Privacy Engineering Coursework Report

Yasmine Hemmati, Sarah de Jong, Bin Sheng Ang, Sean Xiang Peng Ng

1 Introduction

Our objective is to implement data de-identification techniques to protect the privacy of the people in our dataset while preserving its utility for these use cases:

1. To find locations where poor access to education leads to deprivation and more investment in high-quality primary education is needed.
2. To analyse the impact of gender on income.
3. To study the fairness and bias of the credit score algorithm

2 The dataset

We decided that income, on_benefits, home_ownership, and credit_score are sensitive attributes, since they provide information regarding an individual's financial situation. Qualifications and occupations are not deemed sensitive, as they are commonly public information. The quasi-identifiers are area, postcode, dob (date of birth), gender, marital_status, num_children, qualifications, and occupation. The identifier is name.

3 Pseudonymization of the identifier

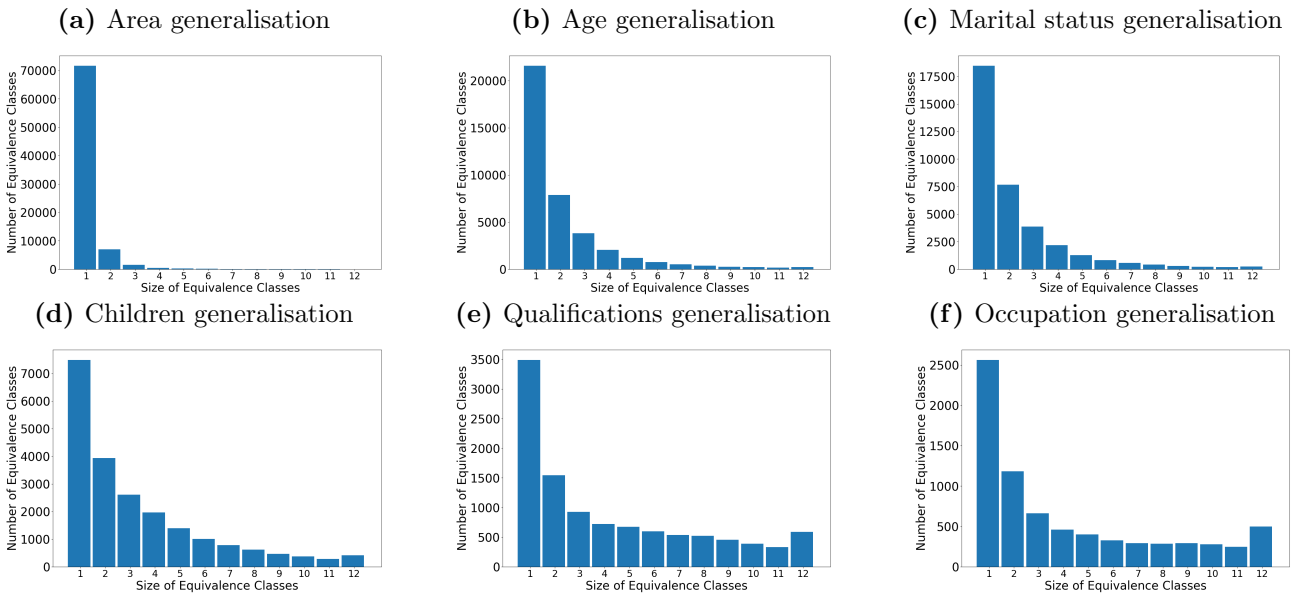
As there are multiple data with identical names, we chose to pseudonymize by combining **name** and **postcode**. Appending the postcode after the name guaranteed uniqueness in the provided dataset. We decided to use the sha3_256 cryptographic hash function with a randomly generated salt, **l#dfbOQ3CUasdsla'; flsafk0q31032U0[1XRP]eqr3**, appended to the end of the combined name and postcode. The long and random salt used virtually prevents a brute-forced attacked. We chose to use the sha3_256 hash function because it provides resistance to length extension attacks and a better performance compared to the previous versions of sha hash function.

