TOWARD A UNIVERSAL PRIVACY AND INFORMATION-PRESERVING FRAMEWORK FOR INDIVIDUAL DATA EXCHANGE

Nicolas Ruiz

UNIVERSITAT
ROVIRA i VIRGILI

Toward a universal privacy and information-preserving framework for individual data exchange

NICOLAS RUIZ

DOCTORAL THESIS
2018

Nicolas Ruiz

Toward a universal privacy and information-preserving framework for individual data exchange

DOCTORAL THESIS

Advisors

Dr. Josep Domingo-Ferrer

and Dr. Krishnamurty Muralidhar

Department of Computer Engineering and Mathematics

UNIVERSITAT ROVIRA i VIRGILI

Tarragona

2018

UNIVERSITAT ROVIRA i VIRGILI

WE STATE that the present study, entitled "*Toward a Universal Privacy and Information-Preserving Framework for Individual Data Exchange*", presented by *Nicolas Ruiz* for the award of the degree of Doctor, has been carried out under our supervision at the Department of Computer Engineering and Mathematics of this university.

Tarragona, 2018

Doctoral Thesis Supervisors

Dr. Josep Domingo-Ferrer

Dr. Krishnamurty Muralidhar

Abstract

Data on individual subjects provide a rich amount of information that can inform statistical and policy analysis in a meaningful way. However, due to the legal obligations surrounding such data, this wealth of information is often not fully exploited in order to protect the confidentiality of respondents. While statistical disclosure control research has historically provided the analytical apparatus through which data on individuals can be disseminated in such a way so as to preserve both privacy and information way, in recent years the literature has burgeoned in many directions, leading to a lack of a comprehensive view on best practices. Against this backdrop, this thesis focuses on establishing some common grounds for individual data anonymization by developing some new universal tools. We begin by proposing some universal measures of disclosure risk and information loss that can be computed in a simple fashion and used for the evaluation of any anonymization method, independently of the context in which they operate. Building on these measures, we then propose a new approach to data anonymization by formulating a general cipher based on permutation keys, which appears to be equivalent to a general form of rank swapping. Beyond the existing methods that this cipher can universally reproduce, it also offers a new, more efficient way to practice data anonymization, based on the ex-ante exploration of different permutation structures. Finally, we extend these new insights to two areas, longitudinal and synthetic data. For the former, we develop a specific anonymization framework, while for the latter it is established that the distinction made in the literature between non-synthetic and synthetic data is in fact artificial.

Resum

Les dades sobre subjectes individuals proporcionen una gran quantitat d'informació que pot ser molt útil per a l'anàlisi estadística i per a la planificació. Tanmateix, a causa de les obligacions legals que envolten aquesta mena de dades, sovint aquesta riquesa d'informació no s'explota totalment per tal de protegir la confidencialitat dels enquestats. Tot i que la recerca sobre el control de la revelació estadística històricament ha proporcionat l'aparell analític a través del qual es poden difondre dades útils sobre persones de manera compatible amb llur privadesa, en els darrers anys la literatura ha anat florint en moltes direccions, cosa que ha dut a una manca de visió de conjunt sobre les millors pràctiques. En aquest context, aquesta tesi se centra a establir un terreny comú per a l'anonimització de dades individuals desenvolupant algunes noves eines universals. Començarem proposant unes mesures universals de risc de divulgació i de pèrdua d'informació que poden calcular-se de manera senzilla i fer-se servir per avaluar qualsevol mètode d'anonimització, independentment del context en el qual operi. Partint d'aquestes mesures, proposem una nova aproximació a l'anonimització de dades mitjançant la formulació d'un xifratge general basat en claus de permutació, que resulta equivalent a una forma general d'intercanvi de rangs. Més enllà de reproduir mètodes existents de forma universal, aquest xifratge també ofereix una manera nova i més eficient de practicar l'anonimització de dades, basada en l'exploració ex ante de diferents estructures de permutació. Finalment, ampliem aquestes noves idees a dues àrees, dades longitudinals i dades sintètiques. Per a les primeres, desenvolupem un marc específic d'anonimització, mentre que per a les segones constatem que la distinció feta a la literatura entre dades no sintètiques i sintètiques és de fet artificial.

Resumen

Los datos sobre individuos proporcionan una gran cantidad de información que puede guiar el análisis estadístico y de políticas de una manera significativa. Sin embargo, debido a las obligaciones legales que rodean dichos datos, esta gran cantidad de información a menudo no se explota completamente para proteger la confidencialidad de los encuestados. Si bien la investigacion en el campo del control de la revelación estadística ha proporcionado históricamente el aparato analítico a través del cual los datos sobre individuos pueden diseminarse de tal manera que se preserve la privacidad y la información, en los últimos años, la literatura ha florecido en muchas direcciones, dando lugar a una falta de visión completa de las mejores prácticas. En este contexto, esta tesis se centra en establecer algunas bases comunes para la anonimización de datos individuales mediante el desarrollo de algunas herramientas universales nuevas. Comenzamos proponiendo algunas medidas universales de riesgo de divulgación y pérdida de información, que pueden computarse de manera simple y utilizarse para la evaluación de cualquier método de anonimización, independientemente del contexto en el que operan. Sobre la base de estas medidas, proponemos un nuevo enfoque para la anonimización de datos mediante la formulación de un cifrado general basado en claves de permutación, que seria equivalente a una forma general de intercambio de rango. Más allá de los métodos existentes, que este cifrado puede reproducir universalmente, también ofrece una forma nueva y más eficiente de anonimizar los datos, basada en la exploración ex ante de diferentes estructuras de permutación. Finalmente, ampliamos estos nuevos conocimientos a dos áreas, datos longitudinales y sintéticos. Para el primero, desarrollamos un marco de anonimización específico, mientras que para el segundo se establece que la distinción hecha en la literatura entre datos sintéticos y no sintéticos es de hecho artificial.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Motivation

Data on individual subjects are increasingly gathered and exchanged. By their nature, they provide a rich amount of information that can inform statistical and policy analysis in a meaningful way. However, due to the legal obligations surrounding these data, this wealth of information is often not fully exploited in order to protect the confidentiality of respondents and to avoid privacy threats. In fact, such requirements shape the dissemination policy of individual data at national and international levels. The issue is how to ensure a sufficient level of data protection to meet releasers' concerns in terms of legal and ethical requirements, while still offering users a reasonable level of information. Over the last decade the role of micro data has changed from being the preserve of National Statistical Offices and government departments to being a vital tool for a wide range of analysts trying to understand both social and economic phenomena. This has raised a range of questions and concerns about the privacy/information trade-off and the quest for best practices that can be both useful to users but also respectful of respondents' privacy.

Statistical disclosure control (SDC) research has a rich history of addressing those issues by providing the analytical apparatus through which the privacy/information trade-off can be assessed and implemented. SDC consists in the set of tools that can enhance the level of confidentiality of any data while preserving to a lesser or greater extent its level of information. Over the years, the literature has burgeoned in many directions. In particular, techniques applicable to micro data, which are the focus of this thesis, offer a wide variety of tools to protect the confidentiality of respondents while maximizing the information content of the data released, for the benefits of society at large. Such diversity is undoubtedly useful but has several major drawbacks.

First, there is a clear lack of agreement and clarity on the appropriate choice of tools in a given context, and as a consequence, no comprehensive view (or at best an incomplete one) of the relative performances of the techniques available. The practical scope of current micro data protection methods is not fully exploited precisely because there is no overarching framework: all methods generally carry their own analytical environment, underlying approaches and definitions of privacy and information.

As a consequence, beyond the choice of method is a second issue that the cross-evaluation of current micro data masking methods is also a challenging task, for at least two reasons. The first is analytical: the evaluation of utility and privacy for each method is metric and data-dependent. As a result, there is no common language for comparing different mechanisms, all with potentially varying parametrizations applied on the same data set or different data sets. Moreover, there is a variety of definitions of privacy and information loss, and picking one is often related to the context in which it is used and can result from an arbitrary choice. The fact that all evaluations can only be practical in nature and context-specific is clearly problematic, not least because this precludes a sound and simple communication on data anonymization as well as a wider democratization of the field that could allow for more data to be disseminated.

Finally, a third issue is related to the variety of parties involved in micro data exchange. Indeed, it is natural to suppose that across parties, different sensitivities to privacy and information will prevail. Some may place greater emphasis on the preservation of privacy, e.g. typically the data releasers, while others may be more concerned with the extent to which information is preserved, e.g. typically the researchers. These sensitivities can additionally differ within groups, e.g. one researcher may have a low sensitivity to information loss and consider a release better than no release at all, while another could simply disregard the data above a certain threshold of loss set according to his intended use of the data.

It is based on these considerations that this thesis will focus on establishing some common grounds for individual data anonymization by developing a new, universal approach relying on permutations. In fact, permutations happen to be the essential principle upon which individual data anonymization can be based. This principle allows the proposal of a universal analytical environment that can be used to evaluate the information/privacy outcomes of any anonymization method applied on any type of data in a universal way. But such an analytical environment can also be used to conduct anonymization directly under the form of a cipher. This cipher can also replicate any methods currently available in the literature, whatever the original technical apparatus of these methods and independently of the nature of the data to which these methods are applied. Finally, this new environment offers the possibility to capture, in a continuous and selectable way, the variety of views across all agents interacting in an individual data exchange

1.2 Structure and contributions of this thesis

This thesis and its contributions are organized as follows:

- Chapter 2 presents a state of the art on individual data protection and transaction, notably the broad approaches available in the SDC literature. It also outlines the main commonalities and differences with cryptography. A description of the recent functional equivalence in anonymization for ex-post evaluation, as established by the permutation-based paradigm and upon which this thesis relies, is also proposed.

- Chapter 3 presents a context-oriented, specific contribution to the protection of indvidual data, with the goal of preserving positive skewness. While many economic variables are distributed according to a heavy tailed, asymmetric form that makes the normality assumption unsuitable, several popular perturbation techniques use this assumption nevertheless. The multiplicative masking method proposed, based on lognormal distributions, allows for the generation of perturbed data that are similar to the original data to a degree that is selected by the user, depending on his requirements regarding the protection of individuals away from the mean. This noise-based method is classical in its approach. But in addition to having its own potential range of application, it serves to illustrate that whatever the specific context upon which anonymization is meant to operate, and the context of this method is rather specific, anonymization all boils down to permutations.
- Chapter 4 explores the first consequences of the permutation-based paradigm in anonymization. It proposes some universal measures of disclosure risk and information loss that can be computed in a simple fashion and used for the evaluation of any anonymization method, independently of the context under which they operate. In particular, they exhibit distributional independence. The construction of these measures allows for the notions of dominance in disclosure risk and information loss to be introduced in data anonymization, which formalise the fact that different parties involved in micro data release can have different sensitivities to privacy and information, and can inform as to which methods can be used to reach a consensus among all parties involved. These two notions of dominance can in fact identify which methods, under any tastes for privacy and information, will always perform better than others. A graphical representation of disclosure risk and information loss is also introduced.

- Chapter 5 develops a new approach to data anonymization by proposing a general cipher based on permutation keys, which appears to be equivalent to a general form of rank swapping. Beyond the existing methods that this cipher can universally reproduce, it also offers a new way to practice data anonymization based on the exploration of different permutation structures. This cipher can be used to perform anonymization in an ex-ante way instead of being engaged in several ex-post evaluations and iterations to reach the protection and information properties sought after. The subsequent study of the cipher's properties additionally reveals certain new insights as to the nature of the task of anonymization taken at a general level of functioning. Finally, to make this cipher operational, permutation menus in data anonymization are introduced, where the measures developed in Chapter 4 are used ex-ante for the calibration of permutation keys. To justify the relevance of their use in an ex-ante context, a theoretical characterization of these measures is also proposed.

- Chapter 6 tackles the specific issue of longitudinal data anonymization. Despite the fact that the SDC literature offers a wide variety of tools suited to different contexts and data types, there have been very few attempts to deal with the challenges posed by longitudinal data. This Chapter develops a general framework and some associated metrics of disclosure risk and information loss, tailored to the specific challenges posed by longitudinal data anonymization. To do so, it builds on a permutation approach where the effect of time on time-variant attributes can be seen as an anonymization method that can be captured by temporal permutations. This approach allows the analytical alignment of the specificities of longitudinal data with the cipher developed in Chapter 5.

- Using the insights of the preceeding chapters, Chapter 7 aims at challenging the information and privacy guarantees of synthetic data. It shows that in fact any synthetic data set can always be expressed as a permutation of the original data, in a way similar to non-synthetic SDC techniques. This result offers applications for the disclosure risk assessment of synthetic data but also beyond. For one thing, it is always possible to release synthetic data sets with the same privacy properties but with an improved level of information, because the marginal distributions can always be preserved without increasing risk. On the privacy front, it leads to the consequence that the distinction made in the literature between non-synthetic and synthetic data is not so clear-cut. The subsequent simulation of an attack on synthetic data shows that the practice of releasing several synthetic data sets for a single original data set entails privacy issues that do not arise in non-synthetic anonymization (where typically only one anonymized data set is released). Indeed, the multiple releases can lead to better privacy guarantees, by confusing the attacker, or instead facilitate attribute disclosure by narrowing the range of the possible values that the attacker is trying to retrieve.

- Finally, Chapter 8 gathers a summary of the results presented in this thesis and the list of publications supporting them. Some guidelines for future research are also proposed.

2 STATE OF THE ART

2.1 Introduction

In this Chapter we review the general approach to individual data transaction and the basics of SDC methods, as well as their commonalities and differences with cryptography. We then turn to the description of the recent functional equivalence in anonymization for ex-post evaluation established by the permutation paradigm, upon which the bulk of this thesis is based.

2.2 Transaction on individual data

A general and standard way of describing a transaction of individual data is to consider two types of agents: a data releaser, that supplies individual data, e.g. public administrations, enterprises, and data users, who demand individual data, e.g. researchers, public administrations, enterprises. The former typically gathers, under some suitable forms, a micro data set that is data collected from individuals. The latter will have various needs in terms of information and seek the data in order to conduct a potentially large variety of tasks. Note that in this simple setting we assume trustworthiness on the supplier side, meaning that the data releaser knows the identities of the respondents who contributed to the data set out of their good will. Moreover, we do not restrict the set of potential tasks to be conducted by the data users, which thus can range from simple data mining tasks such as frequencies counts and computation of the mean and median of a distribution, to more elaborate tasks such as econometric techniques. This is equivalent to considering that the data releaser is not equipped with sufficient technical knowledge to conduct the different tasks that the users have in mind. Thus, data are released without being tailored to specific needs.

The delivery of the micro data set by the data releaser to the data users, via any potential channel, is what characterizes a transaction of individual data. The users then go away with the data to perform some tasks on them without any further interaction with the releaser. As such and as previously defined in the literature, the transaction is a standard non-interactive one [21]. Naturally, other types of transaction are possible: for example, under the assumption that the data releaser has sufficient technical knowledge, data mining tasks could be performed on the data by the releaser upon request of the users, and the former will communicate the outputs of the tasks to the latter. For such an interactive transaction, differential privacy has gained strong momentum in the literature to conceptualize and tackle the issues that could arise in terms of privacy protection. However, some questions remain unresolved, such as the quality of the output that is delivered to the users in terms of information [48]. Moreover, and because in an interactive transaction a mechanism is in place between the releaser and the users in order to perform the tasks, it is ultimately outputs that are delivered to the users, not data *per se*. As a result, one has to make some untenable assumptions about the users' needs, by inevitably restraining them or similarly assuming a very expert data releaser that can perform any kind of task. As noted, this is not what we will assume in this thesis, not least because such assumptions would lead to unrealistic, or at most, highly specific forms of data transactions. Given these limitations, the scope of this thesis is thus voluntarily narrowed to the non-interactive exchange of data sets.

Non-interactive data transactions immediately raise the pressing question of privacy, even more so than in other forms of exchange. In modern societies with pervasive data collection, it is a matter of general interest to grant access to individual data, but not to the detriment of privacy, a fundamental right for all individuals. The exchange of individual data in their original form, as collected by the releaser, generally entails a violation of individual privacy given the sensitive information that the data can contain. This is why privacy legislation that prevails in most countries precludes the dissemination of data that are linkable to individuals, or allows the recovery of only some of their characteristics. So, in order to prevent any disclosure of individuals' information/identity, data have first to be anonymized through the application of suitable statistical disclosure control (SDC) techniques.

2.3 Statistical disclosure control

SDC research has a long and rich history in providing data releasers with a set of tools for anonymizing individual data under various settings [26]. In a nutshell, for non-interactive data exchange, the overall approach of SDC is for a data releaser to modify the original data set in some ways that reduce disclosure risk while altering the information that it contains as little as possible. At a general level, SDC techniques can be classified into two main approaches:

- *Privacy-first*: the method is applied with the primary goal of complying with some pre-requisites on the level of privacy, judged as acceptable and under which data exchange can take place.
- *Utility-first*: the method is applied with the primary goal of complying with some pre-requisites on the level of information, judged as valuable enough to make data exchange worthwhile.

The privacy-first approach shares certain features with cryptography. Indeed, the act of protecting privacy through anonymization can be conceived as a form of encryption, where it is the individuals’ identities that are encrypted. However, the utility-first approach establishes a first fundamental difference with cryptography, as it would be pointless to release micro data that contain no information at all. So, while the very goal of cryptography is to release a cypher text that discloses nothing whatsoever about the underlying plaintext, the purpose of individual data exchange is to release data (i.e. the cypher text) considered as safe as possible in terms of privacy, while purposefully leaking some information (and generally the more the better). A second fundamental difference lies in the types of agents involved and how the transaction operates.

