A Case Study of Anonymization of Medical Surveys

Michele Gentili
Eurecat - Technology Centre of Catalonia
Barcelona, Spain 08018
michele.gentili93@gmail.com

Sara Hajian
Eurecat - Technology Centre of Catalonia
Barcelona, Spain 08018
sara.hajian@eurecat.org

Carlos Castillo
Eurecat - Technology Centre of Catalonia
Spain 08018
chato@chato.cl

ABSTRACT

Health data anonymization is a hot topic, on which both the medical and the computer science communities have made a great effort to provide a safer and trustful way of sharing data among research centers and hospitals. The main challenge in data anonymization is to provide a proper trade off between the utility of the resulting data/models and protecting individual privacy. In this paper we present a real anonymization case, with particular emphasis on choices that have to be made to carry it on, and difficulties experienced using a data set with many dimensions, and not well distinguishable features. We present our approach for evaluating disclosure risks and methods for anonymising high-dimensional medical survey data and measuring the utility of the transformed data.

CCS CONCEPTS

•Security and privacy →Privacy protections; Usability in security and privacy;

KEYWORDS

Health privacy, Data anonymization

1 INTRODUCTION

People have concern about disclosing their personal information, or having their personal information processed for secondary purposes. For example, individuals often cite privacy and confidentiality concerns and lack of trust in researchers as reasons for not having their health information used for research purposes [4]. One study found that the greatest predictor of patients' willingness to share information with researchers was the level of trust they placed in the researchers themselves. A number of studies have shown that attitudes toward privacy and confidentiality of the census are predictive of people's participation [7]. These trust effects are amplified when the information collected is of a sensitive nature. On the other hand, organizations, such as hospitals, need to release micro data[1] for research and other public benefit purposes. However, sensitive personal information (e.g., disease of a specific person) may be revealed in this process. Conventionally, a naïve data anonymization is carried out by column removal of the identifying attributes, assuming a simple privacy attack model in which a single attribute is involved (e.g. DNI, Person Name). However attacks can be more complex and involve multiple attributes, due to the existence of quasi-identifiers in the released micro-data [3]. Quasi-identifiers are sets of attributes (e.g., zipcode, gender, date of birth), which can be joined to information obtained from diverse sources (e.g., public voting registration data, social networks) in order to reveal the identity of individual records. To address this threat, many algorithms have been proposed such as the well known k-anonymity model [9], i.e. or every record in a released table there should be at least k other records with the identical values in the quasi-identifier attributes, and l-diversity [6] to protect data when there are many identical sensitive values in the same set of quasi-identifiers attributes. Records with identical quasi-identifier values constitute an equivalence class. K-anonymity is commonly achieved by generalization (e.g., by showing only the area code instead of the exact phone number) or suppression (i.e., hide some values of the quasi-identifier), which inadvertently lead to information loss. l-diversity prevents uniformity and background knowledge attacks by ensuring that at least l sensitive attributes values are well-represented in each equivalence class (e.g., the probability to associate a tuple with a sensitive value is bounded by $1/l$). Still, the data should remain as informative as possible, in order to be useful in practice. Hence a trade-off between privacy and information loss emerges.

Table 1: Anonymized data - output example

Age	GENDER	ETHNIC	STUDIES
*	*	*	*
42	1	1	7
40	1	1	7
45	1	1	7
[48, 51[	2	1	<=3
[48, 51[	1	1	<=3
*	*	*	*
[60, 63[	2	6	<=6
58	1	1	5
*	*	*	*

European Regulation. To justify our work proposal and our general principles are based on the latest European regulations[2]. It

[1] Specific data per each individual, not statistics or summaries

[2] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016

DH '17, July 2–5,2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07…$15.00.
DOI: http://dx.doi.org/10.1145/3079452.3079490

Table 2: Feature Categories

Category	Attributes
Demographic	BIRTH_DAY , SEX , ETHNIC , STUDIES , ...
Medical	SPORT_FREQUENCY , WEIGHT , HEIGHT , BMI ...
Diseases	CANCER , DIABETES , ECZEMA , LUPUS ,...

worth noting that the disclosure risk is not related only to mathematical scores or uniqueness percentages, but it has to be assessed taking into account the actual means and technologies reasonably likely to be available, costs and amount of time required for identification and technological developments.

Scenario. The hospital needs to share data among trusted research centers. In this scenario, we don't have to face with the publication of given data, where the perfect privacy would requires not to release any sensitive information. We seek to increase the cost of re-identification process, dealing with the trade-off between data privacy and cost of making the attack unfeasible.

