Birju Patel

Dr. Garrison

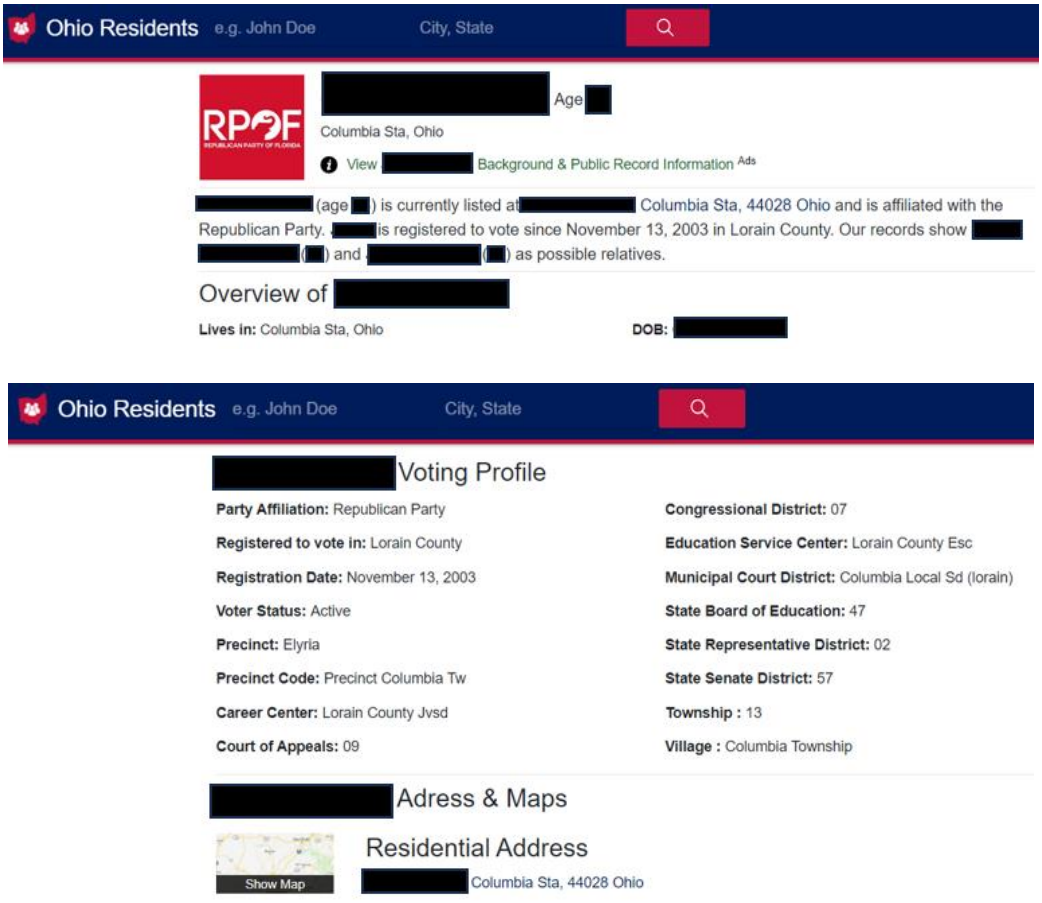
April 19, 2024

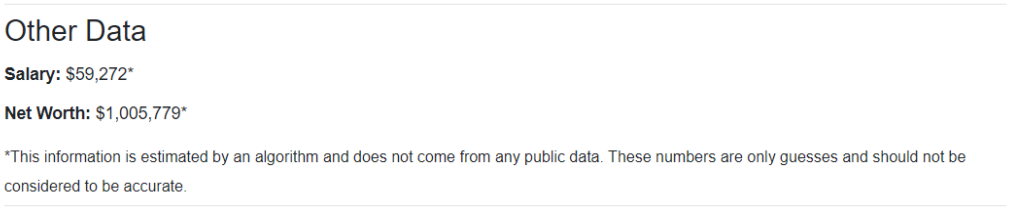
Project 3 Report

Problem Statement

All US states maintain a list of voters that are eligible to vote in the upcoming election. These rolls are made available to the public so that political campaigns can reach out and inform voters on local issues. In many states, organizations can request this data by contacting the election office, so the office may verify that they are indeed a legitimate campaign, but Ohio does not. Instead, the Secretary of State's office publishes the raw rolls on the internet, revealing the name, birthdate, address, political party, and voting history of every registered voter in every county to everyone.

An organization calling itself the Ohio Resident Database has built a website that leverages this data. It allows users to search residents by name, and presents the public information in an easy-to-use manner. The site cross references the data with other public records sources, and even does analytics to estimate values like annual income and net worth. Below are screenshots of a search result, with the personally identifiable information blacked out.





Ohio residents have expressed concerns about the availability of this data. Most understand that it is a public record, but feel that it should not be made as easily accessible as it is. One Reddit user laments that "just because it's a matter of public record doesn't mean those records should be plastered on the internet for everyone to see." They also worry that they might have been passed up for jobs in the past because potential employers could see that they are a Democrat. While the site has an opt-out feature, many users complained that it does not work, and that they had to email the site administrators multiple times to get their information removed. One particularly frustrating issue is that users must disclose this data to vote in general elections, and must declare their party preference to vote in primary elections. Users concerned about their privacy might be deterred from voting due to these overbroad disclosures.

Proposed Solution

The Ohio election office may counter that their policy increases government transparency. Because they post this data publicly, campaigns no longer have to waste time wading through bureaucratic red tape to get access to voter information. This makes it easier and faster for small campaigns to get off the ground. A well-crafted data release policy would publish information that is useful to campaigns while also protecting individual resident's privacy.

