IT UNIVERSITY OF COPENHAGEN

Final Project - Report - Phase 1

Security and Privacy

BSSEPRI1KU

Alexis Serruys (serr@itu.dk)
Pedro Prazeres (peca@itu.dk)

BSc in Data Science
IT University of Copenhagen
November, 2024

Contents

1 Anonymization Methods

To keep the voters' information confidential, we employed several anonymization techniques on the dataset.
We selected methods like PRAM (Post Randomization Method), removal of specific variables, data skewing, binning, group shifting, masking, noise addition, and record removal based on k-anonymity. Our objective was to protect privacy without rendering the data unusable for analysis, while retaining as many entries as possible.
We created two versions of the anonymized dataset. The first version turned out to be too difficult to deanonymize, so we decided to retry by dialing down PRAM and other techniques a bit. This way, we keep the data secure but still allow for some level of deanonymization for the next group to work on during Phase 2 of the project.

1.1 PRAM

We used the Post Randomization Method (PRAM) to add some uncertainty to specific fields: `evote`, `zip`, `citizenship`, and `party`.
PRAM works by randomly changing some data values based on a set probability, which can help prevent a potential adversary from identifying individuals in the dataset.

Variable	PRAM Probability
evote	15%
zip	20%
citizenship	5%
party	5%

Table 1: PRAM probabilities applied to different variables

1.1.1 Probabilities

We intentionally kept these probabilities very low for this exercise. This approach allows us to preserve most of the original data for analysis while still enhancing privacy and giving the next group a chance to deanonymize the data. Using higher probabilities would have made the data safer but less usable, so we found what we consider a good middle ground.

1.2 Binning

Binning was an important part of our strategy to reduce data specificity. By grouping data into broader categories, we made it harder to identify individuals but still kept the data useful for analysis. We did this for attributes like `age`, `education`, and `citizenship`.

1.2.1 Age

For age, we grouped people into broader ranges. For instance, we combined ages 18 to 29 into one group labeled "18-29" instead of a single-year category. For older age groups, we used a wider category like "70+" to include everyone 70 and older. This was necessary because there were only two individuals aged 90 or above, which would have made them easily identifiable. These groupings make it harder to pinpoint someone's exact age.
We also added noise by adding a 10% chance that someone's age group might shift up or down one bin. This small adjustment helps protect individual identities without changing the overall age distribution too much.

1.2.2 Education

We grouped education levels into broader categories like "Basic Education", "Vocational and Short-Cycle Education", and "Higher Education". This way, we protect privacy

by not revealing specific education details but still keep useful distinctions for analysis. Three categories were chosen as two were too destructive to keep the data usable.

1.2.3 Citizenship

For citizenship, we used a straightforward *true/false* classification. This helps reduce the risk of identifying someone based on their `Danish_citizenship` status while still keeping the data useful for our analysis.

1.3 Masking

We masked the `party` variable, and randomly morphed the 2 entries we had marked as "Invalid Vote". This helps prevent re-identifying individuals through rare or unique party affiliations. Masking makes the quasi-identifiers less effective without taking away too much of the data's usefulness. Masking was also applied to the `zip` variable, assigning a random `zip_region` to each zip code.

1.4 Skewing the Data to Maintain Gender Parity

To ensure we kept the balanced ratio of men to women in the dataset, we adjusted the gender distributions when needed. While this is a form of PRAM, our main goal was to keep gender representation fair and consistent for accurate analysis.

1.5 Variable Removal

We decided to remove some highly sensitive quasi-identifiers like `dob` (date of birth) or direct identifiers like `name` because they posed too much of a privacy risk and weren't essential for our statistical analysis. We thought about hashing these variables, but removing them entirely was a better way to protect privacy without losing important analytical value.

1.6 k-Anonymity

To further ensure privacy, we applied k-anonymity with a threshold of $k = 2$. This means each combination of key attributes appears at least twice in the dataset, making it harder to re-identify individuals. We chose $k = 2$ as a balance between privacy and data usability. A higher k would have been safer but might have made the data less useful for the purpose of the exercise.

