Security and Privacy Final Project

Group Random 0

Janusz Wilczek, Sofia Terenziani, Thomas Charton-Jaeg and Gareth Thomas

How to make sure nobody knows you didn't vote

IT University of Copenhagen

15th November 2023

1 Introduction

Within the domain of data science, the balance between privacy preservation and analytical utility is often confronted when dealing with sensitive datasets. One such challenge lies in the anonymization of datasets, where the goal is to obscure identifiable information, introducing challenges for any entity to establish links between specific records and individuals, while retaining the necessary information for meaningful analyses.

This assignment involves the anonymization of a dataset originating from a recent election, specifically focusing on voter attributes and affiliation. The provided dataset features the attributes of 200 voters involved in the election and includes direct identifiers, quasi-identifiers, and sensitive-non identifying data, as detailed in Figure 1.

While obfuscating private and identifying data was paramount, preserving the dataset's utility for statistical insights remained a key directive in the anonymisation process and in informing our decisions regarding the risk/utility trade-off present in all privacy adjacent problems.

	name	sex	evote	dob	zip	education	citizenship	marital_status	party
72	Chavez, Marlyn	Female	1	1977-05-03	2300	Upper secondary education	Denmark	Never married	Green

Figure 1: Voter attributes (Fictional example)

2 Statistical Analysis

In this assignment, we performed analyses on the original dataset prior to the anonymization process to assess the potential impact of the anonymization on the utility of the data. By examining the non-anonymized dataset, we aimed to identify specific analyses that might be feasible with the raw data but could be restricted or altered post-anonymization.

Due to the categorical nature of our data, we chose the Chi-Squared test when performing statistical significance analyses. Chi-squared tests contrast the expected attribute frequencies resulting from a null hypothesis with the observed values to calculate a p-value indicating the probability of the null hypothesis being true.

2.1

"*Is there a significant difference between the political preferences as expressed in the survey and the election results for both electronic and polling station votes?*"

The Chi-Square test is applied to investigate if there is a significant difference between political preferences expressed in the survey and the actual election results for both electronic and polling station votes.

The results led to the failure to reject the null hypothesis. Consequently, we conclude that there is no statistically significant difference between the political preferences as expressed in the survey and the actual election results for both electronic and polling station votes. In simpler terms, the survey results align closely with the election outcomes.

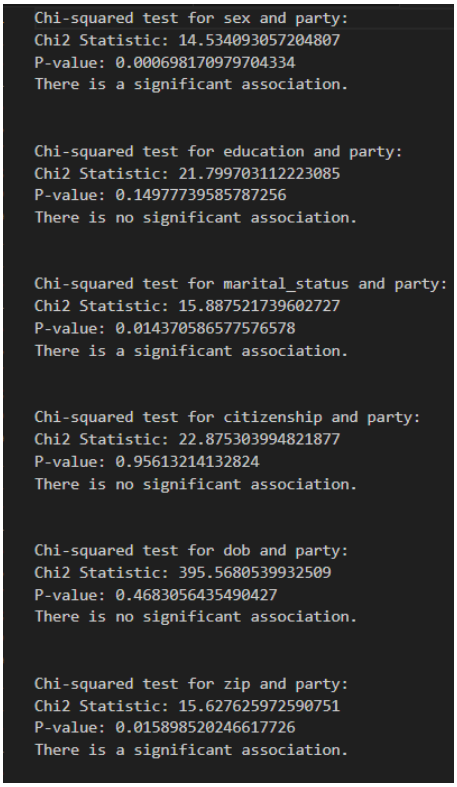
```
Chi-squared statistic: 0.4480239378867181
P-value: 0.7993055674690487
Fail to reject the null hypothesis: There is no significant difference between the election and survey results.
```

2.2

"*Is there a significant difference between political preferences of the voters depending on their demographic attributes recorded in the survey (that is, age, gender, education level. . .)?*"

In exploring political preferences based on attributes (day of birth, gender, education level, citizenship, marital status, and zip), the Chi-Square test serves to assess whether there is a significant association between these variables. This analysis aims to identify potential patterns or correlations between political leanings and individual attributes.