In cryptography, a sender encrypts a message and the receiver decrypts it with the appropriate key, while an attacker tries to intercept the message and to decipher it using cryptanalysis techniques. In an individual data exchange, first, there is ideally no decryption phase: the data user takes the released data set as given for his analysis needs. Second, while in cryptography there is a clear distinction between sender, receiver and attacker, in an individual data exchange the receiver can also be an attacker. Indeed, a malevolent user could potentially try to re-identify individuals in a data set and the data releaser has no way of preventing this after the exchange takes place (nor would an ex-ante screening of the users to identify the reliable ones preclude, in principle, that they become attackers). Finally, a third difference is that the re-identification of individuals, which constitutes an attack in data anonymization, carries a different meaning than an attack in cryptography. Indeed, while in the latter case the single objective is generally to retrieve the full plaintext, in the former this is not necessarily so: the re-identification of at least one individual can be considered as a successful attack. Thus, the cryptographic viewpoint of an attack in data anonymization is about identifying some individuals (or retrieving some information about them) but not necessarily all of them, i.e. some of the plaintext but not necessarily all of it. To summarize, while in principle micro data are not meant to be deciphered, the releaser must sufficiently encipher the data so as to prevent any re-identification of individuals, while at the same time ensuring that the data contain a sufficient level of information to be meaningful to most users. Here lies the fundamental trade-off in individual data exchange that is not present in cryptography: encryption, i.e. privacy preservation, versus information leakage. The goal of SDC techniques is to manage this trade-off in a meaningful and practical way.

To achieve this goal, a wide variety of tools is available. In terms of operating principles, such tools can be classified as follows:

- *Non-perturbative*: The level of details in the data are reduced or suppressed before release. Sampling (only a sub-sample of the original data is released), global recoding (some continuous variables are discretized and/or some categorical variables are coarsened) and local suppression (a combination of variables judged as unsafe are deleted) form the bulk of non-perturbative approaches.
- *Perturbative*: data are altered before release yet it must be ensured that the altered data do not depart significantly from the original data, so that information loss does not reach a level that makes the release worthless to users. Noise-based methods (e.g. noise addition, multiplicative noise such as the method presented in Chapter 3), cluster-based methods (where records are clustered into small aggregates of size at least k, e.g. microaggregation, univariate or multivariate), rank-based methods (where the values of selected variables are exchanged among individuals according to some criteria, e.g. rank swapping) are amongst the most popular approaches for perturbative disclosure control.
- *Synthetic*: the data released are simulated with the constraint that certain statistics and relationships across variables should be preserved. It represents a departure from non-perturbative and perturbative approaches in the sense that, generally, a synthetic release does not contain any original data, while for the former the original data, albeit altered, are disseminated. Chapter 7 notably shows that this distinction appears to be artificial and that synthetic data can in fact always be thought of in terms of the original data.

Over the years, research in SDC has led to the development of a wide variety of tools, suited for many circumstances and spanning several possible types of data across different fields. This diversity of available techniques is undoubtedly an asset but it entails certain drawbacks. As mentioned in Chapter 1, the lack of an overarching framework upon which the trade-off between utility and disclosure risk can be assessed is problematic because it leads to an absence of consensus regarding "best practices". In fact, the current state of the literature, while high in quality, offers at best techniques that are tied to the context upon which they operate. For example, comparing the level of utility and privacy achieved by different methods on different data sets is an awkward task as different metrics and/or different parametrizations are largely heterogeneous, so that no common ground exists for comparison. This is generally why only ad-hoc comparisons can be conducted [12]. Additionally, each metric embodies distributional dependence and this feature has a significant impact on the performance evaluation of SDC methods across data sets [35]. Moreover, even in a utility (resp. privacy)-first approach, it is advisable to check the value of privacy (resp. utility) achieved by a method before data dissemination, which thus always lead to the limitation of context-dependence discussed above.

To address these issues and the need to generalize the concepts used in SDC, a recent contribution to the literature proposed a general functional equivalence based on permutations to describe any data masking method (see [39] and its subsequent development in [12]). This equivalence forms the building block upon which disclosure risk and information loss can in fact be measured in a universal fashion (Chapter 4), but also constitutes a general method in itself to conduct data anonymization (Chapter 5).

2.4 The permutation-based paradigm

The permutation paradigm in data anonymization starts from the observation that any anonymized data set can be viewed as a permutation of the original data plus a non-rank perturbative noise addition. It thus establishes that all masking methods can be thought of in terms of a single ingredient, i.e. permutation. This result clearly has far reaching conceptual and practical consequences, in the sense that it provides a single and easily understandable reading key, independent of the model parameters, the risk measures or the specific characteristics of the data, to interpret the utility/protection outcome of an anonymization procedure.

To illustrate this equivalence, we use a toy example which consists (without loss of generality) of five records and three attributes X=(X_1, X_2, X_3) generated by sampling N(10,10^2), N(100,40^2) and N(1000,2000^2) distributions, respectively. Noise is then added to obtain Y=(Y_1, Y_2, Y_3), the three anonymized version of the attributes, from N(0,5^2), N(0,20^2) and N(0,1000^2) distributions, respectively. One can see that the masking procedure generates a permutation of the records of the original data (Table 2.1).

Table 2.1 An illustration of the permutation paradigm

Original dataset X			Masked dataset Y		
X_1	X_2	X_3	Y_1	Y_2	Y_3
13	135	3707	8	160	3248
20	52	826	20	57	822
2	123	-1317	-1	122	248
15	165	2419	18	135	597
29	160	-1008	29	164	-1927

Rank of the original attribute			Rank of the masked attribute		
X_{1R}	X_{2R}	X_{3R}	Y_{1R}	Y_{2R}	Y_{3R}
4	3	1	4	2	1
2	5	3	2	5	2
5	4	5	5	4	4
3	1	2	3	3	3
1	2	4	1	1	5

Now, as long as the attributes' values of a data set can be ranked, which is obvious in the case of numerical and categorical ordinal attributes, but also feasible in the case of nominal ones [14], it is always possible to derive a data set Z that contains the attributes X_1, X_2 and X_3, but ordered according to the ranks of Y_1, Y_2 and Y_3, respectively, i.e. in Table 2.1 re-ordering (X_1, X_2, X_3) according to (Y_{1R}, Y_{2R}, Y_{3R}). This can be done following the post-masking reverse procedure outlined in [39]. Finally, the masked data Y can be fully reconstituted by adding small noises (E_1, E_2, E_3) (small in the sense that they cannot re-rank Z while they can still be large in absolute values) to each observation in each attribute (Table 2.2).

Table 2.2 Equivalence in anonymization: postmasking reverse mapping plus noise addition

Original dataset X			Reverse mapped dataset Z		
X_1	X_2	X_3	Z_1	Z_2	Z_3
13	135	3707	13	160	3707
20	52	826	20	52	2419
2	123	-1317	2	123	-1008
15	165	2419	15	135	826
29	160	-1008	29	165	-1317

Noise E			Masked dataset Y(=Z+E)		
E_1	E_2	E_3	Y_1	Y_2	Y_3
-5	0	-459	8	160	3248
0	5	-1597	20	57	822
-3	0	1256	-1	122	248
2	0	-229	18	135	597
0	-1	-610	29	164	-1927

By construction, Z has the same marginal distributions as X, which is an appealing property. Moreover, as will be discussed in Chapter 5, the small noise addition turns out to be irrelevant: re-identification can only come from permutation, as by construction noise addition cannot alter ranks. Reverse mapping thus establishes permutation as the overarching principle of data anonymization, allowing the functioning of any method to be viewed as the outcome of a permutation of the original data, independently of how the method operates. This result has been explicitly proposed by its authors for the ex-post evaluation of anonymization, but not as a new technique for conducting anonymization. As we will see, it can in fact be viewed and operationalized as a new, general framework for anonymization (Chapter 5).

To conclude, it should be mentioned that this result may seem surprising, and one might ask why the fundamental principle of data anonymization ultimately appears to be as simple as permutation. After all, in cryptography, permutation ciphers and their cryptanalyses have been known for centuries. They are easy to detect because they do not affect individual symbols' frequencies (the equivalent of this in the permutation paradigm being the preservation of marginal distributions). In fact, as will be discussed in Chapters 4 and 5, it turns out that the obvious weakness of a permutation cipher in standard cryptography shows up as a strength in data anonymization, in that the degree of permutation performed allows controlling for the amount of information that is leaked. Moreover, because the permutation paradigm proposes one single universal language for data anonymization, it allows introducing some measures of disclosure risk and information loss that can be used in any context, that are flexible enough to capture the variety of views that can occur in a data exchange (Chapter 4). While these measures are originally proposed for the ex-post evaluation of the outcomes of any anonymization techniques on any data, they can in fact be used equally validly ex-ante to perform anonymization (Chapter 5).

3 A MASKING METHOD TO PRESERVE THE SKEWNESS OF INDIVIDUAL DATA

3.1 Introduction

As we saw in the previous Chapter, a possible approach for data perturbation consists in the matching of the original data with random noise terms. This can be performed in various ways, from a simple additive structure to non-linear transformations, applicable to both categorical and numerical variables. However, most of the perturbation techniques focus on continuous variables and so will the methodology presented in this Chapter.

In practice, popular perturbation techniques [6,8,38] use an additive structure for noise application, where error terms are randomly drawn from a normal distribution, the latter being data-dependently parameterized in such a way that the resulting distribution of the perturbed values have the same first and second order moments as those in the original data. As information on these two moments is sufficient to fully identify a normal distribution, this implies that if the original values follow a normal law then the original and the perturbed values will have exactly the same distribution. The loss of statistical information is thus low, in that only the values of the data points of the underlying distribution are altered but their overall shape is not. Such a high degree of preservation is made possible by the use of the Gaussian framework. Apart from its peculiar properties, the choice of additive noise methods is motivated by the fact that normality underlies many statistical and econometric tools, extending thus the usefulness and audience for these techniques.

Additive noise methods nevertheless have some drawbacks. The most obvious and crucial is the amount of information that is lost when the original data do not follow a normal law. In this case, analysis performed on perturbed data could produce quite different results from those performed on the original set. In particular, the Gaussian framework implies a strong assumption of symmetry in the original distribution. Clearly, for numerous economic variables, this assumption is too strong to be tenable.

In fact, micro data often exhibit positively skewed distributions, as in the case of household income and wealth. Recent studies relying on a growing stream of research on income inequality [41,3] have pointed out that in most developed countries top incomes contribute disproportionately to the overall level of income inequality in a country. As a result, skewness matters, and perturbation methodologies preserving it are of central interest for SDC, despite its lack of treatment in the literature (see [33] for an exception). In such cases, lognormal distributions appear to display a reasonable approximation for a large range of economic variables [28,31]. As such, Gaussian perturbation methods would be of limited utility when applied to such distributions for at least two reasons:

- First, the sum of skewed and non-skewed distribution provides an identifiable distribution in very rare cases [22,29]. Thus perturbed datasets will, in most cases, follow unknown and unidentifiable distributions.

- Second, as the presence of observations far from the mean leads to a skewed distribution, it follows that adding noise drawn from a normal distribution to those observations will only weakly perturb them. As an example, very large firms in business surveys will still be subject to high disclosure risk after perturbation, hence raising the issue of protection and confidentiality.

This Chapter presents a new multiplicative masking method that preserves positive skewness of the original data based on lognormal distributions. This method allows users to generate perturbed data that are similar to the original data to a degree that is selected by the user. The methodology preserves confidentiality constraints in particular for observations away from the mean, by permuting them in the sample during the perturbation process. Despite the fact that this method aims at offering a solution in a rather specific data-context, its outcome can ultimately be appraised in term of permutations. The contributions in this Chapter have been published in [47].

3.2 Methodology

This section describes the proposed methodology for preservation of asymmetric distribution based on the identification of sufficiency conditions for lognormal distributions. To fully appraise the departure from additive Gaussian methods, we first describe the latter using a methodology proposed in [36], showing how it is possible to generate perturbed data that preserves the distribution of the original data set with a selectable degree of similarity.

3.2.1 The Muralidhar-Sarathy (MS) hybrid generator

Let’s assume that X is a confidential variable that we want to perturb, and that S is a non-confidential variable with a low level of identification risk. Without loss of generality, it is assumed that the means of X and S are equal to zero. Let σ_{XX}^2, σ_{SS}^2 and σ_{SX}^2 be respectively the variance of X, S and the covariance between X and S. We will denote by Y the perturbed value of X generated by the following equation (where

$y_i, x_i, s_i\ \forall i = 1, \dots, n$ are the values of Y, X and S variables for the i^th^ respondent in the dataset):

$$y_i = \left[(1-\alpha)\frac{1}{n}\sum_{i=1}^{n} x_i - \beta\frac{1}{n}\sum_{i=1}^{n} s_i\right] + \alpha x_i + \beta s_i + u_i\ \forall i = 1, \dots, n$$

α and β are coefficients and u_i is a random term generated from a normal distribution $N(0, \sigma_{uu}^2)$, satisfying $\frac{1}{n}\sum_{i=1}^{n} x_i u_i = \frac{1}{n}\sum_{i=1}^{n} s_i u_i = 0$ ($x_i$and s_i are orthogonal to u_i). This equation shows that α can be interpreted as a similarity parameter between Y and X. When α=0, X and Y are completely dissimilar. For α=1 Y equals X and no perturbation is added. Thus, the choice of α allows the user (e.g. National Statistical Offices in the case of records from official sources) to control for the degree of similarity between the original and the perturbed variable that will be disseminated.

The conversion of X into Y through the preceding equation adds ‘noise’ to the original variable X. In fact, it is easy to verify that $E(y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$ and thus that X and Y will have the same expectation: the first moment of X’s distribution is then preserved. To preserve the second moment, the following condition must be satisfied:

$$\sigma_{XX}^2 = \sigma_{YY}^2 = E[(\alpha x_i + \beta s_i + u_i)(\alpha x_i + \beta s_i + u_i)]$$

$$= \alpha^2\sigma_{XX}^2 + \beta\sigma_{SS}^2 + \sigma_{uu}^2 + 2\alpha\beta\sigma_{SX}^2$$

Finally, in order to preserve the covariance between the confidential and non-confidential variables, the following equation must also hold:

$$\sigma_{SX}^2 = \sigma_{SY}^2 = \alpha\sigma_{SX}^2 + \beta\sigma_{SS}^2$$

$$\Leftrightarrow\ \beta = (1-\ \alpha)\frac{\sigma_{SX}^2}{\sigma_{SS}^2}$$

Combining the two preceding equations above, we obtain the following restriction for σ_{uu}^2:

$$\sigma_{uu}^2 = (1-\alpha^2)\left[\sigma_{XX}^2 - \frac{(\sigma_{SX}^2)^2}{\sigma_{SS}^2}\right]$$

The term $\left[\sigma_{XX}^2 - \frac{(\sigma_{SX}^2)^2}{\sigma_{SS}^2}\right]$ is always greater than or equal to zero. Thus, the necessary and sufficient condition to have σ_{uu}^2>0 is that -1≤α≤1. As a negative α induces a negative correlation between the original and the perturbed value, this case is ignored in the following, i.e. we will focus only on 0≤ α≤1 to fulfil the above restrictions.

When α is set to 1, X=Y and no perturbation is added; when α=0, Y is not a function of the (confidential) value X but only of the non-confidential variable S and of an error term. The intermediary cases where 0<α<1 therefore create a hybrid dataset, as the released variable is a combination of its original value, of the non-confidential variable S and of a noise term. Through this method, users can thus choose to which extent they want to protect their initial release. This procedure is perfectly secure in the sense that no reverse engineering is possible as the hybridation is performed using a random draw for u_i. A direct consequence of this algorithm is that users can choose to communicate transparently their chosen degree of dissimilarity: in other terms, knowledge of α provides access to the value of σ_{uu}^2 but not to the u_i values themselves.

While it can be argued that this method implies significant information loss, statistical information is actually preserved to a greater degree than with other approaches [20]. In particular, the MS method preserves the first two moments of variable X's distribution, these moments being the necessary and sufficient conditions for the identification of a normal distribution; it follows that if the distribution of X is normal, then Y will have exactly the same distribution as the original, undisclosed variable. Moreover, by using a non-confidential variable in the perturbation process, this method allows preserving the covariance between the confidential variable X and the non-confidential variable S.

As appealing as this framework is, it relies on the pivotal normality assumption. Normality underlies many statistical analyses commonly used (such as regressions and hypothesis tests), and ensures that analysis based on the masked data will lead to the same results that one would have obtained with the original data. But the methodology is rather limiting if, rather than being interested in using the data for econometrics and inference, users are interested in the intrinsic features of the distribution, e.g. to compute descriptive statistics such as fractiles or measures of dispersions. In this case, perturbation using additive Gaussian noise loses its usefulness as additional features of the original distribution have to be preserved in a privacy-safe way, in particular skewness, which conveys substantial and relevant information.

