Measuring Biases in a Data Broker's Coverage

Levi Kaplan
Northeastern University
USA

Alan Mislove
Northeastern University
USA

Piotr Sapieżyński
Northeastern University
USA

ABSTRACT

In the absence of a standardized form of identification in the United States, various businesses and organizations turn to the data broker industry to confirm the identity of, or perform background checks on, their clients. Often, potentially life-changing decisions depend on a successful and accurate match between the client's identity and data broker records; for example, decisions about housing, credit, employment, and—more recently—even access to vaccines against a global pandemic, can all be based in part on information from data brokers. However, the data brokers provide little transparency and it is notoriously difficult for researchers to study these companies at scale. In this work, we develop a measurement methodology to understand the coverage of one such data broker: Experian. We demonstrate that Experian's coverage of adults in North Carolina by is not only far from perfect, but is also worse for individuals who are more likely to be in historically disadvantaged groups. Our results indicate that younger populations as well as ethnic minorities and those living in lower income areas are less likely to be present in data broker databases, and even if they are, their data is more likely to be inaccurate than for white individuals and those living in more wealthy locations. These biases can potentially further exacerbate real-life societal divides along ethnic and economic lines, as they make access to essential life opportunities even more difficult for the most vulnerable populations.

1 INTRODUCTION

Data brokers are corporate entities whose business model is based on collecting, analyzing, and reselling data about individuals. They obtain their information from a variety of sources (e.g., public records, loyalty cards, web tracking, etc), and combine it to build rich profiles of individuals: their financial details, education and employment history, health status, and even religious beliefs, political views, and ethnicity. The data brokers then either sell the raw or derived data, or they provide data-based services, such as estimates of creditworthiness [2, 19]. From identity verification [10], credit scoring [12], and personalized advertisements [9] to housing and employment decision support [21, 37], many aspects of daily life are now mediated by data brokers [2].

Despite their ubiquity, there are a number of concerns surrounding the data broker industry. In the U.S., individuals have limited rights regarding the data about them: outside of a few areas with special legal protections (e.g., credit scores), they are not asked for consent to data collection, have no right to view data about them, cannot always petition to have errors corrected, and have no right to ask that their data be removed [2]. Worse yet, data brokers have been shown to have a poor security posture, and have been the victims of multiple data leaks in recent years [8, 20], affecting hundreds of millions of people worldwide. Finally, data brokers are notoriously opaque [41], making it difficult to analyze and understand the coverage and accuracy of their data.

While prior work has demonstrated that data brokers can have significant inaccuracies [41], there is an additional concern that has become more acute as life opportunities are increasingly mediated by data brokers. Specifically, historically disadvantaged groups are often less likely to show up in "official" databases: those with lower socio-economic status are less likely to own property [13], be registered to vote [5], or have access to mainstream financial services [25]. Thus, if data brokers use these sources, could historically disadvantaged groups be *even further* disadvantaged, as access to life opportunities is now increasingly dependent on data brokers in whose databases they are less likely to appear?

In this work, we use a unique opportunity provided by one data broker to understand their coverage at finer detail than was previously possible. Specifically, we identified one data broker, Experian, which until recently offered "self-service" web interfaces for marketers to (a) buy custom mailing lists of personally identifiable information (PII), including physical addresses, of people who match specified attributes, and (b) append attributes to existing lists of PII.

We first use the data append interface to study the coverage and accuracy of data broker data for users in different racial groups. We selected four samples of 8,930 individuals with four different self-reported races from voter records in North Carolina, created PII lists, and asked Experian to "append" their birth year. Because the birth year is also in voter records, this methodology allowed us to measure both the coverage (how many records successfully had data appended?) as well as the accuracy of the matches (how many appended birth years were correct?) that Experian provides, without having to purchase any data about these users that was not already contained in the publicly available voter records. Our measurements show that there are stark differences in data quality along the lines of race, ethnicity, age, and economic status. For example, the data we purchased on white non-Hispanic Americans was 25% more likely to be accurate than that on Hispanic Americans of any race. Further, only 32% of Hispanic voters below 26 were correctly represented by Experian, compared to 65% of those above 54.

Finally, we perform a logistic regression on this data to disambiguate how the factors of race and ethnicity, age, gender, and poverty contribute to the problem. We show that the racial differences persist even when we control for age, gender, and poverty.

Taken together, our results demonstrate that Experian's coverage and accuracy may have significant discrepancies across different races, with non-white individuals generally being less likely to be covered and having less accurate data. We note that our work has a number of limitations, as it only studies a single data broker, only uses only one of their many services, and that particular service is not used for credit, housing, and employment decisions (we could not get access to those services). However, our work

sheds more light on the opaque industry of data brokers, and suggests that further scrutiny as well as a search for more reliable and less intrusive alternatives are needed, as these services become increasingly important in deciding which users receive important life opportunities.

2 RELATED WORK

We next detail related work on data brokers, the effect they have on determining life outcomes, and auditing approaches.

