IT UNIVERSITY OF COPENHAGEN

Final Project

Security and Privacy

BSSEPRI1KU

Anna Maria Gnat (agna@itu.dk)
Bogdan Cristian Mihaila (bomi@itu.dk)
Josefine Nyeng (jnye@itu.dk)
Miranda Speyer-Larsen (spey@itu.dk)

BSc in Data Science
IT University of Copenhagen
November, 2024

Contents

1 Introduction

Governments have recently experimented with new voting technologies to make elections more accessible and efficient. This report examines a scenario where a country introduced an option of online voting in some municipalities along with traditional voting using paper ballots. The election results showed a different distribution for electronic votes, raising questions about potential manipulation as well as demographic differences between voters choosing online voting and the traditional method. To understand these differences and assess the influence of demographics on voting behavior, a survey was conducted, collecting data on voter preferences and behaviors. Given the sensitive nature of the data, proper anonymization was crucial before public release.

Dataset

The datasets used in this project are derived from three primary sources:

- Private Survey Data: This dataset contains raw survey responses from voters, including their demographic attributes, voting method (online or paper), and political preferences.
- Public Population Register: This dataset provides publicly available demographic information of the population in the municipality. Attributes include name, gender, date of birth, ZIP code, marital status, education level, and citizenship.
- Public Election Results: This dataset includes the aggregated election results, specifying how many votes each party received, split by voting method (online or paper).

These datasets provide a comprehensive view of both voter preferences and demographic attributes, allowing for a detailed analysis of voting patterns and the impact of demographic characteristics on the choice of voting channel.

Importance of anonymization

The primary challenge in working with survey data, particularly one involving political preferences, lies in safeguarding voter privacy. Anonymization is crucial in this context for several reasons:

Firstly, the survey contains sensitive data, including political preferences, personal demographics, and voting methods. If such data were linked to specific individuals, it could lead to significant privacy violations, putting individuals at risk of discrimination, political targeting, or social stigmatization.

Secondly, many countries have stringent data protection regulations, such as the GDPR, which mandate that personal data must be anonymized before being shared or used for secondary purposes. An effective anonymization process helps meet these regulatory requirements and ensures responsible handling of citizens' information.

Lastly, anonymization aims to minimize disclosure risk—the risk that individuals can be re-identified from anonymized datasets by linking data attributes. This project specifically evaluates disclosure risks by simulating an adversary attempting to deanonymize the dataset using additional publicly available information. Balancing utility and privacy is key: while data anonymization should ensure privacy, it should also retain as much utility as possible for meaningful analysis.

In this project, both perturbative and non-perturbative anonymization methods are applied to achieve an optimal balance between data privacy and utility. Disclosure risk metrics are used to evaluate the robustness of the anonymized dataset and determine whether the risk of re-identification is acceptable in comparison to the losses in data utility.

2 Anonymizing the dataset

To anonymise our dataset we employed 3 techniques: Removing direct identifiers, Global Recoding, and Post-Randomization Method.

2.1 Removing direct identifiers

Direct identifiers are attributes that allow for the identification of an individual with no additional data. That can be, for example, full name, CPR number, passport number, etc. In the data from the survey, there is one unique identifier - the full name of the voter. Since this information will not be used by us in any analysis, we decided to remove it from the dataset. This ensures that individuals can not be directly identified based on their name.

2.2 Global recoding

Global recoding is an anonymization method that takes the values of a certain attribute and combines them. For categorical variables, that means creating more general categories, while for numerical values it means creating intervals. In our dataset, we have used the global recoding on the date of birth attribute. For this, we have first converted the specific date of birth into age. Then we computed the intervals based on quantiles, ensuring that each interval had a similar amount of observations. This resulted in the following intervals: 18 – 30; 30 – 38; 38 – 48; 48 – 58; 58 – 70; 70 – 101.

These age groups still provide us with valuable information when looking at political preferences in different age demographics, while increasing the anonymity of the dataset.

We also applied global recoding to the Marital Status attribute so that instead of 4 categories we only had 2: Married and Not married. We felt this still provided enough information for analysis purposes while still reducing specificity to lower the risk of re-identification.

2.3 Post-Randomisation Method

Post-Randomisation Method (PRAM) is a perturbative anonymization method that replaces values of categorical variables according to some pre-defined transformation matrix. This transition matrix defines the probabilities of each value of the categorical variable staying the same or being changed to one of the other possible values. This means that while adversaries may attempt to re-identify records within the dataset, they can never be sure that the listed value for the variable is accurate or has been altered.

For this reason, we decided to apply pram to the Sex variable in our dataset using the pram function from the R package, sdcMicro. We set the pd parameter to 0.7, meaning each record has a 70% of retaining the original sex and a 30% chance of being swapped to the opposite sex.

3 Disclosure risks

After anonymizing our dataset we computed two different disclosure risks: the k-anonymity and the global average risk.

3.1 k-anonymity

To access the re-identifiability of the records, we computed the number of k-anonymity violations. k-anonymity ensures that, for each record, there are at least k-1 other records that share the same values for the quasi-identifiers. Thus we want as few k-anonymity violations as possible.

