UNIVERSIDADE
BEIRA INTERIOR

Anonimização de Dados em Educação

Wilson Gama dos Santos

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutora Maria Paula Prata de Sousa
Coorientador: Prof. Doutora Maria Eugénia Ferrão

Dezembro de 2020

Dissertação elaborada no Instituto de Telecomunicações - Delegação da Covilhã e no Departamento de Informática da Universidade da Beira Interior (UBI) por Wilson Gama dos Santos, Licenciado em Engenharia Informática pela Universidade da Beira Interior, sob orientação da Doutora Maria Paula Prata de Sousa, Investigadora do Instituto de Telecomunicações e Professora Auxiliar do Departamento de Informática da UBI, e coorientação da Doutora Maria Eugénia Ferrão, Investigadora do Centro de Matemática Aplicada à Previsão e Decisão Económica e Professora Auxiliar com Agregação do Departamento de Matemática da UBI, e submetida à UBI para discussão em provas públicas.
Trabalho integrado em projeto parcialmente financiado por: - FCT/MCTES através de fundos nacionais e quando aplicável cofinanciado por fundos comunitários no âmbito do projeto UIDB/50008/2020 e CEMAPRE/REM - UIDB/05069/2020.

Acknowledgments

At the end of an important phase of my life, I would like to express my deepest gratitude for everyone that made this possible.

First, and foremost, I would like to thank professor Maria Paula Prata de Sousa and Maria Eugénia Ferrão for their support and guidance. All the precise feedback that was provided and all the time spent to help me through this work was essential to conclude this dissertation.

A special thank to my parents and brother for their unconditional support and always being present when I needed, and for that, I am very grateful.

Last but not least, I would like to thank my professors, colleagues and friends that helped me through this path. Thank you all.

Resumo

Hoje em dia é possível observar que tanto a preocupação com a privacidade dos dados pessoais como a quantidade de dados recolhidos estão a aumentar. Estes dados, recolhidos e armazenados eletronicamente, contêm informação relacionada com todos os aspetos das nossas vidas, informação essa muitas vezes sensível, tal como registos financeiros, atividade em redes sociais, rastreamento de dispositivos móveis e até registos médicos. Consequentemente, torna-se vital assegurar a proteção destes dados para que, mesmo se tornados públicos, não causem danos pessoais aos indivíduos envolvidos. Para isso, é necessário evitar que registos nos dados sejam associados a indivíduos reais. Apesar de atributos, como o género e a idade, singularmente não conseguirem identificar o individuo correspondente, a sua combinação com outros conjuntos de dados, pode levar à existência de um registo único no conjunto de dados e consequente associação a um individuo. Com a anonimização dos dados, é possível assegurar, com variados graus de proteção, que essa associação a um individuo real seja evitada ao máximo. Contudo, este processo pode ter como consequência uma diminuição na utilidade dos dados. Com este trabalho, exploramos a terminologia e algumas das técnicas que podem ser utilizadas no processo de anonimização de dados. Mostramos também os efeitos dessas várias técnicas tanto na perda de informação e utilidade dos dados, como no risco de re-identificação associado, quando aplicadas a um conjunto de dados com informação pessoal recolhida a alunos que conluíram o ensino superior. No final, e uma vez feita a apresentação dos resultados, é feita uma análise e discussão comparativa dos resultados obtidos.

Palavras-chave

Anonimização de dados, k-anonimato, ℓ-diversidade, t-proximidade.

Resumo alargado

No mundo digital toda a atividade humana deixa um rasto de dados que constitui um recurso cada vez mais valioso para avaliação e definição de estratégias nos mais variados domínios. A partilha desses dados, sendo socialmente importante, implica o respeito pela privacidade individual e, portanto, a sua anonimização. As atuais leis e regulamentos sobre privacidade oferecem orientações limitadas para lidar com um vasto leque de tipos de dados, ou com técnicas de re-identificação.

Esta dissertação tem como principal objetivo explorar e aplicar técnicas utilizadas na anonimização de dados. De forma mais especifica, pretende-se: 1) explorar a terminologia utilizada na área da anonimização; 2) rever algumas das técnicas existentes atualmente e em que é que as suas diferenças contribuem para diferentes tipos de dados e proteção; 3) aplicar os modelos de privacidade k-anonymity, ℓ-diversity e t-closeness, a um conjunto de dados com informação pessoal de alunos que frequentaram o ensino superior; 4) apresentação dos resultados da perda de informação, risco de re-identificação e utilidade dos dados, obtidos para cada variante dos modelos de privacidade; 5) discussão dos resultados e comparação entre os modelos de privacidade. De forma a alcançar estes objetivos, esta dissertação apresenta várias contribuições descritas ao longo de 6 capítulos.

O primeiro capítulo contém o âmbito onde esta dissertação se enquadra. São também descritos os objetivos, bem como as principais contribuições deste trabalho para a área de anonimização de dados. No final do capítulo é então definida a estrutura deste documento.

O segundo capítulo apresenta um enquadramento legal segundo o Regulamento Geral sobre a Proteção de Dados da União Europeia, seguido de uma revisão da terminologia utilizada na área de anonimização de dados. Para além disso, são ainda apresentados alguns dos modelos de privacidade existentes, focando com mais detalhe o k-anonymity, o ℓ-diversity e o t-closeness, uma vez que serão utilizados no estudo prático presente nesta dissertação. São ainda apresentados os modelos de utilidade, nomeadamente o ANOVA que será utilizado na componente prática deste trabalho.

O terceiro capítulo apresenta uma revisão da literatura, começando por analisar trabalhos que realcem a dificuldade em determinar se um conjunto de dados está verdadeiramente anonimizado. Para além disso, são apresentadas várias técnicas atuais de anonimização, baseadas no k-anonimity, mas que o refinam seja para melhorar a utilidade, ou para o otimizar para casos de uso específicos. São também apresentados alguns trabalhos que descrevem técnicas de anonimização baseadas na introdução de aleatoriedade, machine learning e quando aplicadas a data mining. São ainda analisados trabalhos que avaliam o desempenho dos vários algoritmos de anonimização. No final do capítulo, são apresentadas algumas das ferramentas existentes para aplicar o processo de anonimização, sendo que, com mais detalhe se descreve a ARX Data Anonymization Tool.

Nos dois capítulos seguintes são aplicados os modelos de k-anonymity, ℓ-diversity e t-closeness, a um conjunto de dados com informação pessoal de alunos que frequentaram o ensino superior seguido da apresentação dos resultados da perda de informação, risco de re-identificação e utilidade dos dados, obtidos para cada variante dos modelos de privacidade e posterior discussão dos resultados e comparação entre os modelos de privacidade. Por último, é apresentada a conclusão onde se revelam os principais resultados e no final faz-se também uma perspetiva de possível trabalho futuro.

Abstract

Interest in data privacy is not only growing, but the quantity of data collected is also increasing. This data, which is collected and stored electronically, contains information related with all aspects of our lives, frequently containing sensitive information, such as financial records, activity in social networks, location traces collected by our mobile phones and even medical records. Consequently, it becomes paramount to assure the best protection for this data, so that no harm is done to individuals even if the data is to become publicly available. To achieve it, it is necessary to avoid the linkage between records in a dataset and a real world individual. Despite some attributes, such as gender and age, though alone they can not identify a corresponding individual, their combination with other datasets can lead to the existence of unique records in the dataset and a consequent linkage to a real world individual. Therefore, with data anonymization, it is possible to assure, with various degrees of protection, that said linkage is avoided the best we can. However, this process can have a decline in data utility as consequence. In this work, we explore the terminology and some of the techniques that can be used during the process of data anonymization. Moreover, we show the effects of said techniques on information loss, data utility and re-identification risk, when applied to a dataset with personal information collected from college graduated students. Finally, and once the results are presented, we perform an analysis and comparative discussion of the obtained results.

Keywords

Data anonymization, k-anonymity, ℓ-diversity, t-closeness.

Contents

List of Figures

List of Tables

Acronyms

GDPR	*European General Data Protection Regulation*
ENADE	*Exame Nacional de Desempenho dos Estudantes*
SINAES	*Sistema Nacional de Avaliação da Educação Superior*
OLA	*Algorithm and Optimal Lattice Anonymization*
API	*Application Programming Interface*
SQL	*Structured Query Language*

Chapter 1

Introduction

These days the amount of data collected about individuals is proceeding at an increasing rate. The amount of personal data that is collected electronically is related to all aspects of our lives, such data is sometimes sensitive personal information which includes financial records, activity in social networks, location traces collected by our mobile phones, network providers and even medical records. The analysis and study of this data from a general perspective, meaning not focused on as specific individual's data, allows socially and technologically important advances in many fields such as health care decision support, computational criminology, and terrorism informatics [8] [9]. However, this potential comes with a cost, the sensitive data collected usually needs to be published and shared, causing it to be available to be used by third parties as they please.

There is a whole range of risks related to data sharing, and the disclosure of sensitive data may lead to personal harm on targeted individuals, especially when linked with different sources. A study showed [10] a real-life privacy threat on William Weld, former governor of the state of Massachusetts in the United Stated of America, where an individual's name in a public voter list was linked with his record in a published medical database through the combination of zip code, date of birth and gender. Each of these attributes alone cannot uniquely identify a record owner, but their combination usually joins a small number or even a unique record owner. To achieve this identification success, the "attacker" must know two important pieces of prior information: the victim record in the released data, and the quasi-identifier of the victim. This knowledge can be obtained in many ways, in the example given, the "attacker" knew for a fact that the victim in question was hospitalized, therefore, the medical record of the victim would appear in the release patient database [11]. As for the rest of the victim's data, is, most of the time, very easy to obtain, (namely zip code, date of birth, and gender). Research showed that 87% of the US population had reported characteristics that made them unique based on the combination of just three data points, attributes known as quasi-identifiers: Zip code, gender and date of birth [12].

Even when data is believed to be anonymized, it might be done poorly as shown by Panduragan [13] where hash values to anonymize the license plate numbers of taxis in New York city were used, which was easily reversed to its original identifiable value as the licenses seven digit format only allowed two million possibilities. This meant that it was easy to link each number to its anonymized data, revealing sensitive information on the drivers like exact routes taken by the driver, gross income, and even where they lived. The previous examples, and others [14] [15] [16] [17] [18], show how important, and difficult, reliable data anonymization really is to achieve. So, there is a growing understanding of the risks related to the safety of our data. This also led to emergence of the European General Data Protection Regulation (GDPR) [19] [20] that mandates a strengthening protection of private and personal data, and in recent years there has been a considerable amount of research on ethical, legal and social issues on data sharing, e.g [21] [22].

1.1 Motivation

Sharing data brings many benefits to society, being for scientific advancements, assessing policies and improving services. Current privacy laws and regulations offer, arguably, limited guidance when dealing with the wide range of types of data and even in how to deal with new techniques of re-identifying data. Any organization or researchers are presented with many open questions when attempting to anonymize their data, and should never restfully believe that formerly accepted techniques are still secure and in acceptance with the current regulations. In the case of researchers, there is a vast variability on their knowledge of data privacy and their statistics, which might impact their decision when it comes to publish their data. It cannot be expected of them to only publish perfectly anonymized datasets, but instead facilitate a conscious decision. Thus, it becomes paramount not only clarifying the data privacy laws and guidelines, but also making the anonymization process easier, with more intuitive and easier to access anonymization tools and software.

Sometimes, even when data is believed to be anonymous it might be susceptible for re-identification as Sweeney recently showed in [23], where the authors were able to put real names of law school students to records produced by four protocols that were referred to as being popular ways to make personal information anonymous. The four protocols failed to provide the protection promised, with about half of the records being unique when none should have been.

As previously mentioned, anonymization has become important due to the GDPR. However, by definition, anonymization requires the removal of certain characteristics from data. This means that information that could be important for certain evaluations in research and studies can be lost, as a result, it becomes paramount to minimize data loss and find a reasonable balance with the risk of re-identification. This process, to be achievable, varies immensely depending on the type and the use cases of the data, as shown in this blog article [24].

1.2 Objectives

In this work, the objective is to study various anonymization techniques, particularly when applied to a data set with real educational private data. The main objective of this dissertation is to perceive how the application of said anonymization techniques contribute to a reasonable and comfortable reduction of the risk of re-identification, and how the desired outcome affects the information loss and data utility. More specifically, it is intended to explore the terminology used in the field of anonymization and elaborate a comprehensive review of the techniques currently available. Next, to apply the k-anonymity, ℓ-diversity and t-closeness privacy models to a dataset containing personal information from college graduate students. Finally, to present the results obtained for each privacy model and perform an analysis and comparative discussion of the obtained results.

1.3 Main Contributions

The work developed in this dissertation provided the following contributions to the field of data anonymization:

- The state of the art with regard to data anonymization;
- A practical example with application of various anonymization techniques and the consecutive analysis of the loss and utility of information, as well as of the re-identification risk;
- Two scientific publications, one containing a sensivity study of data anonymization procedure applied to a real open government data available from the Brazilian higher education evaluation system, and a second one illustrating a process of anonymization, comparing to several models of privacy, the loss of information and the usefulness of that dataset resulting from the anonymization:
  - W. Santos, G. Sousa, P. Prata and M. E. Ferrão, ”Data Anonymization: K-anonymity Sensitivity Analysis,” 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9141044.
  - P. Prata, M. E. Ferrão, W. Santos, G. Sousa. ”Garantia de Privacidade Versus Utilidade dos Dados em Anonimização: um estudo no ensino superior”, Accepted for publication in RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, December 2020.

1.4 Thesis Organization

The remainder of this thesis is organized as follow:

- Chapter 2 – Gives an introduction about the various concepts and the various anonymization methods available as well as how to evaluate the loss of information and associated risk.
- Chapter 3 – The state of the art is presented in the area of data anonymization.
- Chapter 4 – Application of selected anonymization techniques, and respective analysis regarding data loss, re-identification risk and data utility.
- Chapter 5 – Discussion of the results obtained.
- Chapter 6 – Summarizes the work presented in this dissertation, adding final conclusions. Finally some directions on possible future work are outlined.

Chapter 2

Fundamental Concepts

In this chapter, we review the concepts related to anonymization. Starting in section 2.1 with an introduction to the General Data Protection Regulation , and presenting the relevant regulation to the subject. Section 2.2 reviews the terminology used in the literature around the subject of anonymization, and also manifestly used in this dissertation. Similarly, section 2.3 introduces methods used in the anonymization process, called privacy models, and proceeds to meticulously describe those that were used for the purpose of the dissertation. Finally section 2.5 summarizes the most relevant conclusions of this chapter.

2.1 GDPR - General Data Protection Regulation

Since the introduction of the *European General Data Protection Regulation* (GDPR) there has been much discussion and research in order to help corporations and organization with activity inside the European Union to comply with the recent regulation. Recital 26 of the GDPR [25] defines anonymized data as “data rendered anonymous in such a way that the data subject is not or no longer identifiable.”. This definition clearly states that anonymized data must be cleared of any identifiable information, making it impossible to reveal and disclose information on a discreet individual. This means that when done properly [26] anonymization places the processing and storage of the anonymized data outside the scope of the GDPR.

However, to determine whether a specific person is identifiable the GDPR does not leave it completely clear, only stating that “account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly” [25], and also taking into account “the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing” [25].

The ambiguity of these statements has already caused some controversies and monetary fines, for example, the case of the Danish taxi service Taxa 4×35 [27], where it was concluded that “Taxa 4×35 failed to meet the high standard that the GDPR sets for data anonymization” [27] and citing the above mentioned Recital 26 and arguing that it was “relatively easy” to look up a phone number and match it to an individual on the database, and thus the TAXA dataset was not anonymous. [28]

To help clarify GDPR compliance, a website GDPR.eu[29] was created, where it is provided an extensive checklist [30] focusing on Lawful basis and transparency, data security, accountability, governance and privacy rights to limit organization's exposure to regulatory penalties.

2.2 Terminology

Due to the extensive number of terms used in the literature in the context of anonymization, this section explains the meaning of some terms used throughout the document:

- **Categorical data** – Represent types of data which may be divided into groups. Categorical data can take on numerical values (such as "0" indicating male and "1" indicating female), but those numbers have no mathematical meaning. Examples of categorical data are race, gender and educational level.
- **Numerical data** – Information that is something that is measurable. It is always collected in number form. Examples of numerical data are a person's heigh and grade point average.
- **Anonymization** – The processing of personal data in such a way that the data subject/individual is not or no longer identifiable, seeking to hide the identity and/or the sensitive data of record owners/individuals, assuming that sensitive data must be retained for data analysis. To this end, any explicit identifiers of records must be removed [31] [32].
- **Pseudonymization** – The "processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information" [33].
- **Attribute** – One type of information found on the data set. Usually a column. Examples of attributes are name, gender, and address.
- **Record** – Represents data about one individual.
- **Data set** – Collection of records. Conceptually similar to a table in a relational database or data-sheet.
- **Identifying attributes** – Attributes from the records associated with a high risk of re-identification and not necessary for analyses. Should be removed from the anonymized dataset. Examples of identifying attributes are Social Security numbers or names which explicitly identify record owners.
- **Quasi-identifying attributes** – Attributes from the records required for analyses that in combination with quasi-identifiers from other data sets can be taken advantage

of for re-identification. These attributes can be recoded to guarantee the data fulfills the privacy criteria [11]. Examples are gender, date of birth and ZIP codes.

- **Sensitive attributes** – Attributes that consist of sensitive person specific information which individuals are not willing to be linked with, and, if disclosed could cause harm to data subjects. These are required for analyses but may be subject to further constraints. Examples of these attributes are disease, salary and disability status.

- **Insensitive attributes** – These attributes are not associated with any privacy risk and so will be kept unmodified.

- **Suppression** –- Replacing certain values of the attributes by an asterisk '*'.

- **Generalization** – A generalization is a form of abstraction whereby common properties of specific instances are formulated as general concepts or claims.

- **Generalization hierarchies** – View of the structure from the bottom up, which leads to a more generalized or abstracted view of the higher classes [34], for categorical or continuous attributes, figure 2.1 shows age being generalized from the exact age in years to the respective age groups, for example an individual with the age of 26 will have his age recoded to "20-60", rather than revealing the precise value. At the top of the hierarchy an asterisk can be found, which is the highest generalization possible, as it includes all the possible values. Other examples: ZIP codes can be generalized by dropping, at each generalization step, the least significant (rightmost) digit; Postal addresses can be generalized to the street (dropping the door number), the to the city or even the state [35]. This Generalization hierarchies are usually applied to quasi-identifiers.

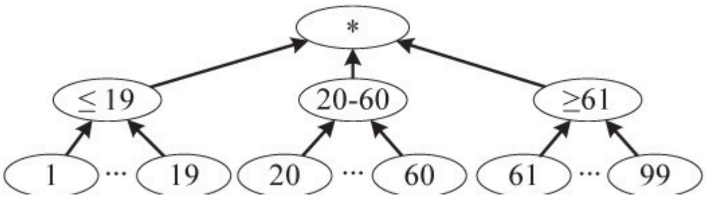


Figure 2.1: Hierarchies for attributes age and gender [1].

- **Generalization level** – Refers to the specific abstraction level in a generalization hierarchy. For example, as seen in figure 2.2 ZIP code was generalized by dropping the least significant (rightmost) digit (head(4)) would be referred as a level 1 generalization. Attribute "age" has three levels of generalization, level 0 is the exact age in years, level 1 is age below and above or equal to 50, and level 2 that includes all possible values. As for the attribute "gender", there are only two levels. The highest level for each attribute is always "*" as it includes all the possible values for that attribute.

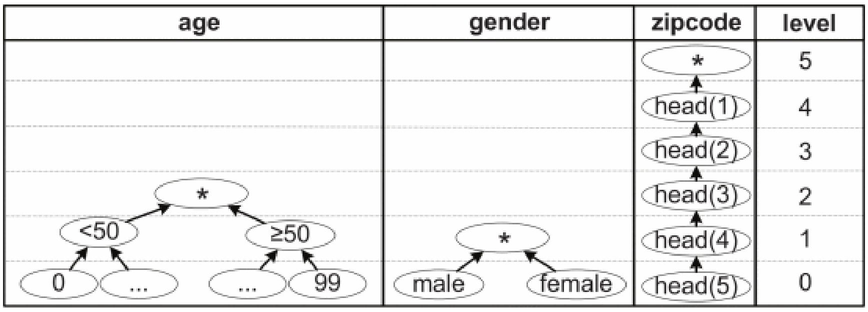


Figure 2.2: Generalization hierarchies and levels. [2]

- **Attribute disclosure** – Occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release.
- **Membership disclosure** – Occurs when the presence of an individual's personal information in a dataset is confirmed.
- **Identity disclosure** – Occurs when an individual is linked to a particular record in a dataset.
- **Equivalence class** – Each group of indistinguishable records in a data set.
- **Adversary** – Entity trying to re-identify one or multiple individuals using the data set supposedly anonymized.
- **Journalist attack model** – The attacker targets a specific individual and it is assumed that he already knows that data about the individual is contained in the dataset [36].
- **Prosecutor attack model**– the attacker targets a specific individual but it is not expected that he possesses background knowledge about membership [36].
- **Marketer attack model** – The attacker does not target a specific individual but he aims at re-identifying a high number of individuals. An attack can therefore only be considered successful if a larger fraction of the records could be re-identified [36].

2.3 Privacy Models

A wide range of privacy models addresses different threats, namely the previously defined membership disclosure, attribute disclosure, and identity disclosure. Depending on the intent and background knowledge of possible attackers, different approaches can be considered to evaluate the risk of the different privacy models, such as the prosecutor model, journalist model, and the marketer model. A Syntactic model enforces restrictions on the structure of data, statistical models estimate risks in relationship to a larger underlying population or the success probabilities of attacks while semantic models have more direct relationships to mathematical notions of privacy. [7]

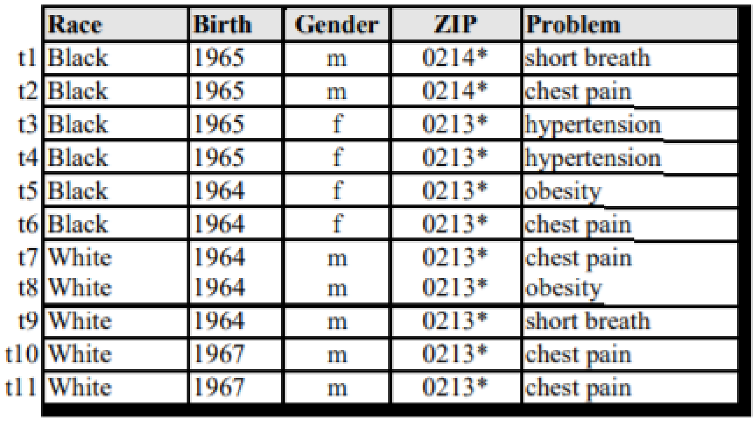
Table 2.1 shows an overview of various privacy models, framing the type, disclosure, and attacker models for each of them. The upcoming subsections will succinctly describe three of these privacy models, k-anonymity, ℓ-diversity, and t-closeness, as they are used for studying anonymization methods and effects later in this dissertation. These three privacy models were chosen because k-anonymity is the most well known privacy model, ℓ-diversity is an improvement to k-anonymity, providing protection against attribute disclosure, and finally t-closeness, since it also is an improvement from ℓ-diversity by maintaining the distribution of the sensitive attributes.

Table 2.1: Overview of privacy models. [7]

Privacy model	Type	Disclosure model	Attacker model
δ-Presence [37]	Syntactic/statistical	Membership	Journalist
k-Anonymity [3]	Syntactic/statistical	Identity	Prosecutor
k-Map [3]	Syntactic/statistical	Identity	Journalist
ℓ-Diversity [4]	Syntactic/statistical	Attribute	Prosecutor
t-Closeness [38]	Syntactic/statistical	Attribute	Prosecutor
δ-Disclosure privacy [39]	Syntactic/statistical	Attribute	Prosecutor
β-Likeness [40]	Syntactic/statistical	Attribute	Prosecutor
(ε ,δ)-Differential privacy [41]	Semantic	All	All

2.3.1 K-anonimity

K -anonymity[3] is a privacy model with the intent of protecting datasets from re-identification. A dataset is k-anonymous if each and every record cannot be distinguished from at least k-1 other records, regarding the quasi identifiers. Consider the table 2.3, where the quasi identifiers are “Race”, “Birth” Gender” and “ZIP” and complies with (k=2)-anonymity. Therefore, each of the tuples corresponding to a quasi-identifier appears at least twice. That is, for every record in the table, from t1 to t11 there is at least another record with the same value for “Race”, “Birth”, “Gender” and “ZIP”. If we were to consider a k=5 anonymity, for every record in the table, from t1 to t11 there must be at least 4 (k-1) records with the same value for “Race”, “Birth”, “Gender” and “ZIP”, which for the example table 2.3, are non existent. We can then conclude that table 2.3 is 2-anonymous, but it is in fact not 5-anonymous.



	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2.3: Example of k-anonymity, where k=2 [3].

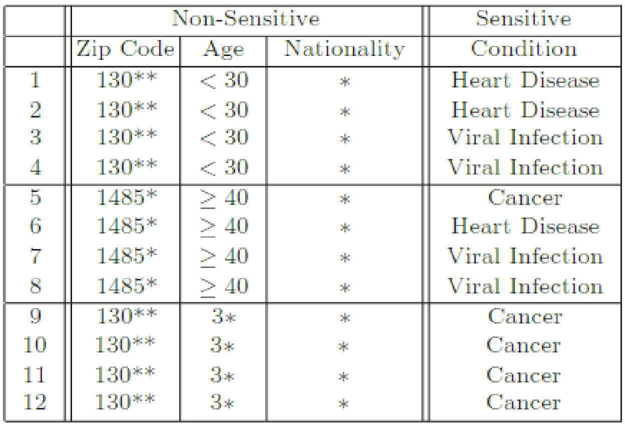
Sweeney and Samarati define k-anonymity as follows:

> ”Let T(A1,...,An) be a table and QI be the quasi-identifier associated with it. T is said to satisfy k-anonymity wrt QI if and only if each sequence of values in T[QI] appears at least k occurrences in T[QI]” [3] p.564

There are several algorithms to implement k-anonymity that have been developed [42] including, Datafly [10], Incognito [43] and Mondrian [44] , however results demonstrated that “there is no best anonymization algorithm for all scenarios, but the best performing algorithm in a given situation is influenced by multiple factors.” [42] For example, Incognito was shown to be time consuming and memory intensive in regards to the number of quasi-identifiers and Mondrian’s data utility was shown to be significantly impacted by the data distribution and the mechanism used for partitioning.