Data broker coverage Previous work has shown data brokers to collectively have data on a significant fraction of the population. For example, Venkatadri et al. [41] showed a combination of data brokers achieve 90% coverage of U.S. Facebook users. Given that three quarters of U.S. individuals have a Facebook account [18], this represents a majority of Americans. While the combined coverage is high, it does vary between individual data brokers and, for all of them, appears to be worse for counties with higher poverty rates [41], a conclusion our results corroborate. Importantly, previous research investigated the coverage of data brokers via Facebook's advertising platform, and was therefore limited in the level of detail it could examine data brokers data with. Our current work addresses this shortcoming by avoiding the need to use such a third party.

Low coverage can lead to barring important services from already vulnerable populations. Experian and other data brokers aggregate information for reporting credit scores, which are used for identification and risk assessment for a number of services, including employment, insurance, and loans. Missing or inaccurate credit scores can lead to being barred from employment [35, 37], paying more for car insurance [3, 14], and being offered loans with higher interest rates [1]. More recently, Experian has been used for identity verification for those wanting to receive a COVID-19 vaccine; imperfect coverage by data brokers has resulted in some being barred from inoculation [17].

Data broker accuracy Data brokers are used by 90% of landlords [21] to perform background checks on would-be tenants and 47% of employers to check the credit score of applicants [37]. Unfortunately, the data offered by the brokers is not always accurate. For example, The Markup recently described how background checks run by data brokers have shown convictions that should be expunged or sealed [22]. Further, prior work [41] found at least 40% of the attributes data brokers had on people to be inaccurate or no longer accurate, and found the errors to be present for people who had other accurate attributes. That work analyzed in-depth surveys on the accuracy of multiple attributes from 200 users who installed a browser plugin and used it for over a month, potentially producing a biased sample. Here, we focus instead on measuring the accuracy of just one attribute but for tens of thousands of users, selected at random from public voter records. Other researchers studied accuracy of data broker inferences through the lens of browser cookies, rather than PII, to identify individuals. They found mulitple data brokers whose gender inference accuracy was below 50% for adult male subjects, compared to a random guess accuracy of 50% with a binary gender label [27].

Additionally, the U.S. Federal Trade Commission estimated in 2013 that roughly 10 million U.S. individuals had an error on a credit report—controlled by companies like Experian—which was severe enough to cause higher borrowing costs [1]. Requesting to see credit reports and correcting any inaccuracies can be a time consuming, exhausting process; in some cases individuals have spent years trying to correct flawed reporting [26]. Some data brokers provide individuals access to their data in the system [36], but these reports contain carefully curated (and often incorrect [23, 33]) data and do not include the inferences drawn from them [30].

Auditing data brokers Data brokers have proven difficult to audit due to limited regulation and their desire to keep the scope and content of the data they collect hidden from public scrutiny. While a number of studies have developed tools to measure online data aggregators [15, 16, 31, 32, 38, 39], offline data brokers have remained difficult to study. Despite this, prior research has found a few opportunities. Previous work [41] used Facebook's "partner categories" to investigate the sources of data obtained by offline data brokers. From there, they used transparency enhancing advertisements (Treads) [40] to deduce accuracy, surveying 300 participants on the data attributes that Facebook and the data broker had on them. In this work, we leverage a different approach, instead using an interface provided by the data broker itself to gather data on the number of individuals present in their database and the accuracy of their data. Our approach allows us to perform the study on a larger scale (with 36,000 users instead of 300) and include race into our analysis.

3 METHODOLOGY

We now detail the methodology used to collect and analyze the data in this paper. In brief, we use three separate data sources: two Experian web interfaces (data append and custom mailing lists), North Carolina voter records, and the U.S. Census. We explore the ethical considerations regarding all of the data collection in Appendix B.

3.1 Data Append

Until April 30, 2021, Experian offered a "data append" web interface. In this interface, a customer would first specify a list of attributes they wished to purchase, ranging from basic demographics (age, sex, race) to financial information (net worth, household income) to "market segments" (new parents, empty nesters, etc). The different attributes had different prices per record (from $0.007 per record to over $0.20 per record), and the minimum purchase was $500.[1] After selecting the attributes, the customer would upload a CSV file that contained the PII of the individuals on whom they wished to buy data. Upon payment, Experian's system would return the "appended" file containing the purchased attributes.

Additionally, the file also contained information that described how "certain" each record's match was. These included, from weakest to strongest, Non-Match, Geographic Match, Household Match, and Person Match. In our experiments, we treat anything other than Person Match as not matching. Finally, the file also described the certainty of the purchased attribute, in our case birth age, from weakest to strongest: None, Estimated, Exact. In our experiments we only verify the correctness of the Exact matches. Note that the

[1]For our experiments, the cost for purchasing the date of birth was $517.94 for 35,717 records, or $0.015 per record.

results are robust to various relaxations of these definitions, as we show in the Appendix.

Voter records We used publicly-available voter records from North Carolina as the source of PII to study the data append interface. These records are published on the web by the Board of Elections [29], and they contain each registered voter's name, address, sex, age, as well as self-reported race and ethnicity. When using these records, we only selected voters who were listed as Active. Note, that the Board of Elections releases the file every week, thus keeping the age reported in the voter records up-to-date. Also note, that "Hispanic" and "non-Hispanic" are descriptions of ethnicity, while "Asian", "Black", and "white" are descriptions of race; we use these categories as they are the ones provided in the U.S. Census and voter records. For the measurements we select only non-Hispanic Asian, Black, and white individuals and refer to these lists by the race. Our "Hispanic" list contains Hispanic individuals of any race.

