially undesirable, such as the case of marketing textbooks to students, this study highlights how the lack of transparency by Facebook to the public and to its advertisers about how its ad platform can potentially discriminate by race and ethnicity may be exploited by discriminatory advertisers while undermining the goals of non-discriminatory ones. For example, discriminatory advertisers may already know that the "African-American Culture" targeting option contains fewer White users than the "African-American (US)" option Facebook removed in 2020. Discriminatory advertisers may also be using similar proxy variable techniques to the ones tested in this study based on racially stereotypical names and ZIP codes to create biased Lookalike and Special Ad audiences. On the other hand, non-discriminatory advertisers may be unintentionally choosing similar targeting settings as discriminatory ones but being unaware of how Facebook's ad platform is carrying out racially and ethnically biased targeting on their behalf.

The data currently disclosed by Facebook’s Ad Library falls short of what’s needed to detect racial discrimination by an advertiser. As of January 2021, Facebook’s Ad Library is limited to publishing some data about political, housing, employment and credit-related ads [109]. For political ads, Facebook publishes metadata about how much was spent, how many viewed the ad, and who saw the ad in terms of gender and state (Appendix C). Facebook does not release data on the Ad Library about which targeting options, such as the racially-affiliated interest groups studied here, were used. It also does not release information about the racial and ethnic breakdown of who saw an ad. For housing, employment, and credit-related ads, no metadata about who was targeted nor who saw the ad is released (Appendix C). This is especially problematic in light of the final version of the U.S. Department of Housing and Urban Development’s (HUD) “Implementation of the Fair Housing Act’s Disparate Impact Standard” published on September 24, 2020. This regulation required a plaintiff to present evidence of a “robust causal link” in order to bring a disparate impact discrimination lawsuit in the first place [110]. This study demonstrates how Facebook’s targeting options, Lookalike Audiences, and Special Ad Audiences can be used to discriminate by race or ethnicity, but Facebook’s Ad Library doesn’t currently release any data to document a “robust causal link” between how an advertiser is using Facebook’s tools and the discriminatory impact on who sees their ads.

Currently, Facebook’s anti-discrimination efforts are concentrated on limiting the targeting options for “Special Ads” related to housing, employment, or credit. Facebook disables the usage of certain sensitive targeting options for Special Ads such as multicultural affinity groups in 2020 and cultural interest groups in 2021 (Appendix C). Facebook only allows an advertiser to add more attributes as Detailed Targeting options and does not allow an advertiser to exclude any attribute (Appendix C). Facebook also provides notices on its ad planning tool to encourage advertisers to not

discriminate when targeting a Special Ad (Appendix C). Finally, Facebook does not allow Special Ads to use regular Lookalike audiences but rather they have to use Special Ad audiences, which Facebook designed to not use sensitive demographic attributes such as “age, gender or ZIP code” in considering which users to include [2]. However, this study has found that in 2020 and 2021 Special Ad audiences can become racially biased at similar rates to Lookalike audiences when using the same demographically homogenous customer list to create both types of audiences. Regular ads not related to housing, employment, or credit do not face any of these restrictions or notices.

In order to better tackle discrimination in the future, Facebook can leverage its data and analytical capabilities to better detect potential racial and ethnic discrimination for both “Special” and regular ads. Right now, Facebook’s ad planning tool already provides daily reach estimates given any combination of different ad targeting options, Custom audiences, Lookalike audiences, or Special Ad audiences, which is how I collected data about Facebook’s ad platform for this study. As a potential feature, Facebook can enrich its estimated reach report by displaying the demographic distribution of who will see an ad on the basis of race, ethnicity, gender, age, geography, and other categories. If a particular Custom, Lookalike, or Special Ad audience is racially or ethnically biased, Facebook can flag those audiences when they are first created in order to notify the advertiser and potentially limit their usage.

In order to carry out these digital audits for potential racial and ethnic biases in an advertiser’s target audiences, Facebook has a number of ways to collect or infer racial and ethnic data about its users. One option is similar to the North Carolina voter registration form, which asks a user to voluntarily provide their race and ethnicity. Facebook currently requests gender and date of birth on its account sign up page and includes an option for a “Custom” gender where a user can select their preferred pronouns and textbox for a preferred gender label. Facebook could adapt this approach to

collect racial and ethnic data directly from its users. It's possible that many would find it unappealing to give Facebook more data about themselves given past controversies over how Facebook handled user data such as the Cambridge Analytica scandal [111]. Another option is for Facebook to infer the data indirectly by following other examples in the tech industry such as Airbnb's Project Lighthouse, which was launched in 2020 to study the racial experience gap for guests and hosts on Airbnb [112]. Project Lighthouse used a third party contractor to assess the perceived race of an individual based on their profile picture and name [112]. Another approach is to use the name alone to infer race and ethnicity by using algorithmic approaches such as the `ethnicolr` Python library [78] or the U.S. Census Bureau's Frequently Occurring Surnames dataset [79]. In fact, there is precedence at Facebook for doing exactly this type of analysis. In December 2009, Facebook researchers, Lars Backstrom, Jonathan Chang, Cameron Marlow, and Itamar Rosenn published a paper on the race and ethnicity of Facebook users from January 2006 to January 2009 by comparing the last names of users to the U.S. Census' Frequently Occurring Surnames dataset [113]. They found that Facebook was becoming increasingly diverse over time by having more African-American and Hispanic users, which was reported at the time in *The Wall Street Journal* as "Facebook Touts Diversity of Its Members" [114].

How "Fairness Through Unawareness" Doesn't Prevent Algorithmic Discrimination

Finally, this study has ramifications beyond Facebook in terms of how to detect and address the issue of algorithmic discrimination in an increasingly digital world. Many of the anti-discrimination changes that Facebook has implemented in recent years to its advertising platform are examples of trying to achieve "fairness through unawareness" [63], the idea that discrimination is prevented by eliminating the use of protected class variables or close proxies. For example, Facebook explicitly created the Special Ad Audiences tool – as an alternative to Lookalike Audiences – to not use sensitive attributes such as "age, gender or ZIP code" in considering which users are similar enough to the

source audience to get included [2]. However, this study demonstrates that Special Ad audiences based on African-Americans or Whites can be biased towards the race that is more dominant in the customer list used to create the audience, just like the corresponding Lookalike audiences. In fact, even though Facebook designed its Special Ad Audience tool to explicitly not use ZIP codes as part of its algorithm, in 2021, the sample shares of African-American voters were 12 percentage points higher for Special Ad audiences based on African-Americans with stereotypically African-American ZIP codes versus African-Americans from anywhere in North Carolina (Figure 2.28). Likewise, in 2021, the sample shares of White voters were 10 percentage points higher for Special Ad audiences based on Whites from >90% White ZIP codes versus Whites from anywhere in North Carolina (Figure 2.29).

Statistics research has labelled this phenomenon as the Rashomon effect or the multiplicity effect [115]. This means that given a large dataset with many variables, there exists a large number of potential models that can perform approximately to equally as well as a prohibited model that uses protected class variables [63]. Thus, even though the Special Ad Audiences algorithm for finding similar users to a customer list does not use demographic attributes in the same way as the Lookalike Audiences algorithm, the two algorithms may end up making functionally comparable decisions on which users are considered to be similar enough to get included.

In recent years, there have been regulatory efforts to promote “fairness through unawareness” as a means of protecting companies from the liability of a discrimination lawsuit. This study illustrates that Facebook’s “fairness through unawareness” changes such as its Special Ad Audiences tool doesn’t necessarily prevent discrimination on the platform, though it may have prevented the ability for plaintiffs to successfully sue Facebook for discrimination if the proposed federal policies were implemented. On August 19, 2019, the initial language of the “Implementation of the Fair Housing Act’s Disparate Impact Standard” rule by the U.S. Department of Housing and Urban

Development's (HUD), which was drafted in response to the Supreme Court's *Texas v. Inclusive Communities* decision, stated that a defendant may successfully argue its model is not discriminatory if the model does not rely on "factors that are substitutes or close proxies for protected classes under the Fair Housing Act" [116]. After public comments criticized this language, it was removed in the final rule published on September 24, 2020 [110]. This study found that Facebook's Special Ad Audience tool can exhibit racial and ethnic biases even though it doesn't rely on sensitive attributes that are likely related to protected classes, which is the defense criteria established in the initial language of HUD's disparate impact rule.

References

1. Wagner K. Digital advertising in the US is finally bigger than print and television. Vox. February 20, 2019.
2. Facebook. Updates To Housing, Employment and Credit Ads in Ads Manager. 2019. .
3. Angwin J and Parris Jr T. Facebook Lets Advertisers Exclude Users by Race. ProPublica. October 28, 2016. <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>.
4. Angwin J, Tobin A, and Varner M. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. ProPublica. November 21, 2017. <https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin>.
5. Sapiezynski P, Ghosh A, Kaplan L, Mislove A, and Rieke A. Algorithms that “Don’t See Color”: Comparing Biases in Lookalike and Special Ad Audiences. 2019. <http://arxiv.org/abs/1912.07579>.
6. NATIONAL FAIR HOUSING ALLIANCE; FAIR HOUSING JUSTICE CENTER, INC.; HOUSING OPPORTUNITIES PROJECT FOR EXCELLENCE, INC.; FAIR HOUSING COUNCIL OF GREATER SAN ANTONIO v. Facebook. United States District Court, Southern District of New York. <https://nationalfairhousing.org/wp-content/uploads/2019/03/2018-06-25-NFHA-v.-Facebook.-First-Amended-Complaint.pdf>.
7. Sherwin G. How Facebook Is Giving Sex Discrimination in Employment Ads a New Life. ACLU. 2018. <https://www.aclu.org/blog/womens-rights/womens-rights-workplace/how-facebook-giving-sex-discrimination-employment-ads-new>.
8. Communications Workers of America vs. T-Mobile US, Inc. Amazon.com, Inc., Cox Communications, Inc., Cox Media Group, LLC, and similarly situated employers and employment agencies, DOES 1 through 1,000. United States District Court, Northern District of California. <https://doi.org/https://www.onlineagediscrimination.com/sites/default/files/documents/og-cwa-complaint.pdf>.
9. HUD Public Affairs. HUD CHARGES FACEBOOK WITH HOUSING DISCRIMINATION OVER COMPANY’S TARGETED ADVERTISING PRACTICES. Department of Housing and Urban Development. 2019. https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035.
10. Statt N. Facebook signs agreement saying it won’t let housing advertisers exclude users by race. The Verge. July 24, 2018.

- <https://www.theverge.com/2018/7/24/17609178/facebook-racial-discrimination-ad-targeting-washington-state-attorney-general-agreement>.
11. Timber C and Stanley-Becker I. Black voters are being targeted in disinformation campaigns, echoing the 2016 Russian playbook. The Washington Post. August 26, 2020. <https://www.washingtonpost.com/technology/2020/08/26/race-divisions-highlighted-disinformation-2016/>.
 12. Frenkel S and Barnes J. Russians Again Targeting Americans With Disinformation, Facebook and Twitter Say. The New York Times. September 1, 2020. <https://www.nytimes.com/2020/09/01/technology/facebook-russia-disinformation-election.html>.
 13. Horwitz J. Political Groups Elude Facebook's Election Controls, Repost False Ads. The Wall Street Journal. November 1, 2020. <https://www.wsj.com/articles/political-groups-elude-facebooks-election-controls-repost-false-ads-11604268856>.
 14. Stanley-Becker I. Disinformation campaign stokes fears about mail voting, using LeBron James image and boosted by Trump-aligned group. The Washington Post. August 20, 2020. https://www.washingtonpost.com/politics/disinformation-campaign-stokes-fears-about-mail-voting-using-lebron-james-image-and-boosted-by-trump-aligned-group/2020/08/20/fcadf382-e2e2-11ea-8181-606e603bb1c4_story.html?itid=lk_inline_manual_2.
 15. DiResta R et al. The Tactics & Tropes of the Internet Research Agency, New Knowledge. 2019. <https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/787lea6d5bafbf19/optimized/full.pdf#page=1>.
 16. Wagner K and Nix N. Facebook Scorned by Advocacy Groups After Zuckerberg Meeting. Bloomberg. July 7, 2020. <https://www.bloomberg.com/news/articles/2020-07-07/facebook-denounced-by-civil-rights-group-over-speech-policies>.
 17. Hsu T and Lutz E. More Than 1,000 Companies Boycotted Facebook. Did It Work? The New York Times. 1. August 2020. <https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html>.
 18. Murphy L and Relman Colfax. Facebook's Civil Rights Audit – Final Report. 2020.
 19. Sandberg S. Making Progress on Civil Rights – But Still a Long Way to Go. Facebook. 2020. <https://about.fb.com/news/2020/07/civil-rights-audit-report/>.
 20. Wagner K. Facebook Limits Ad Targeting That Some Linked to Race. Bloomberg. August 11, 2020. <https://www.bloomberg.com/news/articles/2020-08-11/facebook-further-limits-advertisers-ability-to-target-by-race>.

21. Newitz A. Facebook's ad platform now guesses at your race based on your behavior. *Ars Technica*. March 18, 2016.
22. Weise E. Facebook sued for housing and employment bias. *USA Today*. November 7, 2016. <https://www.usatoday.com/story/tech/news/2016/11/07/facebook-sued-housing-and-employment-bias/93424824/>.
23. SUZANNE-JULIETTE MOBLEY, KAREN SAVAGE, VICTOR ONUOHA, on behalf of themselves and all others similarly situated v. Facebook. United States District Court, Northern District of California. 2016. <https://pdfserver.amlaw.com/nlj/FacebookFairHousing.pdf>.
24. Angwin J and Tobin A. Fair Housing Groups Sue Facebook for Allowing Discrimination in Housing Ads. *ProPublica*. March 27, 2018. <https://www.propublica.org/article/facebook-fair-housing-lawsuit-ad-discrimination>.
25. Gosselin P. New Allegations Added to Lawsuit on How Facebook's Targeting Tools Helped Advertisers Exclude Older Workers. *ProPublica*. May 30, 2018. <https://www.propublica.org/article/new-allegations-lawsuit-on-how-facebook-targeting-tools-exclude-older-workers>.
26. Tobin A. Facebook Promises to Bar Advertisers From Targeting Ads by Race or Ethnicity. *Again*. *ProPublica*. July 25, 2018. <https://www.propublica.org/article/facebook-promises-to-bar-advertisers-from-targeting-ads-by-race-or-ethnicity-again>.
27. Rosenberg M. Facebook must stop advertisers from excluding people from viewing ads for housing, jobs and more. *Seattle Times*. July 24, 2018. <https://www.seattletimes.com/business/technology/facebook-must-end-discriminatory-ad-practice-under-deal-with-washington-attorney-general/>.
28. Boyd R et al. Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election using Linguistic Analyses. 1–9. 2018. <https://doi.org/10.31234/osf.io/ajh2q>.
29. Ribeiro F N et al. On microtargeting socially divisive ads: A case study of Russia-linked Ad campaigns on Facebook. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. 140–149. 2019. <https://doi.org/10.1145/3287560.3287580>.
30. Lecher C. Trump campaign using targeted Facebook posts to discourage black Americans from voting. *The Verge*. October 27, 2016. <https://www.theverge.com/2016/10/27/13434246/donald-trump-targeted-dark-facebook-ads-black-voters>.

31. Sanz C. Misinformation targeted Latino voters in the 2020 election. ABC News. November 21, 2020. <https://abcnews.go.com/Politics/latino-voters-misinformation-targets-election-2020/story?id=74189342>.
32. Romero L. Florida's Latino voters being bombarded with right-wing misinformation, experts and advocates say. ABC News. October 20, 2020. <https://abcnews.go.com/Politics/floridas-latino-voters-bombarded-wing-misinformation-advocates/story?id=73707056>.
33. Ryan-Mosley T. "It's been really, really bad": How Hispanic voters are being targeted by disinformation. MIT Technology Review. October 12, 2020. <https://www.technologyreview.com/2020/10/12/1010061/hispanic-voter-political-targeting-facebook-whatsapp/>.
34. Rodriguez S and Caputo M. "This is f---ing crazy": Florida Latinos swamped by wild conspiracy theories. Politico. September 14, 2020. <https://www.politico.com/news/2020/09/14/florida-latino-disinformation-413923>.
35. Wong Q. Facebook ad boycott: Why big brands "hit pause on hate." CNET. July 30, 2020. <https://www.cnet.com/news/facebook-ad-boycott-how-big-businesses-hit-pause-on-hate/>.
36. Guynn J. What civil rights groups want from Facebook boycott: Stop hate speech and harassment of Black users. USA Today. July 7, 2020. <https://www.usatoday.com/story/tech/2020/07/07/facebook-ad-boycott-racism-harassment-hate-african-americans/5385514002/>.
37. Fryer R G and Levitt S D. The causes and consequences of distinctively Black names. *Quarterly Journal of Economics*. Vol 119. No 3. 767–806. 2004. <https://doi.org/10.1162/0033553041502180>.
38. Imai K and Khanna K. Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Political Analysis*. Vol 24. No 2. 263–272. 2016. <https://doi.org/DOI: 10.1093/pan/mpw001>.
39. Bertrand M and Mullainathan S. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*. Vol 94. No 4. 991–1013. 2004. <https://doi.org/10.1257/0002828042002561>.
40. Kang S K, DeCelles K A, Tilcsik A, and Jun S. Whitened Résumés: Race and Self-Presentation in the Labor Market. *Administrative Science Quarterly*. Vol 61. No 3. 469–502. March 17, 2016. <https://doi.org/10.1177/0001839216639577>.
41. Quillian L, Pager D, Hexel O, and Midtbøen A H. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the*

- National Academy of Sciences. Vol 114. No 41. 10870 LP – 10875. October 10, 2017. <https://doi.org/10.1073/pnas.1706255114>.
42. Gaddis S M. An Introduction to Audit Studies in the Social Sciences BT - Audit Studies: Behind the Scenes with Theory, Method, and Nuance. Gaddis S M, Editor. Springer International Publishing. Cham. 2018. 3–44.
 43. WHITE A R, NATHAN N L, and FALLER J K. What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials. *American Political Science Review*. Vol 109. No 1. 129–142. 2015. <https://doi.org/DOI:10.1017/S0003055414000562>.
 44. Hanson A and Hawley Z. Where does racial discrimination occur? An experimental analysis across neighborhood and housing unit characteristics. *Regional Science and Urban Economics*. Vol 44. 94–106. 2014. <https://doi.org/https://doi.org/10.1016/j.regsciurbeco.2013.12.001>.
 45. Capps K and Rabinowitz K. How the Fair Housing Act Failed Black Homeowners. *City Lab*. April 11, 2018.
 46. Richardson J, Mitchell B, and West N. Home Mortgage and Small Business Lending in Baltimore and Surrounding Areas. 2015.
 47. Glantz A and Martinez E. For people of color, banks are shutting the door to homeownership. *Reveal News*. February 15, 2018.
 48. Chetty R, Hendren N, Jones M, and Porter S. Race and Economic Opportunity in the United States: An Intergenerational Perspective. 24441. 2019.
 49. Turner Lee N, Resnick P, and Barton G. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. 2019.
 50. Zarsky T. Understanding Discrimination in the Scored Society. *Washington Law Review*. Vol 89. No 4. 2014.
 51. Sweeney L. Discrimination in online Ad delivery. *Communications of the ACM*. Vol 56. No 5. 44–54. 2013. <https://doi.org/10.1145/2447976.2447990>.
 52. Buolamwini J and Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. Vol 81. 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>.
 53. Hannak A, Soeller G, Lazer D, Mislove A, and Wilson C. Measuring Price Discrimination and Steering on E-Commerce Web Sites. in *Proceedings of the 2014 Conference on*

- Internet Measurement Conference*. 2014. 305–318.
<https://doi.org/10.1145/2663716.2663744>.
54. Kay M, Matuszek C, and Munson S A. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015. 3819–3828.
<https://doi.org/10.1145/2702123.2702520>.
 55. Kulshrestha J et al. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017. 417–432.
<https://doi.org/10.1145/2998181.2998321>.
 56. Robertson R E, Jiang S, Joseph K, Friedland L, Lazer D, and Wilson C. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact.* Vol 2. No CSCW. November 2018. <https://doi.org/10.1145/3274417>.
 57. Datta A, Datta A, Makagon J, Mulligan D K, and Tschantz M C. Discrimination in Online Advertising: A Multidisciplinary Inquiry. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. Vol 81. 20–34.
<http://proceedings.mlr.press/v81/datta18a.html>.
 58. Chen L, Ma R, Hannák A, and Wilson C. Investigating the Impact of Gender on Rank in Resume Search Engines. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018. 1–14. <https://doi.org/10.1145/3173574.3174225>.
 59. Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, and Wilson C. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017. 1914–1933. <https://doi.org/10.1145/2998181.2998327>.
 60. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuter*. October 10, 2018. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
 61. Bolukbasi T, Chang K-W, Zou J, Saligrama V, and Kalai A T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. in *NIPS*. 2016. NIPS. <https://www.microsoft.com/en-us/research/publication/quantifying-reducing-stereotypes-word-embeddings/>.
 62. Obermeyer Z, Powers B, Vogeli C, and Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Vol 366. No 6464. 447 LP – 453. October 25, 2019. <https://doi.org/10.1126/science.aax2342>.

63. Xiang A. Reconciling Legal and Technical Approaches to Algorithmic Bias. *Tennessee Law Review*. Vol 88. No 3. 2021.
64. Federal Reserve. *Consumer Compliance Handbook*. 2017.
65. Feldman M, Friedler S A, Moeller J, Scheidegger C, and Venkatasubramanian S. Certifying and Removing Disparate Impact. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015. 259–268. <https://doi.org/10.1145/2783258.2783311>.
66. Zafar M B, Valera I, Gomez-Rodriguez M, and Gummadi K P. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research*. Vol 20. No 75. 1–42. 2019. <http://jmlr.org/papers/v20/18-262.html>.
67. Facebook Newsroom. Improving Enforcement and Promoting Diversity: Updates to Ads Policies and Tools. Facebook. 2017. <https://about.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools/>.
68. Tobin A and Merrill J. Besieged Facebook Says New Ad Limits Aren't Response to Lawsuits. *ProPublica*2. August 23, 18AD. <https://www.propublica.org/article/facebook-says-new-ad-limits-arent-response-to-lawsuits>.
69. Facebook Business. Keeping Advertising Safe and Civil. Facebook2. 2018. <https://www.facebook.com/business/news/keeping-advertising-safe-and-civil>.
70. Sherwin G and Bhandari E. Facebook Settles Civil Rights Cases by Making Sweeping Changes to Its Online Ad Platform. *ACLU*. 2019. <https://www.aclu.org/blog/womens-rights/womens-rights-workplace/facebook-settles-civil-rights-cases-making-sweeping>.
71. National Fair Housing Alliance. Facebook Settlement: Civil Rights Advocates Settle Lawsuit with Facebook: Transforms Facebook's Platform Impacting Millions of Users. 2019. <https://nationalfairhousing.org/facebook-settlement/>.
72. ACLU. Facebook Agrees to Sweeping Reforms to Curb Discriminatory Ad Targeting Practices. 2019. <https://www.aclu.org/press-releases/facebook-agrees-sweeping-reforms-curb-discriminatory-ad-targeting-practices>.
73. Berman G. STATEMENT OF INTEREST OF THE UNITED STATES OF AMERICA in National Fair Housing Alliance v. Facebook, Inc. UNITED STATES DISTRICT COURT SOUTHERN DISTRICT OF NEW YORK. New York. 2018. <https://www.justice.gov/crt/case-document/file/1089231/download>.
74. Tse A and Winslow S. UNITED STATES' STATEMENT OF INTEREST in ONUOHA v. FACEBOOK, INC. UNITED STATES DISTRICT COURT NORTHERN DISTRICT OF CALIFORNIA. San Jose. 2018. <https://www.justice.gov/crt/case->

document/file/1112561/download.

75. Abril D. The Facebook ad boycott ended months ago. But some big companies continue the fight. *Fortune*. November 7, 2020. <https://fortune.com/2020/11/07/facebook-ad-boycott-big-brands-lego-clorox-verizon-microsoft-hp/>.
76. Newton C and Schiffer Z. What a damning civil rights audit missed about Facebook. *The Verge*. July 10, 2020. <https://www.theverge.com/interface/2020/7/10/21318718/facebook-civil-rights-audit-critique-size-congress>.
77. Dwoskin E, Tiku N, and Kelly H. Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show. *The Washington Post*. December 3, 2020. <https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/>.
78. Sood G and Laohaprapanon S. Predicting Race and Ethnicity From the Sequence of Characters in a Name. May 5, 2018. <https://arxiv.org/abs/1805.02109>.
79. U.S. Census Bureau. Frequently Occurring Surnames from the 2010 Census. *Genealogy Data*. 2010. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html.
80. Yoon S, Falzon L, Anderson N B, and Davidson K W. A look at the increasing demographic representation within behavioral medicine. *Journal of Behavioral Medicine*. Vol 42. No 1. 57–66. 2019. <https://doi.org/10.1007/s10865-018-9983-y>.
81. Yeung N, Lai J, and Luo J. Face Off: Polarized Public Opinions on Personal Face Mask Usage during the COVID-19 Pandemic. October 31, 2020. <http://arxiv.org/abs/2011.00336>.
82. Shiffer-Sebba D and Behrman J. Gender and Wealth in Demographic Research: A Research Brief on a New Method and Application. *Population Research and Policy Review*. 2020. <https://doi.org/10.1007/s11113-020-09603-w>.
83. Nations J M and Martin I W. Racial Context and Political Support for California School Taxes. *Social Science Quarterly*. Vol 101. No 6. 2220–2237. October 1, 2020. <https://doi.org/https://doi.org/10.1111/ssqu.12869>.
84. Bertolero M A et al. Racial and ethnic imbalance in neuroscience reference lists and intersections with gender. *bioRxiv*. 2020.10.12.336230. January 1, 2020. <https://doi.org/10.1101/2020.10.12.336230>.
85. U.S. Census Bureau. American Community Survey (ACS). Our Surveys & Programs.

2019. <https://www.census.gov/programs-surveys/acs>.
86. U.S. Census Bureau. U.S. Census Bureau QuickFacts: North Carolina. QuickFacts. 2020. <https://www.census.gov/quickfacts/fact/table/NC/PST045219>.
 87. U.S. Census Bureau. Annual Estimates of the Resident Population by Sex, Age, Race Alone or in Combination, and Hispanic Origin for the United States: April 1, 2010 to July 1, 2019. Newsroom. 2019. <https://www.census.gov/newsroom/press-kits/2020/population-estimates-detailed.html>.
 88. Wiggers K. IBM releases Diversity in Faces, a dataset of over 1 million annotations to help reduce facial recognition bias. VentureBeat. January 29, 2019. <https://venturebeat.com/2019/01/29/ibm-releases-diversity-in-faces-a-dataset-of-over-1-million-annotations-to-help-reduce-facial-recognition-bias/>.
 89. Lohr S. Facial Recognition Is Accurate, if You're a White Guy. The New York Times. February 9, 2018. <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>.
 90. Emerging Technology from the ArXiv. Racism is Poisoning Online Ad Delivery, Says Harvard Professor. MIT Technology Review. February 4, 2013. <https://www.technologyreview.com/2013/02/04/253879/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>.
 91. Corbett-Davies S, Pierson E, Feller A, Goel S, and Huq A. Algorithmic Decision Making and the Cost of Fairness. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017. 797–806. <https://doi.org/10.1145/3097983.3098095>.
 92. Kleinberg J, Ludwig J, Mullainathan S, and Rambachan A. Algorithmic Fairness. AEA Papers and Proceedings. Vol 108. 22–27. 2018. <https://doi.org/10.1257/pandp.20181018>.
 93. Hardt M, Price E, and Srebro N. Equality of Opportunity in Supervised Learning. in *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016. 3323–3331.
 94. Zafar M B, Valera I, Gomez Rodriguez M, and Gummadi K P. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. in *Proceedings of the 26th International Conference on World Wide Web*. 2017. 1171–1180. <https://doi.org/10.1145/3038912.3052660>.
 95. Angwin J, Larson J, Mattu S, and Kirchner L. Machine Bias. ProPublica. May 23, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

96. Dieterich W, Mendoza C, and Brennan T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. 2016.
97. Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*. Vol 5. No 2. 153–163. June 1, 2017. <https://doi.org/10.1089/big.2016.0047>.
98. Kleinberg J. Inherent Trade-Offs in Algorithmic Fairness. in *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 2018. 40. <https://doi.org/10.1145/3219617.3219634>.
99. Dwork C and Ilvento C. Fairness Under Composition. in *ITCS*. 2019.
100. Sharad Goel J M H M I S. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. *Proceedings of the Sixth International AAI Conference on Weblogs and Social Media*. 1–8. 2012. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPDFInterstitial/4660/4975%5Cnpapers2://publication/uuid/F3306966-AB04-4CF1-ACCE-A7EF49E1282C>.
101. De Bock K and Van Den Poel D. Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*. Vol 98. No November. 49–70. 2010. <https://doi.org/10.3233/FI-2010-216>.
102. Ying J J, Chang Y, Huang C, and Tseng V S. Demographic Prediction Based on User's Browsing Behaviors. *Research.Nokia.Com*. Vol 2012. No 1. 1–6. 2012. <http://research.nokia.com/files/public/mdc-final241-ying.pdf%5Cnfile:///Users/orasanen/Documents/Papers/Ying/research.nokia.com>.
103. Sweeney L. Online ads roll the dice. *Tech@FTC*. 2014. <https://www.ftc.gov/news-events/blogs/techftc/2014/09/online-ads-roll-dice>.
104. Messias J, Vikatos P, and Benevenuto F. White, Man, and Highly Followed: Gender and Race Inequalities in Twitter. in *Proceedings of the International Conference on Web Intelligence*. 2017. 266–274. <https://doi.org/10.1145/3106426.3106472>.
105. Vikatos P, Messias J, Miranda M, and Benevenuto F. Linguistic Diversities of Demographic Groups in Twitter. in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. 2017. 275–284. <https://doi.org/10.1145/3078714.3078742>.
106. Fairlie R. Have we finally bridged the digital divide? Smart phone and Internet use patterns by race and ethnicity. *First Monday*. Vol 22. 2017.
107. Tsetsi E and Rains S A. Smartphone Internet access and use: Extending the digital divide and usage gap. *Mobile Media & Communication*. Vol 5. No 3. 239–255. June 13, 2017. <https://doi.org/10.1177/2050157917708329>.

108. Cox D, Navarro-Rivera J, and Jones R. Race, Religion, and Political Affiliation of Americans' Core Social Networks. PRRI. Vol 5. 1–5. 2014.
<https://www.ppri.org/research/poll-race-religion-politics-americans-social-networks/>.
109. Facebook. Ad Library. 2020. .
110. Department of Housing and Urban Development. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard. 2020.
<https://www.federalregister.gov/documents/2020/09/24/2020-19887/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard>.
111. Confessore N. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. The New York Times. April 4, 2018.
<https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>.
112. Airbnb. A new way we're fighting discrimination on Airbnb. Airbnb Resource Center. 2020. <https://www.airbnb.com/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201>.
113. Chang J, Rosenn I, Backstrom L, and Marlow C. ePluribus: Ethnicity on Social Networks. 2010. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1534>.
114. Taylor M. Facebook Touts Diversity of Its Members. The Wall Street Journal. December 18, 2009. <https://www.wsj.com/articles/BL-DGB-9541>.
115. Semenova L, Rudin C, and Parr R. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. August 5, 2019. <https://arxiv.org/abs/1908.01755>.
116. Department of Housing and Urban Development. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard. 2019.
<https://www.federalregister.gov/documents/2019/08/19/2019-17542/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard>.

Chapter 3

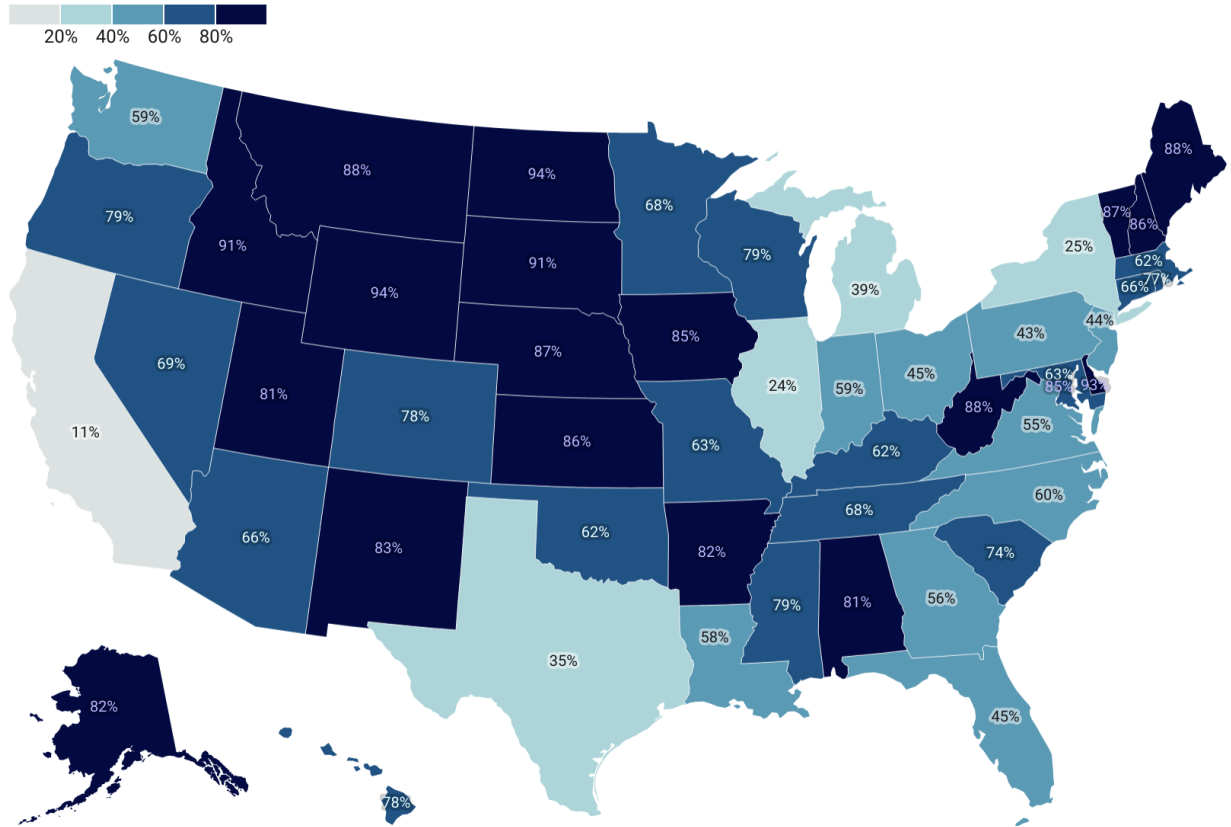
How Were Social Security Numbers Assigned?

Jinyan Zang

Highlights

- I build upon earlier research to propose my own hypothesis about SSN assignment as following a nested loop protocol
- For Americans born between 1989 and 2011, they have SSNs most vulnerable to prediction based on their state of birth and date of birth, due to the Social Security Administration's Enumeration At Birth program
- For SSNs in the Death Master File, I am able to accurately predict the first 5 digits 48% of the time and the first 6 digits 11% of the time
- States with smaller populations were the most vulnerable: I am able to accurately predict the first 5 digits of the SSN in 19 states including DC more than 80% of the time, and for 5 states – Delaware, Idaho, North Dakota, South Dakota, and Wyoming – more than 90% of the time
- It's time for public policy to focus on solutions that can replace SSNs with alternatives that are designed to be strong authenticators from the start

Predictive Accuracy of First 5 Digits of SSN By State (1989 - 2011)



Predictive Accuracy of First 5 Digits of SSN By State (1989 - 2011).

Abstract

Social Security Numbers (SSNs) serve dual purposes in the US. They are used as identifiers, which are “unique data used to represent a person’s identity and associated attributes”, but also as authenticators, “the means used to confirm the identity of a user, process, or device”. However, they were never designed to be strong random authenticators in the first place. Before the Social Security Administration (SSA) started randomizing SSNs in 2011, SSNs were assigned based on a protocol, partially released by the SSA, which generates a coded number based on an individual’s state for the first 3 digits and serial numbers for the remaining digits. For Americans born after 1989 when the Enumeration At Birth program began giving SSNs to newborns, I test whether this assignment protocol created a new vulnerability by effectively encoding an individual’s state and date of birth into their SSN. I build upon earlier research to propose my own hypothesis about SSN assignment as following a nested loop protocol. I test my hypothesis using the Death Master File, which is a SSA dataset that contains the SSNs of deceased individuals.

Results summary: I find strong evidence that my proposed SSN assignment protocol was used in all 50 states and DC between 1989 and 2011. Using regression models based off my hypothesis, I am able to predict the first 5 digits accurately 48% of the time and the first 6 digits accurately 11% of the time. There was significant variation between states with smaller population states being the most vulnerable. I am able to accurately predict the first 5 digits of the SSN in 19 states including DC more than 80% of the time, and for 5 states – Delaware, Idaho, North Dakota, South Dakota, and Wyoming – more than 90% of the time. Since the coding of personal information in one’s SSN is not obvious upon first glance, we tend to think of our SSN as a secret number while our state of birth and date of birth are not secrets. But I show how this misconception creates another vulnerability when using SSNs as

authenticators. Thus, it's time for public policy to focus on solutions that can replace SSNs with alternatives that are designed to be strong authenticators from the start.

Introduction

The first Social Security Numbers (SSNs) were issued on December 1, 1936 [1]. One unique number in order to keep track of the wages earned and the Social Security benefit to be received by an individual at retirement.

However, these numbers soon spread to usages for other purposes beyond Social Security. In 1943, President Franklin Roosevelt issued Executive Order 9397 which stated:

“Hereafter any Federal department, establishment, or agency shall, whenever the head thereof finds it advisable to establish a new system of permanent account numbers pertaining to individual persons, utilize exclusively the Social Security Act account numbers” [2].

Thus, the Internal Revenue Service (IRS) started using it for taxes in 1961. That same year, SSNs became the ID number for federal employees. Medicare used it for enrollment in 1965. The Veterans Administration started using it in 1966. The Department of Defense used it as the military ID number starting in 1969. Food stamps started using it in 1977. Housing and Urban Development (HUD) programs started using it in 1988. One federal program after another started adopting the SSN in order to track the individuals they service [3, 4]. State governments also used the SSN on driver’s licenses and marriage licenses [3].

Why is the Social Security Number so useful?

Because it is supposed to be one unique number for each individual, it works perfectly as an **identifier**, which is “unique data used to represent a person’s identity and associated attributes” [5]. First names and last names can also serve as identifiers, but many people share the same name. Thus, if two individuals both named “John Smith” requested assistance from a government agency, the government would need to request additional information to distinguish the two John Smiths apart

from each other. But if the government knew their Social Security Numbers, then there wouldn't be a de-duplication problem to resolve in the first place [4].

The attractiveness of Social Security Numbers soon spread to the private sector, especially in financial services. Under the 1970 Bank Records and Foreign Transactions Act, all banks, savings and loan associations, credit unions and broker/dealers in securities are required to obtain the SSNs of all of their customers [6]. Beyond banks, other financial services companies such as the three major credit bureaus, Equifax, Experian, and TransUnion, also use SSNs to distinguish the credit histories of different individuals.

As usage of SSNs spread, financial firms and government agencies weren't simply using them as identifiers, but also as **authenticators**, "the means used to confirm the identity of a user, process, or device" [1]. This means if an individual submits a credit card application to a bank or submits their tax return to the IRS and includes their SSN on their form in addition to other identifiers such as their name and date of birth, the bank or the IRS will check to see if the given SSN is actually associated with the given name and date of birth. If the check is successful, then the bank or the IRS may assume that the individual has confirmed their identity successfully and continue processing the submitted form.

