Report Phase 2

Piotr Grabowski, Peter Tropko, Abel Szemler, Jakub Kozanyi

November 2024

Introduction

This report outlines the methods and techniques employed to deanonymize the anonymized survey data. The primary objective was to identify political preferences for as many survey participants as possible by linking the anonymized dataset with publicly available records.

Methodology

Data Preparation

To achieve the deanonymization, we used these four datasets:

1. **Anonymized Survey Data** (`anonymised_dataP.csv`): Containing anonymized records of participants' political preferences and limited demographic attributes.
2. **Public Population Register** (`public_data_registerP.xlsx`): A publicly available list of individuals in the population with demographic attributes.
3. **Published Election Results** (`public_data_resultsP.xlsx`): Aggregate election outcomes by demographic categories.
4. **Survey Participants List** (`survey_listP.txt`): A leaked list of survey participants' names.

The datasets were preprocessed to standardize field names and formats, ensuring compatibility for comparison.

Record Linkage Process

We took the following steps to identify matches between the anonymized dataset and the public register:

Feature Selection

We identified attributes common to both datasets and used them for comparison:

- **Sex**: Gender of the individual.
- **Citizenship**: Country of citizenship.
- **Marital Status**: Whether the individual is in a relationship.
- **E-voting Status**: Whether the individual voted online.
- **ZIP Code**: Encoded geographic information.

ZIP Code Mapping

To address potential mismatches in ZIP code anonymization, We created a mapping between public ZIP codes and anonymized ZIP codes based on frequency analysis.

Candidate Pair Generation

Using the `recordlinkage` library, we applied a full indexing approach to create all possible pairs of records between the anonymized dataset and the public register. This ensured no potential matches were excluded at this stage.

Feature Comparison

We compared each pair of records across the selected attributes:

- Exact matching was performed for `sex`, `citizenship`, `marital status`, and `e-voting`.
- ZIP codes were matched using the mapped values.

A similarity score was calculated for each pair, representing the number of attributes that matched exactly.

Threshold-Based Filtering

We retained pairs with a total similarity score of 5 out of 5 (perfect match) as final matches. This strict threshold minimized the risk of false positives, prioritizing accuracy over quantity.

Result Compilation

For each matched record, we assigned the corresponding political preference from the anonymized dataset to the public record. This yielded the inferred political preferences of survey participants.

Results

The above methodology resulted in **46 matches** between the anonymized dataset and the public register. We verified these matches by ensuring consistency across all attributes. Political preferences for these individuals were successfully inferred and are summarized in the final dataset.

Anonymization Constraints

- The anonymized dataset included only a single age category (40-59), limiting the discriminatory power of age-based matching.

Dataset Resolution

- Certain attributes, such as marital status and e-voting, were not diverse enough to generate unique matches in all cases.

Conclusion

Our deanonymization process successfully identified political preferences for 46 survey participants by leveraging feature-based record linkage, ZIP code mapping, and strict thresholding for similarity scores.