2.3.2 ℓ-Diversity

ℓ-Diversity [4] is a privacy model, improved from k-anonymity, to protect data against attribute disclosure by ensuring that each sensitive attribute has at least ℓ “well represented” values in each equivalence class. This privacy model was implemented due to the fact that k-anonymity can create groups that leak information due to the lack of diversity in the sensitive attribute. As seen in table 2.3, which is k=4 anonymous and where Zip Code Age and Nationality are quasi-identifiers, however an attacker with previous knowledge of a victim’s partial zip Code, Age and the victim’s presence in the dataset, could conclude the presence of a condition/disease which was sensitive information.



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Figure 2.4: Example of ℓ-Diversity[4].

Considering a *q*block* as the ℓ-diversity equivalent of an equivalence class, Machanavajjhala, Kifer, Gehrke and Venkitasubramaniam define the ℓ-diversity principle as follows:

> ”A q*-block is ℓ-diverse if it contains at least ℓ well-represented values for the sensitive attribute S. A table is ℓ-diverse if every q*-block is ℓ-diverse.” [4] p. 16

A more specific way to understand the improvements from k-anonymity is considering a determined and targeted attack that will use all the available resources and methods including ones based on probability to identify an individual. Probabilities such as: men having less breast cancer occurrences and Japanese having very low incidence of heart disease [4] p. 4. This might seem specific and too probabilistic but identifying specific individuals from datasets like public figures and/or politicians (as shown before) will have these kinds of details into account. Having an ℓ-diverse dataset will help to mitigate these very specific cases. And only then can we achieve a safely anonymized dataset.

Two instantiations of the ℓ-diversity principle exist: entropy ℓ-diversity and recursive ℓ-diversity [4]. Machanavajjhala, Kifer, Gehrke and Venkitasubramaniam define entropy ℓ-diversity as follows:

> ”A table is Entropy ℓ-Diverse if, for every q*-block
>
> $$-\sum_{s\in S} p_{(q*,s)} log(p_{(q*,s')}) \geqslant log(\ell)$$
>
> where $p_{(q*,s)} = \frac{n_{(q*,s)}}{\sum_{s'\in S} n_{(q*,s')}}$ is the fraction of tuples in the q*-block with sensitive attribute value equal to s.” [4] p. 17.

And recursive ℓ-diversity:

> ”In a given q*-block, let r_i denote the number of times the ith most-frequent sensitive value appears in that q*-block. Given a constant c, the q*-block satisfies recursive (c,ℓ)-diversity if $r_1 < c(r_\ell + r_{\ell+1} + ... + r_m)$. A table T* satisfies recursive (c, ℓ)-diversity if every q*-block satisfies recursive ℓ-diversity. We say that 1-diversity is always satisfied.” [4] p. 18.

2.3.3 t-Closeness

This privacy model is a further refinement of ℓ-diversity by additionally maintaining the distribution of the sensitive attributes. The authors Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian defines t-closeness [38] as:

> ”An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.” [38] p.2

An example to better understand the improvements from l-diversity is considering a salary attribute, for instance, in table 2.2, for the first group. The sensitive attribute salary is 3K 4K and 5K and although it is diverse, they are all very close and relatively low values which would lead to the disclosure of sensitive information for all three individuals, in this example a low income salary. This leakage of information occurs because while l-diversity actually ensures diversity in the sensitive attributes for each group, it does not take into account the semantical closeness of these values.

Table 2.2: t-closeness explanatory example.

	ZIP Code	Age	Salary	Disease
1	620**	2*	1K	gastric ulcer
2	620***	2*	2K	gastritis
3	620**	2*	3K	stomach cancer
4	271**	>40	5K	gastritis
5	271**	>40	12K	flu
6	271**	>40	8K	bronchitis
7	640**	3*	11K	pneumonia
8	640**	3*	12K	bronchitis
9	640**	3*	13K	gastric ulcer

To achieve t-closeness requirements the Earth Mover’s [45] distance is generally used to transform one distribution to another by moving distribution mass between each other. To calculate the Earth Mover’s distance between the two distributions there are three cases to consider:

- **Ordered Distance** – Applied to numerical values, the distance between two values of the attribute is based on the number of values between them in the total order.
- **Equal Distance** – Applied to categorical attributes, defines the distance between any two categorical attributes to be 1.
- **Hierarchical Distance** – Applied to categorical attributes, the distance between two values is defined by a given generalization hierarchy.

2.4 Utility Models

The evaluation of the utility of an anonymized dataset should be based on the intended use. The closer the results obtained from anonymized and the original dataset are, the more we can consider that the utility was preserved. Utility models typically evaluate data utility by quantifying the amount of information loss, for example, by measuring differences or similarities between the input and the output dataset [7]. Some examples of utility models include the sum of squared errors, average distinguishability [44], nonuniform entropy [46] and Discernibility [47]

In this dissertation, we use the analysis of variance (ANOVA) model to statistically compare the results between the original and the anonymized datasets. We present the model specification with two factors and the respective interaction, which can be generalized, with additive terms, to the number of factors and interactions referent to the analysis in question. Considering a sample of size n (i=1, ..., n), the model's equation is the following:

$$y_{ipk} = \mu + \gamma_p + \delta_k + \beta_{pk} + e_{ipk}$$

where y_{ipk} represents the final classification of i-th record belonging to group p of factor γ and also of group k of factor δ. Meaning, γ_p represents the first factor, δ_k represents the second factor and β_{pk} denotes the effect of interaction between the two factors, $p = 1, ..., P; k = 1, ..., K$. It follows that factor γ has P groups, factor δ has K groups and there are PK interaction subgroups. The random term of the model is represented by e_{ipk}, with the following assumptions: normal distribution with zero mean, homoscedasticity or homogeneity of variance, elements independent of each other. More detail about the model can be found in the work by Scheffé [48]

We used the SPSS [49] to estimate the parameter of the ANOVA model applied to several datasets. From the results reported by the statistical software, we chose the statistics F and the respective p-value [48] to show the impact that the privacy model may have on the final results. Despite the use of such methods being commonly applied for statistical inference purposes, we explicitly make clear that, for many reasons, their use in the context of this Dissertation does not intent to contribute for the consequences on such debate.

2.5 Conclusion

This chapter presented a comprehensive review of the concepts behind anonymization and the current algorithms, known as privacy models, used to achieve the targeted anonymization standard. First, we reviewed the GDPR and its relevant regulations its clearness, supported with a pratical real life example. Then a full description of the literature relevant to the subject of anonymization and that was used during the conception of this dissertation. Similarly, we presented methods used in the anonymization process, called privacy models, and proceeded to particularly analyze in detail k-anonymity,l-diversity, and t-closeness, as these were used in the practical experiment of this dissertation. Finally, we examined the utility model ANOVA as it was also used later in practical experiment.

Chapter 3

Related Work

In this chapter we review the state of the art related to Anonymization, different techniques, their results, associated data loss and re-identification risk. We also detail some of the existing tools used for data anonymization.

3.1 Saying it's Anonymous Doesn't Make It So

A recent paper "Saying it's Anonymous Doesn't Make It So" [23] by Sweeney et al. analyses law school graduate's hypothetical data. The paper focus on demonstrating that four protocols presented by an experienced group of data privacy practitioners, *The Sander Team*, which claimed that those four protocol protected privacy and were technically reasonable to implement, were not. The protocols were presented by *The Sander Team* when trying to obtain data collected by the State Bar of California that would be a good source data for a research on how race-based law school affirmative action policies relate to law school outcomes. The dataset contained individual-level data on the race, law school, year of graduation, bar exam score, bar passage result and score. The request for that data was denied after a trial at the Superior Court of California on the basis of proposed protocols, which were provided in a form that protected the privacy of applicants and that numerous countervailing interests outweighed the public's interest in disclosure. Therefore, the paper uses hypothetical, but a detailed and accurate version of the data to illustrate the flaws with the proposed protocols. The authors assess the re-identification risk in each of the four protocols.

The first protocol known as "11-Anonymity Protocol", is based on k-anonymity where k is 11, however it only enforces k-anonymity across certain data fields. It also applies generalization to the race variable. The second protocol, called the "Plus Protocol", took the 11-Anonymity Protocol and made further changes to it namely law school names and scores to make the values less specific. The third protocol, named the "Enclave protocol", does not focus on imposing the privacy by making changes to the dataset. Nevertheless, it determines where and how the dataset is to be shared. Once the dataset is constructed, it is only available in a secure, sequestered physical room ("safe room"). Finally, the fourth protocol, called the "Standardized Protocol", constructs a statistical database from the original dataset. School names are removed, race is generalized to 4 values and the scores are also removed. This was the least desired protocol as it heavily reduced data utility.

The results of the re-identification risk showed the real danger in the flaws of the used "11-Anonnymity" protocol. With the information collected online, the authors were able to identify real life individuals, were the "anonymized" dataset to be publicly released. This re-identification would be easily done through various methods. The same conclusions apply to the other protocols. The Enclave Protocol is particularly vulnerable to re-identifications of targeted individuals, even with the lesser risk of large-scale re-identification. As for the "The Standardized Protocol" the authors found that the technical knowledge required to execute was not reasonable. The conclusion is that none of the protocols provided k-anonymity protection. The authors gave many examples of re-identification cases and strategies, and explicitly note that more examples could be made. This leading to the conclusion that models like k-anonymity can provide guarantees of protection, but they have to be properly implemented.

Another recent article [50], proposes a generative copula-based method to quantify the likelihood for a re-identification attempt to be successful. They show that, even if the dataset is heavily incomplete, it may not satisfy the modern standards for anonymization defined by the GDPR.

In [51] Gregory E. Simon et al. describe a framework for assessing and mitigating risk of re-identification when sharing research health data, emphasizing that the risk of re-identification depends on external resources and datasets that might be available, and even the motivation of a potential adversary, factors that might be unknown at the time of data release. It is also concluded that the risk is disproportionately higher with people with rarer health conditions and members of minority racial or ethnic groups.

3.2 Models Based on k-anonymity

In this section we present some works that consider k-anonymity as a basis to obtain anonymized data. Starting with the most well known privacy models, Rajendran et al. [52] present a detailed description of k-anonymity, ℓ-diversity and t-closeness comparing the advantages and disadvantages of each. They refer that k-anonymity is effective when preserving against identity disclosure. The ℓ-diversity model provides a greater distribution of sensitive attributes, but can be redundant and laborious to achieve. Finally, t-closeness identifies the semantic closeness of attributes, a limitation of ℓ-diversity, but it necessitates that sensitive attributes spread in the equivalence class to be close to that in the overall table.

In [53], Jang discusses a method based on deep anonymization detection to find an appropriate solution to reduce information loss when applying k-anonymity and ℓ-diversity to big data. The decision for deep anonymization is done by considering a domain data characteristic, data receiver's purpose, and data importance. This decision then influences the information distortion applied to the dataset.

The article by Esquivel-Quirós et al. [54] studies k-anonymity privacy model by performing an evaluation of privacy preserving and data utility of the anonymized dataset using machine learning algorithms. Between the various experiments the authors vary the number of quasi identifier attributes and the value of k. The result analysis shows loss in the data utility as the value of k increases, concluding that it is difficult to establish good parameters.

In the realm of health data, since it handles the most sensitive and private information about the population, there is an extensive amount of research about the topic. In [55] a biomedical dataset was used to evaluate the risk and utility of anonymized data using ℓ-diversity, t-closeness and β-likeness. For measuring the utility of output data, the authors used a general-purpose model that captures the granularity of output data. The authors discuss how hard it is to achieve a reasonable trade-off; however, "when data is only moderately skewed, both β-likeness and ℓ-diversity can yield significantly better risk-utility trade-offs than the baseline approaches." [55]. Additional cases and methods to anonymize heath data can be found in [56] and [57].

In [58] the authors present an empirical risk analysis on a real world k-anonymised dataset. They convey that if the adversary has only knowledge of some or all quasi-identifiers attributes, the risk of re-identification is usually lower than the worst-case risk for the majority of the records. And on the contrary, the risk can be significantly higher with the knowledge of other non quasi-identifier attributes, thus raising the importance for the careful selection of quasi-identifier attributes, and the assumptions of the adversary's prior knowledge.

Some dedicated solutions are developed with k-anonymity as a foundation, providing anonymization solutions for very specific areas of activity. For example, Zhang et al. [59] propose a privacy-preserving solution for continuous location based services through multi-level caching and spatial k-anonymity designed to improve the user location privacy. In [60] is proposed a secure e-voting scheme based on k-anonymity. In [61], Schwee et al. evaluates anonymization techniques applied to data from inexpensive IoT sensors, showing that only the case with a combination of k-anonymity and of suppression can disable the attack for re-identifying the dataset.

The k^m-anonymity model ensures that any attacker who knows up to m items of a target record cannot use that knowledge to identify more than k individuals in the dataset. This [62] work presented by Gkountouna in 2014, describes a k^m anonymity approach for continuous numerical data. The aim is to provide protection against identity disclosure and significantly limit the information loss by not using a fixed *a priori* generalization hierarchy, but the anonymization algorithm dynamically explores different possible anonymization levels. This solution has a significantly larger set of possible generalization levels than the standard k-anonymity. Therefore, to deal with such complexity of the optimal anonymization, the authors opted for a heuristic solution that selects the best generalization level at each step. The results showed an increase in data utility.

The authors of [63] propose an anonymization technique to achieve another probabilistic relaxation of k^m-anonymity. The suggested technique does not rely on the classification of sensitive or quasi-identifier attributes, neither pre-defined generalization hierarchies but rather on “more general constraints describing the desired output”. The idea of this approach is to perform anonymization only if k-anonymity is violated. The assessment of data utility is done by measuring the distance between the original dataset and the anonymized version. The authors conclude that the relaxation method was important to achieve scalability with large datasets and improve the utility of the anonymized data.

Other approaches also consider the k^m-anonymity method, in [64] , Terrovitis et al. also define a new version of k^m-anonymity that also relies on generalization instead of suppression, they conclude that although their proposed solution was “optimal but not scalable” hence “not practical for large, realistic databases”. The authors also aim, in the future, at extending the model to ℓ-diversity. Another approach to anonymize data was proposed in [65], where the privacy constrains are specified pre anonymization. In [66], the k^m-anonymity model has been applied to trajectory data by using distance based generalization.

3.3 Probabilistic Anonymization - Random Based

The paper [67], by Liu, addresses the anonymization of social networks data with the combination of k-anonymity and randomization methods. They assess data utility by using various utility metrics, including measuribg the similarity between the original data and anonymous data. The authors achieve an increase in data utility while keeping the same k privacy level. The experimental results show that the proposed algorithm modifies the original data less than other randomization methods. Still with social networks data Ying et al. [68], suggest adding using randomization to improve privacy protection,

In the paper [69] by Avraam et al.the authors developed a method that adds normally distributed random noise to the input dataset, according to the user specified variance. It then proceeds to calculate the re-identification risk of the anonymized data taking advantage of the method proposed in [70] by Goldstein and Shlomo, by estimating “the probability of an attacker being successful in identifying their individual of interest within the anonymized dataset.”[69]. In summary the method used by the algorithm consists on calculating the Eclidean distances between each row in the true dataset and all rows in the noisy dataset, in order to determine the closeness between each of the original records and every record in the noisy dataset. The records are ranked by identifying the position of the closest record. Similarly, the algorithm also calculates the Euclidean distance between a copy of the original dataset and each row of the original dataset. The intent is to rank them in order of distance. Finally, the algorithm proceeds to identify the distances previously calculated at the location before identified as the closest record, and calculates the difference between the two into a vector, *h-ranks*. According to the authors, the average value of the *h-ranks* provides a metric of disclosedness, where the larger average of *h*, the greater the level of protection against identity disclosure. The authors conclude that a probabilistic anonymization procedure can reduce the disclosure risk to acceptable levels, in addition to retaining data utility despite suffering "some loss of statistical efficiency when compared with analysis on the true data."[69].

3.4 Other Approaches

A scalable method for the safe sharing of health data was proposed in [5], where the user accesses a proxy which redirects to the interface of the proposed analytics solution running on top of the sensitive data. The proxy assures every user is properly authenticated and authorized, preventing the user from breaching privacy. Figure 3.1 shows a data-flow diagram for the basic design of the proxy.

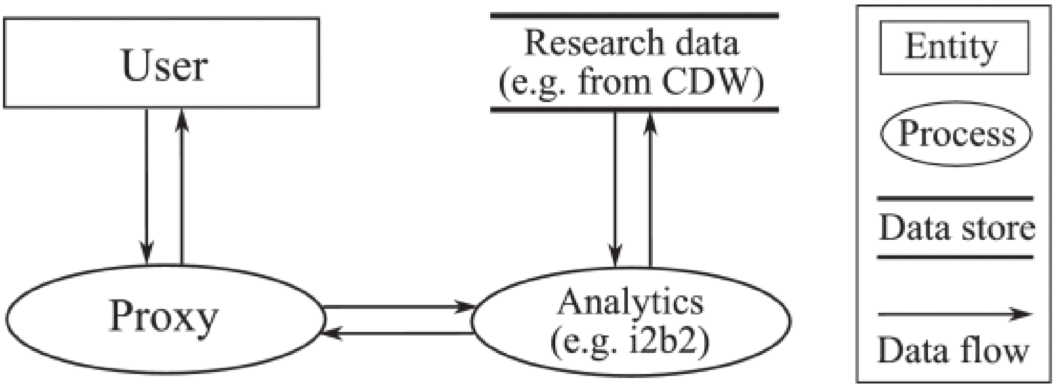


Figure 3.1: Data-flow diagram for the basic design of the proxy[5]

A Machine learning aided anonymization was proposed by Shaham et al. [71], where the authors developed a method to preserve privacy of users publishing spatio-temporal trajectories.

An overview of data privacy techniques that are applicable to data mining is surveyed by Mendes and Vilela [72], further describing the most common approaches throughout the collection, publishing and distribution phases of data. In addition, the authors document various metrics to measure the level of privacy, including k-anonymity based privacy models, and data utility which includes parameters such as information loss.

In [73] a novel framework is proposed for generating synthetic health patient records that have similar characteristics to the real patient records, suggesting that the framework could be used as a safe, legal and ethical solution for the sharing of health data.

3.5 Performance Evaluation

In [42], the authors perform a systematic comparison of three k-anonymization algorithms, Datafly, Incognito, and Mondrian, to measure their efficiency and data utility on a real and in a synthetically generated dataset. Data utility is assessed by examining the amount of records preserved and with the help o statistical classifiers. The paper demonstrated significant performance difference between the evaluated algorithms; however, none of them outperformed all the others across all the evaluated metrics. The authors of [74] compare Datafly, Improved Heuristic Greedy Algorithm, Samarati's Algorithm and *Algorithm and Optimal Lattice Anonymization* (OLA) Algorithm, they also conclude no algorithm outperforms the other, with Datafly having less execution time and information loss but giving an optimum local solution and suggesting that data publishers should know in prior the application of the data being used. Previously, El Emam and C. Álvarez [75] showed a comparison between their proposed algorithm OLA, and the three other k-anonymity de-identification algorithms previously mentioned (Datafly, Samarati, and Incognito). The authors discuss that Datafly and Samarati tended to have higher information loss than the proposed algorithm; however, OLA had the exact information loss as Incognito but was found to be significantly faster.

When dealing with big data the computation performance of the algorithms is an important factor, for example in [76] the authors evaluate the execution time of t-closeness with ARX [36] on differently sized datasets. The method used in ARX [36] utilizes high-performance data structures reducing execution times of the anonymization processes by up to a factor of two.

3.6 Anonymization Tools

In this section we present some of the most well-known tools currently used in data anonymization. We start with a detailed description of ARX [36], which is the anonymization tool used in this dissertation, and then we present an overview of some of the other available tools.

3.6.1 ARX Data Anonymization Tool

ARX [36] is an open source [77] software of anonymizing sensitive personal data, it transforms structured tabular data and can be used to remove/suppress direct identifiers and enforce constraints on quasi-identifiers. It also supports methods for removing sensitive attributes from disclosure and semantic privacy models. ARX supports a variety of privacy models including k-anonymity, ℓ-diversity and t-closeness. Records can be made less unique as it supports generalization based on user specified hierarchies. ARX provides a graphical frontend with various visualizations and wizards to help import and prepare the dataset for anonymization.

The tool divides the anonymization process into four phases, as figure 3.2 shows. 1) configuring privacy models, utility measures and transformation methods, 2) exploring the solution space, 3) analyzing data utility and 4) analyzing privacy risks

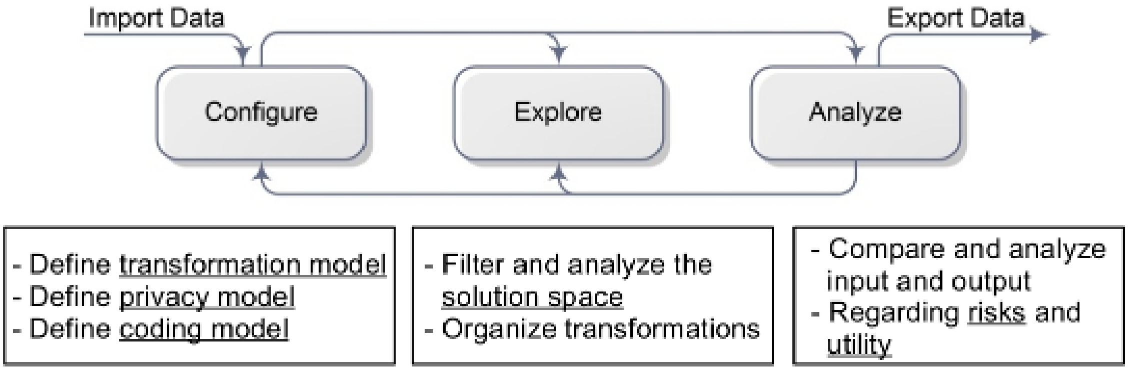

Import Data
Export Data
Configure
Explore
Analyze
- Define transformation model
- Define privacy model
- Define coding model
- Filter and analyze the solution space
- Organize transformations
- Compare and analyze input and output
- Regarding risks and utility

Figure 3.2: ARX anonymization process.[6]

During the configuration phase the input dataset is imported and the transformation rules can be specified such as selecting the privacy models, categorizing each attribute as sensitive, quasi-identifier or insensitive, building or importing the generalization hierarchies and determining the max suppression level. There are different methods for creating generalization hierarchies corresponding to different types of attributes, the masking-based that allows for the creation of hierarchies for a broad spectrum of attributes, interval-based for variables with a ratio scale, order-based for variables with an ordinal scale, and date-based, that can be used for dates.

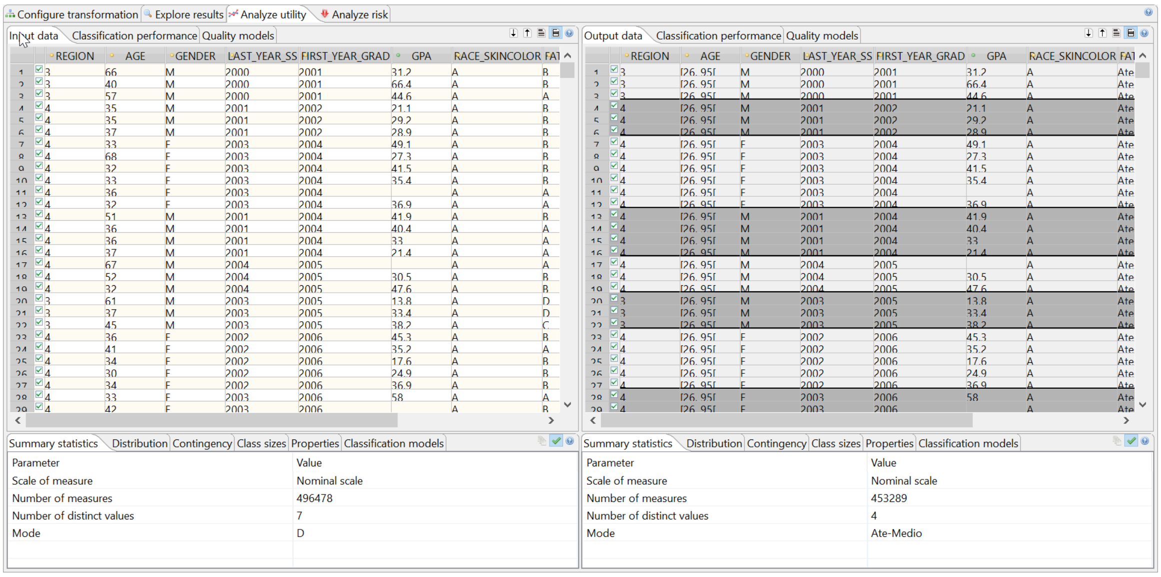
As for the available supported privacy models, the full list can be seen in table 3.1. Most models support weights that should be assigned to those attributes where a specified importance is desirable, reducing the loss of information for the attributes with higher weight. The data import wizard supports csv files, excel spreadsheets and some relational database systems. Here it is also possible to rename, remove, reorder columns and select the datatype.

Table 3.1: Privacy models supported by ARX.

Privacy model
δ-Presence [37]
k-Anonymity [3]
k-Map [3]
ℓ-Diversity [4]
t-Closeness [38]
δ-Disclosure privacy [39]
β-Likeness [40]
(ε ,δ)-Differential privacy [41]

Throughout the anonymization process a solution space with the potential transformations is generated. For each solution candidate the risk thresholds are displayed, and it is stated whether they are met or not. It is also possible to apply the various transformations and explore their effects on the dataset, namely the loss and risk, hence being called the exploration phase.

During the analysis phase the suitability of the selected transformation is confirmed. To help with the analysis the original and the transformed datasets are displayed side by side. Moreover, on the interface, the horizontal and vertical scroll bars of both datasets are synchronized. A demonstration of said interface can be seen in figure 3.3. Furthermore, some graphical and numerical representations can be analyzed, namely, empirical distribution with the histogram showing the frequency of the various attributes, a contingency view showing a heat map of two selected attributes, equivalence classes with an information summary of the dataset, and lastly the data utility models.



