Final project

Consider following (hypothetical) scenario: for the latest local election, an unnamed country decided to experiment with electoral technologies and to provide a possibility to cast a vote online for voters in a selected municipality. The rest of the voters voted on paper in their polling station, as usual. The results of the election have shown that the votes cast electronically were distributed differently than the votes cast on paper. In order to investigate the possibility of election manipulation, as well as to study the demographics of voters who decided to cast their vote online, a survey was conducted. For the sake of transparency, the statistical agency conducting the survey decided to release the results to the public. However, due to sensitive nature of the survey (in particular, the political preferences of the voters), the released dataset had to be anonymised beforehand.

In this project you will take the role of both the statistical agency (the anonymisers) and the adversary who attempts to learn the political preferences of the population from the released dataset (the deanonymisers). The project would therefore consist of two parts.

Part 1 (2 weeks)

You will receive the following files (in all the file names, X stands for the letter assigned to your group):

- "private_dataX.xslx" representing the non-anonymised survey data,
- "public_data_registerX.xslx" representing the public population register,
- "public_data_resultsX.xslx" representing the published results of the election.

A description of the attributes in the datasets (same for all the groups) is furthermore available on LearnIT.

During the first part, you will work with the non-anonymised dataset ("private_dataX.xslx") containing the attributes as outlined above. Your task would be:

1. Anonymise and submit the anonymised version of the dataset. In this, you should consider following requirements on utility and risk of the anonymisation:
   - One should be able to use the dataset to investigate the following research questions (it is up to you to decide, what kind of analyses exactly should be possible to analyse these questions).

        - (A) Whether the political preferences as expressed in the survey differ from the actual election results (“public_data_resultsX.xslx”) for both electronic and polling station votes
        - (B) Whether there are significant differences between political preferences of voters who cast their vote electronically and the ones who voted on in the polling station
        - (C) Whether these differences still persist if one accounts for various demographical factors
    - The adversary is assumed to have access to the data from the population register (“public_data_resultsX.xslx”) and to the results of the election (“public_data_resultsX.xslx”)
2. Prepare auxiliary data that would allow an attacker to learn further disclosures from your anonymised dataset. This data should not include the political preferences directly (i.e. no posting of survey answers with names intact), but you still need to make sure, that someone with access to the auxiliary data can learn the political preferences of at least some of the voters. [Optional: think about how you would further anonymise the data, if you wanted to protect against disclosures based on your suggested auxiliary data as well]
3. Prepare a report on your anonymisation process following the statistical disclosure process from the lecture (see also sdcpractice.readthedocs.io/en/latest/process.html). Your report will consist of two parts
    1. Private part (visible only to the teachers and the TAs): you will (1) outline the risk and utility measures you have chosen for your anonymisation process, (2) describe the methods you applied and the reasoning behind your decisions, (3) describe, how exactly you would use the non-anonymised and the anonymised dataset to answer each one of the research questions (A)-(C), (4) describe in details, how the auxiliary data you prepared can be used for disclosure, as well as specify the voters (by their names) whose political preferences would be disclosed.
    2. Public part (will be visible to other students): you will report on your anonymisation methods, specifying the data attributes that have been changed, and which anonymisation methods were used for changing the data.

Your submission should consist of (1) the anonymised dataset (as .csv or Excel file), (2) the auxiliary data (as .csv or Excel file), and (3) two parts of the report (as separate PDFs). Submission deadline is: Tuesday 10. November, 14:00. NOTE: no extensions on the submission of the anonymised datasets will be granted, as these are necessary for the second part of the project.

Part 2 (1 week)

You will get access to the anonymised datasets, the auxiliary data and the public part of the report from another group assigned as your “opponent”. Your task would be to learn the political preferences of as many voters as possible, and prepare a report on which techniques you used for this. As further auxiliary knowledge, you will have access to the population register and the public results of the opponent group dataset.

Your submission should consist of the list of reconstructed preferences for the voters in the form <name, party> (as .csv or Excel file), the report (as PDF) and the code you used for your disclosure attacks. Submission deadline is: Tuesday 17. November, 23:59.

Project workshop: each group will present the result of their work on both anonymisation and disclosure attacks. More information on the presentation schedule and contents will follow.

The submissions for each one of the two parts and the participation in the workshop are mandatory.