3.2 Logistic regression modeling

Given the inequities in the society at large, demographic variables are correlated with wealth. In order to disabiguate the contributions of the various factors to the differences in coverage, we perform a logistic regression. We describe each individual in the dataset using the following independent variables: birth age (in years), female (true or false), Asian, Hispanic, Black, fraction of residents in poverty in the zipcode of residence. We standardize the birth age and poverty rate variables to enable easier interpretation of the coefficients. We then attempt to predict whether Experian fails to accurately represent that individual, i.e. their match is not a 'Person Match', or the age estimate is not 'Exact', or the reported age is more than 1 year different from the ground truth. For easier readability we guide the user through the interpretation of the model coefficients. First, we can translate the Intercept coefficient β_I to base risk of not being represented, using the following formula $\frac{e^{\beta}}{1+e^{\beta}} \approx 0.40$. This means that when all variables are equal to 0, i.e. for a white man of average age living in a zip-code with average povery rates, the risk of inaccurate representation is 40%. Further, we can interpret the sign of other coefficients as the direction of changes to the risk - positive coefficients mean increased risk, negative coefficients indicate lower risk. To compute the relative risk change when a value of a variable changes from 0 to 1 we use the following formula: $e^{\beta} - 1$ for positive coefficients and $1 - e^{\beta}$ for negative coefficients. For example, the coefficient of 0.6021 associated with the 'Hispanic' indicator in 1 means an 83% increase of risk when other variables are 0 (i.e. compared to a white man living in a zip-code with the average poverty rate). Finally, note that age and poverty rates are normalized, which means that a unit change corresponds to a change by a standard deviation, i.e. 16.3 years of age or 7.3 percentage points of poverty rates. Hence, the coefficient of -0.3811 associated with age indicates a 32% decrease of risk when age is increased by 16.3 years while other variables are equal to 0.

We report the performance of the model using the Area Under ROC curve (receiver operating characteristic curve). The easy interpretation of the metric in the context of our model is as follows: given two individuals where one is accurately represented and the other is not, how often does the model associate higher risk with the latter.

4 RESULTS

We now turn to present our analysis. First, we describe the Experian coverage and accuracy differences for different races and ethnicities, as well as age and location, among registered voters of North Carolina. Next, we aim to verify whether the trends shown with North Carolina data hold for the rest of the country.

4.1 Minorities represented less accurately

We first measure differences in coverage (whether Experian claims to have data on a given individual) and data accuracy (whether the reported birth age matches the ground truth) between races and ethnicities. The marketing materials for Experian's Identity Verification claim that it uses a "proprietary demographic database to immediately validate and correct important patient information: name, address, Social Security number, date of birth, phone number and county" [11]. We do not have access to the Identity Verification service and instead use the Data Append service, but it is not unreasonable to hypothesize that the two share a common data source, if not the same demographic database. In the following experiment, we assume that if a person's information, for example their birth date, is found in the broker's database but it does not match it, the identity verification may fail and they could still be denied access to opportunities.

First, we obtain publicly available North Carolina voter rolls where residents self report their race, ethnicity, full home address, and date of birth. We deem this data as reliable ground-truth, as it is self-reported by individuals who face potential criminal charges for mis-representation. Then, we select 8,930 voters with up-to-date registration from each of the following race and ethnicity groups: Asian non-Hispanic, Hispanic of any race, African American non-Hispanic, and white non-Hispanic. Our choice of individuals is random with the constraint that each age range has the same number of individuals in each racial group (for age ranges 18–24, 25–34, ... 85–94). We upload the resulting list of 35,720 names and home addresses to the Data Append service and purchase the Estimated Age of each person. Note that we do *not* use the date of birth as a key for Experian to identify the individuals in their database. Instead, we purchase this information from Experian to compare it to the ground truth obtained from the voter records.

Once we have obtained the results from Experian, we calculate the rates of highest certainty matches per race; within those we calculate the rates of individuals with the age reported correctly (within one year of the age reported in the voter records). Figure 1 summarizes the results. We note differences among races and ethnicities both in terms of coverage and accuracy. Hispanic voters of any race have highest rate of failing to match based on name and street address at 27% compared to white voters at 18%. That means, on average, a Hispanic voter is 50% more likely not to be identified in Experian data than an average white voter. Furthermore, there are notable disparities in accuracy even for the individuals whom Experian claimed to find a match for. The age of 17% of white voters is wrong by more than a year, compared to 23% of Hispanic and Asian voters. We used the G-test of goodness-of-fit [24] and

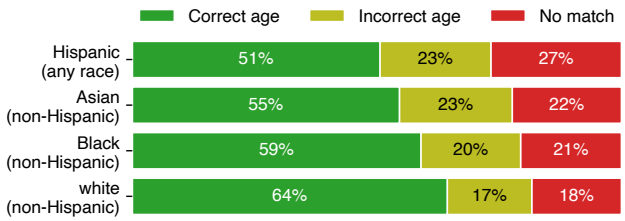

Figure 1: Differences in coverage and data accuracy among adult North Carolina population of different races. One in two Hispanic or one in three white voters are not represented or are represented inaccurately.

rejected the null hypotheses that the Correct/Incorrect/No-match counts of each pair of two races came from the same distribution, at $p_{val} < 0.001$ level, with Bonferroni correction for multiple testing. We opted for the G-test instead of the more popular chi-squared test because the latter is prone to false rejection of the null hypothesis for large sample sizes [24].