		REGION	AGE	GENDER	LAST_YEAR_SS	FIRST_YEAR_GRAD	GPA	RACE_SKINCOLOR	FAT
1	☑	3	66	M	2000	2001	31.2	A	B
2	☑	3	40	M	2000	2001	66.4	A	B
3	☑	3	57	M	2000	2001	44.6	A	A
4	☑	4	35	M	2001	2002	21.1	A	B
5	☑	4	35	M	2001	2002	29.2	A	B
6	☑	4	37	M	2001	2002	28.9	A	B
7	☑	4	33	F	2003	2004	49.1	A	B
8	☑	4	68	F	2003	2004	27.3	A	B
9	☑	4	32	F	2003	2004	41.5	A	B
10	☑	4	33	F	2003	2004	35.4	A	B
11	☑	4	36	F	2003	2004		A	A
12	☑	4	32	F	2003	2004	36.9	A	A
13	☑	4	51	M	2001	2004	41.9	A	B
14	☑	4	36	M	2001	2004	40.4	A	A
15	☑	4	36	M	2001	2004	33	A	A
16	☑	4	37	M	2001	2004	21.4	A	A
17	☑	4	67	M	2004	2005		A	A
18	☑	4	52	M	2004	2005	30.5	A	B
19	☑	4	32	M	2004	2005	47.6	A	B
20	☑	3	61	M	2003	2005	13.8	A	D
21	☑	3	37	M	2003	2005	33.4	A	D
22	☑	3	45	M	2003	2005	38.2	A	C
23	☑	4	36	F	2002	2006	45.3	A	B
24	☑	4	41	F	2002	2006	35.2	A	A
25	☑	4	34	F	2002	2006	17.6	A	B
26	☑	4	30	F	2002	2006	24.9	A	B
27	☑	4	34	F	2002	2006	36.9	A	B
28	☑	4	33	F	2003	2006	58	A	A
29	☑	4	42	F	2003	2006		A	B

Parameter	Value
Scale of measure	Nominal scale
Number of measures	496478
Number of distinct values	7
Mode	D

		REGION	AGE	GENDER	LAST_YEAR_SS	FIRST_YEAR_GRAD	GPA	RACE_SKINCOLOR	FAT
1	☑	3	[26, 95[	M	2000	2001	31.2	A	Ate
2	☑	3	[26, 95[	M	2000	2001	66.4	A	Ate
3	☑	3	[26, 95[	M	2000	2001	44.6	A	Ate
4	☑	4	[26, 95[	M	2001	2002	21.1	A	Ate
5	☑	4	[26, 95[	M	2001	2002	29.2	A	Ate
6	☑	4	[26, 95[	M	2001	2002	28.9	A	Ate
7	☑	4	[26, 95[	F	2003	2004	49.1	A	Ate
8	☑	4	[26, 95[	F	2003	2004	27.3	A	Ate
9	☑	4	[26, 95[	F	2003	2004	41.5	A	Ate
10	☑	4	[26, 95[	F	2003	2004	35.4	A	Ate
11	☑	4	[26, 95[	F	2003	2004		A	Ate
12	☑	4	[26, 95[	F	2003	2004	36.9	A	Ate
13	☑	4	[26, 95[	M	2001	2004	41.9	A	Ate
14	☑	4	[26, 95[	M	2001	2004	40.4	A	Ate
15	☑	4	[26, 95[	M	2001	2004	33	A	Ate
16	☑	4	[26, 95[	M	2001	2004	21.4	A	Ate
17	☑	4	[26, 95[	M	2004	2005		A	Ate
18	☑	4	[26, 95[	M	2004	2005	30.5	A	Ate
19	☑	4	[26, 95[	M	2004	2005	47.6	A	Ate
20	☑	3	[26, 95[	M	2003	2005	13.8	A	Ate
21	☑	3	[26, 95[	M	2003	2005	33.4	A	Ate
22	☑	3	[26, 95[	M	2003	2005	38.2	A	Ate
23	☑	4	[26, 95[	F	2002	2006	45.3	A	Ate
24	☑	4	[26, 95[	F	2002	2006	35.2	A	Ate
25	☑	4	[26, 95[	F	2002	2006	17.6	A	Ate
26	☑	4	[26, 95[	F	2002	2006	24.9	A	Ate
27	☑	4	[26, 95[	F	2002	2006	36.9	A	Ate
28	☑	4	[26, 95[	F	2003	2006	58	A	Ate
29	☑	4	[26, 95[	F	2003	2006		A	Ate

Parameter	Value
Scale of measure	Nominal scale
Number of measures	453289
Number of distinct values	4
Mode	Ate-Medio

Figure 3.3: ARX interface with synchronized scroll bars.

After analyzing the data, the risk can be inspected, these include re-identification risks for the methods described in section 2.2: journalist, prosecutor and marketer risk.

All the described features are also accessible via an *Application Programming Interface* (API), including loading datasets, adding generalization hierarchies and execution of the available privacy models.

3.6.2 Other Anonymization Tools

Other anonymization tools worth noting but not used in this work include:

- Amnesia [78] – Open source and developed in Java with an installable version and a limited online version, both with the same user interface. It supports loading datasets from csv files and defines the data type of each variable. Generalization hierarchies can be imported or auto generated, the auto generation will consider the variable type for each attribute and build it according to its distinct or range values. The only privacy model supported is the k-anonymity. Once the algorithm is executed, we are presented with the solution space with the various potential transformations, then the result anonymized dataset can be compared with the original dataset, side by side, and finally it can be exported to a csv file.

- sdcMicro [79] – Open source and developed in R programing language is used for the generation of anonymized (micro)data, i.e. for the creation of public- and scientific-use files. Most functionalities of the package are also available via an interactive graphical user interface.

- Aircloak Insights[80] – A commercial tool that operates under a proxy linking analyst and the sensitive data set. It processes normal *Structured Query Language* (SQL) queries to an SQL or NoSQL big data store, and the results are returned ensuring they are aggregated and fully anonymized. The full platform deployment consists of two separate components: Insights Air and Insights Cloak. Insights Cloak analyzes and anonymizes any sensitive data requested and runs inside a secure perimeter. Insights Air is the web control center which offers full control over the data, with built-in authentication and authorization controls.

- UTD Anonymization Toolbox [81] – Open source and developed in Java contains three different privacy models, k-anonymity, ℓ-diversity and t-closeness.

3.7 Conclusion

This chapter reviewed the state of the art, starting by analyzing works that highlight the difficulty in determining whether a dataset is truly anonymized, and proceeding to present various of the current anonymization techniques, namely the ones based on k-anonymity that further improve it either for privacy or data utility reasons, or to optimize it to specific cases of utilization, techniques based on randomization of the data, techniques based on other approaches such as securing data behind proxies and machine learning anonymization. Then, some articles regarding the performance of various algorithms are analyzed. Finally, we detail some of the existing tools used for data anonymization, particularly ARX Data Anonymization Tool.

Chapter 4

Example of Application to the ENADE dataset

In this chapter we review the ENADE dataset, and discuss why it was selected for this study. We also analyse the contents in the dataset and proceed with the classification of the attributes. Finally, for each stage of the study we measure the loss of data and the re-identification risk as well as an evaluation of the data utility.

4.1 ENADE Data

The ENADE takes place every year in Brazil since 2004. The purpose of the exam is to evaluate the higher education graduates' performance. It is part of *Sistema Nacional de Avaliação da Educação Superior* (SINAES). The exam is mandatory and in the form of questionnaires covering several cognitive domains depending on the area of studies. Each year a subgroup of disciplinary areas is evaluated so that whole evaluation cycle occurs over a triennium.

Only the data from the year 2018 was considered for analyses in which 548,127 students were involved. The data is publicly available online for download on the ENADE website [82] and contains 137 variables for each record. For the purpose of this dissertation we considered: the student's general score, i.e. grade point average (GPA), sociodemographic variables such as Gender, Age, self-declared Race/skin color, Mother's education, Father's education and Household income. The higher education institution and program identification codes (respectively University id and Program id), Region, Year of high school conclusion (YHSC) and Year of beginning graduation (YBG) are also included for analyses. Table 4.1 presents the selected variables as well as their respective scales as listed in the data dictionary. The abbreviations of the variables used from now on will be given in brackets.

Before being subject to the study, the data was pre-processed in order to remove some less plausible values. The process consisted of eliminating the values of the first year of graduation and the last year of secondary school that led us to the conclusion of negative values for the number of years needed to finish the graduate studies or the number of years to start the graduation. The cases where the starting year of graduation coincided with the last year were also removed, because that is not possible, since in Brazil the academic year agrees with the civil year. Furthermore, and since it was also incoherent, we ended up eliminating the cases that had the value of first year of entrance in graduating studies greater than 2018.If the value for research variables were all missing data, the respective records were also suppressed at this stage. The whole process resulted in the elimination of 41,447 records from the downloaded 2018 dataset. In the end, there were 506,680 records which will be considered the original dataset for the remainder of the document.

Table 4.1: Selected variables, abbreviated names and scales.

Variable	Scale
University id	Between 1 and 23 410
Program id	Between 1 and 5 001 389
Region	1 = North (N) 2 = Northeast (NE) 3 = Southeast (SE) 4 = South (S) 5 = Central-West (C-W)
Age	Between 4 and 94

Gender	M = Male
	F = Female

Year of high school conclusion (YHSC)	AAAA = Between 0 and 2,686
Year of beginning graduation (YBG)	AAAA = Between 1,973 and 2,099
Grade point average (GPA)	Minimum = 0; Maximum= 93.7

Race / Skin Color (Race)	A = White
	B = Black
	C = Yellow
	D = Pardo
	E = Indigenous
	F = Not declared

Mother's education (Mother Edu)
Fathers's education (Father Edu)
	A = None
	B = 1st – 5th grade
	C = 6th – 9th grade
	D =Secondary school
	E = Graduation
	F = Post-graduation

Household income (Income)	A = Up to 1.5 minimum wages
	B = 1.5 to 3 minimum wages
	C = 3 to 4.5 minimum wages
	D= 4.5 to 6 minimum wages
	E = 6 to 10 minimum wages
	F = 10 to 30 minimum wages
	G = Above 30 minimum wages

Although the dataset is publicly available, assuming the data is real, and despite direct and highly identifying information like name and social/government ids not being present in the released dataset, it can hardly be considered anonymized. As it was mentioned and discussed previously in this dissertation, the grouping of the various attributes that are present, namely quasi-identifiers, can disclose some sensitive information, and, in the worst case scenario, accurately link a real person to a specific record, disclosing the information from all the attributes present in that record, revealing details about the participants that were not meant to be available to the public.

The immense number of records in the dataset, around half a million, may give a false sense of security, transmitting the wrong idea that unique records must be rare. This was not in accordance with the results obtained in this study. As it will be shown later, even with an anonymization of k=2, where all the unique records are removed, the number of records removed is very high, and it is worth noting that in this study not all variables were used, which would increase the number even more.

As a practical example let us consider a random individual named Bob, knowing that Bob finished his degree in 2018, he would have participated in ENADE and, consequently, be present in the dataset. We will also consider that we know some basic information from Bob, such as sex, age, the university where he studies, etc. In the dataset there will be information for directly identifying the university, and the course. There will also be a lot of personal information on each individual such as, age, gender, marital status, race, household details, etc. For anyone that personally knows Bob (colleagues, teachers, friends) it is reasonable to assume that they would have knowledge of the mentioned information. Based on this study's results, there is a high probability that there is only one record corresponding to that basic information. This means that for anyone that personally knows Bob there is a high chance that it is possible to obtain all the answers Bob gave in the ENADE questionnaire. This is only taking into account any previous knowledge in order to link an individual with a record, at the same time, there is also the risk of possible linkage to other datasets where multiple individuals could be identified. This makes the ENADE dataset a good dataset for this study, as it is composed of real data from a number of commonly used attributes in the field of anonymization.

To assess data utility after the anonymization process, we apply the ANOVA model to the original dataset, with GPA as the dependent variable and *Region, Gender, YHSC, YBG, Race, Mother Edu, Fathers Edu* and *Income* as the fixed factors.

Table 4.2 presents the statistic results of test F and the p value for the original dataset under analysis. From table 4.2 we attest that the factors YBG, Race, Father Edu and Income are statistically relevant to a level of significance of 5% (p value < 0.05). Likewise, with the exception of the interaction between *YHSC * Father Edu*, the results show that all interactions between the fixed factors are also statistically relevant. Moreover, according to the obtained results we can not reject the null hypothesis to the fixed factors *Region, YHSC, Gender and Mother Edu*, and the interaction between *YHSC and Father Edu*. In the table, the terms which in the original dataset are differenciated groups in their relation with the dependent variable, GPA, are highlighted in grey.

Table 4.2: ANOVA with fixed factors and interaction for the original dataset.

Variable	**F**	**p value**
Region	0.612	0.654
YHSC	1.218	0.113
YBG	3.038	0.000
Gender	0.250	0.617
Race	2.762	0.017
Father Edu	6.819	0.000
Mother Edu	0.035	0.999
Income	8.111	0.000
YBG * Father Edu	1.429	0.005
Gender * Father Edu	10.277	0.000
Father Edu * Income	3.299	0.000
YHSC * Father Edu	1.038	0.332
Father Edu * Mother Edu	17.574	0.000
Race * Father Edu	1.616	0.027
Region * Father Edu	4.547	0.000
YBG * Gender	1.804	0.019
YBG * Income	1.794	0.000
YHSC * YBG	1.484	0.000
YBG * Mother Edu	1.257	0.047
YBG * Race	1.762	0.000
Region * YBG	6.505	0.000
Gender * Income	14.269	0.000
YHSC * Gender	3.200	0.000
Gender * Mother Edu	20.517	0.000
Gender * Race	18.080	0.000
Region * Gender	9.507	0.000
YHSC * Income	2.978	0.000
Motherr Edu * Income	5.685	0.000
Race * Income	3.856	0.000
Region * Income	3.822	0.000
YHSC * Mother Edu	1.189	0.026
YHSC * Race	1.231	0.010
Region * YHSC	1.984	0.000
Race * Mother Edu	3.939	0.000
Region * Mother Edu	3.379	0.000
Region * Race	7.985	0.000

4.2 Study 1 – K-anonymity Sensitivity Analysis

In the first study, the anonymization is conducted applying the k-anonymity privacy model when varying K=2, ..., 5. Table 4.3 shows the variable's classification. All the variables were classified as quasi-identifier except for GPA and Income that were classified as insensitive. The quasi-identifier attributes were classified as such due to the particularity of those variables that in combination with information from external sources can be taken advantage of for re-identification. As for the insensitive attributes, the GPA is later used to assess data utility and Income is considered as a sensitive variable, but since k-anonymity does not treat sensitive attributes we classify it as insensitive so the variable will be kept unmodified. Is is also worth noting that at this point no generalization is applied to any variable.

Table 4.3: Variables selected for Study 1 (k-anonymiy without generalization) and respective classification.

Variable	Classification
University id	Quasi-identifier
Program id	Quasi-identifier
Region	Quasi-identifier
Age	Quasi-identifier
Gender	Quasi-identifier
YHSC	Quasi-identifier
YBG	Quasi-identifier
GPA	Insensitive
Race	Quasi-identifier
Mother Edu	Quasi-identifier
Father Edu	Quasi-identifier
Income	Insensitive

For each value of k, a k-anonymous dataset was obtained. This means that for k=2 every record has at least another k-1 identical record, based on the quasi-identifiers. This is valid for every value of k studied, k=2, ..., 5. Any record not satisfying that condition will be suppressed.

The anonymization process was conducted using ARX by providing it with information on which attributes are quasi-identifiers or insensitive. After that, the loss of information is calculated based on the number of records that were suppressed, information that is available in ARX once the anonymization process is complete.

Next the re-identification risk is evaluated, as k-anonymity groups the records into equivalence classes of k records, the maximum risk will always be 1/k, meaning the probability of identifying a record. As for the average risk it is calculated by obtaining the average risk of the equivalence classes and it is also available in ARX, once the anonymization process is complete. This processes is replicated for the succeeding studies.

4.2.1 Results - Information Loss, Risk and Data Utility

Table 4.4 presents the number of records in the dataset resulting from the anonymization process when varying K=2, ..., 5. It also shows the percentage of suppressed records. As can be seen, the suppression percentage of records was more than 90% across all k variations in this study. For the highest k tested, k=5, there is an extreme loss of data, where of the 506,680 only 7,586 are left after the anonymization process.

Table 4.4: Number of records after anonymization and percentage of suppressed records for Study 1 (k-anonymity without generalization).

K = 2	**K = 3**	**K = 4**	**K = 5**
48 951	19 775	11 342	7 586
90.34%	96.10%	97.76%	98.50%

As table 4.4 shows, as k increases, the suppressed records also increase. Even for the lowest k, the percentage of suppressed records surpasses 90%, meaning that only less than 10% of the records have at least one other record with the exact same values for the quasi-identifier attributes.

Table 4.5 presents the maximum and average re-identification risk for k varying k=2, ..., 5, calculated as described in the previous section. As for the column k=1 it refers to the original dataset, before k-anonymity is performed, where the maximum re-identification risk is 100 and the average risk is 94.16, which is also considerably high. As it can be seen, the risk declines as k increases, and the average risk of re-identification decreased from more than 90% to near 14%.

Table 4.5: Maximum and average re-identification risk for Study 1 (k-anonymity without generalization).

k = 1	**K = 2**	**K = 3**	**K = 4**	**K = 5**
Maximum risk				
100%	50%	33.3%	25%	20%
Average Risk				
94.16%	39.56%	24.18%	17.37%	13.59%

Table 4.6 presents the results of applying the ANOVA model to the datasets obtained for each k in Study 1. From table 4.6 we verify that for k=2 the factors *Father Edu and Income* and the interactions *Father Edu * Income, Father Edu * Mother Edu, Region * Father Edu, YBG * Mother Edu, YHSC * YBG, YBG * Race, Region * YBG, Gender * Income, Region * Gender, Region * Income, Region * YHSC, Race * Mother Ed* are statistically relevant to a level of significance of 5%. The number of statistically relevant factors decreases for k=3 where only *YBG* and the interactions *YBG * Father Edu, YBG * Mother Edu, Region * YBG, Region * YHSC and Region * Mother Edu* are statistically relevant. The number keeps declining with *Race, YBG * Father Edu, YHSC * Father Edu, YBG * Mother Edu and Region * Mother Edu* for k=4. Finally for k=5, *Gender * Father Edu, Region * YHSC and Region * Mother Edu* are the statistically relevant interaction terms. Comparing the results for k=2, 3, 4 and 5 we highlight that there is no consistency in the factors that are statistically relevant, considering that not a single factor or even interactions possess this property across all of the values of k. In the following tables the values of p for the statistically relevant terms are highlighted.

The comparison of the ANOVA results from table 4.6 with the ANOVA results of the original dataset, presented in table 4.2, will support the assessment of data utility preservation after the anonymization process. The result states that of the 31 factors that were statistically relevant to explain the dependent variable (GPA) in the original dataset, 14 of them keep that property for k=2, decreasing to 6 for k=3, 5 for k=4 and 3 for k=5.

Table 4.6: ANOVA with fixed factors and interaction for Study 1 (k-anonymiy without generalization).

	k=2		k=3		k=4		k=5	
Variable	**F**	**p value**	**F**	**p value**	**F**	**p value**	**F**	**p value**
Region	1.377	0.239	0.861	0.486	0.221	0.927	1.366	0.243
YHSC	1.307	0.088	0.708	0.876	0.677	0.894	1.031	0.420
YBG	1.595	0.093	3.440	0.002	1.556	0.156	0.225	0.952
Gender	1.877	0.171	0.068	0.794	0.984	0.321	0.002	0.967
Race	1.152	0.330	0.793	0.498	3.306	0.019	0.121	0.886
Father Edu	2.211	0.050	1.726	0.125	1.184	0.314	0.110	0.990
Mother Edu	1.102	0.357	1.272	0.273	0.923	0.465	1.661	0.156
Income	2.218	0.038	0.847	0.533	1.986	0.064	1.544	0.160
YBG * Father Edu	1.043	0.400	2.048	0.012	2.548	0.018	0.456	0.768
Gender * Father Edu	0.862	0.506	0.458	0.767	1.365	0.244	2.925	0.020
Father Edu * Income	1.739	0.008	0.738	0.834	0.903	0.609	0.741	0.813
YHSC * Father Edu	1.007	0.462	0.883	0.719	1.677	0.022	0.802	0.685
Father Edu * Mother Edu	1.770	0.020	1.271	0.217	1.280	0.229	0.475	0.875
Race * Father Edu	0.969	0.488	1.070	0.378	0.801	0.524	0.124	0.946
Region * Father Edu	1.682	0.029	0.655	0.840	0.573	0.888	0.371	0.960
YBG * Gender	0.603	0.796	1.015	0.407	1.702	0.164	0.827	0.437
YBG * Income	1.158	0.201	1.169	0.235	1.069	0.371	0.818	0.667
YHSC * YBG	1.551	0.000	1.345	0.067	0.927	0.528	1.773	0.131
YBG * Mother Edu	1.548	0.023	2.097	0.012	2.623	0.015	1.704	0.146
YBG * Race	2.496	0.001	1.128	0.342	1.783	0.168	0.045	0.956
Region * YBG	5.417	0.000	4.081	0.000	1.671	0.124	1.267	0.275
Gender * Income	2.939	0.007	1.247	0.278	0.843	0.537	0.901	0.493
YHSC * Gender	0.813	0.755	0.670	0.860	1.358	0.165	0.806	0.645
Gender * Mother Edu	1.635	0.147	0.459	0.766	0.383	0.821	0.611	0.655
Gender * Race	1.725	0.141	0.266	0.767	1.900	0.150	0.264	0.608
Region * Gender	2.380	0.049	0.995	0.409	1.406	0.229	1.069	0.370
YHSC * Income	1.039	0.347	0.885	0.816	0.787	0.941	0.855	0.811
Mother Edu * Income	0.978	0.498	0.741	0.830	1.104	0.327	1.116	0.319
Race * Income	1.267	0.172	0.650	0.845	1.263	0.245	1.381	0.199
Region * Income	1.645	0.024	1.049	0.396	1.188	0.239	0.865	0.653
YHSC * Mother Edu	0.650	0.998	0.801	0.855	1.344	0.125	0.813	0.687
YHSC * Race	0.803	0.881	0.532	0.983	0.693	0.844	0.685	0.802
Region * YHSC	2.374	0.000	2.369	0.000	1.270	0.174	3.318	0.000
Race * Mother Edu	2.047	0.005	1.753	0.081	1.656	0.141	0.908	0.475
Region * Mother Edu	1.459	0.084	1.840	0.021	2.350	0.005	1.973	0.046
Region * Race	0.961	0.494	0.682	0.688	1.334	0.254	1.679	0.152

4.3 Study 2 – K-anonymity Sensitivity Analysis With Generalization

Considering that “University id” directly identifies the higher education institution , and “Program id” directly identifies the course inside the institution (the same course has different id's on different institutions, therefore "Program id also identifies the institution), they are no longer considered quasi-identifiers. Also, for the purpose of anonymization, "Region" can be considered a generalization of the institution (University id), after all, a region can contain multiple universities, but a university will always be in one region. The full variable classification can be seen on table 4.7.

Table 4.7: Variables selected for Study 2 (k-anonymity with generalization) and respective classification.

Variable	Classification
University id	Identifying
Program id	Identifying
Region	Quasi-identifier
Age	Quasi-identifier
Gender	Quasi-Identifier
YHSC	Quasi-identifier
YBG	Quasi-identifier
GPA	Insensitive
Race	Quasi-identifier
Mother Edu	Quasi-identifier
Father Edu	Quasi-identifier
Income	Insensitive

In addition, three variables were generalized, namely Age, Mother Edu and Father Edu. Figure 4.1 shows a representation of the applied generalization. Any record with age inferior to 26 is recoded to "<26", and with age superior or equal to 26 is recoded to ">=26". Father and mother education are grouped 2 by 2, meaning that answers A and B are recoded to group "1", C and D to "2" and E and F to "3". This simply groups education levels up to the 5th grade to group "1", between the 6th grade and secondary school to group "2" and any superior graduation to group "3".

For study 2 was considered the generalization level 1, since level 2 would be suppression. Again, the anonymization is conducted by varying K=2, ..., 5.

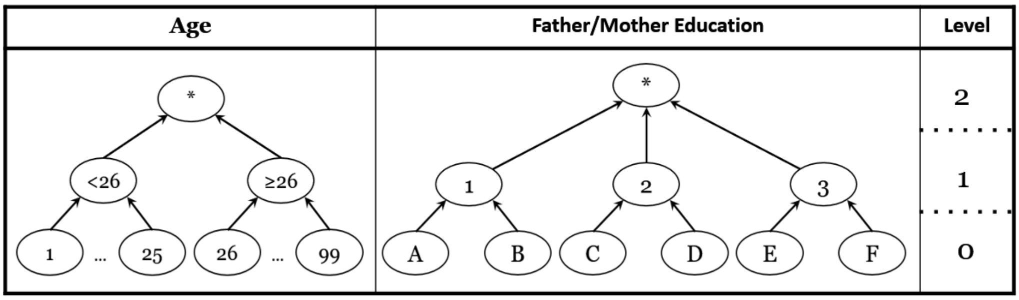


Figure 4.1: Generalization levels of variables Age and Father/Mother Education.

4.3.1 Results - Information Loss, Risk and Data Utility

Table 4.8 presents the number of records in the datasets resulting from the variation of K=2, ..., 5. It also shows the percentage of suppressed records. As it can be seen the suppression percentage was less than 15% in this study.

Table 4.8: Number of records after anonymization and percentage of suppressed records for Study 2 (k-anonymity with generalization).

K = 2	**K = 3**	**K = 4**	**K = 5**
481 447	463 141	448 447	435 4157
4.98%	8.59%	11.49%	14.07%

As table 4.8 shows, as k increases, the suppressed records also increase. However, the number of records suppressed is relatively low, because with a higher level of generalization it becomes easier to form equivalence classes and match k-1 identical records for each record.

Regarding the re-identification risk, again, the maximum risk will always be 100/k. As for the average risk it is calculated similarly to the first study, by obtaining the average risk of the equivalence classes. Table 4.9 presents the maximum and average re-identification risk for k varying k=2, ..., 5. Column k=1 refers to the original dataset. As can be seen the risk declines as k increases, and with generalization it is possible to decrease the risk of re-identification from more than 90% to near 4%. Even for the lowest k, k=2, the average re-identification risk is only 7.03%, decreasing to 3.80% for k=5.

