Anonymization Report for Election Survey Data

Security and Privacy

BSSEPRI1KU

IT UNIVERSITY OF CPH

BSc in Data Science

Costel Gutu, cogu@itu.dk Oleksandr Adamov, olea@itu.dk
Vladislav Konjusenko, vlko@itu.dk

November 12, 2024

1 Transformation and Anonymization Techniques

To ensure the confidentiality of sensitive voter information, we applied several anonymization techniques, following best practices for privacy protection. These transformations focused on reducing the detail of certain variables while preserving the dataset's analytical utility.

1.1 Variable Transformation

Education Mapping (Reducing Detail by Recoding): To simplify educational levels and reduce the granularity of educational attainment, we recoded education into two broader categories: "High school or lower" and "Higher education." This reclassification reduces the risk of identifying individuals by their specific educational backgrounds.

- **Mappings Applied**:
  - "Vocational Education and Training (VET)," "Primary education," and "Upper secondary education" were grouped as "High school or lower."
  - "Vocational bachelor's educations," "Masters programmes," "Short cycle higher education," "Bachelor's programmes," and "PhD programmes" were grouped as "Higher education."

Age Grouping (Top-Coding and Binning): We grouped ages into broader intervals, specifically "20-40," "40-60," and "60+." This approach reduces the risk of re-identification by concealing specific ages, which can often be highly identifying, while retaining the age information needed for demographic analysis.

Citizenship Simplification (Reducing Detail by Recoding): We simplified citizenship data by creating a binary category, "Denmark" versus "Other." This aggregation decreases the disclosure risk for individuals belonging to minority nationalities, especially within smaller demographic groups.

1.2 Direct Identifier Removal (Removing Variables)

To prevent direct linkage to identifiable individuals, we removed specific identifiers, such as names and ZIP codes. This action is essential for confidentiality, aligning with standard anonymization practices to eliminate direct connections to individuals.

1.3 Marital Status Recoding (Reducing Detail by Recoding)

To further reduce specificity, we combined less common categories in the marital status variable:

- **Before Generalization**: "Married/separated," "Divorced," "Widowed," and "Never married."
- **After Generalization**: "Married/separated" and "Not married" (encompassing "Divorced," "Widowed," and "Never married").

This generalization minimizes the specificity of marital status, increasing anonymity for individuals in less common categories.

2 Disclosure Risk Metrics: K-Anonymity and L-Diversity Evaluation

2.1 K-Anonymity

To assess the dataset's resistance to re-identification, we calculated K-anonymity. Initially, the raw dataset showed a K-anonymity value of 0, indicating high re-identification risk for certain demographic profiles.

- **Anonymized Dataset Results**: After anonymization, the minimum K-anonymity value improved to 2. This value, while low, shows increased privacy protection compared to the raw data.

2.2 L-Diversity

We evaluated the L-diversity for sensitive attributes, particularly political preference, ensuring demographic groups contained diverse values.

- **Results**: In the raw dataset, 19 out of 48 groups met the 2-diversity criterion (39.58%). After anonymization, 15 out of 24 groups (62.50%) met the 2-diversity threshold. This improvement indicates that our anonymization techniques effectively enhanced the dataset's resistance to inference attacks on sensitive attributes.

3 Results of the Analyses: Chi-Square Test

To assess the impact of anonymization on data utility, we conducted Chi-square tests on both the raw and anonymized datasets. This evaluation focused on the relationship between demographic attributes and both political preferences and voting methods (e-vote vs. paper).

- **Voting Preference Analysis**:
  - **E-Vote Chi-Squared Test**: Chi-Squared Value: 3.44, p-value: 0.064
    Interpretation: There is no significant difference between survey e-votes and actual e-votes.
  - **Paper Vote Chi-Squared Test**: Chi-Squared Value: 0.17, p-value: 0.678
    Interpretation: There is no significant difference between survey paper votes and actual paper votes.
- **Note**: The Chi-square test results for both E-Vote and Paper Vote comparisons remained consistent before and after anonymization, indicating that the anonymization process did not affect the integrity of these comparisons.
- **Chi-Squared Test on Quasi-Identifiers (Raw Data)**:
  - Sex and Political Preference: $p = 0.0171$
    Significant difference in political preference based on sex.
  - Voting channel and Political Preference: $p = 0.0181$
    Significant difference in political preference based on voting channel.
  - Education and Political Preference: $p = 0.0747$
    No significant difference in political preference based on education.

  - Citizenship and Political Preference: p = 0.9910
    No significant difference in political preference based on citizenship.
  - Marital Status and Political Preference: p = 0.0987
    No significant difference in political preference based on marital status.

- **Chi-Squared Test on Quasi-Identifiers (Anonymized Data)**:
  - Sex and Political Preference: p = 0.0199
    Significant difference in political preference based on sex.
  - Voting channel and Political Preference: p = 0.0123
    Significant difference in political preference based on voting channel.
  - Education and Political Preference: p = 0.1077
    No significant difference in political preference based on education.
  - Citizenship and Political Preference: p = 0.7271
    No significant difference in political preference based on citizenship.
  - Marital Status (Generalized) and Political Preference: p = 0.6014
    No significant difference in political preference based on marital status (generalized).

These Chi-square test results demonstrate that anonymization preserved the key statistical relationships in the dataset with only minor variations, indicating that the anonymized dataset retains valuable analytical insights.

4 Reflections on Disclosure Risk vs. Utility Trade-Off

Balancing disclosure risk and data utility is a central challenge in anonymization. Here we reflect on this trade-off:

- **Disclosure Risk**: The raw dataset had a K-anonymity of 0, indicating a high risk of re-identification. Following anonymization, the dataset reached a minimum K-anonymity of 2, offering a moderate improvement in privacy protection. Additionally, the increase in L-diversity, with 62.50% of groups meeting the 2-diversity criterion compared to 39.58% in the raw data, demonstrates enhanced resistance to attribute inference.
- **Utility Loss**: The anonymization process involved generalizing certain categories, such as education and marital status, to ensure confidentiality. Despite these modifications, the data retained its statistical integrity, as indicated by the Chi-square test results pre- and post-anonymization. This preservation of statistical relationships allows meaningful insights into demographic patterns and voting behavior.
- **Evaluation**: Overall, the trade-off between disclosure risk and data utility is satisfactory. The transformations applied provide a level of privacy that makes the dataset suitable for controlled public release, without excessively compromising data utility.

In conclusion, our anonymization approach successfully protected individual privacy while preserving analytical insights. The anonymized dataset remains a valuable resource for studying electoral patterns with minimized confidentiality risks.