IT UNIVERSITY OF CPH

Final Project - Part 1

Analysis and anonymization of election data

Julia Bijak
jbij@itu.dk

Krzysztof Michal Parocki
krpa@itu.dk

Renata Katarzyna Sapeta
resa@itu.dk

Christian Sneftrup Fleischer
chfl@itu.dk

13th november 2024

Presented for the Course:
BSSEPRI1KU Security and Privacy,
5th Semester of BSc, Data Science
IT University of Copenhagen
Denmark

1 Introduction

This report gives insights on our approach to balancing privacy and data utility in a hypothetical election scenario where online voting was piloted in one municipality. Given concerns over potential discrepancies between online and in-person votes, we conducted statistical analyses on survey data to assess variations in political preferences, demographics, and voting channels.

To ensure data privacy, we applied anonymization techniques, carefully evaluating disclosure risks and comparing analytical results from raw and anonymized datasets to measure utility loss and privacy gains. Our report discusses each stage of the process, including anonymization methods, risk assessment, and insights gained into maintaining data integrity while protecting sensitive information. Some information about our methods are omitted in this report to not reveal additional information about the anonymized dataset.

2 Anonymisation

To begin the anonymization task, we first defined quasi-identifiers (key variables) to be date_of_birth (dob), zip, sex, marital_status and citizenship. All of those variables are present in the public_data_registry dataset, and so can be used for re-identification. Education is not present in the public dataset, and so we assumed it cannot be used as a key variable. We defined the sensitive variable of utmost importance to the potential attackers as the party preference, and to a lesser extent e-vote.

As the first step, we dropped the direct identifier from the dataset, i.e., the name column. Another column we decided to remove was zip_code, as its correlation with party preferences can be computed from the public_data_results instead, making use of the entire population instead of the sample in the survey. This limited the dimensionality of our anonymised dataset, making the privacy protection easier.

We applied a range of anonymization techniques on the resulting dataset, such as generalization (casting foreign citizens to a Non-Danish category to limit identification of non-frequent nationalities), ranging (age groups) and a mix of suppression and swapping on some of the columns (we will omit the information about which ones they were). We restricted swapping to only within each party group to keep correlations and aggregate statistics intact. We decided not to use PRAM as that would result in modifying the data entries significantly, and so severely impacting the utility of the dataset.

To calculate the disclosure risks we used the **measure_risk** function from the *sdcMicro* library. The risks are as of below in Figure 1.

```
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 5 (2.500%)
  - 3-anonymity: 15 (7.500%)
  - 5-anonymity: 39 (19.500%)
```

Figure 1: Disclosure risk measure

3 Data Analysis

After the anonymization we inspected the three different data sets again to get an overview over the attributes and the results. As part of our inspection we did some analysis to look into some significance statistics regarding different aspects of the data sets.

3.1 Political preferences and election results

We started with checking if there is a significant difference between the political preferences as expressed in the survey and the official election results. The resulting values were 4.58 chi-square and and 0.60 p-value, indicating that there wasn't a significant difference between the two, and we could use the sample as a valid representation of the general population in the municipality.We observed the same outcome for both the original dataset and the anonymized dataset. The Chi-Square statistic and p-value are the same across both as the general party preferences were not influenced by the anonymization process.

3.2 Political preferences and demographic attributes

We looked into finding a significant **difference between political preferences of the voters depending on their demographic attributes** recorded in the survey.
We used chi-squared for assessing whether a difference between two categorical variables is due to chance or a relationship between them. The higher the chi-squared, the greater the difference from what would be expected under the null hypothesis.
To confirm the significance of these results, we calculated p-value, where the null hypothesis is that there is no association between the two variables. We can disregard it if the p-value is low (usually below 0.05).

Party Preference vs.	Chi2		p-value	
	Original	Anonymized	Original	Anonymized
Gender	4.70	4.04	0.09523	0.13287
Marital Status	23.90	20.53	0.00055	0.00222
Citizenship	64.67	9.91	0.00000	0.00704
Education	48.00	48.00	0.00001	0.00001
Age Category	20.97	20.97	0.00185	0.00185
Zip Code	12.32	-	0.05511	-

Table 1: Significance test on Preferred Party and demographics

From the above table, we can see that there is a significant difference between political preference of voters grouped by one's martial status, citizenship, age and education in the original dataset. We achieve similar results for the anonymized data, except for citizenship. This difference emerges due to combining all the minorities in Denmark into one category. As we decided to not include zip code in the anonymized data, we cannot compare the results of the two. We visualized the distributions of chosen demographic attributes in Figure 2. We can see from it (and the table above) that generally, the methods used for anonymization (even those perturbative) were not destructive - the correlations to party preferences were retained for all key variables.

