Security Phase 1

Benjamin Storm Larsen
bsla@itu.dk

Yuliia Storm Larsen
yuls@itu.dk

November 13, 2024

Abstract

In this work we aim to investigate the datasets according to the guidelines of this assignment that dictates the comparison analysis of the anonymised dataset statistics versus the original data. Using different metric to see deeper into the data gave us detailed information and outlined the process plan.

1 Anonymize methods

Performing multiple steps that have been used to anonymise the data for external reporting. Several methods has been applied to minimize the risk of identification of voters with quasi-indentifiers. The anonymization process began by addressing highly unique identifiers, which are the most critical in mitigating re-identification risks. Due to the small sample size, the name column reflects a high degree of re-identification risks. Consequently it is necessary to discard the column. The name column does not contribute to the data analysis, and is not used within it. Following the names column is the gender of the voter, which, given the small sample size, reflects a binary selection of male or female. We have decided not to apply any methods of anonymisation to this column. Age is what could contribute to an increase in re-identification risks, especially with smaller data samples. We

decided to apply a generalization function to the age, essentially grouping ages into 4 groups of ages. This will result in a decrease in the data resolution, but with the counter effect of increasing anonymity. This can be done using the cut function from the pandas library, and by selecting appropriate bins sizes.

The embedded risks of zip codes that can highly correlate with the other identifiers it is combined with. The problem also carries a greater risk if the population of the postal is small. We have determined that, the zip code carries valuable information in terms of the analysis of the voters. Therefore we made use of *pseudo-nymization* to mask the zip-codes with pseudo random numbers. While the approach hides the actual zip-code it is still possible to find the underlying zip codes. However, this will increase the workload of a potential infiltrator.

Educational information poses a great risk, and cannot be left unnoticed. We deemed that a highly naive over generalization was necessary to reduce the risk of re-identification. Given the educational levels and specialization, we grouped university-related degrees together and mapped the reaming values to "other". A consequent of the generalization is the direct impact on the analysis.

The citizenship of the voter can essentially be split into two values which is: "danish" and "other". Given that the majority of the votes are casted by people with danish citizenship there is a lower risk of re-identification. But for the other nationalities where the might only be a few it increases the risk level tremendously. Arguably, we could also completely discard the column as the small sample size of counties will not represent a true distribution of votes.

The voters marital status can be a powerful quasi-identifier. With names included it can enhance the possibility of finding two married people, as they likely share similar data, such as age, zip code and citizenship. However, as the name column is already discarded this does not pose a large risk.

2 Disclosure risk metrics

Calculating the risk of re-identification of the anonysmied is a crucial part of evaluating the methods applied in the previous part. The information in the section is not gonna be verbose, the primary focus will be laid on the evaluation calculation used. Our approach used the k-anonymity calculations. This method uses grouping to group voters sharing the same attributes, which the function ranks by count. The function takes the parameter k, which signifies the acceptable count of groups. As an example of the parameter 5, if there is group size below that count the dataset

then does not meet the privacy requirement. The downside is that it does not protect against homogeneity attacks, as it groups people together regardless of the similar quasi-identifiers.
This is why we also included the l-diversity test. This test aims to battle the short comings of the k-anonymity test. This test ensures that for each grouping there is at least "l" different values for the sensitive attributes. With theses to test we could iteratively apply new methods and change parameters to reach an optimal balance between value and privacy.

3 Results of analysis on the anonymised dataset

The analysis conducted a comparison between the voters political preferences and their demographic parameters. The analysis is based on the ANOVA method. The findings reveal that male and female voters does not share the same distribution of voting for the two parties. It is difficult to conclude if this result is expected or not. The test on the citizenship is as explained prior not sufficient for the private survey data. However, since the anonymization grouped this column to danish citizen and other foreigners the was possible although with some underlying uncertainty. It showed that there is no particular difference between the native group and forefingers.
The educational level of the voter would be typically be expected to influence the political preferences, as observed in real-world scenarios. However. this is not the case for our dataset. This could be due to the dataset being randomly generated, with no underlying correlation or settings to reflect a real-world dataset. As a result, educational level does not appear to depict the voters political preferences in this analysis. It is important to note that the education column has been applied an high degree of generalization during the anoymisation process, which could have contributed to the lack of correlation between the two groups.

The marital status also does not have a significant affect on the choice of party among the voters. However, given a real-world context we would not assume this to be a strong correlation with the political preferences.
The results portray an interesting image of the voters and their demographic correlation to where their vote lies.

The analysis of voter preferences, based on demographic groups, reveals interesting patterns when comparing both the private survey and the anonymized public survey results. These findings confirm that our anonymization process and methods was successful, effectively balancing the privacy concerns while keeping valuable

resolution of the data. The expected outcome of the analysis should optimally reflect each other. However, evidently some of the anonymization methods over generalized part of the dataset. In the case of the education level, the generalization reached an extreme that can be seen in the results, as both groups now have a similar mean across political preferences. The marital status of the voters has shown to matter less in terms of political preferences.

4 Comparison between raw and anonymised dataset analysis

In nature of this reports focus on anomysation we will not disclose direct results from the analysis performed on the private data. However, comparing the two is a necessity to discussing the value of the results. The trends across the two datasets analysis remained to a certain degree the same. The one noticeable change is the educational level. As discussed previously the data might have been generalized to a too high of a degree. And comparing the results to the private show that in there in reality is a difference in the political preference of the voters based on their educational level. This aligns with the expectation and real-world statics of the education of a voter and their political preference.
It can be argued whether pseudo-anymization masking has the desired effect. The underlying distribution will remain the same, it is only visually that the numbers does not correlate to the prior data. However, the real anomysation is questionable. As an example the zip codes, if the zips were simply sorted and given an index from top to bottom the they sequence could easily be found. However, the masking has not change to the underlying distribution and even mask the data visually.

5 Conclusion

In conclusion, we can say that the anonimsation did not decrease the value of the analysis results to an extent of making the analysis useless. The iterations with the testing methods of k-anonymity and l-diversity assisted in balancing between the value and privacy of the analysis.