After conducting a Chi-Squared test between the political preferences of voters and their attributes, we found a significant correlation between gender and party affiliation, zip code and party, and marital status and party. However, no significant association was observed for education level and party, or between day of birth and party. These findings provide valuable insights into the interplay between certain demographic factors and political preferences among the voters of the provided dataset.

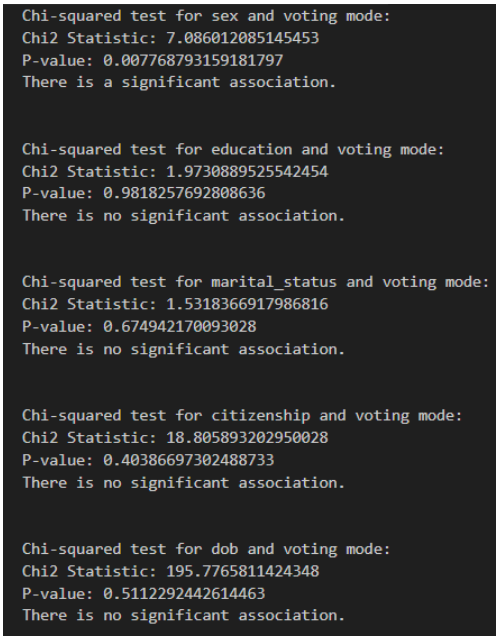


2.3

"*Is there a significant difference between voter's choice of the voting channel (that is, if they decide to vote either online or in person) depending on their demographic attributes recorded in the survey?* "

The Chi-Square test is applied to investigate if there is a significant difference in the voter's choice of voting channel (online or in person) based on demographic attributes recorded in the given dataset.

After conducting a Chi-Squared test between the voters' choice of voting channel and their attributes, we only found a significant correlation between gender and voting mode.



3 Anonymisation

After performing our analyses on the unmodified source data, we ran these on our anonymised data to give an estimate of the utility loss engendered by the anonymisation process.

In order to quantify the risk reduction provided by our anonymisation methods, we used k-diversity to investigate the groupings of individuals with similar quasi-identifiers. We then extended this investigation using l-Diversity to ascertain the diversity of sensitive values relative to each quasi-identifier.

3.1 Anonymisation by Generalisation

Our methodology primarily revolved around generalizing our data by minimising the number of unique records and concatenating smaller attribute groupings that could lead to unwanted disclosures. The aim of generalisation involves generalizing or modifying quasi-identifiers to reduce the risk of re-identification while still maintaining the utility of the data.

In the survey dataset, when ignoring names, all rows were unique. As A result of our generalisation, the number of distinct rows decreased to 151 number of unique samples based on quasi-identifiers and 83 samples with the same values for quasi-identifiers. The fact that there are fewer unique combinations after generalization compared to the original number of records indicates that the generalization process has reduced the granularity of the quasi-identifiers. However, having 83 groups with the same values for quasi-identifiers means that there is still a risk of re-identification within groups of individuals, as each group represents a set of records that are indistinguishable based on quasi-identifiers.

When including vote information, the sensitive attribute, for each participant, there are 162 unique samples based on quasi-identifiers, with 67 samples sharing the same values for quasi-identifiers.

Original	Anonymised
Full Name	Name removed
Date of Birth	Age grouped in bins of 20 years
Citizenship	Grouped into Danish and Non-Danish
Education	Reduced categories to 3 groups
Marital status	Grouped widowed and Divorced
Evote	Remained unchanged

Table 1: Anonymisation method summary

High school education	University education	Vocational education
Upper secondary education	PhD programes	Vocational bachelors educations
Primary education	Bachelors programes	Vocational Education and Training (VET)
	Masters programes	Short cycle higher education

Table 2: Education groupings

3.1.1 Name and Date of Birth

These attributes provided direct and quasi-direct identifiers respectively for our voters, a significant cause for concern. The name was simply removed, however, to preserve the utility provided by dates of birth we chose instead to replace them with age, then compromised by generalizing further and grouping in bins of 20 years.