1.6.1 Graph of k-Anonymity

We created a graph (Figure 1) showing how the number of removed entries changes as k increases. This visual helps explain why we settled on $k = 2$, highlighting the trade-off between protecting privacy and keeping enough data for meaningful analysis.

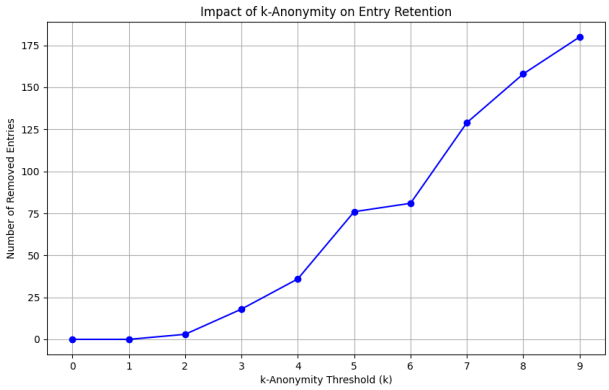

Figure 1: Impact of k-Anonymity on Entry Retention

We expected the graph to follow more of an exponential curve, where the number of removed entries would rise rapidly as k increased. However, the current graph shows a mostly linear increase up to $k = 6$, with a sharp jump only at $k = 7$ and $k = 8$.
We hypothesized that this pattern was likely due to the small size of the dataset and that in a larger one, we would expect the curve to be steeper, reflecting a more exponential trend as increasing k would continuously require more and more entries to be suppressed.

2 Disclosure Risk Metrics

To ensure the maintenance of voter privacy in the dataset, three key metrics were examined: **k-anonymity**, **l-diversity**, and **t-closeness**. These metrics assist in assessing the likelihood that an individual could be identified or that sensitive information about them could be learned from the anonymized data.

2.1 k-Anonymity

2.1.1 Definition

k-Anonymity ensures that each unique combination of certain identifying attributes (called quasi-identifiers) appears in at least k different records in the dataset. Essentially, any individual is indistinguishable from at least $k - 1$ others based on those attributes.

2.1.2 Selected Quasi-Identifiers

The following attributes were chosen as quasi-identifiers due to their potential to identify individuals if not properly anonymized:

- Sex
- Electronic vote status (`evote`)
- Age group
- Zip region
- Education category
- Marital status

2.1.3 Calculation Method

The dataset was grouped based on these quasi-identifiers, and the number of records in each group was counted. The group sizes varied from 1 to 3.

2.1.4 Results

- Groups with only one record ($k = 1$) pose a risk because an individual can be singled out.
- Groups with 2 or 3 records are better, but increasing the group size would enhance privacy.

Quasi-Identifier Combination	Group Size
Female, 0, 18–29, Region 1, Higher Education, Married/Separated	3
Male, 0, 70+, Region 1, Basic Education, Never Married	2
Male, 0, 50–59, Region 3, Vocational and Short-Cycle Education, Married/Separated	1
Male, 0, 50–59, Region 4, Basic Education, Divorced	1
Male, 0, 50–59, Region 4, Vocational and Short-Cycle Education, Divorced	1
⋮	⋮

Table 2: k-Anonymity Results

2.1.5 Discussion

Some groups have only one individual, which could pose a re-identification risk. While this may not be critical for the current exercise, it is important to acknowledge these potential privacy issues.

2.2 l-Diversity

2.2.1 Definition

l-Diversity extends k-anonymity by ensuring that within each group of quasi-identifiers, there are at least l different values for the sensitive attribute (in this case, political party). This helps protect against someone deducing sensitive information even if they cannot identify a specific individual.

2.2.2 Calculation Method

The number of unique political parties (the `party` attribute) within each quasi-identifier group was counted. The diversity ranged from 1 to 3.

2.2.3 Results

- The average l-diversity was approximately 1.06, which is low and indicates that many groups lack diversity in political party affiliation.
- Groups with $l = 1$ present a high risk because if an individual's quasi-identifiers are known, their political party can be inferred.

2.2.4 Discussion

To protect against attribute inference attacks, increasing the variety of political party affiliations within each group is necessary. While acceptable for the purposes of this exercise, it is important to acknowledge these issues and keep them in mind when working with privacy-sensitive data.

