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 (date of data collection 1/7/2024). 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 phases.

The submissions in both phases of the project will occur in groups, with only one submission per group. The submissions will be done using the Peer Feedback activity.

Phase 1

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

- "private_dataX.xslx" as the raw (i.e. non-anonymised) survey data,
- "public_data_registerX.xslx" as the public population register,
- "public_data_resultsX.xslx" as the published results of the election.

A description of the attributes in the datasets (same for all the groups) is furthermore available here or appendix at the bottom.

Your tasks during the first phase will be:

Step 1: Analyse the data in the raw dataset answering the following questions (note, you are free to choose the suitable methods of the analysis yourself, based on your knowledge of e.g. applied statistics or other courses of your study):

(A) 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?

(B) 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...)?

(C) 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?

Step 2: Anonymise the raw dataset and calculate the disclosure risks (you can use any suitable metrics, e.g. among these that were discussed in the course or presented in optional readings). In these calculations, assume that your anonymised dataset, the public data (that is, the files public_data_registerX.xslx and public_data_resultsX.xslx) as well as your submitted report (see Step 4 below) will be available to a potential adversary, but no other data will be available.

Step 3: Conduct the same analyses as in Step 1, but this time on the anonymised dataset. Note the differences in the analysis outcome compared to the analyses you have performed on raw dataset. If the differences are too large (according to your own evaluation), return to Step 2 and redo the anonymisation.

Step 4: Write a report describing:

1. The methods you used to anonymise the dataset in Step 2. You can use the external reports from the case studies here: https://sdcpractice.readthedocs.io/en/latest/appendices.html#case-study-1-external-report as examples.
2. The disclosure risk metrics you calculated in Step 3. Note, depending on the metrics you used, reporting on them might reveal additional information about the dataset; in that case, feel free to omit details to prevent such disclosure.
3. The results of the analyses conducted in Step 4.
4. Your own reflections on whether the trade-off between disclosure risk you achieved and the utility losses (i.e. the difference between the analyses conducted in Step 1 and Step 4) is good enough. Note, you don't have to include the actual results of the Step 1 analyses in the report.

Your report should be at most 10 pages long, using 12pt font size.

How to submit: Your submission should consist of the following files, uploaded as an attachments in the Peer Feedback activity:

(a) the anonymised dataset as a .csv file named "anonymised_dataX.csv" with X as the letter assigned to your group

(c) the report as a PDF file named “reportX.pdf” with X as the letter assigned to your group

The deadline for the submission of the first phase results is **Tuesday, November 13th, 23:59**.

Phase 2:

In this phase you will be assigned an opposing group among those who submitted their anonymised data in the first phase. You will simulate an attacker in a scenario when a list of survey respondents from that group has been leaked (note that this scenario simulates a violation of the assumption made by anonymisers during the first phase), and your goal is to use this data to learn information about them from the anonymised dataset. You will therefore have access to the following information (in all the file names, Y stands for the letter assigned to your **opposing** group):

- "anonymised_dataY.csv" as the anonymised survey data,
- "public_data_registerY.xslx" as the public population register,
- "public_data_resultsY.xslx" as the published results of the election.
- "survey_listY.xslx" as the names of people who took part in the opposing group’s survey (that is, whose data is included in the anonymised dataset)
- "reportY.pdf" as the report about the anonymisation submitted by the opposing group

Your task will be two-fold:

- Given the data above, learn the the political preferences of as many voters as possible
- Prepare a report describing the techniques you used (at most 5 pages using 12pt font). You don't have to provide the exact code you used for deanonymising the dataset, but your description should be detailed enough so that someone reading the report can replicate your results given the same dataset.
- Provide your feedback for the report of your opposing group. Your feedback, in particular, should address the following questions:
  - Is the overall language of the report clear and understandable?
  - Is the methodology for analysing the data (i.e. answering questions (A)-(C)) reasonable and well-described?
  - Is the methodology of the anonymisation well-described?
  - Is the achieved trade-off between privacy and security of the dataset reasonable?
  - What are the overall strengths and weaknesses of the report?

How to submit: Provide the results of your deanonymisation (i.e. the list of voters together with their political preferences that you were able to identify), your description of the used deanonymisation techniques and your feedback to the opposing group as your assessment to the Peer Feedback activity.

The deadline for the submission of the second phase results is **Tuesday, November 20th, 23:59**.

Project workshop:

Each group will present the result of their work on both anonymisation and disclosure attacks during the workshop on **Friday, November 22nd**. The presentation will include information on their own anonymisation, including but not limited to.

- Description of anonymisation methods they chose and discussion on why these methods were chosen
- The effects of anonymisation on the utility of the dataset, e.g. as the resulting difference between the results of the analyses run on research questions (A)-(C) on anonymised vs non-anonymised dataset

More information on the presentation schedule and contents will follow.

The project will not be graded, however, the submissions for each one of the two parts and the participation in the workshop are mandatory.

Frequently asked questions about the project will be answered in the forum.

Appendix: Dataset attributes

Attribute	Survey	Population Register
name	Last and first	
sex	Female, Male	
dob	Date of birth (DD/MM/YYYY)	

zip	ZIP code of the voter’s address (2100, 2200, 2300, or 2400)	
evote / last_voted	Whether the voter cast their vote electronically: – 0: vote cast on paper (polling station) – 1: vote cast electronically	How the voter cast a vote during the most recent e – 0: vote cast on paper (polling station) – 1: vote cast electronically – 2: no vote (abstained)
party	How the voter has voted: Red party, Green party, Invalid vote (spoiled ballot)	-
marital_status	Marital status: Never married, Married/separated, Divorced, Widowed	
education	Level of education: Primary education, Upper secondary, VET, Short cycle higher education, Vocational bachelor's, Bachelor's, Master's, PhD, Not stated	-
citizenship	Name of the country	

Last modified: Friday, 1 November 2024, 14:11