3.2.2 A sufficient multiplicative masking method for lognormal distributions

Using the same notations as before, we let X follow a lognormal distribution with parameters $\mu_X > 0$ and σ_{XX}^2:

$$X \longmapsto LN(\mu_X, \sigma_{XX}^2)$$

where, by definition of a lognormal distribution, $\mu_X = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$ and $\sigma_{XX}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \mu_X)^2$. The first and second order moments of X are thus respectively:

$E(X) = exp\left(\mu_X + \frac{\sigma_{XX}^2}{2}\right)$ and $V(X) = [exp(\sigma_{XX}^2) - 1]exp(2\mu_X + \sigma_{XX}^2)$

The same assumptions apply for the perturbation u, assumed to be independent of X and with parameters $\mu_u = \frac{1}{n}\sum_{i=1}^{n} \ln u_i > 0$ and $\sigma_{uu}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln u_i - \mu_u)^2$:

$$u \longmapsto LN(\mu_u, \sigma_{uu}^2)$$

with $E(u) = exp\left(\mu_u + \frac{\sigma_{uu}^2}{2}\right)$ and $V(u) = [exp(\sigma_{uu}^2) - 1]exp(2\mu_u + \sigma_{uu}^2)$.

The perturbed value of X, Y is generated through the following equation, a homothetic function:

$$Y = X^{\alpha}u^{1-\alpha} \text{ with } 0 \leq \alpha \leq 1$$

As for the MS hybrid generator, α is also a similarity parameter: when α is set to 1, X=Y and no perturbation is generated; when α=0, Y is not a function of the confidential value X but only of the lognormal noise. The intermediary cases 0<α<1 create convex combinations of confidential values and noises.

The properties of lognormal distribution ensure that the α power distribution of X also follows a lognormal law [29]:

$$X^{\alpha} \longmapsto LN(\alpha\mu_X, \alpha^2\sigma_{XX}^2)$$

and the same applies for the 1- α power of u:

$$u^{1-\alpha} \longmapsto LN((1-\alpha)\mu_u, (1-\alpha)^2\sigma_{uu}^2)$$

Given independency of u and X, Y has thus the following distribution:

$$Y \longmapsto LN(\alpha\mu_x + (1-\alpha)\mu_u, \alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{uu}^2)$$

with the associated two first moments being: $E(Y) = exp\left(\alpha\mu_x + (1-\alpha)\mu_u + \frac{\alpha^2\sigma_{XX}^2+(1-\alpha)^2\sigma_{uu}^2}{2}\right)$ and $V(Y) = [exp(\alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{uu}^2) - 1]exp[2(\alpha\mu_X) + (1-\alpha)\mu_u + \alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{uu}^2]$.

We can now derive the necessary and sufficient conditions that will ensure that Y has the same distribution as X. Unlike the additive framework, we cannot proceed by preserving the first two moments of Y. More generally any set of k-order moments with k≥1 is not isomorphic to any lognormal law: we can invariably find other laws (lognormal or not) that have the same moments [29]. To achieve sufficiency we have to consider the logarithmic transformation of Y:

$$lnY \longmapsto N(\alpha\mu_X + (1-\alpha)\mu_u, \alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{uu}^2)$$

Being now in a Gaussian case, we can derive conditions for the first two moments:

$$\alpha\mu_X + (1-\alpha)\mu_u = \mu_X \Leftrightarrow \mu_X = \mu_u$$

$$\alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{uu}^2 = \sigma_{XX}^2 \Leftrightarrow \sigma_{uu}^2 = \frac{1-\alpha^2}{(1-\alpha)^2}\sigma_{XX}^2$$

As σ_{uu}^2 ≥0, we also have $1-\alpha^2 \geq 0$ and thus 0≤α≤1, confirming α as a well-defined similarity parameter. Using the sufficiency conditions at the logarithmic level and exponentiating ln Y, we find that u must have the following lognormal distribution:

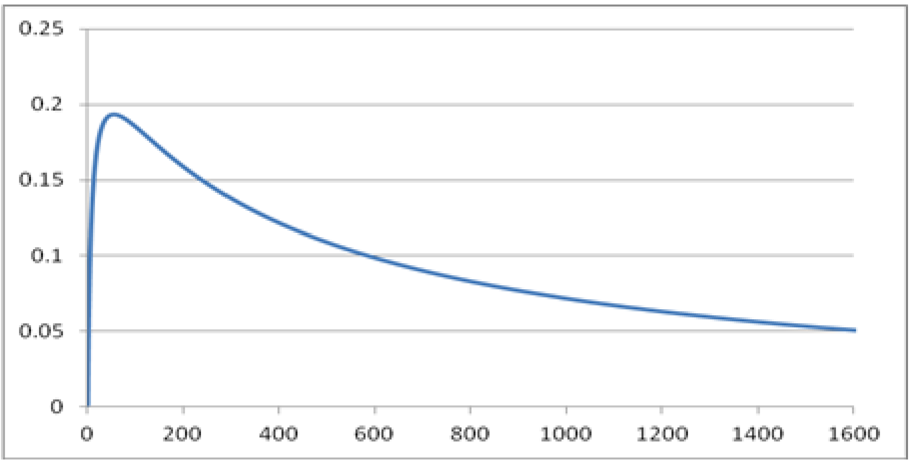
$$u \longmapsto LN(\mu_X, \frac{1-\alpha^2}{(1-\alpha)^2}\sigma_{XX}^2)$$

As exponentiation establishes a one to one correspondence (i.e. it is a bijective mapping), the sufficiency conditions at the logarithmic scale ensure sufficiency at the original variable scale. Thus, this perturbation method preserves the features of the original distribution including its skewness, but allows the similarity of data points to be selected. As shown in the following section, this method is also confidentiality efficient, in particular for observations far from the mean.

3.3 Numerical validation

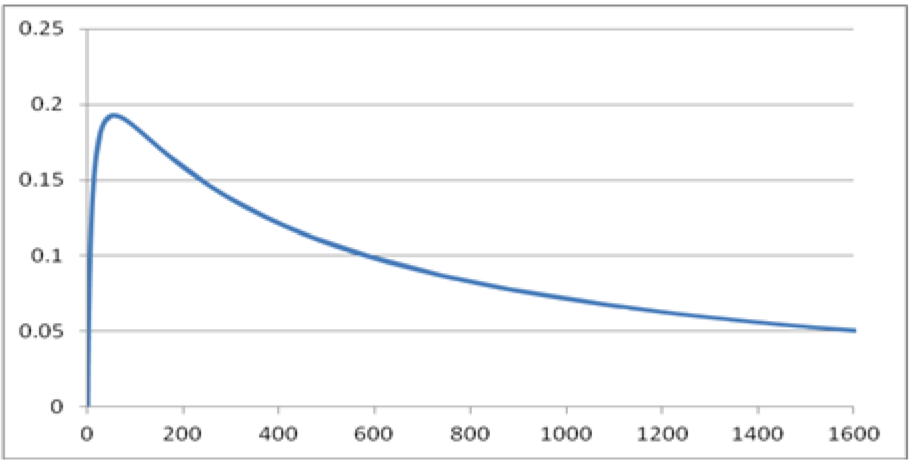
We simulated a vector consisting of one thousand data points drawn from a lognormal distribution with parameters 4 and 2, i.e. a deliberately highly skewed distribution. Figure 3.1 shows the density of the original distribution.

Figure 3.1: Density of original data.



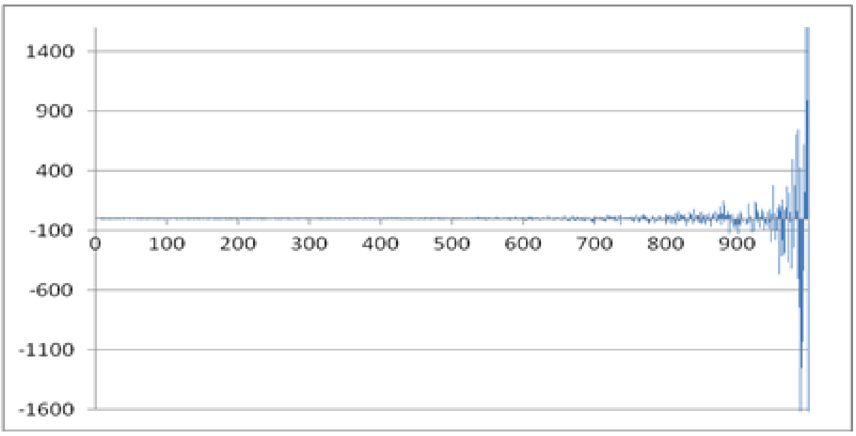
When α=0.9, the distribution of the perturbed data exactly matches that of the original data: as shown in Figure 3.2, the density of the former is strictly identical to the latter.

Figure 3.2: Density of perturbed data with alpha =0.9.



As was established in the previous section, perturbed distributions will remain the same as the original one for 0≤α≤1. Thus, the multiplicative masking method preserves the initial data structure. Data points are nevertheless altered in an interesting way, in particular for confidentiality purposes. Figure 3.3 depicts the changes that occur in the absolute values for each point (ranked in ascending order on the x-axis according to their original values).

Figure 3.3: Absolute differences between original and perturbed data for alpha =0.999.



One immediately sees that, for a small value of the dissimilarity parameter, most of the data points that are close to the mean are very close to the original values while, due to the multiplicative structure used, values that are far away from the mean are substantially altered. And as high values are those where disclosure risk is higher, this pattern of perturbation is that which is most appropriate. For lower values of α, and thus greater dissimilarity, perturbations start to spread along the distribution from the upper to the lower tails, as can be seen in Figures 3.4, 3.5 and 3.6.

Figure 3.4: Absolute differences between original and perturbed data for alpha =0.95.

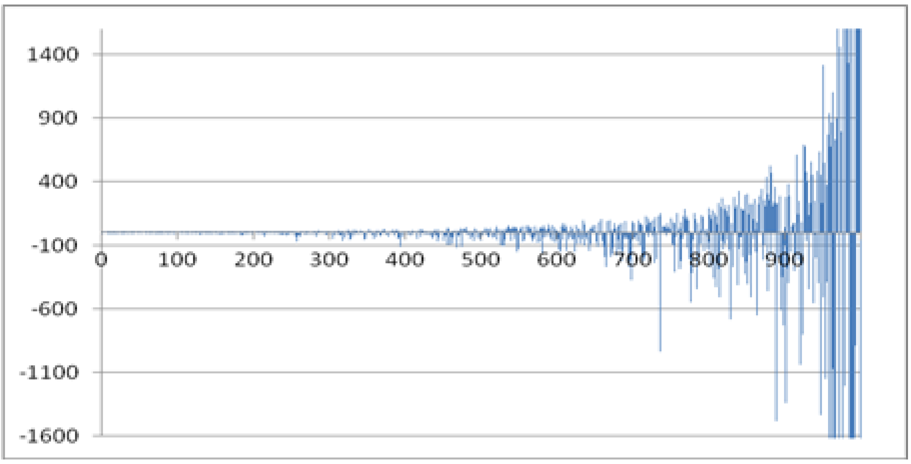


Figure 3.5: Absolute differences between original and perturbed data for alpha =0.9.

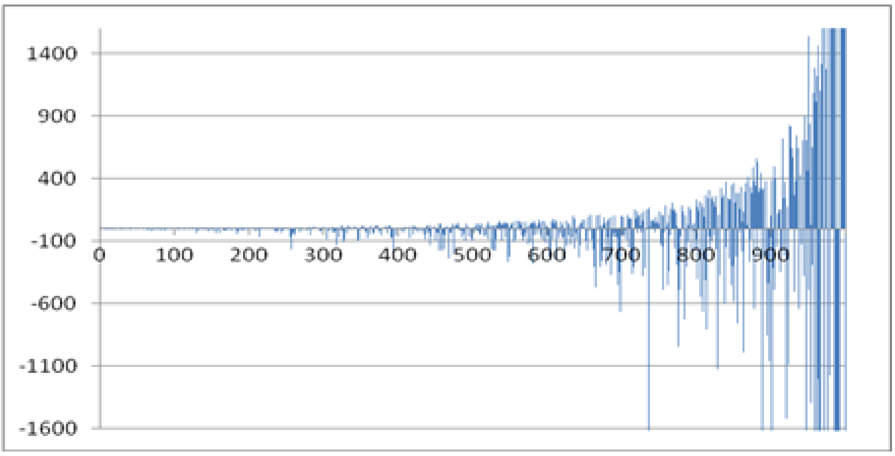
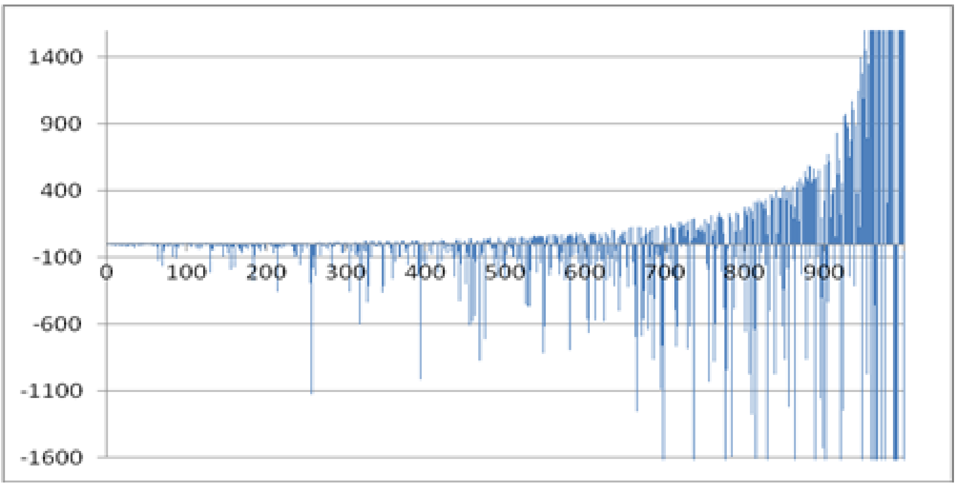


Figure 3.6: Absolute differences between original and perturbed data for alpha =0.7.



As perturbations can both reduce and increase values of different data points, the ranking of data points is likely to change during the process, thus increasing data protection against disclosure risk (in particular, observations away from the mean could now near it, and conversely). As shown in Figures 3.7 and 3.8, the more dissimilarity is introduced, the more swaps occur in the data ranking, i.e. the more observations are permuted.

Figure 3.7: Initial vs. perturbed ranks for alpha =0.95.

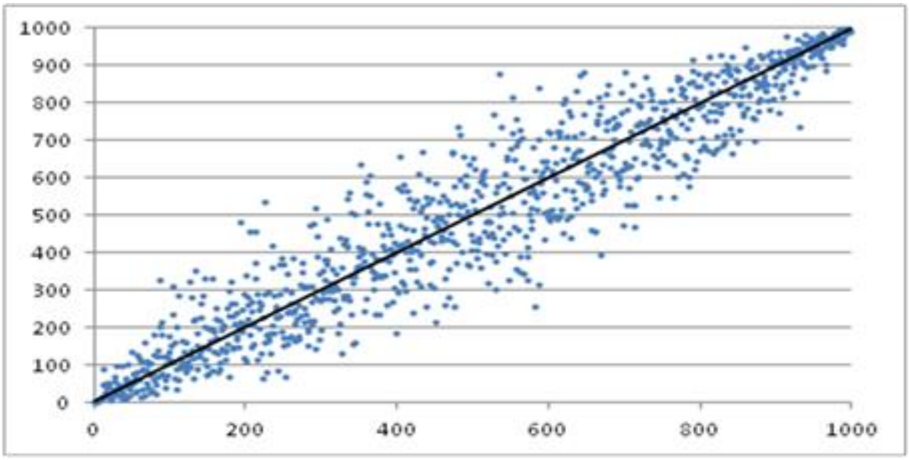
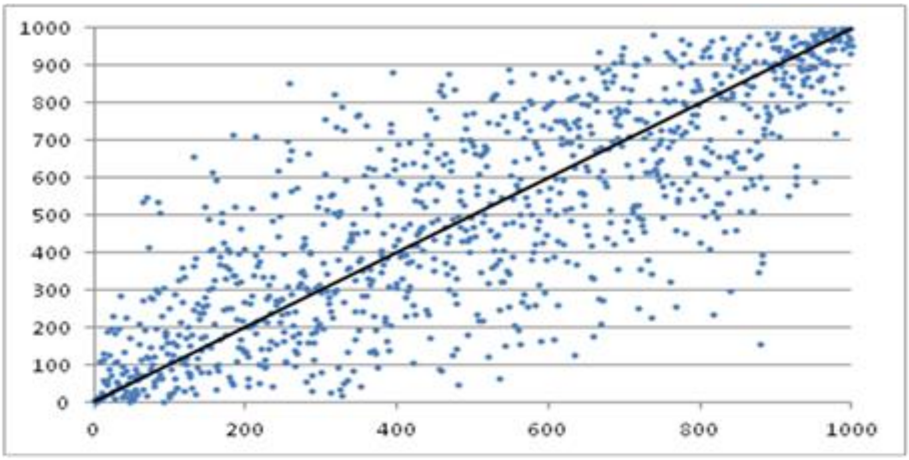


Figure 3.8: Initial vs. perturbed ranks for alpha =0.7.



Permutations reinforce the fact that greater dissimilarity lowers disclosure risk for the disseminated microdata perturbed by this method. Data points that are further away from the sample mean can be more easily identified due to two distinct problems: the classic issue of protection of the value recorded, plus a distance effect i.e. while perturbed, an observation away from the mean could again face high disclosure risk by still remaining far from it. Permutations circumvent this additional problem. This mechanism happens to be an alternative way to describe the method proposed.