There's a feedback loop between the widespread usages of SSNs by different services and how that reinforces future usages of SSNs to link datasets together using one number that most likely exists in many places. For example, SSNs can be used as identifiers to link records about an individual from multiple financial, housing, criminal history, and other datasets together for a background check. It's also a convenient authenticator to request an individual to fill out a 9-digit number on a credit card, rental, job, or other form, especially online, when compared to verifying other forms of government IDs such as driver's licenses or passports. In the US today, potential customers do not

have to provide their SSNs when requested, but businesses then are able to refuse to service the customer, unless there are specific federal or state laws that regulate the transaction [7]. In *Cassano v. Carb* (2006), the US Court of Appeals, Second Circuit ruled that the “the Constitution does not provide a right to privacy in one's SSN... we decline to expand the constitutional right to privacy to cover the collection of SSNs” [8]. In this case, the court ruled against an employee seeking to not provide their SSN to their employer due to fears of identity theft.

However, the same widespread usage of SSNs, especially as authenticators, has contributed to the identity theft problem. In 2019, the cost of identity theft was \$16.9 billion and impacted 5.1% of Americans [9].

What are the vulnerabilities to using SSNs as authenticators?

First, the same number should not be used as an identifier and an authenticator [10]. The issue is that an identifier can become more widely known and discoverable as more services start using it. However, this also undermines its strength as an authenticator, since besides the individual, now many other services also know of their SSN. In recent decades, the government has passed laws to limit the sharing of datasets with SSNs. For example, after 1990, datasets released by the federal government should not disclose SSNs [7]. In 2000, the amended Driver's Privacy Protection Act required state departments of motor vehicles to obtain consent from individuals in order to release their SSNs [7]. The 1999 Gramm-Leach-Bliley Act regulated the sharing of personally identifiable information (PII) by financial institutions, the 1996 Health Insurance Portability and Accountability Act (HIPAA) regulated the sharing of PII in health data, and the Family Educational Rights and Privacy Act (FERPA) regulated the sharing of PII by educational institutions [7]. However, these laws are limited in scope and don't address the issue of unintended release of SSNs through data breaches.

Second, large scale data breaches have made many SSNs available to identity thieves. For example, the 2017 Equifax data breach possibly revealed personal information including the SSNs of 145.5 million individuals [11]. In prior research, I have found that dark web marketplaces sell breached datasets and SSNs for as low as \$1 per SSN [12]. As the threat of data breaches and identity theft grows, Americans have become far more wary of giving out their SSNs. Research on how many respondents to the Census Bureau's General Social Survey provided their SSN, which was an optional field in the contact information section, found a decrease from 60% in 1993 to 17% in 2008 [13].

Third, Social Security Numbers, before 2011, were never designed to be strong random authenticators in the first place. According to the World Bank's 2019 *ID for Development* report, random numbers are ideal for authenticators when compared to serial numbers (numbers assigned sequentially) or coded numbers (numbers that contain an individual's attributes such as birth year, gender, nationality, location, or more) [10]. This is because random numbers (1) reveal no personal information, (2) are more secure by making it harder for an attacker to guess, and (3) are immutable by not needing to be updated over time like coded numbers if an individual's gender or location changes [10]. However, before SSA started randomizing SSNs in 2011 [14], SSNs were assigned based on a protocol, partially released by the SSA, which generates coded and serial numbers [15]. The first three digits, the Area Number (AN), are a coded number for the state where the individual is applying from, and the middle two digits, the Group Number (GN), and the last four digits, the Serial Number (SN), are serial numbers that follow a sequence for assignment over time created by the SSA. What is not publicly confirmed by the Social Security Administration is the relationship between the Area Number, the Group Number, and the Serial Number in terms of how they get assigned over time. However, given enough SSNs to put into sequential order per state, this assignment protocol can

reveal the relationship between an individual's state and date of SSN assignment and their SSN. But how can someone learn when did an individual apply for an SSN?

In 1989, the SSA started the Enumeration At Birth (EAB) program, which was an anti-fraud program that integrated the application for SSNs into the birth certification process. By 1995, 50% of all new SSNs being assigned were given to newborns [16, 17]. Prior to this program, most individuals applied to receive an SSN when they started working as an adult and needed an SSN in order to track their contributions into Social Security. However, in 1988 Congress passed legislation requiring SSNs for children 2 years old or older to be claimed as dependents on tax returns, and in 1990, new legislation lowered that requirement to children 1 years old or older [4], which created more of an incentive for parents to claim SSNs for their children at birth. Thus after 1989, for most individuals their state of assignment and date of assignment was likely their state of birth and date of birth.

In 2009, Acquisti and Gross analyzed a database of Social Security Numbers published by the SSA of deceased individuals and demonstrated how it's possible to predict SSNs based on an individual's state of birth and date of birth, which is publicly accessible information that can be found online such as on social media websites [18]. They were able to accurately predict the first 5 digits for individuals born between 1989 and 2003, 44% of the time and all 9 digits 0.9% of the time.

Based on their analysis of SSNs and its assignment to individuals over time, Acquisti and Gross propose that the SSN Assignment Protocol is as follows [18]:

“The combined SSN assignment scheme consists of ANs transitioning first; after 9,999 ANs associated with a certain combination of GN and GN, the next AN in the issuance scheme is assigned; then, when all ANs assigned to a state or territory are exhausted, the next GN in the scheme is assigned.”

However, they don't actually fully demonstrate that this protocol was followed by all states over the years; nor do they actually exploit the implied sequential assignment of SSNs described by their proposed protocol to improve their SSN prediction accuracy.

Thus, in this study I propose an alternative SSN assignment protocol, which is based on the Acquisti and Gross hypothesis, and demonstrate that it was used by all 50 states and DC between 1989 and 2011. Using an individual's state and date of birth, I was able to predict the first 5 digits accurately 48% of the time and the first 6 digits 11% of the time using a 2013 dataset of the SSNs of decedents.

This means that the third vulnerability described above from having a non-random SSN as an authenticator can still potentially harm Americans even if they were fortunate enough to avoid the first two vulnerabilities. An individual can follow recommended best practices and be wary of sharing their SSN unless absolutely necessary [13], and the businesses and government agencies that have their SSNs may have strong security in place to prevent data breaches. If that individual was born after the start of the Enumeration At Birth program in 1989 and before the SSA started randomizing all 9 digits of the SSN on June 25, 2011 [14], then they could have an SSN that an identity thief could predict based on their state of birth and date of birth. This attack becomes more potent if the last 4 digits of their SSN, which are the hardest to predict, are already known to the attacker due to the common practice, endorsed by the Internal Revenue Service (IRS), of obscuring the first 5 digits while revealing the last 4 digits of an SSN on many forms and documents [19].

How can an American respond to this vulnerability? Should they start hiding their birthday celebrations or stop showing hometown pride? That may not even make a big difference since their state of birth and date of birth are often already available in many datasets and on social media [12, 18]. Since the coding of personal information in one's SSN is not obvious upon first glance, we tend to think of our SSN as a secret number while our state of birth and date of birth are not secrets. But I show how this misconception creates another vulnerability when using SSNs as authenticators. Given all of these vulnerabilities, it's time for public policy to focus on designing solutions that can replace Social Security Numbers with alternatives designed to be strong authenticators from the start.

Background

History of SSNs

When the Social Security system was being established after the passage of the Social Security Act in 1935, the federal government needed a way to track the earnings of 26 million workers as part of their “lifetime working record” [20]. Multiple options were considered. Using names would create “endless perplexities” given the number of people with similar names such as 294,000 Smiths, 227,000 Johnsons, and 165,000 Browns over the age of 65 [20]. Another option was fingerprints, which were already used by the War and Navy departments and the Veterans Administration [20]. However, there were concerns that would be an unpopular solution given the “connotations attaching to it from police usage” [20]. The Social Security Board didn’t want American workers to think they were being treated like criminals. Thus, eventually the Social Security Board agreed to create unique account numbers, our Social Security Numbers [20]. But with SSNs, there were still concerns about how it can be dehumanizing and empower the federal government to limit the privacy and freedom of Americans [20]. For example, the Republican National Committee (RNC) chairman John D. M. Hamilton in 1936 charged that eventually Americans would need to wear “dog tags” showing their SSN [20]. Newspapers made the comparison between dog tags of SSNs and being drafted even though there was no war [20]. The Social Security Board responded by accusing the RNC of spreading “deliberate falsehood” and a “hostile campaign to confuse, deceive, and scare the people of this country by threats, coercion and by misleading statements” [20]. However, the criticism likely contributed to the Board choosing to use a paper card for Social Card Cards rather than a metal token, which would have been more durable and error proof but potentially more similar to dog tags [20]. They also tried to emphasize in their communications that the SSN is simply a means of tracking one’s “account” and not the “person” in order to minimize the “charge of regimentation” [20]. The Social Security Board

also justified using SSNs since Title VIII of the Social Security Act stated that “an identifying number will be assigned to each employer and to each employee” [20].

So how were the different parts of the Social Security Number assigned?

Social Security Numbers have 3 parts that follow an XXX-YY-ZZZZ structure. Based on what has been released by the SSA, we know the following about how each part of the SSN is assigned.

The first 3 digits (XXX) are the Area Numbers (ANs) assigned to each state according to Table 3.1.

0xx	1xx	2xx	3xx	4xx	5xx	6xx	7xx
001-003 NH	135-158 NJ	212-220 MD	303-317 IN	400-407 KY	501-502 ND	600-601 AZ	750-751 HI
004-007 ME	159-211 PA	221-222 DE	318-361 IL	408-415 TN	503-504 SD	602-626 CA	752-755 MS
008-009 VT		223-231 VA	362-386 MI	416-424 AL	505-508 NE	627-645 TX	756-763 TN
010-034 MA		232-236 WV	387-399 WI	425-428 MS	509-515 KS	646-647 UT	764-765 AZ
035-039 RI		237-246 NC		429-432 AR	516-517 MT	648-649 NM	766-772 FL
040-049 CT		247-251 SC		433-439 LA	518-519 ID	650-653 CO	
050-134 NY		252-260 GA		440-448 OK	520-520 WY	654-658 SC	
		261-267 FL		449-467 TX	521-524 CO	659-665 LA	
		268-302 OH		468-477 MN	525-525 NM	667-675 GA	
				478-485 IA	526-527 AZ	676-679 AR	
				486-500 MO	528-529 UT	680-680 NV	
					530-530 NV	681-690 NC	
					531-539 WA	691-699 VA	
					540-544 OR		
					545-573 CA		
					574-574 AK		
					575-576 HI		
					577-579 DC		
					585-585 NM		
					587-588 MS		
					589-595 FL		

Table 3.1. Area Numbers assigned to each state 1989 – 2011 [21, 22].

The middle 2 digits (YY) are the Group Numbers (GNs) which are not assigned sequentially 01 to 99 but rather following the order below (Table 3.2) [15].

Assignment Order	1	2	3	4	5	6	...	50	51	52	53	54	55	...	99
GN	01	03	05	07	09	10	Evens	98	02	04	06	08	11	Odds	99

Table 3.2. Group Number assignment order.

The last 4 digits (ZZZZ) are the Serial Numbers (SNs) that are assigned sequentially from 0001 to 9999 [15].

What is not publicly confirmed by the Social Security Administration is the relationship between the Area Number, the Group Number, and the Serial Number in terms of how they get assigned over time.

In the beginning, Social Security Number applications were processed at post offices, then starting in July 1937 at regional Social Security offices, and finally, in 1961, all new SSN assignment was centralized to a Social Security office in Baltimore and done via computers starting in 1972 [4]. In 1989, the SSA started the Enumeration At Birth (EAB) program, which was an anti-fraud program that integrated the application for SSNs into the birth certification process. By 1995, 50% of all new SSNs being assigned were given to newborns [16, 17]. Starting on June 25, 2011, the SSA started randomizing all 9 digits for new SSN assignments [14]. According to the SSA, randomization will “protect the integrity of the SSN” and “extend the longevity of the nine-digit SSN nationwide”, since Area Numbers are no longer designated for different states [23]. SSA acknowledged that randomization “will help protect an individual's SSN by making it more difficult to reconstruct an SSN using public information” [23], which is the attack first described by Acquisti and Gross in 2009 [18].

Acquisti and Gross Hypothesis for SSN Assignment Protocol

In their 2009 paper, Acquisti and Gross hypothesized that SSNs were assigned in a nested loop pattern starting with Group Numbers, then Area Numbers, and last Serial Numbers [18] as shown in the pseudo-code below and visualized for Massachusetts, which has the ANs 010 to 034. As shown in Figure 3.1, the first assigned SSN is 010-01-0001, and assignment would continue with the same 010-

01 (AN-GN pair) until 010-01-9999. Then the next AN would be assigned while keeping the GN the same so the SSN would be 011-01-0001 incrementing the SN until 011-01-9999. The last possible SSN for GN 01 would be 034-01-9999. Afterwards, the next SSN would increment the GN and restart the nested loops with the first possible AN and SN again resulting in 010-03-0001. This assignment pattern can continue until we reach the end of the loop for the last possible GN, 99, the last possible AN, 034, and the last possible SN, 9999, forming the SSN 034-99-9999. Since Massachusetts has 25 unique Area Numbers, this means there are approximately 25 million combinations of possible SSNs assigned to Massachusetts.

For y in $(GN_1, GN_2, GN_3, \dots)$:

For x in $(AN_1, AN_2, AN_3, \dots)$:

For z in $(SN_1, SN_2, SN_3, \dots)$:

$$SSN = "AN_x - GN_y - SN_z"$$

Example SSN Assignment Protocol for Massachusetts

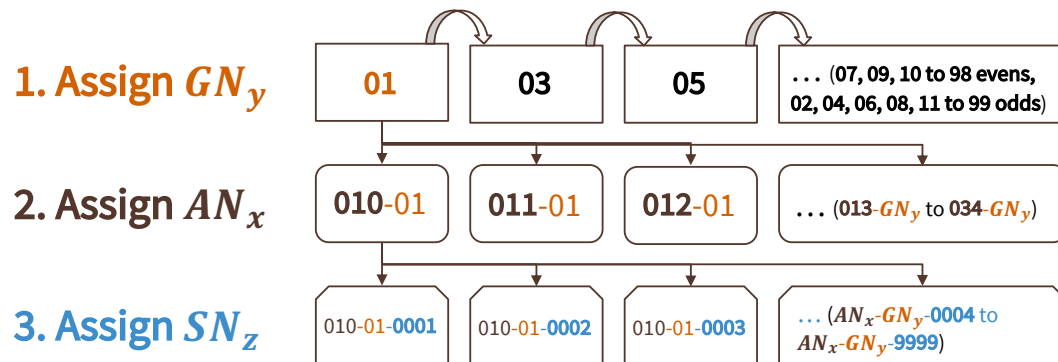


Figure 3.1. Acquisti and Gross hypothesis for SSN assignment protocol and example SSN assignment pattern for Massachusetts. Massachusetts has ANs (010 to 034). The figure above shows example ANs if the GN was 01 and example SNs if the GN was 01 and the AN was 010. Since

it follows a nested loop pattern, the next GN to be assigned would be 03, 05, etc. and each new GN would go through the nested set of ANs and SNs below it.

While Acquisti and Gross provide examples supporting their hypothesis in their paper, they don't comprehensively show it for all states across time. In addition, the primary prediction algorithm in their paper disregards the likely assignment order of SSNs based on their hypothesis, which likely reduced the accuracy of their predictions.

Acquisti and Gross used a 2-step process for predicting SSNs for a target state of birth and date of birth [18].

1. Choose the **modal** ANGN within a variable window (vw) of days around the target date

$$vw_{i,y} = 0.8 * \frac{365}{(\text{number of births in state } i \text{ during year } y)/9,999}$$

The motivation for choosing 0.8 multiplier according to Acquisti and Gross is that:

“Specifically, for each state and year, we extracted the number of births from NCHS data and calculated how many days during that year and in that state it would take, on average, to assign 9,999 SSNs (the number which marks the switch from one ANGN combination to the next). We then calculated 80% of that number of days, rounded that up to the closest odd integer, and used the resulting number as the window of days for all calculations performed in that state in that year...We selected windows of days 20% shorter than the number of days it would theoretically take to transition from one ANGN to the next one (under the simplifying assumption that all SSNs were assigned under EAB to newborns in the state), in order to reduce the number of such windows that would fall at the overlap between two or more assigned ANGNs.”

The calculated variable window is rounded up to the nearest odd number and then subtract 1 and divide by 2 to get the number of days before and after the target date to include in the variable window.

2. Regress SNs against dates of birth and ANGN dummies within the variable window and predict the SN using the target date of birth and the modal ANGN from the Step 1

$$SN_i = \alpha + \beta_1 dob_{i,vw} + \sum_{j=1}^{all\ ANGNs\ in\ vw} \beta_j ANGN_j + \epsilon_{i,vw}$$

If the predicted SN for the target date of birth is < 1 or > 9999 then bound the prediction at 0001 and 9999 respectively.

In Step 1, Acquisti and Gross takes only the modal ANGN within the variable window, which is an inefficient prediction method that discards useful information, since ANGNs don't even need to appear in the order described by their hypothesis in order to be predicted. Each ANGN just needs to be assigned for a given time period in its own cluster to become the modal choice. But knowing the likely ANGN for a given target date doesn't provide additional information for the most likely ANGN in future dates with their method.

There's also lost accuracy in Step 2, especially if the target date falls within the overlap between two ANGN combinations. Since the SN predictions are bounded at 0001 or 9999 if the predicted SN is below 1 or above 9999, it's not possible to modify the ANGN by 1 to the previous or next ANGN combination in the sequence and then take the remainder SN below 1 or above 9999 as the new SN for the modified ANGN.

Methods

This paper is the first to take a comprehensive approach to test an extension of the Acquisti and Gross hypothesis for all states and DC for SSNs assigned between 1989 and 2011. The overall approach is to first convert all SSNs within the dataset to a modified SSN index that reflects the

sequential order of assignment according to my hypothesis. Then I regress the SSN index values on dates of birth for each state, and use the regression results to predict SSNs – after converting back from predicted SSN index values to SSNs – for a given state of birth and date of birth.

Arriving at the Zang hypothesis for SSN Assignment Protocol

In examining the 2013 DMF data we find that for 15 states not all ANs were being used before the GN changed: Arizona, Arkansas, Colorado, Florida, Georgia, Hawaii, Louisiana, Mississippi, Nevada, New Mexico, North Carolina, South Carolina, Tennessee, Utah, and Virginia. These 15 states have two sets of ANs, and there appears to be a Step 0 where the state exhausts all possibly SSNs in the first set of ANs before starting over with the second set (Figure 3.2). For example, in Colorado, the last SSN assigned using the first set of ANs was 524-99-9999 and the next SSN assigned starting with the second set of ANs would be 650-01-0001. This means, in contrast to the Acquisti and Gross hypothesis, a higher group number, 99, was assigned before a lower group number, 01, for the same state because the AN set changed which restarted the loop for iterating through GNs. The Zang hypothesis can also be generalized to apply to all states including states that only have just one set of ANs.

Zang hypothesis for SSN Assignment Protocol

For i in $\{(AN\ set\ 1: AN_{1,1}, AN_{1,2}, \dots), (AN\ set\ 2: AN_{2,1}, AN_{2,2}, \dots)\}$:

For y in $(GN_1, GN_2, GN_3, \dots)$:

For x in $(AN_{i,1}, AN_{i,2}, AN_{i,3}, \dots)$:

For z in $(SN_1, SN_2, SN_3, \dots)$:

$$SSN = "AN_{i,x} - GN_y - SN_z"$$

Example SSN Assignment Protocol for Colorado

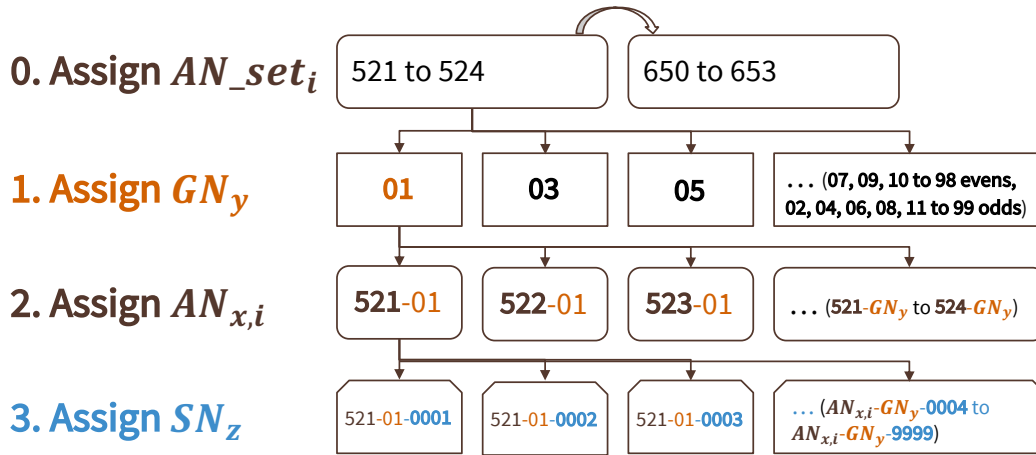


Figure 3.2. Example SSN assignment pattern for Colorado based on the Zang Hypothesis. Since Colorado has 2 sets of ANs, 521 to 524 and 650 to 653, SSN assignment according to Hypothesis 2 exhausted all possible SSN combinations using the first set of ANs before moving onto the second set. Thus 524-99-9999 was the last SSN assigned using the first set of ANs and 650-01-0001 was the next SSN assigned starting with the second set of ANs. The figure above shows the possible SSN assignments with the first AN set with AN being 521 and GN being 01.

Approach Details

I have a 3-step approach for data preparation, model training, and SSN prediction and analysis.

Step 1. Data Preparation

I downloaded the 2013 Death Master File (DMF) that contained the SSNs of deceased individuals who passed away before or during 2013 [24]. Since I'm only interested in predicting the SSNs of individuals after the start of the Enumeration at Birth program and before the start of SSN randomization, I subset the DMF to only include the 260,665 individuals born on or between 1/1/1989 and 6/24/2011.

I then converted the ANGN combination into an index value based on the predicted order of assignment according to the Zang hypothesis.

First, I converted GNs into GN index values based on the known GN assignment order from the Social Security Administration (Table 3.2).

Then, for each state, I created a GNAN index based on the Zang hypothesis and the observed ranges of GNs assigned in combination with the relevant sets of ANs. For example, for Colorado, for the first set of ANs, 521 to 524, GN index values ranging from 84 to 99 were observed; for the second set of ANs, 650 to 653, GN index values ranging from 1 to 32 were observed. This means that 521-84 would be the first observed AN-GN index value during the 1989 – 2011 study period from the first set of ANs which iterates until 524-99, the 64th assigned GNAN index value, before switching to the second set of ANs starting with 650-01 until 653-32 (Table 3.3).

GNAN Index	AN-GN Index
1	521-84
2	522-84
3	523-84
4	524-84
5 to 64	521-85 to 524-99
65	650-01
66	651-01
67	652-01
68	653-01
69 to 192	650-02 to 653-32

Table 3.3. GNAN Index to AN-GN Index mapping for Colorado. For GNAN index 1 to 64, the first AN set of 521 to 524 is iterated through in combination with GN index values of 84 to 99. For GNAN index 65 to 192, the second AN set of 650 to 653 is iterated through in combination with GN index values of 01 to 32.

Finally, in order to calculate the SSN index value, I calculated

$$SSN_Index = GNAN_Index * 10,000 + SN$$

Thus, for an SSN from the state of Colorado with a GNAN index of 1 and an SN of 1234, the SSN index value would be 11234. Since the maximum SN is only 9999 before the GNAN index value increments, the SSN index formula linearizes the SSN according to the assignment protocol predicted by the Zang hypothesis. Figure 3.3 shows the predicted assignment of SSNs converted into SSN index values for Colorado. While the majority of the observations, likely due to the Enumeration at Birth program, increments at a predictable rate forming a black line, there are some observations far above the line. For example, Point A is an outlier most likely of an individual who didn't receive their SSN at birth in 1990 but rather much later in life around 2005, when other newborns in 2005 were receiving their own SSNs through the EAB program. In addition, in 1989 to 1994, the early years of the EAB program, there appears to be a ramp-up process with noisier data due to some individuals likely receiving their SSNs a few months or years after their birth, resulting in more points slightly above the line.

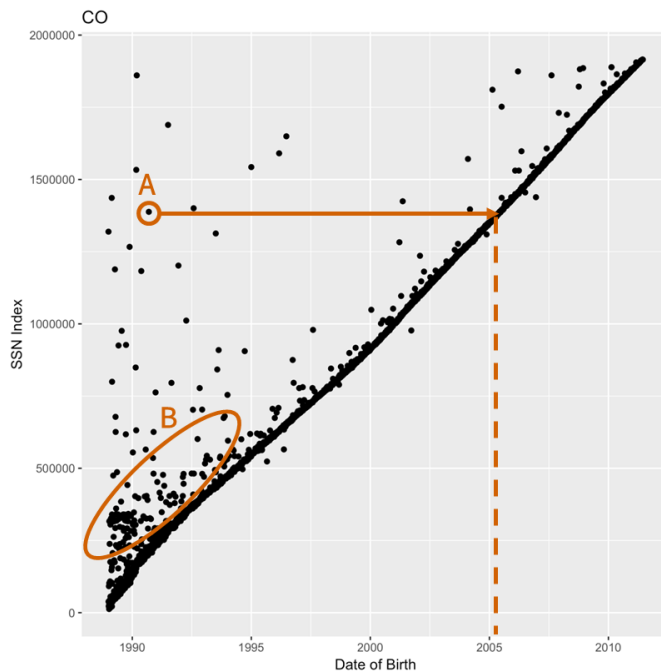


Figure 3.3. SSN index values by Date of Birth for Colorado. Point A is an outlier most likely of an individual who was born in 1990 but wasn't assigned an SSN until 2005 around the same time as other newborns in 2005 were receiving their own SSNs through the Enumeration At Birth (EAB) program. The observations within Area B is likely due to the ramp-up for EAB when not every child received their SSN at birth, but with some children receiving their SSNs a few months or years later.

In order for my prediction to be reliable using an individual's date of birth, focusing on only individuals who likely received their SSNs through the Enumeration At Birth program as newborns or young children in the US, I wanted to exclude outlier data points similar to Point A from Figure 3.3, far above the bulk of the post-EAB observations. To identify these outliers, I used the Cook's Distance formula to exclude outliers with a Cook's Distance value $> 4 / \#$ of observations, which is a common benchmark for outlier detection of influential points that may bias a regression [25]. To estimate the Cook's Distance value for each point, I regressed SSN index value on date of birth and quadratic and cubic transformations of date of birth to allow for some non-linearity in SSN assignment rate over time. Figure 3.4 shows the points identified as outliers by the Cook's Distance rule for Colorado. Upon visual inspection for all states and DC, the Cook's Distance metric appears to perform well at eliminating desired outliers while still being a conservative measure, with only 2.33% of observations identified as outliers.

$$\text{If Cook's Distance}_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p + 1)\hat{\sigma}^2} > \frac{4}{N}, \text{ then influential outlier}$$

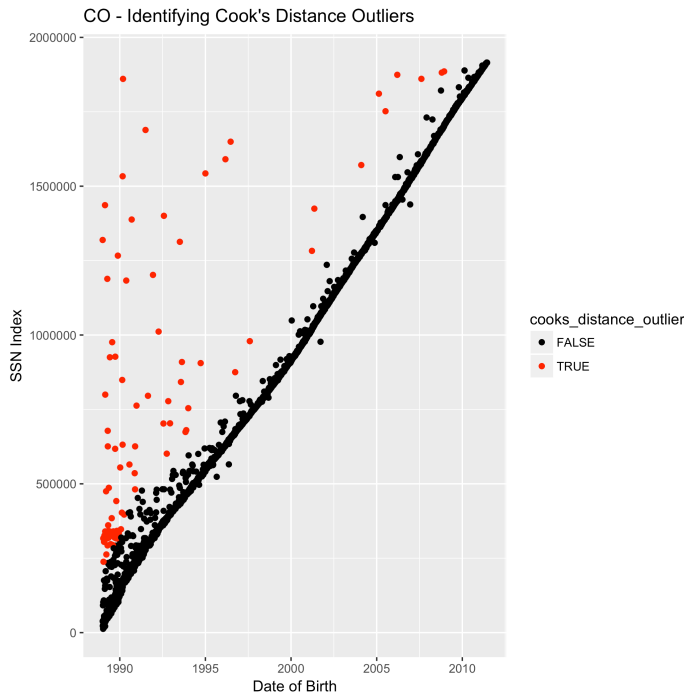


Figure 3.4. Identifying Cook's Distance Outliers in Colorado. Red points are outliers with Cook's Distance values $> 4 / \#$ of observations.

I compared the foreign born percentage of the population in each state from the 2000 Decennial Census versus the percentage of SSNs identified as Cook's distance outliers in each state from 1995 – 2011, after the ramp-up of the EAB program, which is shown in Appendix A. I didn't find a significant relationship between the two variables, which may be due to confounders such as states with high foreign born populations also being more likely to have undocumented foreign born residents who are included in the Census data but not in the SSN data.

Step 2. Model Training

I used a 5-fold cross validation approach in order to test the robustness of my model.

Since SSN assignment rate within the same state may fluctuate over time due to changes in birth rates, I used a robust loess regression, also known as a robust locally weighted polynomial regression, in order to fit locally optimal regressions for model training using 5% of the data around

each x , or date of birth, and a tricubic kernel that reduces the weight of points further away from x when calculating the local regression [26].

At each x , I find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ to minimize:

$$\sum_{i=1}^n p(y_i - \beta_0 - \beta_1(x - x_i) - \beta_2(x - x_i)^2)K\left(\frac{x - x_i}{h}\right), \text{ where } x = \text{date of birth}$$

where the resulting estimate is:

$$\hat{m}(x) = \sum_{j=0}^2 \hat{\beta}_j x^j$$

I used Iteratively Reweighted Least Squares algorithm using Tukey's biweight function for the reweighting to reduce the weight for any remaining outliers. $K(\cdot)$ is a compacted support kernel with a local weight function that downweights x_i that are far away from x . The kernel function is tricubic:

$$K(x) = \begin{cases} (1 - |x|^3)^3 & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1 \end{cases}$$

The parameter h in the kernel is the bandwidth, which I specified to 5% of the data to be within the support of $K(\cdot/h)$.

I used the loess function in the stats library in R for model training.

Step 3. SSN Prediction and Analysis

Using the loess model based on the training data, I predicted SSN index values for dates of birth for each state in the test data. I then do a reverse look-up to convert the SSN index value into an SSN in order to analyze the accuracy of the SSN predictions. All results are reported as the average across the 5-fold cross validation, and since people likely die randomly with regards to their dates of birth and thus enter the DMF, the pseudo-out-of-sample testing of the 5-fold cross validation is likely reflective of the what the results of the prediction would be for the whole population. Figure 3.5 shows the close fit from the loess regression based on the training data for Colorado as compared to the remaining test data points in Fold #1.

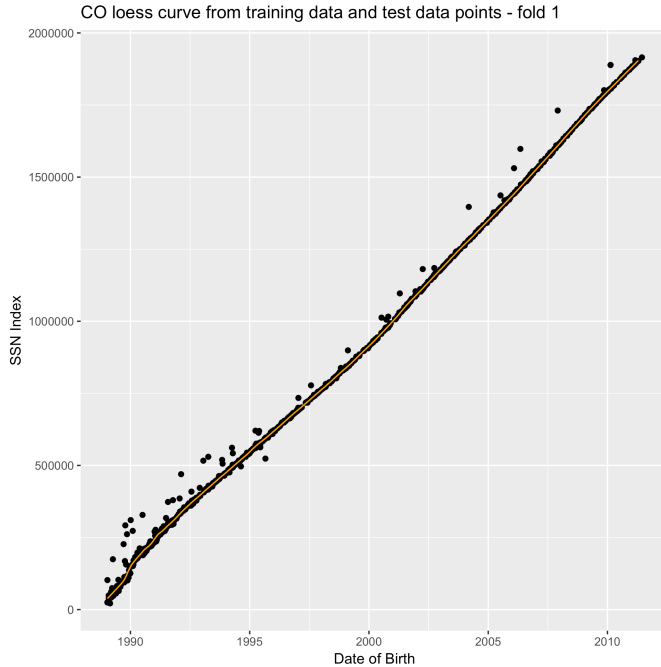


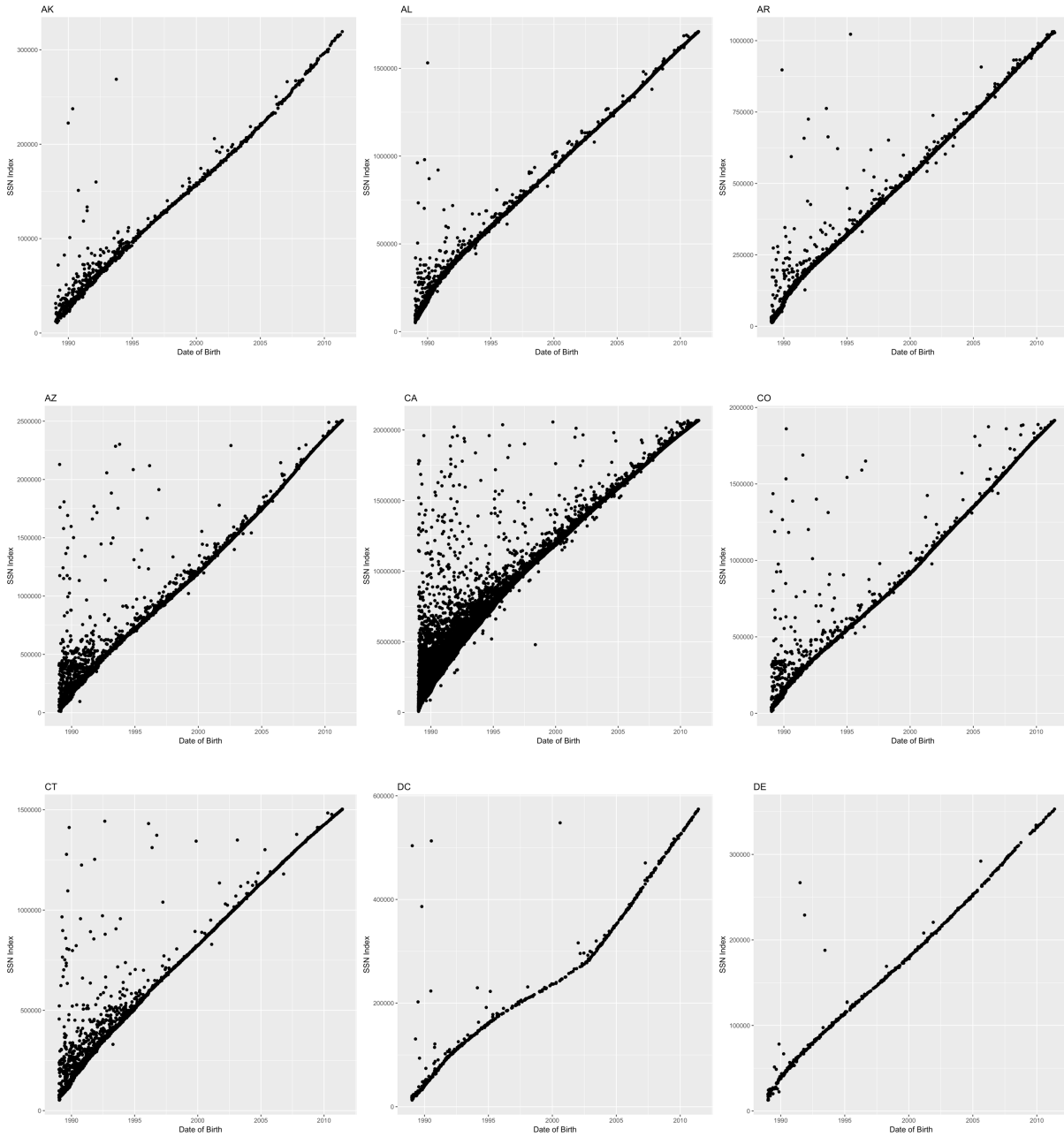
Figure 3.5. Loess curve from the training data and the test data points for Colorado in Fold #1. The orange curve is from the loess regression using the 80% of the dataset subsample that is the training data for Fold #1 and the black points are the 20% of data subsample that is test data.