Table 4.9: Maximum and average re-identificationrisk for Study 2 (k-anonymity with generalization).

k = 1	**K = 2**	**K = 3**	**K = 4**	**K = 5**
Maximum risk				
100%	50%	33.3%	25%	20%
Average Risk				
42.23%	7.03%	5.54%	4.42%	3.80%

The results of ANOVA applied to the datasets obtained in Study 2 are presented in table 4.10. Comparing them with the ANOVA results of the original dataset, presented in table 4.2, we quantify that, for k=2, of the 31 factors that were statistically relevant in the original dataset, 27 kept that property. The non-fulfilled being *Race, Father Edu, YBG * Mother Edu, YHSC * Race*, and 1 factors, *YHSC*, gaining statistical relevance not present in the original dataset. Similarly, for k=3, 27 kept statistically relevant, the non-fulfilled being *Father Edu and YBG * Mother Edu, YHSC * Mother Ed and YHSC * Race*. Moreover, for k=4, 25 kept statistically relevant, the non-fulfilled being *Race, Father Edu , YBG * Mother Ed, YHSC * Mother Edu, YHSC * Race and Race * Mother Edu*, and 1 factor, *YHSC*, gaining statistical relevance. Finally, for k=5, 26 terms kept statistically relevant, the non-fulfilled being *Father*

*Edu, YBG * Mother Edu, YHSC * Mother Ed, YHSC * Race and Race * Mother Edu*, and 1 factor, *Mother Edu*, gaining statistical relevance.

Across all the values of k, the ANOVA results are relatively consistent, with mostly the same factors losing their statistical relevance from the original dataset, namely *YBG * Mother Edu, YHSC * Mother Ed and YHSC * Race*. It is worth mentioning that these factors are the ones that, having a p value <= 0.05, have the highest values (0.047, 0.026 and 0.010, respectively).

Table 4.10: ANOVA with fixed factors and interaction for Study 2 (k-anonymity with generalization).

	k=2		**k=3**		**k=4**		**k=5**	
Variable	**F**	**p value**	**F**	**p value**	**F**	**p value**	**F**	**p value**
Region	1.191	0.312	0.304	0.875	0.064	0.993	0.706	0.587
YHSC	1.594	0.005	1.176	0.195	1.442	0.029	1.093	0.314
YBG	3.384	0.000	1.814	0.021	2.991	0.000	2.965	0.000
Gender	0.016	0.899	0.011	0.917	1.672	0.196	2.478	0.115
Race	1.125	0.344	2.853	0.014	2.156	0.056	3.996	0.001
Father Edu	0.817	0.442	1.029	0.357	1.242	0.289	2.027	0.132
Mother Edu	1.508	0.221	2.315	0.099	0.796	0.451	5.725	0.003
Income	4.865	0.000	5.019	0.000	4.685	0.000	4.149	0.000
YBG * Father Edu	2.003	0.001	2.086	0.001	2.132	0.001	2.205	0.001
Gender * Father Edu	10.079	0.000	8.959	0.000	10.026	0.000	10.752	0.000
Father Edu* Income	7.385	0.000	6.989	0.000	6.728	0.000	6.603	0.000
YHSC * Father Edu	1.168	0.139	0.982	0.524	0.873	0.766	0.911	0.672
Father Edu* Mother Edu	49.892	0.000	47.913	0.000	37.678	0.000	34.343	0.000
Race * Father Edu	3.030	0.001	2.914	0.001	2.459	0.006	2.548	0.009
Region * Father Edu	6.429	0.000	7.010	0.000	7.501	0.000	6.735	0.000
YBG * Gender	1.920	0.013	2.414	0.003	2.464	0.003	3.025	0.000
YBG * Income	1.749	0.000	1.765	0.000	1.532	0.001	1.847	0.000
YHSC * YBG	1.771	0.000	1.934	0.000	1.932	0.000	2.017	0.000
YBG * Mother Edu	1.073	0.357	1.109	0.320	1.096	0.340	1.296	0.155
YBG * Race	1.952	0.000	2.125	0.000	2.248	0.000	2.293	0.000
Region * YBG	7.493	0.000	8.745	0.000	9.532	0.000	10.114	0.000
Gender * Income	20.990	0.000	20.565	0.000	19.421	0.000	17.466	0.000
YHSC * Gender	3.030	0.000	3.033	0.000	3.085	0.000	2.972	0.000
Gender * Mother Edu	48.525	0.000	46.805	0.000	41.029	0.000	38.773	0.000
Gender * Race	16.996	0.000	13.724	0.000	11.402	0.000	9.041	0.000
Region * Gender	9.626	0.000	9.810	0.000	9.087	0.000	9.658	0.000
YHSC * Income	2.944	0.000	2.731	0.000	2.629	0.000	2.491	0.000
Mother Edu * Income	9.643	0.000	9.246	0.000	9.143	0.000	9.340	0.000
Race * Income	6.535	0.000	6.069	0.000	5.808	0.000	5.661	0.000
Region * Income	4.061	0.000	3.707	0.000	3.850	0.000	3.620	0.000
YHSC * Mother Edu	1.280	0.046	1.038	0.388	1.011	0.452	1.136	0.222
YHSC * Race	1.106	0.161	1.037	0.367	1.137	0.145	0.993	0.504
Region * YHSC	2.035	0.000	2.105	0.000	2.151	0.000	2.184	0.000
Race * Mother Edu	2.034	0.026	2.039	0.026	1.795	0.056	1.838	0.065
Region * Mother Edu	4.385	0.000	4.575	0.000	4.390	0.000	4.743	0.000
Region * Race	6.429	0.000	6.087	0.000	5.654	0.000	6.111	0.000

4.4 Study 3 – ℓ-diversity

To apply the ℓ-diversity privacy model, it is required, by definition, that at least one variable is classified as sensitive. From the ENADE dataset, *household income* was chosen to be the sensitive attribute, as it might be information that an individual does not want to be disclosed, thus, after k-anonymity being applied, ℓ-diversity will guarantee a diverse set of *household income* values on each equivalence class. All the same attributes as study 2 remain as quasi-identifiers, with identical generalization hierarchies for Age, "Mother's education" and "Father's education". The full variable classification can be seen on table 4.11.

Here the anonymization is conducted by varying ℓ=2, ..., 5 and c=2, 3, 4, on recursive-(c,ℓ)-diversity.

Table 4.11: Study 3 (ℓ-diversity) selected variables and classification.

Variable	Classification
University id	Identifying
Program id	Identifying
Region	Quasi-identifier
Age	Quasi-identifier
Gender	Quasi-identifier
YHSC	Quasi-identifier
YBG	Quasi-identifier
GPA	Insensitive
Race	Quasi-identifier
Mother Edu	Quasi-identifier
Father Edu	Quasi-identifier
Income	Sensitive

4.4.1 Results - Information Loss, Risk and Data Utility

Table 4.12 presents the number of records in the dataset resulting from the variation of ℓ=2, ..., 5 and for each value of ℓ varying c=2, 3, 4, on recursive-(c,ℓ)-diversity. It also shows the percentage of suppressed records. As can be seen, the suppression percentages increased as ℓ increased, and decreased when c increases. Raising the value of c means raising the number of times that the value of the most frequent sensitive attribute may occur in each equivalence class. The exact mathematical formula can be examined in chapter 2 section 2.3.2, where the definition of recursive ℓ-diversity is present.

Table 4.12: Number of records after anonymization and percentage of suppressed records for Study 3 (c,ℓ-diversity).

(2,2)-diversity	**(2,3)-diversity**	**(2,4)-diversity**	**(2,5)-diversity**
404 308	332 420	209 333	99 753
20,20%	34,39%	58,69%	80,31%

(3,2)-diversity	**(3,3)-diversity**	**(3,4)-diversity**	**(3,5)-diversity**
416 925	364 163	266 401	155 406
17,71%	28,13%	47,42%	69,33%

(4,2)-diversity	**(4,3)-diversity**	**(4,4)-diversity**	**(4,5)-diversity**
421 136	377 919	295 827	189 409
16,88%	25,41%	41,61%	62,62%

Regarding the risk, as ℓ-diversity groups the records into groups of ℓ records, much like k, the maximum risk will always be 1/ℓ. As for the average risk it is calculated as mentioned in section 4.2.1. As can be seen in Table 4.13, the average risk decreases as ℓ increases, the change is very much in line with the re-identification risk result's obtained in Study 2. Regarding ”c”, there is a small increase in the risk as ”c” increases, but the change is not noteworthy. This is highly expected as this is just the average re-identification risk from all the equivalence classes formed by the ℓ-diversity privacy model algorithm, i.e., by definition, it is not supposed to decrease the re-identification risk. What it reduces, though, is the attribute disclosure risk. In the studied case, ℓ-diversity assures that the attribute ”*Income*” is well enough diverse for the given (c,ℓ). It is out of the scope of this work to analyse attribute disclosure risk; However, from the ℓ-diversity definition, we can be assured that most of the records in a given equivalence class have a different value for the attribute ”*Income*”, which certainly reduces the risk of disclosure of the ”*Income*” value.

Table 4.13: Maximum and average re-identification risk for Study 3 (c, ℓ)-diversity) .

(2,2)-diversity	**(2,3)-diversity**	**(2,4)-diversity**	**(2,5)-diversity**
6.23%	3.87%	2.65%	1.63%

(3,2)-diversity	**(3,3)-diversity**	**(3,4)-diversity**	**(3,5)-diversity**
6.68%	4.41%	3.08%	2.06%

(4,2)-diversity	**(4,3)-diversity**	**(4,4)-diversity**	**(4,5)-diversity**
6.76%	4.56%	3.25%	2.32%

Maximum risk			
50%	33.3%	25%	20%

The results of ANOVA applied to the datasets obtained in Study 3 are presented in table 4.14 with c =2, table 4.15 with c=3 and table 4.16 with c=4. Comparing the overall results for ℓ-diversity with c=2,3 and 4 , from tables 4.14, 4.15 and 4.16 respectively, with the ANOVA results of the original dataset, shown in table 4.2, we quantify that, for ℓ=2, of the 31 factors that were statistically relevant in the original dataset, invariably 25 of them keep that property, for ℓ=3, between 26 and 27 kept statistically relevant, for ℓ=4, between 23 and 25 kept statistically relevant and finally, for ℓ=5, between 15 and 20 kept statistically relevant. Moreover, across all the values of ℓ, expect for ℓ=5, the ANOVA results are relatively consistent, with mostly the same factors losing their statistical relevance from the original dataset, namely *Race * Father Edu, YBG * Mother Edu, YHSC * Mother Ed and YHSC * Race.* It is noteworthy that these factors are the ones that, having a p value <= 0.05, have the highest values. Finally, we denote that although most of the factors retain their statistical relevance, this property is slightly greater for c = 3 and 4. Nevertheless, we consistently identify a considerable increase in data utility, for all the values of c, as the value of ℓ decreases. Interestingly, the term *Race * Father Edu*, which has statistical significance, loses its statistical relevance for ℓ values 3, 4 and 5, and *YHSC * Mother Edu* keeps it only for ℓ=5.

Table 4.14: ANOVA with fixed factors and interaction for Study 3 (ℓ-diversity) with c=2.

	ℓ=2		ℓ=3		ℓ=4		ℓ=5	
Variable	**F**	**p value**	**F**	**p value**	**F**	**p value**	**F**	**p value**
Region	1.100	0.354	0.660	0.620	0.267	0.899	0.097	0.983
YHSC	1.431	0.026	1.002	0.468	1.121	0.272	1.092	0.325
YBG	2.881	0.000	1.998	0.008	2.182	0.006	2.611	0.002
Gender	0.065	0.798	0.414	0.520	0.868	0.352	1.179	0.278
Race	1.138	0.338	3.532	0.003	2.829	0.015	0.817	0.514
Father Edu	0.175	0.839	0.192	0.825	0.073	0.929	0.047	0.954
Mother Edu	1.957	0.141	9.434	0.000	3.019	0.049	0.240	0.787
Income	3.995	0.001	2.945	0.007	4.741	0.000	2.378	0.027
YBG * Father Edu	1.883	0.002	2.163	0.001	1.694	0.022	1.760	0.038
Gender * Father Edu	9.695	0.000	4.533	0.011	2.543	0.079	3.831	0.022
Father Edu* Income	6.480	0.000	5.114	0.000	3.905	0.000	2.624	0.002
YHSC* Father Edu	1.148	0.167	1.096	0.271	0.929	0.627	0.929	0.604
Father Edu* Mother Edu	40.030	0.000	30.848	0.000	5.814	0.000	1.580	0.177
Race * Father Edu	2.269	0.012	1.549	0.124	1.135	0.335	0.483	0.821
Region * Father Edu	6.423	0.000	4.743	0.000	3.934	0.000	2.102	0.032
YBG * Gender	2.042	0.008	1.892	0.026	1.949	0.025	0.985	0.450
YBG * Income	1.659	0.000	1.644	0.000	1.092	0.271	0.963	0.565
YHSC* YBG	1.743	0.000	1.868	0.000	1.471	0.000	1.205	0.069
YBG * Mother Edu	1.221	0.185	1.036	0.414	0.992	0.470	1.562	0.088
YBG * Race	1.914	0.000	2.077	0.000	2.219	0.000	0.892	0.589
Region * YBG	7.078	0.000	7.162	0.000	4.415	0.000	1.666	0.021
Gender * Income	20.033	0.000	17.933	0.000	16.357	0.000	10.798	0.000
YHSC* Gender	2.970	0.000	2.870	0.000	2.059	0.000	2.363	0.000
Gender * Mother Edu	43.828	0.000	39.875	0.000	27.388	0.000	11.329	0.000
Gender * Race	12.588	0.000	8.378	0.000	4.793	0.001	2.722	0.028
Region * Gender	8.116	0.000	9.516	0.000	8.009	0.000	1.930	0.102
YHSC* Income	2.753	0.000	2.429	0.000	1.654	0.000	1.380	0.000
Mother Edu * Income	9.041	0.000	8.501	0.000	4.605	0.000	1.968	0.023
Race * Income	5.441	0.000	3.841	0.000	1.784	0.007	1.038	0.411
Region * Income	3.569	0.000	2.986	0.000	2.754	0.000	2.125	0.001
YHSC* Mother Edu	1.183	0.125	1.301	0.049	1.118	0.254	1.716	0.003
YHSC* Race	1.028	0.385	1.007	0.460	1.035	0.389	0.749	0.889
Region * YHSC	1.955	0.000	1.731	0.000	1.278	0.024	1.148	0.185
Race * Mother Edu	1.815	0.053	1.172	0.308	0.894	0.520	0.126	0.993
Region * Mother Edu	3.819	0.000	3.425	0.001	3.116	0.002	1.213	0.286
REGION * Race	5.809	0.000	4.841	0.000	1.874	0.018	0.914	0.548

Table 4.15: ANOVA with fixed factors and interaction for Study 3 (ℓ-diversity) with c=3.

	ℓ=2		ℓ=3		ℓ=4		ℓ=5	
Variable	**F**	**p value**	**F**	**p value**	**F**	**p value**	**F**	**p value**
Region	0.829	0.506	0.367	0.832	0.408	0.803	0.384	0.820
YHSC	1.453	0.021	1.375	0.048	1.133	0.254	1.140	0.251
YBG	2.896	0.000	2.260	0.002	2.749	0.000	2.513	0.002
Gender	0.011	0.917	0.077	0.782	2.223	0.136	0.002	0.964
Race	1.096	0.360	3.129	0.008	4.181	0.001	1.780	0.130
Father Edu	0.990	0.372	0.484	0.616	1.703	0.182	0.575	0.562
Mother Edu	2.558	0.077	5.826	0.003	4.257	0.014	0.987	0.373
Income	4.362	0.000	2.571	0.017	4.752	0.000	4.071	0.000
YBG * Father Edu	1.868	0.002	1.861	0.006	1.873	0.008	1.287	0.184
Gender * Father Edu	9.240	0.000	5.940	0.003	3.855	0.021	0.879	0.415
Father Edu* Income	7.152	0.000	5.168	0.000	4.175	0.000	3.085	0.000
YHSC* Father Edu	1.210	0.092	0.971	0.551	1.156	0.193	0.658	0.969
Father Edu* Mother Edu	45.951	0.000	36.634	0.000	14.382	0.000	4.347	0.002
Race * Father Edu	2.753	0.002	1.999	0.035	1.003	0.431	0.603	0.754
Region * Father Edu	6.807	0.000	6.054	0.000	4.132	0.000	3.224	0.001
YBG * Gender	1.951	0.011	1.926	0.023	1.824	0.039	1.053	0.396
YBG * Income	1.680	0.000	1.556	0.001	1.392	0.013	0.878	0.765
YHSC* YBG	1.771	0.000	1.901	0.000	1.584	0.000	1.299	0.007
YBG * Mother Edu	1.008	0.454	0.934	0.555	1.413	0.095	1.027	0.425
YBG * Race	1.884	0.000	2.218	0.000	3.215	0.000	1.793	0.013
Region * YBG	7.144	0.000	8.193	0.000	5.899	0.000	3.487	0.000
Gender * Income	20.120	0.000	19.947	0.000	19.120	0.000	12.509	0.000
YHSC* Gender	3.066	0.000	2.944	0.000	2.079	0.000	1.904	0.001
Gender * Mother Edu	46.790	0.000	40.374	0.000	32.622	0.000	21.770	0.000
Gender * Race	15.302	0.000	9.173	0.000	6.604	0.000	3.489	0.007
Region * Gender	9.372	0.000	9.367	0.000	8.840	0.000	3.390	0.009
YHSC* Income	2.862	0.000	2.509	0.000	1.991	0.000	1.347	0.000
Mother Edu * Income	9.143	0.000	9.357	0.000	6.230	0.000	3.163	0.000
Race * Income	6.107	0.000	4.881	0.000	2.486	0.000	1.367	0.108
Region * Income	3.812	0.000	3.018	0.000	3.236	0.000	2.883	0.000
YHSC* Mother Edu	1.192	0.115	1.093	0.277	1.249	0.093	1.384	0.040
YHSC* Race	1.033	0.369	1.040	0.359	1.124	0.183	0.913	0.675
Region * YHSC	2.104	0.000	1.897	0.000	1.421	0.001	1.382	0.006
Race * Mother Edu	1.778	0.059	1.788	0.065	1.051	0.395	0.213	0.973
Region * Mother Edu	4.258	0.000	3.873	0.000	3.204	0.001	3.020	0.002
REGION * Race	6.360	0.000	4.153	0.000	3.693	0.000	0.941	0.516

Table 4.16: ANOVA with fixed factors and interaction for Study 3 (ℓ-diversity) with c=4.

	ℓ=2		ℓ=3		ℓ=4		ℓ=5	
Variable	**F**	**p value**	**F**	**p value**	**F**	**p value**	**F**	**p value**
Region	0.754	0.555	0.295	0.881	0.340	0.851	0.289	0.886
YHSC	1,440	0.023	1,363	0.053	1,099	0.302	1,417	0.042
YBG	2,872	0.000	2,222	0.003	2,743	0.000	3,026	0.000
Gender	0.018	0.893	0.166	0.684	1,991	0.158	0.140	0.709
Race	0.974	0.432	3,135	0.008	4,399	0.001	3,697	0.005
Father Edu	1,014	0.363	1,455	0.233	1,579	0.206	0.522	0.593
Mother Edu	2,664	0.070	4,472	0.011	7,497	0.001	1,501	0.223
Income	4,232	0.000	2,608	0.016	2,991	0.006	5,101	0.000
YBG * Father Edu	1,925	0.001	1,815	0.009	2,250	0.001	1,331	0.141
Gender * Father Edu	8,986	0.000	7,619	0.000	4,256	0.014	1,819	0.162
Father Edu* Income	7,316	0.000	5,537	0.000	4,911	0.000	3,361	0.000
YHSC* Father Edu	1,180	0.124	0.981	0.525	1,122	0.241	0.929	0.620
Father Edu* Mother Edu	46,065	0.000	39,863	0.000	20,203	0.000	5,653	0.000
Race * Father Edu	2,903	0.001	2,305	0.014	1,731	0.086	0.957	0.468
Region * Father Edu	6,815	0.000	6,567	0.000	5,069	0.000	3,253	0.001
YBG * Gender	1,898	0.014	2,094	0.012	2,441	0.004	1,390	0.169
YBG * Income	1,696	0.000	1,602	0.000	1,449	0.006	1,060	0.340
YHSC* YBG	1,783	0.000	1,932	0.000	1,814	0.000	1,312	0.004
YBG * Mother Edu	0.957	0.534	0.877	0.636	1,364	0.118	0.908	0.581
YBG * Race	1,848	0.000	2,089	0.000	2,979	0.000	2,234	0.001
Region * YBG	7,224	0.000	8,673	0.000	6,942	0.000	4,226	0.000
Gender * Income	20,568	0.000	19,589	0.000	17,927	0.000	14,400	0.000
YHSC* Gender	3,072	0.000	2,999	0.000	2,470	0.000	2,265	0.000
Gender * Mother Edu	47,634	0.000	42,468	0.000	36,435	0.000	23,993	0.000
Gender * Race	15,989	0.000	8,841	0.000	5,374	0.000	4,279	0.002
Region * Gender	9,650	0.000	8,843	0.000	10,102	0.000	6,936	0.000
YHSC* Income	2,890	0.000	2,534	0.000	2,095	0.000	1,576	0.000
Mother Edu * Income	9,334	0.000	8,925	0.000	7,748	0.000	5,134	0.000
Race * Income	6,376	0.000	5,141	0.000	2,935	0.000	1,424	0.082
Region * Income	3,878	0.000	3,059	0.000	3,307	0.000	2,737	0.000
YHSC* Mother Edu	1,187	0.120	1,099	0.266	1,203	0.134	1,595	0.004
YHSC* Race	1,052	0.306	1,059	0.307	1,155	0.130	1,058	0.345
Region * YHSC	2,104	0.000	1,988	0.000	1,568	0.000	1,252	0.039
Race * Mother Edu	1,827	0.051	1,709	0.081	1,202	0.293	0.964	0.462
Region * Mother Edu	4,344	0.000	4,171	0.000	3,352	0.001	2,064	0.036
REGION * Race	6,503	0.000	4,554	0.000	4,135	0.000	1,339	0.163

4.5 Study 4 – t-closeness

To apply the t-closeness privacy model it is also required, by definition, that at least one variable is classified as sensitive. Again, following study 3, *household income* will be the sensitive attribute. The variable classification is the same as the previous study, table 4.11

In this study the anonymization is conducted by varying k=2,5 and for each value of k, t=0.3 and t=0.15, using the Equal Distance as the Earth Mover's distance measure.

4.5.1 Results - Information Loss, Risk and Data Utility

Table 4.17 presents the number of records in the dataset resulting from k=2 and k=5, and for each value of k t=0.3 and t=0.15. As it can be seen when varying k and maintaining t=0.3, there is a relatively slight increase in the number of suppressed records. Although in the extreme case of k=5 and t=0.5 almost 44% of the records were suppressed. This is due to a lower t value, as the model will try to obtain a closer match with the distribution of the attribute *household income* in the original dataset. The value t=0.3 was found to be a good enough compromise between the loss of data and the diversity of *household income* values in each equivalence class.

Table 4.17: Number of records after anonymization and percentage of suppressed records for Study 4 (t-closeness).

k = 2, t = 0.3	**k = 2, t = 0.15**	**k = 5, t = 0.3**	**k = 5 t = 0.15**
402 256	300 986	363 078	283 087
20.61%	40.60%	28.34%	44.13%

Table 4.18: Maximum and average re-identification risk for Study 4 (t-closeness).

<table>
<tr><th>k = 2, t = 0.3</th><th>k = 2, t = 0.15</th><th>k = 5, t = 0.3</th><th>k = 5 t = 0.15</th></tr>
<tr><td colspan="4"><i>Maximum risk</i></td></tr>
<tr><td colspan="2">50%</td><td colspan="2">20%</td></tr>
<tr><td colspan="4"><i>Average risk</i></td></tr>
<tr><td>7.19%</td><td>5.60%</td><td>3.93%</td><td>3.67%</td></tr>
</table>

Regarding the re-identification risk, as k-anonymity was used alongside with t-closeness, it again groups the records into equivalence classes of k records with the maximum risk being 100/k and the average risk being calculated by obtaining the average risk of the equivalence classes. As can be seen in Table 4.18, the average risk is not relevantly different from the Study 2 re-identification risk results. This is very much expected as k-anonymity was used alongside t-closeness, and t-closeness does not significantly change the equivalence classes to affect the risk of re-identification. By definition, t-closeness is not supposed to decrease this kind of risk. What it reduces though, is the attribute disclosure risk. In the studied case, t-closeness assures that for each equivalence class, the attribute "*Income*" variation is in accordance with the original dataset. From the t-closeness definition, we can be assured that most of the records in a given equivalence class have a distribution of the value for the attribute "*Income*" that reassembles the original dataset, which should reduce the risk of disclosure of the "*Income*" value.

Table 4.19 presents the results of applying the ANOVA model to the datasets obtained in Study 4. Observing the results from table 4.19 and comparing them with the ANOVA results of the original dataset, presented in table 4.2, we quantify that, for k=2 and t=0.3, of the 31 factors that were statistically relevant in the original dataset, 26 of them keep that property, for k=2 and t=0.015, 22 kept statistically relevant, for k=5 and t=0.3, 26 kept statistically relevant and finally, for k=5 and t=0.15, 24 kept statistically relevant. Across all the values of k and t, the ANOVA results are relatively consistent, with mostly the same factors losing their statistical relevance from the original dataset, namely *YBG * Mother Edu, YHSC * Mother Ed and YHSC * Race*. It is worth mentioning that these factors are part of the ones with a statistical significance greater than 0, with 0.047, 0.026 and 0.010, respectively. Furthermore, the factor *Race * Father Edu*, which also has a statistical significance greater than 0, only loses this property when t=0.015.