3.1.2 Citizenship

Our particular survey sample contained 175 Danish citizens, leaving 25 voters tied to an almost universally unique country of origin. To address this we grouped all non danish voters into their own category.

3.1.3 Education

There were large differences in sample size for the different educational attributes, therefore we reduced the number of groups from 9 to 4 by mapping similar courses of study into new groups, identified in Table 4 below, while maintaining the "not stated" category.

3.1.4 Marital status

In this case, the divorced and widowed categories were significantly smaller than married and never married which lead to our decision to merge them.

3.2 Assessing Anonymisation by Generalisation

3.2.1 k-Anonymity

The reduction in number of unique samples after generalization signifies a reduction in the granularity of quasi-identifiers. However, the presence of 83 groups sharing identical quasi-identifiers indicates that some re-identifiability persists within those groups. This suggests that additional measures may be necessary for measuring individual privacy.

To address this, our methodology incorporates k-anonymity. K-anonymity ensures that each combination of quasi-identifiers is indistinguishable from at least k-1 other combinations, minimizing the risk of re-identification.

Testing out K-anonymity on the generalised dataset; it satisfies only 1-anonymity. This raises a concern regarding the effectiveness of the anonymisation measures applied. Achieving the lowest k-anonymity means that each unique combination of quasi-identifiers remains distinct in the dataset, indicating a lack of anonymity and vulnerability to re-identification for some individuals.

3.2.2 l-Diversity

K-anonymity does have limitations that make it not always sufficient in addressing all privacy concerns, as it does not consider the diversity of sensitive attributes within a group. This could potentially lead to issues in case of homogeneity of sensitive information, in this case, each individual's choice of vote. L-diversity extends the concept of anonymity by ensuring that within each anonymised group, there is sufficient diversity in terms of the values of sensitive attributes. L-diversity aims to provide a higher level of protection by making it more difficult for an adversary to infer sensitive information, even if the adversary can identify the group to which an individual belongs.

Testing out L-diversity on the generalised dataset resulted in a satisfactory outcome only for L = 1. The result indicates that the level of diversity in terms of sensitive variables within each group of individuals is very limited. When only L = 1 is satisfactory, it means that within each group there is only one unique value for the sensitive attribute. This lack of diversity is a risk for individual re-identification in the dataset.

4 Anonymisation

After performing our analyses on the unmodified source data, we ran these on our anonymised data to give an estimate of the utility loss engendered by the anonymisation process.

In order to quantify the risk reduction provided by our anonymisation methods, we used k-diversity to investigate the groupings of individuals with similar quasi-identifiers. We then extended this investigation using l-Diversity to ascertain the diversity of sensitive values relative to each quasi-identifier.

4.1 Anonymisation by Generalisation

Our methodology primarily revolved around generalizing our data by minimising the number of unique records and concatenating smaller attribute groupings that could lead to unwanted disclosures. The aim of generalisation involves generalizing or modifying quasi-identifiers to reduce the risk of re-identification while still maintaining the utility of the data.

In the survey dataset, when ignoring names, all rows were unique. As A result of our generalisation, the number of distinct rows decreased to 151 number of unique samples based on quasi-identifiers and 83 samples with the same values for quasi-identifiers. The fact that there are fewer unique combinations after generalization compared to the original number of records indicates that the generalization process has reduced the granularity of the quasi-identifiers. However, having 83 groups with the same values for quasi-identifiers means that there is still a risk of re-identification within groups of individuals, as each group represents a set of records that are indistinguishable based on quasi-identifiers.

When including vote information, the sensitive attribute, for each participant, there are 162 unique samples based on quasi-identifiers, with 67 samples sharing the same values for quasi-identifiers.