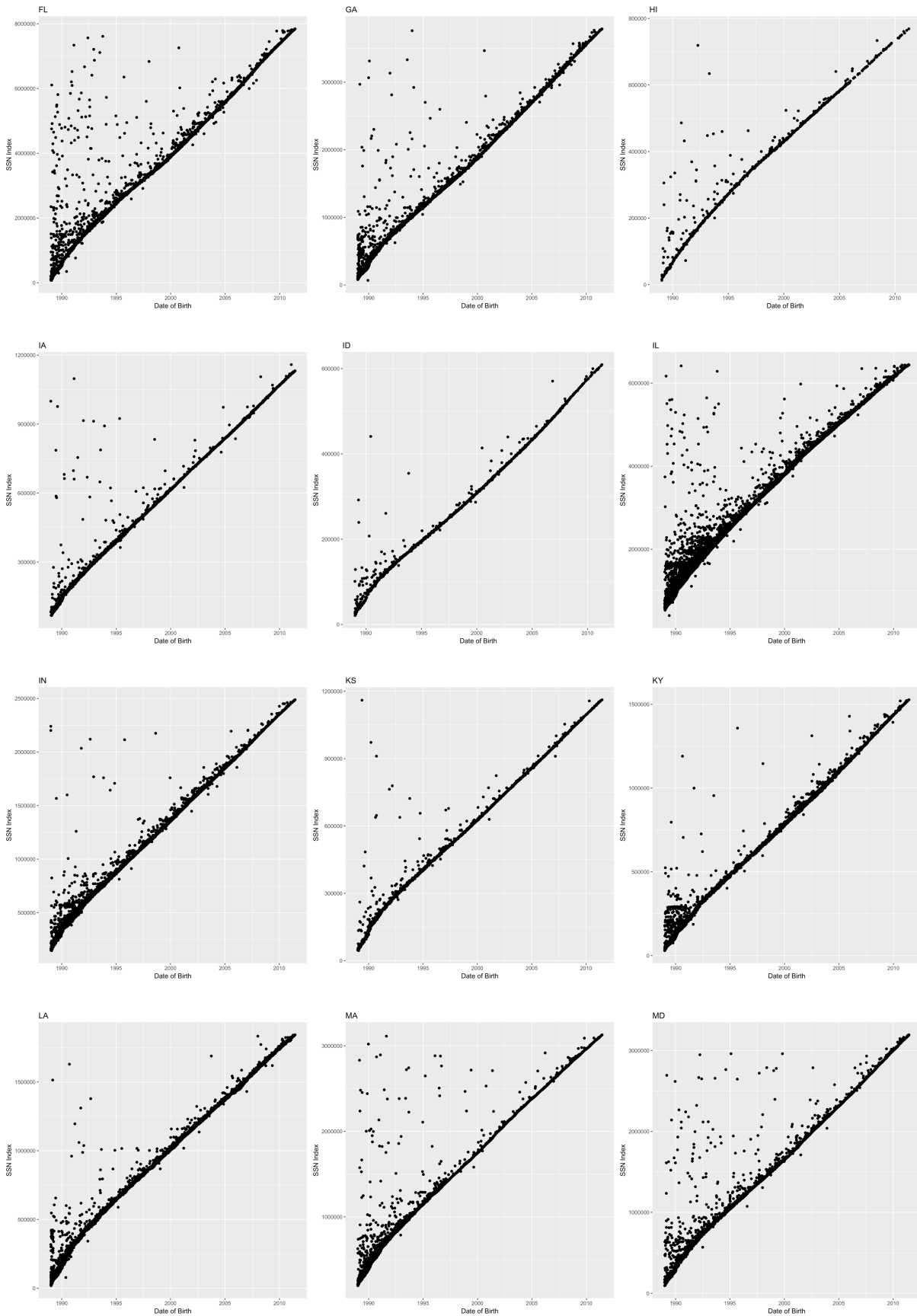
Results

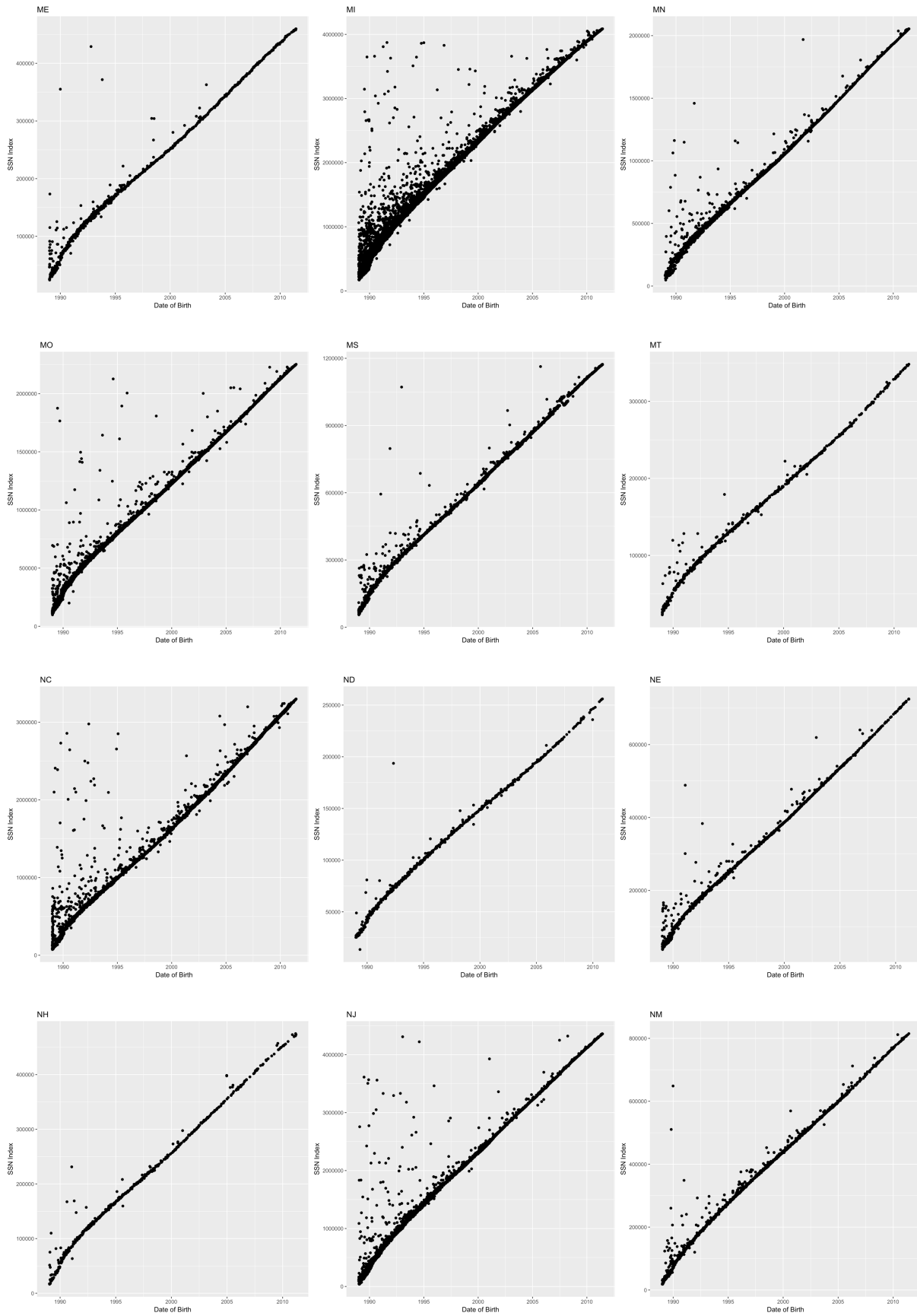
Confirming the Zang Hypothesis as the SSN Assignment Protocol

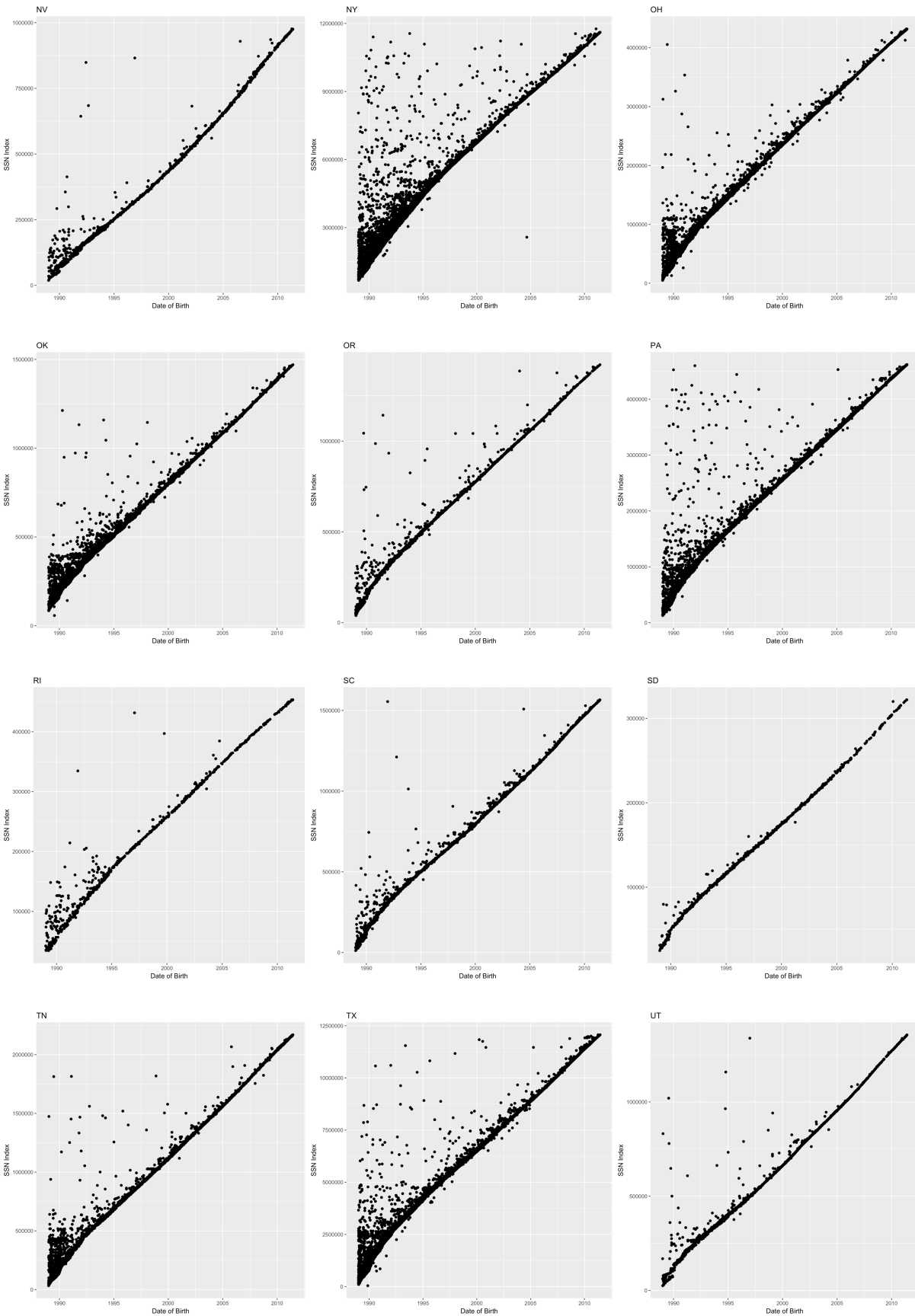
After converting the SSNs of every state into their respective SSN index values according to the data preparation steps described above, which is based on the Zang hypothesis for the SSN assignment protocol, I found that all 50 States and DC exhibited a nearly linear relationship between SSN index values and dates of birth during the study period (Figure 3.6), which provides strong evidence that the Zang hypothesis was the SSN assignment protocol for all of the US from 1989 to 2011. If a state did not follow this protocol, then we would expect irregularities and non-linear gaps when plotting SSN index values against dates of birth, which is a proxy variable for date of assignment

after the start of the EAB program in 1989. For most states, only 2% or less of the SSNs were considered Cook's distance outliers (Appendix A).









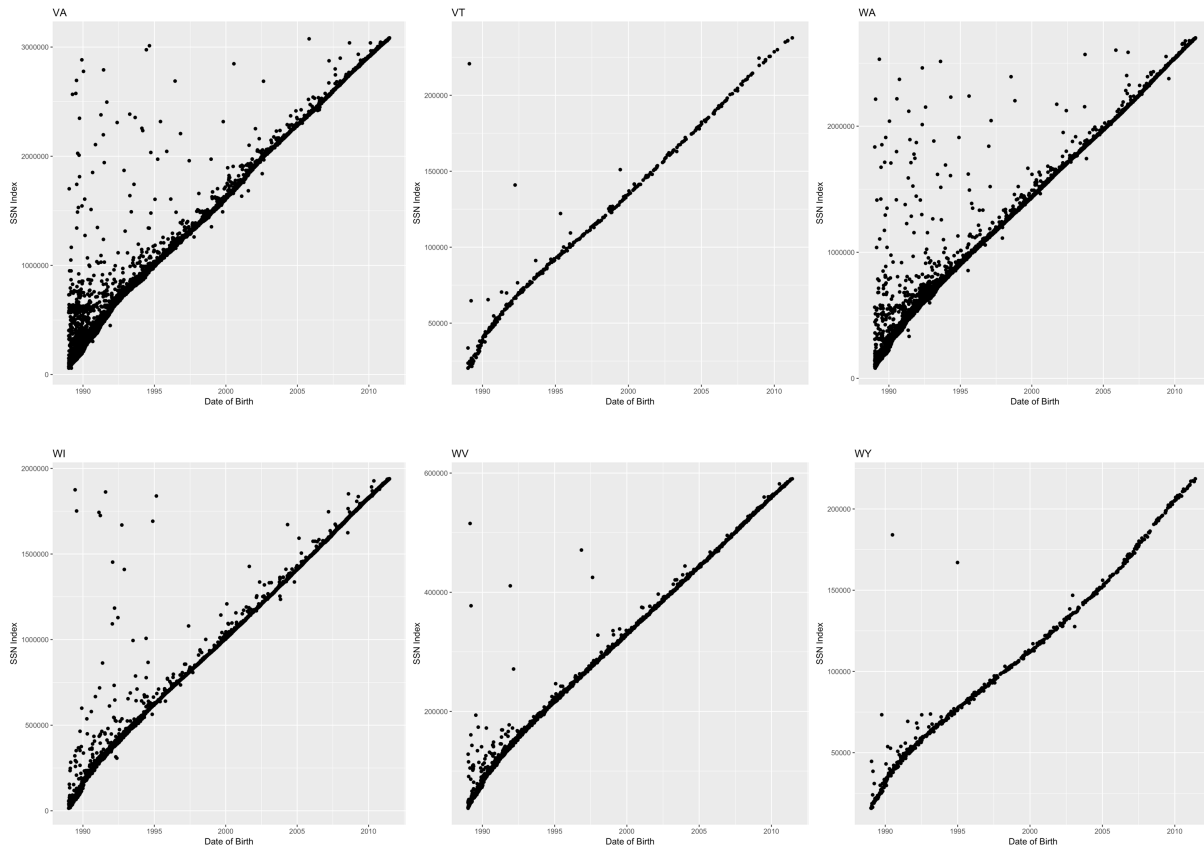


Figure 3.6. SSN Index vs. Date of Birth for all states and DC 1989 – 2011.

Predictive Accuracy

Since every state appeared to have followed the Zang hypothesis SSN assignment protocol, my loess regressions of SSN index values on dates of birth for each state performed well. The median prediction error weighted by annual births per state in terms of the difference in SSN index values for the actual SSN versus the predicted SSN after 5-fold cross validation was 6,724 for 1989 to 2011 and just 3,581 for 1995 to 2011. Since there was a ramp up period in the implementation of EAB, the median prediction error quickly declined from more than 30,000 in 1989 to less than 5,000 after 1996 and continued to decline to less than 2,000 by 2011 when Randomized Assignment starts (Figure 3.7).

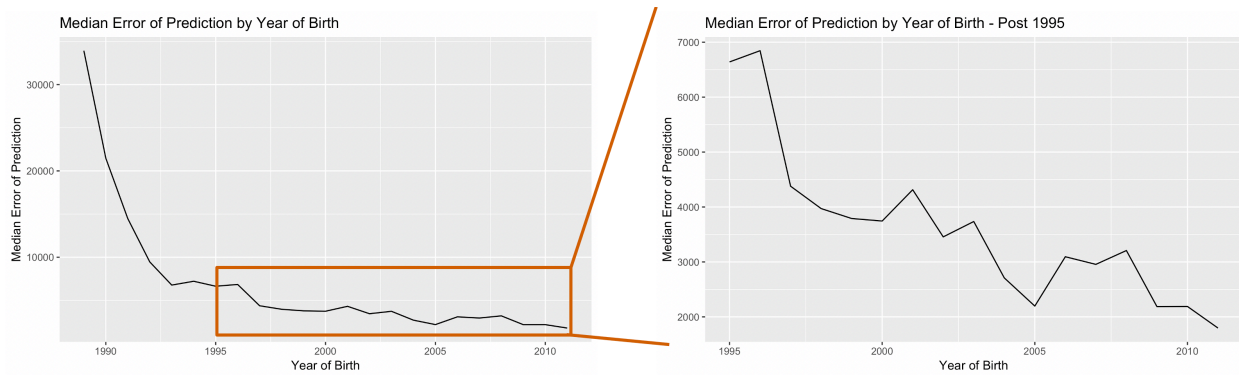


Figure 3.7. Median Prediction Error by Year of Birth from 1989 to 2011 and a zoomed-in look at 1995 -2011.

For the whole study period of 1989 – 2011, the loess models were able to accurately predict the first 5 digits 48.5% of the time and the first 6 digits 10.8% weighted by annual births per state, and for 1995 – 2011, and the loess models were able to accurately predict the first 5 digits 55.2% of the time and the first 6 digits 12.8% of the time. Figure 3.8 shows that predictive accuracy increased from 19.3% for the first 5 digits in 1989 to 66.3% by 2011. The loess models also improved their accuracy for the first 6 digits from 2.8% in 1989 to 16.1% by 2011. The loess models were not very accurate in predicting beyond the first 6 digits.

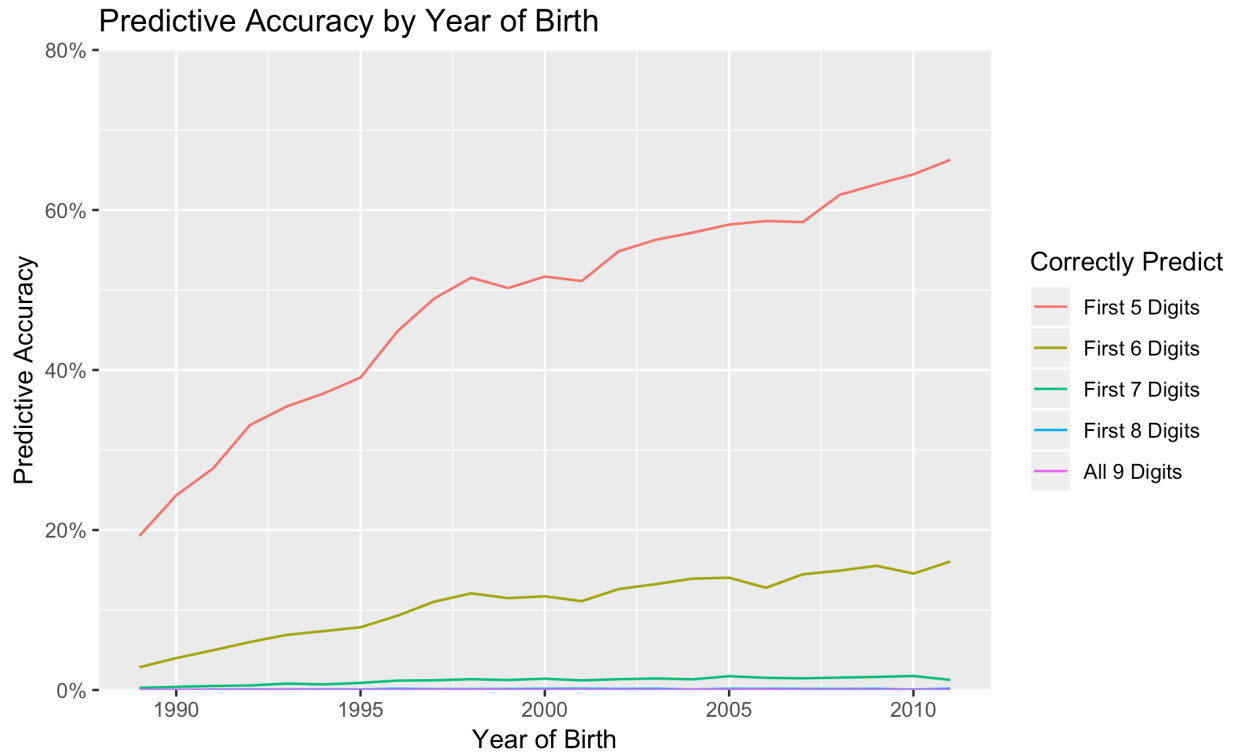


Figure 3.8. Predictive Accuracy by Year of Birth.

There is significant variation between the states in terms of the median error of the loess models over the entire study period from 1989 to 2011 (Figure 3.9). Generally, states with small populations such as Delaware, South Dakota, Wyoming, and Maine had the lowest median errors all below 520, while larger states such as California (21,142), New York (10,034), and Illinois (11,313) had very large median errors.

Median Error of SSN Prediction By State (1989 - 2011)

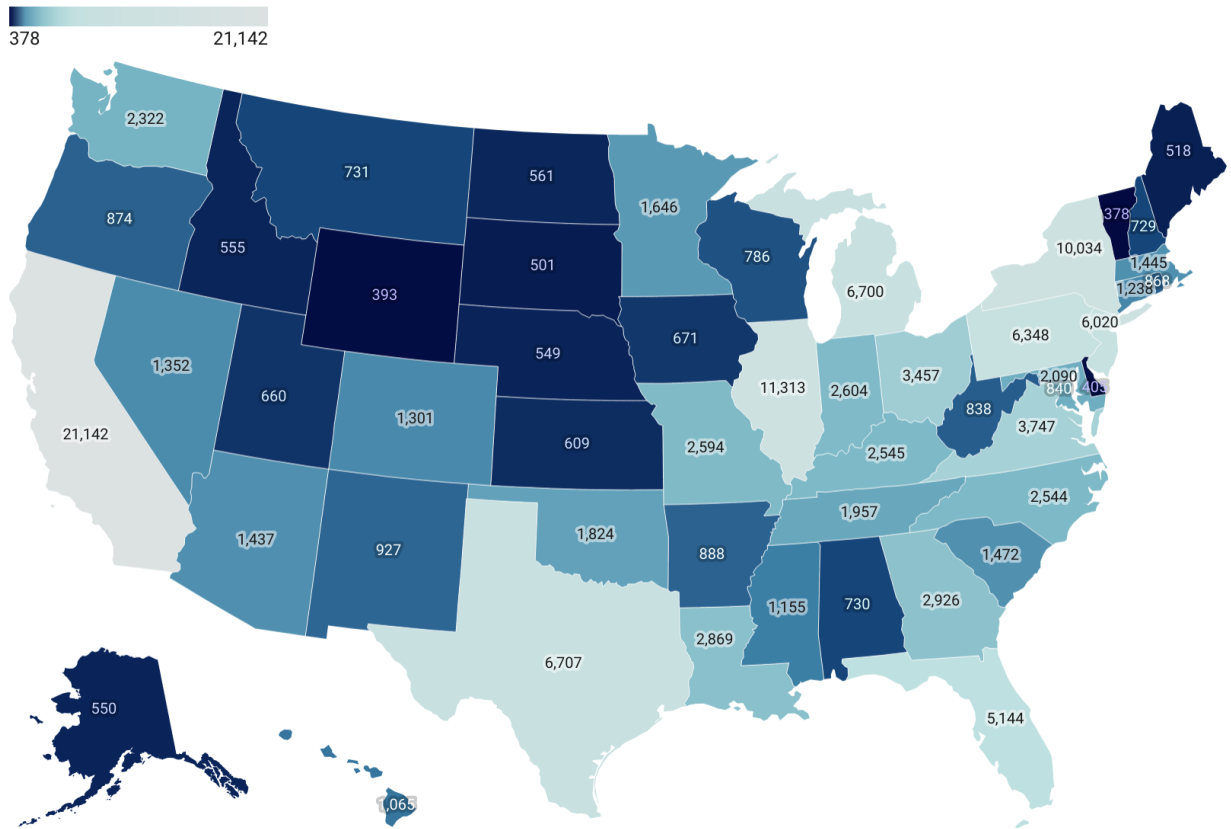


Figure 3.9. Median Error of SSN Prediction By State (1989 – 2011).

Figure 3.10 shows the median error of SSN prediction versus mean daily births by state for 1989 to 2011. The x and y-axis are log-scaled, so there's an exponential relationship between the two variables, which is shown as a linearly positive trend on the plot with higher median errors as daily births increase. The x-axis becomes a potential lower bound on how low median errors of SSN predictions can go even with more data since multiple SSNs are being assigned for each a state per date of birth.

Median Error of SSN Prediction vs.
Mean Daily Births by State (1989 - 2011)

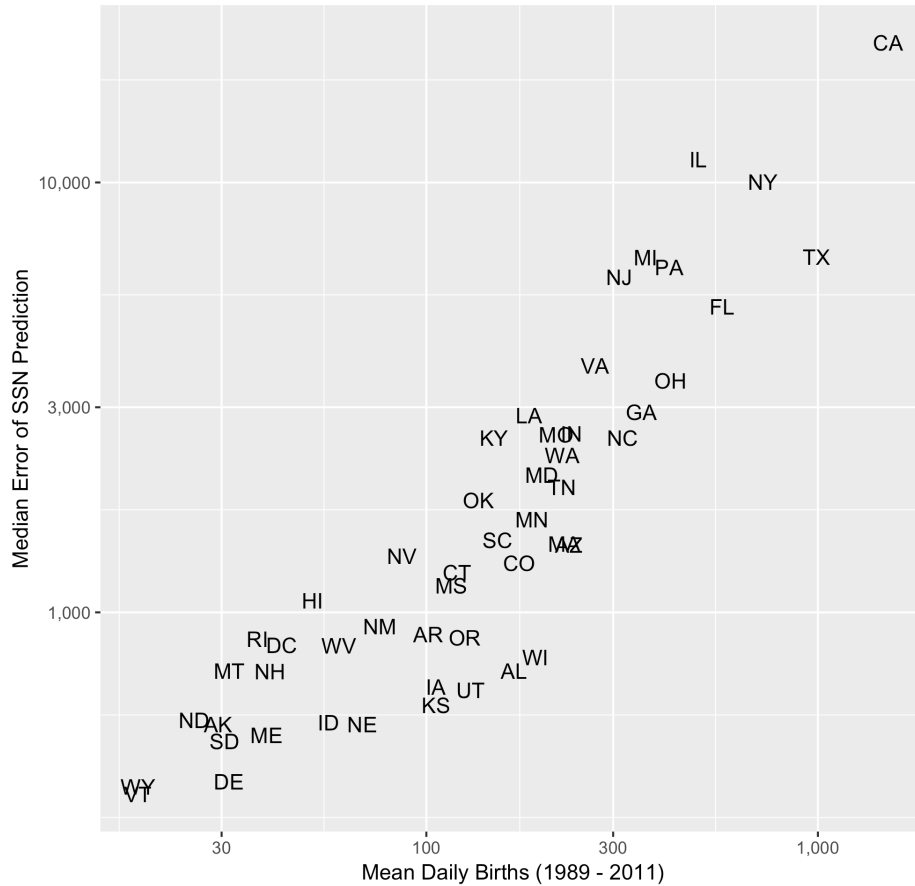


Figure 3.10. Median Error of SSN Prediction vs. Mean Daily Births by State (1989 - 2011). The x- and y-axis are log-scaled.

I was able to accurately predict the first 5 digits of the SSN in 19 states including DC more than 80% of the time, and for 5 states – Delaware, Idaho, North Dakota, South Dakota, and Wyoming – more than 90% of the time (Figure 3.11).

Predictive Accuracy of First 5 Digits of SSN By State (1989 - 2011)

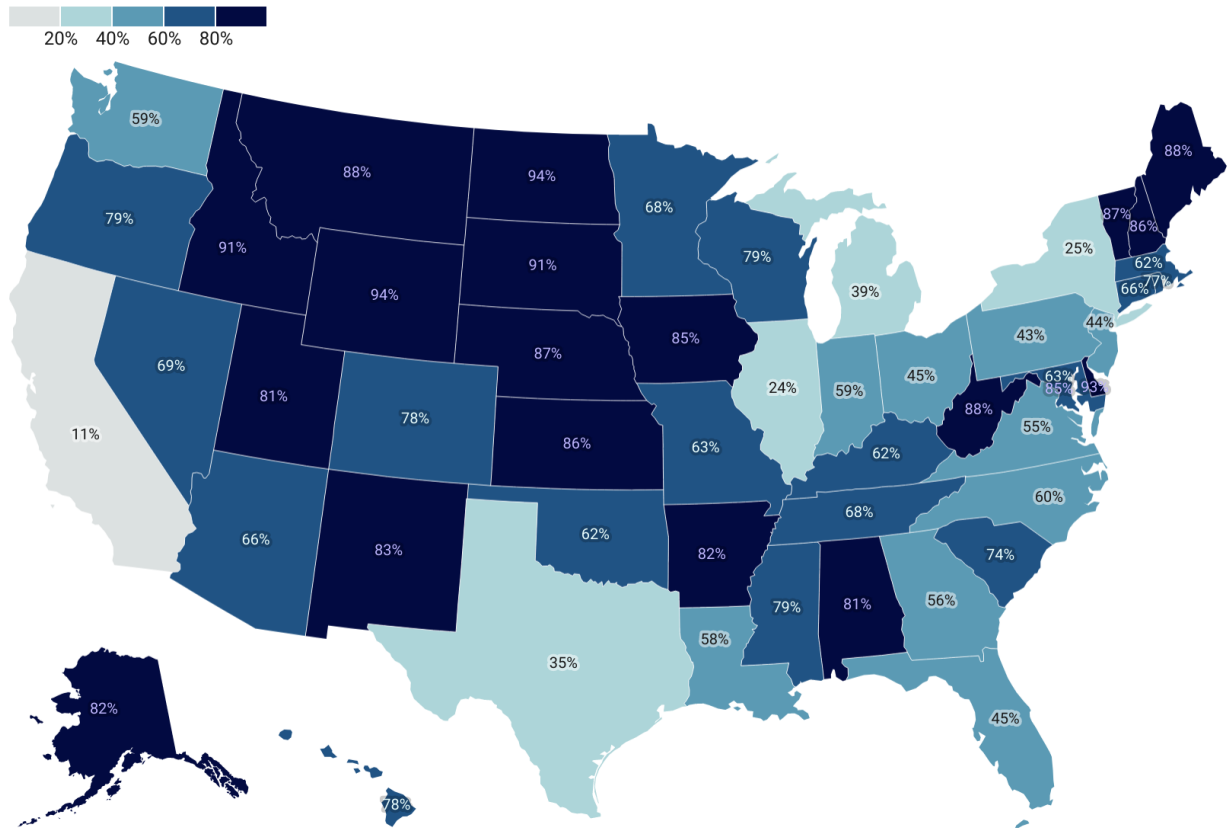


Figure 3.11. Predictive Accuracy of First 5 Digits of SSN By State (1989 – 2011).

I was able to accurately predict the first 6 digits of the SSN in 15 states including DC more than 30% of the time, and for 6 states – Delaware, Idaho, North Dakota, South Dakota, Vermont, and Wyoming – more than 40% of the time (Figure 3.12).

Predictive Accuracy of First 6 Digits of SSN By State (1989 - 2011)

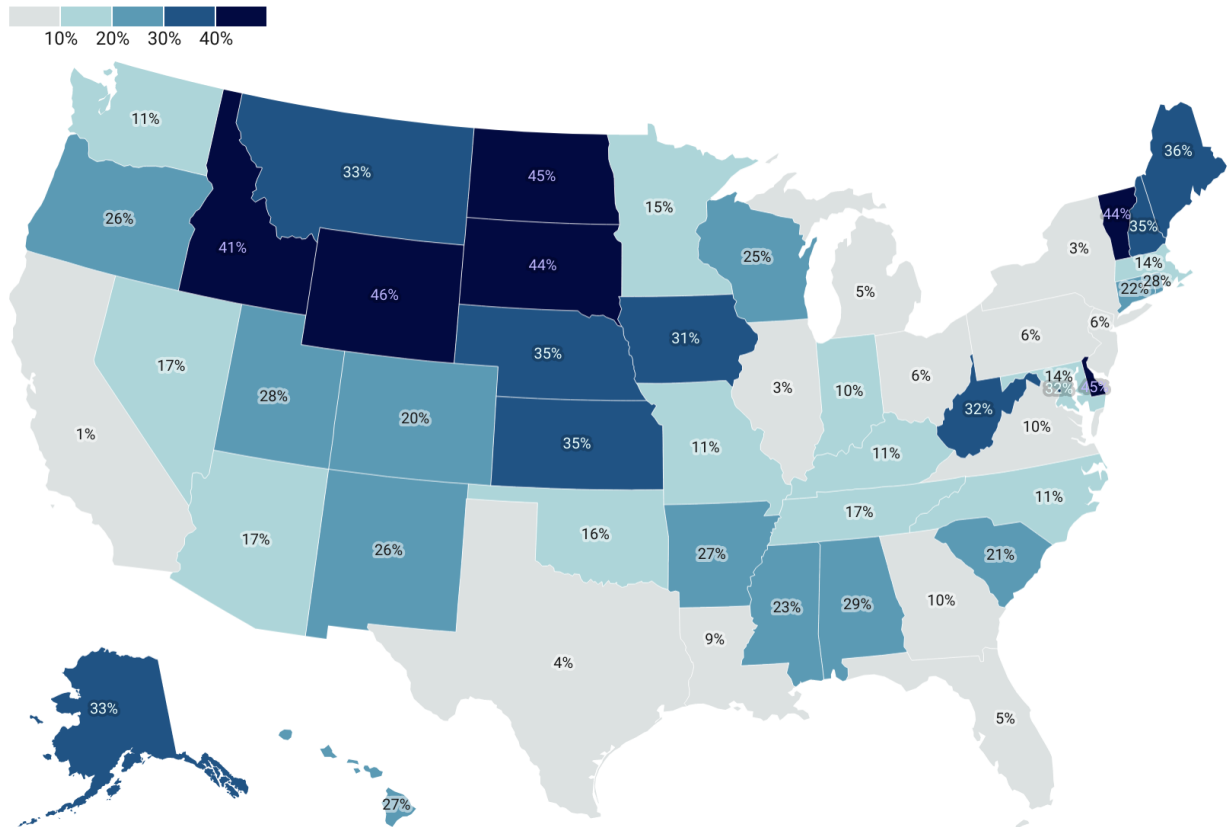


Figure 3.12. Predictive Accuracy of First 6 Digits of SSN By State (1989 – 2011).

When examining median error by state for 1995 – 2011 after Enumeration At Birth becomes more established, there are moderate improvements. While large states such as California and New York still have median errors above 10,000, 24 states had median errors below 1,000. I was able to accurately predict the first 5 digits for 26 states more than 80% of the time and accurately predict the first 6 digits for 19 states more than 30% of the time. Plots for these results are shown in Appendix B.

While more individuals enter the Death Master File each year as they pass away, when running historical simulations of the prediction models trained on subsets of the 2013 Death Master File (DMF) that would only contain individuals who died and entered the dataset from 1999 to 2012, there was a

very fast convergence in predictive accuracy for each birth year cohort towards their current predictive accuracy levels (Figure 3.13). Older cohorts, such as those born in 1990 had much lower predictive accuracy for the first 5 digits since Enumeration At Birth was just beginning. For the 2000 cohort, the additional individuals entering the DMF in 2001 did improve predictive accuracy but it quickly plateaued. There was no significant improvement observed for the 2005 cohort over time, while there was a 6 percentage point improvement for the 2010 cohort when comparing the predictive accuracy of the simulated 2010 DMF versus the simulated 2011 DMF.

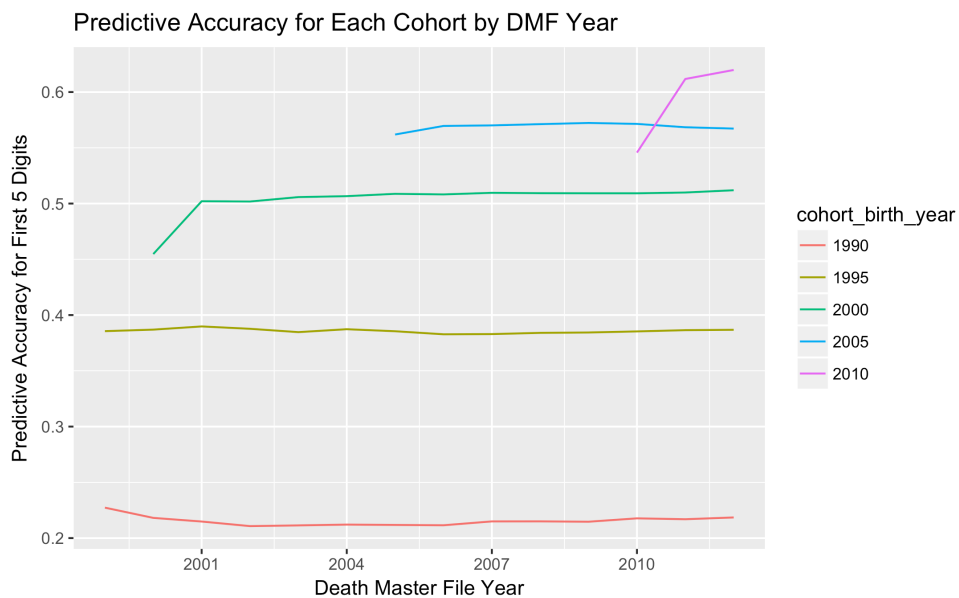


Figure 3.13. Predictive Accuracy for Accurately Predicting the First 5 Digits for Each Birth Year Cohort by DMF Year.

Discussion

The results show the vulnerability in Social Security Numbers as authenticators by uncovering the SSN Assignment Protocol as following the Zang hypothesis of nested loops of Area Number sets, Group Numbers, Area Numbers, and Serial Numbers. This assignment protocol also creates a strong relationship between an individual's SSN and their state of birth and date of birth if they were born

between 1989 and 2011 due to the Enumeration At Birth program. Using loess regression models based on this relationship, I was able to accurately predict the first 5 digits of SSNs 48% of the time and the first 6 digits 11%. There was also significant variation between states with the most vulnerable ones having relatively small populations such as Delaware, Idaho, North Dakota, South Dakota, and Wyoming, where I was able to accurately predict the first 5 digits of the SSNs more 90% of the time. These results raise questions about the how vulnerable to the disruptive effects of identity theft these young people will be when they grow up and enter the financial system and begin employment.

In addition, since the SSN assignment protocol would iterate the fastest through the Serial Numbers, while changing the Area Numbers and Group Numbers (ANGNs) more slowly over time, this means it's harder to accurately predict the last 4 digits for a given date of birth than the first 5 digits. However, it is often common practice to reveal the last 4 digits while obscuring the first 5 digits on forms and documents. For example, according to the IRS, in order "to reduce the risk of identity theft", employers and other taxpayers "may replace the first five digits of the nine-digit number with an asterisk (*) or X on most payee statements" [19]. If the goal is to actually reduce the risk of identity theft, then this practice actually has the opposite effect due to the SSN assignment protocol.

Social Security Numbers are not suitable to be used as authenticators. There are two vulnerabilities that already exist due to their widespread use as identifiers and the many data breaches of datasets that contain SSNs that have made many SSNs accessible to identity thieves. This study demonstrates a third vulnerability in how their assignment protocol effectively encodes the state and date of birth of Americans born over the 22-year period from 1989 to 2011 due to SSA's Enumeration At Birth program. Even the SSA recognized the need to address this vulnerability when it started randomizing all 9 digits for SSNs assigned after 2011, but no solution came for those who were already assigned SSNs [23]. Americans are already wary of sharing their SSNs due to fears of identity

theft [13]. It's time for public policy to address these fears by designing alternatives to SSNs to use as strong authenticators throughout American society.

Potential alternatives to SSNs

Public sector alternatives

In recent years, we've seen examples of the government beginning to limit the use of SSNs as identifiers and authenticators. For example, Medicare originally had SSNs on its ID cards for seniors. In 2015, Medicare started a \$320 million, four year program to issue new Medicare cards that do not show the individual's SSN on them [27]. In 1996, 29 states used the SSN as their driver's license ID number or showed it on their driver's licenses, and there were plans to expand it to all states [17]. Instead, the Intelligence Reform and Terrorism Prevention Act of 2004, "prohibits Federal, State, and local governments from displaying SSNs, or any derivative thereof, on drivers' licenses, motor vehicle registrations, or other identification documents issued by State departments of motor vehicles" [28]. The IRS has spent years working on methods to go beyond relying on SSNs as authenticators on tax returns in order to prevent identity theft tax fraud. Identity thieves would use the SSNs of taxpayers to claim fraudulent tax refunds before the legitimate returns get filed, with estimated costs peaking at \$5.8 billion in 2013 [29]. The IRS has implemented a number of different methods to detect these fraudulent tax returns, and in 2019, the IRS reports that the number of ID theft tax returns has decreased to 13,737 claiming \$184 million in refunds [30]. One of the IRS' response measures is the creation of the Identity Protection PIN (IP PIN) for victims of identity theft to include on their legitimate tax returns. In 2021, the IRS opened up the IP PIN to be available to any taxpayer [31]. In order to sign up, a taxpayer has to verify their address and a financial account number, such as a credit card, student loan, mortgage, or car loan [32]. Then the IRS would send an activation code to

log into the IP PIN website either to a mobile phone number associated with the taxpayer or by mail [32].

Private sector alternatives

In the private sector, we see alternative technologies with device-based authentication, single sign-on solutions, knowledge-based authentication, and verification of driver's licenses and other photo IDs.

Given the ubiquity of mobile phones and smartphones as part of everyday life, device-based authentication methods using SMS or one time pad (OTP) apps have become popular as the second factor for two-factor authentication [33]. In two-factor authentication, a user would often first authenticate with their password and then again with a short code they receive on their phone before being able to log in to a service. Thus, even if an attacker manages to steal or guess the user's password, they still can't log in without knowing the short authentication code. Criticisms of this method include the possibility for a "SIM swapping attack" to allow an attacker to receive the short code on a phone they control and the complicated nature of setting up an OTP app and migrating it to a new phone, which may frustrate less technically savvy users [34, 35].

