Privacy Engineering Coursework MAX 4 PAGES & 1500 WORDS

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

1. Sensitive Information

- **Identifier:** we chose the identifier to be the *phone number* attribute, as it is unique for each data entry.
- **Sensitive attributes:** we decided that *income* is a sensitive attribute, since it is part of the individual's financial situation which we identified as to be protected. Additionally, we identified *distance_to_work* to be a sensitive attribute, as coupling this information with the knowledge of the area where an individual lives in, may allow an attacker to track down an individual with high certainty.
- **Quasi-identifiers:** the quasi-identifiers are *area*, *postcode*, *dob* (date of birth), *gender*, *ethnic_group*, *marital_status*, *qualifications*, *occupation, home_ownership*.

2. Methods Applied

a. Pseudonymisation of the Identifiers

We firstly performed pseudonymisation of the identifier by salting and hashing the phone numbers to enhance protection (since a hash function alone could be brute-forced with a lookup table by knowing that phone numbers are always 11 digits). We used *SHA-256* for the hashing function as it doesn't have collisions and we generated a random salt (from the seed 345) of length 256 to make it computationally infeasible to brute force a rainbow table.

b. Generalisation and Suppression for *k*-anonymity

We used k-anonymity throughout our anonymisation process as a guide for the best parameters and strategies. Through a combination of *non-perturbative methods* (generalisation and suppression), we were able to achieve *4-anonymity* thus ensuring that no combination of quasi-identifiers would uniquely identify an entry. At each step, we evaluated k-anonymity and decided between *suppressing* (deleting a whole attribute or entries) or *generalising* (creating ranges so an attribute is less specific) attributes based on how necessary the data was for each use case. To avoid compromising utility, we aimed to keep the number of generalised records similar across the different value ranges, while ensuring that these were still meaningful and usable 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 → We coupled the postcode and area and generalised them into *region*. We initially tried removing the second half of the postcode, however this still resulted in mostly equivalence classes of size 1. We then tried suppressing the postcode entirely and using only the area provided, but this was still not enough. Our next approach was to use the UK "area code to region" table conversion (https://ideal-postcodes.co.uk/guides/postcode-areas) and, using a UK postcode parsing python library (*uk_postcodes_parsing*), we got the area code from each postcode in the dataset and mapped them to the region table.

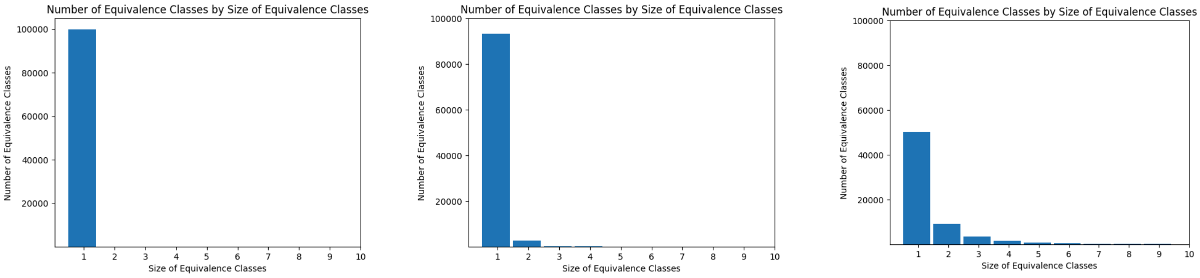

Figure 1: Number of equivalence classes of each size k before generalising (Left), after using area and postcode (Center) and after using Region and suppressing area and postcode (Right)

Although we had concerns about using the region attribute since some values are quite vast areas (e.g. "Wales"), we relied on the fact that the use cases were not specific about the size of the areas, and decided this would not compromise utility too much. Furthermore, we decided that this option would jeopardise the dataset the least, as the alternative would have been to delete up to 20% of the records in order to achieve 4-anonymity.

Age → Since the original full date of birth provided too much information about a single individual and had mostly unique entries, we generalised the date of birth into age groups. We first only kept the year of birth, however, this still resulted in unique data entries. We then decided to group individuals into different age categories, as this would not impact the use cases for which this attribute is relevant. We converted the year of birth to age (assuming the year it was collected in was 2024) and tried different ranges to minimise the number of size 1 equivalence classes, and ended up compromising on 5 ranges: 15-30, 30-45, 45-60, 60+.

