Anonymity, Utility and Risk: an Overview

Sofiane Mahiou

Dr Fintan Nagle

July 11, 2018

Abstract

Data anonymisation is a rapidly-changing field, and Hazy must stay up to date with recent work in order to deliver anonymisation as a service. Here we present an overview of anonymisation and present three main aspects of the field: anonymity, utility and risk. For each subject, we provide introductory definitions as well as more detailed solutions from the reviewed literature. Each algorithm or metric is provided with an example of pseudo-code in the hope of giving a clear and concise approach to the field of anonymisation.

Contents

1 Introduction

In May 2018, new legislation designed to reform the legal framework in order to ensure the rights of EU residents to a private life will be enforced. The General Data Protection Regulation (GDPR) is designed to enable individuals to better control their personal data; as such, each individual has the right to request to be anonymised within any corporation or institution's database. Furthermore, with the exponential growth of personal data Brown (2015) available online comes the need to automatically anonymise all personal information, with as little as possible human intervention.

However, most modern organisations and institutions rely heavily on data to function properly and efficiently Newman (2017), in such a case, simply removing all data for a given customer or for a large set of customers might endanger the well-being of the impacted organisations. The challenge is then to find a way to abide by the GDPR and therefore provide each individual with her right to privacy while keeping most of the useful information that a business might rely on.

There is a wide field of literature on anonymisation, providing multiple potential solutions each with its own advantages and shortcomings . However, such literature tends to focus on a specific aspect of the anonymisation problem. The aim of this document is to present an overview of the various aspects of the anonymisation problem. We highlight three main concepts: **anonymity**, **utility**, and **risk**.

In this document, each concept will be described and various methods that relate to it will be presented in detail.

2 Useful anonymisation concepts

In order to understand the various aspects of anonymisation that are presented in the following sections, it is necessary to be aware of the various **anonymisation techniques** as well as the way information is **categorised** during the anonymisation process.

2.1 Information type

When it comes to protecting personal information stored in a given database, such data tends be categorised in one of the following groups as defined by OECD (2001)

- **Quasi-identifiers** : Pieces of information that are not of themselves unique identifiers, but provide sufficient information so that when combined, can uniquely identify an individual. For example : *gender, age, nationality*
- **Direct identifiers** : Information that on its own can directly identify an individual. For example : *Social Security Number*
- **Sensitive information** : Personal information about a specific individual. This is typically information that is to be protected by avoiding the identification of the individual. A typical example in medical record is the *diagnosis* of a given patient.
- **Insensitive information** : Information that is to be ignored during the anonymisation process.

These categories allow an algorithm to anonymise appropriate attributes while leaving intact the remaining information. In order to allow data analysis, as little insensitive information as possible should be anonymised.

In general, the sensitive attributes in a database are left unchanged. Instead, direct identifiers and quasi-identifiers are anonymised using tools presented below in order to ensure that no individual can be identified.

2.2 Anonymisation techniques

In order to anonymise a set of records containing direct identifiers and quasi-identifiers, several methods have been presented in the literature. Such methods either modify or delete personal information. Radical methods that imply deletion of the information tend to be applied to either direct identifiers or, even after softer anonymisation methods have been applied, to unique outliers within the set of quasi-identifiers.

2.2.1 Perturbation

Data perturbation is a popular anonymisation method Hsu (2015) which guarantees privacy by directly modifying records in a non statistically significant manner. This allows for the data to be anonymised while ensuring that the statistical usefulness of the data remains the same. The main shortcoming of such a technique is that the data within the anonymised data is **no longer accurate or truthful** even if statistically useful; therefore any mechanism that relies on the accuracy of such anonymised data would be put a jeopardy.

As presented by Zhou et al. (2008), several methods of perturbation exist. *Micro-aggregation* as well as *random perturbation* are among the most popular techniques.

- **Micro-aggregation** : Assuming that there exists a distance measure for a given personal attribute such as age, this method works by aggregating together small groups of similar values and replacing all such values by the average of the aggregate group.
- **Random perturbation** : Assuming that a given attribute can be mapped to a certain multi-dimensional space, this method consists of adding a small-variance noise to each entry before mapping it back to the initial value space.

2.2.2 Permutation

Permutation can be seen as a sub-technique of perturbation. It is especially useful when the values of a given attribute cannot easily map to a numerical

multi-dimensional space. In such an instance, an efficient way of adding perturbation to the data is to permute the personal attribute and the value of a similar record. This allows for the same protection of the statistical significance of the data while increasing the anonymity of each record.