Quasi-Identifier Combination	Party Diversity
Female, 0, 18–29, Region 1, Higher Education, Married/Separated	3
Male, 0, 70+, Region 1, Basic Education, Never Married	2
Male, 0, 50–59, Region 3, Vocational and Short-Cycle Education, Married/Separated	1
Male, 0, 50–59, Region 4, Basic Education, Divorced	1
Male, 0, 50–59, Region 4, Vocational and Short-Cycle Education, Divorced	1
⋮	⋮

Table 3: l-Diversity Results

2.3 t-Closeness

2.3.1 Definition

t-Closeness measures how closely the distribution of the sensitive attribute within a group matches its distribution in the entire dataset. The goal is to prevent an attacker from gaining significant information about an individual based on the group's distribution.

2.3.2 Calculation Method

The distribution of political parties within each group was compared to the overall distribution in the dataset using the Wasserstein distance metric. The t-closeness values ranged from 0.20 to 0.35.

2.3.3 Results

- Lower t-closeness values (closer to 0) indicate that the group's distribution closely matches the overall distribution, which is favorable for privacy.
- Higher t-closeness values suggest that the group differs from the overall dataset, increasing the risk of sensitive information being inferred.

Quasi-Identifier Combination	t-Closeness
Female, 0, 18–29, Region 1, Higher Education, Married/Separated	0.35
Male, 0, 70+, Region 1, Basic Education, Never Married	0.34
Male, 0, 50–59, Region 3, Vocational and Short-Cycle Education, Married/Separated	0.25
Male, 0, 50–59, Region 4, Basic Education, Divorced	0.20
Male, 0, 50–59, Region 4, Vocational and Short-Cycle Education, Divorced	0.30
⋮	⋮

Table 4: t-Closeness Results

2.3.4 Discussion

Lower t-closeness values should be aimed for by making group distributions more similar to the overall dataset to enhance privacy. While the current values are acceptable for this exercise, recognizing where improvements can be made is important.

2.4 Conclusion

The current anonymization methods are adequate for the purposes of this exercise, especially since another group will be receiving the anonymized data. However, it is important to be aware of these potential privacy issues. In real-world applications, the following should be considered:

- Generalizing and/or suppressing quasi-identifiers even more, to increase group sizes (k-anonymity);
- Increasing the diversity of sensitive attributes within each group (l-diversity);
- Adjusting the data to make group distributions more similar to the overall dataset (t-closeness).

By acknowledging these flaws, we get a better understanding of how to protect sensitive data from possible adversaries.

3 Data Analysis

3.1 Chi-square Test Results

We performed Chi-square tests to examine the relationships between various categorical variables before and after data anonymization. The variables analyzed include:

- Voting Channel (Polling Station Votes and E-votes)
- Demographic Factors (Gender, Age Group, Education Level)
- Interaction Effects (Voting Channel with Gender, Age Group, and Education Level)

3.1.1 Before Anonymization

Table 5: Chi-square Test Results Before Anonymization

Variable	**Chi-square**	**p-value**
Polling Station Votes	37.87	7.58×10^{-10}
E-votes	20.81	5.08×10^{-6}
Gender	11.14	0.0038
Age Group	39.30	9.39×10^{-5}
Education Level	50.77	1.73×10^{-5}
Voting Channel by Gender	4.06	0.0439
Voting Channel by Age Group	17.53	0.0075
Voting Channel by Education Level	10.17	0.2531

3.1.2 After Anonymization

Table 6: Chi-square Test Results After Anonymization

Variable	**Chi-square**	**p-value**
Polling Station Votes	47.30	6.11×10^{-12}
E-votes	16.55	4.74×10^{-5}
Gender	9.39	0.0520
Age Group	34.92	0.0005
Education Level	23.14	0.0008
Voting Channel by Gender	23.76	0.0006
Voting Channel by Age Group	24.57	0.1371
Voting Channel by Education Level	8.49	0.4859

3.2 Analysis

Comparing the Chi-square results before and after anonymization reveals significant changes:

3.2.1 Voting Channel

- **Polling Station Votes**: The Chi-square value increased from 37.87 to 47.30, and the p-value decreased, indicating a stronger association post-anonymization. This suggests that anonymization may have marginally highlighted the relationship between polling station votes and party preferences.
- **E-votes**: The Chi-square value decreased from 20.81 to 16.55, with a slight increase in the p-value. While still significant, the association appears slightly weaker after anonymization.

