Coursework: Protecting Data

Julie Terrassier, Maria Smirnova, Asia Belfiore, Daniel Peacock, Elsa Polo Laube

CLASSIFICATION OF THE ATTRIBUTES

Firstly, we identified whether attributes were identifiers, quasi-identifiers or sensitive information:

- Identifier: *phone_number* (since it can uniquely identify each person in the dataset)
- Quasi-identifiers: *area*, *postcode*, *dob*, *gender*, *ethnic_group*, *marital_status*, *qualifications*, *occupation*, *home_ownership* (combined, these attributes can identify a person)
- Sensitive information: *income* (since it is part on the individual's financial situation) and *distance_to_work_km*

METHODS APPLIED

Pseudonymisation

We performed pseudonymisation of the identifier by salting and hashing the phone numbers (since a hash function alone could be brute forced by knowing that phone numbers are 11 digits). We used SHA-256 as it does not have collisions and generated a random long salt (length 12) to make it computationally infeasible to brute force. We then use the hashed values as indexes.

Generalisation and suppression

To ensure there were no small equivalence classes, we employed non-perturbative methods (generalisation and suppression): we evaluated entropy and k-anonymity at each step and decided between suppressing or generalising attributes based on how necessary the data was for each use case. To avoid compromising utility, we tried to decrease the loss of information, while staying meaningful to derive statistically truthful conclusions from them.

Area	Postcode	Date of birth	Gender	Ethnic Group	Marital Status	Qualifications	Occupation	Distance to Work	Home Ownership
S	S	G	-	G	S	G	G	G	-
Region		Age	-	-	-	-	-	-	-

Table 1: Method applied to each original attribute (G = Generalisation, S = Suppression) and resulting changed attribute.

Region: Since postcode and area were mostly unique, we coupled and generalised them into regions using a table conversion to map postcodes to their corresponding regions. Although they are quite vast areas (e.g. "Wales"), the use cases mentioned "regions", so this wouldn't compromise utility. Also, this would jeopardise the dataset the least, as the alternative would be to delete 20% of the records for 4-anonymity.

Age: Since the date of birth provides too much information about individuals, we first generalised it into age (this doesn't impact utility since use case 3 only needs the age). We then tried different mappings from ages to age ranges and considered the impact on k-anonymity as well as the loss of information induced (using entropy) to find the best mapping. We compromised on the age ranges: 16-20, 20-30, 30-40, 40-50, 50-60, 60-70 and 70+.

Ethnic group: We generalised ethnic groups into larger groups by keeping only the part before the ":". However, since use case 1 wants to compare pay inequities amongst different racial groups and knowing the dataset is focused on people working in the UK, it made more sense to keep the "White: English, Welsh, Scottish, Northern Irish or British" category separate.

Qualifications: Merged into four categories based on education level, since we considered these to be the most relevant ones to the use cases and it only induced a 5.57% loss of information.

Occupation: Generalised into five groups, taking into account that this attribute might be used to identify pay inequities. Not generalising this attribute, despite its utility, was unfeasible, as we would have had to delete too many records to achieve more than 1-anonymity.

Marital status: Suppressed as it was not relevant to any of the use cases and its presence significantly impacted the number of rows to be deleted to have more than 1-anonymity.

Gender: We chose to keep the gender column because it could be another reason for pay inequality, so the attribute is important to nuance inequalities due to race (for use case 1).

Suppression: After we performed exhaustive generalisation, while minimising the impact on utility, we checked the distribution of the equivalence classes by size and decided to perform suppression to get a reasonable k-anonymous dataset. We found that deleting 10.98% of the records to obtain a 4-anonymous dataset was the best compromise between a high k and a low number of deleted records.

Distance to work: Although a sensitive attribute, the original information was too precise and would have facilitated re-identification. We rounded the distances to two decimals and kept the 0 distance separate to retain utility, since it represents individuals who work from home.

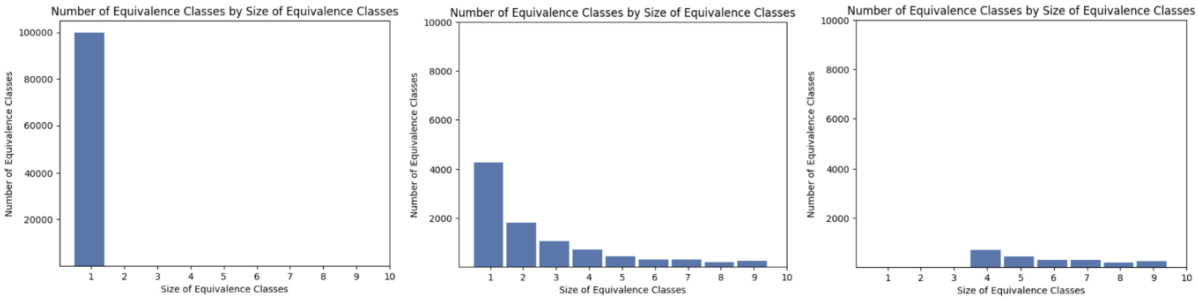

Figure 1: Number of equivalence classes by size before any modification (left), after generalisation (centre) and after suppression (right).