2.2.3 Generalisation

As presented by Martínez et al. (2012), generalisation is an anonymisation method that relies on the existence for a given attribute of an underlying hierarchy, for example

$$address \rightarrow street \rightarrow district \rightarrow city \rightarrow region \rightarrow country.$$

Generalisation then uses such a hierarchy to reduce the specificity of the record and therefore the amount of information that such a record divulges *without making the record inaccurate or untruthful*. Depending on the granularity of the hierarchy, such a method can partially preserve both the statistical utility of the data as well as its accuracy.

2.2.4 Suppression

As its name suggests, suppression is an anonymisation method that consists of ensuring anonymity by deleting personal information from the record. This can be seen as a last resort where neither generalisation nor perturbation technique can be applied to partially preserve the usefulness of the information while protecting the personal information.

3 Anonymity

One of the main aspects of the anonymisation process is to be able to reliably measure the level of anonymity of a given database. Such a measure needs to be normalised, interpretable and realistically implementable. Several solutions have been presented in the literature.

3.1 k-anonymity

Description

As presentend by El Emam and Dankar (2008), the k-anonymity metric aims to identify the highest integer k for which the data verify the following statement :

When it comes to the quasi-identifiers of interest, no record can be distinguished from $k-1$ other records.

In essence that means that the records can gathered together by the values of their quasi-identifiers in groups of size at least k. While this insures that no single person can be identified, this doesn't ensure that sensitive information cannot still be deduced from the anonymised table. For example, all the records within the same *bin* could have the same value for one of the sensitive attributes, giving us information about that attribute for all bin members.

Pseudocode

Algorithm 1 K-anonymity metric

```
1: procedure COMPUTE K-ANONYMITY
2:     groups = {}
3:     for row in data do
4:         if row[qids] ∈ groups then
5:             groups[row[qids]] ← groups[row[qids]] + 1.
6:         else
7:             groups[row[qids]] ← 1.
8:     return min(groups)
```

3.2 l-diversity

Description

l-diversity (Machanavajjhala et al. (2006)) builds on the previously described k-anonymity metric in order to ensure a certain degree of diversity within a certain bin. Several **diversity** measures can be used, including:

- Distinct l-diversity: The number of **unique values** for a given sensitive attributes in each bin. *[the simplest definition]*
- Entropy l-diversity: Measure of the entropy within a given "bin" or "equivalence class" the l-diversity of a given bin is then defined as the exponential of the entropy of the bin. This methods is considered to be the most complex and aims to make sure that each bin is sufficiently disordered.

$$Entropy(b) = \sum_s p(b,s) log(p(b,s))$$
$$l(b) = floor(exp^{Entropy(b)})$$

- Recursive (c-l)-diversity: This metrics is an in-between the two previous ones. Its definition relies on choosing a given value c. a given bin is said to have (c-l)-diversity if the frequency of the most popular value of the bin $f(r_1)$ if it is c times less frequent as the sum of the remaining bin and if the remaining bin [without the most popular value] verifies $(c-(l-1))$-diversity.

$$f(r_1) < c \sum_{i>1} f(r_i)$$

This metric ensures stronger anonymity since it protects against the weakness of k-anonymity; however, this comes at the expense of a greater loss of utility. Furthermore, while l-diversity guarantees that no value is too overly represented within a bin, it doesn't protect against statistical shifts when compared to the overall distribution within the database.

Pseudocode

Algorithm 2 L-diversity metric

```
1: procedure COMPUTE L-DIVERSITY
2:     groups = {}
3:     for row in data do:
4:         groups[row[qids]].append(row).
5:     l_div = inf
6:     for bin in groups do:
7:         l_score = get_l(bin)
8:         l_div = min(l_div, l_score)
9:     return l_div
```

3.3 t-closeness

Description

The t-closeness model extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute. To be more specific, t-closeness relies on the measure of the distance between a global statistical distribution. As such a table is said to have t-closeness if any bin's distance to the overall distribution is no more than t. Several distance measures exist; the following, which we have implemented so far, rely on the earth mover's distance or **EMD**.

- EMD for Numerical Attributes : Let p_i and q_i be the probabilities of the ith smallest unique value, p being the overall distribution and q the distrbution within a specific bin. Then the distance D is computed as

$$D(P,Q) = \frac{1}{m-1} \sum_{i=1}^{i=m} |\sum_{j=1}^{j=i} r_j|$$

  where

$$r_i = p_i - q_i.$$

- EMD for categorical attributes: For categorical information or more generally information that doesn't hold any order to it. two methods

exists but only one has been implemented so far *[equal distance]*. Using the same notations,

$$D[P, Q] = \sum_{i=1}^{m} |p_i - q_i|.$$

Out of all the presented metrics, the t-closeness is the most conservative and therefore leads to to the greatest loss in utility. Furthermore, this metric is also more expensive to compute.