3.2.2 Demographic Factors

- **Gender**: The association was significant before anonymization ($p = 0.0038$) but became marginally non-significant after ($p = 0.0520$), suggesting a weaker relationship post-anonymization.
- **Age Group**: The relationship remains highly significant, though the Chi-square value decreased slightly from 39.30 to 34.92.
- **Education Level**: The Chi-square value dropped from 50.77 to 23.14, but the association remains strongly significant, indicating a reduced yet persistent relationship.

3.2.3 Interaction Effects

- **Voting Channel by Gender**: There was a significant increase in the Chi-square value from 4.06 to 23.76 and a decrease in the p-value from 0.0439 to 0.0006, indicating a stronger interaction after anonymization.
- **Voting Channel by Age Group**: Although the Chi-square value increased from 17.53 to 24.57, the p-value rose from 0.0075 to 0.1371, making the interaction non-significant post-anonymization.
- **Voting Channel by Education Level**: The Chi-square value decreased from 10.17 to 8.49, and the p-value increased from 0.2531 to 0.4859, showing no significant association both before and after anonymization.

3.2.4 Conclusion

Our analysis shows that data anonymization can both enhance and weaken certain associations within the data. Any researchers applying anonymization to data should be aware of these effects when designing studies and interpreting results, and future research should focus on methods that preserve key statistical relationships while ensuring data privacy.

4 Discussion

4.1 DRM discussion

- **k-Anonymity:** Some groups have only one individual, posing a re-identification risk, and therefore should be taken into consideration.
- **l-Diversity:** Low diversity in political party affiliation within groups increases the risk of sensitive information being inferred, so we should consider increasing diversity where possible.

- t**-Closeness:** Moderate differences between group and overall distributions suggest some risk. Those differences could be reduced to improve privacy.

4.2 Analysis discussion

Data anonymization impacts the relationships between voting channels, demographics, and voting preferences in various ways:

- **Strengthened Associations**: The link between polling station votes and party preference became stronger, and the interaction between voting channel and gender significantly increased post-anonymization. This suggests that anonymization might reveal deeper patterns in these areas or, on the other hand, might be altering the results in ways that do not represent reality.
- **Weakened Associations**: The relationship between gender and voting preferences became marginally non-significant, and the interaction between voting channel and age group lost its significance, indicating that anonymization can obscure some of the more nuanced relationships.
- **Stable Associations**: Age group and education level maintained strong, significant associations with voting preferences, although their strength slightly decreased.

These findings highlight the need to balance data privacy with maintaining analytical accuracy. While anonymization protects individual privacy, it can alter the data's statistical properties, potentially affecting research outcomes.

4.2.1 Conclusion on Privacy-Utility Balance

We aimed to protect privacy while keeping our data useful. By using methods like PRAM, binning, masking, and k-anonymity, we successfully reduced the risk of revealing sensitive information. Privacy measures, including k-anonymity, l-diversity, and t-closeness showed strong protection of sensitive data. However, some combinations of demographic attributes still showed low diversity, which means there is a slight risk of inferring certain attributes.
We also maintained most of the key statistical relationships in the data after anonymization, although a few changed a bit. For example, the link between polling station votes and party preferences became stronger, while the connection with gender became weaker. These changes have only a modest impact on how we interpret the data.

We made sure there was at least one easily re-identifiable character

5 References

[Aggarwal and Yu, 2005] Aggarwal, C. C. and Yu, P. S. (2005). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-Preserving Data Mining: Models and Algorithms*, pages 11–52. Springer US.

[Dwork, 2006] Dwork, C. (2006). Differential privacy. *Proceedings of the 33rd International Conference on Automata, Languages and Programming (ICALP)*, 4052:1–12.

[Li et al., 2007] Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 106–115. IEEE.

[Machanavajjhala et al., 2007] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 24–35. IEEE.

[Sweeney, 2002] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, volume 10, pages 557–570. World Scientific.