The compounding problems of lack of coverage and data inaccuracies would potentially lead to failure of identification of close to one in two Hispanic voters among the users in our random sample.

4.2 Other demographics represented less accurately

We further investigate whether there are disparities in how well voters of each race/ethnic group are represented as we vary their age and the economic disenfranchisement of the locale where they reside.

We begin by assigning each person the poverty rate of the ZIP code where they reside. We note that we do not know the financial situation of each individual and we use their ZIP code's level of poverty as a proxy. We then identify quintiles, i.e. five ranges of poverty rates such that there is an equal number of voters in each (one fifth of the total): Q1 below 6.7%, Q2 from 6.7%–11.0%, Q3 from 11.0%–15.6%, Q4 from 15.6%–18.8%, and Q5 18.8% and above. Finally, in each quintile we calculate the fraction of the voters of each race for whom Experian returned a match (Figure 2A) as well as the fraction of the voters of each race whom Experian matched and provided the correct birth age with high confidence (Figure 2B). Note that the numbers in Figure 2A correspond to the sum of the green and yellow areas in Figure 1 (i.e., the total fraction of users that matched regardless of the correctness), while those in Figure 2B correspond to the green area in Figure 1 (i.e., only correct age estimates with high certainty among the matched users).

Figure 2 highlights that the differences observable in Figure 1 are not explained away by the difference in affluence between the races. The downward trends in Figure 2A show that even within each race/ethnicity, individuals living in areas with higher prevalence of poverty are less likely to figure in Experian's data. Further, Figure 2B shows that data accuracy is impacted negatively especially for Hispanic individuals living in the highest poverty areas. In summary, only 45% of Hispanic individuals living in the highest poverty ZIP codes have their correct birth age in the Experian data,

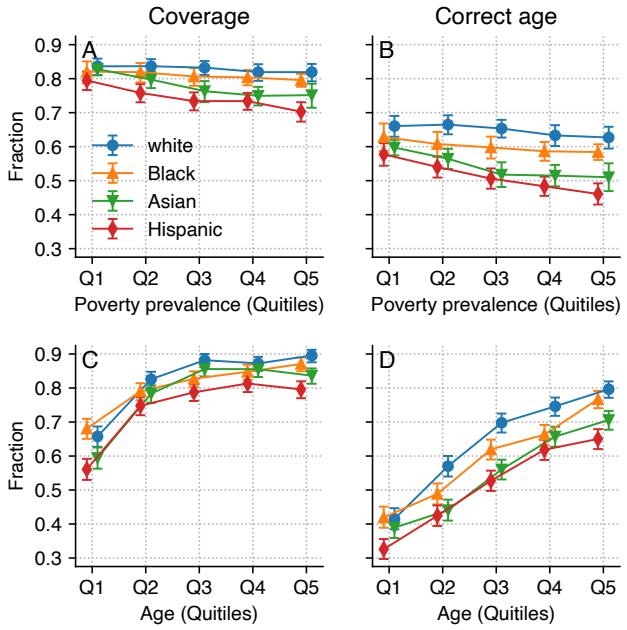

Figure 2: Coverage and accuracy varies not only with race, but also with the poverty levels in the ZIP code where individuals reside (A, B) and their age (C, D). Error bars represent the 99% confidence interval and plots are shifted along the x-axis for readability.

compared to 67% of white individuals living in the lowest poverty ZIP codes.

Next, we follow a similar approach to investigate whether the age of an individual is a factor in match rates. To this end, we identify similar quintiles of age: Q1 from 18–25, Q2 from 26–36, Q3 from 37–46, Q4 from 47–56, and Q5 as 57 and above. The upward trends in Figure 2C show that even within each race/ethnicity, younger individuals are less likely to figure in Experian's data. The coverage rates saturate around the third quantile (37 to 47 year old) and reach nearly 90% for white voters compared to 82% of all white voters and 65% of the youngest white voters. Unlike for coverage, we do not observe the saturation for data accuracy. Figure 2B shows that accuracy continues to grow with the age of an individual. The racial differences presented in Figure 1 still hold, with white voters having the highest coverage and accuracy across ages. In summary, only 32% of Hispanic individuals below 26 years old have their correct birth age in the Experian data, compared to 80% of white individuals above 57 years old.