Single sign-on (SSO) solutions from Facebook, Google, Twitter, Apple, Amazon, and other tech companies can provide identification and authentication to a service provider such as a dating, e-commerce, or gaming app that incorporate them [36]. A user of the service provider would sign-on through their accounts with Facebook, Google, Twitter, or another SSO service, and on the backend the SSO solution would authenticate that user and present their ID string to the service provider [36]. This removes the need for the service provider to build its own user identification and authentication system by setting up usernames, passwords, and possibly two-factor authentication with mobile

phones or security keys. The trade-off is that the user must establish an account with the SSO company first.

Knowledge Based Authentication (KBA) has also become a popular authentication method used by banks and other financial institutions. Credit reporting agencies, Equifax, Experian, and TransUnion, would generate questions for a user to answer based on information in their credit report, such as the names of financial institutions where they have accounts, their former employers, or their old addresses. If a user is able to answer the questions correctly, then they are authenticated, and their credit card or bank account application will be processed. There are multiple criticisms of this method. First, the 2017 Equifax data breach may result in identity thieves being able to answer the KBA questions for 145.5 million Americans [37]. Second, while the data may exist in an individual's credit report, they may not remember the names of old bank accounts, employers, or addresses from years ago, and thus, fail to provide the right answers for KBA [37].

Since 212 million Americans have a state driver's license and even non-drivers can apply for a state ID card, some digital services try to authenticate their users by requesting these state-issued ID documents [38]. For example, online education platform, edx, requires students to take a photo of their ID with their webcam and upload it on their browser for each proctored exam [39]. Some financial technology companies go even further by requiring users to take an "ID Selfie" of their face and their ID in the same photo [40]. A drawback of this method is that not all individuals have a state driver's license or even state ID, especially if they are undocumented [41].

International alternatives

The US can also look to how other countries have tried to create a national ID and authentication system.

One alternative is to use biometrics. For example, India enrolled 1.2 billion Indians into the Aadhar system, which can provide a 12-digit ID number that can be authenticated by the individual's irises or fingerprints [42]. There are criticisms of the accuracy of Aadhar in verifying clouded irises of the elderly or smooth fingerprints of manual laborers [42]. Privacy is another issue in this massive biometric data collection effort by the Indian government. In 2018, the Indian Supreme Court declared that Aadhar is constitutional but its uses need to be restricted to government services only, thus businesses can no longer require it for opening new bank accounts or cellphone plans [43].

Another alternative is to use national ID cards with embedded digital chips for authentication. This technology was used by the Estonian national ID card since 2002 [44]. It uses public-private key cryptography to authenticate Estonians for digital government services [44]. An individual would insert their national ID card into a chip reader, which can then authenticate the card as being genuine and belonging to the individual. A major crisis occurred in November 2017 when security researchers found that a bug in the manufacturing of some national ID cards resulted in generating weak RSA keys that an attacker could guess [45]. As a result, the Estonian government took the drastic step of shutting down access to digital government services such as paying taxes or managing healthcare information for 760,000 Estonians or 58% of the population, who had the vulnerable national ID cards, until they could update the certificates on their cards to fix the problem [45].

Social Security Numbers were never designed to be *de facto* national authenticators, and their reach grew since their launch in 1936 as different government policies pushed government agencies and private firms to integrate SSNs into their processes. This study demonstrates how millions of young Americans born between 1989 and 2011 have vulnerable SSNs that effectively encoded their state of birth and date of birth into their number. We're seeing a variety of alternative authentication

solutions in the public sector, private sector, and internationally. It's time for public policy to address this issue with alternative technologies that are designed to be strong authenticators from the start.

Acknowledgments

I would like to acknowledge the Spring 2016 FRSEMR 42R and GOV 2430 and the Spring 2017 FRSEMR 42R, GOV 1430, and GOV 2430 students at Harvard University, whom I had the privilege of working with as a Teaching Fellow on class projects studying Social Security Numbers using a variety of other methods. You can find details about their projects at <https://aboutmyinfo.org>

- Spring 2016 – FRSEMR 42R Technology to Save the World: Christian Bailey-Burke, Boris Davidov, Brian Lai, Tim Maounis, Sharanya Pulapura, Matthew Li, Ibrahim Syed, Tiffany Yu, and Sarah Zia
- Spring 2016 – GOV 2430 Data Science to Save the World: David Chang, Sarah Chapin, Zack Chauvin, Santiago Fajer Botaya, Vinay Iyengar, Nishant Kakar, Peter Karas, Jacqueline Martinez, Omar Mawloud, Neel Patel, Samuel Plank, Andrew Raftery, Alina Ranjbaran, John Henry Ronan, Elizabeth Rosenblatt, Avi Saraf, Anna Sato, Samantha Udolf, Brandon Wang, Lily Zhang, Yuning Zhang, and Benjamin Zhou
- Spring 2017 – FRSEMR 42R Technology to Save the World: Robert Collins, Louis Delano, Eugenio Donati, Becina Ganther, Harshita Gupta, Carl Sibley, and Hirsh Sisodia
- Spring 2017 – GOV 1430 and GOV 2430 The Politics of Personal Data: Alexandra Abrahams, Stephanie Antonian, Alla Baranovsky, Harold Begg, Holly Breuer, Darcey Carr, Richa Chaturvedi, Martin Chorzempa, Liam Cleary, Avika Dua, Anjali Fernandes, John Gilheany, Priscilla Guo, Ankit Gupta, Gregory Hewett, Emily Houlihan, Anna Liu, Leo Liu, Neel Mehta, Ezinne Nwankwo, Harry Oppenheimer, Thalia Orphee, Matias Rojas, Aizhan Shorman, Kate Steinman, Tayjus Surampudi, Aron Szanto, Alexandra Thaler, Keenan Venuti, Saranya

Vijayakumar, Jessica Wang, Gilbert Wassermann, Elizabeth Yemane, David Yoo, David Wang,
and Sally White

References

1. Social Security Administration. The First Social Security Number and the Lowest Number. <https://www.ssa.gov/history/ssn/firstcard.html>.
2. Social Security Administration. Executive Order 9397 Numbering System for Federal Accounts Relating to Individual Persons. .
3. Malone K and Smith R. XXX-XX-XXXX. NPR. March 14, 2018. <https://www.npr.org/transcripts/593603674>.
4. Puckett C. The Story of the Social Security Number. Social Security Bulletin. Vol 69. No 2. 2009. <https://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html>.
5. Kissel R. Glossary of Key Information Security Terms. NISTR 7289 Revision 2. 2013.
6. Social Security Administration. Social Security History. 2005. <https://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html>.
7. Swendiman K and Lanza E. The Social Security Number: Legal Developments Affecting Its Collection, Disclosure, and Confidentiality. 2014.
8. US Court of Appeals S C. Cassano v. Carb. 2006. <https://casetext.com/case/cassano-v-carb>.
9. Soto G. The Unexpected Costs of Identity Theft. Experian. September 30, 2020. <https://www.experian.com/blogs/ask-experian/what-are-unexpected-costs-of-identity-theft/>.
10. The World Bank. ID4D Practitioner’s Guide (English). 2019.
11. Franceschi-Bicchierai L. Equifax Was Warned. Motherboard. October 26, 2017. <https://www.vice.com/en/article/ne3bv7/equifax-breach-social-security-numbers-researcher-warning>.
12. Sweeney L, Yoo J S, and Zang J. Voter Identity Theft: Submitting Changes to Voter Registrations Online to Disrupt Elections. Technology Science. No 2017090601. 2017. <https://techscience.org/a/2017090601/>.
13. Kim J, Shin H-C, Rosen Z, Kang J, Dykema J, and Muennig P. Trends and Correlates of Consenting to Provide Social Security Numbers: Longitudinal Findings from the General Social Survey (1993–2010). Field Methods. Vol 27. No 4. 348–362. February 23, 2015. <https://doi.org/10.1177/1525822X15572334>.
14. Social Security Administration. Social Security Number Randomization. 2011. <https://www.ssa.gov/employer/randomization.html>.
15. Social Security Administration. The SSN Numbering Scheme.

- <https://www.ssa.gov/history/ssn/geocard.html>.
16. Long W. Social Security numbers issued: A 20-year review. *Social Security Bulletin* 1. Vol 56. No 1. 1993. <https://www.ssa.gov/policy/docs/ssb/v56n1/v56n1p83.pdf>.
 17. Social Security Administration. Report to Congress on Options for Enhancing the Social Security Card. 1997.
 18. Acquisti A and Gross R. Predicting Social Security numbers from public data. *Proceedings of the National Academy of Sciences of the United States of America*. 2009. <https://doi.org/10.1073/pnas.0904891106>.
 19. Internal Revenue Service. Truncated Taxpayer Identification Numbers (TTIN). 2018. <https://www.irs.gov/government-entities/federal-state-local-governments/truncated-taxpayer-identification-numbers>.
 20. Igo S. *The Known Citizen: A History of Privacy in Modern America*. Harvard University Press. 2018.
 21. Morse S. *Decoding Social Security Numbers in One Step*. 2007. <https://stevemorse.org/ssn/ssn.html>.
 22. Social Security Administration. *Social Security Number Allocations*. 2005.
 23. Social Security Administration. *Social Security Number Randomization Frequently Asked Questions*. <https://www.ssa.gov/employer/randomizationfaqs.html>.
 24. Alciere T. *Cancel These Funerals*. 2013. <http://cancelthesefunerals.com/>.
 25. Fox J. *Regression Diagnostics: An Introduction*. SAGE Publications. 1991.
 26. Lee J S and Cox D D. Robust smoothing: Smoothing parameter selection and applications to fluorescence spectroscopy. *Computational Statistics and Data Analysis*. 2010. <https://doi.org/10.1016/j.csda.2009.08.001>.
 27. Pear R. New Cards for Medicare Recipients Will Omit Social Security Numbers. *The New York Times*. April 20, 2015. https://www.nytimes.com/2015/04/21/us/new-law-to-strip-social-security-numbers-from-medicare-cards.html?_r=2.
 28. Social Security Administration. President Bush Signs Public Law 108-458, the Intelligence Reform and Terrorism Prevention Act of 2004. *Social Security Legislative Bulletin*. Vol 108. No 27. 2005. https://www.ssa.gov/legislation/legis_bulletin_010705.html.
 29. US Government Accountability Office. *IDENTITY THEFT AND TAX FRAUD: Enhanced Authentication Could Combat Refund Fraud, but IRS Lacks an Estimate of Costs, Benefits and Risks*. 2015.

30. Treasury Inspector General for Tax Administration. Results of the 2019 Filing Season. 2020.
31. Internal Revenue Service. Get An Identity Protection PIN (IP PIN). 2021. <https://www.irs.gov/identity-theft-fraud-scams/get-an-identity-protection-pin>.
32. Internal Revenue Service. Secure Access: How to Register for Certain Online Self-Help Tools. 2021. <https://www.irs.gov/individuals/secure-access-how-to-register-for-certain-online-self-help-tools>.
33. Duo. Two-Factor Authentication (2FA) from Duo. 2021. <https://duo.com/product/multi-factor-authentication-mfa/two-factor-authentication-2fa>.
34. Barrett B. How to Protect Yourself Against a SIM Swap Attack. Wired. August 19, 2018. <https://www.wired.com/story/sim-swap-attack-defend-phone/>.
35. Kingsley-Hughes A. Still using Google Authenticator? Here's why you should get rid of it today. ZDNet. 2020. <https://www.zdnet.com/article/using-google-authenticator-heres-why-you-should-get-rid-of-it/>.
36. Newman L. Think Twice Before Using Facebook, Google, or Apple to Sign In Everywhere. Wired. September 21, 2020. <https://www.wired.com/story/single-sign-on-facebook-google-apple/>.
37. US Government Accountability Office. DATA PROTECTION: Federal Agencies Need to Strengthen Online Identity Verification Processes. 2019.
38. U.S. Department of Transportation Federal Highway Administration. Federal Highway Administration, Highway Statistics, DL-1C. 2015.
39. Chatigny D. Taking a picture of your photo ID with Software Secure. edX Help Center. 2021. <https://support.edx.org/hc/en-us/articles/360000488887-Taking-a-picture-of-your-photo-ID-with-Software-Secure>.
40. Coinbase. How to Take an ID Selfie. Coinbase Help. 2021. <https://help.coinbase.com/en/coinbase/managing-my-account/verify-my-identity/how-to-take-an-id-selfie->.
41. Johnson K R. Driver's Licenses and Undocumented Immigrants: The Future of Civil Rights Law Symposium: Pursuing Equal Justice in the West. Nevada Law Journal. Vol 5. No 1. 213-239. <https://heinonline.org/HOL/P?h=hein.journals/nevlj5&i=223>.
42. Sudhir K and Sunder S. What Happens When a Billion Identities Are Digitized? Yale Insights. 2020. <https://insights.som.yale.edu/insights/what-happens-when-billion-identities-are-digitized>.
43. Doshi V. India's top court upholds world's largest biometric ID program, within limits. The Washington Post. September 26, 2018.

https://www.washingtonpost.com/world/asia_pacific/indias-top-court-upholds-worlds-largest-biometric-id-program-within-limits/2018/09/26/fe5a95b0-c0ba-11e8-92f2-ac26fda68341_story.html.

44. Parsovs A. Estonian Electronic Identity Card: Security Flaws in Key Management. in *USENIX Security Symposium*. 2020.
45. Cimpanu C. Estonia Cancels 760,000 Electronic ID Cards Because of Crypto Flaw. BleepingComputer. November 4, 2017.
<https://www.bleepingcomputer.com/news/government/estonia-cancels-760-000-electronic-id-cards-because-of-crypto-flaw/>.

Chapter 4

Building A Collocation Detection System Using A Wi-Fi Sensor Array for COVID-19 Contact Tracing in A University Setting

Jinyan Zang and Latanya Sweeney

Highlights

- TraceFi is a Wi-Fi based collocation detection system that uses a sensor array to accurately detect mobile devices within 6 feet of each other for possible use in contact tracing without the burden of requiring a user to install an app in order to participate
- We tested multiple machine learning models in a TraceFi pilot across 12 different spaces in 3 different buildings under regular use conditions and found XGBoost models had a peak sensitivity of 91% and a peak specificity of 86%, with a high median sensitivity of 77% and a high median specificity of 81%
- TraceFi can be used for accurate real-world collocation detection for contact tracing to determine whether 2 devices were within 6 feet for 15 minutes or more and is the first Wi-Fi technology to do so

- We engaged with stakeholders around the university to incorporate their concerns around the ease of adoption, accuracy, cost, and privacy into how we propose including TraceFi data into the contact tracing data flow
- We designed a system for using TraceFi data for contact tracing according to Fair Information Practices that seek to preserve the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive

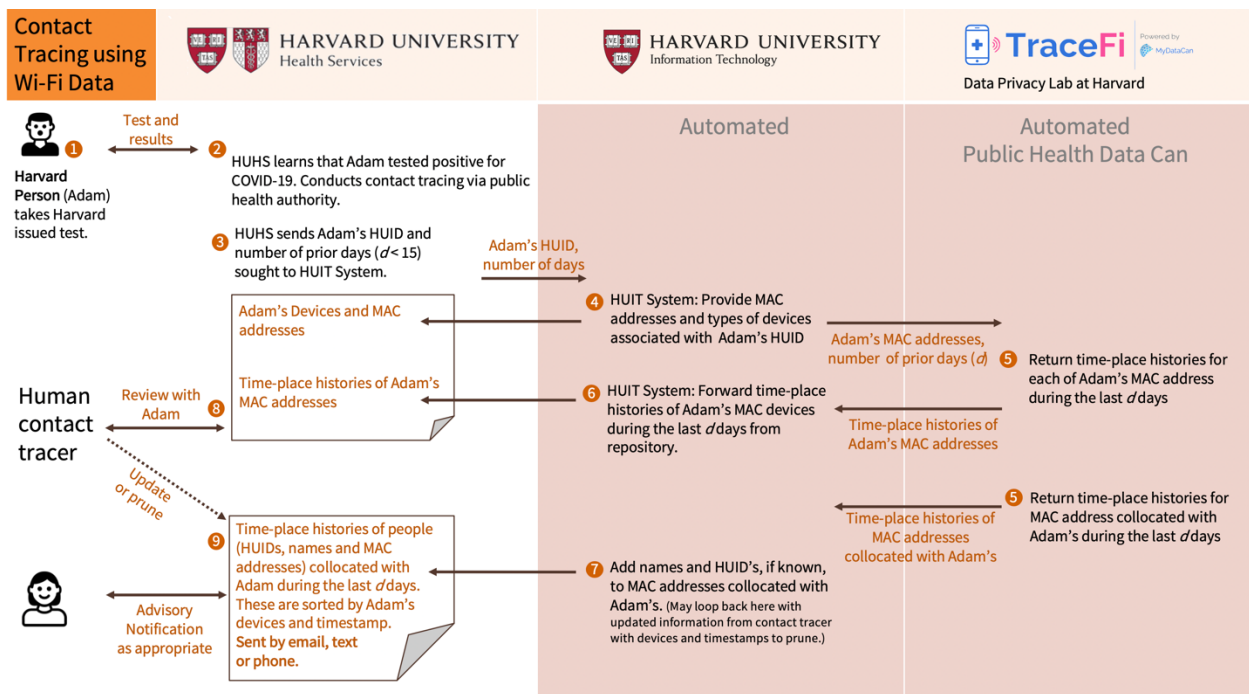


Diagram of Steps for How A Human Contact Tracer Would Be Able to Use TraceFi Data.

Abstract

COVID-19 highlights the risk of exposure from close contact between individuals. The Google-Apple Exposure Notification (GAEN) API, a popular digital contact tracing technology, presents challenges for implementation of contact tracing in a university setting due to its requirement for a decentralized app. Thus, we developed TraceFi, a Wi-Fi based collocation detection system that uses a sensor array to accurately detect mobile devices within 6 feet of each other for use in contact tracing without the burden of requiring a user to install an app. TraceFi builds on existing research and implementations of Wi-Fi based indoor location prediction that uses Received Signal Strength data from Wi-Fi packets to locate a mobile device within a building. TraceFi addresses the propagation of uncertainty problem that arises from using conventional Wi-Fi based location predictions for collocation detection. We tested TraceFi's machine learning models based on Wi-Fi fingerprint data in a pilot study in 12 different spaces in 3 different buildings. We also engaged with stakeholders around the university to incorporate their concerns around the ease of adoption, accuracy, cost, and privacy into how we propose including TraceFi data into the contact tracing data flow.

Results summary: The pilot results found that XGBoost models used in TraceFi had a peak sensitivity of 91% and a peak specificity of 86%, with a high median sensitivity of 77% and a high median specificity of 81% across all 12 spaces. This pilot demonstrates that TraceFi can achieve high accuracy in identifying collocated devices and supports its use in contact tracing systems. Based on stakeholder feedback, we designed a system for using TraceFi data for contact tracing according to Fair Information Practices that seek to preserve the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive.

Introduction

Contact tracing of COVID-19 positive individuals is an effective method of containing the spread of the disease [1]. When a person tests positive to COVID-19, a human contact tracer interviews them and asks questions about places they have been and the people with whom they have had contact. The goal is to identify and notify others who were likely infected so that they can take precautions to not further spread the disease.

Recent research has shown that effective contact tracing may be able to control community outbreaks if more than 70% of contacts can be traced [2]. Another benchmark for effective contact tracing according to *Roadmap to Pandemic Resilience* report at Harvard University published on April 20, 2020 is for exposed contacts of a COVID-19 positive individual to be notified within 12 hours of a positive test result [1]. According to CDC guidelines, a COVID-19 close contact is defined as an individual who is within 6 feet for 15 minutes or more of a person who tested positive; the 15 minutes does not have to be contiguous [3].

Research has found that contact tracers using real-time location data are able to double the number of contacts identified versus self-reported contacts alone [4, 5]. Even a single individual who uses technology to document their location and proximity information can dramatically enhance contact tracing efforts; if they become infected, they can provide detailed recollection. So how can contact tracers leverage technology and data to identify potential close contacts to reach out to within that 12-hour time window?

In 2020, we saw the rise of digital contact tracing technologies that track the location of mobile phones that most people carry with them everywhere. Nations such as Singapore with its TraceTogether mobile app were early adopters. In April 2020, Google and Apple formed a partnership to create the Google-Apple Exposure Notification API or GAEN, which was first released in May [6].

GAEN used Bluetooth signal strength to estimate the distance between two devices. While the technology is innovative and potentially useful for detecting if two devices are collocated (i.e. within 6 feet of each other), GAEN also presented its own challenges for implementation of contact tracing in a university setting. First, GAEN is an API that would need to be incorporated into an app installed by both collocated device owners in order to function properly. This creates an adoption hurdle since GAEN apps would have to be created and promoted in order to increase GAEN's effectiveness in a campus environment. Most GAEN apps were created by governmental authorities, with 20 states in the US representing 45% of the population announcing interest in GAEN by July 2020 [7]. Thus, there's a question of should non-state entities such as a private university even try to build their own GAEN apps that may compete with their local state government's app for public attention. Second, Google and Apple enforced strict criteria on ensuring GAEN apps on the Play Store and the App Store are "decentralized", meaning that private data about one's COVID-19 exposure stays on one's phone rather than in a centralized database accessible for contact tracing [8]. This means that a GAEN app can show an exposure notification alert on the devices owned by someone who was potentially exposed to COVID-19, but a contact tracer can't use GAEN apps to learn who got exposed and reach out to them.

Is it possible to build a digital contact tracing technology that overcomes the operational challenges that GAEN poses to effective contact tracing for use in a university setting?

After working closely with many stakeholders around Harvard University from May to November of 2020, we found there were four primary areas of concern regarding a digital contact tracing technology: ease of adoption, accuracy, cost, and privacy.

- The ideal digital contact tracing technology should be easy to adopt, since requiring users to install and run a contact tracing app on their phone may pose a significant behavioral barrier.

One simulation study found that 56% of a population would need to use contact tracing apps in order to ensure enough opportunities of exposure occur when both collocated devices have the apps running in order to slow down the spread of COVID-19 [9, 10].

- The ideal digital contact tracing technology should also be accurate in determining when two devices are close enough to each other for potential COVID-19 exposure. A false negative occurs when an individual who was a close contact by the CDC guidelines was not identified by the technology as a close contact. A false positive occurs when an individual who was not a close contact but the technology identified the person as being one. There are adverse effects in both types of cases. For false negatives, a potentially infected individual is not notified and thus may spread COVID-19 to others. For false positives, a person may quarantine needlessly or suffer other harms from unnecessary fear of exposure. Over time an unreliable system that creates too many false positives may create apathy among its users due to constantly receiving false alarms. An ideal technology should minimize both false positives and false negatives, though the relative importance of each measurement may depend on the stakeholder.
- The ideal digital contact tracing technology should be low cost to implement. Given the tight budgets brought on by the COVID-19 pandemic, cost is another major concern.
- The ideal digital contact tracing technology should preserve privacy as much as possible. Different stakeholders had varying views on how much privacy should be preserved. For potential data subjects, such as students on campus, having an ability to opt-out of being tracked is highly important. For contact tracers in University Health Services, replicating the decentralized approach of GAEN may be going too far since it would prevent them from learning who got exposed to COVID-19. Thus, we sought to approach the ideal privacy model

of preserving the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive.

Wi-Fi emerged as a potentially promising alternative for a digital contact tracing technology that can address the primary concerns of university stakeholders. The Wi-Fi protocol itself requires all Wi-Fi devices, such as smartphones, to be constantly emitting data, management, or control packets, which can be sensed by any nearby Wi-Fi antennas on the same channel without disturbing the emitting device or its associated Access Point (AP). No changes are needed to the device and no special software needs to be installed. Using the metadata from these packets, such as the Received Signal Strength (RSS), a sensor can predict the distance between itself and the device. When data from multiple sensors are combined, a Wi-Fi based system can predict the location of the emitting device with approximately 6 feet or 2 meters of error under ideal circumstances [11] and with 10 to 16 feet or 3 to 5 meters of error under more real-world conditions using leading vendor solutions [11, 12]. However, this poses a significant propagation of uncertainty problem when simply using Euclidean distance between the predicted locations of a device pair in order to predict if they are collocated within 6 feet of each other or not (Figure 4.1).

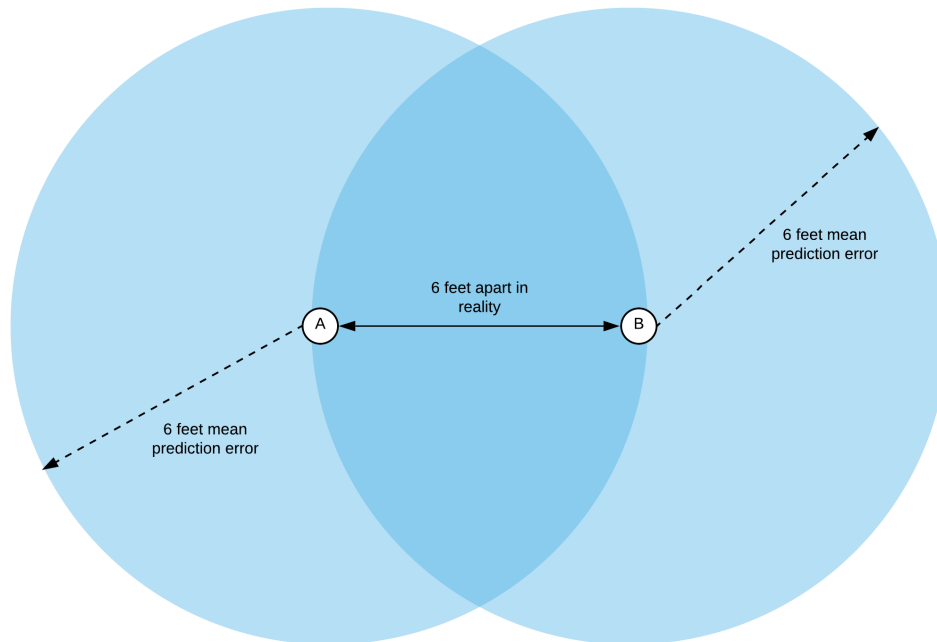


Figure 4.1. The Propagation of Uncertainty Problem with Wi-Fi Collocation Detection Using Predicted Locations. Device A and B are located 6 feet apart in reality. However, even under ideal circumstances, the mean prediction error for device A and B’s locations using Wi-Fi signal strength is 6 feet. Thus, there’s a significant propagation of uncertainty problem in simply using Euclidean between the predicted locations of a device pair in order to predict if they are collocated within 6 feet of each other or not.

We found that it is possible address this propagation of uncertainty problem while using Wi-Fi as a digital contact tracing technology.

Thus, this paper describes how we built a collocation detection system using a Wi-Fi sensor array, which we call TraceFi, in order to target the stakeholder concerns of ease of adoption, accuracy, and cost. This paper will provide an overview of TraceFi and the results of tests we conducted in pilot buildings with more technical details discussed in a separate forthcoming paper in *Technology Science*. This paper also describes how we engaged with stakeholders around the university to create

a contact tracing data flow that incorporates TraceFi data while addressing privacy concerns. Finally, in the Discussion section, we use the four primary concerns described by stakeholders to compare the proposed TraceFi digital contact tracing technology to GAEN, the dominant alternative in the marketplace.

Background

GAEN

How does GAEN work?

Each mobile device is given a unique random ID string which changes every 10 to 20 minutes and is broadcasted over Bluetooth to nearby devices [6]. GAEN apps on each device will collect a log of all ID strings, timestamps, and Bluetooth signal strengths that a device has come into contact with. If at a later date, a potential close contact tested positive for COVID-19, then with confirmation from their local public health authority and consent from the individual, their GAEN app would upload the ID strings associated with their device over the last 14 days to a cloud server. Everyone else's GAEN apps will periodically download from the server a list of ID strings associated with positive cases in their region. If that list contains an ID string that the device has come into contact with in the last 14 days, then the GAEN app would show an exposure notification alert to the device owner that they may be exposed to COVID-19. The GAEN app should also provide information about next steps from their local health authority that the exposed individual can take. Figure 4.2 shows how Google and Apple described GAEN through a scenario with two individuals, Alice and Bob.

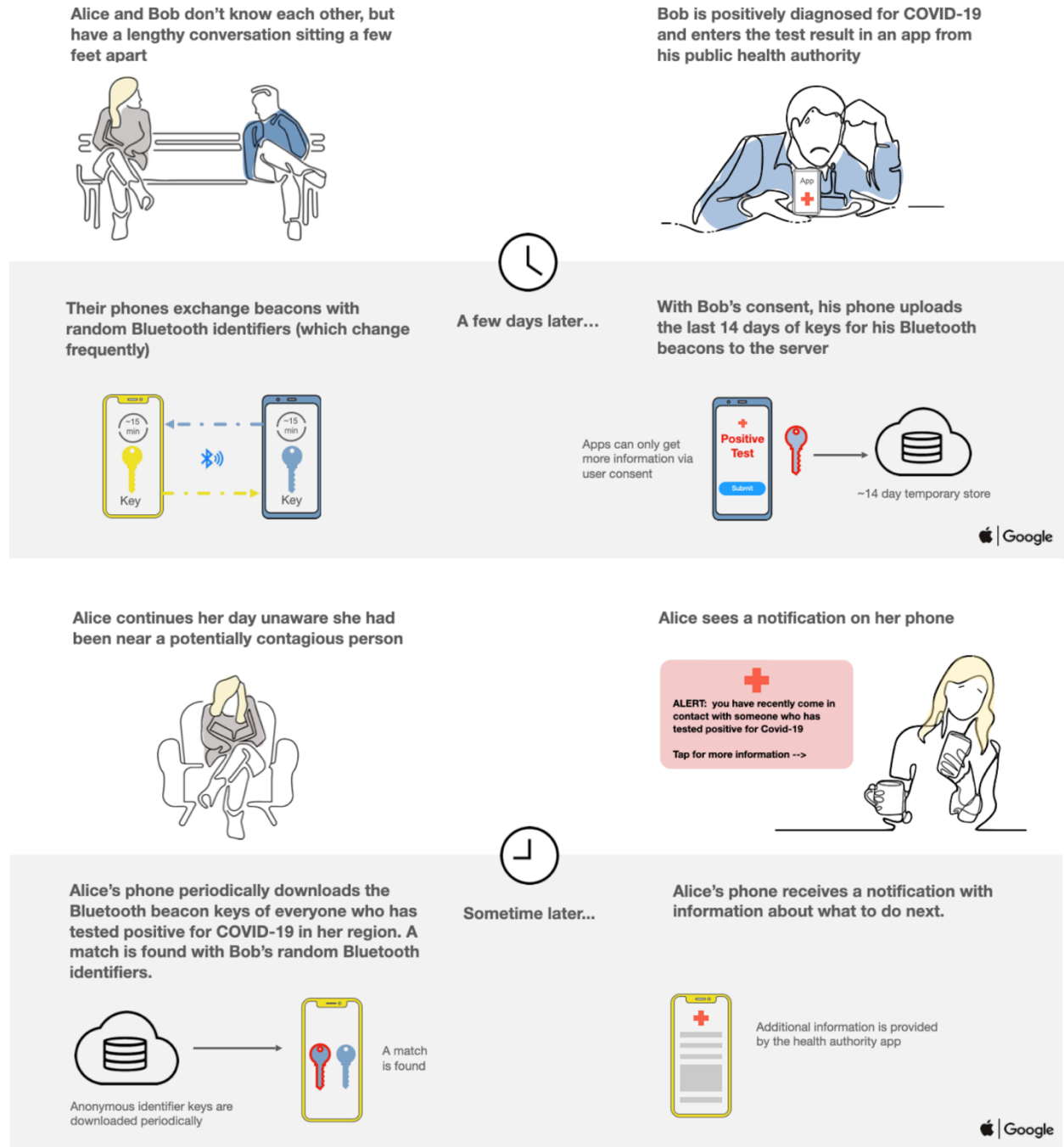


Figure 4.2. An Explanation of How GAEN Apps Work from Google and Apple [13]. When Alice and Bob are in close contact with each other, their phones exchange random identifiers through Bluetooth. A few days later when Bob tests positive for COVID-19, with confirmation from his local public health authority and his consent, his GAEN app would upload the last 14 days of his

identifiers to a server. Alice’s phone would periodically download lists of identifiers associated with positive cases in her region and determine Bob’s identifier as matching the log on her phone. Alice’s phone would then show her an exposure notification alert and information about next steps from her local public health authority.

There have been multiple criticisms of GAEN as a digital contact tracing solution.

First, its reliance on Bluetooth signal strength may not accurately identify if two devices are close contacts according to CDC guidelines of being within 6 feet of each other. GAEN uses the received signal strength indicator (RSSI) as a proxy for how far apart two devices are with higher strength representing closer proximity. However, RSSI can fluctuate due to noise, Bluetooth antenna layout, transmitter strength, battery settings, and other device-related attributes [14]. As a result, Google and Apple have created calibration files for more than 12,000 types of mobile devices, but 97% of them are listed as low confidence of correct calibration [14]. In addition, real world factors such as having a phone in a pocket, having signals bounce off walls, having signals being absorbed by carpets and furniture, and more can significantly reduce the accuracy of collocation detection [14]. Researchers have found that signal strengths can increase instead of decrease as expected when two devices move from 6 feet to 12 feet apart in a tram [15] or in a supermarket with nearby metal shelves [16]. Thus, Google and Apple allow the public health authorities building a GAEN app to set the RSSI thresholds for triggering exposure notification alerts. However, research on the thresholds set by the Swiss and Germany apps had a 100% false negative rate while the Italian app had a 50% false negative rate and a 50% false positive rate [15].

Second, Google and Apple enforced a strict requirement for decentralized exposure notification on GAEN apps approved for the Play Store and the App Store, which meant that contact tracers at public health authorities can’t learn the identity of device owners who received GAEN

exposure notifications [17]. Governments, such as France and the UK, that tried to build more centralized apps were not allowed to use the GAEN API to have their apps run in the background on the device [18, 19, 20]. Instead, their apps would have required the user to constantly have the app running in the foreground and the phone unlocked, which is far more cumbersome [20]. French junior minister for digital affairs, Cédric O, stated, “It is highly abnormal that you are constrained as a democratic state in your technical choice because of the internal policies of two private companies” [21]. GAEN apps that use Bluetooth are also not able to track users with GPS or other location services on the device [22]. According to Google and Apple, enforcing a requirement for decentralization allows them to provide a stronger privacy guarantee to the users of GAEN apps. Users can be reassured that “user identity will not be shared with other users, Apple and Google as part of this process” [22]. Google and Apple also argue that decentralization means that it will be more difficult for governments to conduct surveillance and for data breaches to expose private data en masse [23].

Third, since GAEN apps require users to install and run them on their devices in order to be effective, it imposes a significant adoption burden that public health authorities have to overcome. As of October 3, 2020, only 10 states in the US had a GAEN app, with download rates up to 11% of a state’s population in the case of Virginia [24]. Internationally, Ireland’s COVID Tracker GAEN app has seen one of the highest adoption rates at 35% of the population as of October 2020 [25].

Location prediction with Wi-Fi

Wi-Fi based indoor location prediction is a well-established technology for tracking devices and individuals in commercial spaces such as shopping malls [26, 27, 28]. There are 4 types of data used for Wi-Fi based location prediction: Time of Arrival (ToA), Time Difference of Arrival (TDoA), Angle of Arrival (AoA), and Received Signal Strength (RSS), with RSS being the most popular method since it doesn’t require time synchronization or line of sight between the device and the Access Points [29].

The Received Signal Strength (RSS) is related to the distance between the transmitting device and a receiving Wi-Fi sensor, with stronger signal strength representing closer proximity [29]. Similar to Bluetooth, significant differences in Wi-Fi RSS may be observed for devices of similar distances apart due to environmental, device and other factors.

The use of RSS data, Cell of Origin, trilateration, and fingerprinting are the most common methods, with fingerprinting providing the most accurate location predictions [11, 29].

RSS Fingerprinting usually has 2 phases: an initial phase to collect reference points within the coverage area and a deployment phase to predict the location of an observed device given known fingerprint data at a given location [11].

The TraceFi approach described in this paper focuses on the use of machine learning algorithms with RSS data to make collocation predictions of whether two Wi-Fi devices are within 6 feet of each other. Machine learning models have been used previously to predict the point location of a physical device using RSS fingerprints; common methods include weighted K Nearest Neighbors (wKNN), Decision Trees, Support Vector Machines, Neural Networks, and other machine learning methods [11, 29, 30]. But the prediction errors on these earlier approaches were approximately 6 feet under ideal conditions [11]. More real-world conditions of leading vendor solutions have greater error: 10 to 16 feet or 3 to 5 meters of error with a properly calibrated model due to the density and placement of sensors and the test environment [11, 12]. These prior efforts suggest that using the Euclidean distance of the predicted locations of two Wi-Fi devices to accurately predict their collocation within 6ft of each other would not be reliable due to the propagation of uncertainty problem. As reported later in this writing, the TraceFi approach reliably achieves its goal by focusing on collocation and not point location prediction, using the latest machine learning algorithms and the introduction of a novel invention.

Wi-Fi fingerprints can be collected in one of two ways depending on how one wants to design a location prediction system. One method is for a device-centric system where a device, such as a smartphone, can collect RSS data from nearby APs at each fingerprinting location. Thus, when devices observe similar RSS data in the future, they can use the trained models to predict its location. A second method is for a network-centric system where sensors connected the Wi-Fi network is collecting RSS data from nearby clients at each fingerprinting location to train the location prediction models. The TraceFi pilot described used a network-centric system.

Another Wi-Fi based approach proposed by UMass Amherst researchers uses the connection of multiple devices to the same Access Point (AP) for collocation detection. This provides coarse location and collocation information with a large and significant number of false positives because commonly used commercial APs having ranges of 600 feet [31]. Devices located hundreds of feet apart would be identified as being collocated because they are associated with the same AP. The TraceFi pilot described in this study uses an array of independent Wi-Fi sensors.

Other digital contact tracing technologies

Other digital contact tracing technologies have their own trade-offs such as GPS and ultrasound.

GPS was used in some non-GAEN apps such as Iceland's Rakning C-19 app, which was launched in April 2020. It collected GPS data for possible sharing with a contact tracer and has been downloaded by more than 38% of the Icelanders [32, 33]. GPS location accuracy on smartphones is typically within 16 feet or 4.9 meters outdoors [34], and GPS is readily available on most consumer smartphones. However, GPS location accuracy is much lower indoors, where COVID-19 exposure risk is higher due to less airflow than outdoors, at 19.8 feet or 26.3 feet, or 6 – 8 meters, depending on the building's material [35]. Due to its significant location accuracy errors, using GPS to calculate the

collocation of two devices within 6 feet of each other would be between 2 and 3 times lower than what is required to do reasonable contact tracing [14].

A mobile app could also use ultrasound time-of-flight to measure distance and exchange identification tokens with nearby collocated devices for potential contact tracing. NOVID launched in April 2020 using ultrasound and was deployed at many colleges around the country [36]. When testing under different conditions, the NOVID team found that its technology had a sensitivity of 55% and a specificity of 99.6% based on the CDC guideline of collocation at ≤ 6 feet apart [36]. In a separate contact tracing project on campus, we worked with the NOVID team to incorporate the NOVID app as part of the MyDataCan personal data repository system that we created for the campus community. The details of that project will be described in a future paper. Since NOVID also requires an identification token exchange similar to GAEN apps, this means that both devices would need to have NOVID installed and running in the background when they are collocated.

The university context

There is a clear risk in students, staff, and affiliates living, learning, and working on campus and spending extended periods of time in close contact indoors to spread COVID-19 to each other [37]. At the same time, there is an opportunity to design a technology model as a “local” solution to cover the campus community that can quickly reach a high effectiveness rate for contact tracing without the same barriers faced by more “global” technology solutions at the city, state, or national level that require high rates of adoption in disparate social and technological settings.

We were able to leverage important aspects of the university setting for this project.