Our objective is to present a real case data anonymization, presenting all the steps and choices that the complete process has to undergo, ensuring that the dataset is ready for future proper analysis with some privacy guarantee. Example of output results is presented in Table 1.

The rest of this paper is organized as follows: Section 2 presents the brief analysis of the data and the data anonymization tool, then we present the results of privacy risk evaluation in Section 3. Our data anonymization process is introduced in Section 4, followed by a brief analysis of data utility in Section 5 and finally we present our recommendations for anonymization of medical datasets in Section 6.

2 DATA AND TOOLS

The data that we used in this study comes from a survey made by a Spanish hospital. The dataset has 11,000 records and 121 attributes. As standard procedures require, name and surname of patients have been replaced by a unique ID. Moreover, we preprocessed and cleaned the data and mapped the categorical values to a numerical representation. Either categorical or numerical attributes present the demographic, medical and disease information about each patient. Table 2 shows some examples of these attributes in each category.

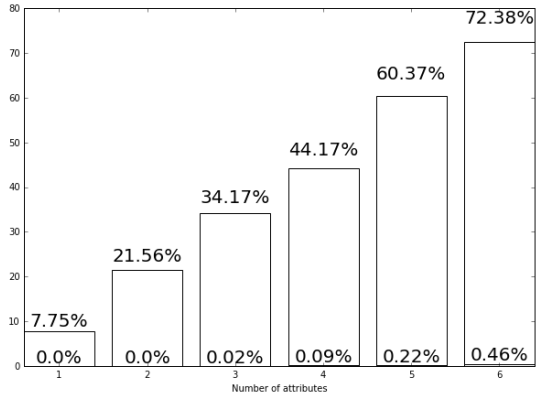
Table 5 and 6 in the appendix present the summaries of the dataset. Table 5 displays the number of unique values, i.e., the domain of each attribute, the most frequent value, its frequency and finally the fraction of missing values. Table 6 presents common statistics and percentiles of numerical attributes in the dataset. In order to enhance an intuitive data visualization, and since the diseases (cancer, diabetes, etc...) present almost the same statistical characteristics, in the table they have been summarized into just one record showing average values among them. Almost all of the diseases have as the most frequent (96%-99% ca) value the absence of the disease (i.e. 0), and they present only a few missing values. It is worth noting that on average each row of the tables present less than 1% of missing values, except for some of them that reach the 7-8%. As it will be discuss later, it is an important issue for current anonymization algorithm and software that treat Null values as an extra value of the domain of the attribute.

Tool. We used the ARX[8], open-source software developed at the Technical University of Munich to compute the pitman risk and anonymize data at 3-anonymity level. For data preprocessing, data analysis and visualization we wrote our own code using a variety of basic and open-source, statistical and graphical Python's packages.

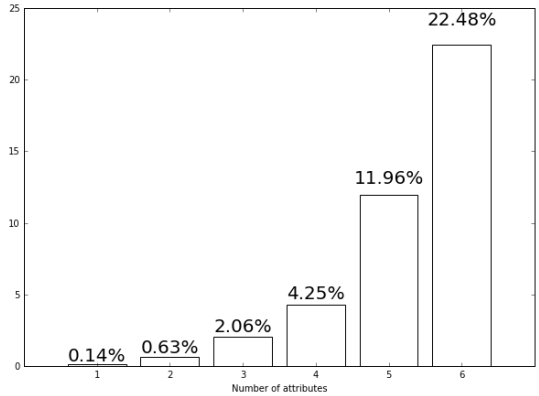
3 DATA PRIVACY RISK

We used different techniques and measures for evaluating the disclosure risks of the medical survey dataset described in Section 2 . First, we looked at the minimum and the maximum value of the percentage of unique values in the data given all the possible combination of a fixed number of attributes. So for t in [1,...,no. of attributes] at each step we find a combination of t-attributes that have the highest percentage of unique values. To carry on our analysis, clearly, we have replaced the date of birth information with the age, since the day and the month are not relevant for investigation made on this data.

The bar plot in Figure 1a shows values related to the whole dataset, as we increase the number of features we are considering in each combination, the percentage of unique records increases



(a) Complete Data set



(b) Grouping age, residential info removed

Figure 1: Attribute uniqueness

rapidly, as a matter of fact taking into account only 6 attributes almost all the records are unique. Lastly the value at the bottom of the bars represent the minimum uniqueness percentage that we can reach always considering 1,2 ... 6 attributes. Looking at this first graph, we can observe that the survey data has a high disclosure risk (i.e., 72% of patients are uniquely identifiable with only 6 attributes). Municipality and Country were the attributes that contribute the most in increasing the disclosure risk. However, not all the attributes add the same amount of risk, therefore we can find combination of 6 attributes where only 0.46% of the records are unique.