4.3 Disambiguating demographic variables

Finally, we build a Logistic Regression model to allow for more systematic disambiguiation of the factors associated with lower coverage. Table 1 shows the coefficients of the model as well as risk change computed as described in the Methods section. The `Risk change` column indicates the relative risk change that corresponds to a change in the variable value from 0 to 1, while other variables are equal to 0. Note that `Age` and `Poverty in Zipcode` variables are

Table 1: Dependent variable: Risk of wrong or missing data

Feature	Coefficient	std. err.	Risk change
Intercept	-0.405^{***}	±0.025	40.0%
Poverty in Zipcode	0.099^{***}	±0.011	+10.4%
Age	-0.381^{***}	±0.011	-31.7%
Female	-0.049^{*}	±0.022	-4.8%
Hispanic	0.602^{***}	±0.031	+82.6%
Asian	0.438^{***}	±0.031	+54.9%
Black	0.226^{***}	±0.032	+25.3%
AUC ROC = 0.626			

*p<0.05; **p<0.01; ***p<0.001

normalized; in these two cases a zero value corresponds to the mean and a "unit increase" is an increase of one standard deviation.

This analysis further supports our results, to show that even holding age, gender, and poverty rates constant, ethnicities other than white still suffer from lower coverage. In the most extreme case, a Hispanic man is 82.6% more likely to not be represented accurately compared to a white man if they both live in an average-poverty zip code and are of average age.

In summary our results show that age and poverty levels are predictive of coverage and accuracy for individuals in all ethnic/race groups. Regardless of race, in our sample, younger individuals and those living in locations with higher prevalence of poverty are less likely to be present and correctly represented in Experian's databases.

5 DISCUSSION

Data brokers are known to be opaque to the people whose data they trade and to the researchers who try to shed light on their operations. This work, while limited in scope, offers a glimpse into the coverage and accuracy differences across race, ethnicity, age, and economic status in the data of one of the biggest data brokers: Experian. We found that across ages and locations, the white non-Hispanic voters of North Carolina have better coverage and data accuracy than Asian, Black, and Hispanic voters. Within each race/ethnic group, younger individuals and those living in areas with higher poverty rates are less likely to be represented in Experian's data. Our further regression analysis indicates that these effects likely persist across the country. While we cannot give a firm explanation for the root causes of this situation we speculate that it is related to banking and credit use. Individuals without established credit history are less likely to appear in the databases of data brokers (whose original business was credit scoring). This is in line with our results that show lower coverage of younger adults (who do not *yet* have a credit history), as well as those living in areas with higher prevalence of poverty and/or higher fraction of residents of color [34]. Taken together, our results show that those who rely on data brokers for identify verification might disproportionately reject individuals who are already more likely to be vulnerable.

5.1 Limitations

The generalizability of our study is limited by a number of factors.

First, we only studied one of multiple data brokers active in the U.S. and it is conceivable that the biases we observed are less dire for other brokers. Another data broker, Equifax, offers very similar services, but obtaining the API credentials required a conversation with their employee; once we revealed the purpose of the study we were not granted access. Nevertheless, prior work indicates that the problems of unequal representation are likely to persist. Previous research placed Experian's coverage of Facebook users about the average for U.S. brokers, lower than Acxiom and Datalogix, but ahead of Epsilon [41]. That same work reported that even the broker with the highest coverage suffered from under-representation of younger users and those in lower income ZIP codes. Further, we only focused on one *marketing* service offered by the broker, which, by law, is based on a separate database from services used for credit, housing, and employment decisions. Still, this data can still be used for the purposes of identity verification, which is not covered by the Fair Credit Reporting Act (FCRA) [7]. In fact, if we were to attempt an audit of the accuracy of credit, housing, or employment tools, it could be seen as using the data for other than statutory purposes, and thus, as a violation of the FCRA. We hope that in the future there would be legal mechanisms in place allowing researchers to investigate such systems more freely.

Second, we only purchased data about voters in North Carolina and the scale of the presented problems could be different in other states in the U.S., and in other countries. We also assumed the age reported in the voter records as ground truth but it is likely that not all of the records are accurate. Erroneous data in the voter records would lower the apparent overall accuracy of Experian but would not created the demographic biases we observe in the study. A study design with paid participants who reveal their information only for the purpose of the study could offer a possibility for a bigger scale study of coverage and accuracy, while ensuring a possibly better ground truth.

Despite these limitations, this work contributes to the ongoing discussion on the perils of relying on data broker information for access to opportunities. While we focused on only one data broker, previous research shows that Experian is not unlike other players in terms of coverage limitations and biases. Since other data brokers offer similar services, extending our methods (including the use of the Census and the voter records) to these companies is not a technical challenge. Instead, it is a matter of convincing other brokers to sell or otherwise make available the information for the stated purpose of bias measurements.

5.2 Implications

Data brokers such as Experian now play an important role in determining people's access to life opportunities. Unfortunately, their data is far from perfect, as evidenced both in this paper and in prior work. Furthermore, coverage and accuracy issues appear to disproportionately affect individuals who are already more likely to be in historically disadvantaged groups. Given these facts, we hope our research will help forward the discussion on what needs to be done to alleviate these problems.

Overall, there is a clear need for increased transparency and easier avenues for recourse for incorrect data. Today, individuals have few rights to access or correct data about them (outside of

a few areas with special legal protections). Whenever data brokers are used to confirm identities or to verify an individual's history, that individual should be informed of the data that is used. The individual should be able to contest the decision or request correction—regardless of that individual's digital literacy levels—if any data proves inaccurate. Furthermore, parties who still rely on data brokers for critical decisions should pay close attention to cases of failed identity verification and offer alternative methods without a penalty to the individual.