Rank swaps, however, can also be a drawback, as the swapping of ranks will perturb the covariances with other variables. In fact, the lower α is, the lower the correlation between the original and the perturbed variable will be (Table 3.1); this will also imply higher perturbation of covariance with other variables.

Table 3.1: Correlation coefficients between the original and perturbed variable for different similarity degrees.

α	0.999	0.95	0.9	0.8	0.7
Correlation coefficient	0.99	0.60	0.41	0.24	0.18

The MS hybrid generator outlined in the previous section automatically preserves some covariances, at least for the non-confidential variable used in the perturbation equation. However, it remains mute for covariances with other confidential variables external to the equation. Through its similarity parameter, the multiplicative method presented here allows preserving the covariance with any other variables, but with a trade-off as to the degree of protection that one wants to achieve in the disseminated data. This trade-off represents an inherent limitation to the multiplicative masking structure. For example, one cannot adapt the perturbation process by introducing a non-confidential variable in order to exactly preserve some set of covariances: a necessary condition to do that would be that the non-confidential variable also follows a lognormal distribution. But a heavy-tailed non-confidential variable is a very unlikely configuration. In other cases, the use of the perturbation method with any non-lognormal distribution would induce a distribution of the perturbed variable having a different functional neither exact nor closed form, or being too cumbersome an approximation to be tractable in a simple disclosure control environment [30].

3.4 Conclusion

When using SDC techniques to generate perturbed data, the/an analysis performed on the altered datasets should yield results that are identical or at least very close to those that would have been obtained using the original data. The assumption of normality in the distribution of the original variable and in the error term is a convenient way to achieve this objective. Unfortunately, many economic variables are distributed according to a heavy tailed, asymmetric form that makes the Gaussian framework limited. Moreover, and as underlined in many recent studies [41], fat tails are important for economic analysis as their impact could be substantial. It should nevertheless be noted that data points generating a heavy tailed distribution are often scarce in microdata sets, especially those that come from survey-based data (except if specific oversampling procedures are used).

Two reasons account for this under-representation of high values. The first is simply due to the sampling scheme, as observations away from the mean are less likely to be observed in surveys. The second is that, as observations away from the mean face a higher disclosure risk than data points closer to it, control of these risks forces data producers to rely on top coding, i.e. values above a certain amount are automatically censored to that amount. As a result, a survey's skewedness is only a partial measure of the true population skewedness. In this case, one can still reasonably assume that normality is a sufficient assumption for surveys' data perturbation, but further research will have to be conducted to determine the relative performances of these additive masking methods when the original data differ from a normal distribution.

The case of register-based microdata is quite different from that of surveys, as all of the population is generally included. In this case, skewness is likely to occur very often, and our methodology will perform better than methods such as the MS hybrid generator. Moreover, as only heuristic rules are possible in practice for preserving covariances (one being, for example, choosing a degree of similarity between 0.99 and 0.95 that will protect observations away from the mean while sufficiently preserving the covariance), register-based data are favoured; due to their nature and the fact that they are not originally collected for analytic purposes, fewer variables are available than in a survey for covariance computations.

In conclusion, this Chapter has presented a simple technique that allows data producers to generate perturbed datasets according to a selectable degree of similarity when the underlying distribution is positively skewed, using the properties of lognormal distribution. Despite the fact that this method is meant to be applicable in a particular data-context, it must be emphasized that ultimately its outcome can be mainly characterized permutations. Obviously, this echoes the findings of the permutation-based paradigm outlined in Chapter 2, i.e. whatever the analytical apparatus of method and the features of the data to be anonymized, anonymization can always be appraised through permutation. The next two Chapters aim at developing this insight, first for the ex-post evaluation of anonymization, and second, for performing anonymization directly through the ex-ante selection of permutation patterns.

4 UNIVERSAL MEASURES OF DISCLOSURE RISK AND INFORMATION LOSS

4.1 Introduction

As we saw in Chapter 2, the permutation paradigm is not considered by its authors as a new anonymization method per se (a statement that can be reconsidered, see Chapter 5), but aims at offering the potential to evaluate all available techniques through the same lens. The development of a set of appropriate measures of disclosure risk and information loss based on permutation distances, however, remains to be seen. This is the objective of this Chapter, which explores the first consequences of the permutation paradigm. Notably, it proposes some universal measures of disclosure risk and information loss that can be computed in a simple fashion and can be used for the evaluation of any anonymization method, independent of the context under which they operate. The construction of these measures also allows introducing the notions of dominance in disclosure risk and information loss in data anonymization, which formalise the fact that different parties involved in micro data release can each have different sensitivities to privacy and information, and can inform about the methods that can be used to reach a consensus among all parties. These two notions of dominance can characterize which methods, under any tastes for privacy and information, will always perform better than others. The contributions in this Chapter have been published in [45].

4.2 A class of universal measures of disclosure risk based on permutation distances

We start by observing that from the permutation-based paradigm, it is always possible to retrieve post-anonymization for any method applied on any data, the overall amount and distances of permutations performed. Thus, for a given attribute j, permutation distances can be retrieved and collected under the form of a vector of rank displacement r_j, i.e. a vector measuring for each record the amount of rank shifting that occurred. Note that to avoid some unnecessary technical difficulties, in what follows zero values in r_j (i.e. no permutation took place) will be assigned, without loss of generality, an infinitesimally small value ε>0. An illustrative example of rank displacement vectors for three attributes is:

$$r_1 = \begin{pmatrix} \varepsilon \\ \varepsilon \\ \varepsilon \\ \varepsilon \\ \varepsilon \end{pmatrix} \quad r_2 = \begin{pmatrix} 3 \\ \varepsilon \\ \varepsilon \\ 1 \\ -4 \end{pmatrix} \quad r_3 = \begin{pmatrix} \varepsilon \\ 2 \\ 2 \\ -2 \\ -2 \end{pmatrix}$$

Now, r_j has to be evaluated in some way for assessing disclosure risk based on permutation distances. A natural choice is to gauge r_j by assigning a magnitude, taking its Euclidean norm and adopting the rule that the higher the norm, the lower the disclosure risk (as the larger will be the permutation distances contained in r_j). But other cases are possible. In general, any L(p)-norm is acceptable: for example, for r_1, r_2 and r_3, the ∞-norm (or Chebyshev distance) would give ε, 4 and 2, respectively. This variety of choice to evaluate vectors generally depends on the problem at hand, as one will select a L(p)-norm adapted to the meaning of the object that is meant to be quantified. In the case of a vector of permutation distances, it is not clear why a Euclidean length would be more suitable and meaningful than a Chebyshev length, or why all the norms in-between can or cannot be considered. Thus, there can be a fundamental arbitrariness in this choice. However, we argue that in the permutation paradigm, such choice can be given an intuitive interpretation in terms of disclosure risk.

To further illustrate this arbitrariness, consider the following example: if in r_3 the third record is now permuted one rank more and the second one rank less, r_3 will be viewed as identical to r_2 according to the ∞-norm. It is, however, not totally clear if the situation has really improved in terms of disclosure risk for the third attribute. On the contrary, it can be reasonably thought that the new situation is more problematic, as having a record permuted only one time increases the disclosure risk in a way that may not be offset by the additional permutation of an already sufficiently permuted record. In fact, being able to evaluate if the situation has improved necessitates a notion of aversion to disclosure risk, which, to the best of the author's knowledge, is not present or formalized in the literature on SDC. The permutation paradigm allows introducing this notion in a simple way:

> ***Definition 4.1****: In the permutation paradigm, aversion to disclosure risk is the preference toward less permuted records for the evaluation of this risk.*

Aversion to disclosure risk accounts for the fact that different data releasers or subjects can have potentially different appreciations of disclosure risk (alternatively, this can also be viewed as different levels of privacy awareness). Some releasers may consider that achieving a certain average level of permutation is sufficient, while from a contributing subject’s point of view, or from the point of view of other data releasers (say, for example, when multiple releasers are involved in the release of a data set), this could be judged as insufficient. Because the permutation paradigm reduces the relevant information needed for the evaluation of any method to permutation, aversion to disclosure risk can be modelled by assuming that different permutation distances have different weights. On the one hand, a strongly averse data releaser/subject may put relatively more weight on the lowest permutation distances achieved; on the other hand, a weakly averse releaser/subject may consider different permutation distances the same way and focus only on the average amount of permutations.

Indeed, existing measures of disclosure risk generally entail some implicit assumptions regarding how the risk is assessed. This can be illustrated by considering the formula for rank order correlation coefficient, previously used in the permutation paradigm for the assessment of disclosure risk [39,12], which for a non-masked attribute X_j and its reverse mapped version Z_j can be written as (where d_i is the difference between the ranks of each record):

$$\rho_{X_j,Z_j} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

It is apparent that the rank order correlation coefficient implies specific preferences on the permutation distances, as the square of the ranks' differences magnifies the impact of large permutations compared to small ones. One could even argue that the rank order correlation coefficient is not an appropriate measure, as for the assessment of disclosure risk it is small, not large, permutation distances that matter. For example, according to ρ_{X_j,Z_j} an anonymization method permuting only one record 10 times will be judged as having reduced disclosure risk more than another method permuting 3 records 5 times. Again, it is difficult to rank the two situations in terms of disclosure risk. To overcome this issue, the following proposition establishes a measure of disclosure risk sensitive to different aversions, with an adjustable degree of focus on small permutation distances:

Proposition 4.1*: For any attribute j=1,…,p of a data set $Y_{(n,p)}$, a quantitative measure of disclosure risk in the permutation paradigm is given by:*

$$D_j(\alpha) = \left[\frac{1}{n}\sum_{i=1}^{n} abs(r_{j(i)})^{\alpha}\right]^{1/\alpha} \quad for\ \alpha \le 1\ and\ \alpha \neq 0$$

$$and\ D_j(\alpha) = \prod_{i=1}^{n} abs(r_{j(i)})^{1/n}\ \ for\ \alpha = 0$$

where $r_{j(i)}$ denotes the elements of r_j and α the parameter of aversion to disclosure risk.

$D_j(\alpha)$ makes use of a power mean (see [25] for a discussion of its various properties) for the aggregation of the components of r_j, with the parameter α substantiating the notion of aversion to disclosure risk. The arithmetic mean becomes a special case (α=1) of $D_j(\alpha)$, which forms a natural starting point by computing the average level of permutation distances. In that case, all distances are given the same weight and there is a one-to-one substitution between them, e.g. two records permuted two ranks are equivalent to one record permuted four ranks. From this benchmark, the more α decreases, the more weight is given to the smallest permutation distances. The more α approaches -∞, the more $D_j(\alpha)$ converges towards the smallest permutation distance in r_j. As a result, for a given r_j and $\alpha' < \alpha$, we have $D_j(\alpha') \leq D_j(\alpha)$: the lower is α, the stronger is the aversion to disclosure risk. Note that as a general case of averages, $D_j(\alpha)$ is independent of the number of records, which eases the comparison across different data sets of different sizes. Moreover, for an attribute observed over n records, the maximum permutation distance for a record is abs(n-1). Thus, re-scaling $D_j(\alpha)$ by 1/n-1 will produce a measure of risk that ranges between 0 and 1, which is an appealing property for performing comparisons and quantifying the utility/privacy trade-off [26].

One might be tempted to think that the notion of aversion to disclosure risk adds an unnecessary layer of complexity to the evaluation of this risk. We maintain that it provides a better grasp of the reality of individual data exchange (see Chapter 2). In the current state of the literature, it is not a notion that can be made analytically tractable in a straightforward way for all methods (or as we saw, is embodied implicitly rather than explicitly). But in the permutation paradigm, permutation distances are the only meaningful quantities under scrutiny, which makes natural the fact that these distances can be judged by different individuals differently. Given the number of parties involved in data dissemination, e.g. numerous data releasers and respondents, it is very unlikely that all of them will have the same judgment. The $D_j(\alpha)$ measures are a way to incorporate this diversity. In practice, by computing the measure for several α, a data releaser can, for example, communicate about the prevention against disclosure risk through different points of view. This circumvents the issue involved in the empirical assessment of disclosure risk [35], where a score based on different measures of disclosure risk is computed using an ad-hoc weighting scheme. Under such an approach, weights can drive the overall assessment that is made. But using the current proposal, a single measure can be computed on a continuum of weights which all carry an interpretation in terms of disclosure risk.

The measure $D_j(\alpha)$ can also be used to characterize in an unambiguous way which data anonymization methods perform better than others through the concept of disclosure risk dominance that we introduce below. The concept of dominance comes originally from the notion of stochastic dominance [32], which is widely used in economics and finance. It can, however, be applied to any distribution, which is done here for the distribution of permutation distances. To the best of the author’s knowledge, this is the first time it is considered in the context of data anonymization:

> ***Definition 4.2****: For an attribute j, an anonymization method A is said to dominate (i.e. unanimously performs better than) another method B for the protection against disclosure risk if it holds that* $D_j(\alpha)' \leq D_j(\alpha) \; \forall \; \alpha \leq 1$ *(where* $D_j(\alpha)$ *(resp.* $D_j(\alpha)'$*) are the measures of Proposition 4.1 computed from A (resp. B)).*

Disclosure risk dominance characterizes anonymization methods that will consistently ensure greater levels of permutation distances (and thus levels of protection against disclosure risk) from the mean to the bottom of their distribution. In practice, that means that whatever the aversion the agents involved in the data dissemination may have, a dominant method will ensure unanimity regarding its performance against disclosure risk.

Obviously, dominance may not always be reached in practice. For example, a method A can happen to dominate B over -4$\leq \alpha \leq 1$ but being dominated by B over -$\infty \leq \alpha < -4$. In that case, that means that the use of A is advisable for small up to medium disclosure risk aversion, while for strong aversion B is more advisable. As a result, one can learn about the relative performance of methods by investigating where dominance holds but also where it ceases to hold.

One final remark on $D_j(\alpha)$ is in order. The domain of variation of the disclosure risk aversion parameter has been set to range from one and below, which does not define a L(p)-norm strictly speaking. In fact, it would be $D_j(\alpha)$ with $\alpha > 1$ that would rigorously define a L(p)-norm, up to a factor $\sqrt[\alpha]{n}$ [4], leading to a standard notion of distance for the vector r_j. However, we argue that in the context of data anonymization, the interpretation of the parameter α is not suited to that case. With $\alpha > 1$, the more α increases, the more weight is given to the largest permutation distances (and the more α approaches $+\infty$, the more $D_j(\alpha)$ converges towards the largest permutation distance in r_j, i.e. a Chebyshev distance is computed). That would mean that large permutations make up for the bulk of protection against disclosure risk, but it is small permutations that can lead to greater disclosure risk. As a result, $D_j(\alpha)$ makes use of the aggregation structure of a p-norm but does not define one strictly. This has no incidence on the validity and interpretation of the measure.

In this section, the measures $D_j(\alpha)$ and the concept of dominance have been introduced with the aim of offering a more granular view of disclosure risk, with an easy-to-grasp notion of disclosure risk aversion. Given that in the permutation paradigm all the necessary information is reduced to permutation distances, they provide a common and understandable language for performing meaningful comparisons of anonymization methods, independently of their analytical environment or the distributional features of the data. The class of $D_j(\alpha)$ measures formalizes the tool for such comparisons and is very general in its scope, in that it allows incorporating different judgments about disclosure risk and characterising methods that can be viewed as unanimously superior to others.

4.3 A class of universal measures of information loss based on relative permutation distances

A key feature of the permutation paradigm is that it preserves exactly the marginal distributions of the data (as Z is simply a permutation of X; see Chapter 2). Thus, information loss can only come from the alteration of the dependency among attributes. Thus to achieve an exact preservation of multivariate distributions (here bivariate distributions), the same permutation patterns must be applied to some block of attributes. In fact, any multivariate anonymization method can be viewed as a block permutation of attributes. It is a simpler view by comparison to the current multivariate anonymization methods available in the literature, which can be analytically complex [26]. Of course, the exact preservation of a multivariate distribution may impinge on the level of privacy achieved by the anonymized data. Additionally, it has been previously empirically established that obtaining a safe anonymized data set that is resistant to an attack *via* record linkage necessitates an amount of masking (or equivalently, of permutations) proportional to the dependency between the attributes of the original data set [13]. Expressed in the permutation paradigm, this means that the permutation patterns must be more dissimilar.

In practice then, the question turns out to be more about the extent of preservation of multivariate distributions and an inescapable trade-off: the less preservation there is, the more the anonymized data set will be judged as safe. For a dataset with a strong dependence between its attributes, the trade-off may be particularly severe. But for a dataset with weak attributes dependence it is also a non-trivial issue, as (while less likely to occur in practice) an anonymization method can create an artificial dependence between the attributes, which in a way is also a loss of information. For example, it is possible that two completely independent attributes in the original data happen to be, through a peculiar permutation, both ranked in increasing order of magnitudes in the anonymized version, fooling the data user as to the real strength of the relationship.