Firstly, we began by defining the quasi-identifiers that the adversary may use when attempting to de-anonymize the data. In the publicly available dataset, the adversary would have access to the following variables: name, sex, date of birth, zip, citizenship, and marital status. Since we re-coded the date of birth to be age groups, the adversary will have use this as a quasi-identifier instead of date of birth. Moreover, since we applied PRAM to the sex variable, it will be virtually useless as a quasi-identifier since the adversary cannot be sure the value is correct. Thus we assumed the quasi-identifiers to be: age group, citizenship, marital status, and zip.

We computed the number of violations of k-anonymity for k = 2,3 and 5, in order to assess the robustness of your anonymization process across a range of levels of anonymity. We decided to omit the actual values in this report to avoid revealing additional information about the dataset.

3.2 Global risk

The global risk is an aggregate risk measure of the re-identification risk for the complete dataset. We decided to use the average risk as we wanted to focus on minimizing overall risk rather than considering the outlier cases of each individual. We computed the average risk by taking the mean of the individual risks for each record. The individual risk is the likelihood that an individual can be re-identified based on the quasi-identifiers used above. Our anonymized dataset had an average risk of 0.295. This seems respectably low, however we do acknowledge that the average risk can be misleading, as some individuals may still have very large individual risks and end up being in more danger of re-identification than others.

4 Analysis

To determine whether the anonymization process preserves the overall distribution and characteristics of the data, we conducted some analysis on our anonymized dataset.

4.1 Comparison between datasets

Political preferences across survey and results data

The first analysis we conducted was to investigate if there was a significant difference between the political preferences as expressed in the survey and the election results for both electronic and polling station votes. To do this we conducted a Chi-square test of independence, which tests whether two variables are related based on their contingency table. The null hypothesis is that the variables are un-related i.e. that there is no significant difference between the survey's political preferences and the election results. Thus, a p-value below the generally accepted significance level of 0.05 indicates that we can reject the null hypothesis, implying that there is a significant difference. We conducted this test to both for the Polling Station votes and the E-votes.

The p-value was found to be 0.8223 for Polling Station votes and 0.4241 for E-votes, which are both not below the significance level, 0.05. Thus the null hypothesis can not be rejected, which implies that there is no significant difference between political preferences in the survey data and results data for either voting method.

4.2 Analysis of the demographic attributes

To investigate how demographic attributes influence the political preference and the voting channel used, we also employed a Chi-square test to compare the values across the following demographic attributes: Age, Sex, Marital Status and Education. The null hypothesis is that there is no relationship between the political preferences/voting channel and a given demographic attribute. We use the same significance level of 0.05 as above.

Demographic attributes vs. Political preference

Table 1 below displays the results for each Chi-square test comparing the political preferences across each demographic attribute. As shown, the p-value for Sex is above the significance level and thus implies that there is not a significant difference between political preference across different genders. However, the p-values for Age Group, Marital Status, and Education are below the significant level, indicating that these attributes do impact the political preference of the voters.

Category	Chi-square Statistic	p-value
Age Group	39.4953	2.078e-05
Sex	1.7660	0.4135
Marital Status	20.0162	0.0027
Education	34.7005	0.0043

Table 1: Chi-square Test Results

Demographic attributes vs. Voting channel

Table 2 shows the preferences in the voting channel compared to the demographic attributes. As displayed, the p-value for every demographic attribute is above the significance level which implies that we do not have sufficient evidence to conclude that demographics influence the voting channel preferences.

Category	Chi-square Statistic	p-value
Age Group	6.7213	0.2422
Sex	1.0203	0.3124
Marital Status	2.3438	0.5041
Education	8.7840	0.3608

Table 2: Chi-square Test Results

5 Reflections

This project gave us an opportunity to work with data anonymization and assess the effects of different techniques on both privacy protection and data utility. We find data anonymization to be extremely important in two ways.

First, the protection of the privacy of individuals is essential for ethical research practices. Good research relies on good data collection, and so there is a great need for the trust between the researchers and the general public to be preserved.

Secondly, the balance between utility and anonymization can be challenging to preserve. Understanding proper techniques and being critical of the results ensures that the correlations and conclusions drawn from the data remain valid even after anonymization. In our approach, we aimed to maintain this balance,

while prioritizing utility. All the analyses from Section 4 were conducted on both original and anonymized datasets. This allowed us to assess the impact of anonymization on the information quality. Although some information loss was unavoidable, we found that our anonymization preserved the key correlations, allowing the conclusions from the statistical tests to remain the same and thus accurately reflect the original data. Looking at our disclosure risks, it becomes clear that we likely could have applied more anonymization, while still maintaining a respectable level of utility.

We have also found that sometimes anonymization methods such as generalization, can be a useful tool in data analysis. In the case of this project, we noticed that the original dataset noted no correlation between age and political preferences. However, generalizing that attribute to age groups, we see that there is a correlation, as outlined in section 4.2. Even though that result is not present in the original dataset, it provides valuable information in the context of this research and so we decided to keep that change.

Overall, while this project was a fun test of our abilities, it did come with some challenges. Because the next step of this workshop involved another group de-anonymizing our dataset, our anonymization had to include some compromises. This trade-off was challenging, as we needed to protect privacy yet maintain data quality for meaningful analysis and not make it too hard for the next group.