Through the introduction of General Data Protection Regulation (GDPR) in 2018, the European Union severely limited the legal area in which data brokers can operate: data collection and storage requires explicit consent from every affected individual, and the collected data can only be used for the purpose named in the consent with only few exceptions [6]. While the current U.S. laws permit the full range of data broker services, there is a growing concern about individual privacy reflected in some legislative initiatives. For example, statute AB 1202 signed into California Civil Code in 2019 requires data brokers to register as such as well as regulates and tracks their data trades [4]. If this trend continues one could expect more accuracy in the data because of increased transparency in data provenance. However, because of the more transparent yet limited ways of obtaining data, one might not expect a similar boost in data coverage. This situation further underscores the need to decrease reliance on data brokers for critical decisions. For example, in the end, the problem of vaccine access thwarted by the faulty data broker identify verification system [17] was not solved by collecting more data, but instead by providing vaccines to all interested individuals, regardless of their identity.

REFERENCES

[1] 2013. In FTC Study, Five Percent of Consumers Had Errors on Their Credit Reports That Could Result in Less Favorable Terms for Loans. *Federal Trade Commission* (2013). https://www.ftc.gov/news-events/press-releases/2013/02/ftc-study-five-percent-consumers-had-errors-their-credit-reports

[2] 2014. Data Brokers a Call for Transparency and Accountability. https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.

[3] 2015. The Secret Score Behind Your Rates. *Consumer Reports* (2015). https://www.consumerreports.org/cro/car-insurance/credit-scores-affect-auto-insurance-rates/index.htm#creditmap

[4] 2018. AB-1202 Privacy: data brokers. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB1202.

[5] 2018. Voting and Voter Registration as a Share of the Voter Population, by Race/Ethnicity. https://bit.ly/3unwPcp.

[6] 2018. What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr/?cn-reloaded=1.

[7] 2019. *Kidd v. Thomson Reuters Corp., No. 17-3550 (2d Cir.).*

[8] 2020. Equifax Data Breach Settlement. *Federal Trade Commission* (2020). https://www.ftc.gov/enforcement/cases-proceedings/refunds/equifax-data-breach-settlement

[9] 2021. The best data unlocks the best marketing. https://www.experian.com/marketing-services/targeting/data-driven-marketing.

[10] 2021. Experian Identity Verification. https://www.experian.com/decision-analytics/identity-management.

[11] 2021. Experian Product Sheet - Identity Verification. https://www.experian.com/content/dam/marketing/na/healthcare/brochures/identity-verification.pdf.

[12] 2021. My Credit Score. https://www.experian.com/blogs/ask-experian/credit-education/score-basics/my-credit-score/.

[13] 2021. Quarterly Residential Vacancies and Homeownership, First Quarter 2021. https://www.census.gov/housing/hvs/files/currenthvspress.pdf.

[14] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2017. Car Insurance Companies Charge Higher Rates in Some Minority Neighborhoods. *Prosperity Now* (2017). https://prosperitynow.org/files/PDFs/Credit_Fact_File_07-2016.pdf

[15] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. 2016. Tracing information flows between ad exchanges using retargeted ads. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 481–496.

[16] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. 2015. I always feel like somebody's watching me: measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. 1–13.

[17] Samantha Cole. 2021. Vaccine Site Uses Credit History to Verify Patients' Identities. *Vice* (2021). https://www.vice.com/en/article/y3gq9j/nyc-vaccine-site-credit-history-experian-identity-rejected

[18] John Gramlich. 2019. 10 facts about Americans and Facebook. *Pew Research* (2019). https://www.pewresearch.org/fact-tank/2019/05/16/facts-about-americans-and-facebook/

[19] Yael Grauer. 2018. What Are 'Data Brokers,' and Why Are They Scooping Up Information About You? *Vice* (2018). https://www.vice.com/en/article/bjpx3w/what-are-data-brokers-and-how-to-stop-my-private-data-collection

[20] Miguel Helft. 2011. After Breach, Companies Warn of E-Mail Fraud. *The New York Times* (2011). https://www.nytimes.com/2011/04/05/business/05hack.html?_r=1

[21] Lauren Kirchner. 2020. Can Algorithms Violate Fair Housing Laws? *The Markup* (2020). https://themarkup.org/locked-out/2020/09/24/fair-housing-laws-algorithms-tenant-screenings

[22] Lauren Kirchner. 2020. When Zombie Data Costs You a Home. *The Markup* (2020). https://themarkup.org/locked-out/2020/10/06/zombie-criminal-records-housing-background-checks

[23] Kalev Leetaru. 2018. The Data Brokers So Powerful Even Facebook Bought Their Data - But They Got Me Wildly Wrong. *Forbes* (2018). https://www.forbes.com/sites/kalevleetaru/2018/04/05/the-data-brokers-so-powerful-even-facebook-bought-their-data-but-they-got-me-wildly-wrong/?sh=7ccdd2193107

[24] John H McDonald. 2009. *Handbook of biological statistics*. Vol. 2. sparky house publishing Baltimore, MD.