To assess information loss, a first avenue is to compare the rank order correlations between attributes j and j' in the anonymized data and the original data set [39]. The most likely case is that the former will be lower than the latter, indicating an alteration of the attributes' relationship and thus a loss of information by a weakening of the dependence (but in less likely cases the reverse can also happen). For such comparison, the original level of rank order correlation provides the starting point from which information loss is assessed. As a result, it will differ according to each couple of attributes considered, which is rather inconvenient. Also, and for the same reason outlined above, an implicit and specific weighting structure is given to large ranks differences when using rank order correlation. Again, different data users can have different views about distances when assessing information loss. As for disclosure risk, this can be formalized through the concept of aversion to information loss (or stated otherwise, of information awareness):

> ***Definition 4.3****: For two attributes j and j' in the permutation paradigm, aversion to information loss is the preference toward large relative permutation distances for the evaluation of this loss.*

Thus a more general approach is to consider the degree of similarity between the permutations that took place for the two attributes and to allow different weights for different *relative* distances. To do so, it can be observed that a vector $\Delta(r_k)$ of differences between the vectors r_j and $r_{j'}$ is a vector of dissimilarity between the anonymization procedures that have been applied to the couple of attributes k=(j, j') (with $j\neq j'$). When each of the components of $\Delta(r_k)$ are equal to zero (here again zero values in $\Delta(r_k)$ will be assigned, without loss of generality, an infinitesimally small value $\varepsilon>0$), j and j' have been permuted the same way; the permutation patterns applied to them are identical, despite the fact that the anonymization methods used can be different in practice. There is no loss of information as the joint distribution of j and j' is preserved. But when Δ(r_k) has some non-zero elements, information has been modified. This leads to the following proposition:

Proposition 4.2*: For two attributes j and j' of a data set $Y_{(n,p)}$, a quantitative measure of information loss in the permutation paradigm is given by:*

$$I_k(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n} abs(\Delta r_{k(i)})^{\theta}\right]^{1/\theta} \quad for\ \theta \geq 1$$

where $\Delta r_{k(i)}$ denotes the elements of Δ(r_k) and θ the parameter of aversion to information loss.

The measure $I_k(\theta)$ bears strong analytical similarities to D_j(α), but while the latter is concerned with average or small permutation distances across records for a given attribute, the former considers average or large *relative* permutation distances between two attributes across records. Note that this measure delivers a diagnosis independently of the direction of the alteration of dependence between attributes, i.e. if dependence has been weakened or strengthened as a result of anonymization. $I_k(\theta) = 0$ means no information loss, while for a given θ, the larger $I_k(\theta)$ is, the more the relationship between attributes has been altered (and thus the more information has been lost in the process). It thus provides a general measure of information loss than can be applied to any anonymization methods. Note that $I_k(\theta)$ is a power mean but also denotes strictly a L(p)-norm of the vector Δ(r_k) up to the factor $\sqrt[\theta]{n}$. This factor allows performing a comparison independently of the size of the data set. Moreover, for two attributes with n records each, the maximum relative permutation distance for a record is n-1. Thus, re-scaling $I_k(\theta)$ by 1/n-1 will produce a measure of information loss that ranges between 0 and 1, which is convenient for comparison with D_j(α) as it can also range on the same scale (see above).

$I_k(\theta)$ aims at measuring the extent of dissimilarity that anonymization introduced for j and j', with θ capturing different emphasis on relative permutation distances; the greater θ, the stronger the focus on large distances. In a similar fashion to disclosure risk, aversion to information loss accounts for the fact that different agents involved in data dissemination can each have different perceptions of information loss. Typically, this aversion is likely to be stronger for data users than for data releasers. The parameter θ formalizes such diversity in tastes. As for $D_j(\alpha)$, it can also be used to unambiguously rank couples of anonymization methods (or the same anonymization method with two different parametrizations) that perform better than others, by introducing the concept of dominance in information:

> ***Definition 4.4****: For two attributes j and j', two anonymization methods A and B are said to dominate (i.e. perform better than) two other methods C and D for the preservation of information if it holds that* $I_k(\theta)' \leq I_k(\theta)\ \forall\, \theta \geq 1$, *where* $I_k(\theta)'$ *(resp.* $I_k(\theta)$ *) are the measures of Proposition 4.2 computed on A and B (resp. C and D)).*

Information dominance characterizes anonymization methods that, when applied to two attributes, will consistently ensure lower levels of relative permutation distances (and thus a greater preservation of information) from the mean to the top of their distribution. In practice, this means that whatever the aversion to information loss agents involved in data dissemination may have, a dominant couple of methods compared to others will ensure unanimity regarding its performance in terms of information preservation.

Beyond establishing which couple of methods does best in preserving information, $I_k(\theta)$ and information dominance can also be used to tune the extent of information to be preserved. Under a different scenario of aversion to information loss, two anonymization methods can be evaluated ex-post in terms of information preservation through $I_k(\theta)$ and then be re-run to obtain the desired information loss. The permutation paradigm simplifies the implementation of multivariate scenario and the quantification of information loss in comparison to the current techniques available.

4.4 Experimental investigation

The goal of this section is to illustrate the use and effectiveness of the universal measures of disclosure risk and information loss developed above. The anonymization methods considered are some of the most popular, namely: independent additive noise, multiplicative noise and rank swapping. This selection is also representative of some of the diversity of principles used in microdata masking [26]. The experimental data set used is two attributes of the Census data set, observed over 1080 records. This data set has been taken to evaluate the properties of anonymization techniques in terms of disclosure risk and information loss numerous times in the literature [7].

The experiment proceeded as follows:

I. First, we generated the masked version of the data set using: additive noise with standard deviations equal to 50% of the standard deviations of the two attributes; multiplicative noises drawn from a uniform distribution within the range (0.75,1.25); rank swapping [26] with a swapping distance of 30%. For noise-based methods, the noise terms are generated independently for each attribute.

II. We then reverse-mapped the masked data to compute the level of absolute and relative permutations.

III. From these levels, we computed the universal measures of disclosure risk and information loss $D_j(\alpha)$ and $I_k(\theta)$ for a quasi-continuum of aversion parameters, that is, by increments of 0.01. The results are displayed directly in the form of curves, with the aversion parameters on the x-axis and the value of $D_j(\alpha)$ (resp. $I_k(\theta)$) for the evaluation of disclosure risk (resp. information loss) on the y-axis. Notably, this allows drawing conclusions using the dominance concepts developed above.

IV. Given that all the methods considered involved randomness, this experiment was replicated 100 times; the results thus report the average values of $D_j(\alpha)$ and $I_k(\theta)$ over the 100 replications.

Figures 4.1 and 4.2 display the universal measures of disclosure risk for aversion parameters ranging from 1 to -3, for the two attributes respectively. While additive noise and rank swapping offer a similar average level of absolute permutation distances (i.e. for $D_j(1)$) for both attributes, when the focus is progressively strengthened on the low permutation distances, the performances of the two methods happen to diverge rapidly, with noise addition offering no protection while data swapping consistently ensures permutation across all records. In fact, data swapping appears to strictly dominate noise addition as the level of absolute permutation achieved by the former is always greater than the latter for any level of risk aversion. As for multiplicative noise, it is dominated by noise addition and data swapping for both attributes.

Figure 4.1: Disclosure risk assessment for the first attribute.

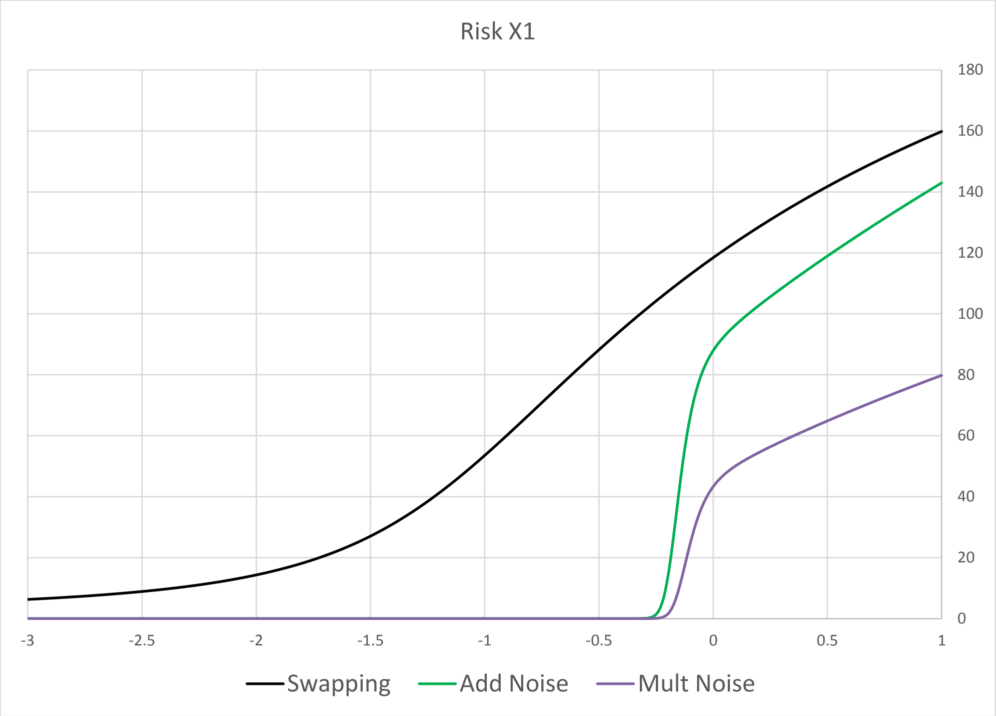


Figure 4.2: Disclosure risk assessment for the second attribute.

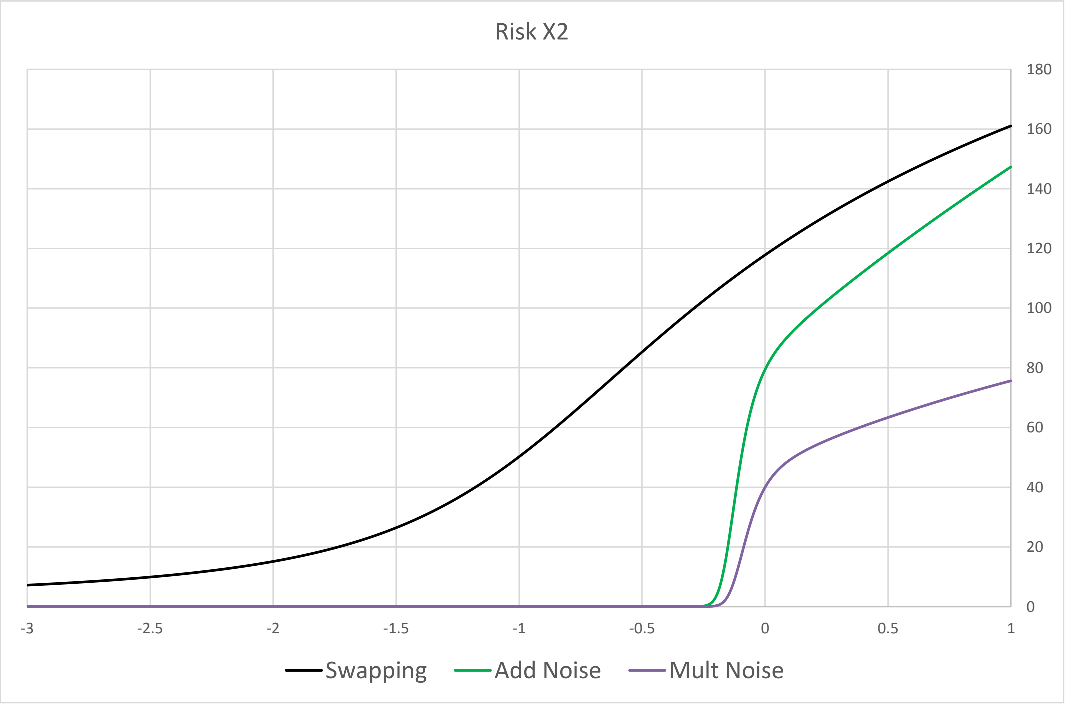
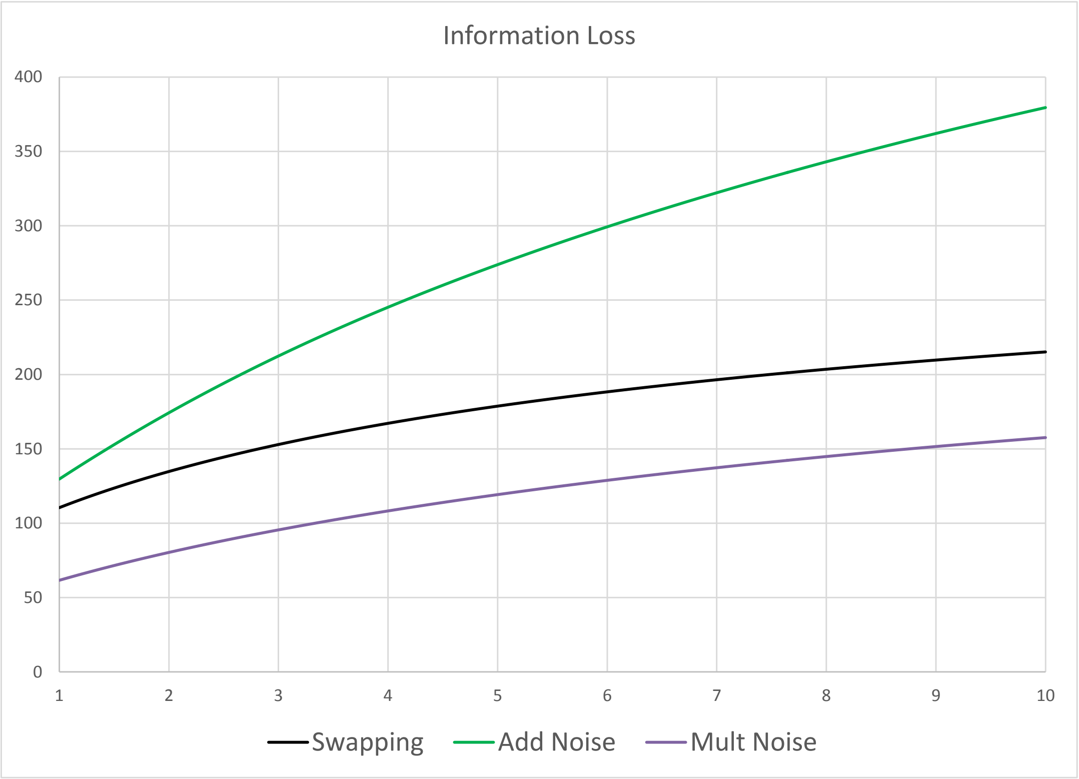


Figure 4.3 displays the universal measure of information loss for aversion parameters ranging from 1 to 10. As the three curves for the three methods do not intersect, some dominance rules hold again. In fact, multiplicative noise dominates data swapping by providing the lowest levels of relative permutation distances across the range of aversion parameters, while data swapping dominates additive noise but appears to be dominated by multiplicative noise. From these results, we can conclude that by providing better protection against disclosure risk and better preservation of information, rank swapping appears to outperform additive noise as an anonymization method in general, that is, whatever the degrees of aversion to disclosure risk and information loss substantiated by the parameters α and θ. On the contrary, the comparisons with multiplicative noise involve some trade-offs, as while being dominated by the two other methods for the protection against disclosure risks, it consistently provides lower levels of information loss.

Figure 4.3: Information loss between the two attributes.



As a result, the two classes of measures developed in this Chapter allow both the evaluation and comparison of any method. Given the fact that permutation appears to be the core principle of data anonymization, comparisons based on $D_j(\alpha)$ and $I_k(\theta)$ can be performed independently of the types of methods considered and the data upon which they are applied. In that sense, they embody a universal scope of application, while currently existing measures happen to be tied to their underlying parametrizations and the distributional feature of the data to be anonymized. To the best of the author's knowledge, this is the first time that such measures have been proposed in the literature. In the experiment considered, additive noise can be ruled out as an effective procedure. A large scale ranking of the variety of the methods currently available is an avenue for future research.

4.5 Measures of disclosure risk and information loss at the data set level

The class of disclosure risk measures introduced in Section 4.2 operates by attributes taken in isolation. While this is a standard approach, one may also be interested in having a quantification of the overall disclosure risk for a data set of p attributes. This kind of measure is in a way complementary to an assessment of disclosure risk attribute by attribute: while the latter is necessary to have a detailed view of the level of protection applied, which is likely to vary according to each attribute's specificity and sensitivity, having a global view of the anonymized data set can be useful, not least for communication purposes. Considering as a starting point the measure $D_j(\alpha)$, which as outlined above bears close similarity with a L(p)-norm (i.e. a vector norm), for a data set with p attributes a possible overall measure can be constructed from a L(p,q)-norm (i.e. a matrix norm, see [23]):

***Proposition 4.3**: For a data set $Y_{(n,p)}$, an overall quantitative measure of disclosure risk in the permutation paradigm is given by:*

$$D(\alpha,\beta) = \left[\frac{1}{p}\sum_{j=1}^{p} D_j(\alpha)^{\beta}\right]^{\frac{1}{\beta}} \; for\ \alpha \leq 1, \beta \leq 1\ and\ \beta \neq 0$$

$$and\ D(\alpha,\beta) = \prod_{j=1}^{p} D_j(\alpha)^{1/p} \; for\ \alpha \leq 1\ and\ \beta = 0$$

$D(\alpha,\beta)$ operates in two stages: it first measures disclosure risk for each attribute with $D_j(\alpha)$, then summarizes these p measures into a single one. Equivalently, it first aggregates the columns of the matrix formed by the collection of the p vectors of rank displacements r_j and then aggregates the p measures. $D(\alpha,\beta)$ is based on the expression of a L(p,q)-norm but does not define one strictly due to the $\sqrt[\alpha]{n}$ and $\sqrt[\beta]{p}$ factors and also the range of variation of $(\alpha;\beta)$: following the same reasoning as for α in $D_j(\alpha)$, β is set to range from one and below. This constraint is attached to the interpretation that can be given to the parameter β in the context of data anonymization. $\beta = 1$ is the benchmark case where all attributes in the data set are weighted equally: from a disclosure risk perspective, all attributes matter the same way. But when β decreases, more weight is given to the lowest protected attributes in the dataset; in the limit case with $\beta \to -\infty$, the overall disclosure risk of the data set is assessed through the perspective of the least protected attribute (i.e. the one having the lowest $D_j(\alpha)$ value). As for α in $D_j(\alpha)$, β in $D(\alpha, \beta)$ substantiates the variety of preferences in disclosure risk that users or releasers may have, but here this variety is expressed across attributes in the context of an overall diagnosis of disclosure risk for a data set.