Naïve Anonymization. Consequently, the second bar plot is made by removing attributes of geographical data, Municipality, Country..., and bucketing the age within a set of 3 years, due to its high utility a higher generalization cannot be accepted. It is worth noting that with this easy procedure we are drastically reducing the frequency of unique records. So the first conclusion is that data presents a high level of disclosure risk, however it can be addressed, after having deeply understood the role of each attribute in both data utility and also disclosure risk during the anonymization process.

Super Population Model. A more formal approach is to assess the disclosure risk is through the usage of a super population model. It is based on the assumption that an attacker doesn't know whether a person has attended the survey or not. Indeed, even if a record is unique in the questionnaire, it may not be unique in the database used to disclose the data, i.e. any public survey on the Spanish population. So that even if a unique record can be found on the released data, it will not lead to a unique record in the data set with the personal identifiers, thus the unique record still remains anonymous. Therefore, the true risk has to be based on the dataset with information on all the possible people who may have filled out the survey, however, this super population database is unknown. According to [2], we decided to use the *Pitman risk estimator* to have a concrete idea of how dangerous a possible release could be. Within our attributes selection, the Pitman score is: 0.4574, this means that even if we take a bigger sample of the super population, that is just increment the size of the survey, still we will have a lower bound of unique records that is 45.75% of the total population, so almost half of the entire population is unique through these attributes. Under the k-anonymity it is clear that all the scores go to 0, because if a record is not unique in the given data it will remain within a class of at least the same size in the super population dataset.

4 DATA ANONYMIZATION

4.1 Attributes Selection

As mentioned before, a significant issue that arose in the first place is to define a discriminatory on the attributes typology, that is to decide whether they are identifiers, quasi-identifiers (QI), sensitives or not-relevant for the anonymization process. The trigger point is to define the background knowledge of the attacker, indeed not all attributes should be considered in a potential pattern for disclosure sensitive information. For instance, BMI[3] cannot be known by an attacker, thus it won't be considered in the privacy risk model, but it will be released as it is due to its high data utility. So, Identifiers and not relevant are the first to be treated. Any Personally identifiable information (PII), has to be removed and any attribute not relevant in the disclosure risk evaluation, neither known by an attacker nor a sensitive attribute, can be removed before the anonymization process and then added again to the dataset as it is. Selecting the attributes and their characterization in terms of quasi identifier, sensitive and not relevant for the anonymization , is actually, far from being trivial. Issues come from the fact that there's no objective nor absolute attribution of them and no unique background shared by attackers.Indeed considering all possible attack environments, each attribute can be considered known and so used for disclosing the data. To decide whether an attribute is QI or sensitive, three factors have been taken into account. The foremost is the level of knowledge required to know that kind of information, truly, the knowledge of the attacker over a feature, cannot be considered as a binary variable, but for instance it's possible to distinguish between a public knowledge (age) and the one that is known only by close persons(the sport frequency). The second is the level uniqueness values in that attribute and the last is the importance of feature for analysis purposes. So, due to the high uniqueness risk of the geographical attributes and their low impact in the research carried out on this data, we decided to work with a selection of the original set of attributes, removing the aforementioned and those feature that cannot be known by an adversary. It's worth noting that at this point we are assuming that all the remaining attributes are either sensitive or QI, so if a feature is not relevant for any of the 3 cases just mentioned, then it is considered to be sensitive. Three examples may clarify the process:

(1) *Age*: it's almost a public information, quite unique (25 classes) at most 6% of record per each class and deeply relevant for research purposes.
(2) *Income*: it can be found through *curriculum vitae* or other sources, boundary class with less then 10% of records, not so important for research purposes.
(3) *Disease*: Private info, although thanks to social networks it's not impossible, unbalanced data (generally only 4% of records present that illness, key feature for the analysis.

Thus *Age* and *Income* will be treated as QI, further more, given its importance, we have placed an upper bound to the size of the *Age* buckets of three years, whereas all the *Diseases* will be considered as Sensitive attributes. Still in the set up step, once defined how to use all the attributes, generalization tree have to be defined. A tree is made of leaf that represent features values and nodes that are generalization of their children up to the root, that is the missing value.

