Poster: Measuring Biases in a Data Broker’s Coverage

Levi Kaplan
Northeastern University
kaplan.l@northeastern.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Piotr Sapiezynski
Northeastern University
p.sapiezynski@northeastern.edu

1 EXTENDED 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 the coverage of U.S. adults by Experian 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.

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. From identity verification, credit scoring, and personalized advertisements to housing and employment decision support, many aspects of daily life are now mediated by data brokers.

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. 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, affecting hundreds of millions of people worldwide. Finally, data brokers are notoriously opaque, making it difficult to analyze and understand the coverage and accuracy of their data.

While others have demonstrated that data brokers can have significant inaccuracies, 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, be registered to vote, or have access to mainstream financial services. 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 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. 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.

Our results using the data append services are limited only to North Carolina (where voter records with race are available), so we then use Experian’s mailing lists interface to see whether our results hold across the entire U.S. We develop a measurement methodology that involves creating a script to programmatically query the mailing list building interface to obtain the total number of matching records for a specific set of attributes. We then collect over 32,000 measurements to obtain the number of individuals in each U.S. ZIP code according to the broker’s data, and compare this data to the 2019 U.S. Census American Community Survey, which we view as “ground truth”. Our measurements show that Experian’s coverage is lower in the ZIP codes with higher populations of Asian and Hispanic residents, as well as those below the age of 26.

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, and only uses two of their many services (in particular, the two that we could get access to). However, our work sheds more light on the opaque industry of data brokers, and suggests that further scrutiny is needed as these services become increasingly important in deciding which users receive important life opportunities.