Along the same lines, an overall measure of information loss for a data set can be constructed. Assuming that if in $Y_{(n,p)}$ its p attributes are to be masked, there are $j(j-1)/2$ potential sources of information loss (i.e. k distinct couples of attributes). Aggregating all these sources can be done by taking the norm of the matrix formed by the collection of the $j(j-1)/2$ relative permutation distances vectors $\Delta(r_k)$, which gives:

> ***Proposition 4.4****: For a data set $Y_{(n,p)}$ with p attributes to be protected against disclosure risk, an overall quantitative measure of information loss in the permutation paradigm is given by:*

$$I(\theta,\pi) = \left[\frac{1}{j(j-1)/2}\sum_{k=1}^{j(j-1)/2} I_k(\theta)^{\pi}\right]^{\frac{1}{\pi}} \quad for\ \theta \geq 1\ and\ \pi \geq 1$$

$I(\theta,\pi)$ also operates in two stages: it first measures information loss for every possible distinct couples of attributes, then summarizes these $j(j-1)/2$ measures into a single one. Equivalently, it first aggregates the columns of the matrix formed by the $j(j-1)/2$ vectors of *relative* rank displacement $\Delta(r_k)$ and then aggregates the collection of $j(j-1)/2$ measures. $I(\theta,\pi)$ is also based on the expression of a L(p,q)-norm and in fact does define one up to the $\sqrt[\theta]{n}$ and $\sqrt[\pi]{j(j-1)/2}$ factors. In particular, the range of variation of π is interpretable in term of information loss. $\pi = 1$ is the benchmark case where every couple of attributes in the data set are weighted equally and matter the same way in terms of information loss. When π increases, more weight will be given to the couple of attributes with the largest information loss; in the limit case with $\pi \to +\infty$, the overall information loss of the data set is assessed from the perspective of the least preserved couple of attributes (i.e. the ones having the highest $I_k(\theta)$ value). As for θ in $I_k(\theta)$, π in $I(\theta, \pi)$ substantiates the variety of preferences in information loss that users or releasers can have, but here such variety is expressed across attributes in the context of an overall diagnosis of information loss for a data set.

4.6 Conclusion

In this Chapter, we have derived two general classes of disclosure risk and information loss measures, which we argued are easy to compute for most methods and data sets, and which are meaningful. These two classes are based on the aggregative structure of p-norms (albeit they do not always define p-norms strictly), and the degrees of these norms can be harnessed with an interpretation in terms of aversion. In the case of disclosure risk, the aversion translates to different emphases on the lowest permutation distances achieved among records for one attribute. For information loss, the aversion translates in different emphases on the highest relative permutation distances among records between two attributes. While data releasers and users would similarly like to achieve the unattainable ideal of data with maximum protection against disclosure risk and minimal information loss, in practice, they are likely to have different judgments regarding utility/risk trade-offs. The measures developed in this Chapter allow both the incorporation of this diversity and, importantly, communication about them, notably with the new graphical representations of risk and information proposed. In addition, these measures allow the derivation of unanimity of judgments following the concepts of dominance introduced.

5 A GENERAL CIPHER FOR INDIVIDUAL DATA ANONYMIZATION

5.1 Introduction

The permutation paradigm unambiguously establishes a common ground upon which any anonymization method can be evaluated. However, this paradigm was not originally considered by its authors as a new anonymization method *per se*, but instead as a way to evaluate any method applied to any data set. This statement can be reconsidered. As will be proposed in this Chapter, the fact that it provides a post-anonymization common ground makes it also suitable for an ex-ante approach to data anonymization where, in fact, anonymization can be performed directly from permutation. This is the objective of this Chapter, which develops a new approach to data anonymization by proposing a general cipher based on permutation keys, bringing SDC closer to cryptography, and which appears to be equivalent to a general form of rank swapping [10,24]. Beyond the existing methods that this cipher can universally reproduce, it also offers a new way to practice data anonymization based on the exploration of different permutation structures. This cipher can be used to perform anonymization in an ex-ante way instead of being engaged in several ex-post evaluations and iterations to reach the protection and information properties sought after. The subsequent study of the cipher's properties additionally reveals some new insights into the nature of the task of anonymization taken at a general level of functioning. Finally, to make this cipher operational, this Chapter proposes the introduction of permutation menus in data anonymization, where the universal measures of disclosure risk and information loss proposed in the preceding chapter are used ex-ante for the calibration of permutation keys. To justify the relevance of their use, a theoretical characterization of these measures is also proposed. The contributions in this Chapter are currently under review (see [46]).

5.2 Definition and properties of a cipher for data anonymization

5.2.1 Data anonymization as a cipher

We start with a first proposition, which constitutes a direct consequence of the permutation-based paradigm:

> ***Proposition 5.1****: For a data set $X_{(n,p)}$ with n records and p attributes (X_1,..,X_p), its anonymized version $Y_{(n,p)}$ can always be written, regardless of the anonymization methods used, as:*
>
> $$Y_{(n,p)} = \left(P_1X_1, \dots, P_pX_p\right)_{(n,p)} + E_{(n,p)}$$
>
> *where P_1,..,P_p is a set of p permutation matrices and $E_{(n,p)}$ is a matrix of small noises.*

Proposition 5.1 highlights the fact that because permutation appears to be the overarching principle ruling data anonymization, the functioning of any method can be expressed as a set of permutation matrices, plus a matrix of small noises. Despite the large heterogeneity in the methods currently available, e.g. rank-based, noise-based, cluster-based, they can essentially all be viewed as the application of permutation matrices to the original data set. This proposition forms the basis upon which a cipher for data anonymization can be built. However, it remains limited in the sense that the permutation keys are not isolated. Indeed, except in the particular case where all the pairwise correlations across the p attributes are equal to one, the set of $P_1,..,P_p$ matrices will not measure the amount of permutation. To do so, each attribute needs first to be sorted in increasing order, which can be viewed as preliminary permutations, then the levels of permutations aimed at anonymizing the data set are introduced, and finally the sorting is undone through the inverse permutation matrix of the first step. This leads to the following proposition:

> ***Proposition 5.2****: For a data set $X_{(n,p)}$ with n records and p attributes ($X_1,..,X_p$), its anonymized version $Y_{(n,p)}$ can always be written, regardless of the anonymization methods used, as:*
>
> $$Y_{(n,p)} = \left(A_1^T D_1 A_1 X_1, \dots, A_p^T D_p A_p X_p\right)_{(n,p)} + E_{(n,p)}$$
>
> *where $A_1,..,A_p$ is a set of p permutation matrices that sort the attributes in increasing order, $A_1^T,.., A_p^T$ a set of p permutation matrices that put back the attribute in the original order, $D_1,..,D_p$ is a set of permutation matrices for anonymizing the data and $E_{(n,p)}$ is a matrix of small noises.*

Proposition 5.2 describes the fundamental functioning of any anonymization method, with the permutation keys made explicit. Proceeding attribute by attribute, each is first permuted to appear in increasing order, then the key is injected, and finally it is reordered back to its original form by applying the inverse of the first step (which in the case of a permutation matrix is simply its transpose). A small noise is also eventually added. Clearly, we have that $P_j = A_j^T D_j A_j \; \forall j = 1, \dots, p$ with $D_1,..,D_p$ subsuming the properties of any anonymization method by capturing the amount of permutation performed. For example, considering the following permutation matrix D_j applied to a given attribute j:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

and counting line by line how this matrix departs from the identity matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

i.e. how the 1's have been shifted by assigning a negative (resp. positive) sign for a right shifting (resp. left shifting), one can conclude that the first record has been moved 4 ranks down, the fourth 3 ranks up and the fifth 1 rank up, while the second and third records have been left in their original positions. These simple computations are a way of describing the functioning of any anonymization method, but in the language of permutation.

Proposition 5.2 thus considers data anonymization at a general level of operation and, following the permutation paradigm, contains all currently existing methods. Interestingly, its nature is similar to the functioning of rank swapping, where data are first sorted in increasing order, permuted within a limited range and then re-ranked according to their original values [26,24]. For example, consider the following permutation matrix for one attribute and 6 records:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

This matrix, when applied using *Proposition 5.2*, is a permutation key for rank swapping with a swapping distance equal to one. Thus, data swapping has a functioning that can in fact describe any anonymization method, while it is the swapping distance selected that constrains the structure of the permutation keys. Other methods, such as noise- or cluster-based, will lead to a different permutation structure, but ultimately they all boil down to a form of general rank swapping. However, working directly with permutation keys allows uncovering some permutation patterns that may not be mirrored by currently known techniques, which can potentially extend the set of anonymization tools available.

Now that a general key structure has been made explicit, we can define a cipher for data anonymization:

Proposition 5.3*: The three-tuple $\Gamma = (P, K, E)$ with the following conditions satisfied:*

- *P is a finite set of possible original and anonymized data sets of $n \geq 2$ records and $p \geq 1$ attributes*
- *K is the keyspace, a finite set of possible key groups k, each containing p permutation-based keys*
- *For each key groups $k \in K$ there exists a group of p permutation-based encryption rules $\varepsilon_k \in E$, where each group $\varepsilon_k: P \rightarrow P$ is a function such that $\varepsilon_k(x) = y$ for $\forall x, y \in P$*

is a cipher for data anonymization.

This proposition derives from *Proposition 5.2* and establishes the whole task of data anonymization as a cipher composed of three entities. The first one is the set of possible data sets *P* (i.e. the set of plaintexts in cryptography) of n records and p attributes, e.g. *$(X_1,..,X_p)$*, which also defines the set of possible anonymized data sets (i.e. the set of cyphertexts). The cipher is thus endomorphic [51]. It is indeed valid to define a cipher for data anonymization in the particular endomorphic case because, as outlined above, the essential principle of data anonymization is permutation. One can also add some small noises, which are in principle required to recompose *exactly* the outcome of some methods (for example noise-based ones). But the small noises will not change any ranks and thus will not provide any additional protection against disclosure risk. Instead, they will alter the data in a small but unnecessary way that could be detrimental to information. For example, adding small noises will not exactly preserve the marginal distributions of a data set, though such preservation remains a desirable feature of any anonymization tools. Stated otherwise, in data anonymization it is desirable and somewhat intuitive to expect that any information loss must have as a counterpart improved protection. This is not the case for these non-rank perturbative small noises, as only permutations matter. Consequently, as they do not provide any additional protection but instead lead to superfluous information loss, small noises can be disregarded from the definition of the cipher. Thus, and as for permutation ciphers in cryptography, the sets of plaintexts and cyphertexts are the same (while adding small noises would have made the two sets generally different).

The second entity of the cipher is the keyspace K. Here, it is important to note that a key is not defined as a single element, which is generally the case in cryptography but, following *Proposition 5.2*, as a group of p keys, i.e. (D_1,..,D_p), with p being the number of attributes in the data set to be anonymized. Otherwise put, each attribute is equipped with its own key, i.e. a permutation matrix, but this is the group of these p keys that forms the key used for anonymizing the whole data set. As will be made clear below, the relative properties of the elements within the key group can be used to assess information loss, a feature that differentiates data anonymization from standard cryptography.

Finally, the third element is the set of encryption rules, whereas for the keys an encryption rule is a collection of p specific rules for each attributes. From *Proposition 5.2*, those rules are given by e.g. $\left(A_1^T D_1 A_1, \ldots, A_p^T D_p A_p\right)$, and thus are all based on the products of permutation matrices. However, one crucial departure from standard cryptography is that no decryption rules are postulated and nor are they necessary. As noted in Chapter 2, individual data exchange does not require decryption per se. Once data have been anonymized with the desired levels of disclosure risk and information loss, they are meant to be released and used anonymized. The fact that decryption is not necessary considerably reduces the potential practical difficulties in implementing the cipher. For example, the problem of key exchange as in symmetric-key cryptography does not exist here. Moreover, in principle, one does not need to select only injective encryption functions to accomplish decryption in an unambiguous manner, albeit in practice it can be noted that because data anonymization relies on permutation, the encryption functions will necessarily be injective [51]. In any case, in the context of data anonymization, this concept appears to be irrelevant.

5.2.2 Some general principles in data anonymization

Having defined a cipher that streamlines the permutation paradigm in data anonymization and that can universally mimic any masking method, we can now characterize some of its properties that will *de facto* pervade the task of data anonymization in general. We start by a first property that establishes data independence in anonymization:

> ***Property 5.1****: Because it can be defined as a cipher, individual data anonymization can always be performed independently of the data to be anonymized. In particular, the distinction between a utility and privacy-first approach is fundamentally unnecessary.*

This first property is a simple but nonetheless pivotal consequence that stems from the possibility of formulating the task of data anonymization as a cipher. It means that the keys, and thus protection, can be handled and calibrated independently of the data. This may be counter-intuitive to certain SDC practitioners, as most of the existing techniques and their performances are linked to the data upon which they are applied. For example, for multiplicative noise injection with a given parametrization, changes in the distributional characteristics of the data may have a large impact on the level of protection [47]. More generally, the parameter values of a given method may be a poor indicator of the protection level achieved, as it is the conjunction of these parameters and the distributional characteristics of the data that will ultimately deliver the protection level. This explains why a round of trial and error is generally necessary in data masking. Even in a privacy-first approach, ex-post disclosure risk analysis is advised to check if a sufficient level of protection has been effectively achieved. The permutation paradigm, and in this Chapter its formulation as a cipher, solves this issue, as the permutation keys can be calibrated ex-ante with a given level of protection and thus of information that the encryption will automatically apply to, but independently of, the data. In particular, it turns out that both privacy and utility can be targeted simultaneously and one does not have to choose an approach ex-ante and check the other one (or even the two) ex-post.

Originally, the permutation paradigm was proposed to put the comparisons of different methods (and their different parametrizations) across different data sets on a common ground [39]. Thus, its main goal was the simplification of post-anonymization comparisons. But in fact nothing precludes, conceptually and practically, thinking about data anonymization only in terms of permutations. In turn, that means that permutation levels, and thus permutation keys, can be calibrated ex-ante to carry out anonymization instead of being retrieved ex-post to assess the effect of an anonymization method. Thus whatever the large heterogeneities in the analytical apparatus of SDC methods available, they all appear to have an underlying, common permutation-based structure that is independent of the data upon which they are applied.

> ***Property 5.2****: Information loss in data anonymization can only come from the alteration of the dependency among attributes, as the cipher Γ requires a permutation key per attribute.*

This property narrows the notion of information loss in data anonymization. As developed in Chapter 4, given the fact that the overarching principle of data anonymization is permutation, marginal distributions are necessarily always preserved as small noise additions in the reverse mapping procedure are unnecessary. Although they can still be considered, small noise additions are not a fundamental step for recreating the protection outcomes delivered by a method. As a result, the preservation of marginal distributions (non-disclosive in nature), a feature that could appear at first glance as a stringent requirement, is in fact implicitly fulfilled by any anonymization method. This property may also address some recurrent users' concerns about the way data have been modified during the anonymization process, where the addition of noise is sometimes viewed as non-acceptable by some users [35]. But in fact, any method can ultimately preserve marginal distributions and thus can always be analyzed on the anonymized data set in the same way as on the original data. In the cipher Γ this fact is made clear by each attribute being equipped with its own permutation key, leaving the attributes' distribution, taken in isolation, unchanged. Information loss can thus only occur from a change in the dependency among attributes, i.e. how attributes will be permuted relative to each other.

> ***Property 5.3**: The compounding of two or more anonymization methods is always an inefficient procedure as the cipher Γ is idempotent.*

Relying on permutation Γ is idempotent, i.e. $\Gamma \times \Gamma = \Gamma$. To see this, assume two unspecified anonymization methods applied sequentially on a given data set. Clearly, each of them has an underlying permutation structure, i.e. they can be expressed respectively as $\Gamma_1 = (\mathrm{P}, \mathrm{K}_1, \mathrm{E}_1)$ and $\Gamma_2 = (\mathrm{P}, \mathrm{K}_2, \mathrm{E}_2)$. The product cipher of Γ_1 and Γ_2, denoted $\Gamma_1 \times \Gamma_2$, is defined to be the cipher $(\mathrm{P}, \mathrm{K}_1 \times \mathrm{K}_2, \mathrm{E})$ [50]. But, the product of two permutation matrices is always a permutation matrix [4]. Therefore, there is no point in encrypting the data set first with the key K_1 and then with K_2, as it could have been done directly using a permutation key equal to the product of K_1 and K_2. In terms of anonymization, that means that compounding two methods necessitates two steps but cannot provide more protection than directly using a single step. Instead of targeting a protection level that is known to be reachable by the successive application of two methods (say, for example, additive noise addition then micro-aggregation), one can calibrate a group of permutation keys to reach this level directly. Consequently, the successive application of different methods is inefficient and anonymization can never reach different outcomes beyond the ones authorized within the set of all permutation keys.