Figure 2 shows an example of generalization for the *Income*, a reader may not agree with the asymmetry structure, however it is important to keep in mind that our activity is compelled by data, thus structures are an attempt to reduce as much as possible the level of the generalization by creating homogeneous dense classes.

4.2 Global and Local Recoding

Due to the presence of a huge multidimensional dataset, and taking into account that the sharing of data is only among recommended

[3]body mass index

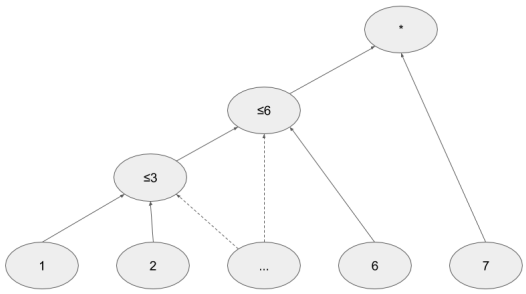


Figure 2: Generalization tree Income

institutes we only want to preserve a 3-Anonymity guarantee, noting that this is the choice of data owner, and for the moment we drop the *l*-diversity requirement. Since sensitive attributes are binary and a very unbalance feature, the presence of disease is rare, it will require an ad hoc solution. After having imposed the requirement, the software now looks for a suitable solution, trying to minimize the information loss. Instead of generalizing entire columns, it tends to remove outliers, because it is less 'expensive' in terms of data utility, Table 3 shows the first 10 lines of the global recoding, where values that are too unique have just been set as unknown.

After having brute force anonymized the data, the algorithm tries to insert again those values up to some generalization level that still preserve the k-anonymity imposed at the beginning. A global minimum is not guaranteed, however we go from a 40% of missing values to 20%, as can be seen in Table 4. So for instance, row 5 can be returned moving just to a higher level *age,studies* and *income*.

5 DATA UTILITY

A common measure of the information loss is to measure the distance between the anonymized data versus the original one[5]. All the attributes have been treated as equally important as a potentially identifying columns. So the total information loss has been given by the mean over the information loss of each columns, that is the sum of the distance for each entry between the two datasets. The distance is then given by the ratio of the new level of generalization or the extent of the new interval and the old one. For further detail refer to [5]. In our case, we had a loss of: 6.1%, thus on average we have lost 6.1% of the variance present in the data for achieving 3-anonymity.

6 DISCUSSION

Logic Rule. To decide whether an attribute is QI or sensitive, a logic rule can be defined. Given some parameters on each attribute: level of knowledge required to be used by an attacker, the sensitive and utility measures; the output is the category into which they belong. Further more, for the cases that are more tricky, because they are both necessary for the analysis and well known by a possible attacker (for instance the age), the choice can be offered to hospital experts to define by themselves how to balance the trade off between security and utility, introducing for instance constraint on the maximum generalization level possible to reach.

Null values. *Null values* are an important issue for current anonymization algorithms and software that treat them as an other possible value for the attribute. So, records with *Missing value* must belong to a class that is k-anonymous and *l*-diverse, further more taking into account that the percentage of *Missing value* is very low, it becomes one of the most important factors that bring it to a very high generalization of values. Investigate missing values through an extension of generalization algorithms and show that NULL aware generalization algorithms lead to a decrease in information loss than standard algorithms [1].

Data utility measurement. It is a key issue in the anonymization process, in particular not only as a cost function for the optimization process, but mainly to provide a proper justification to the owner of the data, to encourage the anonymization of the data. As we showed before, we have used a simple distance measurement to have a clue of the loss of information, however in the next studies we will define other distances, that won't be based on the input/output data but on the results obtained by standard statistical and machine learning algorithms.

ACKNOWLEDGMENTS

This work was partially supported by the EU H2020 innovation action programme, TYPES project (grant No 653449) and Catalonia Trade and Investment Agency (ACCIÓ). The authors would like to thank Dr.Fabian Prasser, TUM University for providing a huge support in implementing all the process with the free software ARX.