4 k-Anonymity

Implementing **k-anonymity** ensures that each dataset record is indistinguishable from at least $k-1$ other records based on every set of quasi-identifiers. An **equivalence class** represents records with identical quasi-identifier values. Thus, k-Anonymity guarantees that an attacker cannot single out a person in the dataset and identify them. If all individuals in an equivalence class share the same sensitive attribute or if the distribution over sensitive attributes is skewed, the attacker can still learn sensitive information about the individual despite k-anonymity. To address these privacy concerns, we consider ℓ-diversity and t-closeness, as discussed in Section
To achieve k-anonymity, we utilised both generalisation and suppression methods.In decision-making, we consistently weigh the utility versus privacy trade-off. Given that we have both age and area, our first step was to remove the date of birth and postcode columns from the dataset (suppression), since this makes people uniquely identifiable. 97358 equivalence classes and 95426 uniquely identifiable people remain. We assessed the utility of each column and evaluated the impact of various column generalizations on the required number of dropped rows for different k-anonymity levels. We will now discuss the generalisations we applied to variables that can have impact on our use cases. The distribution over the equivalence classes for $k \leq 12$ is shown in Figure 1.

Area generalisation: We generalised the areas based on the International Territorial Levels (ITL) published by the UK Office for National Statistics. This will highly impact the utility of our dataset for use case 1, but still retain some location information. After this generalisation, 81432 equivalence classes and 71623 uniquely identifiable people remain, see Figure 1a.

Figure 1. Equivalence classes after generalisation of each size k for $k \leq 12$



Age generalisation: We generalised age into 4 groups: 16-29, 30-44, 45-59, 60+. 39622 equivalence classes and 21572 uniquely identifiable people remain, see Figure 1b. Though this largely impacts the number of equivalence classes, even to achieve 3-anonymity we would have to drop almost half of our data and we decided we needed to generalise more.

Marital status generalisation: We grouped *divorced* and *widowed* into *previously married* and retained *married* and "never married,". 36850 equivalence classes and 18455 uniquely identifiable people remain, see Figure 1c.

Number of children generalisation: We generalised the number of children into 3 groups: 0, 1-3, and 4+. 22472 equivalence classes and 7485 uniquely identifiable people remain, see Figure 1d.

Qualifications generalisation: We group levels 1, 2, and 3 into *Lower Education* and *Apprenticeship* and *BA* into *Higher Education*. 12827 equivalence classes and 3489 uniquely identifiable people remain, see Figure 1e.

Occupation generalisation: We generalised the occupations from 10 to 7 groups by grouping 4 and 5 (*Administrative and secretarial* and *Skilled trades*), grouping 6 and 7 (*Caring, leisure and other* and *Sales and customer service*), and grouping 8 and 9 (*Process, plant and machine operatives* and *Elementary*). 2561 unique individuals and 9911 equivalence classes remain, see Figure 1f.

From Figure 1f we see that to even achieve 3-anonymity, a significant number of rows must be dropped. However, we must balance the loss of utility from dropping rows with the loss from further generalization. We are able to achieve 9-anonymity by dropping 17039 rows. We decide to pick $k = 9$ as we believe it is a good trade-off between minimising dropped rows while avoiding excessive generalization. Opting for a lower k not only compromises privacy but also complicates the implementation of l-diversity. In section 6.4, we discuss in more detail the effect of k on the number of rows dropped and utility.

5 ℓ-Diversity

A dataset has **ℓ-diversity** if every equivalence class contains at least ℓ distinct values for the sensitive attributes, used to prevent homogeneity attacks. Note that we need to consider each sensitive attribute separately to safeguard against information exposure to attackers.

Making each equivalence class 2-diverse w.r.t. home_ownership and on_benefits would require dropping 49353 rows for $k = 9$. However, we consider information the most sensitive if it reflects negative societal norms, i.e. learning someone is on benefits is more sensitive than that they are not. Hence, we dropped equivalence classes with everyone on benefits or without home ownership. For $k = 9$, this requires dropping 4195 rows. Subsequent sections will address the impact on utility and the decision not to prevent additional homogeneity/skewness attacks.

6 Evaluation on Utility

6.1 Use Case 1

For use case 1, there is an obvious loss of utility from the de-identification, due to our generalisation of districts to the 10 large ITL areas and type of lower education people achieve. To measure the utility of use case 1, we generalised the 318 districts in the original dataset into the 10 ITL areas for easy comparison with our de-identified dataset. We then computed the proportion of people on benefits and of Lower Education and ranked them in order of deprivation. In this evaluation, we consider someone to be deprived if they are on benefits. We consider utility to be significantly impacted if the size of the intersection of the top 3 most deprived areas before and after de-identified is less than 2 as this results in drawing very different conclusions.

6.2 Use Case 2

For use case 2, similar to the impact of generalising area in use case 1, there is a loss of utility from de-identification. This is due to generalisation of occupations which inhibits a more fine-grained analysis. Hence to analyse the impact of utility on the 2nd use case, we first generalised the occupation of the original dataset and thereafter employed the student's t-test for both datasets, under the null hypothesis that males and females in the same occupation class have the same mean.