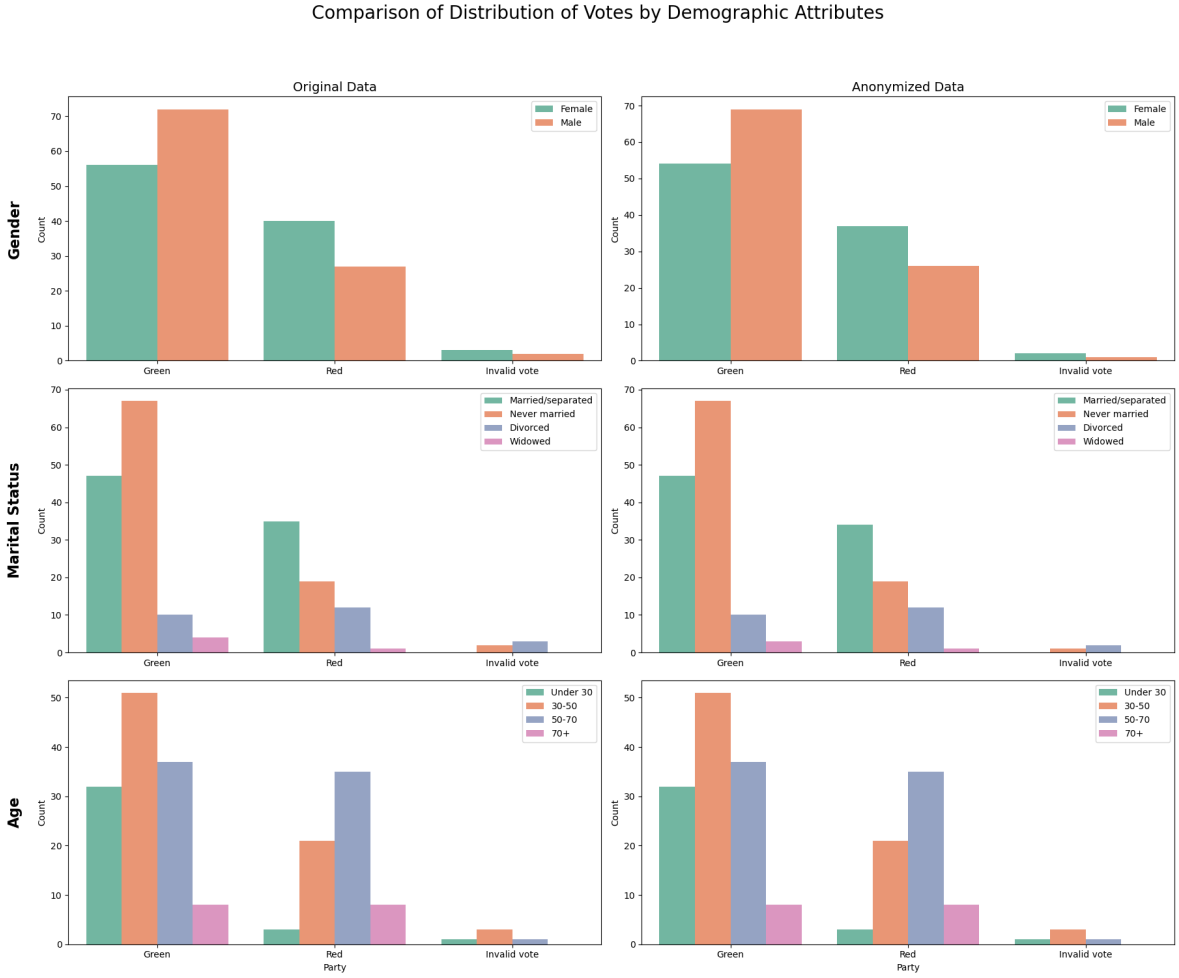

Figure 2: Plots of demographics in the original vs anonymized data depending on preffered party

3.3 Voting channel and demographic attributes

We measured if there is a significant difference between voter's choice of voting channel depending - again - on their demographic attributes recorded in the survey. We took similar measures as for the previous step, namely, chi-squared and p-value that should tell us about the correlations.

Voting Channel vs.	Chi2		p-value	
	Original	Anonymized	Original	Anonymized
Gender	0.36	0.30	0.55116	0.58284
Marital Status	6.96	6.15	0.07316	0.10460
Citizenship	4.88	0.0	0.76987	1.0
Education	5.87	5.87	0.55488	0.55487
Age Category	6.07	6.07	0.10785	0.10785
Zip Code	3.02	-	0.38887	-

Table 2: Significance test on Voting Channel and demographics

From the above results, we can conclude that there isn't a significant difference of casted e-votes between voters grouped by the demographics (p-values well above 0.05). Most of the results achieve similar scores for both original and anonymized datasets, except for citizenship. Yet again, this stems

from generalizing the minorities into one category. We did not include zip code in the anonymized dataset. Finally, we can see from it the Figure 3 below that generally, the methods used for anonymization (even those perturbative) were not destructive and the correlations to party preferences were retained for all key variables.