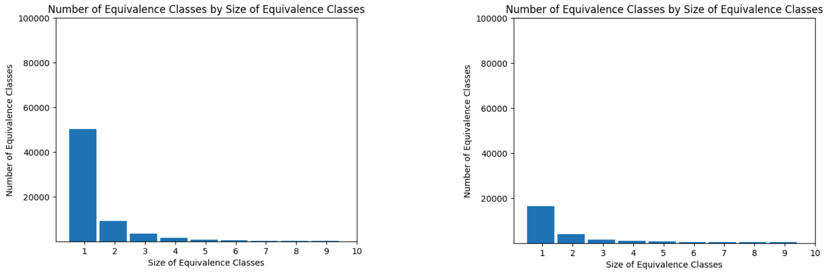

Figure 2: Number of equivalence classes of each size k before (Left) and after (Right) generalising dob to age groups.

Ethnic Group → We generalised all ethnicities into 5 classes by grouping every ethnic background into its main ethnicity (eg. any background under White (Irish, Roma, English, etc.) was grouped into white, and so on).

Qualifications → We merged the qualification attribute into two 2 classes: Level 4+ education (i.e. Higher Education) and under, as we considered these to be the most relevant ones to the use cases.

Occupation → We were initially against generalising the occupation attribute, however, we realised that this was not a feasible approach, as it would have required us to delete a significant amount of records in order to achieve k>1 anonymity. We decided to generalise the occupations into 5 groups based on the skill level and field of occupation: '*Management and Professional Occupations*', '*Administrative and Support Occupations*', '*Skilled and Technical Occupations', 'Service and Care Occupations' and 'Unemployed*'.
Marital Status → Although we initially considered generalising the marital status into two groups ('*Formerly or Currently Married*' and '*Never Married*'), we decided to suppress this attribute as it was not centrally relevant to either of the use cases and since its presence significantly impacted the number of rows that had to be deleted to achieve k>1 anonymity.

TODO: image

Figure 3: Final distribution of equivalence classes after all generalisation and suppression steps have been applied (Right).

Distance to Work → Although a sensitive attribute, we found that the original information that *'distance_to_work_km'* revealed was too precise and would have facilitated the re-identification of individuals when coupled with the 'region' attribute. We decided to generalise the distances into groups based on kilometre ranges. We tested ranges of 3,5,7 and 10 kilometres to check the resulting equivalence classes, and decided to use ranges of 5km, but keeping the 0 distance as a separate class, since it represents individuals who work from home (or are unemployed).

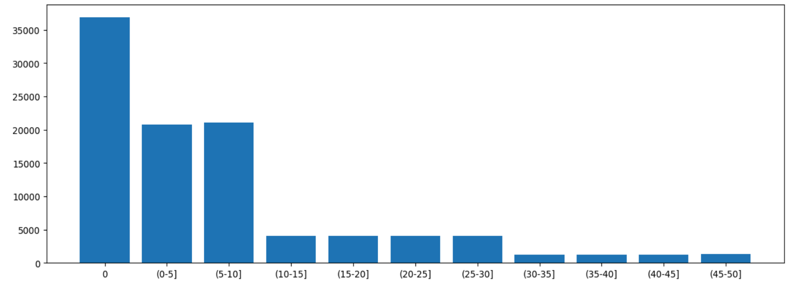

Figure 4: Distribution of entries within given ranges for 'distance to work (km)' (Left).

c. Privacy parameters

- **Entropy:** We used the *Entropy* as a guide to evaluate and choose different *k* values for k-anonymity. The entropy of the original unchanged dataset was approximately 16.6, with almost all unique entries. After performing generalisation and suppression we achieved the Entropies described below in Table 2. To quantify the amount of data lost from the anonymization process, we used the ratio of the Entropies of the unchanged (original) dataset and the anonymised one with different k-values:

	Original Dataset	**2-Anonymous**	**4-Anonymous**	**5-Anonymous**	**7-Anonymous**	**10-Anonymous**
Entropy: $H(D)$	16.61	16.27	15.76	15.57	15.28	14.92

Entropy Ratio $H(D)/H(D_k)$	1	0.980	0.949	0.937	0.920	0.898

Table 2: Calculated Entropies (Row 1) and Entropy Ratios (Row 2) for each k-anonymous Dataset.