6.3 Use Case 3

To evaluate the impact on utility for use case 3, we decided to use information entropy to quantify the reduction in information. This is because nearly every column could influence fairness and bias in the credit score algorithm analysis, depending on the researcher. For example, one researcher might argue that a married person having a higher credit score is fair because they can rely on their partner financially, while another might argue it is biased, as credit scores should be based on an individual's financial capability.

6.4 Privacy Parameters Choosing

k	2	3	4	5	6	7	8	9	10	11
Dropped for k-anonymity	2561	4925	6911	8751	10761	12723	14767	17039	19658	22428
Dropped for l-diversity	6852	6294	5949	5681	5351	5074	4683	4185	3844	3434
Significant Impact	N	N	N	N	N	N	N	N	Y (2)	Y (2)

Table 1: Number of rows dropped after applying k-anonymity and ℓ-diversity and impact on utility

From Table 1, we found that 9-anonymity had the best trade-off between privacy and utility as we were able to draw similar conclusion as the original dataset using the de-identified dataset. This is because 10-anonymity impacted use case 2's utility as it resulted in different conclusions drawn from the dataset.

Using the 9-anonymity dataset and applying our utility evaluation methods, we found that:

(a) Original Dataset

Area	Proportion
East Midlands	0.809191
West Midlands	0.808411
North East	0.804627
Yorkshire and the Humber	0.801070
Wales	0.798287
North West	0.794777
South West	0.785903
East of England	0.767637
South East	0.743353
London	0.683931

(b) Anonymised Dataset

Area	Proportion
Wales	0.973846
North East	0.957627
East Midlands	0.955514
East of England	0.936929
West Midlands	0.919218
North West	0.908700
South West	0.887486
Yorkshire and the Humber	0.885235
South East	0.795799
London	0.726810

Table 2: Proportion of People on Benefits and with Lower Education per Location

Comparing the tables in Table 2, the rankings of the proportions are slightly shuffled in our de-identified dataset. Although both East Midlands and North East areas remain in the top 3, when attempting to identify the area needing the most investment in our de-identified dataset, we might be misled into believing that the area requiring the most investment is Wales, when, it is East Midlands. We decided this is a reasonable trade-off between utility and privacy for **use case 1**.

	Student's t-test p-values for male and female income						
Occupation class	1	2	3	4 and 5	6 and 7	8 and 9	10
Original Dataset	0.766	0.698	0.581	0.029	0.157	0.982	0.985
De-identified Dataset	0.172	0.268	0.784	0.018	0.067	0.528	0.359

Table 3: p-value comparing male and female income for each occupation class before and after anonymisation, where the occupation class corresponds to: 1 - Managers, directors and senior officials, 2 - Professional, 3 - Associate professional and technical, 4 - Administrative and secretarial, 5 - Skilled trades, 6 - Caring, leisure and other, 7 - Sales and customer service, 8 - Process, plant and machine operatives, 9 - Elementary, 10 - No Occupation

From Table 3 and using a significance level 0.05, we can draw the same conclusion for all the occupation classes from the original and de-identified dataset. This shows we were able to retain most of the utility of the dataset for **use case 2**.

The entropy ratio, for **use case 3**, calculated was ≈ 0.683, which signified we are able to retain most of the information in our de-identified dataset. We decided this is a reasonable trade-off to ensure better privacy, as the de-identified dataset still has enough utility for this use case.

7 Remaining Privacy Risks

7.1 Homogeneity and skewness attacks

We did not implement a strategy to prevent a homogeneity/skewness attack for income and credit_score. We explored using t-closeness to limit the distance between equivalence class income distributions and the entire dataset. However, we found large differences in income distribution between different equivalence classes, which is understandable due to the small size of the classes and the strong relationship between income and occupation. Achieving t-closeness therefore required dropping many rows, and we deemed it infeasible to maintain utility. For similar reasons, we also did not implement t-closeness for on_benefits and home_ownership.

7.2 Inference attack

We dropped equivalence classes in which everyone is on benefits or everyone does not own a home (i.e. homogeneous in the negative sense). This causes gaps in the dataset, potentially allowing an attacker to deduce sensitive information by analysing patterns or gaps in the released data. For instance, if we drop a certain equivalence class with occupation 5, but there are many people (much more than k) with the same quasi-identifiers with occupation 4 and 6 in the de-identified dataset of which the majority is on benefits, the attacker might expect occupation 5 to be dropped due to everyone being on benefits. Note though that they cannot determine conclusively whether there were $\leq k$ people with these quasi-identifiers or these people were on benefits. These patterns might allow experienced attackers or machine learning algorithms to infer those classes.