REFERENCES

[1] Margareta Ciglic, Johann Eder, and Christian Koncilia. 2016. Anonymization of Data Sets with NULL Values. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV*. Springer, 193–220.
[2] Fida Kamal Dankar, Khaled El Emam, Angelica Neisa, and Tyson Roffey. 2012. Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making* 12, 1 (2012), 66.
[3] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. 2007. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 758–769.
[4] Lawrence O Gostin, Laura A Levit, Sharyl J Nass, et al. 2009. *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research.* National Academies Press.
[5] Vijay S Iyengar. 2002. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 279–288.
[6] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3.
[7] Thomas S Mayer. 2002. Privacy and confidentiality research and the US census bureau recommendations based on a review of the literature. *Survey methodology* (2002), 01.
[8] Fabian Prasser and Florian Kohlmayer. 2015. Putting statistical disclosure control into practice: The ARX data anonymization tool. In *Medical Data Privacy Handbook*. Springer, 111–148.
[9] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

7 APPENDICES

Table 3: Global Recoding outcome

	Age	SEX	ETHNIC	LATERALITY	MARRIED	DIVORCE	STUDIES	INCOME	WORK	ADOPTION	PREGNANCY
0	*	*	*	*	*	3	*	*	*	2	*
1	42	1	1	2	2	3	7	3	1	2	2
2	40	1	1	2	2	3	7	3	1	2	2
3	45	1	1	2	2	3	7	4	1	2	2
4	*	*	*	*	*	3	*	*	*	2	*
5	*	*	*	*	*	1	*	*	*	2	*
6	*	*	*	*	*	1	*	*	*	2	*
7	*	*	*	*	*	1	*	*	*	2	*
8	58	1	1	2	2	3	5	4	1	2	2
9	*	*	*	*	*	*	*	*	*	2	*

Table 4: Local Recoding outcome

	Age	SEX	ETHNIC	LATERALITY	MARRIED	DIVORCE	STUDIES	INCOME	WORK	ADOPTION	PREGNANCY
0	*	*	*	*	*	3	*	*	*	2	*
1	42	1	1	2	2	3	7	3	1	2	2
2	40	1	1	2	2	3	7	3	1	2	2
3	45	1	1	2	2	3	7	4	1	2	2
4	[48, 51[	2	1	2	2	3	<=3	<=3	3	2	2
5	[48, 51[	1	1	2	4	1	<=3	<=3	1	2	2
6	*	*	*	*	*	1	*	*	*	2	*
7	[60, 63[	2	6	2	4	1	<=6	<=3	1	2	2
8	58	1	1	2	2	3	5	4	1	2	2
9	*	*	*	*	*	*	*	*	*	2	*

Table 5: Categorical Attributes

	Unique	Most Frequent	Frequency	Missing
BIRTH_DAY	6631	*	<0.1%	0.4%
COUNTRY_BIRTH	62	SPAIN	96.4%	0.2%
MUNICIPAL_BIRTH	1371	BARCELONA	42.3%	4.9%
SEX	2	2	59.1%	0.5%
ETHNIC	7	1	83.1%	0.6%
LATERALITY (right-left hand)	3	2	89.5%	<0.1%
MARRIED (single,widow ...)	6	2	65.9%	<0.1%
STUDIES(Primary, middle school ...)	8	7	35.7%	<0.1%
INCOME (<18k, <31k ...)	6	2	29.5%	<0.1%
WORK (yes,no,retired ...)	9	1	70.9%	<0.1%
SPORT_WALKING (yes, no, can't)	3	1	94.0%	<0.1%
SPORT_FREQUENCY(yes,no)	2	2	51.9%	<0.1%
WEIGHT_VARIACION_YEAR	3	1	53.7%	<0.1%
WEIGHT_AT_BIRTH(<2,5kg,<3kg...)	6	3	31.6%	<0.1%
ADOPTION	3	2	99.6%	<0.1%
PREGNANCY_MULTIPLE	3	2	97.4%	<0.1%
DISEASE (CANCER,DIABETES...)	2	0	96.8%	<0.1%
NO_DISEASE(healthy)	2	0	61.1%	<0.1%

Table 6: Numerical Attributes

	Mean	Std	Min	25%	50%	75%	Max	Missing
AGE	51.03	7.59	40	45.00	51.00	57.00	65.00	0.4%
AVG_WEIGHT	75.36	15.75	42	64.00	74.00	85.00	146.00	0.5%
AVG_HEIGHT	165.38	14.44	141.6	159.00	165.20	172.50	200.00	0.5%
AVG_PULSE_RATE	73.87	12.22	30	67.00	73.00	81.00	126.00	0.5%
MAX_WEIGHT	77.85	17.75	42	65.00	76.00	88.00	190.00	0.8%
MIN_WEIGHT	68.83	16.42	37	59.00	68.00	79.00	165.00	1.8%
FATHER_AGE	72.26	16.60	1	68.00	76.00	82.00	103.00	2.4%
BMI	27.24	4.94	16.93	24.03	26.61	29.83	50.47	0.5%