Pseudocode

Algorithm 3 t-closeness metric

```
1:  procedure COMPUTE T-CLOSENESS
2:      groups = {}
3:      p = {}
4:      q = {}
5:      for row in data do:
6:          groups[row[qids]].append(row).
7:          p[row[sensitive_column]] += 1
8:          q[row[qids]][row[sensitive_column]] += 1
9:      t_close = 0
10:     for bin in groups do:
11:         t_close = max(t_close, get_t(p, q[bin]))
12:     return t_close
```

4 Utility

No matter what anonymity metric is chosen, there will always be a certain loss of data utility. However, evaluating such a utility loss can end up being challenging as the term *utility*, depending closely on business context and what other information is in the public domain, can be unclear. As such, several utility metrics exist.

4.1 Classification metric

Description

The idea behind the classification metric (Iyengar (2002)) is to model the loss of information by the loss of accuracy when trying to predict a given sensitive attribute using the generalised quasi-identifiers. As such, the classification metric tries to predict the loss of performance that would occur in a classification task should this generalisation setting be selected.

$$y(g) = |g| - \sum_{r \in g} I(r, mode(g))$$

$$CM = \frac{\sum_{i=1}^{h} y(g_h)}{|T|}$$

This metric varies between 0 and 1, with 1 representing the highest information loss.

Pseudocode

Algorithm 4 classification metric

```
1: procedure COMPUTE CLASSIFICATION METRIC
2:     groups = {}
3:     for row in data do
4:         groups[row[qids]][row[sensitive_col]] += 1.
5:     CM = 0
6:     table_size = 0
7:     for bin in groups do
8:         bin_size = sum(bin)
9:         CM += bin_size − max(bin)
10:        table_size += bin_size
11:    CM = CM / table_size
12:    return CM
```

4.2 Multiple-use metric

Description

The most general case for anonymisation is that in which the data is being disseminated for multiple uses. For such a case Iyengar (2002) offers the following metric as it aims to give an overall idea of the loss of granularity in the data. Iyengar (2002) presents two version of this metric.

- Categorical data: Let M be the number of unique values for the sensitive attribute. and lets note M_b the number of unique values within a given bin b. then the information loss is computed as the **the loss of granularity**

$$l_b = \frac{M_b - 1}{M - 1}.$$

- Numerical data: Let U be the upper bound and L the lower bound of the sensitive attribute, while U_b shall be the upper bound and L_b the lower bound of a given bin b. The information loss is then computed as the **the loss of granularity**

$$l_b = \frac{U_b - L_b}{U - b}.$$

As for the overall information score, several possibilities are available, the simplest one being a weighted average where each bin is weighted by its size.

Pseudocode

Algorithm 5 multiple-use metric

procedure MULTIPLE-USE METRIC
    if $mode \in numerical$ **then**
        $min_g = \{\}$
        $max_g = \{\}$
        $min_t = -inf$
        $max_t = inf$
    else
        $values_g = \{\}$
        $values_t = []$
    $size_g =$
    $size_t = 0$
    for row **in** data **do**
        $r\_sens = row[sensitive\_col]$
        $bin = row[qids]$
        if $mode = numerical$ **then**
            $min_g[bin] = min(min_g[bin], r\_sens)$
            $min_t = min(min_t, r\_sens)$.
            $max_g[bin] = max(max_g[bin], r\_sens)$
            $max_t = max(max_t, r\_sens)$.
        else
            if $r\_sens \notin values_t$ **then**
                $values_t.append(r\_sens)$
            if $r\_sens \notin values_g[bin]$ **then**
                $values_g[bin].append(r\_sens)$
        $size_t \mathrel{+}= 1$
        $size_g[bin] \mathrel{+}= 1$
    $avg_{score} = 0$
    for bin **in** groups **do**
        if $mode = numerical$ **then**
            $avg_{score} \mathrel{+}= \frac{size_g[bin]}{size_t} \frac{max_g[bin] - min_g[bin]}{max_t - min_t}$
        else
            $avg_{score} \mathrel{+}= \frac{size_g[bin]}{size_t} \frac{|values_g[bin]| - 1}{|values_t| - 1}$
    return avg_{score}

4.3 The monotonous entropy information loss metric

Description

The Shannon entropy is one of the oldest and most robust measures of information content. It can be employed to assess the amount of information that is lost by anonymising a table. The usual entropy metric is not necessarily monotonic, as presented by Gionis and Tassa (2009); in other words, the information loss can decrease when generalising. Instead, we can rely on the **monotonous entropy** measure, which ensures that information loss increases when anonymity increases.