> ***Property 5.4****: The cipher Γ is pure. Therefore, an adversary attacking an anonymized data set will always face the same kind of cryptanalytic problem, whatever the method used for anonymization.*

Attacks on a data set to re-identify individuals are generally and realistically conceptualized through record linkage, which can be used in the context of any anonymization method and disclosure scenario [15]. Many different record linkage attacks have been suggested in the literature (see for example [16] for an in-depth comparison between distance-based and probability-based procedures), but *Property 5.4* reduces the type of attacks that can take place on individual data to the same cryptanalytic problem. Because the cipher Γ is both endomorphic and idempotent, it is pure. But in a pure cipher, all keys are essentially the same [50,27]: whatever key is selected for encryption, an attacker will in fact calculate the same ex-post probabilities of the plaintext. In data anonymization, this translates into the fact that different masking methods ultimately deliver the same kind of challenge for an attacker. Consider, for example, two arbitrary noise-based and rank-based methods, say additive noise addition and rank swapping. Because additive noise aims at altering the magnitude of the data, one could intuitively think that a distance-based record linkage attack would turn out to be more efficient than a rank-based attack, while the reverse would be true for data swapping. Yet this is not the case. Because the functioning of any method can always be fundamentally described by an alteration of ranks through a pure cipher, it is ultimately rank-based record linkage attacks that are relevant for both, and in fact, for any anonymization methods.

Indeed, from a heterogeneous selection of methods it has been recently and experimentally remarked in the literature that rank-based record linkage attacks appear to seemingly and consistently outperform distance-based attacks [34]. While no firm explanation was proposed as to why this is the case, we believe that *Property 5.4* suggests a response. However, it must be noted that this proposition does not convey any additional elements about how to define an adversary, notably which kind of background knowledge one must be empowered with to lead to a reasonable and realistic attack scenario, which is a long-standing issue in the literature [12]. What *Property 5.4* claims is just that whatever the background knowledge assumed, the task of cryptanalysis is always the same and must be based on ranks.

5.2.3 Remarks on the maximum-knowledge attacker model and the validity of the Kerckhoff's principle in data anonymization

The issue of an attacker's background knowledge has been recently pushed further in the literature through the notion of a maximum-knowledge attacker [12], which defines an attacker who knows both the original data set and its entire corresponding anonymized version. This is a rather extreme configuration, unlikely to be mirrored by concrete situations, but it remains however conceptually very insightful, as anonymization that can pass the test of such a situation will in fact be able to pass any test. Note also that this concept provides an additional justification for the irrelevance of small noise additions in data anonymization, as a maximum-knowledge attacker can eliminate the small noise matrix of *Proposition 5.2* (being able to perform reverse mapping himself), which leaves him to uncover the permutation keys only [12].

The concept of a maximum-knowledge attacker is the equivalent of a known-plaintext attack in cryptography. Other types of attack exist but carry less meaning in an individual data exchange. A cyphertext-only attack, where only the anonymized data set is available, is the opposite of a known-plaintext attack, and while the latter may be seen as too stringent, the former is too naïve [12]. As for chosen plaintext and cyphertext attacks, they are relevant only in cases in which the attacker can interact with the cipher. Note that a maximum-knowledge attacker, observing both the original data set and its anonymized version, has nothing to gain in terms of information. One can view his attempt as purely malicious, trying to discredit the data releaser by revealing his permutation keys.

Now, given the assumption that such a person might exist, this leads to one question: given his power, is the task faced by a maximum-knowledge attacker so difficult? The answer relies on a consideration that has not been made explicit in the formulation of the cipher Γ: the record tracking numbers. Generally, data releasers can follow which anonymized record derives from which original record through a number that does not carry any information of any sort and is unaffected by encryption. Moreover, when the data are released, all numbers can be modified or deleted. But these numbers, known for practical purposes by the data releaser but not by the maximum-knowledge attacker, act in fact as a mask for the permutation keys. To make this clear, Table 5.1 illustrates the attacker's perspective, using a toy example.

Table 5.1: Point of view of a maximum-knowledge intruder.

	Original dataset X				Masked dataset Y		
ID	X_1	X_2	X_3	ID	Y_1	Y_2	Y_3
1	13	135	3707	[redacted]	8	160	3248
2	20	52	826	[redacted]	20	57	822
3	2	123	-1317	[redacted]	-1	122	248
4	15	165	2419	[redacted]	18	135	597
5	29	160	-1008	[redacted]	29	164	-1927

As previously mentioned, it is clear that the attacker can reverse-map the data and eliminate the small noise addition. In this example he has now to retrieve the permutation key (made of three permutation matrices). In fact, he is already observing some permutation matrices, but those are masked by his ignorance of the tracking numbers, which marks the limit of his knowledge. More explicitly, for each attribute he is observing the product $BA_j^T D_j A_j$: because he has no clue as to who is who between the original and the anonymized data, this is equivalent to assuming that, compared to the data releaser who obviously knows each and every term in the product $BA_j^T D_j A_j$, the attacker is facing an additional, unknown layer of permutation expressed by B. He is therefore only observing the resulting permutations patterns from the product but not its decomposition. More precisely, despite his knowledge of A_j and its transpose, the matrix D_j that he is trying to recompose is masked by B. As B is also a permutation matrix, the attacker is observing an unknown permutation of the encryption keys. As a result, even with his postulated power, due to B the attacker cannot avoid undertaking record linkage because $n!$ possible permutation keys by attributes exist, and only one will be the correct key.

The fact that the knowledge of the permutation keys will necessarily be hidden when the cipher Γ is used makes the Kerckhoff's principle fully relevant in data anonymization [12]. This principle states that the encryption method must be made available to the public while only the key must be kept secret. In data anonymization, the relevant key ultimately happens to be permutation, no matter how anonymization is practiced. Thus, that the cipher Γ has been used to protect the data can be made public, with the permutation keys remaining secret. Such a claim will not weaken the privacy guarantee offered by a data releaser but will contribute to greater clarity in individual data exchange, even in an environment comprised of maximum-knowledge intruders.

To summarize this section, we formulated data anonymization as an all-purpose cipher that is able to replicate the core functioning of any anonymization method. The formulation in terms of a cipher allows deriving some properties which, while standard in cryptography, when applied in the context of data anonymization, deliver some general guiding principles that, to the best of the author's knowledge, have not been identified so far in the literature. Surely, additional principles could be derived. In particular, one could note that the cipher Γ is, theoretically speaking, a one-time pad [50]. A direct consequence of this is that in principle, perfect secrecy could be achieved in data anonymization [51]. However, this possibility is a theoretical curiosum which has no empirical validity for at least two reasons, which we believe illustrate well the fundamental differences between cryptography and data anonymization. The first is that, as noted earlier, the notion of decipherment for individual data is not the same as in cryptography. While in the latter it took place when all the plaintext had been uncovered, in the former it is the amount of correct matches in a record-linkage attack that matters, i.e. which pieces of plaintext have been uncovered, and it does not have to be all of them. So, even in a one-time pad some correct matches could still be claimed. Thus the notion of perfect secrecy has no real meaning in data anonymization, except if one requires that all records must be re-identified to qualify a data set as not secure. This is rather unrealistic.

The second reason is that, for Γ to be strictly qualified as a one-time pad then the key selection should be truly random. While in cryptography this is fully acceptable, in data anonymization it is not. In addition to providing some privacy guarantees to individuals in the original data, the anonymized data should also meet data users' needs by providing some information. As a result, some structures and constraints must be applied to the permutation keys for the released data to be meaningful. The fact that in data anonymization the keys must be selected with both protection and information in mind precludes randomly generating them. In fact, this raises the question as to what should be the guidelines to calibrate the keys of the cipher in order to make it concretely usable. This will be discussed in the following section.

5.3 Calibration of the cipher's keys

In Chapter 4, power means have been proposed for the ex-post evaluation of disclosure risk and information loss, i.e. after having performed reverse-mapping for any method applied on any data set. But nothing precludes, neither conceptually nor practically, their use as ex-ante measures. In fact, it is one of the proposals of this Chapter to use power means as a guidance to calibrate the cipher's keys, as power means can be used equally effectively ex-ante or ex-post. However, before developing this notion, we provide a novel theoretical characterization of power means which, we believe, offers a powerful justification for their ex-ante use.

5.3.1 A theoretical characterization of power means

Power means satisfy a set of basic properties and are already well-known outside the field of data anonymization [25]. Here, and in the context of this Chapter, denoting a distribution of permutation distances by p=(p_1,…,p_n), being *relative or absolute*, J(p,α), the power mean of parameter α for the evaluation of p, satisfies the following:

- **Neutrality in evaluation (NE)**: if q is a permutation of p, then J(q,α)= J(p,α)

This condition ensures that all the information used to evaluate p is considered equally.

- **Size independence (SI)**: if q=(p,p,…,p) is a m-duplicate of p (with m≥2), then J(q,α)= J(p,α)

This condition connects the comparability of J(p,α) across data sets of different sizes, by establishing the ground for comparison on a per record basis.

- **Normalization (NO)**: if p_i= p_j=a for i,j=1,…,n, then J(p,α)=a

Normalization ensures that if all the permutation values in p are equal, then J(p,α) is equal to this permutation value.

- **First degree homogeneity (FD)**: if q=λp for a scalar λ>0 J(q,α)= λ J(p,α)

If the levels of permutation are magnified by the same scalar, so is the power mean.

- **Continuity (CO)**: J(p,α) is continuous

A standard assumption, continuity makes sure that the power mean does not change abruptly for small variations in p.

- **Sub-domain coherency (SC)**: For p’ and p of the same size and q and q’ of the same size, if J(p’,α) > J(p,α) and J(q’,α) = J(q,α), then J((p’,q’),α) > J((p,q),α)

Sub-domain coherency establishes that if the absolute or relative permutation distances from two sub-data sets change in a way that leads to an increase in the power mean in one and remains unaltered in the other, then the overall power mean must increase. Stated otherwise, if absolute permutation distances increase in one sub-set but remain unchanged in the rest of the data set, then protection against disclosure risk must increase on the overall data set. Along the same lines, if relative permutation distances increase in one sub-set but remain unchanged in the rest of the data set, then information loss must increase in the overall data set.

The fact that the class of power means satisfies (NE), (SI), (NO), (FD), (CO), (SC) is trivial. However, less trivial is the fact that this is the only class of measures to do so:

> ***Theorem 5.1:*** *An aggregative structure for the evaluation of disclosure risk and information loss satisfies (NE), (SI), (NO), (FD), (CO) and (SC) if and only if it is a power mean.*

Proof: *For necessity, we left the proof to the reader. For sufficiency, we start by assuming a function J(.) that satisfies (NE), (SI), (NO), (FD), (CO) and (SC). In what follows, permutation distances can be defined in relative or absolute terms indifferently.*

Consider the universe of all possible data sets of at least 3 records, i.e. $n \geq 3$, and pick in this universe four of them which, after anonymization, generate four distributions of permutation distances: p and q of size $m<n$, and p' and q' of size $m'=n-m$. Then, assume that $J(p,p') \geq J(q,p')$. (SC) precludes having $J(p) < J(q)$, which thus implies $J(p) \geq J(q)$. If this inequality holds strictly, then by (SC) we have $J(p,q') \geq J(q,q')$. But if inequality is not strict, then by (SC) $J(p,q') < J(q,q')$ does not hold because $J(p,q,q') < J(q,q',p)$ would contradict (NE). As a result, we have $J(p,p') \geq J(q, p') \Rightarrow J(p,q') \geq J(q, q')$. That means, bearing in mind that J(.) is assumed to verify (CO), that J(.) is strictly separable in every data set partition, which implies, following [5], that J(p) can be expressed as:

$$J(p) = Z_n\left(\sum_{i=1}^{n} \Omega_n(p_i)\right)$$

for every p of size n and with $\Omega_n(.)$ *continuous and* $Z_n(.)$ *continuous and strictly increasing.*

So far, what has been demonstrated is that (SC), (NE) and (CO) leads inevitably to a separable function. Now, what follows works along the same line as [2], which uses separabality to characterize power means.

By (NO) we have $a = Z_n(\sum_{i=1}^{n} \Omega_n(a))$ *for a>0, which leads to* $Z_n^{-1}(a) = n\Omega_n(a)$. *Assuming* $H_n = Z_n^{-1}(a)$ *with* $H_n(.)$ *continuous and strictly increasing,* $J(p)$ *can be rewritten as:*

$J(p) = H_n^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} H_n(p_i)\right)$ *for every p of size n≥3*

From this last equation assume $H = H_4$ *and m=4n. We can write:*

$$H\big(J(p)\big) = H\left[H_m^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} H_m\big(H^{-1}(H(p_i))\big)\right)\right]$$

$= \Theta_m^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} \Theta_m(H(p_i))\right)$ *with* $\Theta_m(.) = H_m\big(H^{-1}(.)\big)$ *strictly increasing and continuous*

Once again, we have $\Theta_m(a) = a$ *and in particular* $\Theta_4(a)$. *From here set p with n=2, p' its 2-duplicate and p'' its m-duplicate. (SI) implies (with in what follows* $w_i = H(p_i)$*):*

$$H\big(J(p'')\big) = \Theta_m^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} \Theta_m\big(H(p_i'')\big)\right)$$

$$= \Theta_m^{-1}\big(0.5 * \Theta_m(w_1) + 0.5 * \Theta_m(w_2)\big) = H\big(J(p')\big)$$

$$= \Theta_4^{-1}\big(0.5 * \Theta_4(w_1) + 0.5 * \Theta_4(w_2)\big) = 0.5 * (w_1 + w_2)$$

Thus, $\Theta_m(.)$ *must satisfy:*

$$0.5 * \Theta_m(w_1) + 0.5 * \Theta_m(w_2) = \Theta_m(0.5 * (w_1 + w_2))$$

This last equation is a Jensen's functional equation having the following solution [1]:

$\Theta_m(b) = a_m * b + c_m$ *for some scalars* a_m *and* c_m.

This solution implies for m=4n:

$$H\big(J(p)\big) = \frac{1}{m}\sum_{i=1}^{m} H(p_i)$$

Now, for a given data set with n≥1 and its four-duplicate, with p and p' the respective distribution of permutation distances, it holds by (SI) that

$$H\big(J(p)\big) = H\big(J(p')\big) = \frac{1}{n}\sum_{i=1}^{n} H(p_i) = \frac{1}{4n}\sum_{i=1}^{m} H(p_i')$$

In turn, this implies that:

$$J(p) = H^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} H(p_i)\right]$$

Now, consider a data set with two observations and a scalar $\vartheta > 0$. *By (FD) and the equation above it holds that (with in what follows* $w_i = H(p_i)$, *meaning that* $H^{-1}(w_i) = p_i$*):*

$$H\big[\vartheta H^{-1}\big(0.5 * H(p_1) + 0.5 * H(p_2)\big)\big] = 0.5 * H(\vartheta p_1) + 0.5 * H(\vartheta p_2)$$

$$\Rightarrow H[\vartheta H^{-1}(0.5 * w_1 + 0.5 * w_2)] = 0.5 * H\big(\vartheta H^{-1}(w_1)\big) + 0.5 * H\big(H^{-1}(w_2)\big)$$

$$\Rightarrow H^{\vartheta}[H^{-1}(0.5 * w_1 + 0.5 * w_2)] = 0.5 * H^{\vartheta}\big(\vartheta H^{-1}(w_1)\big) + 0.5 * H^{\vartheta}\big(H^{-1}(w_2)\big)$$

with $H^{\vartheta}(a) = H(\vartheta a)$ *for a>0*

Now, assuming $L^{\vartheta}(a) = H^{\vartheta}\big(H^{-1}(a)\big)$ *we have:*

$$L^{\vartheta}(0.5 * w_1 + 0.5 * w_2) = 0.5 * L^{\vartheta}(w_1) + 0.5 * L^{\vartheta}(w_2)$$

Following [1] the solution to this Jensen's functional equation is:

$L^{\vartheta}(b) = x^{\vartheta} * b + y^{\vartheta}$ for some scalars x^{ϑ} and y^{ϑ}.