Table 4.19: ANOVA with fixed factors and interaction for Study 4 (t-closeness).

	k = 2. t = 0.3		**k = 2. t = 0.15**		**k = 5. t = 0.3**		**k = 5 t = 0.15**	
Variable	**F**	**p value**	**F**	**p value**	**F**	**p value**	**F**	**p value**
Region	1.725	0.141	0.517	0.723	0.729	0.572	0.426	0.790
YHSC	1.180	0.184	0.671	0.956	1.070	0.351	1.049	0.385
YBG	4.036	0.000	3.136	0.000	3.085	0.000	3.210	0.000
Gender	0.008	0.929	0.007	0.932	1.300	0.254	0.663	0.416
Race	0.473	0.797	1.058	0.381	4.583	0.000	2.681	0.020
Father Edu	0.232	0.793	0.619	0.539	2.252	0.105	0.689	0.502
Mother Edu	3.589	0.028	1.686	0.185	5.944	0.003	5.944	0.003
Income	5.093	0.000	3.367	0.003	3.015	0.006	2.179	0.042
YBG* Father Edu	1.872	0.002	1.373	0.094	1.786	0.012	1.706	0.025
Gender * Father Edu	10.622	0.000	4.642	0.010	11.322	0.000	4.973	0.007
Father Edu * Income	7.957	0.000	4.107	0.000	6.994	0.000	3.427	0.000
YHSC * Father Edu	0.933	0.650	1.049	0.362	0.858	0.776	1.081	0.321
Father Edu * Mother Edu	47.724	0.000	31.747	0.000	32.705	0.000	24.871	0.000
Race * Father Edu	2.675	0.003	1.598	0.100	2.320	0.017	1.828	0.067
Region * Father Edu	5.087	0.000	3.156	0.001	5.386	0.000	3.477	0.001
YBG* Gender	2.068	0.006	3.463	0.000	2.853	0.001	4.041	0.000
YBG* Income	1.502	0.001	1.405	0.007	1.643	0.000	1.456	0.007
YHSC * YBG	1.844	0.000	1.740	0.000	2.102	0.000	1.957	0.000
YBG* Mother Edu	0.956	0.534	0.882	0.641	1.266	0.176	1.431	0.091
YBG* Race	1.897	0.000	1.857	0.000	2.265	0.000	2.491	0.000
Region * YBG	7.387	0.000	7.571	0.000	9.892	0.000	9.150	0.000
Gender * Income	17.684	0.000	9.901	0.000	15.434	0.000	8.913	0.000
YHSC * Gender	2.984	0.000	2.393	0.000	3.008	0.000	2.238	0.000
Gender * Mother Edu	46.709	0.000	24.206	0.000	39.954	0.000	22.141	0.000
Gender * Race	17.172	0.000	10.923	0.000	8.055	0.000	6.912	0.000
Region * Gender	9.972	0.000	8.236	0.000	10.055	0.000	8.320	0.000
YHSC * Income	2.863	0.000	2.204	0.000	2.506	0.000	2.056	0.000
Mother Education * Income	10.057	0.000	7.114	0.000	9.434	0.000	6.973	0.000
Race * Income	6.304	0.000	4.694	0.000	5.864	0.000	4.409	0.000
Region * Income	2.445	0.000	0.955	0.525	2.027	0.002	0.909	0.590
YHSC * Mother Edu	1.051	0.356	1.159	0.165	1.165	0.186	1.234	0.120
YHSC * Race	1.051	0.310	0.870	0.875	0.997	0.492	0.927	0.695
Region * YHSC	2.082	0.000	1.940	0.000	2.241	0.000	1.936	0.000
Race * Mother Edu	2.286	0.011	1.525	0.123	1.532	0.140	0.535	0.831
Region * Mother Edu	4.336	0.000	2.992	0.002	4.754	0.000	3.338	0.001
REGION * Race	6.145	0.000	4.927	0.000	6.372	0.000	5.017	0.000

4.6 Conclusion

This chapter described the ENADE dataset and its contents. The adequacy of the data to the study was promptly discussed, concluding it to be a suitable dataset for the proposed study, as being composed of real data from a number of commonly used attributes in the field of anonymization. Each attribute used in the study was then classified accordingly. The first study proposes a simple anonymization of the dataset with k-anonymity, by varying k from 2 to 5. The second study adds a technique of generalization with the intent of assessing the benefits with the usage of the said technique and originating an anonymized dataset with a reasonable dimension. The third and fourth study apply more complex models of anonymization ℓ-diversity, and t-closeness, which are improvements to the k-anonymity procedure, namely to protect against attribute disclosure. For each study we evaluate the information loss by quantifying the number of records suppressed after the anonymization process and we present the re-identification risk. Finally, we assess data utility by applying the ANOVA model to each dataset and analyse the results.

Chapter 5

Discussion

In this chapter we will analyze the results obtained and presented in chapter 4. The objectives with these studies were to assess and compare the results of data loss, re-identification risk and data utility, when applying the privacy models k-anonymity, ℓ-diversity and t-closeness to a real world dataset. Starting with the side by side analysis of studies 1 and 2, figure 5.1 presents a graphic comparing the percentage of suppressed records for Study 1 (k-anonymity without generalization) and Study 2 (k-anonymity with generalization), for each value of k=2,...,5.

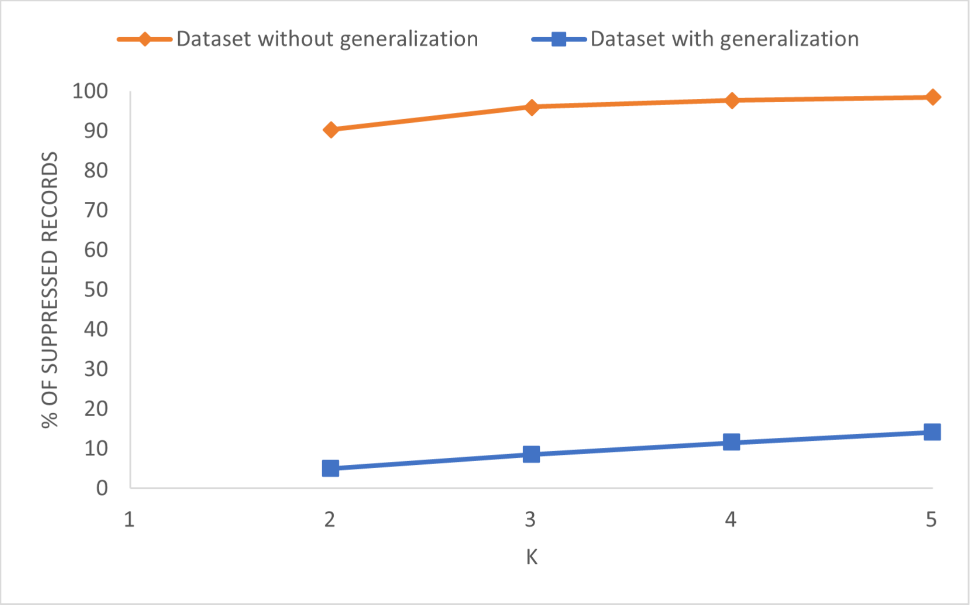


Figure 5.1: Percentage of suppressed records for Study 1 (k-anonymity without generalization) and Study 2 (k-anonymity with generalization).

Regarding the re-identification risk, and as previously mentioned, k-anonymity groups the records into equivalence classes of k records, thus, the maximum risk will always be 100/K, so it is the same for both studies. Figure 5.2 depicts the percentage of maximum and average re-identification risk for Study 1 (k-anonymity without generalization) and Study 2 (k-anonymity with generalization), as can be seen, the re-identification risk becomes substantially lower with generalization, for the same k. This can be explained by a larger number of records having the same attribute values post generalization, lowering the average risk considerably.

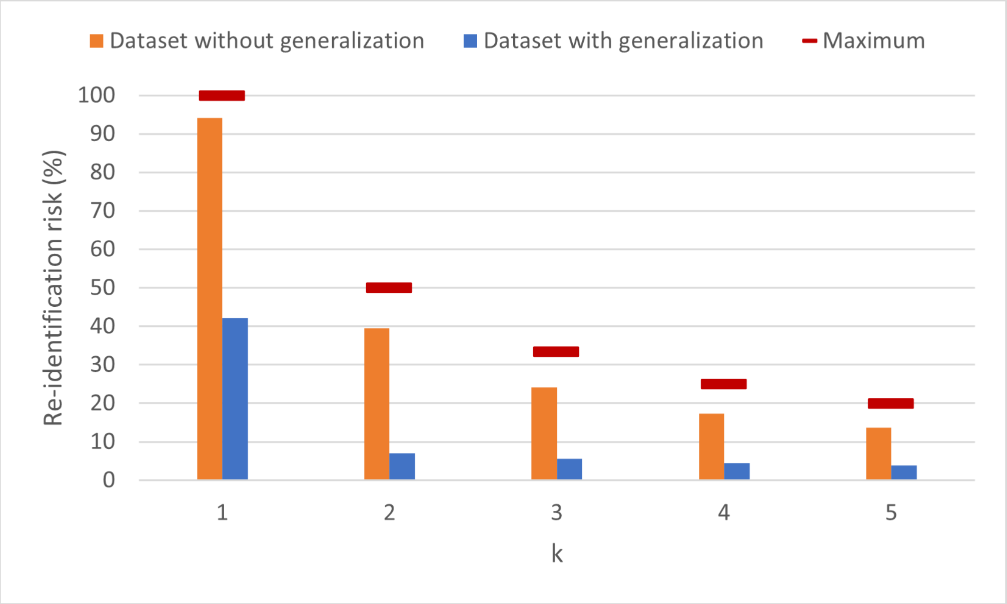


Figure 5.2: Percentage of maximum and average re-identification risk for Study 1 (k-anonymity without generalization) and Study 2 (k-anonymity with generalization).

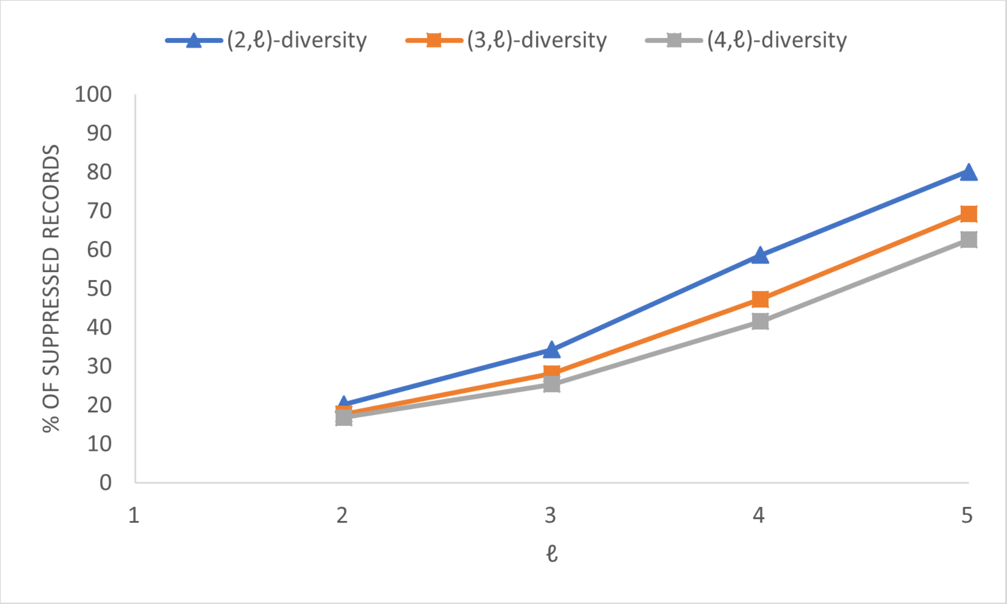


Figure 5.3: Percentage of suppressed records for Study 3 (ℓ-diversity).

Figure 5.3 presents a graphic comparing the percentage of suppressed records for the values of c=2, 3, 4 when varying ℓ=2,...,5. The percentage of suppressed records increases greatly when increasing the value of ℓ. Consequently, when reaching ℓ=5 the percentage of suppressed records consistently surpasses 62%, resulting in a heavy loss of data. Inspecting the repercussion of ”c”, the data confirms that the suppression decreased when c increases; however, there seems to be a slight reduction in the loss when ”c” increases from 3 to 4, when comparing with the increment of c from 2 to 3. With this result we can conclude that a recursive-(c=2,ℓ)-diversity leads to a severe loss, particularly for ℓ=5 where the highest percentage of suppressed records was obtained in this study. Looking at the graphic in 5.3, we can easily compare each recursive-(c,ℓ) result with the equivalent k-anonymity result in figure 5.1.

Now comparing the information loss results for Study 3 (ℓ-diversity) with the equivalent k-anonymity in figure 5.1, the minimum loss percentage obtained in k-anonymity for k=2 (4.98%), with the minimum loss percentage obtained with recursive-(c,ℓ)-diversity, (4,2)-diversity, the percentage of suppressed records increment by a factor of 3.39. Furthermore, even comparing the highest loss percentage obtained in k-anonymity (14.07%), with the lowest loss percentage obtained in recursive-(c,ℓ)-diversity (c=4, ℓ=2, 16.88%), we can denote a slightly lower loss percentage with k-anonymity. On the other hand, for values of ℓ greater than 2 the loss percentage increases at a higher rate than the k counterpart. This might be explained by the fact that there are only seven different classifications for the sensitive variable *Income*.

Figure 5.4 depicts the percentage of suppressed records for Study 4 (t-closeness) for t=0.3 when varying k=2,5, and for t=0.15 when varying k=2,5. The results manifest an almost insignificant increase in the percentage of suppressed records when increasing k. This is noticeable if compared to the results of ℓ-diversity presented in figure 5.3, where an increment in the value of ℓ caused a noticeable increase in the percentage of suppressed records, a situation not manifested here with t-closeness. Instead, t is the determinant factor for information loss.

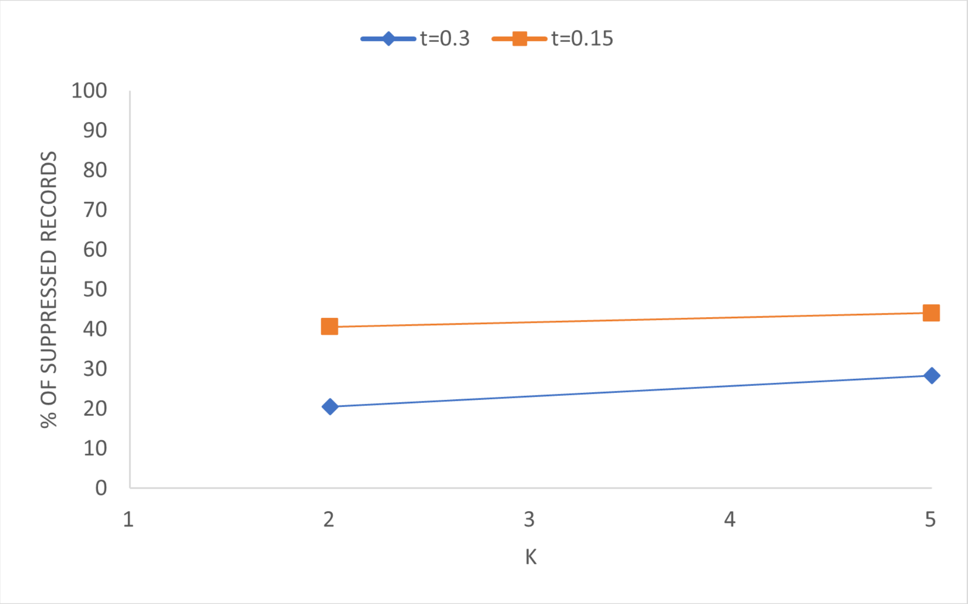


Figure 5.4: Percentage of suppressed records for Study 4 (t-closeness).

Analyzing the overall information loss results we can conclude that K-anonymity privacy model provides a basic anonymization method as it suppresses the most vulnerable individuals, protecting them from re-identification. However, it does not protect against attribute disclosure. In the studied dataset, after k-anonymity was applied, a record could no longer be linked to a specific individual; however, it could disclose the *Income* value if the other records on the same equivalence class had equal or similar *Income* value. This is precisely, as mentioned before, attribute disclosure. This is where ℓ-diversity and t-closeness are useful, as they will ensure a diverse set of values for the sensitive attribute *Income*. Nevertheless, preventing attribute disclosure comes with a cost, more loss of data, as it was shown with the results obtained in Study 3 (ℓ-diversity) and Study 4 (t-closeness). We can also conclude that with ℓ-diversity we document a higher percentage of suppressed records when comparing it with t-closeness. This might be explained by the composition of the dataset itself, as it has significantly less records with a higher *Income* value and thus ℓ-diversity will have difficulty obtaining a diverse enough result from a not very diverse attribute. As for t-closeness, since it is based on the distribution of the original dataset, it will not be as heavily affected. Furthermore, t-closeness will not only guarantee a diverse distribution of the sensitive attribute, but also the closeness of the *Income* values with the original distribution, extending the attribute disclosure protection when compared with ℓ—diversity.

Finally, regarding data utility, we came to the conclusion that study 1 (k-anonymity without generalization) definitively provided the results with the worst data utility preservation. This corroborates the results obtained with the number of suppressed records, since with more records being suppressed that tends to cause more data being lost. The graphic in figure 5.5 illustrates, for each study, the average number of statistically relevant factors and interactions which retained that property following anonymization. As can be seen, with the support from the ANOVA results, Study 2(k-anonymity with generalization) provides the best data utility preservation with an average of 26.25 of the 31 factors retaining their statistical relevance, followed by Study 4 (t-closeness) with 24.5 and Study 3 (ℓ-diversity) with 23.42. We emphasize that, even though, in the case of Study 3 (ℓ-diversity) the number of suppressed records reaches 80%, nearly the same as study 1 (k-anonymity without generalization) with more than 90%, the data utility levels managed to stay very close to the ones obtained from study 2 (k-anonymity with generalization) and study 4 (t-closeness). That is, l-diversity seems to allow the preservation of utility even when there is a high percentage of suppressed records.

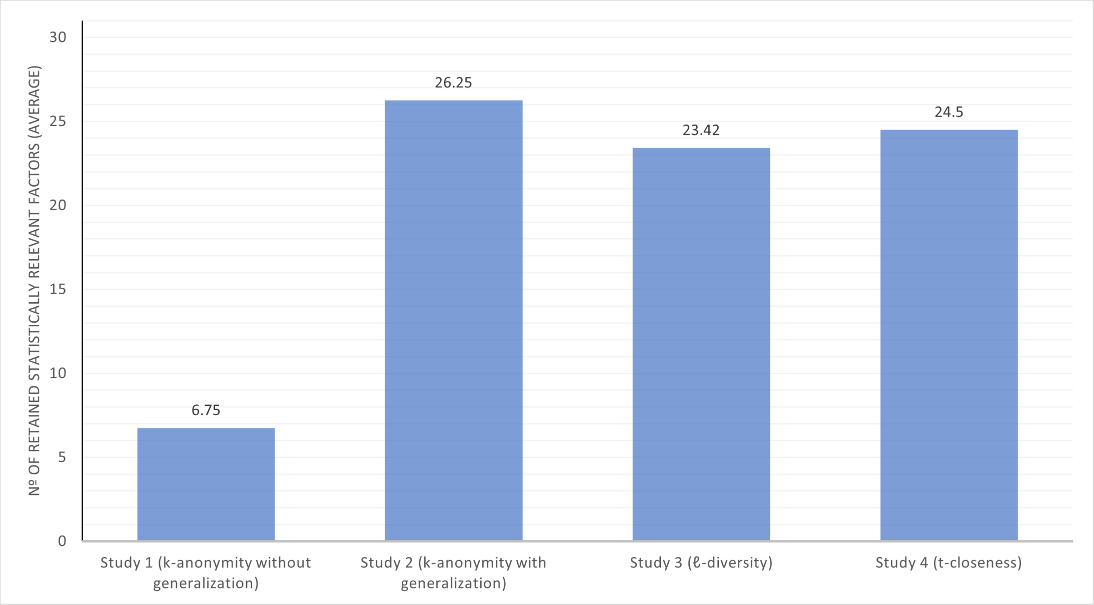


Figure 5.5: Average number of retained statistically relevant factors for each study, with a maximum of 31.

Comparing the data utility results between the studies, we assess that the variables with a *P value* of 0.000 mostly retain they statistical relevance up to a level of significance of 5% (p value < 0.05) in the anonymized datasets, with the particular exception of the factor *Father Edu* which loses this property in every studied case except Study 1 k=2. Another factor worth of note is *Race* that, despite having a level of significance equal to 0.017 in the original dataset, it fails to retain a level of significance in the distinct case of the value 2 of k and ℓ. Furthermore, the two factor interactions that persistently lose their statistical significance are *YBG * Mother Edu, YHSC * Race.* Finally, the only two factors for which we cannot reject the null hypothesis are *Region and Gender*, with *YHSC * Father Edu* also worth mentioning since this is only not the case in two occurrences.

5.1 Conclusion

In this chapter we comprehensively discuss the results obtained throughout the studies, namely regarding information loss, re-identification risk and data utility. The analyzes are supported with a thorough comparison of the results obtained between the studies. In summary, the results corroborate each other, especially the ones obtained from the loss of records and data utility, which the numbers seem to match where a high loss of records translates to a lower data utility, even if unproportional. The results conclude that it is difficult to achieve a reasonable trade-off between privacy and data utility, corroborating the findings present in the previously cited papers [54] and [55].

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this work, our first goal was to study various anonymization techniques with regard to how their application affected: 1) the loss of data; 2) the re-identification risk; 3) data utility.

To achieve such goals, we conducted an extensive and detailed overview of the terminology and methods used in the anonymization process. Doing so, we started by reviewing some academical work highlighting the difficulty in determining whether the anonymization process was effective. Moreover, we concluded that most of the works utilize or are based in the k-anonymity privacy model and also take advantage of statistical based models to assess data utility.

We applied the privacy models k-anonymity, ℓ-diversity and t-closeness to a publicly available and real world dataset containing real data from attributes commonly subject to anonymization. Once the anonymization process was completed, we proceeded to quantify the data loss, i.e. the number of records that were suppressed. Then, we identified the risk of re-identification by comparing the maximum and average risk. Finally, we assessed the data utility, and how much was preserved, by comparing the results between before and after anonymization from the ANOVA statistical model.

In conclusion, the result obtained showed an increase in data loss proportional to the increase of the value k, for k-anonymity and t-closeness, and the increase of the value ℓ for ℓ-diversity. In addition, the application of generalization to designated attributes considerably contributed to a sharply reduction in data loss. Furthermore, the re-identification risk consistently decreased as the values of k and ℓ increased. Finally, the data utility results reveal a considerable deterioration of data utility in the cases where the number of suppressed records surpasses 90%, although, in the case of ℓ-diversity, and despite the number of suppressed records reaching 80%, the data utility levels managed to stay very close to the ones obtained from k-anonymity and t-closeness.

6.2 Future Work

Even though this is outside the scope of this work, the same method could be applied to different privacy models. In addition, this work can be extended to include even more attributes as quasi-identifiers and consequent generalizations, although this will certainly increase the complexity of the analyses especially the data utility assessment. Furthermore, this work could be extended by adding different risk and data utility assessment models, namely a method to identify the attribute disclosure risk.

Bibliography

[1] F. Prasser, F. Kohlmayer, R. Lautenschlaeger, and K. Kuhn, “Arx–a comprehensive tool for anonymizing biomedical data,” *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2014, pp. 984–93, 11 2014. xiii, 7

[2] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, “Highly efficient optimal k-anonymity for biomedical datasets,” in *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2012, pp. 1–6. xiii, 8

[3] L. Sweeney, “K-anonymity: A model for protecting privacy,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, p. 557–570, Oct. 2002, World Scientific Publishing Co., Inc. [Online]. Available: https://doi.org/10.1142/S0218488502001648 xiii, 9, 10, 22

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “L-diversity: Privacy beyond k-anonymity,” *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3–es, Mar. 2007, in Association for Computing Machinery. [Online]. Available: https://doi.org/10.1145/1217299.1217302 xiii, 9, 10, 11, 12, 22

[5] F. Prasser, F. Kohlmayer, H. Spengler, and K. A. Kuhn, “A scalable and pragmatic method for the safe sharing of high-quality health data,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 611–622, 2018. xiii, 19

[6] ARX Data Anonymization Tool Guide. [Online]. Available: https://arx.deidentifier.org/anonymization-tool/ xiii, 21

[7] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, “Flexible data anonymization using arx—current status and challenges ahead,” *Software: Practice and Experience*, vol. 50, no. 7, pp. 1277–1304, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2812 xv, 9, 13

[8] G. Adomavicius and hilin, “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005. 1

[9] H.-c. Chen, R. Chiang, and V. Storey, “Business intelligence and analytics: From big data to big impact,” *MIS Quarterly*, vol. 36, pp. 1165–1188, 12 2012, doi doi:10.2307/41703503. 1

[10] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, p. 571–588, Oct. 2002, World Scientific Publishing Co., Inc. [Online]. Available: https://doi.org/10.1142/S021848850200165X 1, 10

[11] B. C. Fung, K. Wanga, A. W.-C. Fu, and P. S. Yu, “Introduction to privacy-preserving data publishing: Concepts and techniques,” *Chapman & Hall/CRC*, Aug. 2010. 1, 7

[12] L. Sweeney, “Simple demographics often identify people uniquely,” *Carnegie Mellon University*, Jun 2018. [Online]. Available: https://doi.org/10.1184/R1/6625769.v1 1

[13] Panduragan. (2014) On Taxis and Rainbows . [Online]. Available: https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1 1

[14] L. Sweeney, “Only you, your doctor, and many others may know,” *Technology Science*, 2015. [Online]. Available: https://techscience.org/a/2015092903/ 1

[15] Thomas Brewster. (2017) 120 Million American Households Exposed In ’Massive’ ConsumerView Database Leak. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2017/12/19/120m-american-households-exposed-in-massive-consumerview-database-leak/#5e210aaf7961 1

[16] Richie Koch. (2020) Political campaigns and your personal data. [Online]. Available: https://protonmail.com/blog/political-campaigns-and-your-personal-data/ 1

[17] C. Culnane, B. I. P. Rubinstein, and V. Teague, “Health data in an open world,” arXiv 1712.05627, 2017. 1

[18] A. Narayanan and E. W. Felten, “No silver bullet: De-identification still doesn’t work,” 2014. 1

[19] “Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance),” 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj 2

[20] “Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data,” 1995. [Online]. Available: http://data.europa.eu/eli/dir/1995/46/oj 2

[21] J. Kaye, E. Meslin, B. Knoppers, E. Juengst, M. Deschênes, A. Cambon-Thomsen, D. Chalmers, J. Vries, K. Edwards, N. Hoppe, A. Kent, C. Adebamowo, P. Marshall, and K. Kato, “Elsi 2.0 for genomics and society,” *Science*, vol. 336, pp. 673–674, 05 2012, doi:10.2307/41584779. 2

[22] A. Cambon-Thomsen, E. Rial-Sebbag, and B. M. Knoppers, “Trends in ethical and legal frameworks for the use of human biobanks,” *European Respiratory Journal*, vol. 30, no. 2, pp. 373–382, 2007. [Online]. Available: https://erj.ersjournals.com/content/30/2/373 2

[23] L. Sweeney, M. von Loewenfeldt, , and M. Perry, “Saying it’s anonymous doesn’t make it so: Re-identifications of “anonymized” law school data,” 2018. 2, 15