Original	**Anonymised**
Full Name	Name removed
Date of Birth	Age grouped in bins of 20 years
Citizenship	Grouped into Danish and Non-Danish
Education	Reduced categories to 3 groups
Marital status	Grouped widowed and Divorced
Evote	Remained unchanged

Table 3: Anonymisation method summary

4.1.1 Name and Date of Birth

These attributes provided direct and quasi-direct identifiers respectively for our voters, a significant cause for concern. The name was simply removed, however, to preserve the utility provided by dates of birth we chose instead to replace them with age, then compromised by generalizing further and grouping in bins of 20 years.

4.1.2 Citizenship

Our particular survey sample contained 175 Danish citizens, leaving 25 voters tied to an almost universally unique country of origin. To address this we grouped all non danish voters into their own category.

4.1.3 Education

There were large differences in sample size for the different educational attributes, therefore we reduced the number of groups from 9 to 4 by mapping similar courses of study into new groups, identified in Table 4 below, while maintaining the "not stated" category.

4.1.4 Marital status

In this case, the divorced and widowed categories were significantly smaller than married and never married which lead to our decision to merge them.

High school education	University education	Vocational education
Upper secondary education	PhD programes	Vocational bachelors educations
Primary education	Bachelors programes	Vocational Education and Training (VET)
	Masters programes	Short cycle higher education

Table 4: Education groupings

4.2 Assessing Anonymisation by Generalisation

4.2.1 k-Anonymity

The reduction in number of unique samples after generalization signifies a reduction in the granularity of quasi-identifiers. However, the presence of 83 groups sharing identical quasi-identifiers indicates that some re-identifiability persists within those groups. This suggests that additional measures may be necessary for measuring individual privacy.

To address this, our methodology incorporates k-anonymity. K-anonymity ensures that each combination of quasi-identifiers is indistinguishable from at least k-1 other combinations, minimizing the risk of re-identification.

Testing out K-anonymity on the generalised dataset; it satisfies only 1-anonymity. This raises a concern regarding the effectiveness of the anonymisation measures applied. Achieving the lowest k-anonymity means that each unique combination of quasi-identifiers remains distinct in the dataset, indicating a lack of anonymity and vulnerability to re-identification for some individuals.

4.2.2 l-Diversity

K-anonymity does have limitations that make it not always sufficient in addressing all privacy concerns, as it does not consider the diversity of sensitive attributes within a group. This could potentially lead to issues in case of homogeneity of sensitive information, in this case, each individual's choice of vote. L-diversity extends the concept of anonymity by ensuring that within each anonymised group, there is sufficient diversity in terms of the values of sensitive attributes. L-diversity aims to provide a higher level of protection by making it more difficult for an adversary to infer sensitive information, even if the adversary can identify the group to which an individual belongs.

Testing out L-diversity on the generalised dataset resulted in a satisfactory outcome only for L = 1. The result indicates that the level of diversity in terms of sensitive variables within each group of individuals is very limited. When only L = 1 is satisfactory, it means that within each group there is only one unique value for the sensitive attribute. This lack of diversity is a risk for individual re-identification in the dataset.

5 Effects of anonymisaiton

Statistical diversity analyses have been recalculated after anonymization to assess the impact of anonymization on the utility of the data. In comparing the initial analysis of the original dataset, before the anonymization process, with the subsequent analysis performed after anonymization, notable differences emerge. The anonymization process has introduced alterations to the data, influencing the utility and the outcomes of statistical analyses.

Test of the anonymised dataset also fails to reject that there is no significant difference between the election and survey anonymised results, suggesting no significant difference between political preferences in the anonymised dataset and the election results.