First, University Health Services had its own contact tracing team to follow-up on positive cases in the university’s community, and this team was part of the Massachusetts Contact Tracing Collaborative and coordinated with public health authorities. In 2020, the US saw a fairly

decentralized contact tracing effort primarily by state and local governments rather than the federal government [38]. In Massachusetts, the state partnered with Partners in Health to create a coalition of contact tracers working for state and local public health departments as well as other local partners [39]. While many states started rolling out their GAEN apps in 2020, Massachusetts was a relatively late adopter, only beginning tests of the MassNotify GAEN app in April 2021 [40].

Second, the university's Information Technology (IT) department had a robust Wi-Fi network on campus which required all affiliates to register their devices onto the network using their university ID and all guests to provide an email address. This meant that it was possible for the IT department to associate either an affiliate's ID or an email address for a guest to every Wi-Fi device on the campus network.

Methods

From May to November 2020, we engaged in two simultaneous efforts on campus. Effort 1 was the TraceFi pilot, which built and tested a collocation detection system using a sensor array for Wi-Fi based contact tracing in 3 buildings on-campus. An overview of the methods and results will be presented here, with more technical details discussed in a separate forthcoming paper in *Technology Science*. Effort 2 was reaching out and communicating with different stakeholders around the university to incorporate their concerns, especially around privacy, in how we proposed integrating TraceFi data into the contact tracing data flow.

Effort 1. TraceFi pilot

TraceFi's approach

Figure 4.3 describes TraceFi's data flow, which has the following steps:

1. Sensors in the Wi-Fi Sensor Array collect timestamped Received Signal Strength (RSS) data and MAC address about nearby devices.

2. Each sensor regularly sends its collected data to a central repository of “Sensor Data”.
3. Data for the same device from multiple sensors in the “Sensor Data” at around the same timestamp can be input into pre-trained Location Prediction Models, which will predict the location of a device given its sensor data, which may include a sensor array coverage area, a sub-coverage area such as a floor of a building, and/or a specific point-in-space location. The outputs are:
 - a. “Device Location Predictions”
 - b. “Location Predictions” per device pair
 - c. “Sensor Data” per device pair, which is a novel approach of TraceFi to create a large number of additional features per device per device pair for use by the Collocation Detection Models
4. Data about a device pair in “Device Pair Data” can be input into the Collocation Detection Models to predict if the two devices are collocated or not based on the ≤ 6 feet apart CDC guideline to generate the “Device Pair Collocation Data”.
5. A Data User with appropriate access permissions can see and use the “Device Location Predictions” and “Device Pair Collocation Data” from TraceFi for their authorized use case

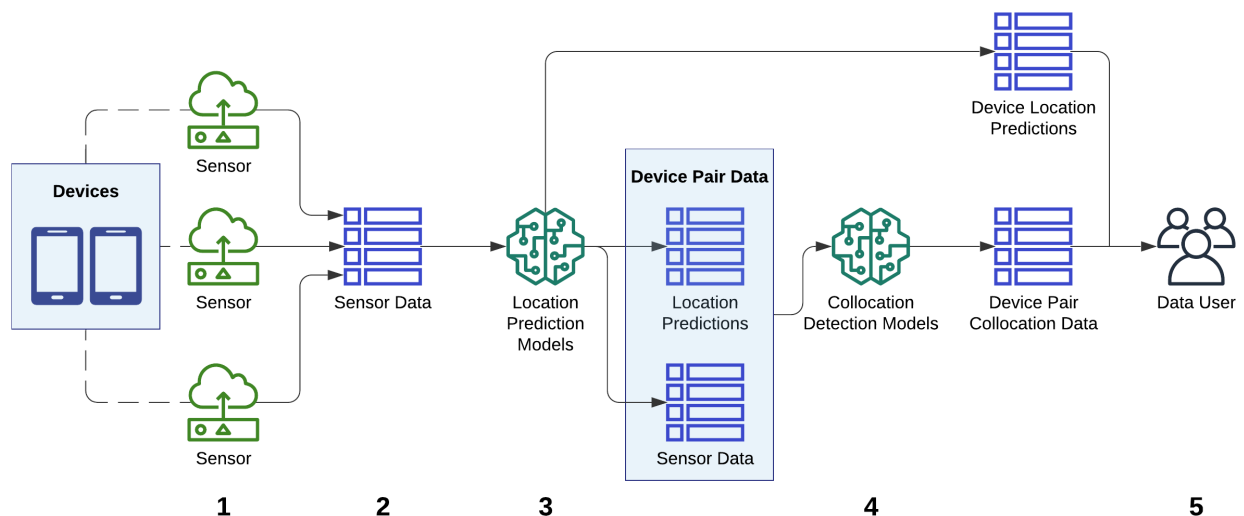


Figure 4.3. Diagram of the TraceFi data flow of a Collocation Detection System using a Sensor Array for contact tracing through 5 steps: (1) sensor array; (2) sensor capture; (3) location prediction; (4) device pair collocation detection; and, (5) authorized user.

Materials and algorithms

A TraceFi pilot was conducted on Harvard University’s campus in July-October 2020 involving 3 buildings. In Buildings A and B, we used a sensor array of Aruba Access Points (APs) which collected the Received Signal Strength (RSS) of Wi-Fi packets from nearby laptops and phones being used for fingerprinting. The RSS data from all APs were accessed through the Aruba Analytics and Location Engine (ALE) API. In Building C, we designed and built our own sensor array using Raspberry Pi 4 computers with dual-band Alfa AWUS036ACS USB Wi-Fi antennas. Every minute, each Raspberry Pi sent the RSS data that it collected on nearby devices to a TraceFi REST API endpoint.

For the TraceFi pilot, we trained and tested models using Euclidean distance, weighted K Nearest Neighbor (wKNN), neural network, random forest, LightGBM [41], XGBoost [42], and soft voting algorithms in order to predict the locations of each device and the collocations of device pairs.

Testing TraceFi’s performance in pilot buildings

The pilot followed the three main steps described below.

Step 1. Collect Wi-Fi fingerprints in pilot buildings

We collected RSS data about known devices at different known locations within the pilot buildings for “Wi-Fi fingerprinting” in order to generate training and test data for Steps 2 and 3. Buildings A and B were lab buildings with a mix of common spaces, offices, and lab spaces across multiple floors. Floors in each building differed by content, configuration and the number and placement of APs. The fingerprinting process consisted of data collection over the course of ~1 week for each building in July 2020. Building C was an IT office building that was unoccupied due to the

pandemic. We configured a conference room on one floor with 8 low-cost Raspberry Pi 4 computers with a dual-band Alfa AWUS036ACS USB Wi-Fi antenna of our design as sensors. We then collected fingerprint data in Building C on September 24, 2020 and October 2, 2020.

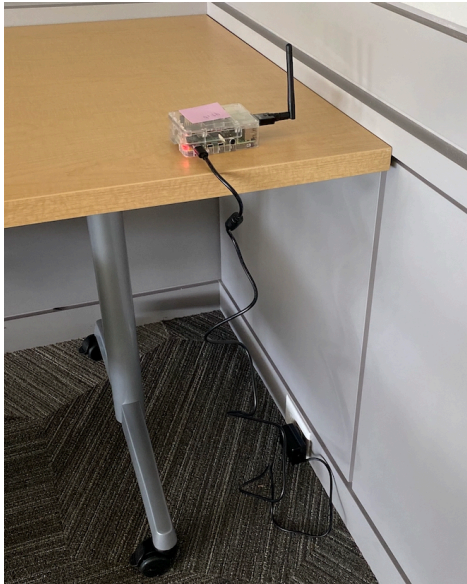


Figure 4.4. Example Raspberry Pi 4 with Alfa AWUS036ACS USB Wi-Fi antenna as a sensor.

The fingerprint process was as follows:

1. We obtained raster versions of the floor plans for each building scaled down to 1 pixel equals 0.2 square feet (Figure 4.5).
2. We marked data collection (i.e. “fingerprinting”) locations on the floor plan by laying out sticky notes in a 4’ x 4’ or 5’ x 5’ grid through spaces accessible to the team, with each sticky note marking the (x, y) pixel coordinates on the relevant floor plan raster map file (Figure 4.6). We also added ~15% additional “test point” fingerprinting locations off the grid to use for testing the performance of our location prediction and collocation detection models.
3. At each fingerprinting location, we placed a cart with a laptop (either a dual-band Macbook or Windows laptop) on the top level and a smartphone (either a dual-band Android Motorola E5

or iPhone 11) on the bottom level and used the Aruba ALE API or the Raspberry Pis to record the RSS data for 2 minutes for Wi-Fi packets sent by the target devices that were observed by the sensor array (Figure 4.7). When the 2 minutes of fingerprinting was over, we moved the cart with the devices to the next fingerprinting location.

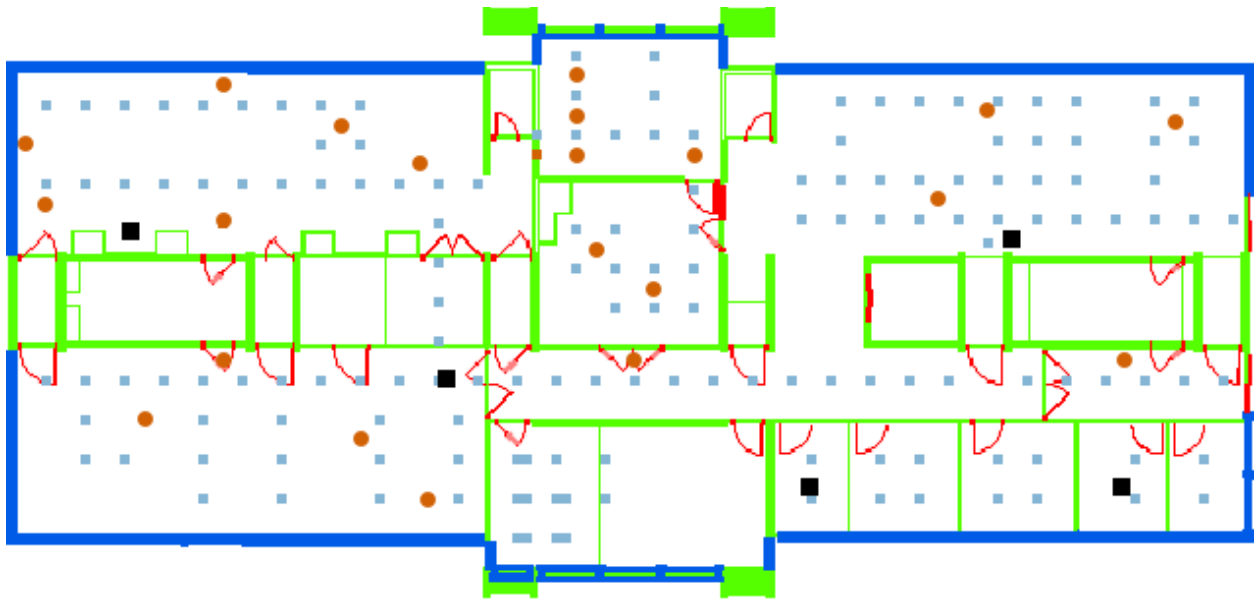


Figure 4.5. Example floor plan of Building A. Blue lines are windows, green lines are walls, and red lines and arcs are doors. Black squares are sensors. Light blue small squares are fingerprints that were training points. Orange circles are fingerprints that were test points.

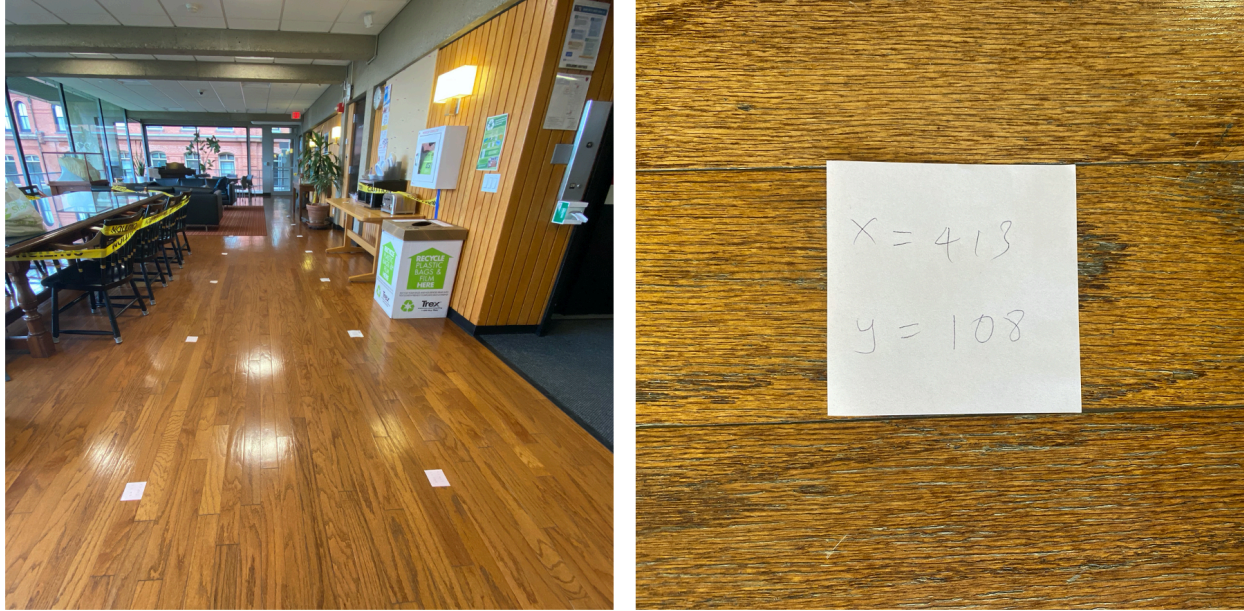


Figure 4.6. Sticky notes for data collection locations in Building A (Left). Each sticky note contained the (X, Y) pixel coordinates of that location on the relevant floor plan raster map file (Right).



Figure 4.7. Carts with 1 laptop and 1 phone on each over a fingerprinting location in Building B.

Step 2. Train and test location prediction models using fingerprinting data

Because our coverage area in the pilot buildings included fingerprint locations within different coverage areas, we found through testing that the approach with the highest predictive accuracy and lowest error is a **two-step approach** where:

1. Predict the coverage area, whether that's a specific floor in the case of Buildings A and B or a specific room in the case of Building C.
2. Predict the point-in-space location, or the (x, y) variables on the relevant floor plan raster map file that represent a device's location in pixel values, using a machine learning regression model based off training points within the predicted coverage area

Step 3. Train and test collocation detection models using fingerprinting data

For the device pair data for collocation detection, we created a dataset of sensor data and predicted locations for each fingerprint pair for each floor. For n fingerprints collected on a floor in Step 1, each floor's dataset had $\binom{n}{2}$ unique permutations of fingerprint pairs. This creates a significant increase in the number of observations for model training for Step 3, since for example, 400 fingerprints on a given floor, a typical number in Building B, would generate 159,600 permutations. We then sampled with replacement the fingerprint pairs that are collocated in order to have balanced collocation classes per floor. Each model we tested had to predict whether the fingerprint pair is within a collocation definition of ≤ 6 feet apart as a classification problem.

We trained and tested a simple Euclidean distance calculation of the predicted x and y , Neural Network, Random Forest, LightGBM and XGBoost models for each floor in each building. We also trained and tested a meta-learner soft voting model that uses the predicted probabilities of Neural Network, LightGBM, and XGBoost models as base learners to generate a consensus prediction.

Testing model performance

We focused on measuring the sensitivity and specificity of each model given their common usage in the medical and public health fields and implications for contact tracing effectiveness.

$$\textit{sensitivity} = \frac{\textit{number of true positives}}{\textit{number of true positives} + \textit{number of false negatives}}$$

$$\textit{specificity} = \frac{\textit{number of true negatives}}{\textit{number of true negatives} + \textit{number of false positives}}$$

Sensitivity measures how many true collocations were detected relative to those being missed as false negatives, which is a priority in the public health context in order to do effective contact tracing to ensure all collocated and exposed individuals are notified. Specificity measures how many true non-collocations were detected relative to false alarms or false positives, which may create distress for an individual who's being unnecessarily notified of their COVID-19 exposure and apathy to exposure notifications in the long run.

Effort 2. Stakeholder outreach


We reached out and communicated with different stakeholders at the university continuously from May to November 2020 in order to solicit feedback and understand their concerns.

- We hosted 13 town halls on Zoom where we explained how the TraceFi pilot worked and how TraceFi could be used by contact tracers, solicited feedback from attendees, and answered their questions. We had 12 town halls for each of the undergraduate dorms on campus where some students were invited back to live on-campus in the fall. We also had a town hall for the researchers and faculty who worked in the TraceFi pilot buildings. Since many of the attendees of these town halls were potential device owners who would be tracked by TraceFi, it became quickly apparent the importance of offering privacy protections such as an accessible opt-out option and clearly demarcating which areas on campus would be monitored by a TraceFi sensor array.

- In each of the pilot buildings, we placed posters describing TraceFi at every location within the coverage areas of the pilot.
- We had daily meetings with a team from the university's IT department to discuss the TraceFi pilot and how the Data Privacy Lab, which would set-up and maintain TraceFi, would work with IT and UHS to integrate TraceFi into a contact tracer's data flow while addressing the concerns around privacy.
- We had regular contact with University Health Services to understand the needs and concerns of contact tracers.
- We had regular contact with university administration and leaders of related efforts on pandemic response to understand the broader needs, resources, and efforts of the university.
- We consulted attorneys at the Office of the General Counsel to ensure all proposals related to TraceFi comply with relevant laws and policies.
- We published information about the TraceFi pilot and other digital contact tracing technologies on campus at <https://covidtech.harvard.edu/>
 - This includes creating an example demonstration of how data can flow between UHS, IT, and TraceFi (Figure 4.8).

Search COVID-19 Positive Individual's Movement History


Next

 **HARVARD UNIVERSITY**
Health Services

Search Movement History

Name

HUID



 **HARVARD UNIVERSITY**
Information Technology

Receives "POST" request from HUHS user

Variable	Value
name	Adam Smith
huid	12345678

Looks up MAC addresses of Wi-Fi devices associated with HUID

device_name	mac_address
Adam_phone	01:23:AC:D2:F1:2C
Adam_laptop	22:12:A1:D4:91:3C

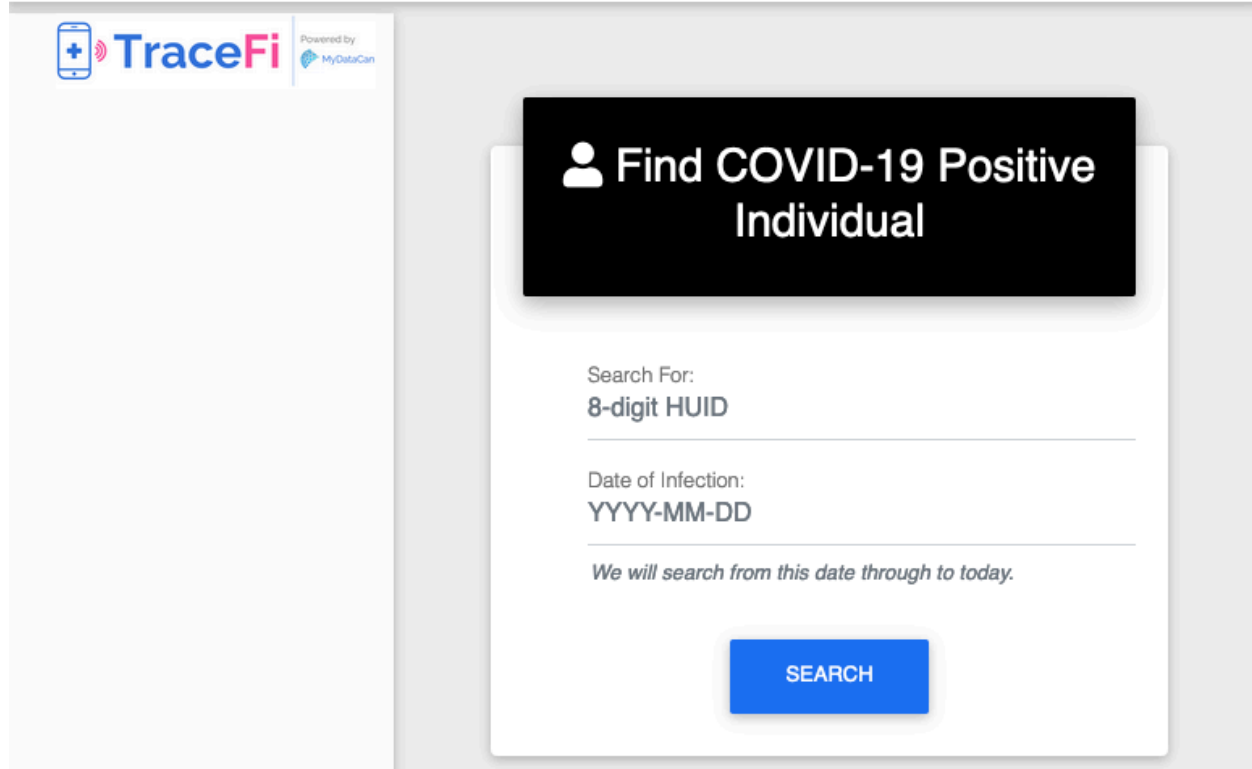
 **TraceFi** | Powered by  MyDataCan

Receives "POST" request from HUIT

Variable	Value
mac_addresses	[01:23:AC:D2:F1:2C, 22:12:A1:D4:91:3C]

Figure 4.8. Screenshot of Example Demonstration of How Data Can Flow Between UHS, IT, and TraceFi on <https://covidtech.harvard.edu/>.

- We partnered with the IT department to create an example TraceFi contact tracer dashboard (Figure 4.9).



TraceFi
Powered by MyDataCan

Find COVID-19 Positive Individual

Search For:
8-digit HUID

Date of Infection:
YYYY-MM-DD

We will search from this date through to today.

SEARCH

Figure 4.9. Screenshot of Example TraceFi Contact Tracer Dashboard on

<https://covidtech.harvard.edu/>.

- We responded to the university newspaper reporting about TraceFi in August 2020 [43] and also wrote a letter to the editor explaining the latest updates on digital contact tracing technologies in September 2020 [44].
- We also considered how TraceFi would integrate with MyDataCan, a personal data repository system that we at the Data Privacy Lab were also creating for the campus community as part of its pandemic response. The details of that project will be described in a future paper. The primary integration details between the two projects are that it was possible to design an option for users to opt-out of TraceFi through their MyDataCan account and also to obtain a copy of their TraceFi data on MyDataCan.

Results

Effort 1. TraceFi pilot

Step 1. Collect Wi-Fi fingerprints in pilot buildings

Over a few weeks in July 2020, we collected 3,127 fingerprints in Buildings A and B and on September 24, 2020 and October 2, 2020, we collected 249 fingerprints in Building C using laptops and phones.

Step 2. Train and test location prediction models using fingerprinting data

We tested 4 machine learning algorithms to build models to classify a device's floor or coverage area within a building based on the RSS values observed by nearby sensors as the first step in our **two-step approach**. XGBoost generally performed well at 93% accuracy in Building A, 98% accuracy in Building B, and 90% accuracy in Building C.

As the second step in the **two-step approach**, we trained 4 machine learning models to predict the (x, y) position of a device on a given floor based on the floor plan raster file. XGBoost models performed well with a Root Mean Squared Error (RMSE) of the actual versus predicted location of a device of 11.4 feet in Building A, 13.7 feet in Building B, and 7.6 feet in Building C.

Step 3. Train and test collocation detection models using fingerprinting data

Figure 4.10 shows the collocation classification sensitivity and specificity by model as a box-and-whisker plot based on the metrics for the 12 coverage areas.

The simplest model of calculating the Euclidean distance between the predicted x and y of each device had a low median sensitivity of 32% and a very high median specificity of 96%. Similarly, the Random Forest models had the lowest median sensitivity of 20% but very high median specificity of 99% due to these models being very pessimistic in favoring predictions of non-collocation for nearly all device pairs.

The Neural Network, LightGBM, and XGBoost models performed better with higher sensitivity rates while maintaining fairly high specificity rates. The Neural Network models had a median sensitivity of 64% but a high median specificity of 90%. The LightGBM models had higher median sensitivity of 72% but lower median specificity of 83%. XGBoost had the highest median sensitivity of 77% and a nearly similar median specificity of 81%. It had the highest peak sensitivity of 91% and peak specificity of 86%. Finally, the soft voting models that used weighted probabilities from Neural Network, LightGBM, and XGBoost models had a lower median sensitivity of 69% than that of the best base estimator, the XGBoost models, but also high median specificity of 89%, which is similar to that of the best base estimator, the Neural Network models.

The sensitivity-specificity trade-off is clearly displayed. For public health contact tracing purposes, high sensitivity and thus XGBoost models may have the highest utility, but in a different context, other machine learning models may be best.

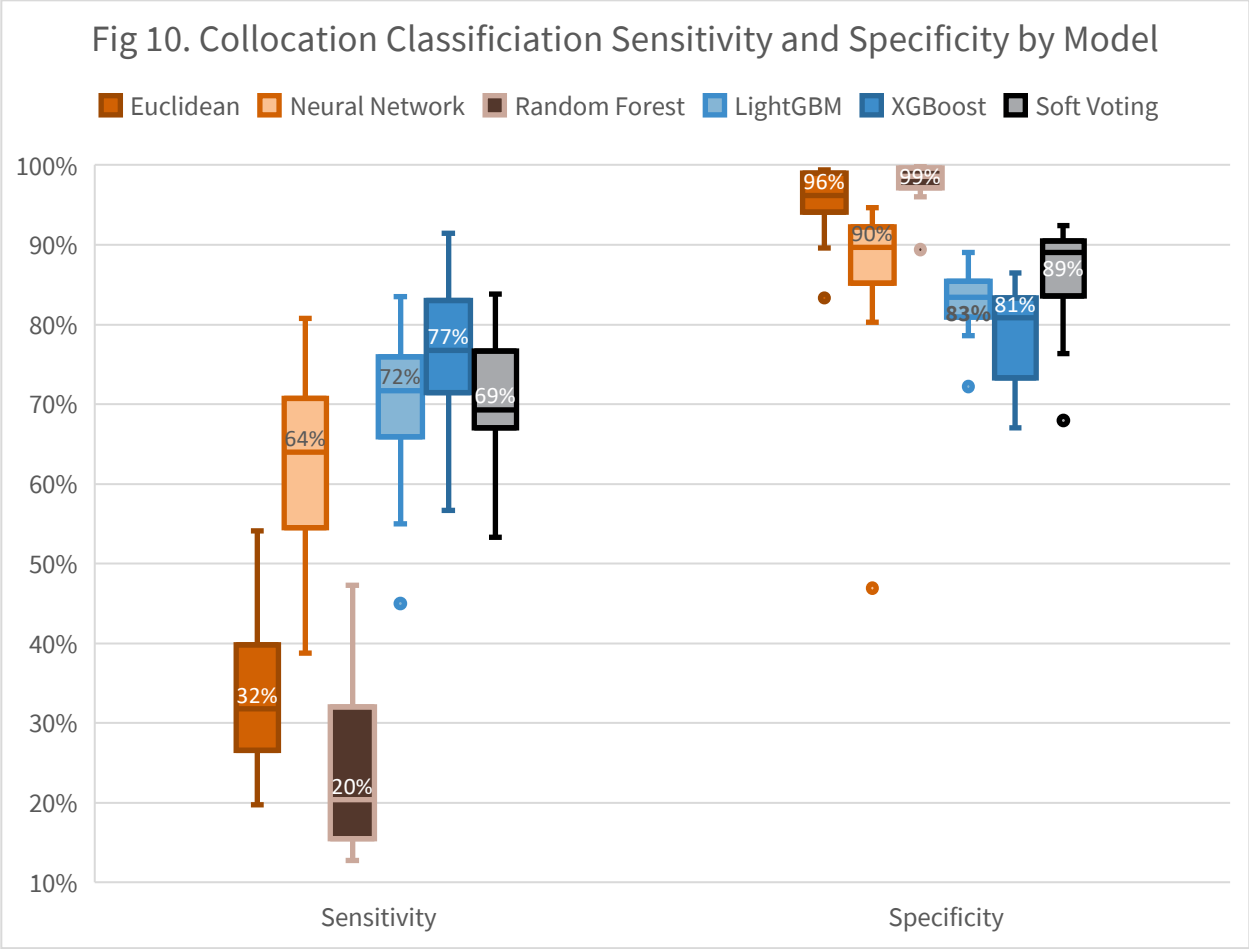


Figure 4.10. Collocation Classification Sensitivity and Specificity by Model for All 12 Floors in Pilot Buildings. This is a box-and-whisker plot that shows the whiskers extending to data points within 1.5 times of the interquartile range and outliers are drawn as additional points.

We also analyzed if the number of sensors that can observe the RSS of each device in a device pair has an impact on the XGBoost collocation prediction models. For sensitivity, it appears that there needs to be at least 3 sensors that can observe each device in a device pair to have a non-zero sensitivity rate of 65% which quickly approaches 76% when there are 5 sensor observations (Appendix A). For specificity, having a mean of 2 sensors observe each device in a device pair resulted in a

specificity of 70% which quickly approaches near a maximum of 80% when there is a mean of 4 sensors observing each device in a device pair (Appendix A).

Effort 2. Stakeholder outreach

Based on the input from our stakeholder outreach, we describe below the proposed flow of data from the TraceFi sensor array to the human contact tracer at University Health Services.

- The TraceFi sensor array captures the signal strengths emitted from mobile devices at a particular time and records the internal MAC addresses of the mobile devices that sent the signals. A MAC address is a unique code assigned to each mobile device by the manufacturer of the device.
- The information from the TraceFi sensor array does not contain a person's name or Harvard ID or any explicit personal identifier.
- The TraceFi sensor data flows into "Raw," a secure data storage container that resides on a system that applies Level 4 (high risk information) protection under Harvard's information security policy. No backups or copies are made of the data. Data are kept no longer than 14 days. On the 15th day from collection, the data are no longer available.
- The TraceFi algorithms read data from the Raw storage, compute proximity and collocations, and write the derived information into "Processed," a secure data storage container that also resides system secured to Level 4 requirements. No backups or copies are made of the data. Data are kept no longer than 14 days. On the 15th day from collection, the data are no longer available. Access to both the Raw and Processed storage containers is locked down and secured. The TraceFi algorithms are the only reader from the Raw storage and the only writer to the Processed storage.

- Harvard University Information Technology has produced a "Dashboard" for a human contact tracer at Harvard University Health Services to use when interviewing patients who have tested positive for and/or have been diagnosed with COVID-19. The human contact tracer enters the person's HUID into the Dashboard, which automatically fetches information from the Processed storage at TraceFi based on the MAC addresses of the person's known devices. The Dashboard then displays the proximity information for the person's devices and any collocations registered by the system.
- The two readers of TraceFi's Processed storage are the Dashboard of the contact tracer and an automated regularly timed function that allows individuals to get their TraceFi data copied into private storage on MyDataCan, a data management and apps platform Harvard has provided for members of the Harvard community. Each access to the Processed storage records in a one-way immutable log that is not readable by TraceFi or the members of the Data Privacy Lab responsible for the operation of TraceFi. The immutable log allows external review to confirm that each instance of access to TraceFi data relates to a specific COVID-19 diagnosis or positive test result.
- Only a human contact tracer at Harvard University Health Services can use the Dashboard to access information from TraceFi. The human contact tracer can receive proximity information relevant to a positive test case and collocation information relative to a positive test case. The retrieved information is sufficient to review the information with the infected person and to notify collocated people as needed. Details are covered under medical and public health confidentiality.

- Security tests and audits are done regularly. Data access is reviewed monthly and reported to the University Electronic Communications Policy Oversight Committee, which prepares public reports of aggregated information.
- Members of the community can opt out of the system, using their MyDataCan dashboards: the TraceFi sensors will not accept or process any information from the MAC addresses of people who have opted out. The TraceFi sensors will also not receive information from any device that has Wi-Fi turned off.

Here is how the proposed approach would work in practice. The numbered steps correlate to the numbers in Figure 4.11.

1. Adam is a Harvard person who tested positive to COVID-19.
2. Adam's test result forwards to a contact tracer at Harvard University Health Services.
3. The human contact tracer enters Adam's name and HUID into the Dashboard that Harvard University Information Technology provided to interface with TraceFi.
4. The Dashboard operation looks up all MAC addresses registered on Harvard's network for Adam's HUID and requests information about those devices from the TraceFi Processed repository.
5. The TraceFi system sends back date-time locations for Adam's MAC addresses. It also sends a list of other MAC addresses of devices that were collocated with Adam's.
6. The human contact tracer can then view Adam's devices on the Dashboard.
7. The human contact tracer can also view the places Adam went on campus and people whose devices were collocated with Adam's, but only those places that are within designated operational areas of TraceFi sensors.

8. The human contact tracer works with Adam to review places and encounters with people. This process may take some time.
9. The human contact tracer will notify people deemed to be at risk to infection.

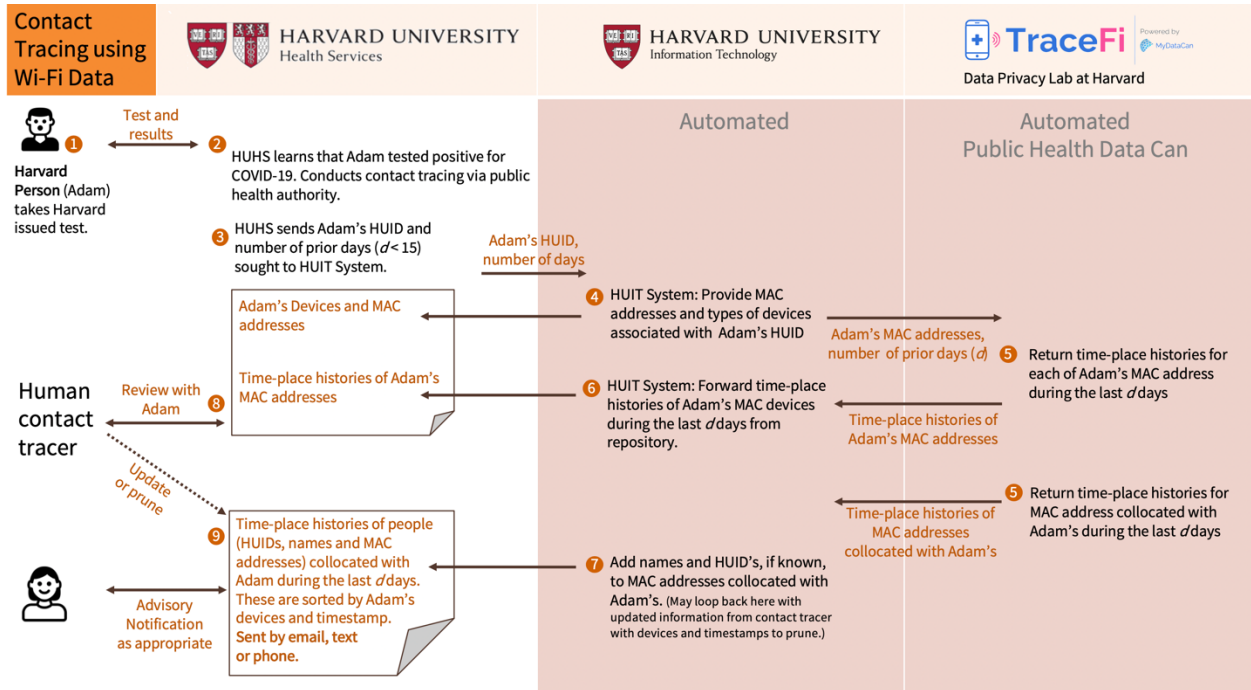


Figure 4.11. Diagram of Steps for How A Human Contact Tracer Would Be Able to Use TraceFi Data.

Importantly, there is a privacy wall between the TraceFi System maintained by the Data Privacy Lab and Harvard University Health Services (HUHS) and Harvard University Information Technology. The TraceFi side of the privacy wall has location information with no identity. The Harvard University Health Services and Harvard University Information Technology side has identity but does not have open access to location information. The only disclosures of location information outside of the TraceFi System at the Data Privacy Lab are to human contact tracers at HUHS, through the Dashboard and, potentially, by automated means to collocated persons who have not opted out of TraceFi.

We designed TraceFi according to the following practices related to the Fair Information Practices and GDPR principles.

- **The minimum collection practice:** collect only what is necessary to accomplish the task.
 - TraceFi sensors only capture signal strength and MAC address and record the date and time of the recording. These three pieces of information (signal strength, MAC address, and date-timestamp) are the only information collected. These are the minimum information needed to help the human contact tracer.
- **The minimum sharing practice:** shared information is the minimal information needed to accomplish the task.
 - The only information shared from TraceFi is with a human contact tracer in the service of a specific positive tested person. TraceFi only provides the date and time, duration, location, if known, and collocations for each of the person's devices. The human contact tracer then reviews exactly this information with the positive tested person to make a determination about who may have been infected.
- **The limited time practice:** keep personal data no longer than is needed to accomplish the task.
 - Contact tracers need only go back for the last 14 days, so only the last 14 days of information is kept in TraceFi and only the last 14 days of information is made available to contact tracers.
- **The access practice:** the person who is the subject of the data should have a copy of their data.
 - A person can receive a copy of their own TraceFi data in their private storage on MyDataCan. This option is available on MyDataCan. This copy is for the person's own

use and is independent of the copy maintained in TraceFi for up to 14 days. The person's copy in MyDataCan is not subject to the 14 day limit.

- MyDataCan is architected around the principle that the person who is the subject of the data is in control of a copy of their data. Each app and service on the MyDataCan platform therefore stores a copy of the person's location information into the person's own private storage on MyDataCan. This copy is under the control of the person, who can opt to share their information with human contact tracers.
- **The accuracy practice:** individuals should be able to make corrections to their own data.
 - The human contact tracer receives proximity and collocation information from TraceFi specific to a case of a positive tested person and for the purpose of reviewing the retrieved information with the positive tested person, who can attest to its accuracy and whose attestations are the basis for the determinations of likely infections to others.
- **The accountability and transparency practice:** the ability to audit any accesses made to the data.
 - Each access to TraceFi data is recorded in an immutable log that is not accessible to the Data Privacy Lab which provides it, but whose contents are available to others for monthly review and audit and reported to the University Electronic Communications Policy Oversight Committee which makes public summaries.
- **The consent practice:** individual participation is optional.
 - A person can opt-out of TraceFi and the TraceFi sensors will not capture information from the person's devices and the person can still use Harvard's Wi-Fi. A person can

also turn Wi-Fi off on their devices (but then those devices would not be able to use Harvard's Wi-Fi).

- **The security practice:** technologies comply with stringent security practices.
 - Both MyDataCan and TraceFi store and process information compliant to Harvard's highest online security level (Level 4). In addition, both systems have routine security audits and tests.

Discussion

The pilot results demonstrate that TraceFi performs very well in multiple buildings with both Aruba APs and Raspberry Pis as the sensor array for collocation detection of device pairs in accordance to the CDC's COVID-19 guidelines of less than 6 feet apart. XGBoost models had a peak sensitivity of 91% and a peak specificity of 86%, with a median sensitivity of 77% and median specificity of 81% across all the diverse floor settings.

A variety of factors likely contributed to the variation in model performance such as a building's material, the location of the sensors, the number and type of barrier objects between each sensor and the device pair, and more. Given the fast-developing field of machine learning and the limits of the pilot study conducted under pandemic conditions, future studies can further examine methods to improve model performance by both optimizing the machine learning models and the physical environment for the sensor array.

Comparing TraceFi and GAEN for contact tracing

We found through stakeholder outreach that there are four primary areas of concern regarding contact tracing: ease of adoption, accuracy, cost, and privacy. Since GAEN apps are the dominant alternative option in the marketplace, how do TraceFi and GAEN compare for each area of concern?