[25] Imani Moise. 2019. African Americans underserved by U.S. banks: study. *Reuters* (2019). https://www.reuters.com/article/us-usa-banks-race/african-americans-underserved-by-u-s-banks-study-idUSKCN1V3081

[26] Gretchen Morgenson. 2014. Held Captive by Flawed Credit Reports. *The New York Times* (2014). https://www.ftc.gov/news-events/press-releases/2013/02/ftc-study-five-percent-consumers-had-errors-their-credit-reports

[27] Nico Neumann, Catherine E Tucker, and Timothy Whitfield. 2019. Frontiers: How effective is third-party consumer profiling? Evidence from field studies. *Marketing Science* 38, 6 (2019), 918–926.

[28] North Carolina Laws, Chapter 163: Elections and Election Laws [n.d.]. North Carolina Laws, Chapter 163: Elections and Election Laws. https://www.ncleg.gov/Laws/GeneralStatuteSections/Chapter163

[29] North Carolina State Board of Elections Data [n.d.]. North Carolina State Board of Elections Data. https://dl.ncsbe.gov/index.html?prefix=data/

[30] Office of Oversight and Investigations Majority Staff. 2013. A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes. https://www.ftc.gov/news-events/press-releases/2013/02/ftc-study-five-percent-consumers-had-errors-their-credit-reports.

[31] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*. 1432–1442.

[32] Javier Parra-Arnau, Jagdish Prasad Achara, and Claude Castelluccia. 2017. Myadchoices: Bringing transparency and control to online advertising. *ACM Transactions on the Web (TWEB)* 11, 1 (2017), 1–47.

[33] Caitlyn Renee Miller. 2017. I Bought a Report on Everything That's Known About Me Online. *The Atlantic* (2017). https://www.theatlantic.com/technology/archive/2017/06/online-data-brokers/529281/

[34] Lisa Rice and Deidre Swesnik. 2013. Discriminatory effects of credit scoring on communities of color. *Suffolk UL Rev.* 46 (2013), 935.

[35] Lea Shepard. 2012. Toward a stronger financial history antidiscrimination norm. *BCL Rev.* 53 (2012), 1695.

[36] Natasha Singer. 2013. Acxiom Lets Consumers See Data It Collects. *The New York Times* (2013). https://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html

[37] Amy Traub and Sean McElwee. 2016. Bad credit shouldn't block employment: How to make state bans on employment credit checks more effective. *Washington, DC: Demos* (2016).

[38] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Measuring the impact of the gdpr on data sharing in ad networks. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. 222–235.

[39] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the porn: A comprehensive privacy analysis of the web porn ecosystem. In *Proceedings of the Internet Measurement Conference*. 245–258.

[40] Giridhari Venkatadri, Alan Mislove, and Krishna P. Gummadi. 2018. Treads: Transparency-Enhancing Ads. In *Proceedings of the Workshop on Hot Topics in Networks (HotNets'18)*. Redmond, WA, USA.
[41] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P Gummadi. 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *The World Wide Web Conference*. 1920–1930.

A ROBUSTNESS

Throughout the paper we say that the data broker accurately identified an individual whenever (1) the returned match was described as "Person Match", as opposed to "Geographic Match", 'Household Match", or "Non-Match"); (2) the reported age information was described as "Exact", as opposed to "Estimated" or "None"; (3) the reported age was within one year of true age. In this section we show that relaxing these requirements results in overall higher reported coverage, but does not erase the differences between races and ethnicities.

Figure A1A repeats the results from the main body of the paper for reference. Figure A1B relaxes the requirement on reported age information quality to include "Estimated" in addition to "Exact". We observe that approximately half of "Estimated" ages for Hispanic users are correct, bringing the overall fraction from 51% to 62%. Approximately a third of white users' estimated ages are correct, resulting in a seven percentage point increase to 71%. Figure A1C further relaxes the requirement on the reported age to fall within one year of the ground truth. Here, we allow for up to four years difference, increasing the total "correct" fraction for each group by approximately two percentage points. Finally, Figure A1D further allows for lower quality matches, i.e. "Geographic Match" and "Household Match" in addition to "Person Match", further increasing the total fractions of both correct and incorrect age estimates.

As we show here relaxing the quality requirements increases the apparent coverage but one should expect that comes at the cost of increased prevalence of false positive identifications.