[24] P. Francis. (2018) Can Anonymized Data Still be Useful? Part Deux. [Online]. Available: https://aircloak.com/can-anonymized-data-still-be-useful-part-deux/ 2

[25] Recital 26 - General Data Protection Regulation. [Online]. Available: https://gdpr-info.eu/recitals/no-26/ 5

[26] Article 29 Working Party - General Data Protection Regulation. [Online]. Available: https://ec.europa.eu/newsroom/article29/news-overview.cfm 5

[27] Data anonymization and GDPR compliance: the case of Taxa 4×35. [Online]. Available: https://gdpr.eu/data-anonymization-taxa-4x35/ 5

[28] Tilsyn med Taxa 4x35’s behandling af personoplysninger. [Online]. Available: https://www.datatilsynet.dk/tilsyn-og-afgoerelser/afgoerelser/2019/mar/tilsyn-med-taxa-4x35s-behandling-af-personoplysninger 5

[29] General Data Protection Regulation - European Union. [Online]. Available: https://gdpr.eu 6

[30] General Data Protection Regulation - Checklist - 2018. [Online]. Available: https://gdpr.eu/checklist/ 6

[31] L. H. Cox, “Suppression methodology and statistical disclosure control,” *Journal of the American Statistical Association*, vol. 75, no. 370, pp. 377–385, 1980. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1980.10477481 6

[32] T. Dalenius, “Finding a needle in a haystack - or identifying anonymous census record.” *Journal of Official Statistics*, p. 329–336, 1986. 6

[33] Article 4, General Data Protection Regulation - European Union, ”Definitions”. [Online]. Available: https://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm 6

[34] Genralization Hiearchies. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/generalization-hierarchy 7

[35] P. Samarati, “Protecting respondents identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001. 7

[36] ARX Data Anonymization Tool . [Online]. Available: https://arx.deidentifier.org/ 8, 20, 21

[37] M. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases,” 01 2007, pp. 665–676, doi:10.1145/1247480.1247554. 9, 22

[38] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115. 9, 12, 22

[39] S. V. Brickell J, "The cost of privacy: destruction of data□mining utility in anonymized data publishing," *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, pp. 70–78, 2008. 9, 22

[40] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *CoRR*, vol. abs/1208.0220, 2012. [Online]. Available: http://arxiv.org/abs/1208.0220 9, 22

[41] R. Bild, K. Kuhn, and F. Prasser, "Safepub: A truthful data anonymization algorithm with strong privacy guarantees," vol. 2018, 01 2018, doi:10.1515/popets-2018-0004. 9, 22

[42] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," *Trans. Data Privacy*, vol. 7, no. 3, p. 337–370, Dec. 2014, IIIA-CSIC. 10, 20

[43] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity." Proceedings of 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD'05), 2005, p. 49–60. 10

[44] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 25–25. 10, 13

[45] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2004. 12

[46] A. Gionis and T. Tassa, "k-anonymization with minimal loss of information," *IEEE Trans. Knowl. Data Eng.*, vol. 21, pp. 206–219, 02 2009. 13

[47] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," vol. 2010, 05 2005, pp. 217– 228. 13

[48] H. Scheffé, *The analysis of variance*. New York and London: Wiley, 1999. 13

[49] Ibm spss software. [Online]. Available: https://www.ibm.com/analytics/spss-statistics-software 13

[50] L. Rocher, J. Hendrickx, and Y.-A. Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications*, vol. 10, 12 2019. 16

[51] G. Simon, S. Shortreed, R. Coley, R. Penfold, R. Rossom, B. Waitzfelder, K. Sanchez, and F. Lynch, "Assessing and minimizing re-identification risk in research data derived from health care records," *EGEMS (Washington, DC)*, vol. 7, p. 6, 03 2019. 16

[52] K. Rajendran, M. Jayabalan, and M. E. Rana, “A study on k-anonymity, l-diversity, and t-closeness techniques focusing medical data,” vol. 17, 12 2017. 16

[53] S. Jang, “A study of performance enhancement in big data anonymization,” in *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, 2017, pp. 1–4. 16

[54] L. G. Esquivel-Quirós, E. G. Barrantes, and F. E. Darlington, “Marco de medición de la privacidad,” *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, pp. 66 – 81, 03 2019. [Online]. Available: http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S1646-98952019000100006&nrm=iso 17, 47

[55] H. Spengler and F. Prasser, “Protecting biomedical data against attribute disclosure,” *Studies in health technology and informatics*, vol. 267, pp. 207–214, 09 2019, doi:10.3233/SHTI190829. 17, 47

[56] K. E. Emam and L. Arbuckle, *Anonymizing health data: case studies and methods to get you started*. O’Reilly, 2013. 17

[57] K. E. Emam and B. Malin, *Concepts and methods for de-identifying clinical trial data*. Trial Data Maximizing Benefits, Minimizing Risk pp. 1–290, 2015. 17

[58] A. Basu, T. Nakamura, S. Hidano, and S. Kiyomoto, “k-anonymity: Risks and the reality,” in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1, 2015, pp. 983–989. 17

[59] S. Zhang, X. Li, Z. Tan, T. Peng, and G. Wang, “A caching and spatial k-anonymity driven privacy enhancement scheme in continuous location-based services,” *Future Generation Computer Systems*, vol. 94, pp. 40 – 50, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17321398 17

[60] Q. Zhao and Y. Liu, “E-voting scheme using secret sharing and k-anonymity,” 01 2017, pp. 893–900. 17

[61] J. Schwee, F. Sangogboye, and M. Kjærgaard, “Evaluating practical privacy attacks for building data anonymized by standard methods,” 04 2019. 17

[62] O. Gkountouna, S. Angeli, A. Zigomitros, M. Terrovitis, and Y. Vassiliou, “km-anonymity for continuous data using dynamic hierarchies,” 09 2014, pp. 156–169, $doi : 10.1007/978 - 3 - 319 - 11257 - 2_13$. 17

[63] G. Acs, J. P. Achara, and C. Castelluccia, “Probabilistic km-anonymity efficient anonymization of large set-valued datasets,” in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 1164–1173. 18

[64] M. Terrovitis, N. Mamoulis, and P. Kalnis, “Privacypreserving anonymization of set-valued data,” *PVLDB*, vol. 1, pp. 115–125, 08 2008, doi10.14778/1453856.1453874. 18

[65] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, “Coat: Constraint-based anonymization of transactions,” *Knowledge and Information Systems*, vol. 28, pp. 251–282, 08 2011, doi:10.1007/s10115-010-0354-4. 18

[66] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis, “Distance-based km-anonymization of trajectory data,” in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 2, 2013, pp. 57–62. 18

[67] P. Liu, Y. Bai, L. Wang, and X. Li, “Partial k-anonymity for privacy-preserving social network data publishing,” *International Journal of Software Engineering and Knowledge Engineering*, vol. 27, p. 1750004, 11 2016. 18

[68] X. Ying, K. Pan, X. Wu, and L. Guo, “Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing,” vol. 10, 01 2009, p. 10. 18

[69] D. Avraam, A. Boyd, H. Goldstein, and P. Burton, “A software package for the application of probabilistic anonymisation to sensitive individual-level data: A proof of principle with an example from the alspac birth cohort study,” *Longitudinal and Life Course Studies*, vol. 9, pp. 433–446, 10 2018, doi:10.14301/llcs.v9i4.478. 18, 19

[70] H. Goldstein and N. Shlomo, “A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets,” *Journal of Official Statistics*, vol. 36, pp. 89–115, 03 2020, doi:10.2478/jos-2020-0005. 18

[71] S. Shaham, M. Ding, B. Liu, Z. Lin, and J. Li, “Machine learning aided anonymization of spatiotemporal trajectory datasets,” in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 1–6. 19

[72] R. Mendes and J. Vilela, “Privacy-preserving data mining: Methods, metrics and applications,” *IEEE Access*, vol. PP, pp. 1–1, 06 2017. 19

[73] J. Yoon, L. N. Drumright, and M. van der Schaar, “Anonymization through data synthesis using generative adversarial networks (ads-gan),” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2378–2388, 2020. 19

[74] D. Patil, R. K. Mohapatra, and K. S. Babu, “Evaluation of generalization based k-anonymization algorithms,” in *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 2017, pp. 171–175. 20

[75] K. Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley, “A globally optimal k-anonymity method for the de-identification of health data,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 16, pp. 670–82, 07 2009, doi:10.1197/jamia.M3144. 20

[76] R. Bild, J. Eicher, and F. Prasser, “Efficient protection of health data from sensitive attribute disclosure,” *Studies in health technology and informatics*, vol. 270, pp. 193–197, 06 2020, doi:10.3233/SHTI200149. 20

[77] ARX Open Source repository on GitHub. [Online]. Available: https://github.com/arx-deidentifier/arx/ 21

[78] Amnesia. [Online]. Available: https://amnesia.openaire.eu/amnesiaInfo.html 23

[79] M. Templ, *Statistical Disclosure Control for Microdata. Methods and Applications in R.*, 05 2017, doi:10.1007/978-3-319-50272-4. 23

[80] Aircloak Insights. [Online]. Available: https://aircloak.com/solutions/features-en/ 23

[81] UTD Anonymization Toolbox. [Online]. Available: http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php 23

[82] ENADE - Exame Nacional de Desempenho dos Estudantes. [Online]. Available: http://portal.inep.gov.br/enade 25

Appendix A

Data Anonymization: K-anonymity Sensitivity Analysis

Data Anonymization: K-anonymity Sensitivity Analysis

Wilson Santos[a,b] , Gonçalo Sousa[a], Paula Prata [a,b], Maria Eugénia Ferrão[a,c]
[a] Universidade da Beira Interior, Covilhã, Portugal
[b] Instituto de Telecomunicações
[c] Centro de Matemática Aplicada à Previsão e Decisão Económica, Lisboa, Portugal.
{wilsongdsantos, goncalosousa291}@gmail.com, pprata@di.ubi.pt, meferrao@ubi.pt

Abstract — **These days the digitization process is everywhere, spreading also across central governments and local authorities. It is hoped that, using open government data for scientific research purposes, the public good and social justice might be enhanced. Taking into account the European General Data Protection Regulation recently adopted, the big challenge in Portugal and other European countries, is how to provide the right balance between personal data privacy and data value for research. This work presents a sensitivity study of data anonymization procedure applied to a real open government data available from the Brazilian higher education evaluation system. The ARX k-anonymization algorithm, with and without generalization of some research value variables, was performed. The analysis of the amount of data / information lost and the risk of re-identification suggest that the anonymization process may lead to the under-representation of minorities and sociodemographic disadvantaged groups. It will enable scientists to improve the balance among risk, data usability, and contributions for the public good policies and practices.**

Keywords - GDPR; personal data protection; ARX; data anonimization; k-anonymity.

I. Introduction

A. Motivation

The data produced by day-to-day human activity have increasing social and economic value for companies and organizations to assess and guide their behaviors and actions. However, the use of such data must respect the privacy of each individual. With the emergence of the European General Data Protection Regulation (GDPR) [1] together with the growth of digitization in every area, data anonymization has become an essential topic in data processing and analysis.

Never before have people generated and recorded so much data. With such a wealth of information it becomes easy to cross several data sources. Sometimes, data that are believed to be anonymous, may however be vulnerable to re-identification as shown in [2]. The authors of that work were able to put real names to the records produced by four protocols that were referred to as being popular ways to make personal information anonymous. Therefore, finding a right balance between data utility and personal privacy is an open issue.

B. Related Work

In the anonymization process it is supposed to identify all the attributes that could be used for linking with external information. Such attributes include all direct identifiers, as name, or social security number, and also indirect or quasi-identifiers. A quasi-identifier is an attribute that linked with other dataset can uniquely identify an individual. The first formal model proposed for microdata anonymization, the k-anonymity model, consists of modifying the quasi-identifiers in order to avoid any data linkage. Sweeney and Samarati define k-anonymity as follows [3] [4]: “Let T(A1,...,An) be a table and QI be the quasi-identifier associated with it. T is said to satisfy k-anonymity wrt QI if and only if each sequence of values in T[QI] appears at least k occurrences in T[QI]” ([4] p. 1013). Several algorithms to implement k-anonymity have been developed [5]. Most of them actuate on quasi-identifier attributes through generalization and suppression operations, in order to create groups of records that share the same quasi-identifier values. Suppression consists in replacing original data by some special value, as for instance “*”. Generalization (also called recoding) consists of a deliberate reduction of data accuracy, as for instance convert a person's age into an age group. At the end, each record is indistinguishable from a group of at least k-1 other records with respect to the set of quasi-identifier attributes. K-anonymity works as the basis for most of anonymization models. Some proposals try to introduce improvements based on the specific contents of data, as avoiding that all the k records of a group have a same sensitive value on one variable [6] [7]. The work presented in [8] performs k-anonymity for a large data set and then recodes sensitive attributes by adding a random, or fuzzy, factor. A software package for probabilistic anonymization is proposed in [9]. Instead of using k-anonymity, they perturb the data through the addition of a random noise.

As important as to anonymize a data set is to assess the re-identification risk. For that purpose, at least three approaches are available [10] [11]: prosecutor risk, journalist risk and marketer risk. In the prosecutor scenario, the adversary is supposed to know that the target is in the data set. In that case the estimates of uniqueness are based in the studied population. In the journalist approach, the adversary doesn’t know for certain that the target is in the data set. In that case, the risk should be calculated using bigger populations, like similar studies or the general population. In the last scenario, marketer risk, the adversary wants to re-identify as many subjects as possible. In [12], it is proposed a statistical model to quantify the likelihood for a re-identification attempt to be successful.

This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/EEA/50008/2020 and by project CEMAPRE/REM - UIDB/05069/2020 - financed by FCT/MCTES through national funds.

They show that, even if the data set is heavily incomplete, it may not satisfy the modern standards for anonymization.

C. Anonymization Tools

There are a number of software tools available to help in the de-identification process and to access re-identification risks. Open source tools include Amnesia [13], a web based application with a Java backend, some tools based on the R language as μ-ARGUS [14] and sdcMicro [15], and Java based tools as Anonimatron [16], a tool compliant with several database systems and ARX [17] the one used in this work. ARX was chosen because it can be used in data sets with up to 50 attributes and millions of records.

D. Contribution and Structure

This paper presents a k-anonymization sensitivity analysis, varying k in the algorithm implemented in ARX software [19] [20]. With worked examples generated from a real dataset made publicly available for the purpose of open government data and accountability – the Enade data. Admitting as a working hypothesis that this set of personal data is protected by law, we assess the risk of re-identification and the loss of data / information for indirect or quasi-identifiers with research value. For instance, some research value variables are: Age, Gender, Race/skin color, Parents' education. Two processes of anonymization are explored: (1) data suppression; (2) data generalization.

The remaining of the paper consists of three sections. The second section presents data characteristics and the sensitivity analysis study design, the third section presents the results and discussion, and finally the conclusion.

II. Methodology

A. Enade Data

The National Student Performance Exam (Enade) takes place every year in Brazil since 2004. It assesses the higher education graduates' performance, taking into account several dimensions and skills [18]. The Enade is part of the Brazilian higher education evaluation system (Sinaes), which is also composed by the programs evaluation and institutional evaluation. The results of the exam and students' answers to the questionnaires provide data to the indicators of higher education quality. Student's participation is compulsory. The assessment instruments cover several cognitive domains depending on the area of studies, but for the purpose of this article we consider a student's general score, e.g. grade point average (GPA). We also consider student's sociodemographic variables such as Gender, Age, self-declared Race/skin color, Mother's education, and Father's education. The higher education institution and program identification codes (respectively University id and Program id), and Region are also included in our analyses. The microdata are available at the INEP site [18]. Each year a subgroup of disciplinary areas is evaluated so that whole evaluation cycle occurs over a triennium. According to INEP site, in the first year, the evaluation includes Baccalaureate programs in Health Sciences and related areas, Agrarian Sciences, Engineering and Architecture and Urbanism, Higher Technology Courses in the areas of Environment and Health, Food Production, Natural Resources, Military and Security. In the second year, the evaluation includes Bachelor courses in the areas of Biological Sciences, Exact and Earth Sciences, Linguistics, Letters and Arts and related areas, Degree courses in the areas of knowledge of Health Sciences; Human Sciences; Biological Sciences; Exact and Earth Sciences; Linguistics, Letters and Arts; Bachelor courses in the areas of knowledge of Humanities and Health Sciences, with courses evaluated in the context of undergraduate degrees; Higher Technology Courses in the areas of Control and Industrial Processes, Information and Communication, Infrastructure and Industrial Production. The third year, Bachelor programs in the Applied Social Sciences and related areas; B.A. programs in the Humanities and related areas. Higher Education programs in Management and Business, School Support, Hospitality and Leisure, Cultural Production and Design.

In 2018, 548,127 students were involved. Table 1 presents the selected variables and the respective scales of measurement as they are listed in the data dictionary.

TABLE I. Selected Variables

Variable	Scale
University id	Between 1 and 23,410
Program id	Between 1 and 5,001,389
Region	1 = North (N) 2 = Northeast (NE) 3 = Southeast (SE) 4 = South (S) 5 = Central-West (C-W)
Age	Between 4 and 94

Gender	M = Male
	F = Female

Year of high school conclusion	AAAA = Between 0 and 2,686
Year of beginning graduation	AAAA = Between 1,973 and 2,099
Grade point average (GPA)	Minimum = 0; Maximum= 93.7

Race / Skin color	A =White
	B = Black
	C = Yellow
	D =Pardo
	E =Indigenous
	F = Not declared

Mother's education
Father's education
	A = None
	B = 1st – 5th grade
	C = 6th – 9th grade
	D =Secondary school
	E =Graduation
	F = Post-graduation

For the purpose of this study the "Number of years needed to start the graduation" is computed by the difference between the "Year of beginning graduation" minus "Year of high school conclusion". The "Number of years needed to finish the graduate studies" is computed by difference between the current year (2018) and the "Year of beginning graduation", plus one. Since some values recorded in the data set were not plausible according to the purposes of ENADE and the Brazilian Educational System, it was necessary to pre-process the dataset.

This data pre-treatment consisted on eliminating the values of the first year of graduation and the last year of secondary school that lead us to the conclusion of negative values for the "Number of years needed to finish the graduate studies" or to the "Number of years to start the graduation". We also deleted the cases where the starting year of graduation coincided with the last year, because that is not possible, since in Brazil the academic year agrees with the civil year. Finally, and since it was also incoherent, we ended up eliminating the cases that had the value of first year of entrance in graduating studies greater than 2018. If the value for research variables were all missing data, the respective records were also suppressed at this stage. This whole process resulted in the elimination of 41,447 records from the downloaded data set. To the resulting data set, with 506,680 records, we will now call the original data set.

B. *Study Design*

The sensitivity analysis considers as input the *K* and as output the relative risk of re-identification, the loss of subjects, and the absolute deviation between descriptive statistics obtained from the original data set and the k-anonymized data sets. The descriptive statistics calculated are the Mean, Median, Mode, Standard Deviation (SD), Skewness, Kurtosis, Coefficient of Variation (CV) and the Interpercentile Range (IPR = P_{90} – P_{10}). In addition, the qualitative variables empirical distribution is analysed.

At the first study, the anonymization is conducted by varying *K*=2, …, 5 and classifying as quasi-identifiers the variables: “University id”, “Program id”, Age, Gender, “Year of high school conclusion”, “Year of beginning graduation”, Race, Mother’s education and Father’s education. The variables Region and the GPA are classified as insensitive, which means not used for anonymization and thus stay untouched. Applying the k-anonymity, with that variable classification, results in a huge loss of information for all values of k.

Considering that the “University id” and “Program id” may be previously pseudo-anonymized, they are not considered as quasi-identifiers in the second study. In addition, we generalize three variables, Age, “Mother’s education” and “Father’s education”. The values of Age are recoded in less than 26 years and equal or greater than 26. For Parents’ education three class intervals are considered: the first includes values A and B of Table I, the second includes values C and D, and the third includes values E and F. Then, the anonymization process is conducted varying *K* from 2 to 5.

III. Results and Discussion

Table II presents the percentage of suppressed records resulting from the anonymization approaches described above. As can be seen, the suppression percentage was more than 90%, in the first study, and was less than 15% in the second.

TABLE II. Percentage of Suppressed Records

K = 2	K = 3	K = 4	K = 5
Dataset without generalization			
90.34	96.10	97.76	98.50
Dataset with generalization			
4.98	8.59	11.49	14.07

Table III shows the valid cases and the suppressed ones for the chosen variables in the first study. As K increases, the suppressed records increase. A huge loss of data / information occurs. For example, the variable Gender in the original data does not have missing values, so that the number of valid cases is 506,580, and the percentage of suppressed cases represents 90.34% when k=2. This loss of data may have serious implications on the good-representativeness of the original population in each anonymized dataset. To enlighten that point, a descriptive analysis is conducted and the results presented in Tables IV and V.

Descriptive statistics in Table IV show that the Mean varies according to the variation of *K*, no matter the variable. Depending on the variable, the Median may or may not remain stable. For example, the Median of “Number of years to complete higher education studies” is 5 in the original dataset, and in any *K* simulation exercise. The Mode remains stable for every variable analyzed and for K simulation exercise. The dispersion statistics, such as IPR, SD and the CV, show that as K increases as the variability sharply decreases. The skewness and kurtosis estimates suggest that each variable distribution changes with *K*, but the pattern of change depends on the variable itself.

So, according to Table IV, both SD and IPR always decrease as the k-value increases. This means that the extreme values are successively eliminated, once they might represent atypical cases, since their low expression in the original dataset. In other words, with the increase of k, the major amount of records is not suppressed, unlike the extreme ones. In conclusion, we get to obtain, with the anonymization process, a less diverse distribution, since the values that stand through the whole process get closer to the Mode, as we can confirm with the Mean and Median values.

Furthermore, for the “Number of years to complete graduate studies” both Kurtosis and Skewness decrease, instead of increasing, as it happens with Age and “Number of years needed to start graduation”. Considering its Mean and Median, we can also notice that, against what happens with the other two variables, both values are closer to the Mode, and SD and IPR, in the original data set are lower. This suggests that these distributions are, originally, more homogeneous than the other two, i.e., the existence of extreme values is less frequent, or their deviation from the Mean value is lower than the other variables.

The empirical distribution of Gender, Race/skin color and Parents’ education is presented in Table V. The comparison between the original distribution and the anonymized sample suggests a complete distortion of results. In fact, the distribution of research value variables, such as Gender or Race/skin color, becomes completely misrepresented.

Tables VI and VII present the deviation between the k-anonymized descriptive statistics and the respective original results. Such differences confirm what we have just described. Most of the descriptive statistics are under-estimated as K increases, and the distribution statistics pattern depends on the variable itself. The relative distortion of the empirical distributions tends to favor female students, self-declared White, whose Parents completed high school or higher education. In other words, as long as we eliminate records through the anonymization process, the racial minorities are sharply decreased or even suppressed, the affluent students become overrepresented, and sociodemographic disadvantaged students under-represented.

TABLE III. VALID AND SUPPRESSED CASES FOR ALL VARIABLES AFTER K-ANONYMIZATION WITHOUT GENERALIZATION.

Variable		Original	k=2	k=3	k=4	k=5
Age	***Valid***	506,680	48,951	19,775	11,342	7,586
	Suppressed	0	457,729	486,905	495,338	499,094
Grade Point Average	***Valid***	431,424				
	Suppressed	75,256				
Number of years to start graduation	***Valid***	496,478	48,113	19,521	11,244	7,532
	Suppressed	10,202	458,567	487,159	495,436	499,148
Number of years to complete graduate studies	***Valid***	496,478	48,113	19,521	11,244	7,532
	Suppressed	10,202	458,467	487,159	495,436	499,148
Gender	***Valid***	506,680				
	Suppressed	---	90.34%	96.10%	97.76%	98.50%
Race / Skin color; Parent's Education	***Valid***	506,680				
	Suppressed	10.70%	91.56%	96.47%	97.90%	98.57%

TABLE IV. STATISTICS OF THE QUANTITATIVE VARIABLES IN EACH DATASET AFTER K-ANONYMIZATION WITHOUT GENERALIZATION.

Variable		Mean	Median	Mode	IPR	SD	Skewness	Kurtosis	CV
Age	***Original***	29.31	26.00	23.00	19.00	8.24	1.52	2.33	0.28
	k=2	24.83	23.00	23.00	9.00	4.73	2.58	7.95	0.19
	k=3	24.18	23.00	23.00	7.00	3.88	2.74	8.56	0.16
	k=4	23.78	23.00	23.00	5.00	3.24	2.97	10.85	0.14
	k=5	23.58	23.00	23.00	5.00	2.88	3.00	10.99	0.12
Grade Point Average		41.90	41.10	37.50	37.8	14.41	0.21	-0.35	0.34
Number of years to start graduation	***Original***	5.13	2.00	0.00	14.00	6.68	1.86	4.00	1.30
	k=2	2.15	0.00	0.00	8.00	4.25	2.84	9.67	1.98
	k=3	1.71	0.00	0.00	6.00	3.71	2.94	9.80	2.17
	k=4	1.40	0.00	0.00	5.00	3.23	3.23	12.81	2.31
	k=5	1.20	0.00	0.00	5.00	2.89	3.32	13.96	2.41
Number of years to complete graduate studies	***Original***	4.67	5.00	5.00	5.00	1.91	2.07	12.05	0.41
	k=2	4.62	5.00	5.00	2.00	1.13	1.54	12.74	0.24
	k=3	4.62	5.00	5.00	1.00	0.86	0.89	9.20	0.19
	k=4	4.62	5.00	5.00	1.00	0.69	0.50	4.86	0.15
	k=5	4.64	5.00	5.00	1.00	0.65	0.47	4.11	0.14

TABLE V. VALID PERCENTAGE FOR EACH CATEGORY OF THE QUALITATIVE VARIABLES AFTER K-ANONYMIZATION WITHOUT GENERALIZATION..