Ease of adoption

Since TraceFi uses a sensor array for collocation detection of two devices rather than using apps installed and running on the devices themselves, it's likely easier for users to adopt than GAEN apps. No changes are needed to the device and no special software needs to be installed. The TraceFi sensor array is using signal strength data that is already being sent out by the standard Wi-Fi protocol. The ease of adoption of TraceFi is important since research simulations have found that 56% of a population would need to adopt digital contact tracing technologies in order to reduce the spread of COVID-19 [9, 10].

Accuracy

The accuracy of TraceFi and GAEN are both mixed.

TraceFi is an innovative technology that tries to address the propagation of uncertainty problem of using traditional Wi-Fi location prediction for collocation detection by training models to evaluate collocation detection as a classification problem instead. We found that XGBoost models had a peak sensitivity of 91% and a peak specificity of 86%, with a median sensitivity of 77% and median specificity of 81% across all the diverse floor settings. However, there were still significant variation in performance between the environments of the 3 pilot buildings. Future studies can examine methods to improve model performance by both optimizing the machine learning models and the physical environment for the sensor array. Appendix B also describes how online learning after TraceFi's deployment can potentially help refine the the location prediction and collocation detection models trained on fingerprinting data.

GAEN apps have been shown to have mixed real-world results as well due to how variation in the characteristics of devices, their positions, and the environment can significantly reduce the accuracy of collocation detection with Bluetooth [14]. Research on the signal strength thresholds for

collocation detection set by the Swiss and Germany GAEN apps had a 100% false negative rate while the Italian app had a 50% false negative rate and a 50% false positive rate [15].

One potential differentiator between TraceFi and GAEN in terms of accuracy is the use of fingerprinting to train models for each sensor array with TraceFi. This means that TraceFi can theoretically be optimized for each environment where a sensor array is deployed. The trade-off is that TraceFi models optimized for one environment may not necessarily translate to another one. GAEN apps are trying to function in any environment where two phones can exchange identifiers over Bluetooth, which is a far more ambitious goal with a trade-off of not being able to optimize according to where the two phones are located.

Cost

TraceFi will likely have higher initial deployment costs in terms of setting up the sensor arrays and doing the fingerprinting, but GAEN apps have their own costs related to promoting adoption by users on campus.

We found in the TraceFi pilot that using low cost \$35 Raspberry Pis for the sensor array in Building C resulted in similar performance as using high cost >\$500 Aruba APs for the sensory array in Buildings A and B. During the pilot, we tested TraceFi's performance while having more than 2,500 sq. ft. of floor coverage per sensor. Nevertheless, there's likely to be significant initial costs to do the Wi-Fi fingerprinting for each TraceFi sensor array. In the pilot, we took a fingerprint reading for 2 minutes approximately every 5 sq. ft, though skipping locations blocked by walls or furniture. Since TraceFi sensor arrays can be deployed independently of one another, this means it's possible to spread out the deployment cost over time as new coverage areas get added.

GAEN apps do not require the same physical infrastructure-related costs as TraceFi, but they may require significant promotion costs in order to encourage adequate levels of adoption. In the US,

state-sponsored GAEN apps have not seen widespread adoption, with Virginia being a leader at 21% as of December 2020 [45].

Privacy

TraceFi and GAEN approach the privacy concern with different goals. With TraceFi, contact tracers are one of the core stakeholders. Thus, we sought to approach the ideal privacy model of preserving the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive. With GAEN apps, Google and Apple's requirement for decentralization means that only exposed individuals would receive exposure notifications and that information is kept private or locally on the device. Thus, contact tracers can't immediately learn the identity of who got exposed through GAEN apps.

In order for TraceFi to achieve its privacy goal, we propose a privacy wall between the TraceFi System maintained by the Data Privacy Lab and Harvard University Health Services (HUHS) and Harvard University Information Technology (HUIT). The TraceFi side of the privacy wall has location information with no identity. The HUHS and HUIT side has identity but does not have open access to location information. The only disclosures of location information outside of the TraceFi System at the Data Privacy Lab are to human contact tracers at HUHS, through the Dashboard and, potentially, by automated means to collocated persons who have not opted out of TraceFi. The privacy wall has the additional benefit of ensuring that the Data Privacy Lab, which maintains TraceFi, doesn't learn any medical information such as which devices belong to owners who tested positive for COVID-19, which resolves any ambiguity with regards to the Health Insurance Portability and Accountability Act (HIPAA). Legal scholars have pointed out how digital contact tracing apps such as the GAEN apps "fall completely outside HIPAA's parameters" [46]. In May 2020, the COVID-19 Consumer Data Protection

Act was introduced though not yet passed in the Senate to address how digital contact tracing technologies should collect and share data [47].

Another differentiator is that we proposed for TraceFi to be opt-out while GAEN apps are opt-in by nature since a user has to choose to install and run them in the first place. With TraceFi, a user can opt out using their MyDataCan dashboards: the TraceFi sensors will not accept or process any information from the MAC addresses of people who have opted out. Similar to what occurred in the pilot buildings, there would also be prominent posters indicating when an area is covered by a TraceFi sensor array. The TraceFi sensors would also not receive information from any device that has Wi-Fi turned off. Since a big advantage of TraceFi over GAEN apps is its ease of adoption, which is highly important for reaching a high enough population adoption rate to reduce the spread of COVID-19, making TraceFi opt-in would significantly reduce that advantage. Stanford professors Michelle Mello and Jason Wang have argued that the ethics of controlling the COVID-19 pandemic supports making contact tracing technologies opt-out [48]. They cite how studies of electronic health record sharing have found that people tend to stick with the default choice, with only 2 to 5% opting out of health information exchange [48]. In the case of contact tracing apps, some surveys found up to 70% of US respondents report that they will probably or definitely use a contact tracing app, which is far more than the actual opt-in rates that's been observed [48]. Mello and Wang argue that contact tracing technologies offer users reciprocal benefits: notification if they come into contact with someone dangerous and assistance in protecting friends and family whom they may have endangered [48]. Thus, providing an opt out would allow “those with strong preferences to act on them while not conflating philosophical objections with simple inertia” [48].

We designed TraceFi to follow the privacy practices described by the Fair Information Practices and GDPR principles, which sought to minimize the collection, sharing, and storage of data

to what's necessary for the purposes of contact tracing. A person can also receive a copy of their own TraceFi data in their private storage on MyDataCan. Data will also be stored according to the university's highest security level (Level 4). Finally, to ensure accountability and transparency, there will be an immutable log that records each access of TraceFi data which is audited and reported to the University Electronic Communications Policy Oversight Committee which makes public summaries. Thus, if these practices were violated in some way, the immutable log can serve as evidence of the violation. With regards to GAEN, Bradford, Aboy, and Lidell argue that GAEN apps fall "within the governance system of the GDPR and ... can be operated in a way that is compatible with the GDPR rules" [46], and many EU nations have built and deployed their own GAEN apps. However, in February 2021, researchers at AppCensus found implementation flaws with how the GAEN API stores its logs on an Android device that may have exposed COVID-19 status, social graph data, and location trails collected by a GAEN app to other third-party apps on the phone [49]. After not addressing the bug report for more than two months, a lawsuit was filed against Google over this issue by two affected users in April 2021 [50].

Conclusion

The university did not end up choosing to implement TraceFi on campus beyond the pilot study. There was a strong set of other pandemic response measures that were implemented to support the re-opening of campus in Fall 2020. For example, individuals on-campus had to attest daily to whether they have potential COVID-19 symptoms on the Crimson Clear web app. There was a vigorous self-administered testing regime that required individuals on-campus to get tested for COVID-19 every three days. Social distancing and mask wearing requirements were in place throughout campus. Only a limited number of students and staff were invited back on campus for the fall, and most classes were virtual.

We demonstrated with the TraceFi pilot that Wi-Fi can be a promising collocation detection technology for COVID-19 contact tracing. We also described how we engaged with stakeholders around the university to incorporate their input in how we proposed including TraceFi into the contact tracing data flow while protecting privacy. TraceFi presents a unique opportunity for a close-knit community such as a university campus, an office park, a cruise ship, a summer camp, and more to build an easily adoptable solution to support contact tracing within their community.

Acknowledgements

The authors extend a tremendous amount of gratitude to the Harvard community for its patience and support during the pandemic while the authors designed and assessed this technology. The authors thank members of Harvard University Information Technology who helped to fingerprint buildings and construct end-to-end systems that used the TraceFi technology: special thanks to Jefferson Burson, Matthew Mazer, David LaPorte, Christopher Curreri, Luke Sullivan, Juliana DiLuca, and James Nelson. The authors also thank Maria Francesconi and Soheyla Gharib at Harvard University Health Services for feedback and discussion on contact tracing, and Brad Frank and Pascal Delpe-Brice in the Data Privacy Lab for support.

The authors give special recognition to the hardworking undergraduate summer research fellows of the Data Privacy Lab who worked on this project: Laurel Carpenter, Juan Guzman, Atuganile Jimmy, and Soheil Sadabadi. The authors want to especially recognize Laurel Carpenter for her tremendous contributions to the TraceFi data collection, processing, and output systems, which allowed the project to move forward on an extremely tight timeline.

Finally, the authors thank the school leadership for the opportunity to do this work: special thanks to Katie Lapp, Giang Nguyen, Claudine Gay, Bradley Abruzzi, Anne Margulies, Christopher Stubbs, Mark Fishman and Alan Garber. The authors also thank Maria-Gabriella Di Benedetto and Luca

De Nardis who, in the spirit of worldwide collaboration during the pandemic, openly shared their notes and codebase from their earlier approaches.

This work was supported in part by the National Science Foundation Grant 1730326.

References

1. Edmond J. Safra Center for Ethics. Roadmap to Pandemic Resilience. 2020.
2. Hellewell J et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*. Vol 8. No 4. e488–e496. April 1, 2020. [https://doi.org/10.1016/S2214-109X\(20\)30074-7](https://doi.org/10.1016/S2214-109X(20)30074-7).
3. CDC. Public Health Guidance for Community-Related Exposure. 2020. <https://www.cdc.gov/coronavirus/2019-ncov/php/public-health-recommendations.html>.
4. Hellmich T R et al. Contact tracing with a real-time location system: A case study of increasing relative effectiveness in an emergency department. *American Journal of Infection Control*. Vol 45. No 12. 1308–1311. 2017. <https://doi.org/https://doi.org/10.1016/j.ajic.2017.08.014>.
5. Ho H J, Zhang Z X, Huang Z, Aung A H, Lim W-Y, and Chow A. Use of a Real-Time Locating System for Contact Tracing of Health Care Workers During the COVID-19 Pandemic at an Infectious Disease Center in Singapore: Validation Study. *J Med Internet Res*. Vol 22. No 5. e19437. 2020. <https://doi.org/10.2196/19437>.
6. Cross J. Apple’s COVID-19 exposure notification API: What it is and how it works in iOS 13.5. *Macworld*. May 27, 2020. <https://www.macworld.com/article/3545329/apples-covid-19-exposure-notification-api-what-it-is-and-how-it-works.html>.
7. Burke D. An update on Exposure Notifications. *Google Blog*. 2020. <https://blog.google/inside-google/company-announcements/update-exposure-notifications/>.
8. Ranisch R et al. Digital contact tracing and exposure notification: ethical guidance for trustworthy pandemic management. *Ethics and Information Technology*. 2020. <https://doi.org/10.1007/s10676-020-09566-8>.
9. Hinch R, Probert W, Nurtay A, Kendall M, and Wyman C. Effective Configurations of a Digital Contact Tracing App: A report to NHSX.
10. University of Oxford. Digital contact tracing can slow or even stop coronavirus transmission and ease us out of lockdown. 2020. <https://www.research.ox.ac.uk/Article/2020-04-16-digital-contact-tracing-can-slow-or-even-stop-coronavirus-transmission-and-ease-us-out-of-lockdown>.
11. Caso G, De Nardis L, Lemic F, Handziski V, Wolisz A, and Benedetto M G DI. ViFi: Virtual Fingerprinting WiFi-Based Indoor Positioning via Multi-Wall Multi-Floor Propagation Model. *IEEE Transactions on Mobile Computing*. Vol 19. No 6. 1478–1491. 2020. <https://doi.org/10.1109/TMC.2019.2908865>.

12. Asuquo P and Udofia K. Performance Evaluation of Ekahau RTLS in Indoor Environments. 2018.
13. Apple. Exposure Notifications - Frequently Asked Questions. 2020.
14. Robinson A and Waldo J. Technical Difficulties of Contact Tracing. 2021.
15. Leith D J and Farrell S. Measurement-based evaluation of Google/Apple Exposure Notification API for proximity detection in a light-rail tram. PLOS ONE. Vol 15. No 9. e0239943. September 30, 2020. <https://doi.org/10.1371/journal.pone.0239943>.
16. Leith D J and Farrell S. Coronavirus Contact Tracing: Evaluating The Potential Of Using Bluetooth Received Signal Strength For Proximity Detection. 1–11. 2020. <http://arxiv.org/abs/2006.06822>.
17. Hoepman J. A Critique of the Google Apple Exposure Notification (GAEN) Framework. ArXiv. Vol abs/2012.05097. 2020.
18. Kissick C, Setzer E, and Schulz J. What Ever Happened to Digital Contact Tracing? Lawfare. July 21, 2020. <https://www.lawfareblog.com/what-ever-happened-digital-contact-tracing>.
19. Lomas N. UK gives up on centralized coronavirus contacts-tracing app — will “likely” switch to model backed by Apple and Google. TechCrunch. June 18, 2020. <https://techcrunch.com/2020/06/18/uk-gives-up-on-centralized-coronavirus-contacts-tracing-app-will-switch-to-model-backed-by-apple-and-google/>.
20. Newton C. Why countries keep bowing to Apple and Google’s contact tracing app requirements. The Verge. May 8, 2020. <https://www.theverge.com/interface/2020/5/8/21250744/apple-google-contact-tracing-england-germany-exposure-notification-india-privacy>.
21. Horowitz J and Satariano A. Europe Rolls Out Contact Tracing Apps, With Hope and Trepidation. The New York Times. June 16, 2020. <https://www.nytimes.com/2020/06/16/world/europe/contact-tracing-apps-europe-coronavirus.html>.
22. Google and Apple. Exposure Notification: Frequently Asked Questions. 2020.
23. Whittaker Z and Etherington D. User identity will not be shared with other users, Apple and Google as part of this process. TechCrunch. April 13, 2020. <https://techcrunch.com/2020/04/13/apple-google-coronavirus-tracing/>.
24. Leswing K. States are finally starting to use the Covid-tracking tech Apple and Google built — here’s why. CNBC. October 3, 2020. <https://www.cnbc.com/2020/10/03/covid-app-exposure-notification-apple-google.html>.

25. Newcomb A. U.S. states are turning to a private Irish company to help stop the spread of COVID. *Fortune*. October 18, 2020. <https://fortune.com/2020/10/18/covid-contact-tracing-apps-us-nearform/>.
26. He S and Chan S-. G. Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons. *IEEE Communications Surveys & Tutorials*. Vol 18. No 1. 466–490. 2016. <https://doi.org/10.1109/COMST.2015.2464084>.
27. Honkavirta V, Perala T, Ali-Loytty S, and Piche R. A comparative survey of WLAN location fingerprinting methods. in *2009 6th Workshop on Positioning, Navigation and Communication*. 2009. 243–251. <https://doi.org/10.1109/WPNC.2009.4907834>.
28. Simonite T. Wi-Fi Trick Gives Devices Super-Accurate Indoor Location Fixes. *MIT Technology Review*. 2015. <https://www.technologyreview.com/2015/10/16/165765/wi-fi-trick-gives-devices-super-accurate-indoor-location-fixes/>.
29. Bai Y B et al. A new method for improving Wi-Fi-based indoor positioning accuracy. *Journal of Location Based Services*. Vol 8. No 3. 135–147. 2014. <https://doi.org/10.1080/17489725.2014.977362>.
30. Xia S, Liu Y, Yuan G, Zhu M, and Wang Z. Indoor fingerprint positioning based on Wi-Fi: An overview. *ISPRS International Journal of Geo-Information*. Vol 6. No 5. 2017. <https://doi.org/10.3390/ijgi6050135>.
31. Trivedi A, Zakaria C, Balan R, and Shenoy P. WiFiTrace: Network-based Contact Tracing for Infectious Diseases Using Passive WiFi Sensing. Vol 1. No 1. 1–23. 2020. <http://arxiv.org/abs/2005.12045>.
32. Johnson B. Nearly 40% of Icelanders are using a covid app—and it hasn’t helped much. *MIT Technology Review*. 2020. <https://www.technologyreview.com/2020/05/11/1001541/iceland-rakning-c19-covid-contact-tracing/>.
33. COVID Iceland. Rakning C-19 App. 2020. <https://www.covid.is/app/en>.
34. U.S. Government. GPS Accuracy. *GPS.gov*. <https://www.gps.gov/systems/gps/performance/accuracy/>.
35. Kjærgaard M B, Blunck H, Godsk T, Toftkjær T, Christensen D L, and Grønbaek K. Indoor Positioning Using GPS Revisited BT - *Pervasive Computing*. 2010. 38–56.
36. Loh P. Accuracy of Bluetooth-Ultrasound Contact Tracing: Experimental Results from NOVID iOS Version 2 . 1 Using Five-Year-Old Phones. 2020.
37. Hubler S and Hartocollis A. How Colleges Became the New Covid Hot Spots. *New York Times*. 2020. <https://www.nytimes.com/2020/09/11/us/college-campus-outbreak-covid.html>.

38. Ollstein A M and Ravindranath M. Getting it right: States struggle with contact tracing push. Politico. May 17, 2020. <https://www.politico.com/news/2020/05/17/privacy-coronavirus-tracing-261369>.
39. Wroe E and Oza S. Maximizing the Impact of Contact Tracing for COVID-19: The Importance of Human-Centered and Equity-Driven Programming. Harvard Health Policy Review. 2021. <http://www.hhpronline.org/articles/2021/3/4/maximizing-the-impact-of-contact-tracing-for-covid-19-the-importance-of-human-centered-and-equity-driven-programming>.
40. DeCosta-Klipa N. Massachusetts is testing a digital COVID-19 exposure app. Boston.com. April 5, 2021. <https://www.boston.com/news/coronavirus/2021/04/05/massachusetts-covid-tracing-app-massnotify>.
41. Ke G et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in Advances in Neural Information Processing Systems 30. Guyon I, Luxburg U V, Bengio S, Wallach H, Fergus R, Vishwanathan S, and Garnett R, Editors. Curran Associates, Inc. 2017. 3146–3154.
42. Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. 785–794. <https://doi.org/10.1145/2939672.2939785>.
43. Cobb S. Harvard to Track Affiliates' Wi-Fi Signals as Part of Contact Tracing Pilot. The Crimson. August 2, 2020. <https://www.thecrimson.com/article/2020/8/2/tracefi-wifi-contract-tracing-coronavirus/>.
44. Sweeney L. Letter to the Editor: The Latest on Contact Tracing Tech at Harvard. The Crimson. September 24, 2020. <https://www.thecrimson.com/article/2020/9/24/letter-sweeney-contact-tracing-tech/>.
45. Griset R. Virginia leads nation in COVID-19 app use. Virginia Business. December 11, 2020. <https://www.virginiabusiness.com/article/virginia-leads-nation-in-covid-19-app-use/>.
46. Bradford L, Aboy M, and Liddell K. COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. *Journal of Law and the Biosciences*. Vol 7. No 1. July 25, 2020. <https://doi.org/10.1093/jlb/ljaa034>.
47. Shachar C. Protecting Privacy In Digital Contact Tracing For COVID-19: Avoiding A Regulatory Patchwork. *Health Affairs*. 2020. <https://www.healthaffairs.org.ezp-prod1.hul.harvard.edu/doi/10.1377/hblog20200515.190582/full/>.
48. Mello M M and Wang C J. Ethics and governance for digital disease surveillance. *Science*. Vol 368. No 6494. 951 LP – 954. May 29, 2020. <https://doi.org/10.1126/science.abb9045>.
49. Reardon J. Why Google Should Stop Logging Contact-Tracing Data. AppCensus Blog. 2021. <https://blog.appcensus.io/2021/04/27/why-google-should-stop-logging-contact-tracing-data/>.

50. Canales K. Google is facing a lawsuit after a privacy flaw in its contact tracing tech exposed Android users' data to third-party apps. Insider. April 28, 2021.
https://www.businessinsider.com/google-lawsuit-contact-tracing-technology-apple-2021-4?utm_source=feedly&utm_medium=webfeeds.

Chapter 5

Conclusion

This chapter shows how the Three Forces Model can be applied to each of the case studies.

Chapter 2 – How Facebook’s Advertising Algorithms Can Discriminate By Race and Ethnicity

Current State

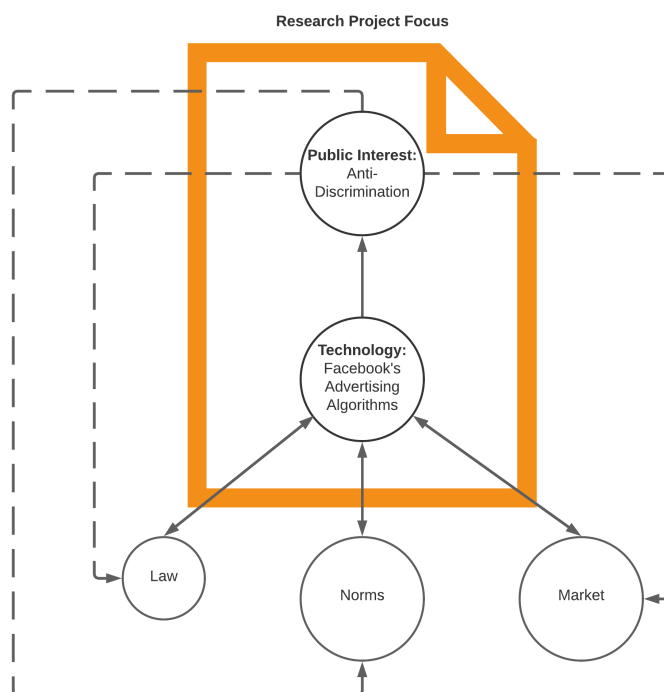


Figure 5.1. The Three Forces Model (Current State) for Chapter 2’s Research Project and Project Focus. The technology is Facebook’s advertising algorithms and the public interest is anti-discrimination. The law is likely exerting a weaker force on Facebook’s advertising algorithms

than norms and the market. The research project focused on how Facebook’s advertising algorithms impacted anti-discrimination.

The technology is Facebook’s different advertising targeting options: its prepackaged “Detailed Targeting” options, its Lookalike Audiences, and its Special Ad Audiences tools. The public interest is anti-discrimination by race and ethnicity.

Law

In the last few years, Facebook has been sued by the National Fair Housing Alliance [1], the ACLU [2], the Communications Workers of America [3], the U.S. Department of Housing and Urban Development [4], and others [5] over issues of discrimination on its advertising platform violating civil rights laws such as the Fair Housing Act and the Civil Rights Act of 1964. On July 24, 2018, Facebook signed a legally binding agreement with the state of Washington to pay \$90,000 in costs and fees to Washington State’s Attorney General’s Office and agree to remove targeting options that may allow advertisers to exclude on the basis of race, religion, sexual orientation, veteran status, and other protected classes for housing, employment, credit, and insurance ads [5]. On March 19, 2019, Facebook settled with the ACLU, the Communications Workers of America, the National Fair Housing Alliance, and others on their multiple lawsuits against Facebook over advertising discrimination by agreeing to create a separate portal for ads regarding housing, employment, and credit [6, 7, 8]. On this portal, Facebook will offer “a much more limited set of targeting options so that advertisers cannot target ads based on Facebook users’ age, gender, race, or categories that are associated with membership in protected groups, or based on zip code or a geographic area that is less than a 15-mile radius, and cannot consider users’ age, gender, or zip code when creating ‘Lookalike’ audiences for advertisers” [8].

Facebook has repeatedly argued in court that it is not liable for discrimination on its platform due to the actions of its advertisers, because it is protected by Section 230 of the Communications Decency Act [9, 10]. According to Facebook, since it is a platform rather than a publisher, then Section 230 means that it should not be liable for the content that it hosts [10]. In 2018, the U.S. Department of Justice has filed Statements of Interest in two different discrimination-related lawsuits against Facebook arguing that it does not believe Section 230 liability protections apply to Facebook's advertising platform [11, 12]. However, Facebook settled both lawsuits before a judge ruled on this issue.

Thus, the law likely exerts the weakest force despite Facebook having faced multiple lawsuits over discrimination and its advertising platform over the last 5 years, because Facebook hasn't faced significant financial or criminal risks from the law over this issue and there's an open question about whether Section 230 means Facebook is ultimately immune from liability for the discrimination of advertisers on its platform.

Norms

Researchers have found that racial and ethnic groups tend to behave differently from each other online. They visit different websites [13, 14, 15, 16], follow different social media [17, 18], and even browse the web using different devices [19, 20]. In addition, researchers have found that Americans tend to have very racially homogenous friend networks [21]. For White Americans, on average 91% of their social network are also White [21]. While for Black Americans, on average 83% of their social network are also Black [21]. Similarly, 75% of White Americans and 65% of Black Americans report having a core social network defined as "people with whom they discuss important matters" being entirely of their own race [21]. Research on algorithmic bias has found that algorithms often learn to replicate existing biases in society through the training data that's used and the

reinforcement from user input [22, 23, 24, 25, 26, 27]. In the case of Facebook’s ad platform, it’s using data from the biggest online social network in the world.

However, just because minority groups may behave differently online, many minority users may not expect to see discriminatory ad targeting online especially for housing, employment, or credit where anti-discrimination laws exist, and even legal areas for discriminatory targeting like political advertising can be controversial. For example, researchers studying the 3,519 ads that Facebook shared with Congress as part of the investigations into Russian interference with the 2016 elections found that many of the ads focused on black identity issues, such as police shootings, BlackLivesMatter, and discrimination [28]. 17 ads even used Facebook’s “African-American (US)” multicultural affinity group [29]. Another study found that across all the Russian-linked ads disclosed by Facebook, 52% had more than double the proportion of African-Americans in their target audience compared to Facebook’s US baseline [30]. In 2020, a significant amount of misinformation targeted Hispanics, particularly Hispanic voters in Florida [31, 32, 33, 34]. Due to the widespread misinformation, often targeting minority users, on Facebook, on June 17, 2020, civil rights groups such as the NAACP, the Anti-Defamation League, Color of Change, and others launched the “Stop Hate for Profit” campaign to pressure major corporate advertisers to stop advertising on Facebook for the month of July 2020 [35].

Market

More than 1,000 major advertisers joined the July 2020 Stop Hate for Profit boycott including Microsoft, Starbucks, Unilever, Target, and more [36]. The organizers of the boycott outlined 10 recommendations for Facebook to adopt such as hiring a C-suite executive to review the company’s products for discrimination, hate, and bias, participating in a regular third-party audit on identity-based misinformation and hate, stopping the amplification of content with ties to hate,

misinformation, or conspiracies, ending the exemption of politicians from fact-checking, and other changes [35].

Facebook responded to the boycott on July 8, 2020 by releasing its civil rights audit [37] and on August 11, 2020 by announcing the removal of its controversial multicultural affinity groups – including “African American (US)”, “Asian American (US)”, and “Hispanic (US – All)” – as ad targeting options [38]. This event in the middle of 2020 presents an interesting opportunity for my study to examine racial and ethnic discrimination on Facebook’s ad platform in January 2020 and January 2021, before and after the boycott.

Ultimately, Facebook does not appear to have suffered significant financial damage as a result of the boycott, though the boycott also coincided with the COVID-19 pandemic, which benefited Facebook as many small and medium businesses increased their online marketing and sales [39]. In addition, as a for-profit company, Facebook wants to maximize profits by serving the needs of advertisers as best as possible, which may reinforce discriminatory ad targeting.

Research project

The research project focused on the dyadic relationship between Facebook’s advertising algorithms and anti-discrimination. The project found that while Facebook’s retirement of multicultural affinity groups in August 2020 has removed one way to target minorities on the platform, its other targeting options, as well as Lookalike and Special Ad Audience tools, can still discriminate by race and ethnicity. While some discriminatory ad targeting may be legal or even desirable, this project demonstrates how there’s a lack of transparency on the discriminatory potential of Facebook’s ad platform which may help cover up the behavior of discriminatory advertisers and undermine the intent of non-discriminatory ones.

- In 2021, Facebook’s “African-American Culture” ad targeting option contained 75% fewer White users than the old “African American (US)” option they removed in the previous year.
- Facebook’s tools to help advertisers find similar users to their existing customers exhibited bias towards including more African-Americans or Whites depending on which racial group was dominant in an advertiser’s customer list, and this was true for the Lookalike Audience tool, as well as the Special Ad Audience tool that Facebook designed to explicitly not use sensitive demographic attributes when finding similar users.
- Lookalike or Special Ad audiences based on customer lists with either stereotypically African-American or White names or ZIP codes would be even more biased towards including more users of that demographic group.
- Similarly, Lookalike audiences based on Asian customer lists can also become biased towards Asians, reaching up to 100% Asian in one case, and Lookalike audiences based on Hispanics over-represented Hispanics versus Non-Hispanics.

Goal state

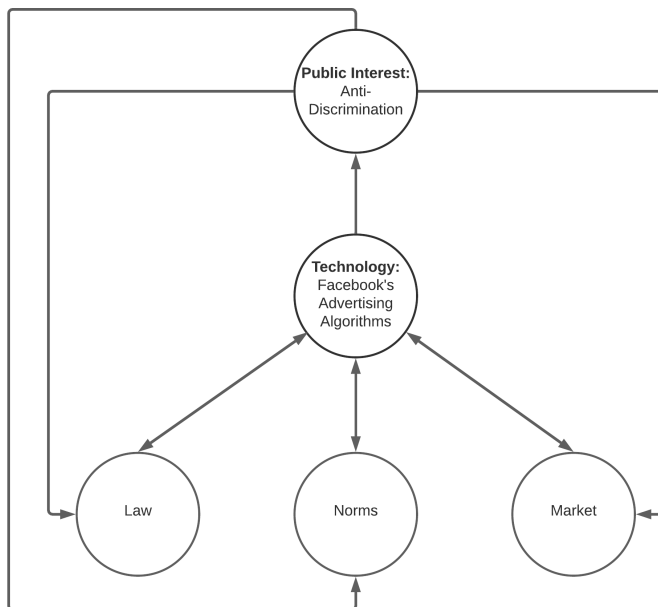


Figure 5.2. The Three Forces Model (Goal State) for Chapter 2’s Research Project. The public interest of anti-discrimination can be an input to how the law, norms, and market affect Facebook’s advertising algorithms.

Law

Chapter 2 demonstrates the need for greater transparency by Facebook on how advertisers are using its targeting tools that have the potential to discriminate by race and ethnicity including for sectors covered by existing anti-discrimination law such as housing, employment, and credit. Regulators can require Facebook’s Ad Library for political, housing, employment and credit-related ads to be updated to show the relevant metadata for establishing a “robust causal link” for racial or ethnic discrimination lawsuits [40]. This is especially important since HUD’s “Implementation of the Fair Housing Act’s Disparate Impact Standard” published on September 24, 2020 now requires a plaintiff to present evidence of a “robust causal link” in order to bring a disparate impact discrimination lawsuit in the first place [41].

In addition, Chapter 2 shows the risks of implementing a “fairness through unawareness” standard, which was initially proposed by HUD on August 19, 2019 [42], that would likely protect companies like Facebook from being successfully sued for discrimination by an algorithm such as its Special Ad Audience tool that does not rely on “factors that are substitutes or close proxies for protected classes under the Fair Housing Act” [42].

While the U.S. Department of Justice has filed Statements of Interest in two different discrimination-related lawsuits against Facebook arguing that it does not believe Section 230 liability protections apply to Facebook’s advertising platform [11, 12], Congress can reform Section 230 to make that position explicit.

Finally, federal regulators such as the Federal Trade Commission (FTC) can audit Facebook's ad platform for discrimination and use their enforcement powers to address violations of Section 5 of the FTC Act, the Fair Credit Reporting Act, and the Equal Credit Opportunity Act [43].

Norms

Facebook released its civil rights audit on July 8, 2020, right after the start of a major boycott of its ad platform by corporate advertisers organized by civil rights groups. One of the demands of the boycott organizers is for Facebook to conduct regular civil rights audits of its platform [35]. Thus, having established a new precedent in 2020, Facebook can use Chapter 2 as an example of how future audits can test its ad platform for racial and ethnic discrimination and go even further in examining why these biases exist and how to address them. In fact, Facebook and many other large US tech companies already participate in the similar norm of releasing annual workforce diversity reports in order to document progress on gender, racial, and other demographic diversity in their labor force.

Chapter 2 studied Facebook's ad platform over two time periods: January 2020 and January 2021. Future civil rights audits by Facebook can continuously examine its different targeting tools for racial and ethnic bias as Facebook itself makes changes to their algorithms. Since Facebook can provide its own auditors with privileged access to its data and systems, these audits can go even further than Chapter 2 in order to study the causes of the bias due to the training data, user feedback, complex interactions between multiple algorithmic systems, and more. As Facebook's Chief Operating Officer Sheryl Sandberg noted in July 2020, "it is the beginning of the journey, not the end" [44].

Since 2014, many large US tech companies, including Facebook, have participated in the related norm of releasing annual workforce diversity reports [45]. Facebook published its first diversity report on June 25, 2014, which showed that its technical workforce had significant gender bias with 85% male vs. 15% female and also significant racial bias with 53% White, 41% Asian, and

only 1% Black [46]. By 2019, Facebook has seen an increase of female tech workers to 23% but a smaller increase of its Black workers to 3.8% [45]. Other major US tech companies who participate in this trend include Apple, Google, Microsoft, Amazon, Twitter, and more [45].

Market

Digital ad platforms like Facebook and its competitors can build discrimination detection into their tools in a transparent way to leverage market forces to compete on the basis of reducing discrimination.

For example, right now Facebook's ad planning tool only provides an estimated reach number given a combination of different ad targeting options, Custom audiences, Lookalike audiences, or Special Ad audiences. In the future, Facebook can enrich its estimated reach report by displaying the demographic distribution of who will see an ad on the basis of race, ethnicity, gender, age, geography, and other categories. If a particular Custom, Lookalike, or Special Ad audience is racially or ethnically biased, Facebook can flag those audiences when they are first created in order to notify the advertiser and potentially limit their usage. Other competitors to Facebook, such as Google or Amazon, may also implement similar features to their own ad planning tools.

Since tech companies don't usually directly ask their users for their race, they could use algorithmic or human evaluators to generate that information for their anti-discrimination tools. For example, they could use the names or pictures of a user. This was the approach taken by Airbnb's Project Lighthouse, launched in 2020 to study the racial experience gap for guests and hosts on Airbnb [47]. Project Lighthouse used a third party contractor to assess the perceived race of an individual based on their name and profile picture [47].

In 2020, the International Counsel for Ad Self Regulation (ICAS) found that the following nations have advertising self-regulation standards that address non-discrimination: Australia,

Belgium, Brazil, Canada, Chile, France, Ireland, India, Netherlands, New Zealand, Peru, Phillipines, Portugal, Romania, Spain, United Kingdom, and South Africa. While self-regulation is likely not a panacea, the US advertising industry should still join that list [48].

Chapter 3 – How Were Social Security Numbers Assigned?

Current state

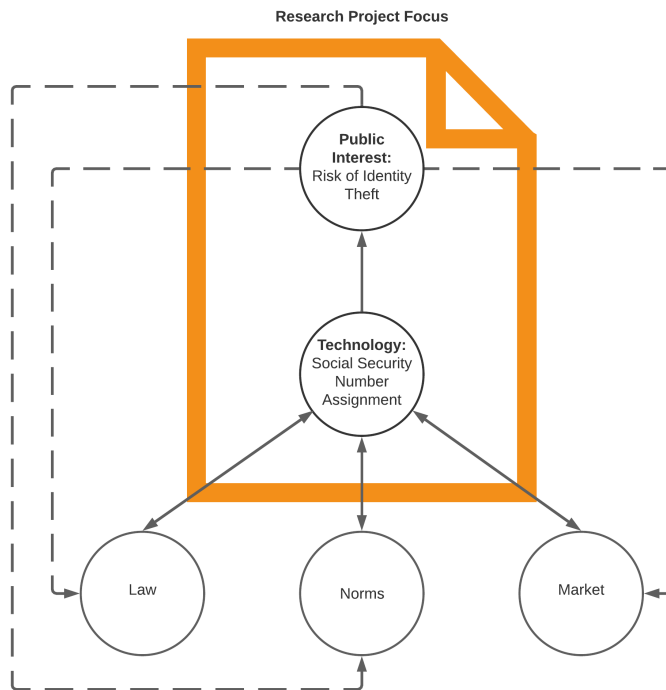


Figure 5.3. The Three Forces Model (Current State) for Chapter 3’s Research Project and Project Focus. The technology is Social Security Number assignment and the public interest is the risk of identity theft. The law, norms, and market are all powerful forces in this case. The research project focused on how Social Security Number assignment impacted the risk of identity theft.

The technology is Social Security Number (SSN) assignment. The public interest is the risk of identity theft.

Law

The first Social Security Numbers (SSNs) were issued on December 1, 1936 [49]. One unique number in order to keep track of the wages earned and the Social Security benefit to be received by an individual at retirement. From the beginning, there were concerns about how it can empower the federal government to limit the privacy and freedom of Americans [50]. For example, the Republican National Committee (RNC) chairman John D. M. Hamilton in 1936 charged that eventually Americans would need to wear “dog tags” showing their SSN [50]. Newspapers made the comparison between dog tags of SSNs and being drafted even though there was no war [50]. In the 1930s, the Social Security Board tried to emphasize in their communications that the SSN is simply a means of tracking one’s “account” and not the “person” in order to minimize the “charge of regimentation” [50].

Social Security Numbers have 3 parts that follow an XXX-YY-ZZZZ structure. The first 3 digits (XXX) are the Area Numbers (ANs) assigned to each state. The middle 2 digits (YY) are the Group Numbers (GNs) which are assigned sequentially based on a custom protocol. The last 4 digits (ZZZZ) are the Serial Numbers (SNs) that are assigned sequentially from 0001 to 9999 [51]. What is not publicly confirmed by the Social Security Administration (SSA) is the relationship between the Area Number, the Group Number, and the Serial Number in terms of how they get assigned over time.

In the beginning, Social Security Number applications were processed at post offices, then starting in July 1937 at regional Social Security offices, and finally, in 1961, all new SSN assignment was centralized to a Social Security office in Baltimore and done via computers starting in 1972 [52]. In 1989, the SSA started the Enumeration At Birth (EAB) program, which was an anti-fraud program that integrated the application for SSNs into the birth certification process. By 1995, 50% of all new SSNs being assigned were given to newborns [53, 54]. Starting on June 25, 2011, the SSA started randomizing all 9 digits for new SSN assignments [55]. According to the SSA, randomization will “protect the integrity of the SSN” and “extend the longevity of the nine-digit SSN nationwide”, since

Area Numbers are no longer designated for different states [56]. SSA acknowledged that randomization “will help protect an individual's SSN by making it more difficult to reconstruct an SSN using public information” [56], which is the attack first described by Acquisti and Gross in 2009 [57] and further tested in Chapter 3.

Since 1936, the law has also played a role in spreading the use of SSNs beyond Social Security. In 1943, President Franklin Roosevelt issued Executive Order 9397, which encouraged federal agencies to use SSNs as account numbers for public services [58]. The Internal Revenue Service (IRS) started using it for taxes in 1961. That same year, SSNs became the ID number for federal employees. Medicare used it for enrollment in 1965. The Veterans Administration started using it in 1966. The Department of Defense used it as the military ID number starting in 1969. Food stamps started using it in 1977. Housing and Urban Development (HUD) programs started using it in 1988. One federal program after another started adopting the SSN in order to track the individuals they service [52, 59]. State governments also used the SSN on driver's licenses and marriage licenses [59]. For the private sector, under the 1970 Bank Records and Foreign Transactions Act, all banks, savings and loan associations, credit unions and broker/dealers in securities are required to obtain the SSNs of all of their customers [60].