We found that the best compromise between preserving utility (less data manipulation) and more robust pseudonymization (lower entropy ratios) was given by the 4-anonymous dataset, which we chose as our final one with a loss of 5% of information through the suppression of 8% of the records overall.

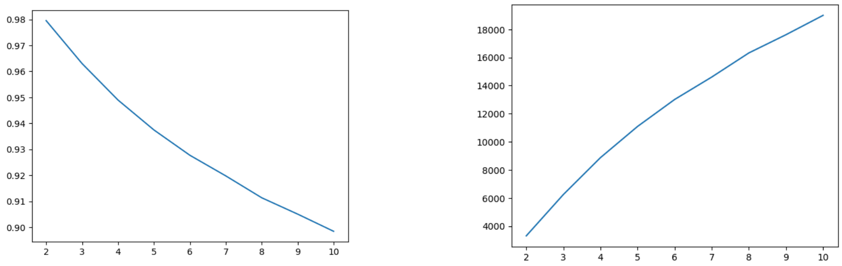

Figure 5: Variance of Entropy Ratios by k-equivalent dataset (Left) and number of records to delete in order to achieve k-anonymous dataset (Right).

- **L-diversity:** We calculated *l*-diversity on the three sensitive attributes and achieved the following l-diverse datasets:

- **T-closeness:**

Final Dataset

k-Anonymous	**l-Diversity**	**t-Closeness**	**Suppressed Records**	**Suppressed Attributes**	**Generalised Attributes**	**Information Lost**
k = 4	l = 3	t = 0.6	8% (8000)	3 + 1 (Marital Status, Area, Postcode + Phone Number)	6 (Age, Occupation, Qualification, Education, Distance to Work, Ethnic Group)	5% (Entropy Ratio=0.95)

Table 3: Overview and Properties on final anonymised dataset.

3. Assessment of privacy

With our modified dataset, we achieved 4-anonymity, a 3-diversity and 0.6-Closeness. Inside each equivalence class, the most unbalanced repartition we obtain for ownership are TODO% (being an owner) and TODO% (otherwise). Also, if we consider only the salary and ignore ownership, the dataset is TODO-diverse. All of these quantities ensure a reasonable amount of privacy.

4. Assessment of utility

We however expect you to explain your reasoning when evaluating utility and **emphasise the limitations of your approach.**

Should we assess each use case?
While trying to increase privacy, we had to be careful to keep in mind some trade-offs, in order to preserve utility. For use case 1, salaries have been split into meaningful and balanced ranges. Moreover, racial groups have been generalised but kept precise. For use case 2, we have exactly the amount of information we need: regions and ownership situations. For use case 3, the ages are divided into reasonable ranges (since they are split into the main ranges of life, regarding work situation), and distances to work ranges are also meaningful, with an emphasis on 0 (= working from home).

Therefore, the entropy achieved with our current dataset is TODO, whereas with the initial dataset it was TODO. The decrease is due to the preservation of privacy, but the entropy still indicates relevance in the information and utility preservation.

5. Remaining privacy risks

Firstly, by knowing one of the sensitive attributes, one can manage to identify an individual. For instance, if you are an employer, you know the salary of your employee and so you could find them in the dataset. Although, it is purposeless because the only information they would get is the ownership situation, which may be already known by an employer.

Also, semantic attacks are still possible. For example, if someone wants to imprecisely determine the level of wealth of an individual, they can just look at the different ranges present and establish if the person has an extreme income or not. This is not something easily solvable: we would need to generalise too much and harm utility.

Likewise, potential skewness attacks still exist. When we discussed privacy, we highlighted that, for some equivalence classes, we could have a probabilistic idea of the salary or the ownership situation of a person.

6. Potential changes/Conclusion

[We need to discuss this]
What would be the risks if they are trusted? Maybe just a leak of the dataset but nothing with trying to identify a person?

If the dataset was only to be shared with trusted researchers, we would degeneralise some attributes to have a more precise dataset, and so to allow the studies to be as accurate as possible. For this, we would keep the 15 different ethnic groups, as well as the initial qualifications and occupations attributes.

A plain date of birth would still be too precise and unnecessary for these use cases, even though the privacy would not matter that much anymore (incorrect if we consider the possibility of a leak and maybe other things too).