Variable		Original	k=2	k=3	k=4	k=5
Region	***N***	5.09				
	NE	18.18				
	SE	45.15				
	S	23.29				
	C-W	8.29				
Gender	***F***	59.50	71.61	75.84	77.72	78.82
	M	40.50	28.39	24.16	22.28	21.18
Race / Skin color	***White***	54.10	71.68	77.16	80.18	82.76
	Black	9.00	2.03	0.68	0.39	0.24
	Yellow	2.40	0.33	0.07	0.04	---
	Pardo	32.40	25.85	22.08	19.40	17.00
	Indigenous	0.30	---	---	---	---
	Not declared	1.90	0.12	---	---	---
Father's Education	***None***	8.20	7.75	9.63	11.74	13.54
	1st – 5th grade	27.00	20.83	19.70	19.95	19.26
	6th – 9th grade	15.40	8.49	4.64	3.27	2.55
	Secondary School	30.80	36.56	34.13	30.18	27.47
	Graduation	13.70	19.41	23.52	25.77	27.18
	Post-graduation	4.80	6.96	8.37	9.08	10.01
Mother's Education	***None***	6.10	7.05	9.29	11.58	13.47
	1st – 5th grade	23.00	16.30	15.81	16.12	15.77
	6th – 9th grade	15.50	8.05	4.93	3.65	2.49
	Secondary School	33.10	38.10	34.25	30.89	28.78
	Graduation	14.40	20.66	25.04	26.83	28.74
	Post-graduation	8.00	9.84	10.68	10.93	10.74

TABLE VI. DEVIATION FROM THE ORIGINAL DATASET FOR THE STATISTICS OF THE QUANTITATIVE VARIABLES WITHOUT GENERALIZATION.

Variable		Mean	Median	Mode	IPR	SD	Skewness	Kurtosis	CV
Age	*k=2*	-4.48	-3.00	0.00	-10.00	-3.51	1.06	5.62	-0.09
	k=3	-5.13	-3.00	0.00	-12.00	-4.36	1.22	6.23	-0.12
	k=4	-5.53	-3.00	0.00	-14.00	-5	1.45	8.52	-0.14
	k=5	-5.73	-3.00	0.00	-14.00	-5.36	1.48	8.66	-0.16
Number of years to start graduation	*k=2*	-2.98	-2.00	0.00	-6.00	-2.43	0.98	5.67	0.67
	k=3	-3.42	-2.00	0.00	-8.00	-2.97	1.08	5.8	0.87
	k=4	-3.73	-2.00	0.00	-9.00	-3.45	1.37	8.81	1.00
	k=5	-3.93	-2.00	0.00	-9.00	-3.79	1.46	9.96	1.11
Number of years to complete graduate studies	*k=2*	-0.05	0.00	0.00	-3.00	-0.78	-0.53	0.69	-0.16
	k=3	-0.05	0.00	0.00	-4.00	-1.05	-1.18	-2.85	-0.22
	k=4	-0.05	0.00	0.00	-4.00	-1.22	-1.57	-7.19	-0.26
	k=5	-0.03	0.00	0.00	-4.00	-1.26	-1.60	-7.94	-0.27

TABLE VII. DEVIATION FROM THE ORIGINAL DATASET RESULTS FOR EACH CATEGORY OF THE QUALITATIVE VARIABLES WITHOUT GENERALIZATION

Variable		k=2	k=3	k=4	k=5
Gender	*F*	12.11	16.34	18.22	19.32
	M	-12.11	-16.34	-18.22	-19.32
Race / Skin color	*White*	17.58	23.06	26.08	28.66
	Black	-6.97	-8.32	-8.61	-8.76
	Yellow	-2.07	-2.33	-2.36	---
	Pardo	-6.55	-10.32	-13.00	-15.40
	Indigenous	---	---	---	---
	Not declared	-1.78	---	---	---
Father's Education	*None*	-0.45	1.43	3.54	5.34
	1st – 5th grade	-6.17	-7.30	-7.05	-7.74
	6th – 9th grade	-6.91	-10.76	-12.13	-12.85
	Secondary School	5.76	3.33	-0.62	-3.33
	Graduation	5.71	9.82	12.07	13.48
	Post-graduation	2.16	3.57	4.28	5.21
Mother's Education	*None*	0.95	3.19	5.48	7.37
	1st – 5th grade	-6.70	-7.19	-6.88	-7.23
	6th – 9th grade	-7.45	-10.57	-11.85	-13.01
	Secondary School	5.00	1.15	-2.21	-4.32
	Graduation	6.26	10.64	12.43	14.34
	Post-graduation	1.84	2.68	2.93	2.74

Table VIII shows the second study results. We intentionally include the subset of variables that showed more severity of misrepresentation in study one. It can be observed that the descriptive statistics of "Number of years to start graduation" are closer to the original dataset, even though they present a little underestimation. The Gender and Race/skin color distributions are also closer to the original ones, but the under-representation of minority groups still remains.

Finally, Table IX presents the average risk of re-identification obtained with the ARX tool for the prosecutor scenario. The risk was assessed for both studied data sets, with and without generalization when varying the value of k. The column for k=1 presents the average risk before k-anonymity is performed. As can be seen after the first stage of anonymization the risk of re-identification decreases from more than 90% in the original data, to approximately 14%. With generalization, it is possible to decrease the risk to near 4%. The risk decline is more sensitive to the growth of k in the first study than in the second. In this one, the risk has an acceptable value of 7% even with k=2.

TABLE VIII. SUMMARY OF RESULTS AFTER GENERALIZATION

			Number of years to start graduation		**Gender**		**Race / Skin color**	
			k=2	*k=5*	*k=2*	*k=5*	*k=2*	*k=5*
Mean			4.74	4.19				
SD			6.17	5.52				
Skewness			1.79	1.75				
Kurtosis			3.50	3.23				
% of valid cases	*Gender*	*F*			60.0	61.0		
		M			40.0	39.0		
	Race / Skin color	*White*					55.50	57.80
		Black					8.50	7.46
		Yellow					1.90	1.20
		Pardo					32.80	32.87
		Indigenous					0.10	0.02
		Not declared					1.30	0.66

TABLE IX. AVERAGE PROSECUTOR RISK OF RE-IDENTIFICATION

K = 1	k = 2	k = 3	k = 4	k = 5
Dataset without generalization				
94.16%	39.56%	24.18%	17.37%	13.59%
Dataset with generalization				
42.23%	7.03%	5.54%	4.42%	3.80%

IV. CONCLUSION

In this work a sensitivity analysis over the k value of the ARX k-anonymization algorithm was performed. Using real data published by INEP, the Institute for the Brazilian educational system evaluation, the impact of varying the value of k on the percentage of suppressed records and the impact on the re-identification risk was assessed. Two main setups were considered: K-anonymization without any generalization and k-anonymization with generalization of three personal attributes, Age of the student and Mother's and Father's education level. Descriptive statistics for all the anonymized data sets were calculated in order to assess the value of the data that remains after each anonymization stage. The results obtained corroborate the conclusion presented recently by Sweeney in [2], "In today's data-rich, networked society, the k constraint must be enforced across all fields or scientific justification provided to exclude a field" ([2], p. 1). In addition, our results confirm that the minorities and socioeconomic disadvantaged groups become under-represented after the anonymization as [2] concluded.

REFERENCES

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) in Official Journal of the European Union, L 119, pp. 1–88, 2016.

[2] L. Sweeney, M. V. Loewenfeldt, M. Perry, "Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data," Technology Science. 2018111301. November 13, 2018. https://techscience.org/a/2018111301.

[3] L. Sweeney. "k-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10 (5), pp 557-570, 2002.

[4] P. Samarati. "Protecting Respondents' Identities in Microdata Release," IEEE Transactions on Knowledge and Data Engineering, Vol. 13 (6), pp. 1010-1027, 2001.

[5] V. Ayala-Rivera, P. McDonagh, T. Cerqueus and L. Murphy, "A Systematic Comparison and Evaluation of K-Anonymization Algorithms for Practitioners," Trans. on Data Privacy, vol. 7(3), pp. 337-370, 2014.

[6] A. Campan, T. M. Truta and N. Cooper, "P - Sensitive K - Anonymity with Generalization Constraints," Transactions on Data Privacy, vol. 3, pp. 65–89, 2010.

[7] M. Al-Zobbi, S. Shahrestani, C. Ruan "Sensitivity-based Anonymization of Big Data," IEEE 41st Conference on Local Computer Networks Workshops, pp. 58-64. 2016.

[8] G. Ursin, S. Sen, J. Mottu and M. Nygard, "Protecting Privacy in Large Datasets - First We Assess the Risk; Then We Fuzzy the Data, Cancer Epidemiology Biomarkers & Prevention," vol 26(8), pp.1219- 1224, 2017.

[9] D. Avraam, A. Boyd, H. Goldstein, P. Burton, "A software package for the application of probabilistic anonymisation to sensitive individual-level data: a proof of principle with an example from the ALSPAC birth cohort study," Longitudinal and Life Course Studies, Vol 9 (4), pp.433-446, 2018. DOI: http://dx.doi.org/10.14301/llcs.v9i4.478.

[10] L. Kniola, "Plausible Adversaries in Re-Identification Risk Assessment" PhUSE Annual Conference, 2017.

[11] F. Prasser and F. Kohlmayer, "Putting statistical disclosure control into practice: the ARX data anonymization tool," in A. Gkoulalas-Divanis and G. Loukides (Ed.s) Medical Data Privacy Handbook. Cham, Switzerland: Springer International Publishing; 2015. p.111–48.

[12] L. Rocher, J. M. Hendrickx, Y. Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature Communications, 2019; 10(1) DOI: 10.1038/s41467-019-10933-3.

[13] Amnesia. A data anonymization tool supported by the Institute for the Management of Information Systems. Available from: https://amnesia.openaire.eu/installation.html.

[14] μ-ARGUS - Anti Re-identification General Utility System. Available from: http://neon.vb.cbs.nl/casc/mu.htm.

[15] sdcMicro - Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation. Available from: https://cran.r-project.org/web/packages/sdcMicro/index.html.

[16] Anonimatron. Available from: https://realrolfje.github.io/anonimatron/.

[17] ARX - Data Anonymization Tool. Available from: https://arx.deidentifier.org/.

[18] INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, "Exame Nacional de Desempenho dos Estudantes (Enade)," 2018. [Online]. Available: http://portal.inep.gov.br/enade. [Accessed: 15-Feb-2020].

[19] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper and K. A. Kuhn, "Flash: Efficient, Stable and Optimal K-Anonymity." 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security,Risk and Trust, pp. 708-717, DOI: 10.1109/SocialCom-PASSAT.2012.52.

[20] J. Eicher, R. Bild, H. Spengler et al."A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models," BMC Med Inform Decis Mak 20, 29 (2020). https://doi.org/10.1186/s12911-020-1041-3.

Appendix B

Garantia de Privacidade Versus Utilidade dos Dados em Anonimização: um estudo no ensino superior

Garantia de Privacidade Versus Utilidade dos Dados em Anonimização: um estudo no ensino superior

Paula Prata [1, 2], Maria Eugénia Ferrão [1, 3], Wilson Santos [1, 2], Gonçalo Sousa [1]

pprata@di.ubi.pt, meferrao@gmail.com

[1] Universidade da Beira Interior, Covilhã, Portugal

[2] Instituto de Telecomunicações (IT-UBI), Covilhã, Portugal

[3] REM - Research in Economics and Mathematics, CEMAPRE

DOI: 10.17013/risti.n.pi-pf

Resumo: No mundo digital, toda a atividade humana deixa um rasto de dados que constitui um recurso cada vez mais valioso, para avaliação e definição de estratégias nos mais variados domínios. A partilha desses dados, sendo socialmente importante, implica o respeito pela privacidade individual e portanto a sua anonimização. As atuais leis e regulamentos sobre privacidade oferecem orientações limitadas para lidar com um vasto leque de tipos de dados, ou com técnicas de reidentificação. Este trabalho pretende ilustrar um processo de anonimização, comparando para vários modelos de privacidade a perda de informação e a utilidade do conjunto de dados resultante. Encontrar o equilíbrio entre privacidade e utilidade é um desafio que pode ser mais facilmente alcançado por quem melhor conhece o significado dos dados e dos objetivos que se pretendem alcançar com eles.

Palavras-chave: Anonimização de dados, k-anonimato; ℓ-diversidade; t-proximidade; ENADE.

Privacy Preserving Versus Utility Preserving in Data Anonymization: a study in higher education

Abstract: In the digital world, all human activity leaves a trace of data that is growingly valued for the evaluation and definition of strategies in varied domains. The sharing of those data, being socially relevant, implies the respect for individual privacy and so, its anonymization. The current laws and regulations about privacy offer limited guidance to deal with the vast range of datatypes or with techniques of re-identification. This work aims at illustrating a process of anonymization, comparing to several models of privacy, the loss of information and the usefulness of that dataset resulting from the anonymization. Finding a balance between privacy and utility is a challenge that can be more easily found by those who know better the meaning of the data and objectives aimed at.

Keywords: Data anonymization; k-anonymity; ℓ-diversity; t-closeness; ENADE.

1. Introdução

Nos dias de hoje, a quantidade de dados sobre a atividade humana que é recolhida e armazenada digitalmente está em constante crescimento. Esses dados podem passar por todos os aspectos da nossa vida, como, por exemplo, a atividade nas redes sociais, rastos de localização recolhidos por telefones móveis, compras *online*, ou registos médicos. Transformar esses dados em conhecimento é uma mais-valia que tem tornado os dados num recurso cada vez mais valioso. O processamento e análise de dados possibilitam avanços socialmente importantes, em campos tão diversos como sistemas de suporte à decisão médica, criminologia computacional, protecção contra terrorismo informático ou marketing direccionado. Todos estes aspectos há muito idealizados (Chen et al., 2012; Adomavicius & Tuzhilin, 2005; Quiñonez et al., 2019) são, cada vez mais, possíveis devido à transversal digitalização da sociedade. O crescente interesse das mais variadas organizações em terem acesso aos nossos dados pode ser traduzido pela frase de Prasser et al. (2020) "The race for innovation has turned into a race for data" (p. 1277). No entanto, todo este potencial de análise de dados tem um custo associado. Os dados recolhidos, incluindo informação sensível, podem ser publicados e partilhados com entidades externas que as poderão usar para fins não previstos originalmente. Existe uma panóplia de riscos associados à partilha de dados pessoais, em especial se esses dados foram posteriormente associados com outras fontes, podendo a divulgação de dados pessoais sensíveis causar danos graves aos indivíduos em causa. Para evitar esses riscos, têm sido criados regulamentos de protecção de dados visando aumentar a garantia de protecção dos dados pessoais, (Directive 95, 1995) assim como existe inúmera investigação sobre os aspectos éticos, legais e sociais da partilha de dados (Kaye et al., 2012; Cambon-Thomsen, 2007). Em particular, com a entrada em vigor do Regulamento Geral sobre a Proteção de Dados, RGPD (GDPR, 2016), este tema está na ordem do dia e tem levado à consciencialização da sociedade para o problema da privacidade dos dados.

Vários exemplos de violação da privacidade têm sido descritos na literatura, como o conhecido caso do Governador do estado do Massachusetts, USA, William Weld que viu os seus dados médicos divulgados publicamente, quando uma base dados de um sistema de saúde foi tornada pública e os seus registos foram cruzados com dados de um caderno eleitoral que continha dados como "zip code", data de nascimento e género (Barth-Jones, 2012). Cada um destes atributos isolado não permite a identificação de um individuo, mas a sua combinação com outras fontes de dados pode levar a um conjunto mínimo de registos (Sweeney, 2002b).

Geralmente, para a reidentificação ser possível, o adversário tem de conhecer *a priori* duas peças de informação: sabe que o registo da vítima está na base de dados e conhece algum atributo quase-identificador. No contexto de anonimização de dados, um adversário é alguém que tenta identificar indivíduos num conjunto de dados, supostamente anonimizado, e um atributo quase-identificador é definido como um atributo que não identifica um indivíduo, mas pode fazê-lo quando associado a outra informação. No caso anterior, o adversário sabia que a vítima tinha estado hospitalizada e os restantes dados foram fáceis de obter (Fung et al., 2010). Este caso teve grande impacto na procura por mecanismos de garantia de privacidade de dados pessoais. Foi demonstrado que 87% da população dos USA pode ser facilmente identificada com apenas três quase-identificadores: "zip code", género e data de nascimento (Sweeney, 2000). Também o caso relatado em (Panduragan, 2014) mostra que dados supostamente anonimizados podem permitir a reidentificação. O número das licenças de cada táxi de Nova Iorque (composto por sete dígitos) foi anonimizado usando valores de dispersão. Os valores foram facilmente revertidos e informação sensível dos taxistas como percursos efectuados, o seu rendimento, e até a sua morada foram revelados. Mais recentemente, o estudo apresentado em (Sweeney et al., 2018) mostrou ser possível identificar univocamente estudantes de uma escola de Direito cujos dados tinham sido anonimizados de forma independente por 4 protocolos, correntemente usados. Muitos outros exemplos mostram quão importante e difícil é efectuar uma correta anonimização, assim como perceber os riscos associados à segurança dos nossos dados (Sweeney, 2015; Culnane et al., 2017; Koch, 2020).

Num processo de anonimização de dados pessoais, um aspeto, tão importante como garantir a privacidade de cada individuo, é garantir que os dados resultantes continuam a ter utilidade. Anonimizar significa retirar algumas características dos dados, e portanto, informação útil para os seus utilizadores pode ser perdida. Anonimizar deve ser um processo iterativo, em que a cada aplicação de um modelo de privacidade, e consequente avaliação do risco de reidentificação, se deve seguir a avaliação da utilidade dos dados obtidos. Todo o processo deve ser repetido, até se alcançar um equilíbrio razoável entre minimizar o risco de reidentificação e manter o máximo de utilidade dos dados (Prasser et al., 2020). Esta última pode ser avaliada pelo cálculo de uma simples proporção dos dados perdidos ou por métodos estatísticos, mais sofisticados, que indiquem em que medida as características dos dados anonimizados se distanciam dos dados originais. Todo o processo de anonimização depende do tipo de dados e do uso dos dados (Francis, 2018) ou propósito da análise de dados.

Neste trabalho, foram estudados, para um subconjunto dos dados públicos do ENADE - Exame Nacional de Desempenho do Estudantes de graduação do Brasil, vários processos de anonimização, comparando os resultados em termos de risco de reidentificação e de utilidade dos dados. Usando uma ferramenta de código aberto, foram aplicados dois modelos de privacidade, ℓ-diversidade e t-proximidade, considerando várias parametrizações, foi avaliado o risco de reidentificação associado e foi avaliada a utilidade dos dados resultantes, através de um modelo de análise de variância com múltiplos fatores principais e interacções de 2ª ordem. A partilha de dados pode trazer vários benefícios à sociedade, seja para avanços científicos, avaliação de políticas ou para melhoria de serviços. Este artigo contribui para a reflexão sobre o *trade-off* entre privacidade e utilidade dos dados. Quando os dados são provenientes de registos administrativos ou de órgãos governamentais, com grande potencial para fins de investigação científica, aspetos normativos e outros decorrentes da aplicação do RGPD podem inviabilizar ou até distorcer os fins da investigação científica. Adicionalmente, constitui uma abordagem exploratória de interesse para investigadores ou organizações que pretendam anonimizar os seus dados, tirando partido do elevado conhecimento do contexto e significado dos dados, e tornando o processo de anonimização tecnicamente explícito. Deste modo, o artigo contribui também para a adoção de práticas informadas e justificadas no processo de anonimização sem, contudo, por em causa os aspetos legais de privacidade impostos pelo RGPD.

Na secção 2 são descritos os modelos de privacidade utilizados, assim como o modelo que está na sua base, o modelo de k-anonimato. A secção 3 apresenta o modelo de utilidade escolhido para o propósito deste trabalho, isto é, o modelo de análise de variância com múltiplos fatores (ANOVA) e a secção 4 refere trabalho relacionado. A secção 5 contém o estudo experimental em três subsecções: descrição dos dados e do seu pré-processamento; a análise de privacidade e discussão dos resultados; a análise de utilidade e discussão dos resultados. Finalmente, a secção 6 apresenta as conclusões.

2. Modelos de Privacidade

As duas principais abordagens de anonimização são a aleatorização e a generalização. A aleatorização consiste em alterar os dados de forma a reduzir a possibilidade de associação entre os dados e o indivíduo. Uma técnica é, por exemplo, a adição de ruído aleatório a algumas variáveis, como proposto em Goldstein e Shlomo (2020). A generalização ou agregação consiste na junção de categorias ou classes de variáveis através de alteração da escala ou ordem de grandeza. Neste trabalho, vamos explorar dois modelos de privacidade baseados em generalização: ℓ-diversidade e t-proximidade. Estes dois modelos são evoluções de um modelo mais simples de privacidade que é o k-anonimato. Os três modelos vão ser descritos nas próximas subsecções. Ao aplicar um modelo de privacidade, pretende-se: reduzir o risco de identificação, isto é, evitar que um indivíduo seja associado a um registo específico; reduzir o risco de ligação, isto é, reduzir a possibilidade de associar dois registos do mesmo indivíduo quer estejam na mesma ou em diferentes bases de dados; reduzir o risco de inferência, isto é, não permitir que, após a anonimização, seja possível deduzir o valor de um atributo a partir dos valores de outros atributos de um dado individuo. Para avaliar o risco de reidentificação, são comuns três abordagens diferenciadas pelo que é suposto o possível adversário conhecer sobre os dados (Prasser & Kohlmayer, 2015; Kniola, 2017): modelo de promotor, em que se supõe que o adversário sabe que o indivíduo que procura está na base de dados; modelo de jornalista, em que o adversário desconhece se o indivíduo está na base de dados; modelo de marketing, em que o adversário quer identificar o maior número de indivíduos possível.

2.1. O Modelo k-anonimato

Um processo de anonimização começa por classificar os atributos do conjunto de dados. Atributos que permitam identificar directamente um indivíduo, como nome ou número de cartão de cidadão, são classificados como identificadores diretos. Atributos que não identificam um indivíduo diretamente, mas que permitam a associação com outros conjuntos de dados, são quase-identificadores. Os restantes atributos podem ainda ser classificados como sensíveis ou não sensíveis. Um atributo é sensível se o seu valor não deve ser descoberto por qualquer adversário, para nenhum indivíduo do conjunto de dados, caso contrário, o atributo será classificado como não sensível. Após a classificação dos atributos, é necessário suprimir ou modificar os atributos diretos. Como vimos nos exemplos apresentados na introdução, isso não é suficiente para evitar a reidentificação. Através de atributos quase-identificadores, é possível ligar os registos com outras bases de dados e identificar indivíduos no conjunto de dados. Para evitar esse risco de ligação, foi proposto o modelo de privacidade k-anonimato (Sweeney, 2002a). Um conjunto de dados é k-anónimo, se cada registo é indistinguível de pelo menos k-1 outros registos, no que diz respeito aos atributos quase-identificadores. Formalmente, k-anonimato é definido da seguinte forma: "Seja a tabela RT(A1, ..., An), e QI_{RT} os quase-identificadores associados a essa tabela. RT satisfaz k-anonimização em relação a QI_{RT} se e só se cada sequência de valores em RT[QI_{RT}] tem no mínimo k ocorrências em RT[QI_{RT}]" (Sweeney, 2002a, p. 564). Para evitar que um indivíduo possa ser univocamente identificado através de ligação a outros conjuntos de dados, o modelo assegura que, para cada combinação dos seus atributos quase-identificadores, existem pelo menos k registos que partilham os mesmos valores. Registos que não verificam esta condição são eliminados.

Foram desenvolvidos inúmeros algoritmos que implementam o k-anonimato, como por exemplo, Datafly (Sweeney, 2002a), Incognito (LeFevre et al., 2005) e Mondrian (LeFevre et al., 2006). Segundo Ayala-Rivera et al. (2014) não existe um algoritmo melhor do que os outros. O melhor algoritmo em cada situação é influenciado por múltiplos fatores, como por exemplo o número de quase-identificadores, ou a distribuição dos dados na base de dados.

2.2. O Modelo ℓ-diversidade

O principal problema do modelo de k-anonimato é permitir a divulgação de informação, devido à falta de diversidade num ou vários atributos sensíveis. Se tivermos um conjunto de k registos, todos com os mesmos valores nos atributos quase-identificadores, e ocorrer que todos eles tenham um mesmo valor para um atributo sensível, então qualquer adversário que conheça um indivíduo que corresponda aos valores dos quase-identificadores irá poder inferir o valor do atributo sensível para esse indivíduo. Diz-se que esse conjunto de registos indistinguíveis constitui uma classe de equivalência. O modelo de privacidade ℓ-diversidade melhora o modelo de k-anonimato, reduzindo o risco de inferência de atributos, ao garantir que cada atributo sensível tem pelo menos ℓ valores distintos representados em cada classe de equivalência. Formalmente, considerando um bloco q que seja uma classe de equivalência relativa aos atributos quase-identificadores considerados, esse bloco q é ℓ-diverso se contém pelo menos ℓ valores distintos para os atributos sensíveis S. Uma tabela é ℓ-diversa se cada bloco q é ℓ-diverso (Machanavajjhala et al., 2007, p. 16). O modelo impõe assim que todos os registos que partilhem os mesmos quase-identificadores devem ter diversos valores para os atributos sensíveis. Existem diversas abordagens que tentam formalizar essa diversidade. A definição de (c, ℓ)-diversidade recursiva garante que o valor mais comum não apareça com demasiada frequência enquanto que os valores menos comuns não aparecem muito raramente. A definição formal é a seguinte: dado um bloco q, seja $r1$ o número de vezes que o valor do atributo sensível mais frequente aparece nesse bloco q; $r2$ será o número de vezes que o segundo valor mais frequente aparece e assim por diante até rm para um atributo sensível que tenha m valores possíveis. Dada uma constante c, o bloco q satisfaz (c, ℓ)–diversidade recursiva se $r1< c(r\ell + r\ell+1, + \dots + rm)$. A tabela T é ($c$, ℓ)–diversa recursiva se cada bloco q satisfaz (c, ℓ)-diversidade recursiva. Para $\ell = 1$ a diversidade é sempre verificada (Machanavajjhala et al., 2007, p. 18).

2.3. O Modelo t-proximidade

O modelo de t-proximidade é um melhoramento da ℓ-diversidade, na medida em que tenta obter classes de equivalência com uma distribuição dos valores dos atributos sensíveis próxima da sua distribuição no conjunto original de dados. Segundo Li et al. (2007, p. 109), uma classe de equivalência é dita como tendo t-proximidade, se a distância entre a distribuição de um atributo sensível nessa classe e a distribuição do atributo em toda a tabela não é mais do que um valor limite t. A tabela é dita como tendo t-proximidade se todas as classes de equivalência têm t-proximidade. Para medir a distância entre as duas distribuições é proposto o uso da métrica *Earth Mover's Distance* (Rubner et al., 2000).