Finally, federal courts have ruled that the right to privacy does not extend to one's SSN. In *Cassano v. Carb* (2006), the US Court of Appeals, Second Circuit ruled that the “the Constitution does not provide a right to privacy in one's SSN... we decline to expand the constitutional right to privacy to cover the collection of SSNs” [61]. In this case, the court ruled against an employee seeking to not provide their SSN to their employer due to fears of identity theft.

Norms

Since SSNs are used as both identifiers and authenticators in the US today, it's normal for Americans to be protective of their SSN while openly sharing relevant information to predict their SSN such as their date of birth and state of birth. Research on how many respondents to the Census Bureau's General Social Survey provided their SSN, which was an optional field in the contact information section, found a decrease from 60% in 1993 to 17% in 2008 [62].

Market

There's a feedback loop between the widespread usages of SSNs by different services and how that reinforces future usages of SSNs to link datasets together using one number that most likely exists in many places. For example, SSNs can be used as identifiers to link records about an individual from multiple financial, housing, criminal history, and other datasets together for a background check. It's also a convenient authenticator to request an individual to fill out a 9-digit number on a credit card, rental, job, or other form, especially online, when compared to verifying other forms of government IDs such as driver's licenses or passports. In the US today, potential customers do not have to provide their SSNs when requested, but businesses then are able to refuse to service the customer, unless there are specific federal or state laws that regulate the transaction [63].

However, the same widespread usage of SSNs, especially as authenticators, has contributed to the identity theft problem. In 2019, the cost of identity theft was \$16.9 billion and impacted 5.1% of Americans [64]. Large scale data breaches have made many SSNs available to identity thieves on dark web marketplaces. For example, the 2017 Equifax data breach possibly revealed personal information including the SSNs of 145.5 million individuals [65]. In prior research, I have found that dark web marketplaces sell breached datasets and SSNs for as low as \$1 per SSN [66].

Research project

The research project focused on the dyadic relationship between Social Security Number assignment and the risk of identity theft. The project found strong evidence that SSNs were assigned in a nested loop protocol based on sets of Area Numbers, Group Numbers, Area Numbers, and then Serial Numbers in all 50 states and DC between 1989 and 2011. This means that Americans born between 1989 – 2011 face an additional SSN-based identity theft vulnerability due to how SSA assigned their SSNs at birth.

- I build upon earlier research to propose my own hypothesis about SSN assignment as following a nested loop protocol
- For Americans born between 1989 and 2011, they have SSNs most vulnerable to prediction based on their state of birth and date of birth, due to the Social Security Administration's Enumeration At Birth program
- For SSNs in the Death Master File, I am able to accurately predict the first 5 digits 48% of the time and the first 6 digits 11% of the time
- States with smaller populations were the most vulnerable: I am able to accurately predict the first 5 digits of the SSN in 19 states including DC more than 80% of the time, and for 5 states – Delaware, Idaho, North Dakota, South Dakota, and Wyoming – more than 90% of the time
- It's time for public policy to focus on solutions that can replace SSNs with alternatives that are designed to be strong authenticators from the start

Goal state

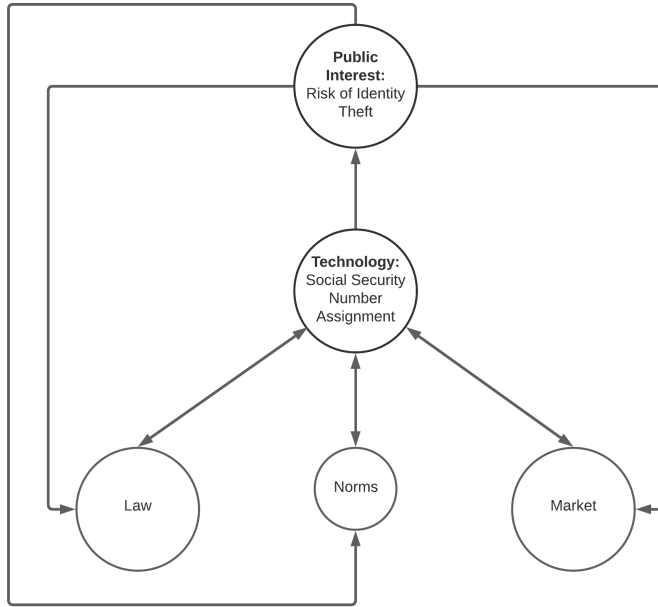


Figure 5.4. The Three Forces Model (Goal State) for Chapter 3’s Research Project. The public interest of reducing the risk of identity theft can be an input to how the law, norms, and market affect the role of Social Security Numbers in society, with a change in norms likely being the weakest force since many Americans are already protective of giving out their SSNs.

Law

In recent years, the federal government has passed laws to limit the sharing of datasets with SSNs. For example, after 1990, datasets released by the federal government should not disclose SSNs [63]. In 2000, the amended Driver’s Privacy Protection Act required state departments of motor vehicles to obtain consent from individuals in order to release their SSNs [63]. The 1999 Gramm-Leach-Bliley Act regulated the sharing of personally identifiable information (PII) by financial institutions, the 1996 Health Insurance Portability and Accountability Act (HIPAA) regulated the sharing of PII in health data, and the Family Educational Rights and Privacy Act (FERPA) regulated the sharing of PII by educational institutions [63]. However, these laws are limited in scope and don’t address the issue of unintended release of SSNs through data breaches nor the vulnerability

documented in Chapter 3 of how SSNs of individuals born between 1989 and 2011 are linked to their date of birth and state of birth.

Instead, US public policy can invest in creating better technology design solutions to replace Social Security Numbers as de facto national identifiers and authenticators. The Discussion section of Chapter 3 describes a variety of alternative authentication solutions in the public sector, private sector, and internationally.

Norms

Changing the current common practice, endorsed by the IRS, of obscuring the first 5 digits of an SSN on a form or paycheck while revealing the last 4 digits, may have a moderate benefit on preventing the revealing of SSNs [67]. But ultimately changing norms is likely the weakest force since many Americans are already protective of giving out their SSNs, but Chapter 3 demonstrates how there's underlying correlations in their SSN and their date and state of birth, which individuals can't fix on their own.

Market

In the public sector, the IRS has implemented a number of different methods to detect fraudulent tax returns that have historically been filed with the SSNs of their victims. The IRS reports that the amount of ID theft tax fraud has decreased from \$5.8 billion in 2013 [68] to \$184 million in 2019 [69]. One of the IRS' response measures is the creation of the Identity Protection PIN (IP PIN) for victims of identity theft to include on their legitimate tax returns. In 2021, the IRS opened up the IP PIN to be available to any taxpayer [70]. In order to sign up, a taxpayer has to verify their address and a financial account number, such as a credit card, student loan, mortgage, or car loan [71]. Then the IRS would send an activation code to log into the IP PIN website either to a mobile phone number associated with the taxpayer or by mail [71].

In the private sector, we see alternative technologies with device-based authentication, single sign-on solutions, knowledge-based authentication, and verification of driver's licenses and other photo IDs, which are described in more detail in the Discussion section of Chapter 3.

While these alternatives have their own trade-offs, they are potentially less vulnerable as strong authenticators than SSNs are today.

Chapter 4 - Building A Collocation Detection System Using A Wi-Fi Sensor Array for COVID-19 Contact Tracing in A University Setting

Current state

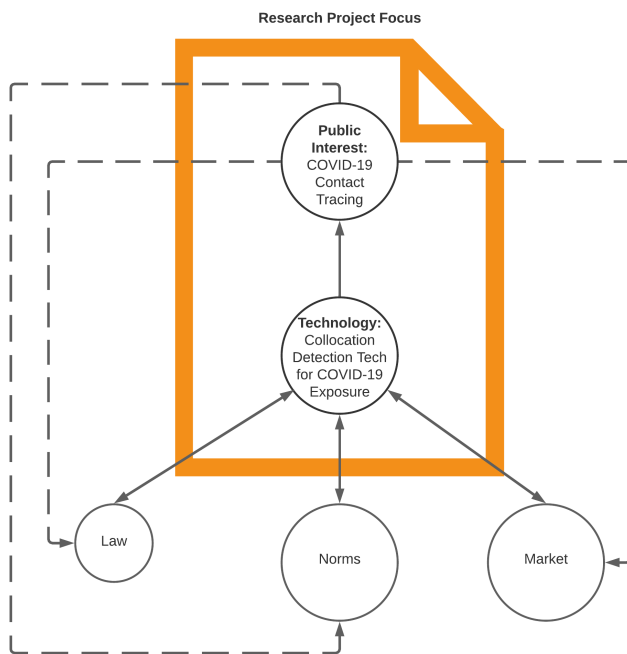


Figure 5.5. The Three Forces Model (Current State) for Chapter 4's Research Project and Project Focus. The technology is collocation detection tech for COVID-19 exposure and the public interest is COVID-19 contact tracing. The law is the weakest force primarily due to the lack of coordination by the US federal government on contact tracing in 2020. The research project

focused on how collocation detection tech for COVID-19 exposure impacted COVID-19 contact tracing.

The technology is collocation detection tech for COVID-19 exposure. The public interest is COVID-19 contact tracing.

Law

In 2020, the US saw a fairly decentralized COVID-19 contact tracing effort primarily by state and local governments rather than the federal government [72]. In Massachusetts, the state partnered with Partners in Health to create a coalition of contact tracers working for state and local public health departments as well as other local partners [73]. While many states started rolling out their GAEN apps in 2020, Massachusetts was a relatively late adopter, only beginning tests of the MassNotify GAEN app in April 2021 [74]. At Harvard, University Health Services had its own contact tracing team to follow-up on positive cases in the university's community, and this team was part of the Massachusetts Contact Tracing Collaborative and coordinated with public health authorities. While the law could have been a strong force in leading contact tracing efforts in the US, in 2020, it was likely the weakest force due to the highly decentralized approach that was taken.

Internationally, Google and Apple's requirement for GAEN apps to be decentralized rather than centralized also undermined their utility for contact tracers hoping to learn who were exposed to COVID-19. French junior minister for digital affairs, Cédric O, stated, "It is highly abnormal that you are constrained as a democratic state in your technical choice because of the internal policies of two private companies" [75].

Norms

Using digital contact tracing technology was a brand-new norm in 2020, which faced challenges due to ease of adoption and privacy concerns. After Google and Apple released the GAEN

API in May 2020 [76], states and countries then had to build and promote GAEN apps to the public. In the US, adoption rates were generally low, with Virginia being a leader at 21% as of December 2020 [77]. This fell far short of the 56% adoption rate that a simulation study found as necessary for slowing down the spread of COVID-19 [78, 79]. Different surveys found a wide range of 17 – 70% of Americans reporting being willing to use contact tracing apps [80]. Mello and Wang argues that it could be ethically justified to have digital contact tracing be opt out rather than opt in due to the public health benefits while still providing “those with strong preferences to act on them while not conflating philosophical objections with simple inertia” [80]. Privacy concerns around who would have access to the data being collected by digital contact tracing apps were also significant. Google and Apple required GAEN apps to be decentralized in order to be in the Play Store and the App Store, meaning that private data about one’s COVID-19 exposure stays on one’s phone rather than in a centralized database accessible for contact tracing [81]. This means that a GAEN app can show an exposure notification alert on the devices owned by someone who was potentially exposed to COVID-19, but a contact tracer can’t use GAEN apps to learn who got exposed and reach out to them.

Market

Google and Apple were able to leverage their duopoly in the market on mobile operating systems and the corresponding app stores to require GAEN apps to be decentralized. GAEN apps that use Bluetooth are also not able to track users with GPS or other location services on the device [82].

Research project

The research project focused on the dyadic relationship between collocation detection tech for COVID-19 exposure and COVID-19 contact tracing. The project built and tested TraceFi, a Wi-Fi based collocation detection system that uses a sensor array to accurately detect mobile devices

within 6 feet of each other, for possible use in contact tracing without the burden of requiring a user to install an app in order to participate.

- TraceFi is a Wi-Fi based collocation detection system that uses a sensor array to accurately detect mobile devices within 6 feet of each other for possible use in contact tracing without the burden of requiring a user to install an app in order to participate
- We tested multiple machine learning models in a TraceFi pilot across 12 different spaces in 3 different buildings under regular use conditions and found XGBoost models had a peak sensitivity of 91% and a peak specificity of 86%, with a high median sensitivity of 77% and a high median specificity of 81%
- TraceFi can be used for accurate real-world collocation detection for contact tracing to determine whether 2 devices were within 6 feet for 15 minutes or more and is the first Wi-Fi technology to do so
- We engaged with stakeholders around the university to incorporate their concerns around the ease of adoption, accuracy, cost, and privacy into how we propose including TraceFi data into the contact tracing data flow
- We designed a system for using TraceFi data for contact tracing according to Fair Information Practices that seek to preserve the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive

Goal state

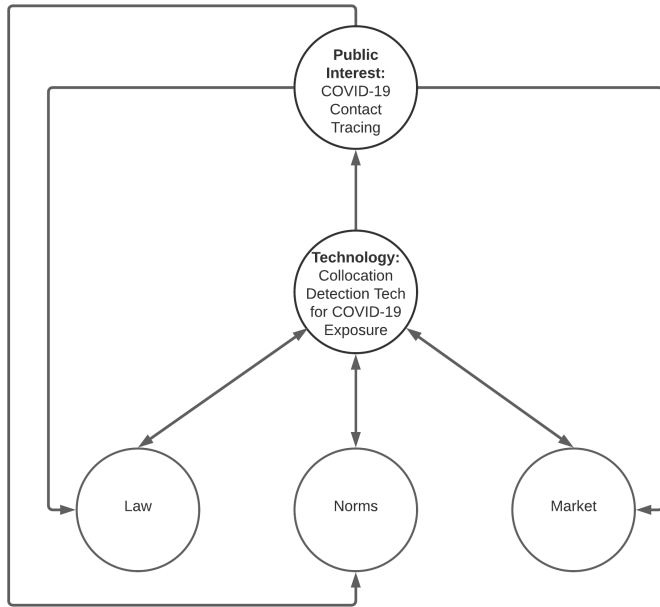


Figure 5.6. The Three Forces Model (Goal State) for Chapter 4’s Research Project. The public interest was an input from the outset to how we considered the law, norms, and market forces in the design of the data flow system for TraceFi’s collocation predictions to support the work of contact tracers at Harvard University Health Services while protecting the privacy of individual device owners.

Law

Legal scholars have pointed out how digital contact tracing apps such as the GAEN apps “fall completely outside HIPAA’s parameters” [83]. In May 2020, the COVID-19 Consumer Data Protection Act was introduced though not yet passed in the Senate to address how digital contact tracing technologies should collect and share data [84]. Ideally, the law should be updated to clarify how HIPAA protections should apply to digital contact tracing technologies. In the case of TraceFi, we propose a privacy wall between the TraceFi System maintained by the Data Privacy Lab and Harvard University Health Services (HUHS) and Harvard University Information Technology (HUIT). The TraceFi side of the privacy wall has location information with no identity. The HUHS and HUIT side has

identity but does not have open access to location information. The only disclosures of location information outside of the TraceFi System at the Data Privacy Lab are to human contact tracers at HUHS, through the Dashboard and, potentially, by automated means to collocated persons who have not opted out of TraceFi.

At Harvard, we were able to work with contact tracers at University Health Services that were also part of the Massachusetts Contact Tracing Collaborative. For organizations that may want to adopt TraceFi but don't have in-house contact tracers, they may still be able to collaborate with their local public health departments and implement a privacy wall along with strict access controls and immutable logs to ensure that only the necessary data for contact tracing is being shared with logs that would record evidence of any violations.

Norms

We designed the data flow system to follow best practices described in the Fair Information Practices and GDPR principles. We sought to minimize the collection, sharing, and storage of data to what's necessary for the purposes of contact tracing. A person can also receive a copy of their own TraceFi data in their private storage on MyDataCan. Data will also be stored according to the university's highest security level (Level 4). Finally, to ensure accountability and transparency, there will be an immutable log that records each access of TraceFi data which is audited and reported to the University Electronic Communications Policy Oversight Committee which makes public summaries. Thus, if these practices were violated in some way, the immutable log can serve as evidence of the violation.

We also proposed making TraceFi opt out in order to ensure ease of adoption. A big advantage of TraceFi over GAEN apps is its ease of adoption, which is highly important for reaching a high enough population adoption rate to reduce the spread of COVID-19. Studies of electronic health

record sharing have found that people tend to stick with the default choice, with only 2 to 5% opting out of health information exchange [80]. Making TraceFi opt out would still allow users to exercise their privacy preferences while ensuring the inertia to not act doesn't undermine effective contact tracing.

Finally, we sought to approach the ideal privacy model for contact tracing of preserving the privacy of the location and collocation data of individuals until they tested positive for COVID-19 or were potentially exposed to someone who tested positive. This meant creating the privacy wall where the TraceFi side has location information with no identity and the HUHS and HUIT side has identity but does not have open access to location information.

Market

Since TraceFi is a new digital contact tracing technology that we developed, we were able to engage with stakeholders throughout the process to ensure that TraceFi would meet the "market" needs of the university.

We sought out stakeholder input throughout the pilot study with 13 digital town halls and regular discussions with partners throughout the university as explained in the Methods section of Chapter 4. We also compared TraceFi to GAEN apps based on stakeholder concerns of ease of adoption, accuracy, cost, and privacy, which is detailed in the Discussion section of Chapter 4.

References

1. NATIONAL FAIR HOUSING ALLIANCE; FAIR HOUSING JUSTICE CENTER, INC.; HOUSING OPPORTUNITIES PROJECT FOR EXCELLENCE, INC.; FAIR HOUSING COUNCIL OF GREATER SAN ANTONIO v. Facebook. United States District Court, Southern District of New York. <https://nationalfairhousing.org/wp-content/uploads/2019/03/2018-06-25-NFHA-v.-Facebook.-First-Amended-Complaint.pdf>.
2. Sherwin G. How Facebook Is Giving Sex Discrimination in Employment Ads a New Life. ACLU. 2018. <https://www.aclu.org/blog/womens-rights/womens-rights-workplace/how-facebook-giving-sex-discrimination-employment-ads-new>.
3. Communications Workers of America vs. T-Mobile US, Inc. Amazon.com, Inc., Cox Communications, Inc., Cox Media Group, LLC, and similarly situated employers and employment agencies, DOES 1 through 1,000. United States District Court, Northern District of California. <https://doi.org/https://www.onlineagediscrimination.com/sites/default/files/documents/og-cwa-complaint.pdf>.
4. HUD Public Affairs. HUD CHARGES FACEBOOK WITH HOUSING DISCRIMINATION OVER COMPANY'S TARGETED ADVERTISING PRACTICES. Department of Housing and Urban Development. 2019. https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035.
5. Statt N. Facebook signs agreement saying it won't let housing advertisers exclude users by race. The Verge. July 24, 2018. <https://www.theverge.com/2018/7/24/17609178/facebook-racial-discrimination-ad-targeting-washington-state-attorney-general-agreement>.
6. Sherwin G and Bhandari E. Facebook Settles Civil Rights Cases by Making Sweeping Changes to Its Online Ad Platform. ACLU. 2019. <https://www.aclu.org/blog/womens-rights/womens-rights-workplace/facebook-settles-civil-rights-cases-making-sweeping>.
7. National Fair Housing Alliance. Facebook Settlement: Civil Rights Advocates Settle Lawsuit with Facebook: Transforms Facebook's Platform Impacting Millions of Users. 2019. <https://nationalfairhousing.org/facebook-settlement/>.
8. ACLU. Facebook Agrees to Sweeping Reforms to Curb Discriminatory Ad Targeting Practices. 2019. <https://www.aclu.org/press-releases/facebook-agrees-sweeping-reforms-curb-discriminatory-ad-targeting-practices>.
9. Tobin A and Merrill J. Besieged Facebook Says New Ad Limits Aren't Response to Lawsuits. ProPublica2. August 23, 18AD. <https://www.propublica.org/article/facebook-says-new-ad-limits-arent-response-to-lawsuits>.

10. Tobin A. Facebook Promises to Bar Advertisers From Targeting Ads by Race or Ethnicity. Again. ProPublica. July 25, 2018. <https://www.propublica.org/article/facebook-promises-to-bar-advertisers-from-targeting-ads-by-race-or-ethnicity-again>.
11. Berman G. STATEMENT OF INTEREST OF THE UNITED STATES OF AMERICA in National Fair Housing Alliance v. Facebook, Inc. UNITED STATES DISTRICT COURT SOUTHERN DISTRICT OF NEW YORK. New York. 2018. <https://www.justice.gov/crt/case-document/file/1089231/download>.
12. Tse A and Winslow S. UNITED STATES' STATEMENT OF INTEREST in ONUOHA v. FACEBOOK, INC. UNITED STATES DISTRICT COURT NORTHERN DISTRICT OF CALIFORNIA. San Jose. 2018. <https://www.justice.gov/crt/case-document/file/1112561/download>.
13. Sharad Goel J M H M I S. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. 1–8. 2012. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPDFInterstitial/4660/4975%5Cnpapers2://publication/uuid/F3306966-AB04-4CF1-ACCE-A7EF49E1282C>.
14. De Bock K and Van Den Poel D. Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*. Vol 98. No November. 49–70. 2010. <https://doi.org/10.3233/FI-2010-216>.
15. Ying J J, Chang Y, Huang C, and Tseng V S. Demographic Prediction Based on User's Browsing Behaviors. *Research.Nokia.Com*. Vol 2012. No 1. 1–6. 2012. <http://research.nokia.com/files/public/mdc-final241-ying.pdf%5Cfile:///Users/orasanen/Documents/Papers/Ying/research.nokia.com>.
16. Sweeney L. Online ads roll the dice. *Tech@FTC*. 2014. <https://www.ftc.gov/news-events/blogs/techftc/2014/09/online-ads-roll-dice>.
17. Messias J, Vikatos P, and Benevenuto F. White, Man, and Highly Followed: Gender and Race Inequalities in Twitter. in *Proceedings of the International Conference on Web Intelligence*. 2017. 266–274. <https://doi.org/10.1145/3106426.3106472>.
18. Vikatos P, Messias J, Miranda M, and Benevenuto F. Linguistic Diversities of Demographic Groups in Twitter. in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. 2017. 275–284. <https://doi.org/10.1145/3078714.3078742>.
19. Fairlie R. Have we finally bridged the digital divide? Smart phone and Internet use patterns by race and ethnicity. *First Monday*. Vol 22. 2017.
20. Tsetsi E and Rains S A. Smartphone Internet access and use: Extending the digital divide and usage gap. *Mobile Media & Communication*. Vol 5. No 3. 239–255. June 13, 2017. <https://doi.org/10.1177/2050157917708329>.

21. Cox D, Navarro-Rivera J, and Jones R. Race, Religion, and Political Affiliation of Americans' Core Social Networks. PRRI. Vol 5. 1–5. 2014. <https://www.prii.org/research/poll-race-religion-politics-americans-social-networks/>.
22. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuter. October 10, 2018. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
23. Buolamwini J and Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. Vol 81. 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>.
24. Wiggers K. IBM releases Diversity in Faces, a dataset of over 1 million annotations to help reduce facial recognition bias. VentureBeat. January 29, 2019. <https://venturebeat.com/2019/01/29/ibm-releases-diversity-in-faces-a-dataset-of-over-1-million-annotations-to-help-reduce-facial-recognition-bias/>.
25. Lohr S. Facial Recognition Is Accurate, if You're a White Guy. The New York Times. February 9, 2018. <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>.
26. Emerging Technology from the ArXiv. Racism is Poisoning Online Ad Delivery, Says Harvard Professor. MIT Technology Review. February 4, 2013. <https://www.technologyreview.com/2013/02/04/253879/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>.
27. Datta A, Datta A, Makagon J, Mulligan D K, and Tschantz M C. Discrimination in Online Advertising: A Multidisciplinary Inquiry. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. Vol 81. 20–34. <http://proceedings.mlr.press/v81/datta18a.html>.
28. Boyd R et al. Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election using Linguistic Analyses. 1–9. 2018. <https://doi.org/10.31234/osf.io/ajh2q>.
29. DiResta R et al. The Tactics & Tropes of the Internet Research Agency, New Knowledge. 2019. <https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/787lea6d5bafbf19/optimized/full.pdf#page=1>.
30. Ribeiro F N et al. On microtargeting socially divisive ads: A case study of Russia-linked Ad campaigns on Facebook. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. 140–149. 2019. <https://doi.org/10.1145/3287560.3287580>.

31. Sanz C. Misinformation targeted Latino voters in the 2020 election. ABC News. November 21, 2020. <https://abcnews.go.com/Politics/latino-voters-misinformation-targets-election-2020/story?id=74189342>.
32. Romero L. Florida's Latino voters being bombarded with right-wing misinformation, experts and advocates say. ABC News. October 20, 2020. <https://abcnews.go.com/Politics/floridas-latino-voters-bombarded-wing-misinformation-advocates/story?id=73707056>.
33. Ryan-Mosley T. "It's been really, really bad": How Hispanic voters are being targeted by disinformation. MIT Technology Review. October 12, 2020. <https://www.technologyreview.com/2020/10/12/1010061/hispanic-voter-political-targeting-facebook-whatsapp/>.
34. Rodriguez S and Caputo M. "This is f---ing crazy": Florida Latinos swamped by wild conspiracy theories. Politico. September 14, 2020. <https://www.politico.com/news/2020/09/14/florida-latino-disinformation-413923>.
35. Wong Q. Facebook ad boycott: Why big brands "hit pause on hate." CNET. July 30, 2020. <https://www.cnet.com/news/facebook-ad-boycott-how-big-businesses-hit-pause-on-hate/>.
36. Hsu T and Lutz E. More Than 1,000 Companies Boycotted Facebook. Did It Work? The New York Times. 1. August 2020. <https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html>.
37. Murphy L and Relman Colfax. Facebook's Civil Rights Audit – Final Report. 2020.
38. Wagner K. Facebook Limits Ad Targeting That Some Linked to Race. Bloomberg. August 11, 2020. <https://www.bloomberg.com/news/articles/2020-08-11/facebook-further-limits-advertisers-ability-to-target-by-race>.
39. Abril D. The Facebook ad boycott ended months ago. But some big companies continue the fight. Fortune. November 7, 2020. <https://fortune.com/2020/11/07/facebook-ad-boycott-big-brands-lego-clorox-verizon-microsoft-hp/>.
40. Facebook. Ad Library. 2020. .
41. Department of Housing and Urban Development. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard. 2020. <https://www.federalregister.gov/documents/2020/09/24/2020-19887/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard>.
42. Department of Housing and Urban Development. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard. 2019. <https://www.federalregister.gov/documents/2019/08/19/2019-17542/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard>.

43. Jillson E. Aiming for truth, fairness, and equity in your company's use of AI. Federal Trade Commission. 2021. <https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>.
44. Sandberg S. Making Progress on Civil Rights – But Still a Long Way to Go. Facebook. 2020. <https://about.fb.com/news/2020/07/civil-rights-audit-report/>.
45. Rooney K and Korram Y. Tech companies say they value diversity, but reports show little change in last six years. CNBC. June 12, 2020. <https://www.cnbc.com/2020/06/12/six-years-into-diversity-reports-big-tech-has-made-little-progress.html>.
46. Williams M. Building a More Diverse Facebook. Facebook. 2014. <https://about.fb.com/news/2014/06/building-a-more-diverse-facebook/>.
47. Airbnb. A new way we're fighting discrimination on Airbnb. Airbnb Resource Center. 2020. <https://www.airbnb.com/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201>.
48. Frankfurt Kurnit Klein & Selz PC. Advertising Self-Regulatory Codes Around the World Generally Prohibit Discrimination, According to New ICAS Report. Lexology. August 3, 2020. <https://www.lexology.com/library/detail.aspx?g=35ea181f-1b71-45d2-a839-28c0ea8b5ebb>.
49. Social Security Administration. The First Social Security Number and the Lowest Number. <https://www.ssa.gov/history/ssn/firstcard.html>.
50. Igo S. The Known Citizen: A History of Privacy in Modern America. Harvard University Press. 2018.
51. Social Security Administration. The SSN Numbering Scheme. <https://www.ssa.gov/history/ssn/geocard.html>.
52. Puckett C. The Story of the Social Security Number. Social Security Bulletin. Vol 69. No 2. 2009. <https://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html>.
53. Long W. Social Security numbers issued: A 20-year review. Social Security Bulletin1. Vol 56. No 1. 1993. <https://www.ssa.gov/policy/docs/ssb/v56n1/v56n1p83.pdf>.
54. Social Security Administration. Report to Congress on Options for Enhancing the Social Security Card. 1997.
55. Social Security Administration. Social Security Number Randomization. 2011. <https://www.ssa.gov/employer/randomization.html>.
56. Social Security Administration. Social Security Number Randomization Frequently Asked Questions. <https://www.ssa.gov/employer/randomizationfaqs.html>.

57. Acquisti A and Gross R. Predicting Social Security numbers from public data. Proceedings of the National Academy of Sciences of the United States of America. 2009. <https://doi.org/10.1073/pnas.0904891106>.
58. Social Security Administration. Executive Order 9397 Numbering System for Federal Accounts Relating to Individual Persons. .
59. Malone K and Smith R. XXX-XX-XXXX. NPR. March 14, 2018. <https://www.npr.org/transcripts/593603674>.
60. Social Security Administration. Social Security History. 2005. <https://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html>.
61. US Court of Appeals S C. Cassano v. Carb. 2006. <https://casetext.com/case/cassano-v-carb>.
62. Kim J, Shin H-C, Rosen Z, Kang J, Dykema J, and Muennig P. Trends and Correlates of Consenting to Provide Social Security Numbers: Longitudinal Findings from the General Social Survey (1993–2010). Field Methods. Vol 27. No 4. 348–362. February 23, 2015. <https://doi.org/10.1177/1525822X15572334>.
63. Swendiman K and Lanza E. The Social Security Number: Legal Developments Affecting Its Collection, Disclosure, and Confidentiality. 2014.
64. Soto G. The Unexpected Costs of Identity Theft. Experian. September 30, 2020. <https://www.experian.com/blogs/ask-experian/what-are-unexpected-costs-of-identity-theft/>.
65. Franceschi-Bicchierai L. Equifax Was Warned. Motherboard. October 26, 2017. <https://www.vice.com/en/article/ne3bv7/equifax-breach-social-security-numbers-researcher-warning>.
66. Sweeney L, Yoo J S, and Zang J. Voter Identity Theft: Submitting Changes to Voter Registrations Online to Disrupt Elections. Technology Science. No 2017090601. 2017. <https://techscience.org/a/2017090601/>.
67. Internal Revenue Service. Truncated Taxpayer Identification Numbers (TTIN). 2018. <https://www.irs.gov/government-entities/federal-state-local-governments/truncated-taxpayer-identification-numbers>.
68. US Government Accountability Office. IDENTITY THEFT AND TAX FRAUD: Enhanced Authentication Could Combat Refund Fraud, but IRS Lacks an Estimate of Costs, Benefits and Risks. 2015.
69. Treasury Inspector General for Tax Administration. Results of the 2019 Filing Season. 2020.
70. Internal Revenue Service. Get An Identity Protection PIN (IP PIN). 2021. <https://www.irs.gov/identity-theft-fraud-scams/get-an-identity-protection-pin>.

71. Internal Revenue Service. Secure Access: How to Register for Certain Online Self-Help Tools. 2021. <https://www.irs.gov/individuals/secure-access-how-to-register-for-certain-online-self-help-tools>.
72. Ollstein A M and Ravindranath M. Getting it right: States struggle with contact tracing push. Politico. May 17, 2020. <https://www.politico.com/news/2020/05/17/privacy-coronavirus-tracing-261369>.
73. Wroe E and Oza S. Maximizing the Impact of Contact Tracing for COVID-19: The Importance of Human-Centered and Equity-Driven Programming. Harvard Health Policy Review. 2021. <http://www.hhpronline.org/articles/2021/3/4/maximizing-the-impact-of-contact-tracing-for-covid-19-the-importance-of-human-centered-and-equity-driven-programming>.
74. DeCosta-Klipa N. Massachusetts is testing a digital COVID-19 exposure app. Boston.com. April 5, 2021. <https://www.boston.com/news/coronavirus/2021/04/05/massachusetts-covid-tracing-app-massnotify>.
75. Horowitz J and Satariano A. Europe Rolls Out Contact Tracing Apps, With Hope and Trepidation. The New York Times. June 16, 2020. <https://www.nytimes.com/2020/06/16/world/europe/contact-tracing-apps-europe-coronavirus.html>.
76. Cross J. Apple's COVID-19 exposure notification API: What it is and how it works in iOS 13.5. Macworld. May 27, 2020. <https://www.macworld.com/article/3545329/apples-covid-19-exposure-notification-api-what-it-is-and-how-it-works.html>.
77. Griset R. Virginia leads nation in COVID-19 app use. Virginia Business. December 11, 2020. <https://www.virginiabusiness.com/article/virginia-leads-nation-in-covid-19-app-use/>.
78. Hinch R, Probert W, Nurtay A, Kendall M, and Wyman C. Effective Configurations of a Digital Contact Tracing App: A report to NHSX.
79. University of Oxford. Digital contact tracing can slow or even stop coronavirus transmission and ease us out of lockdown. 2020. <https://www.research.ox.ac.uk/Article/2020-04-16-digital-contact-tracing-can-slow-or-even-stop-coronavirus-transmission-and-ease-us-out-of-lockdown>.
80. Mello M M and Wang C J. Ethics and governance for digital disease surveillance. Science. Vol 368. No 6494. 951 LP – 954. May 29, 2020. <https://doi.org/10.1126/science.abb9045>.
81. Ranisch R et al. Digital contact tracing and exposure notification: ethical guidance for trustworthy pandemic management. Ethics and Information Technology. 2020. <https://doi.org/10.1007/s10676-020-09566-8>.
82. Google and Apple. Exposure Notification: Frequently Asked Questions. 2020.

83. Bradford L, Aboy M, and Liddell K. COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. *Journal of Law and the Biosciences*. Vol 7. No 1. July 25, 2020. <https://doi.org/10.1093/jlb/ljaa034>.
84. Shachar C. Protecting Privacy In Digital Contact Tracing For COVID-19: Avoiding A Regulatory Patchwork. *Health Affairs*. 2020. <https://www.healthaffairs.org.ezp-prod1.hul.harvard.edu/doi/10.1377/hblog20200515.190582/full/>.

Appendix

Chapter 2

How Facebook’s Advertising Algorithms Can Discriminate By Race and Ethnicity

Appendix A – The Share of Multicultural Affinity Groups (2020) vs. Cultural Interest Groups (2021) in Lookalike Audiences

In 2020, I found that Lookalike audiences had similar shares of users with the “African-American (US)” attribute at 40 – 48% regardless of whether a list of only African-American, Asian, or White voters was used to create the Lookalike audience (Figure 2.32). In 2021, Lookalike audiences based on Asian or White voters had far lower shares of users interested in “African-American Culture” at 13-14% than the 34% share of the Lookalike audience based on African-American voters (Figure 2.32).

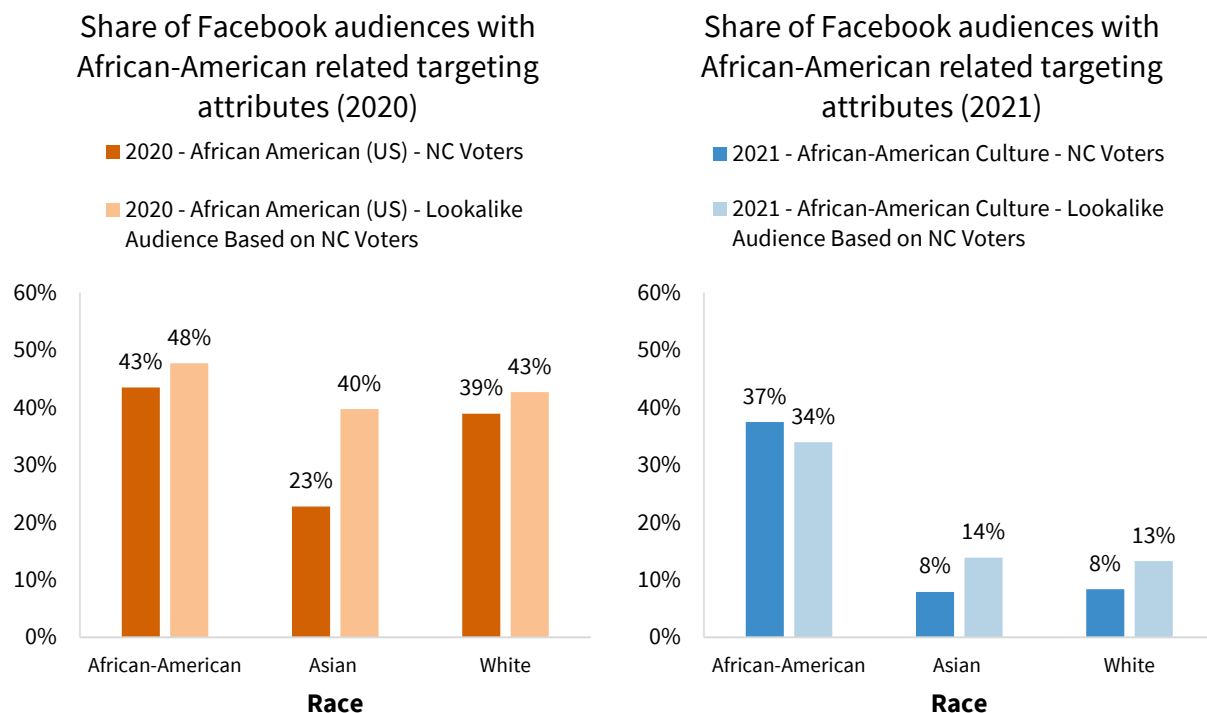


Figure 2.32. Share of Facebook Audiences with African-American Related Targeting Attributes in 2020 and 2021.

I found that the opposite pattern for Facebook’s Asian-American related targeting attributes. In 2020, 2.1% of the Lookalike audience based on Asian voters can be reached by targeting “Asian American (US)” while only 0.2% of the Lookalike audiences based on African-American or White voters (Figure 2.33). In 2021, approximately 2-3% of the Lookalike audiences based on African-American, Asian, or White voters can be targeted by Facebook’s “Asian American Culture” option (Figure 2.33).