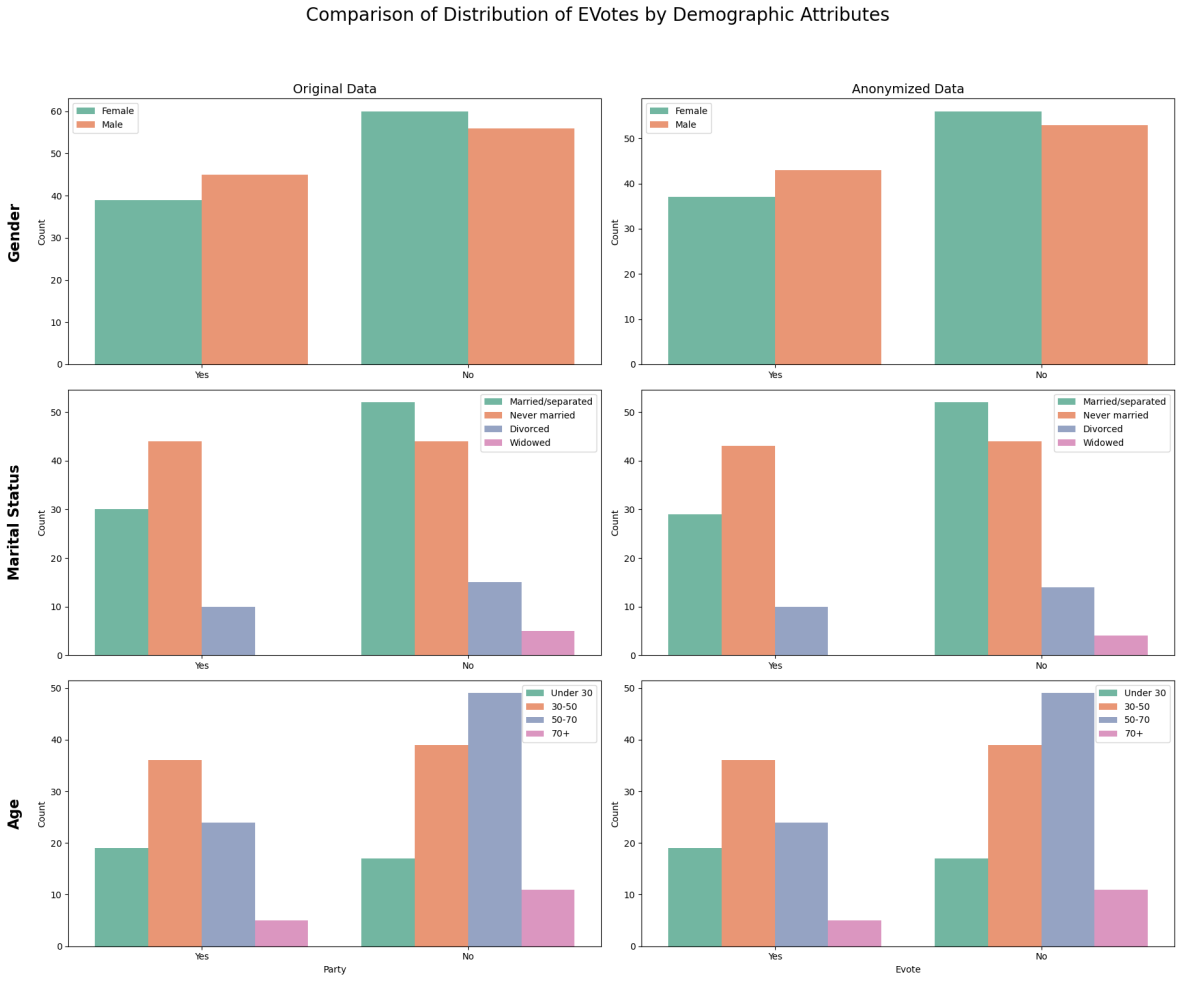

Figure 3: Plots of demographics in the original vs anonymized data depending on casted evote

4 Discussion & Conclusion

Overall, we have achieved solid privacy while preserving most of the utility of the original dataset. We have used non-perturbative methods such as ranging and generalization, losing the details of individual entries but preserving the truthfulness and the general usability. We have also utilized perturbative methods, such as swapping and suppression. However, we implemented swapping only within e-voting and political preference groups, preserving the demographic qualities linked to both of these variables (the same counts by age, marital_status, gender, etc., as indicated by Figure 3).

The only choice we made impacting the utility of the dataset was suppression, however we didn't do a lot of it; eventually, only a handful of records were removed. We might have focused too much on retaining the utility of the anonymized dataset, but we were already satisfied with our k-anonymity results. If we decided to suppress more data, we could have gotten even lower k-anonymity scores. On the other hand, even a handful of suppression records results in data loss limiting the utility of the resulting dataset; therefore, we decided to keep it to the minimum.