Given a table D and generalisation $g(D)$, the *monotonous entropy measure* π_{me} can be computed as

$$\pi_{me}(D, g(D)) = \sum_{i=1}^{n} \sum_{j=1}^{r} Pr(R_i(j)) H(X_j | \bar{R}_i(j))$$
$$\text{where } H(X_j|B_j) = -\sum_{b \in B_j} Pr(X_j = b | X_j \in B_j) log_2 Pr(X_j = b | X_j \in B_j).$$

It is important to note that, unlike with the natural definition of entropy, having **a higher entropy implies a higher loss of information** instead of a greater diversity of information. This is mainly due to the fact that entropy measures the weighted average entropy within equivalence classes; such a measure will naturally increase when the equivalence classes increase in size, as each equivalence class becomes more disorderly. As such, the information loss measure is minimal when no generalisation has been applied and is maximal when all quasi-identifiers have been merged into a single equivalence class.

4.4 Mutual information utility measure

Description

As opposed to the entropy measure presented above, the mutual information metric is a utility measure (Goldberger and Tassa (2010)) and therefore decreases when the anonymisation increases. The goal is thus to maximise it. The mutual information between two random variables is a measure of the information that is disclosed about one of those variables by providing the value of the other.

$$U(g(D)) = I(D, g(D))$$
$$I(D, g(D)) = \frac{1}{r} \sum_{j=1}^{r} I(X_j; \hat{X}_j)$$
$$I(X_j; \hat{X}_j) = \frac{1}{n} \sum_{i=1}^{n} log \frac{Pr(X_j = R_i(j) | X_j \in \bar{R_i}(j))}{Pr(X_j = R_i(j))}$$

As pointed out by Goldberger and Tassa (2010), this metric can also be re-framed as a **mutual information loss metric**, leading to the expression

$$\Pi_{MI}(D, g(D)) = -\frac{1}{nr} \sum_{i=1}^{n} \sum_{j=1}^{r} log Pr(X_j = R_i(j) | X_j \in \bar{R}_i(j))$$

.

4.5 Private mutual information utility measure

Description

One might notice that aside from the classification metric, most utility or information loss metrics presented above focus mostly on the **quasi-identifiers** while ignoring the sensitive information that should be protected. Goldberger and Tassa (2010) offers an alternative that builds on the previous mutual information utility measure. Such a metric, called the **private mutual information utility measure**, aims at making sure that the generalisation of public attributes that are weakly correlated with the private data should be less penalised than generalisation of other public attributes that are strongly correlated with the private data. Again, this metric can either be defined as a utility metric U_PMI that decreases when generalisation increases, i.e.

$$\mathbf{U_{PMI}(g(D)) = I(D, g(D))}$$
$$\text{with } I(D, g(D)) = \frac{1}{nr} \sum_{j=1}^{r} \sum_{i=1}^{n} log \frac{Pr(Y=S_i|X_j \in \bar{R}_i(j))}{Pr(Y=S_i)}$$

or as an information loss measure Π_{PMI} that increases when generalisation increases:

$$\mathbf{\Pi_{MI}(D, g(D)) = -\frac{1}{nr} \sum_{i=1}^{n} \sum_{j=1}^{r} logPr(Y = S_i|X_j \in \bar{R}_i(j))}$$

5 Disclosure risk

While anonymisation metrics provide information about the level of anonymity inherent in the database, they don't provide any explicit information about the probability of reidentification (Domingo-Ferrer and Rebollo-Monedero (2009)). Disclosure risk is a concept that aims to compute the reindentification probability based on plausible scenarios (El Emam et al. (2009)); for example, by assuming a certain access to the database, as well as a certain behaviour from the "intruder".
In the literature, three behaviours are presented: prosecutor, journalist, and marketer. While more details are presented bellow, it is interesting to keep in mind that these three behaviours follow the following rule:

- prosecutor risk will be equal to or larger than journalist risk;
- journalist risk will be equal to or larger than marketer risk;
- prosecutor risk will be equal to or larger than marketer risk.

Each one of these behaviours implies both a certain access to the database as well as an interest for a specific type of re-identification probability.