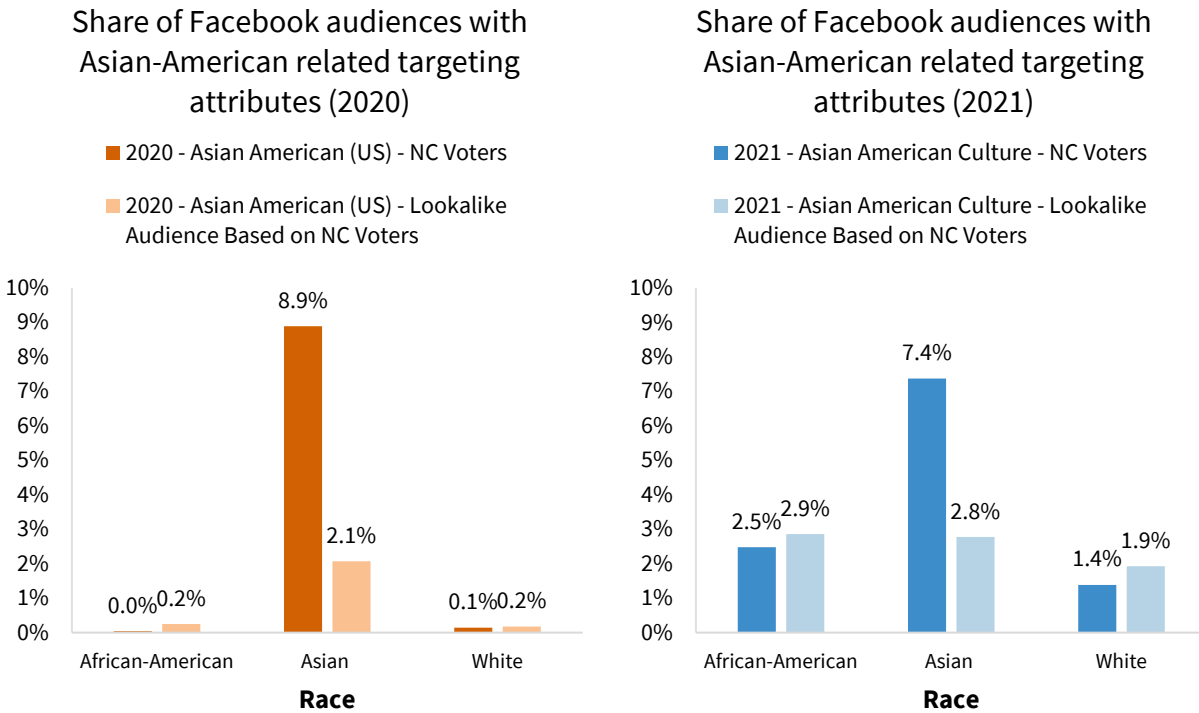


Figure 2.33. Share of Facebook Audiences with Asian-American Related Targeting Attributes in 2020 and 2021.

In both 2020 and 2021, Lookalike audiences based on Hispanic voters were significantly more likely to be reached by Facebook’s Hispanic-American targeting attributes than Lookalike audiences based on Non-Hispanic voters. In 2020, 13% of the Lookalike audience based on Hispanic voters can be targeted with the “Hispanic (US – All)” option, compared to 2% of the Lookalike audience based on Non-Hispanic voters (Figure 2.34). In 2021, those rates are 7% and 2% for the “Hispanic American Culture” targeting option (Figure 2.34).

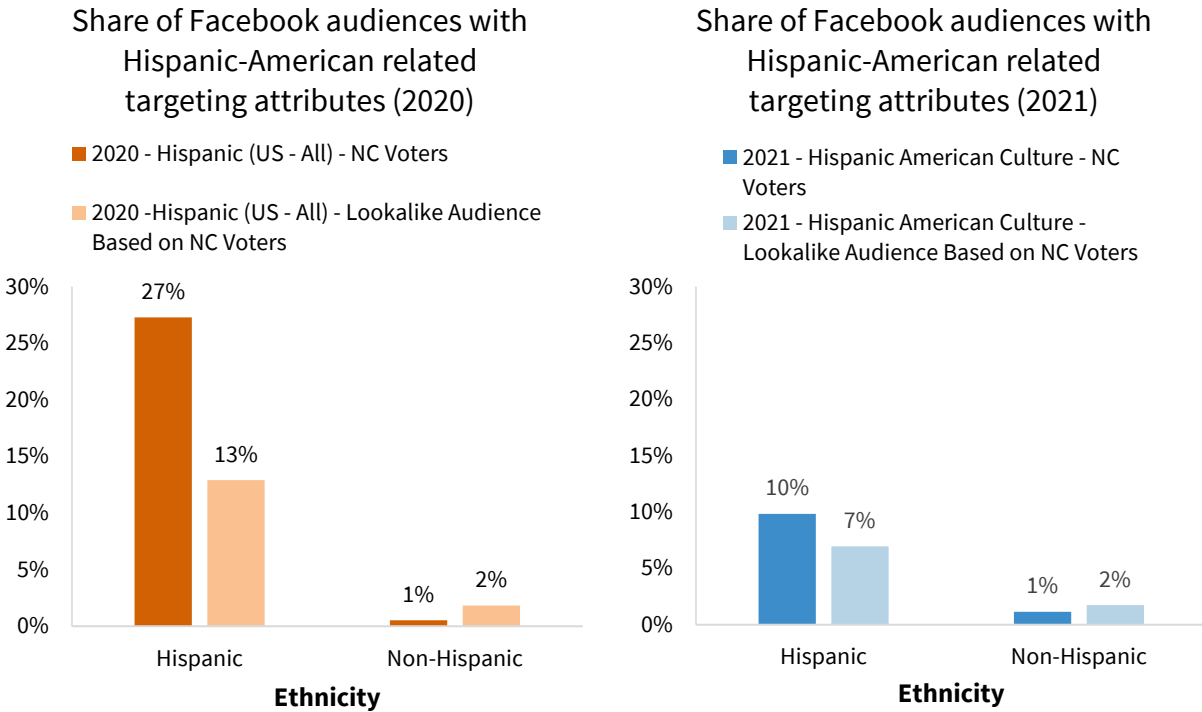


Figure 2.34. Share of Facebook Audiences with Hispanic-American Related Targeting Attributes in 2020 and 2021.

Appendix B – Lookalike Audiences Overlap Analysis

I used Facebook’s audience overlap tool to study how many users are shared between two Lookalike audiences based on different racially-biased or ethnically-biased lists of voters. For example, Figure 2.35 shows that the Lookalike audience based on White voters and the Lookalike audience based on African-American voters shared 19% of the same Facebook users in 2021. Facebook didn’t allow the audience overlap tool to be used for Special Ad audiences.

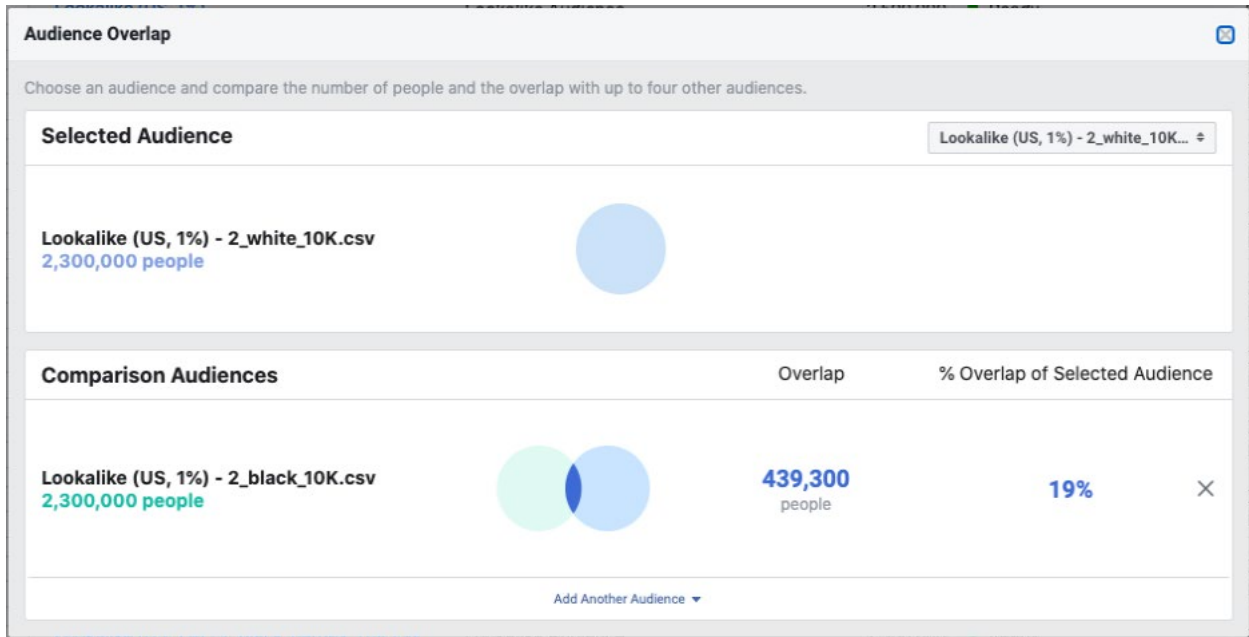


Figure 2.35. Example of Audience Overlap Between a Lookalike Audience Based on White Voters Vs. a Lookalike Audience Based on African-American Voters in 2021.

Lookalike audiences based on voters of different races generally had nearly 1/3 or less of their users be in both groups and the overlap rates decreased from 2020 to 2021. In 2020, 25-26% of the Lookalike audiences based on African-American versus White voters or African-American versus Asian voters overlapped (Figure 2.36). In 2021, that rate decreased to 18-19% (Figure 2.36). Lookalike audiences based on White versus Asian voters had higher overlap rates of 36% in 2020 and 29% in 2021 (Figure 2.36). Finally, the Lookalike audience based on Hispanic voters shared 46% of its users with the Lookalike audience based on Non-Hispanic voters in 2020 and 36% in 2021 (Figure 2.36).

Share of overlap in Lookalike Audiences based on lists of NC voters with different traits

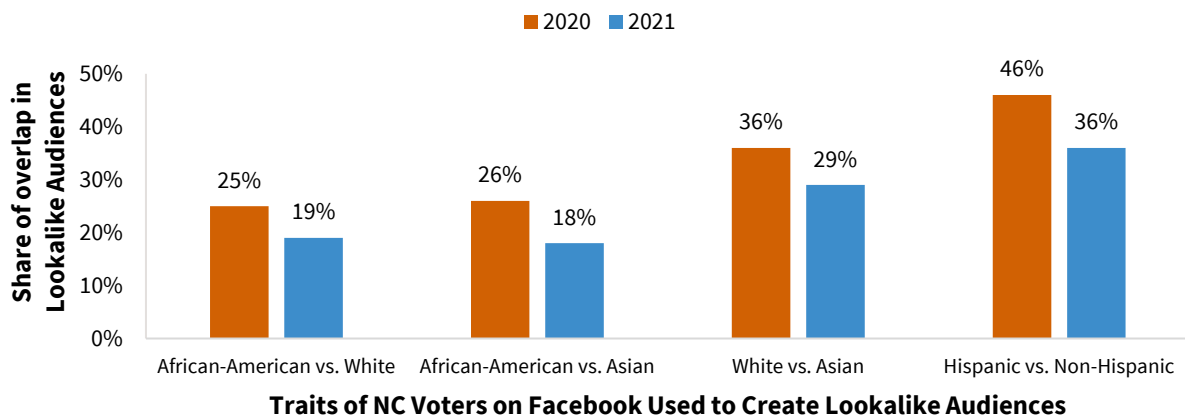


Figure 2.36. Share of Overlap in Lookalike Audiences Based on Lists of NC Voters with Different Traits.

Appendix C – Facebook Ad Library Example Screenshots and Restrictions and Notices for Special Ads Related to Housing, Employment, or Credit

← Search Results

Summary Data

Advertisers often use the same image or video and text to create ad campaigns with different start dates, locations or budgets. This section contains the collective data for 912 ads.

Amount Spent
The estimated total money this advertiser spent on these ads.
[Learn more](#)

Amount Spent
\$2K - \$2.5K (USD)

Impressions
The number of times these ads were seen on a screen. This may include multiple views by the same people.
[Learn more](#)

Impressions
150K - 175K

Who Was Shown These Ads
The age and gender breakdowns of people who saw these ads.

Age Group	Men	Women	Unknown
18-24	5%	2%	0%
25-34	12%	5%	0%
35-44	15%	10%	0%
45-54	13%	12%	0%
55-64	8%	9%	0%
65+	4%	5%	0%

Where These Ads Were Shown
The regions where people who saw these ads are located.

Region	Percentage
Pennsylvania	35%
Wisconsin	16%
North Carolina	10%
Illinois	10%
Michigan	8%
New Jersey	5%
Florida	3%

[See More](#)

Oct 22, 2020
The date this group of ads began running.

912 ads Any filters you applied to the search results are also applied to this group of ads. To adjust the filters, go back to the search results.

Inactive
Oct 31, 2020 - Nov 1, 2020
ID: 349058653014363

Donald J. Trump
Sponsored - Paid for by DONALD J. TRUMP FOR PRESIDENT, INC.
President Trump is coming to town! Get your free tickets now >>>

[HTTPS://WWW.DONALD.JTRUMP.COM/](https://www.donaldjtrump.com/)
BREAKING NEWS: Join President Trump RSVP NOW>>>

[Sign Up](#)

Amount spent (USD): <\$100
Potential Reach: >1M people

[See Ad Details](#)

Inactive
Oct 31, 2020 - Nov 1, 2020
ID: 373334713869779

Donald J. Trump
Sponsored - Paid for by DONALD J. TRUMP FOR PRESIDENT, INC.
President Trump is coming to town! Get your free tickets now >>>

[HTTPS://WWW.DONALD.JTRUMP.COM/](https://www.donaldjtrump.com/)
LAST CHANCE TO GET YOUR TICKETS RSVP NOW>>>

[Sign Up](#)

Amount spent (USD): <\$100
Potential Reach: >1M people

[See Ad Details](#)

Inactive
Oct 31, 2020 - Nov 1, 2020
ID: 632459814095239

Donald J. Trump
Sponsored - Paid for by DONALD J. TRUMP FOR PRESIDENT, INC.
President Trump is coming to town! Get your free tickets now >>>

[HTTPS://WWW.DONALD.JTRUMP.COM/](https://www.donaldjtrump.com/)
LAST CHANCE TO GET YOUR TICKETS RSVP NOW>>>

[Sign Up](#)

Amount spent (USD): <\$100
Potential Reach: >1M people

[See Ad Details](#)


[See More](#)

[About Ads and Data Use](#)

Figure 2.37. Example Political Ad on Facebook’s Ad Library.

Ad Details

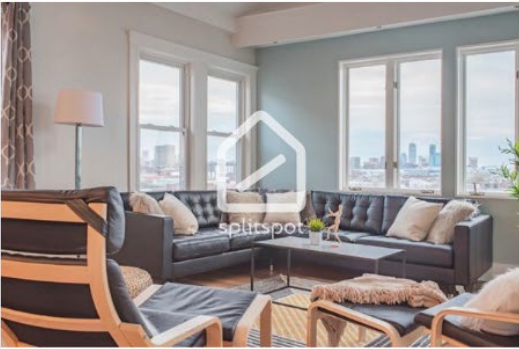
About the Ad



SplitSpot
Sponsored
ID: 870646383754365

SplitSpot makes renting an apartment in Boston quick, easy, and affordable! Follow us and visit our website to get started:

- ✓ Minimal upfront cost. No broker's fee!
- ✓ Flexible leases as short as 4-months!
- ✓ Vet your roommates before you commit!
- ✓ Transfer to any other SplitSpot unit whenever you'd like for no extra cost!




SPLITSPOT.COM
Flexible Boston Room Rentals | SplitSpot
SplitSpot is simplifying renting for both renters and landlords. We offer flexible, month-to-mont... [Learn More](#)

About the Page

[See Ads](#)



SplitSpot

 @SplitSpot
95 likes • Real Estate Service

 @split_spot
106 followers

More info

We are here to make renting in Boston better - for both renters and landlords.

[About Ads and Data Use](#)

Figure 2.38. Example Housing Ad on Facebook's Ad Library.

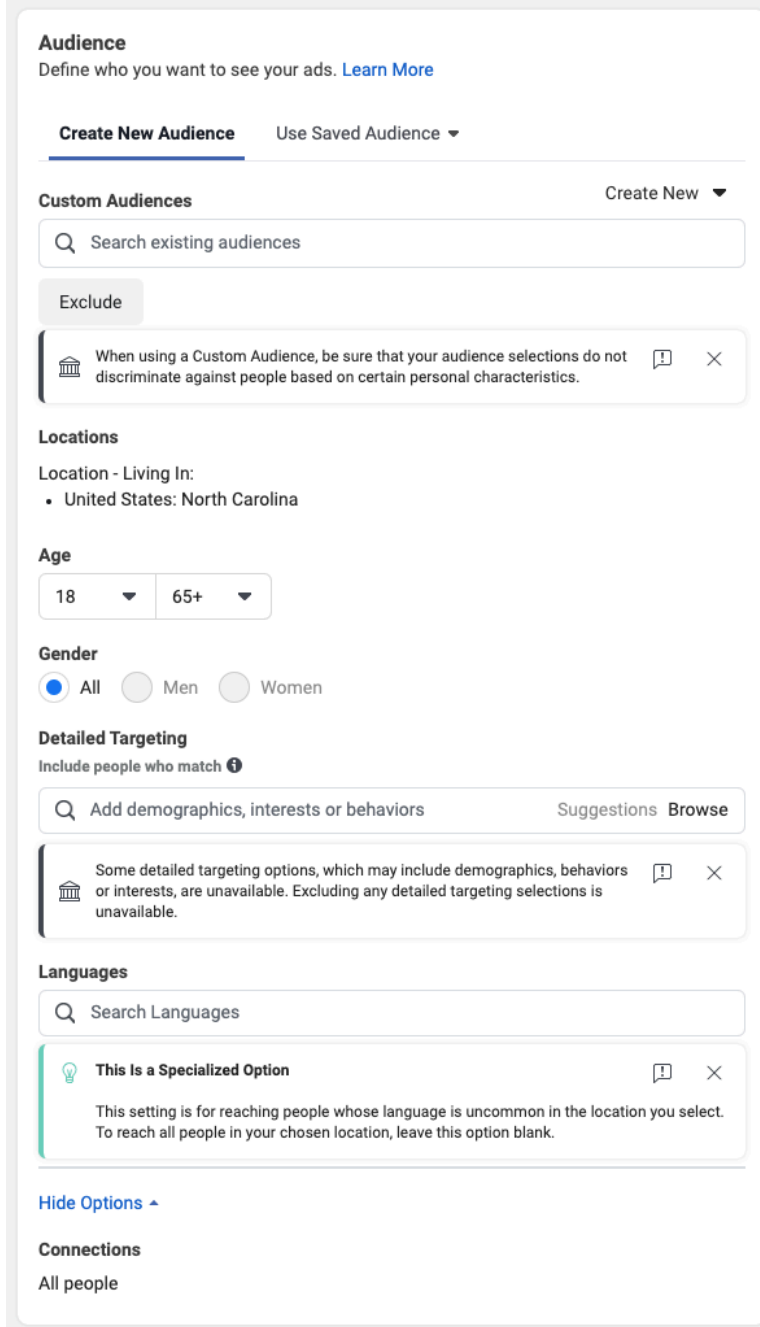


Figure 2.39. Example Anti-Discrimination Restrictions and Notices on Facebook’s Ad Planning Tool for Housing, Employment, or Credit-Related Ads.

Chapter 3

How Were Social Security Numbers Assigned?

Appendix A

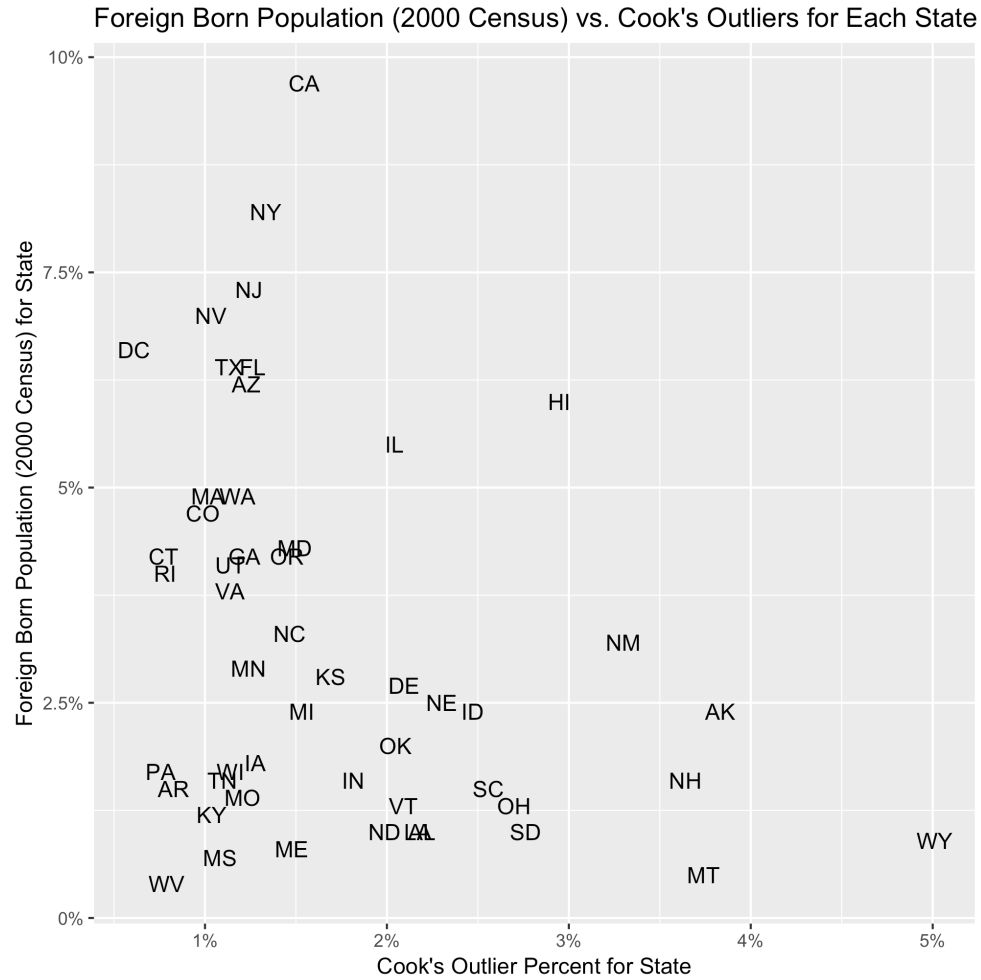


Figure 3.14. Foreign born percentage of the population in each state that entered the US from 1990 to 2000 from the 2000 Decennial Census versus the percentage of SSNs identified as Cook's distance outliers in each state from 1995 - 2011.

Appendix B

Median Error of SSN Prediction By State (1995 - 2011)

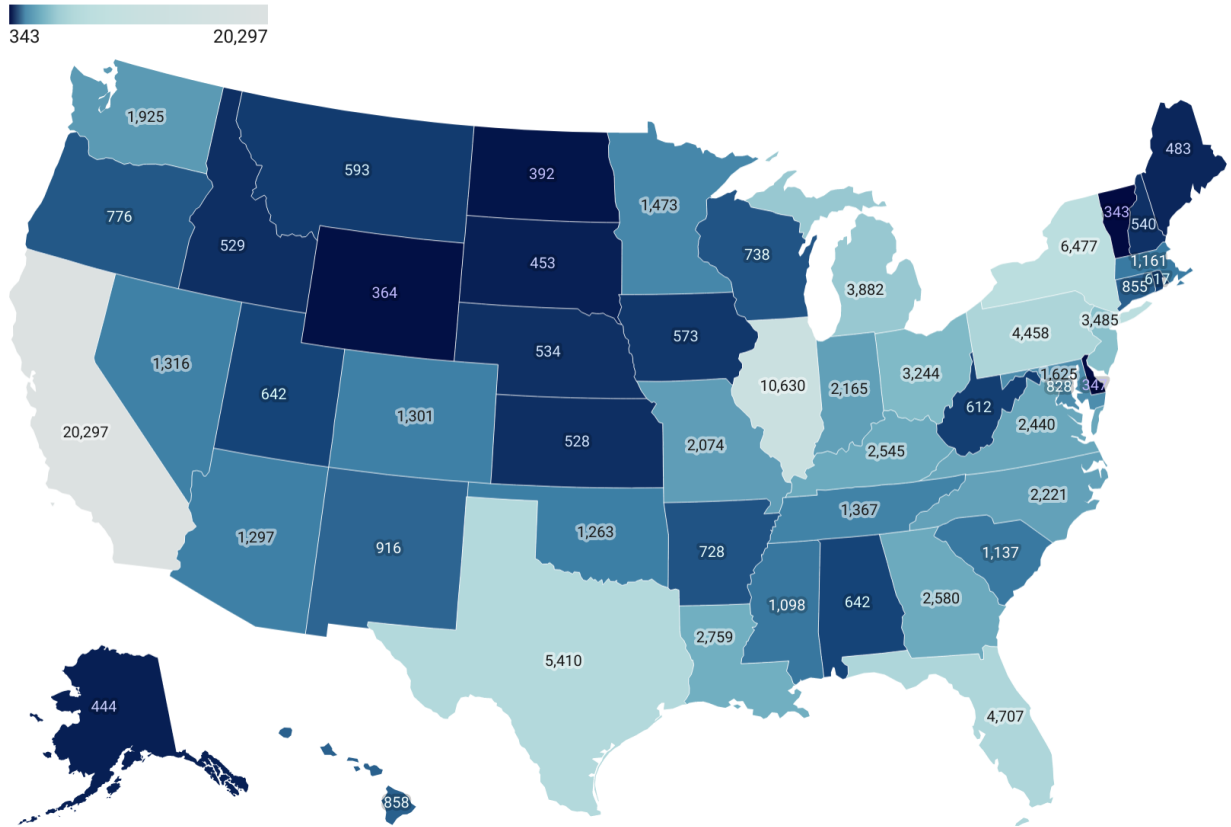


Figure 3.15. Median Error of SSN Prediction By State (1995 – 2011).

Predictive Accuracy of First 5 Digits of SSN By State (1995 - 2011)

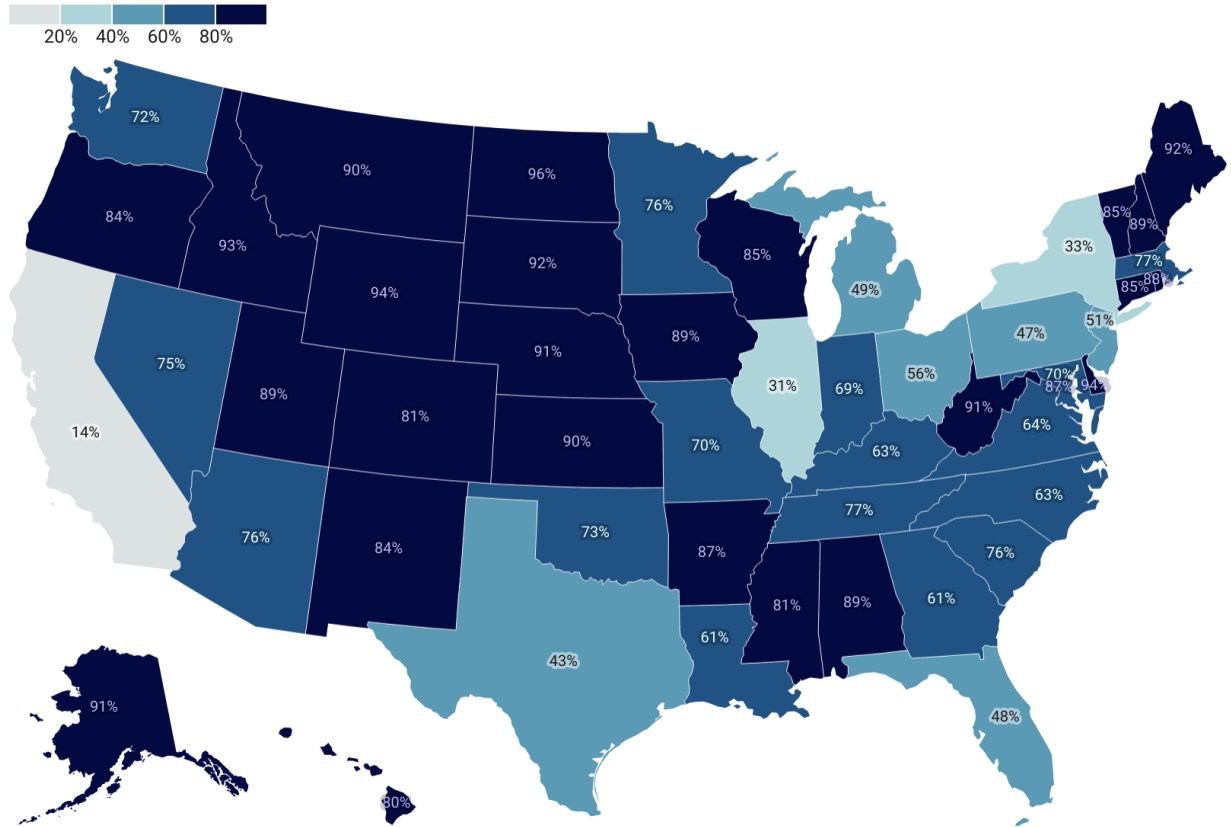


Figure 3.16. Predictive Accuracy of First 5 Digits of SSN By State (1995 – 2011).

Predictive Accuracy of First 6 Digits of SSN By State (1995 - 2011)

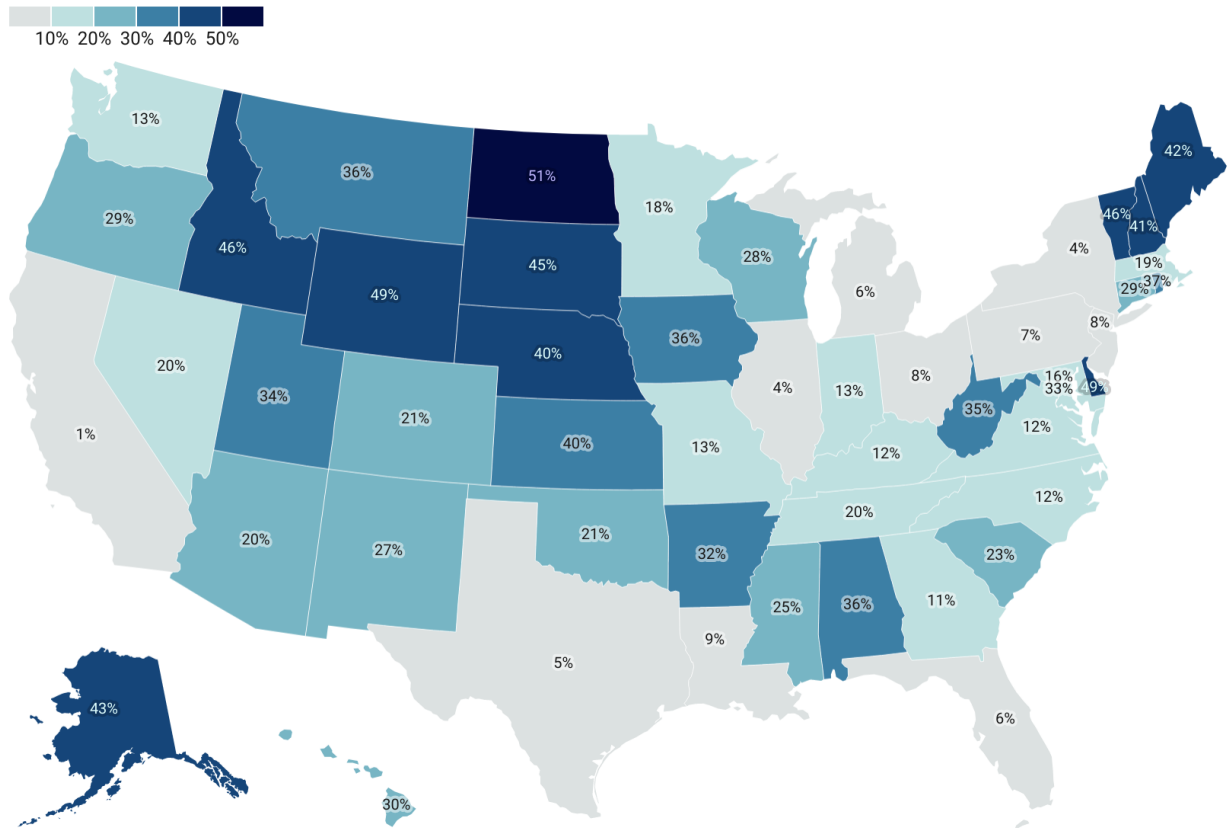


Figure 3.17. Predictive Accuracy of First 6 Digits of SSN By State (1995 – 2011).

Chapter 4

Building A Collocation Detection System Using A Wi-Fi Sensor Array for COVID-19 Contact Tracing in A University Setting

Appendix A

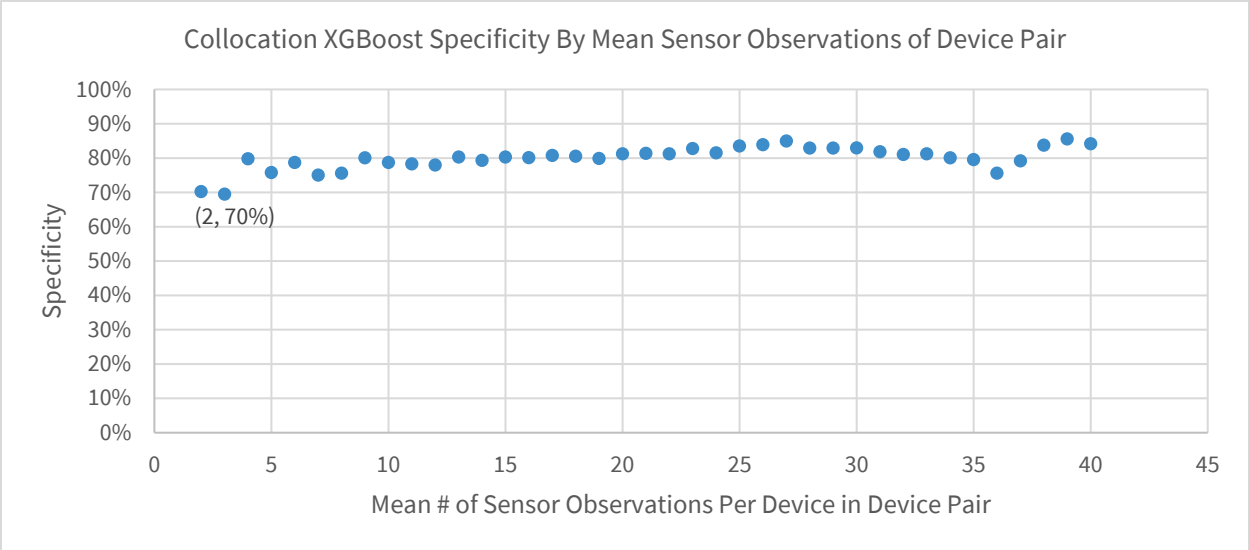
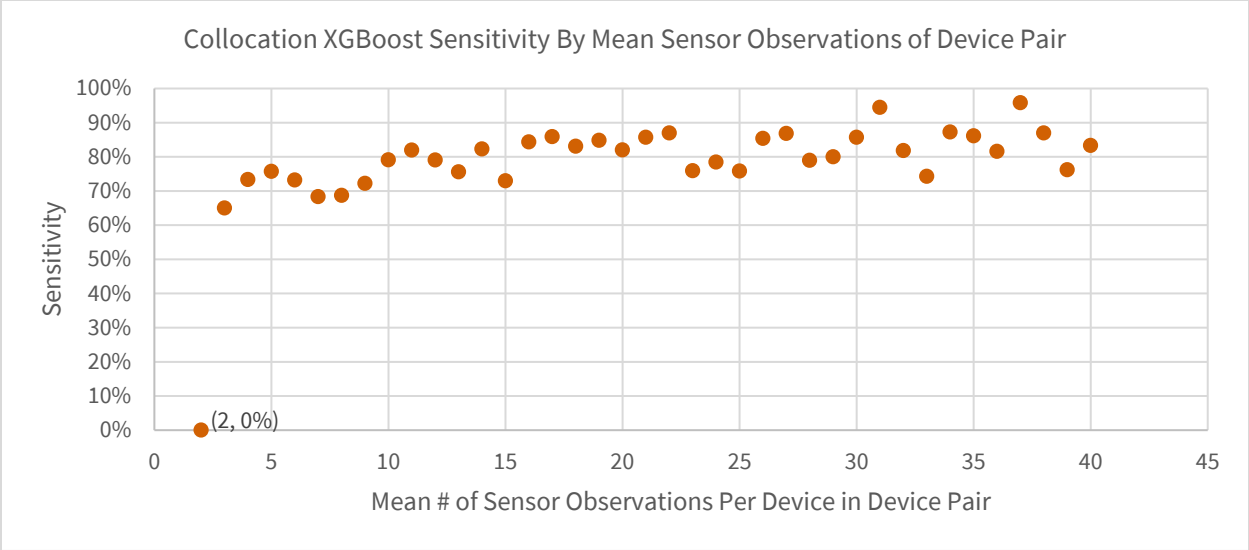


Figure 4.12. Collocation XGBoost Sensitivity and Specificity By Mean Number of Sensor Observations Per Device in Device Pair.

Appendix B

Online learning after deployment can help refine the existing location prediction and collocation detection models through multiple methods shown in Figure 4.13. The location predictions per device can be used as potential new fingerprints to update the location prediction models, since different subsets of sensors may be observing the same device at a given location for each timestamp. The number of possible fingerprints collected in the initial offline phase will always be limited due to how time-consuming it is compared to the significant increase in data after deployment. Similarly, the device pair collocation data can be used to update the collocation detection models. This process also allows the models to adapt over time as old sensors are removed and new sensors are added as long as it's a gradual process. Finally, the data user may learn more information about the accuracy of a

given location prediction or collocation status due to accessing external data sources, talking to the device owner, or other methods and be able to provide that feedback as part of an online learning system for both models.

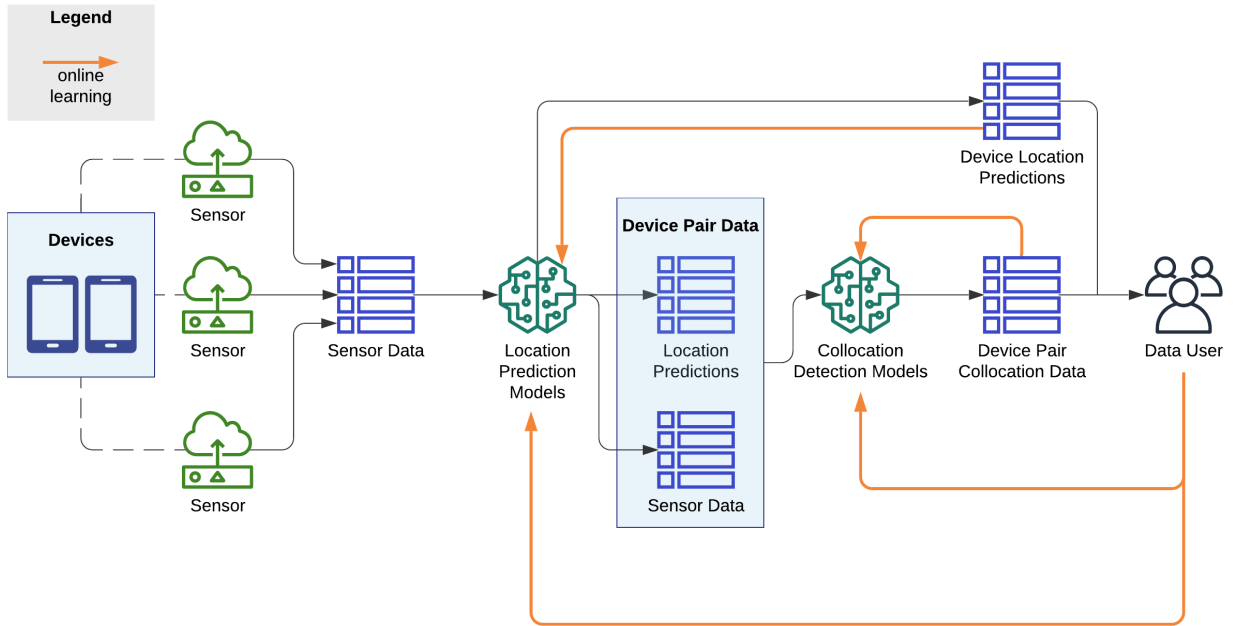


Figure 4.13. Diagram of the possibilities for online reinforcement learning in TraceFi.