ASSESSMENT OF THE PRIVACY LEVEL

Our anonymised dataset is 4-anonymous, 2-diverse and has 0.7-closeness. These metrics guide us in ensuring our dataset is suitably de-identified.

4-anonymity: It means that each equivalence class contains at least four records. This presents a good privacy level since no individual can be identified from the quasi-identifiers.

2-diversity: We evaluated l-diversity to ensure the dataset would not be vulnerable to homogeneity attacks (we needed $l > 1$). To achieve 2-diversity when only considering the attribute *distance_to_work*, we had to remove 124 records (corresponding to a 0.06% loss of information). To achieve 3-diversity, we would have had to remove 599 additional records (0.35% loss of information), which did not seem worth it. For *income*, it already achieved 2-diversity without needing to modify the dataset. To achieve 3-diversity, we had to remove 20 records (0.01% loss of information), which we did since it would not

impact the data significantly but would increase privacy protection. Overall, our de-identified dataset is 2-diverse, but it is 3-diverse when considering the income.

0.7-closeness: In particular, *income* achieves 0.5-closeness and *distance_to_work* achieves 0.7-closeness. Ideally, we would need t-closeness to be much lower to avoid skewness attacks, however, here we made the choice to not modify the dataset further so as to not impede utility.

k-anonymous	l-diversity	t-closeness	Suppressed records	Suppressed attributes	Generalised attributes	Loss of information
k = 4	l = 3	t = 0.7	11,11% (10985 for 4-anonymity + 124 for 2-diversity)	2 (marital status, phone number)	7 (age, occupation, qualifications, distance to work, ethnic group, area, postcode)	33.69%

Table 2: Properties of the final de-identified dataset.

ASSESSMENT OF THE REDUCTION OF UTILITY ENTAILED

We found that the best compromise between preserving utility and more robust anonymization was given by the 4-anonymous dataset, which we chose as our final one with a loss of 33.6% of information through the suppression of about 10.99% of the records overall.

	Original	k=2	k=4	k=5	k=7	k=10
Entropy *H(D)*	16.61	11.42	11.52	10.87	10.66	10.42
Entropy Ratio $H(D_k)/H(D)$	1	0.687	0.664	0.654	0.642	0.627
Number of rows removed	0	4252	10985	13737	17573	22085

Table 3: Calculated entropies, entropy ratios and number of records to be deleted to achieve k-anonymity

To evaluate the reduction of utility, we mostly used entropy. We calculated entropies of the original dataset and of the dataset after performing generalisation and suppression and used the ratio of entropies to quantify the loss of information (Table 3). For some attributes (like ethnic groups), we also applied logic to our generalisation process, to make sure the utility was preserved according to the use cases.

Use case 1: Although incomes were not modified, our generalisation of ethnicities may have created biases towards minorities within each main ethnicity and compromised patterns. We tried to reduce this risk as much as possible, for instance by keeping the ethnic group "White: English, Welsh, Scottish, Northern Irish or British" separate. We chose to keep the *qualifications* and *occupation* attributes since we considered these important in evaluating pay inequities. Finally, we also kept *gender* because it could be another reason for pay inequity, so the attribute is important to nuance inequities due to race.

Use case 2: We were able to retain the type of information needed: regions and home ownership situations. So there is almost no utility reduction in this case.

Use case 3: We divided the ages into reasonable ranges by splitting them into the main stages of life according to work/career situation, to reduce utility as less as possible. Note however, that generalising age alone constituted a 15% loss of information and could have skewed data patterns due to the reduced resolution. The *distance_to_work* attribute however was kept almost as precise as originally, so it does not constitute a big utility reduction.

REMAINING PRIVACY RISKS

Firstly, semantic attacks are still possible: if the income ranges of records in an equivalence class are somewhat similar, an attacker can still determine indicatively the level of wealth of a target individual. This is not easily solvable, as even if we were to generalise income values into ranges (and potentially harming utility) the attacker would still be able to get to the same conclusions.

The dataset is also vulnerable to potential skewness attacks: an attacker could, for example, learn that the distribution of wealth in a certain region is much higher than in other areas, or that incomes are skewed based on the education level, ethnicity or even distance to work.

Finally, there could still be a risk of identification, depending on the combinations of quasi-identifiers known by an attacker. We would have to simulate attacks with every possible combination and find if an individual is uniquely identifiable.

POTENTIAL CHANGES

If we assume no possibility of data leakages and collusions, and if the dataset was only to be shared with trusted researchers, the first thing we would do would be to de-generalise some attributes to have a more precise dataset that allows for more accurate studies and statistical analysis. For instance, we could reduce the number of categories for *occupation*, *qualifications* and *ethnic groups* to increase utility.

Regarding the date of birth, we would still keep it as an age, since the full date of birth is unnecessary for the use cases, but we would not have it in ranges anymore, to improve precision and flexibility in studies.

Additionally, because *distance_to_work* is sensitive and there is no risk of re-identification anymore, we would de-generalise it.

Re-introducing marital status could also be useful, in case of further studies.

However, in reality, there could be leaks. Even if the researchers are trust-worthy, they may for instance be hacked. This implies that we should assume the same level of privacy as if the dataset was to be publicly released.