3. Modelo de Utilidade: Análise de variância

Para o propósito deste artigo usámos o modelo ANOVA com fatores principais e interação de 2ª ordem entre os fatores. Apresentamos a especificação do modelo com dois fatores e respetiva interação, podendo ser generalizado, através de termos aditivos, ao número de fatores e interações referentes à análise em causa. Considerando uma amostra de tamanho *n (i=1, .., n)*, a equação do modelo é a seguinte: $y_{ipk} = \mu + \gamma_p + \delta_k + \beta_{pk} + e_{ipk}$, onde y_{ipk} denota a classificação final do *i-ésimo* estudante que pertence ao grupo p do fator γ e também pertence ao grupo k do fator δ. Ou seja, γ_p representa o primeiro fator, δ_k representa o segundo fator e β_{pk} refere-se ao efeito de interação entre os dois fatores, *p=1,..,P; k=1,...,K*. Decorre que o fator γ tem P grupos, o fator δ tem K grupos e há PK subgrupos de interação. O termo aleatório do modelo é representado por e_{ipk}, com os seguintes pressupostos: distribuição normal com média nula, homocedasticidade ou homogeneidade das variâncias, elementos independentes entre si. Para mais detalhes sobre o modelo ver, por exemplo, Scheffé (1999).

4. Trabalho relacionado

A maioria dos trabalhos experimentais sobre anonimização de dados lida com dados médicos que, pela sua natureza, contêm informação sensível. Em Spengler e Prasser (2019) uma base de dados biomédicos é usada para avaliar o risco e a utilidade dos dados anonimizados usando os modelos de ℓ-diversidade, t-proximidade e β-semelhança. Também para dados médicos, em Lee et al. (2017) é apresentado um modelo de preservação da utilidade e da privacidade baseado em k-anonimização e "h-ceiling" um método que limita a generalização de dados. Na área da educação, Chicaiza et al. (2020) apresenta um estudo sobre análise de dados de aprendizagem usando k-anonimato e modelos de regressão linear para avaliar a utilidade dos dados. Em Santos et al. (2020) a utilidade de dados educacionais k-anonimizados é analisada calculando estatísticas descritivas para vários valores de k. Estudos recentes introduzem modelos de aprendizagem automática para garantir a privacidade dos dados e avaliar a sua utilidade (Eicher et al., 2020; Esquivel-Quirós et al., 2019).

5. Estudo Experimental

Na componente experimental, que descrevemos de seguida, foram usados os dados do Exame Nacional de Desempenho dos Estudantes de graduação no Brasil (ENADE) disponíveis em *http://portal.inep.gov.br/enade*. Na anonimização dos dados foi usada a *framework* de código aberto, ARX (*https://arx.deidentifier.org/)* e para o estudo de utilidade foi usado o *software* estatístico SPSS.

5.1. Conjunto de Dados

Foram considerados para análise os dados do ENADE de 2018, no qual estiveram envolvidos 548 127 estudantes. O grande volume de registos, mais de meio milhão, pode dar uma falsa sensação de segurança, transmitindo a ideia de que registos únicos são raros, mas uma simples k-anonimização do subconjunto de dados apresentados na Tabela 1, para k=2 mostrou um número de registos únicos muito elevado. Apesar de os dados não conterem identificadores diretos, possuem quase-identificadores que poderão permitir a inferência de dados sensíveis ou ainda a associação a registos de outras bases de dados com possível reidentificação, o que justifica o estudo de anonimização realizado. Os atributos seleccionados foram o código da área do curso, região onde funcionou o curso, idade, género, raça/cor e média final do estudante, os níveis de educação da Mãe e do Pai e o rendimento do agregado familiar. Foi ainda calculado o número de anos entre terminar o ensino secundário e iniciar o curso superior, que designámos por "espera ingresso", e foi calculado o número de anos para concluir a graduação, "tempo diploma". A Tabela 1 mostra os nomes das variáveis usadas, a sua descrição e como foram classificadas para efeitos de anonimização.

Tabela 1 – Variáveis seleccionadas e respectiva classificação.

Variável	**Descrição**	**Classificação**
Código Curso	Código da área de enquadramento do curso	Quase-identificador
Região	Código de região de funcionamento do curso	Quase-identificador
Idade	Generalizada nas categorias: [4,26[e [26,95[	Quase-identificador
Género	M ou F	Quase-identificador
Média Final	Média da classificação final obtida pelo estudante	Não sensível
Espera Ingresso	*Anos entre terminar secundário e início superior*	*Quase-identificador*
Tempo Diploma	*Tempo para obtenção do diploma*	*Quase-identificador*
Raça Cor	*Auto declaração*	*Quase-identificador*
Educação Pai	*Generalizada nas categorias: [A,B] [C,D] [E,F]*	*Quase-identificador*
Educação Mãe	*Generalizada nas categorias: [A,B] [C,D] [E,F]*	*Quase-identificador*
Rendimento Familiar	*Número de salários mínimos do agregado familiar*	*Sensível*

Os dados resultantes foram pré-processados, tendo sido removidos registos com valores pouco plausíveis, como, por exemplo, registos em que o ano em que terminavam o ensino superior era inferior a 2018, ou ainda registos cujo valor calculado para o "tempo diploma" dava negativo. O conjunto resultante ficou com 536 466 registos. De seguida, foram generalizadas três variáveis: idade, educação da Mãe e educação do Pai. Os valores da idade foram recodificados em menor de 26 ou maior e igual que 26. Os níveis de educação do Pai e da Mãe foram generalizados em 3 categorias em vez das 6 originais. O dicionário de dados completo pode ser consultado no *site* do ENADE. Finalmente, o atributo rendimento familiar foi classificado como sensível, a média final como não sensível e todos os restantes atributos foram classificados como quase-identificadores.

5.2 Análise de Privacidade

Os dados resultantes do pré-processamento foram anonimizados com (c, ℓ) - diversidade recursiva e com t-proximidade, fazendo variar os valores de c, ℓ *e* t. Para cada uma das parametrizações foi quantificada a percentagem de registos eliminados e foi calculado o risco máximo e o risco médio de reidentificação usando o modelo do prossecutor implementado no ARX.

5.2.1 Anonimização por ℓ-diversidade

A Tabela 2 apresenta os resultados da anonimização por (c, ℓ)-diversidade, fazendo variar o valor de ℓ de 2 a 5 para um valor de $c = 3$. Para cada conjunto anonimizado obtido, apresenta-se o número de registos (dimensão), a percentagem de registos eliminados, o risco médio e máximo de reidentificação. Como se pode observar, ao aumentar o valor de ℓ e portanto ao aumentar o número de registos de cada classe de equivalência a percentagem de registos eliminados aumenta drasticamente, subindo de 34,08% para $\ell = 2$ até 82,85% para $\ell = 5$. Por outro lado, o risco médio reduz gradualmente de 13,27% para 2,78%. Em relação ao risco máximo de reidentificação, ele será de $100/\ell$ uma vez que os registos são agrupados em grupos de ℓ registos com valores iguais para os quase-identificadores. O atributo sensível que está a ser diversificado é o rendimento familiar.

Tabela 2 – Dimensão (N) do conjunto de dados, percentagem de registos eliminados, risco médio e máximo de reidentificação após (c, ℓ)-diversidade, para ℓ a variar de 2 a 5, com c=3.

(3, ℓ) - diversidade	**(3,2)**	**(3,3)**	**(3,4)**	**(3,5)**
N	353 637	264 634	171 107	91 991
Registos eliminados (%)	34,08%	50,67%	68,10%	82,85%
Risco médio (prossecutor)	13,27%	7,52%	4,63%	2,78%
Risco máximo	50%	33.3%	25%	20%

A Tabela 3, apresenta os mesmos valores mas agora para os dados resultantes de (c, ℓ)-diversidade fixando o valor de ℓ em 5, e fazendo variar o valor de c de 2 a 4. Aumentar o valor de c, significa aumentar o número de vezes que o valor do atributo sensível mais frequente pode ocorrer em cada classe de equivalência (ver Secção 2.2). Como se pode observar, a percentagem de registos eliminados diminui de 89,39% para 78,76% quando c aumenta de 2 para 4. Em relação ao risco, este aumenta ligeiramente quando c aumenta, no entanto esse resultado resulta apenas do aumento do número de registos. A avaliação do risco pelo modelo do prossecutor implementada no ARX apenas mede o risco de reidentificação e não o risco de inferência do atributo sensível. A avaliação do risco de inferência do valor do atributo sensível virá a ser tratada num próximo trabalho. Podemos no entanto afirmar que ao introduzirmos a diversidade, o risco de inferência diminui.

Tabela 3 – Dimensão (*N*) do conjunto de dados, percentagem de registos eliminados, risco médio e máximo de reidentificação após (*c*, ℓ)-diversidade, para ℓ = 5 e *c* a variar de 2 a 4.

(*c*, *5*) - diversidade	**(2,5)**	**(3,5)**	**(4,5)**
N	56 935	91 991	113 966
Registos eliminados (%)	89,39%	82,85%	78,76%
Risco médio (prossecutor)	2,24%	2,78%	3,12%
Risco máximo	20%	20%	20%

5.2.2 Anonimização por t-proximidade

Para estudar o modelo de t-proximidade, começamos por definir uma dimensão *k* para as classes de equivalência. O valor de *t* determina a distância entre a distribuição dos valores do atributo sensível nessas classes de equivalência e a distribuição no conjunto original. A Tabela 4 apresenta os resultados para os conjuntos de dados produzidos para k=2 e k=5 fazendo t=0,15 e t=0,3.

Tabela 4 – Dimensão (*N*) do conjunto de dados, percentagem de registos eliminados, risco médio e máximo de reidentificação após t-proximidade (k = 2 e k =5 com t= 0,15 e t= 0,3).

t-proximidade	**k=2, t=0,3**	**k=2, t=0,15**	**k=5, t=0,3**	**k=5, t=0,15**
N	348 519	231 645	259 190	195 235
Registos eliminados (%)	35,03%	56,82%	51,69%	63,60%
Risco médio (prossecutor)	14,65%	10,77%	6,20%	5,78%
Risco máximo	50%	50%	20%	20%

Podemos observar que para um mesmo valor de *t*, a percentagem de registos eliminados aumenta quando *k* aumenta, como seria de esperar. Para o mesmo *k*, a percentagem de registos eliminados diminui quando *t* aumenta. Se exigimos maior proximidade na distribuição dos valores sensíveis, obtemos menos registos. Comparando os resultados de t-proximidade com os obtidos por diversidade, para conjuntos com a mesma dimensão das classes de equivalência, isto é, quando ℓ é igual ao k, podemos observar o seguinte: para k=2, (3, 2)-diversidade tem menos registos eliminados (34,08%) que qualquer dos conjuntos obtidos por proximidade 35,03% para t=0,3 e 56,82% para t=0,15; no entanto para k=5, a diversidade elimina entre 78 a 89% dos registos, enquanto a proximidade elimina no máximo 63,6% para t=0,15. Na próxima secção, iremos fazer a análise de utilidade para o conjunto obtido por (3, 5)-diversidade e para os casos de t-proximidade em que a dimensão das classes de equivalência é igual à do caso anterior, k=5, com t=0,15 e t=0,3. O conjunto obtido por diversidade tem um risco médio de reidentificação baixo (2.78%) e o atributo sensível tem bastante diversidade, no entanto, isso ocorre à custa da supressão de mais de 80% dos registos. Os conjuntos obtidos por proximidade perderem respectivamente cerca de 64% e 52% dos registos originais.

5.3 Análise de Utilidade

O modelo ANOVA foi aplicado aos dados ENADE descritos e ajustado considerando como variável dependente a média final e as restantes variáveis como fatores. A versão 24 do SPSS apresentou problemas de execução com elevado número de variáveis em particular quando cada uma delas tem diversas categorias tal como código do curso. O processador usado foi um Intel(R) Core(TM) i3-7100U CPU @ 2.40GHz com 8 GB de RAM. Esta limitação foi ultrapassada através da selecção de variáveis. Foram considerados 5 fatores: região, idade, género, raça/cor, educação do Pai e educação da Mãe.

As Tabelas de 5 a 8 apresentam os resultados da estatística de teste F e valor de prova, respectivamente para os dados originais, os dados anonimizados através do modelo de privacidade ℓ-diversidade (com c=3 e ℓ =5) e para os dados anonimizados através do modelo de privacidade t-proximidade com k=5 e t=0,15 e t=0,3. Os testes de hipóteses consideram, sob H0, que cada um dos fatores e cada um dos termos de interação são iguais a zero.

Através da análise efetuada à Tabela 5, verificamos que, com excepção do termo principal associado ao fator região, todos os demais termos principais e termos de interacção são estatisticamente significativos ao nível de significância de 5% (valor p < 0,05). Ou seja, de acordo com tais resultados e em presença de todos os termos aditivos, só não é possível rejeitar a hipótese nula para o efeito principal de região. Apesar disso, os termos de interação entre região e idade, região e educação do pai e da mãe, região e sexo, região e raça/cor autodeclarada constituem-se como grupos diferenciadores na sua relação com a variável dependente média final obtida pelo/a estudante. Notamos, adicionalmente, que a maioria dos termos é estatisticamente significativa ao nível de 1%. No entanto, após a anonimização (3, 5)-diversidade, para o mesmo nível de significância, a maior parte das variáveis deixa de ter impacto direto na explicação da variável dependente (Tabela 6). Apenas os fatores raça/cor autodeclarada e educação da mãe continuam como fator estatisticamente diferente de zero, na associação à média final obtida pela/o estudante. Quanto aos termos aditivos de interacção, os resultados também se modificam com o processo de anonimização. Entre os 15 termos de interação, 5 deixam de ser estatisticamente significativos ao nível de significância de 5%.

De forma diferente acontece com as duas parametrizações do modelo de privacidade t-proximidade (Tabelas 7 e 8). Embora registando alterações relativamente à distribuição original, a explicação das variáveis do preditor linear sobre a variável resposta é em tudo mais idêntica aos dados originais. Ora, isto pode sugerir uma distorção menos drástica dos dados por parte deste procedimento de anonimização. Em detalhe, verificamos que, mesmo em tais cenários de anonimização, os resultados nem sempre confirmam os obtidos com os dados originais. Compare-se a título de exemplo o efeito principal de região, que nas Tabelas 7 e 8 se constitui como fator diferenciador da média final do estudante e o termo de interação entre idade e raça/cor autodeclarada que deixa de ser estatisticamente significativo.

Tabela 5 – ANOVA com termos principais e interação, aplicado aos dados originais.

Fonte de variação	**F**	**Valor p**
Região	0,930	0,445
Idade	5,400	0,000
Género	8,139	0,004
Raça Cor	10,825	0,000
Educação Pai	9,760	0,000
Educação Mãe	8,819	0,000
*Idade * Educação Pai*	1,190	0,018
*Idade * Género*	3,079	0,000
*Idade * Educação Mãe*	1,344	0,000
*Idade * Raça Cor*	1,550	0,000
*Região * Idade*	1,559	0,000
*Género * Educação Pai*	17,223	0,000
*Educação Pai * Educação Mãe*	22,695	0,000
*Raça Cor * Educação Pai*	1,560	0,037
*Região * Educação Pai*	7,422	0,000
*Género * Educação Mãe*	28,581	0,000
*Género * Raça Cor*	23,235	0,000
*Região * Género*	10,456	0,000
*Raça Cor * Educação Mãe*	3,878	0,000
*Região * Educação Mãe*	3,885	0,000
*Região * Raça Cor*	12,860	0,000

Tabela 6 - ANOVA com termos principais e interação, (3,5)-diversidade.

Fonte de variação	**F**	**Valor p**
Região	2,135	0,074
Idade	0,384	0,535
Género	2,633	0,105
Raça Cor	9,621	0,000
Educação Pai	2,825	0,059
Educação Mãe	4,570	0,010
*Idade * Educação Pai*	0,483	0,617
*Idade * Género*	16,059	0,000
*Idade * Educação Mãe*	9,673	0,000
*Idade * Raça Cor*	2,296	0,076
*Região * Idade*	2,961	0,019
*Género * Educação Pai*	5,657	0,003
*Educação Pai * Educação Mãe*	14,162	0,000
*Raça Cor * Educação Pai*	1,129	0,341
*Região * Educação Pai*	1,998	0,043
*Género * Educação Mãe*	5,140	0,006
*Género * Raça Cor*	4,080	0,003
*Região * Género*	2,786	0,025
*Raça Cor * Educação Mãe*	1,887	0,079
*Região * Educação Mãe*	1,980	0,045
*Região * Raça Cor*	0,876	0,597

Considerando os casos válidos, os pressupostos do modelo de utilidade foram verificados para todos os conjuntos de dados. Apresentamos na Tabela 9 a assimetria, curtose e desvio padrão referentes à distribuição dos dados originais e à distribuição dos dados anonimizados com ℓ -diversidade (3,5). Tais estatísticas são as necessárias para usar o teste Jarque-Bera (Bera & Jarque, 1981; Greene, 2003) segundo o qual a normalidade da distribuição é testada sob H0. Aplicando o teste, em ambos os conjuntos de dados a hipótese nula não é rejeitada ao nível de significância de 5%. A comparação das estatísticas de distribuição para a variável dependente permitem-nos verificar que com o processo de anonimização a distribuição se altera, e.g. a curtose acentua-se. A alteração da distribuição já era esperada uma vez que no processo os casos extremos/raros são suprimidos ou agregados. Para os restantes conjuntos de dados os resultados conduzem a interpretação semelhante.

Tabela 7 - ANOVA com termos principais e interação, t-proximidade (k=5, t=0,15)

Fonte de variação	F	Valor p
Região	15,937	0,000
Idade	22,903	0,000
Género	7,801	0,005
Raça Cor	29,824	0,000
Educação Pai	1,869	0,154
Educação Mãe	13,390	0,000
*Idade * Educação Pai*	12,456	0,000
*Idade * Género*	14,365	0,000
*Idade * Educação Mãe*	5,224	0,005
*Idade * Raça Cor*	0,432	0,786
*Região * Idade*	5,879	0,000
*Género * Educação Pai*	2,925	0,054
*Educação Pai * Educação Mãe*	7,959	0,000
*Raça Cor * Educação Pai*	1,545	0,136
*Região * Educação Pai*	2,433	0,013
*Género * Educação Mãe*	7,447	0,001
*Género * Raça Cor*	12,115	0,000
*Região * Género*	4,396	0,001
*Raça Cor * Educação Mãe*	2,465	0,011
*Região * Educação Mãe*	2,905	0,003
*Região * Raça Cor*	10,381	0,000

6. Conclusões

Este trabalho analisou, para dados reais do sistema de ensino superior Brasileiro, estratégias para alcançar o equilíbrio entre privacidade e utilidade dos dados no processo de anonimização. Para estes dados verificou-se que, com classes de equivalência de dimensão 5, o que já garante um risco baixo de reidentificação, o modelo de t-proximidade pode levar a uma menor perda de registos do que o modelo de ℓ-diversidade recursiva, garantindo maior utilidade dos dados. Os nossos resultados também permitem verificar que os resultados do modelo de utilidade estão condicionados ao desenho do modelo de privacidade e podem tornar-se inúteis ou mesmo falaciosos. Neste caso, é necessário acautelar as possíveis interpretações substantivas e eventuais contribuições ou recomendações de política e prática, pois poderiam produzir efeito no sentido oposto ao que seria desejável. A comparação das estatísticas de distribuição referentes aos diferentes conjuntos de dados também nos permite afirmar que pressupostos teóricos estabelecidos para o modelo de utilidade podem deixar de se verificar após o processo de anonimização, podendo eventualmente comprometer a inferência estatística e a tomada de decisão subsequente.

Tabela 8 - ANOVA com termos principais e interação, t-proximidade (k=5, t=0,30).

Fonte de variação	**F**	**Valor p**
Região	21,292	0,000
Idade	23,688	0,000
Género	12,181	0,000
Raça Cor	83,183	0,000
Educação Pai	7,913	0,000
Educação Mãe	25,070	0,000
*Idade * Educação Pai*	12,076	0,000
*Idade * Género*	42,438	0,000
*Idade * Educação Mãe*	8,050	0,000
*Idade * Raça Cor*	0,769	0,545
*Região * Idade*	7,254	0,000
*Género * Educação Pai*	8,341	0,000
*Educação Pai * Educação Mãe*	37,334	0,000
*Raça Cor * Educação Pai*	3,309	0,001
*Região * Educação Pai*	3,637	0,000
*Género * Educação Mãe*	28,428	0,000
*Género * Raça Cor*	15,721	0,000
*Região * Género*	6,938	0,000
*Raça Cor * Educação Mãe*	2,107	0,032
*Região * Educação Mãe*	5,365	0,000
*Região * Raça Cor*	13,121	0,000

Tabela 9 – Estatísticas de distribuição

Conjunto de dados	N válido	N omisso	Assimetria	Curtose	Desvio padrão
Original	452 578	83 888	0,217	-0,344	14,392
(3, 5)-diversidade	88 931	3 060	0,102	-0,459	14,739

Agradecimentos

Este trabalho é parcialmente financiado pela FCT/MCTES através de fundos nacionais e quando aplicável cofinanciado por fundos comunitários no âmbito dos projetos UIDB/50008/2020 e CEMAPRE/REM - UIDB/05069/2020.

Referências

Adomavicius G. & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6) pp. 734–749.

Ayala-Rivera, V., McDonagh, P., Cerqueus, T. & Murphy, L. (2014). A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Privacy*. 7(3), pp. 337–370.

Bera, A., & Jarque, C. (1981). Efficient tests for normality, heteroscedasticity, and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, Vol. 7, 313–318.

Barth-Jones, D. (2012). The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. https://ssrn.com/abstract=2076397.

Cambon-Thomsen, A., Rial-Sebbag, E. & Knoppers, B. M. (2007). Trends in ethical and legal frameworks for the use of human biobanks. *Eur. Respiratory Journal*, 30(2), pp. 373–382. https://erj.ersjournals.com/content/30/2/373.

Chen, H, Chiang, R. H. L. & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, 36(4) pp. 1165–1188.

Chicaiza, J., Cabrera-Loayza, Ma. C., Elizalde, R., Piedra, N. (2020). Application of Data Anonymization in Learning Analytics. *In 3rd Int. Conf. on Applications of Intelligent Systems*, ACM, https://doi.org/10.1145/3378184.3378229.

Culnane, C., Rubinstein, B. I. P. & Teague, V. (2017). *Health data in an open world.* arXiv:1712.05627v1 [cs.CY].

Directive 95. (1995). http: //data.europa.eu/eli/dir/1995/46/oj.

Eicher, J., Bild, R., Spengler, H. *et al.* (2020). A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. *BMC Med Inform Decis Mak,* 20(29). https://doi.org/10.1186/s12911-020-1041-3

Esquivel-Quirós, L. G., Barrantes, E. G. & Darlington, F., E. (2019). Marco de medición de la privacidad. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (31), 66-81. https://dx.doi.org/10.17013/risti.31.66-81.

Francis, P. (2018). *Can anonymized data still be useful? part deux*. https://aircloak.com/can-anonymized-data-still-be-useful-part-deux/.

Fung, B. C. M., Wang, K., Fu, A. & Yu, P. S. (2010). *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, Taylor & Francis Group.

GDPR (2016). Regulation (EU) 2016/679, L 119, pp. 1–88. https://gdpr-info.eu/recitals/no-26/.

Goldstein, H. & Shlomo, N. (2020). A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. *Journal of Official Statistics*, Vol. 36, pp. 89–115.

Greene, W.H. (2003). Econometric Analysis (5th edition). New York: Prentice Hall.

Kaye, J., Meslin, E., Knoppers, B., Juengst, E., Deschênes, M., Cambon-Thomsen, A., Chalmers, D., Vries, ., Edwards, K., Hoppe, N., Kent, A., Adebamowo, C. Marshall, P., & Kato, K. (2012). Elsi 2.0 for genomics and society. *Science*, vol. 336, pp. 673–674.

Kniola, L. (2017). Plausible Adversaries. *In Re-Identification Risk Assessment. PhUSE Annual Conference.*

Koch, R. (2020). Political campaigns and your personal data. ProtonMail; https://protonmail.com/blog/political-campaigns-and-your-personal-data/.

Lee, H., Kim, H., Kim, J. W. & Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *Medical Informatics and Decision Making*. 17(104). DOI 10.1186/s12911-017-0499-0.

LeFevre, K., DeWitt, D., & Ramakrishnan, R., (2005). Incognito: Efficient full-domain k-anonymity. In ACM *SIGMOD Int. Conf. on Management of Data*, pp.49–60.

LeFevre, K., DeWitt, D J. & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. *In 22nd Int. Conf. on Data Engineering (ICDE'06).*

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *In IEEE 23rd Int. Conf. on Data Eng.*, pp. 106–115.

Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), article 3, 52 pages. https://doi.org/10.1145/1217299.1217302.

Panduragan. (2014) *On taxis and rainbows.* https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

Prasser, F., Eicher, J., Spengler, H., Bild, R. & Kuhn, K. A. (2020). Flexible data anonymization using ARX—Current status and challenges ahead. *Softw Pract Exper*. Vol. 50, pp. 1277–1304. https://doi.org/10.1002/spe.2812.

Prasser, F. & Kohlmayer, F. (2015) Putting statistical disclosure control into practice: the ARX data anonymization tool. In A. Gkoulalas-Divanis and G. Loukides (Ed.s) Medical Data Privacy Handbook. Cham, Switzerland: Springer International Publishing, p.111–48.

Quiñonez, Y., Lizarraga, C., Peraza, J., & Zatarain, O. (2019). Sistema inteligente para el monitoreo automatizado del transporte público en tiempo real. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (31), 94-105. https://dx.doi.org/10.17013/risti.31.94-105.

Rubner, Y., Tomasi, C. & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2), pp. 99–121.

Santos, W., Sousa, G., Prata, P. & Ferrão, M. E. (2020). Data Anonymization: K-anonymity Sensitivity Analysis. 15th Iberian Conf. on Information Systems and Technologies (CISTI) pp. 1-6, doi: 10.23919/CISTI49556.2020.9141044.

Scheffé, H. (1999). The analysis of variance. New York and London: Wiley.

Spengler, H. & Prasser, F. (2019). Protecting biomedical data against attribute disclosure. Studies in health technology and informatics, 267, pp. 207–214.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Carnegie Mellon Univ., http://dataprivacylab.org/projects/identifiability/.

Sweeney, L. (2002a). K-anonymity: A model for protecting privacy. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5) pp. 557–570.

Sweeney, L. (2002b). Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10(5) pp. 571–588.

Sweeney, L. (2015). Only you, your doctor, and many others may know. *Technology* Science. https://techscience.org/a/2015092903/.

Sweeney, L. Loewenfeldt, M. von & Perry, M. (2018). Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. Technology Science. https://techscience.org/a/2018111301/.