Utilizing Chi-squared tests to assess associations between attributes and political preferences or voting modes, the analysis of the anonymised dataset unveils some different findings than the original significance test. Based on the Chi-squared test results, it appears that there are significant associations between political preferences and certain attributes; gender, age, education level, and marital status all indicate significant associations with party affiliation. This differs from the analysis on the original dataset, which did not see an association between age, which descents from date of birth, and party, nor for education and voted party. The Chi-squared test results on the anonymised dataset also indicate associations between certain attributes and the choice of voting mode (online or in-person) that were not present in the initial analysis. Nonetheless, it must be noted, that the emergence of a relationship between age and party preference can have a positive value from an analytical stand-point, as date of birth is too granular to effectively infer any relationships.

```
Chi-squared test for sex and party:
Chi2 Statistic: 14.534093057204807
P-value: 0.000698170979704334
There is a significant association.


Chi-squared test for age and party:
Chi2 Statistic: 37.2069203607463
P-value: 1.0542858858352217e-05
There is a significant association.


Chi-squared test for education and party:
Chi2 Statistic: 13.685981165813207
P-value: 0.0333475422099726
There is a significant association.


Chi-squared test for citizenship and party:
Chi2 Statistic: 1.876321724748493
P-value: 0.39134691474849054
There is no significant association.


Chi-squared test for marital_status and party:
Chi2 Statistic: 15.887521739602725
P-value: 0.014370586577576568
There is a significant association.
```

The emergence of a relationship in the anonymized data suggests that certain trends in political preferences become more apparent when individuals are generalized into larger, not-specific categories. Generalization has here both enhanced privacy by reducing by a little the risk of re-identification and unveiled new associations between attributes.

Excessive generalization can however lead to a loss of information, thereby affecting the statistical significance of findings. Over-generalization can lead to meaningful relationships being overlooked or misleading associations being introduced. Over-generalisation may in this case also result in relationships that are in fact artifacts of the anonymization process rather than reflections of true data structure.

These disparities underline the impact of even a naive anonymization of the original dataset, and the necessity of considering the implications of measures for preserving privacy on the outcomes of statistical analyses.

```
Chi-squared test for sex and online or in-person vote:
Chi2 Statistic: 7.086012085145453
P-value: 0.007768793159181797
There is a significant association.


Chi-squared test for age and online or in-person vote:
Chi2 Statistic: 19.301925889018047
P-value: 0.0006855338781413511
There is a significant association.


Chi-squared test for education and online or in-person vote:
Chi2 Statistic: 0.7115728059337608
P-value: 0.8704785915429434
There is no significant association.


Chi-squared test for citizenship and online or in-person vote:
Chi2 Statistic: 0.0
P-value: 1.0
There is no significant association.


Chi-squared test for marital_status and online or in-person vote:
Chi2 Statistic: 1.5318366917986816
P-value: 0.674942170093028
There is no significant association.


Chi-squared test for party and online or in-person vote:
Chi2 Statistic: 9.735376362718677
P-value: 0.007691125237786415
There is a significant association.
```

6 Conclusion and Considerations

The anonymization process in this assignment did not yield optimal results. The emphasis on generalization, while contributing to new associations in the dataset, has presented challenges in maintaining the individual privacy of the participants in the dataset. Satisfying only 1-anonymity and 1-diversity indicates limitations in the level of protection and diversity in the anonymized data, and suggests that we should improve our anonymization strategy. However, we were working under the limitation of maintaining same level of utility, which reduced our range of tools and parameters.

Considerations for a more effective anonymization process include exploring other approaches beyond generalization. Approaches such as suppression and perturbation methods may offer a better balance between privacy preservation and data utility. By adopting more sophisticated and diverse anonymization techniques, like perturbation-based methods, we could have attained a better level of privacy protection.

Additionally, a reevaluation of the chosen quasi-identifiers and their impact on record uniqueness could lead to a better anonymization strategy. For example, citizenship does not show a significant association with either party affiliation or voting mode. Therefore, while some attributes do have a statistically significant association with how individuals have voted, others, such as citizenship, may not play a significant role in the utility of the dataset for further analysis.