Political campaigns want to know resident's age and party preference, so that they can target them with messages that are most relevant to them. They also want to know their location, so they can send them mail and have canvassers knock on their doors prior to election day. However, especially in today's divisive political climate, residents do not want people to be able to lookup their political party preference, or else uncover it by doing analysis on the released dataset. The released dataset must give campaigns a general idea of the age, location, and party preferences of voters, while also providing a measure of anonymity to individuals. I will use k-anonymity, l-diversity, and differential privacy to attempt to transform the public datasets to satisfy these properties.

Dataset Description

For each county, the office releases a text document that contains the voter roll. The data is in a CSV format, with each row representing a single voter. Each row contains 126 items, but we are only interested in the voter's name, date of birth, party preference, and address. To reduce computing time, I selected two small counties to do the analysis. I first selected Vinton County, a small rural in central Ohio. The county went for Trump by a 54% margin in 2020, and voted for

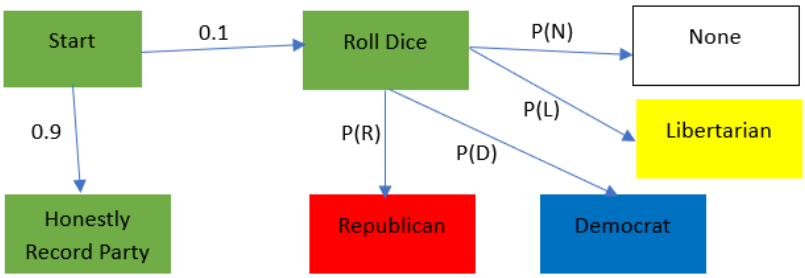
the Republican gubernatorial candidate Mike DeWine by a 57% margin in 2022. The second county I selected was Athens County. This is also a rural county and is along the Ohio-Kentucky border. However, since it contains Ohio University, it typically votes Democratic, going for Biden by a 15% margin in 2020 and for the Democratic gubernatorial candidate Nan Whalen by a 6.5% margin in 2022.

Results

Initially, I made the data k-anonymous with a k value of 5. Instead of releasing the voter's age and address, I generalized and released only their age range (0-24, 25-44, 45-64, 65+) and the street their home was on. If there was a particular group such that they all had the same quasi-identifiers, but the size of the group was less than 5, I deleted these records. Only a small number of records had to be deleted. While this data was k-anonymous, it was still vulnerable to attack. In both datasets, there were several streets where everyone that lived on that street was registered as either a Democrat or Republican. If an attacker learned someone's address from another source, they could cross reference it with this dataset to identify their party preference with certainty.

To solve this issue, I made the data l-diverse. Since there are two major political parties, I have to assure that for each possible quasi-identifier, there is at least one record that is identified as a Democrat, and another as a Republican. To get this level of anonymity, I had to zoom out even further. Instead of reporting a voter's street, I report their municipality. This method guarantees l-diversity, but it makes the dataset less useful to campaigns. Knowing which municipalities to target may be useful for more broad marketing like purchasing television ads or putting up billboards, but canvassers need more fine-grained information when they are going door to door. Knowing which streets are rich in Democrats or Republicans could save canvassers considerable time.

I attempted to find a compromise between the two methods by using differential privacy. I assumed that what campaigns want to know is the number of Democrats and Republicans on each street. I used the following sampling procedure to add randomness to the data. For each row in the dataset, I flip a weighted coin. If the coin lands on heads, I honestly record what that row's party preference was. If the coin lands on tails, I randomly assign the voter as either a Democrat, Republican, Libertarian, or None with some probability. This probability distribution is calculated by the total percentage of the dataset that is of each party.



This random sampling method gives each voter a measure of plausible deniability. If an attacker sees a street with only Republicans, they cannot know if the street truly contains only Republicans, or if it contains mostly Republicans and some Democrats that lied about their party affiliation because of the random nature of the process.

Conclusion

I believe the differential privacy method provides the best tradeoff between individual privacy and usefulness. It allows the end user, which in this case is a political campaign, to know which streets are rich in which type of voter. This will help them deploy campaign resources in the most targeted manner. At the same time, an attacker will not know the true counts of how many voters on each street belong to a particular party. So even if there is a particular street which is reported as consisting of only members of one party, the attacker cannot know if that is truly the case, or if it is an artifact of the random sampling. As a bonus, releasing only the counts, not individual records, significantly compresses the size of the released dataset.

Reflection

During this project, I found a public dataset that leaked people's private information. I researched to see how that dataset was being used by the public, and found an instance of it being used in ways that people objected to. I read their complaints to get a better understanding of why they felt their privacy had been violated. I also considered the reason why the dataset was released, and what information the end users of the dataset wanted from it. Finally, I proposed and implemented a program that used the anonymization techniques we learned in class to solve this problem. I ended up with a solution that I believe balanced privacy and usability.

This project ties in nicely with my course goal of understanding the impact of privacy on society. I was able to see how real people were affected by data privacy. I was also glad to be able to investigate a topic that I am personally interested in, which is politics.

Sources

Ohio Voter Rolls

Ohio Resident Database

Reddit Thread on the Ohio Resident Database

Notes

To recompute the released data, navigate to the root directory and run *python anonymize.py*.

The input data should be stored in the *data* directory. The output data will also be saved there.

You can download your own data from the Ohio Secretary of State's website and copy it into the *data* directory. To run the program on this data, modify the code in *anonymize.py* to read from that file instead of the default files.