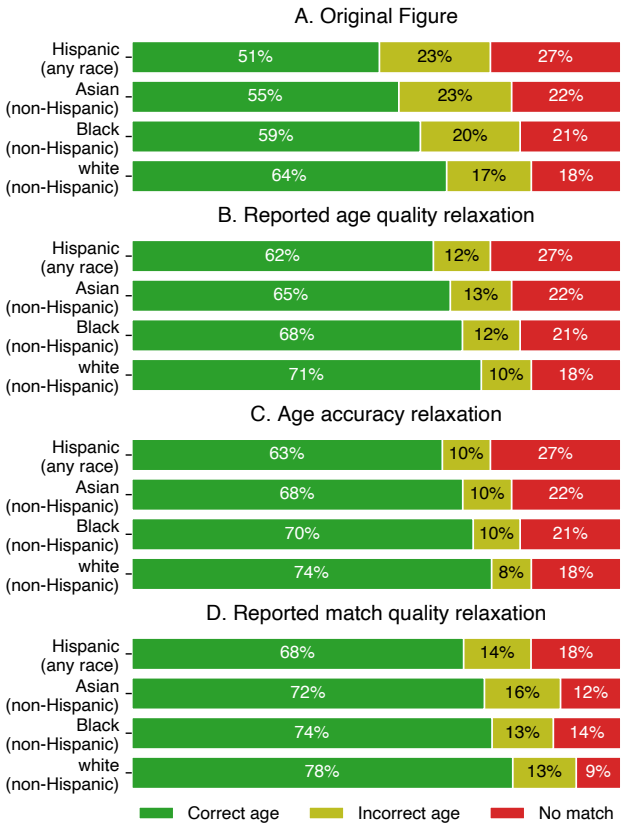

Figure A1: Differences in coverage and data accuracy among adult North Carolina population of different races. One in two Hispanic or one in three white voters are not represented or are represented inaccurately.

B ETHICAL CONSIDERATIONS

We took care to ensure our data collection and analysis was inline with community ethical standards, and minimized harm to individuals and data providers.

First, for *individuals*, we note that we did not interact with individuals in any way, but we did obtain personally identifiable information (PII) which may be considered a risk. In this work, we used PII from North Carolina voter records, which contain registered voters' name, address, self-reported race, and age (among other items). This data is considered "public information" by North Carolina law §163-82.10 *Official record of voter registration* [28] and it is available to any interested party for download from the North Carolina State Board of Elections [29]. To verify the accuracy of data broker matches, we purchased age estimates of some of the voters from Experian. By doing so we did not increase the privacy exposure of these individuals since we were purchasing estimates of information (age) which is already in public voter records. Our use of voter records for building advertising audiences was marked as Exempt by the IRB at our Univeristy (IRB# 18-11-13).

Second, for *data providers*, we took care to ensure that our measurements did not cause significant load on the infrastructure that provides the service. The North Carolina voter records are available for download as a single file, so no repeated calls to the server were needed. The U.S. Census data we use in the paper can be obtained using single API calls per endpoint or using the provided exporting functionality of the data explorer, so the load on their servers was minimal. Finally, the per-ZIP-code resident count estimates from Experian required multiple API calls (one for each ZIP code). To minimize the risk of overburdening the Experian infrastructure, we only used a single thread to query Experian, and we enforced a 5-10 second wait time between successive calls. At no point were we blocked by Experian for quota violations.

Third, a final concern is whether our work is considered Human Subjects Research (as defined by the U.S. Department of Health and Human Services). In brief, human subjects research is defined as research that either "obtains information ... through intervention or interaction" or "obtains identifiable private information". Since we do not interact with any individuals, and we only use information that is in the public domain, the presented work is not considered Human Subject Research. The self-assesment guide from the U.S. National Institutes of Health (NIH), which administers the IRB rules pointed to "most likely considered exempt".

C RESPONSE FROM EXPERIAN

We shared the initial version of this manuscript with Experian. Upon receiving the response we introduced a number of changes and clarifications which we summarize below.

First, in addition to the Data Append service described in this manuscript, the initial version also used a Mailing List service. The mailing list serivice offered an API which returned counts of profiles matching a specified a criterion. We had assumed the returned number referred to individual profiles, but Experian clarified it was effectively the count of households with at least one adult individual matching the criterion. Given that households returned instead of individuals, our analysis did not meausre the biases it purpoted to, so we removed the corresponding section of the manuscript.

Second, Experian pointed out using data collected for the purpose of credit or background reporting for other purposes would be a violation of the Fair Credit Reporting Act (FCRA). Therefore, we stress that our results do not provide direct evidence of any possible bias in the services covered under FCRA. Nevertheless, identity verification services, which we use as one of our motivating examples are not covered under FCRA and could still share the source data with the marketing services.

Third, Experian questioned the validity of using Voter Records as ground-truth, claiming they can be out-of-date. To minimise the risk of out-of-date information we had only selected individuals with an "Active" registration. Further, we note that erroneous data in the voter records might lower the apparent overall accuracy of Experian but would not create the demographic biases we observe in the study.

Finally, Experian disagreed with our comment that US customers are not asked to consent to data collection nor do they have the right to view and correct the collected data. Instead, Experian asserted that "Experian makes it easy for consumers to access the marketing data it maintains about them upon their request and provides them the option to opt-out from its using or selling their personal information for advertising and marketing solicitations". Unfortunately, we found this not to be the case. Not only is the site not indexed by search engines, it requires four clicks from the main company site, and—once found—it requires the customer to reveal many private attributes: full name, Social Security Number, date of birth, full street address, phone number, and email address (all fields obligatory). Understandably, many privacy-conscious customers will not be willing to share all of this information with Experian. Regardless, our statement still holds. While a number of states require data brokers to allow customers to opt out, no such right is granted state-wide, nor is constent being saught from individuals.