5.1 Notations

- U the subset of records for which the anonymised quasi-identifiers as well as the sensitive information has been released
- D the identifying database with the assumption that $U \subseteq D$.
- $|U| = n$ and $|D| = N$.
- $Z = \{z_1, ..., z_p\}$ the set of available quasi-identifiers in the database.
- $|z_i|$ be the number of unique values for this specific attribute.
- $\tilde{J}$ the number of equivalence classes available in D
- $j \in \{1, .., \tilde{J}\}$ a given equivalence class among all possible equivalence class available in U.
- $F_j = \sum_{i \in D} I(X_i = j)$ The frequencies in D for different values of $\tilde{J}$.
- $f_j = \sum_{i \in U} I(X_i = j)$ The frequencies in U for different values of $\tilde{J}$.
- g_j is an equivalence class for U
- G_j is an equivalence class for D

5.2 Marketer risk

As presented by Dankar and El Emam (2010), the marketer risk doesn't focus on the re-identification of a specific individual. Instead, it focuses on the probability of any random disclosed record being re-identified. This justifies the "marketer" name attributed to such a scenario - assuming that the intruder wants to use the information for marketing purposes, there is no need to know exactly who is appropriately identified and who isn't. Instead the intruder simply aims to evaluate how likely it is that the right marketing material is sent to the appropriate individual.

Based on the notations presented bellow, the *marketer risk*, or the probability that a given record is appropriately re-identified, is

$$R_m = \frac{1}{n} \sum_{j \in \tilde{j}} \frac{f_j}{F_j}.$$

In both of these cases the risk measure captures the worst case probability when reidentifying a single record. For marketer risk we are evaluating the expected number (proportion) of records that would be correctly re-identified. Another important difference is that marketer risk does not help identify which records in U are likely to be re-identified. However, with journalist and prosecutor risk measures, it is possible to identify the highest risk records and focus disclosure control action only on those.

5.3 Prosecutor risk

In the prosecutor scenario, two main concept differ from the marketer scenario. Firstly, the *intruder* is assumed to have access to an identifying database D that contains all and only the records from the disclosed anonymised database U (i.e. $D = U$). In this scenario, an intruder wants to re-identify a specific person in a de-identified database. As such, for the least protected equivalence class. the risk is

$$R_p = \frac{1}{min_j(f_j)} = \frac{1}{min_j(F_j)}.$$

This is noticeably different from the marketer approach as it also allows uncovering which set of records are the most vulnerable to de-identification.

5.4 Journalist risk

Journalist risk is also concerned with the re-identification of individuals. However, in this case, the journalist does not care which individual is re-identified. This is due to the assumption that $U \subseteq D$; therefore, it is not certain that any given individual within the identifying database is one of the records in the disclosed database.

$$R_j = \frac{1}{min_{j \in \tilde{J}}(F_j)}$$

In essence the computation of the risk R_j is sensibly similar to the prosecutor case; however the fact that U is only a subset of D has a significant impact as all equivalence classes within D might not be represented in U and will therefore be discarded during the computation.

References

Brown, N. (2015). Healthcare Data Growth: An Exponential Problem.

Dankar, F. K. and K. El Emam (2010). A method for evaluating marketer re-identification risk. In *Proceedings of the 1st International Workshop on Data Semantics - DataSem '10*, pp. 1.

Domingo-Ferrer, J. and D. Rebollo-Monedero (2009). Measuring risk and utility of anonymized data using information theory. *Proceedings of the 2009 EDBT/ICDT Workshops on - EDBT/ICDT '09*, 126.

El Emam, K. and F. K. Dankar (2008). Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association 15*(5), 627–637.

El Emam, K., F. K. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *Canadian Journal of Hospital Pharmacy 62*(4), 307–319.

Gionis, A. and T. Tassa (2009). K-anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering 21*(2), 206–219.

Goldberger, J. and T. Tassa (2010). Efficient anonymizations with enhanced utility. *Transactions on Data Privacy 3*(2), 149–175.

Hsu, D. (2015). Techniques to Anonymize Human Data.

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, pp. 279.

Machanavajjhala, A., J. Gehrke, D. Kifer, and M. Venkitasubramaniam (2006). ℓ-Diversity: Privacy beyond k-anonymity. In *Proceedings - International Conference on Data Engineering*, Volume 2006, pp. 24.

Martínez, S., D. Sánchez, A. Valls, and M. Batet (2012). Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion 13*(4), 304–314.

Newman, D. (2017). Data As A Service: The Big Opportunity For Business.

OECD (2001). Glossary of Statistical Terms.

Zhou, B., J. Pei, and W. Luk (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter 10*(2), 12.