Now, using H(b)=a it holds that:

$$H(\vartheta b) = x(\vartheta)H(b) + y(\vartheta)$$

Following [19] the solution to this functional equation is:

$$H(b) = \begin{cases} g * b^{\alpha} + h\, for\ \alpha = 0 \\ g * ln\, b + h\, for\ \alpha \neq 0 \end{cases}$$

But given that $J(p) = H^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} H(p_i)\right]$ we thus have:

$$J(p, \alpha) = \begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n} p_i^{\alpha}\right)^{\frac{1}{\alpha}}\ for\ \alpha \neq 0 \\ \prod_{i=1}^{n} p_i^{\frac{1}{n}}\ for\ \alpha = 0 \end{cases}$$

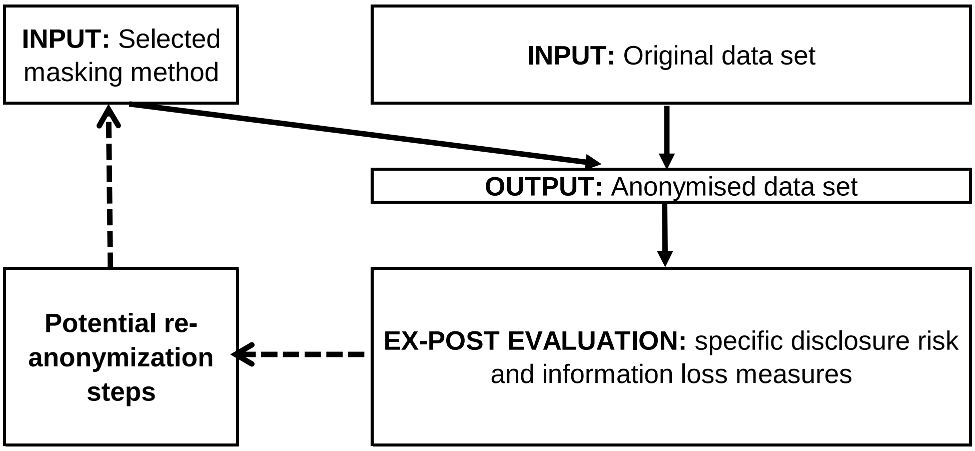
Thus, $J(p, \alpha)$ is a power mean, which completes the proof.

This result establishes power means as the only aggregative structure which, alongside a set of standard properties, satisfies sub-domain coherency. It is a result valid beyond the context of data anonymization, in fact for any vector of any quantity to be evaluated. It must also be emphasized that power means have been previously theoretically characterized in the literature [2], but by postulating at the onset the condition of separability. The result in this Chapter extends this previous work by demonstrating that separability appears to be in fact based on three conditions: neutrality in evaluation, continuity and sub-domain coherency. It is this last condition that is of particular and practical importance for data anonymization, as it turns out that *only power means can coherently cope with anonymization by block of records*.

5.3.2 Ex-ante calibration of permutation and a new approach to data anonymization

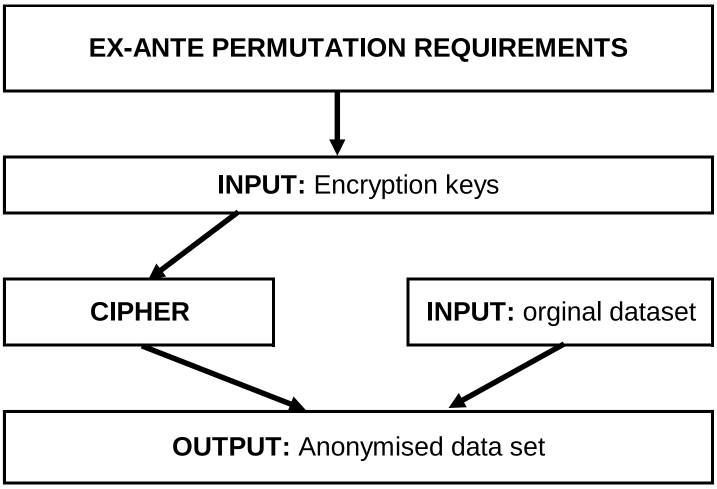
As stated earlier, data anonymization is currently practiced using a variety of methods, often very heterogeneous in nature and with some of them now very well-established in the literature. However, regardless of the many choices available, at a general level they are all used the same way (Figure 5.1). A method is selected with the anonymization practitioner having in mind either a utility-first or a privacy-first approach, and is applied to a data set. The outcome of this is then evaluated using specific measures of disclosure risk and information loss. But as mentioned earlier, because the methods' parameters in themselves are a poor guide to inform about the final levels of privacy and information obtained, as for a given parametrization different outcomes are possible according to the distributional features of the data, a necessary and specific ex-post checking step leads generally to some re-runs before reaching an anonymized version of the data viewed as acceptable. Additionally, because the ex-post checking is specific, the comparison of performances across different methods is an arduous task [35].

Figure 5.1: Current approach to data anonymization.


INPUT: Selected masking method
INPUT: Original data set
OUTPUT: Anonymised data set
Potential re-anonymization steps
EX-POST EVALUATION: specific disclosure risk and information loss measures

We have already seen that the use of power means on absolute and relative permutation distances provides a ground for universal ex-post checking, based on the retrieval of the permutations pattern that a method has generated. But at the conceptual level, the fact of using a method that unavoidably leads to a permutations pattern (plus eventually but unnecessarily a small noise addition), or applying this permutation pattern directly by using the cipher previously developed, is equivalent. These two ways will lead strictly to the same outcome in terms of risk and information. However, the latter appears to be more efficient, as once the permutations pattern has been set, it will be automatically translated into the final, anonymized data set. In fact, this will avoid the empirical ex-post checking stage and some eventual iteration to attain the desired levels of disclosure risk and information loss. This leads to a new approach for the practice of individual data anonymization (Figure 5.2).

Figure 5.2: New approach to data anonymization.



Of course, for this new approach to be practical, it requires thinking about anonymization only in terms of permutation. The permutation paradigm already pointed out that any anonymization method is equivalent to applying permutations. This is in a way a new language for data anonymization. With classical methods it is primarily their parameters (for example, the variance for noise addition or the parameter α in Chapter 3), and their varying strengths the language, which allow translating some targeted levels of disclosure risk or information loss into practice, albeit due to the varying nature of the data this translation is rarely perfect in the end. Now, to set permutation as a language to perform anonymization ex-ante, it is needed to expand its vocabulary so as to provide guidance on how to build the cipher's keys.

As we saw in Chapter 2, a data exchange generally requires two groups of agents: a data provider and the data users. The former wants to disseminate some individual data for some users that are in need of them. But prior to the exchange the provider, equipped with some raw, non-anonymized data, needs to secure them so that no individuals could be reasonably identified, while at the same time providing an acceptable level of information. To achieve this, he will undertake data anonymization himself. Now, we can introduce a new third agent, the permutation provider, whose task is to build some suitable permutation keys. Clearly, this new agent will never need to see the data. He can just work in isolation on the keys, having as information the number of attributes and individuals in the data, signalled by the releaser. However, what the releaser has to do is to formulate some desiderata on how he wants the data to be anonymized. This can be expressed through a permutation menu.

First, and for disclosure risk, the data releaser must advise the amount of permutation for each attribute. For example, for a given attribute, he can advise that he wants all records permuted at least one time, while at the same time a certain average of permutations must be achieved. For other attributes, these constraints can be modified, for example not all individuals must be permuted, or the average amount of permutation can be lower or reinforced, for example every individual must be permuted at least two times and the average amount of absolute permutation must be high. Second, and for information loss, the releaser must notify which couple of attributes are critical in terms of information and must be preserved to a large extent, with a small average of relative permutation distance. The other less valuable couples in terms of information can then be relatively permuted higher on average or within a certain portion of the distribution of relative permutation distances. Obviously, all the requirements in a permutation menu must be formulated simultaneously, as the keys taken in isolation make up for disclosure risk, while it is their relative properties taken by pair that make up for information loss. The data releaser must then formulate all his demands simultaneously to the permutation provider and must pay attention to the coherence of his requests, bearing in mind for example that two attributes cannot be protected with very dissimilar keys if at the same time their joint distribution has to be reasonably preserved. Keeping up with such coherence simply means coping with the unavoidable protection/information trade-off in data anonymization. In fact, *in an ex-ante approach information and privacy must be dealt with simultaneously*.

Now, power means constitute a way to create a permutation menu. For different scenarios of risk and information aversion, different levels of power means can be required ex-ante, from which the permutation provider will reconstitute the permutation keys. Of course, technically speaking it is clear that there may be no unique way to create permutation matrices from various values of power means. This will not affect the overall level of protection and information for the anonymized data set, while of course it could change the property of verifiability by the subjects [12]: for a given set of power means values and the associated levels of protection and information, different keys could lead to a given individual being permuted differently. This is, however, a minor issue. There may also be no permutation keys that can be derived from a set of power means, but this problem can be avoided to begin with by ensuring the coherence of the permutation menu proposed.

While power means is one way of creating a permutation menu for then generating keys, it must be recognized that there may be other ways. However, we just saw that power means are the only measures that are sub-domain coherent, which is a powerful justification for using them. Notably, and as far as big data are concerned, it can offer some obvious practical benefits. For instance, *anonymization can be performed by blocks to ease the computational workload*: when the data are split in m blocks, with some given levels of protection and information on m-1 blocks, the anonymization of the mth block will lead to an increase in protection of the *overall* data set. Such coherence cannot be ensured by other measures.

5.3.3 Examples of permutation menus

We now provide some empirical examples of permutation menus. Those menus are conceived independently of any method, i.e. based on power means guidance only. One might note, however, that Figures 4.1, 4.2 and 4.3 are in fact permutation menus *retrieved* from existing methods. The experimental data set used is, as in Chapter 4, two attributes of the Census data set observed over 1080 records. Let's assume the following:

- For the first attribute, we require that all records must be permuted at least one time and that the average level of absolute permutation must be high (menu 1). Alternatively, we require a low level of average absolute permutation in conjunction with a large chunk of records not being permuted (menu 2).
- For the second attribute, we require quite similar menus with a large chunk of records not be permuted at all, while we also set menu 1 to have an average level of absolute permutation almost twice as high than menu 2.
- As a result, we aim at two different scenarios for information loss. With menu 1, the keys for the two attributes are relatively dissimilar in their profiles, not least because the first key must permute all records while the other not. However, with menu 2 the keys are relatively similar. Consequently, we purposefully relax the constraint of information preservation for menu 1 while menu 2 must preserve it to a great extent.

Figures 5.3, 5.4 and 5.5 display the resulting permutation requirements when one starts from power means desiderata, creates the associated vectors of absolute and relative rank displacements and then generates the underlying permutation matrices. Notably, one can see that in the second menu relative permutation distances are small for whatever scenario of aversion to information loss, while the contrary holds true for the first menu (Figure 5.5). This result is ensured by the similar absolute permutation profiles for the two attributes requested in menu 2 (Figures 5.3 and 5.4). Now, when thinking about data anonymization only in terms of permutation as a universal approach, as we just did, the data can then be anonymized using the created keys and the cipher of *Proposition 5.3*. The ex-post properties in terms of disclosure risk and information loss will be strictly the same as the ones determined ex-ante.

Figure 5.3: Permutation menus for the first attribute.

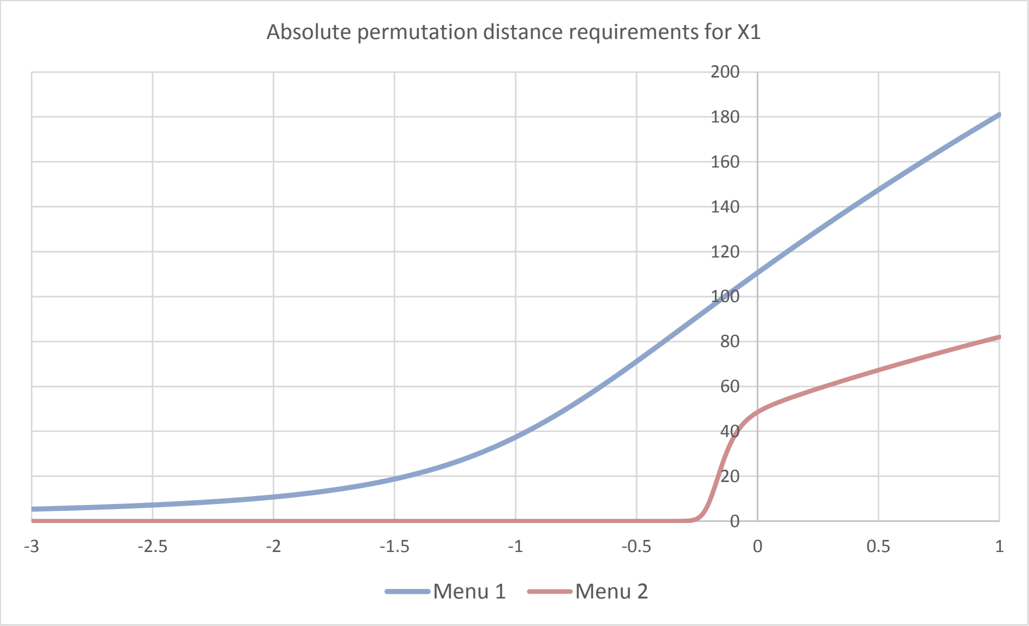


Figure 5.4: Permutation menus for the second attribute.

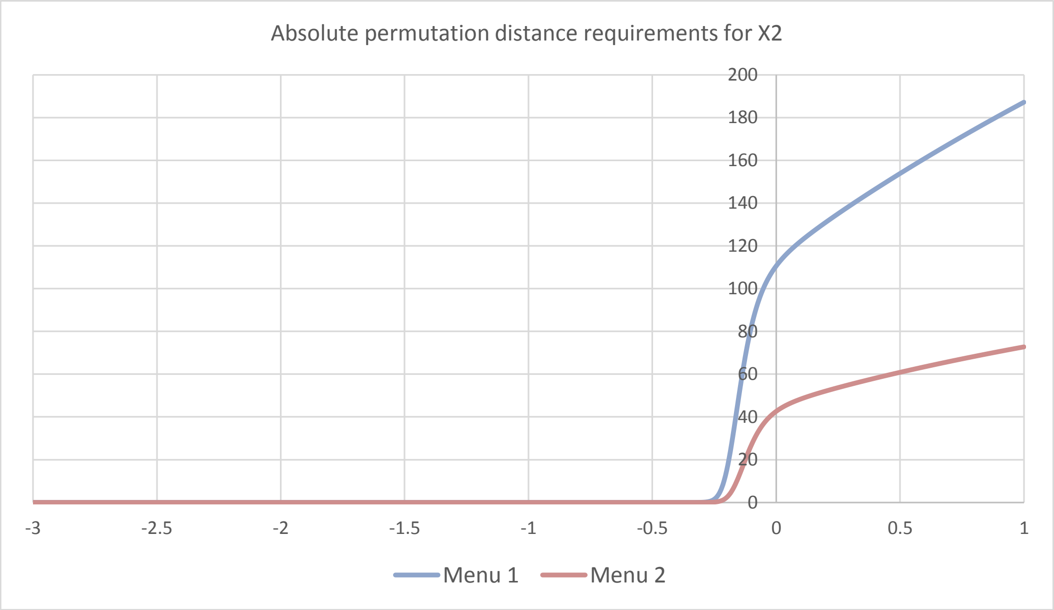
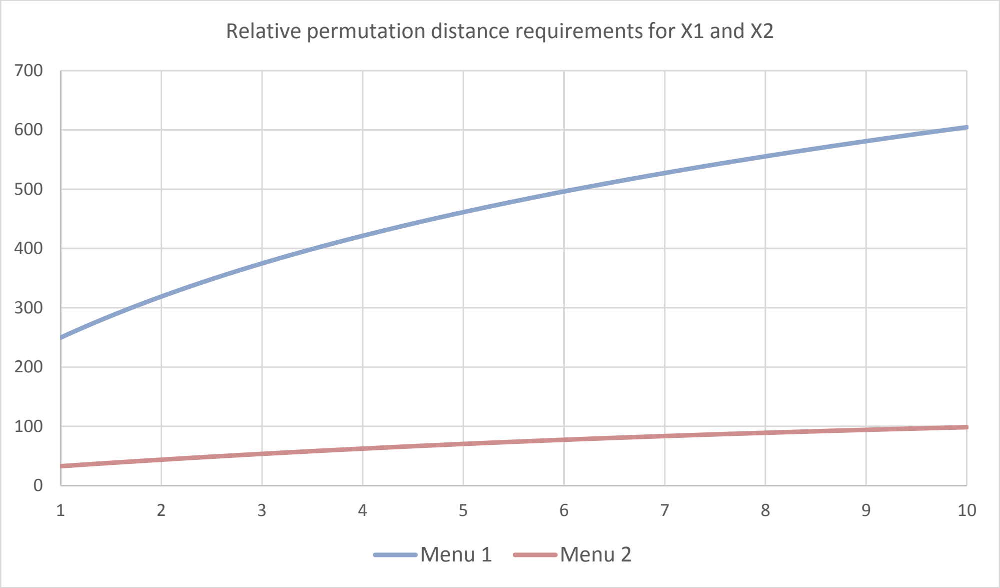


Figure 5.5: Permutation menus for the joint distribution.



5.4 Conclusion

The permutation paradigm was not seeking a new anonymization framework *per se*, but instead tried to establish an analytical environment for the comparison of currently existing methods in a sound and universal way. In this Chapter, we have challenged this limitation of scope by arguing that it can be as effective pre-anonymization as post-anonymization. Borrowing from cryptography, we have developed for the first time a general cipher for data anonymization. This cipher is able to replicate the outcome of any method, and some of its properties outline general lessons for data anonymization. In particular, at a general level of functioning, anonymization can always be performed independently of the data to be anonymized. As a result, beyond being a universal mimicker, the cipher is a tool in itself that can be used through the exploration of permutation structures. We then provided some guidance about how to explore these structures, notably by proposing to calibrate permutation keys using power means, for which we also suggested a new theoretical justification. The tools proposed in this Chapter can allow for a more efficient, ex-ante approach to data anonymization.

6 LONGITUDINAL DATA ANONYMIZATION AS PERMUTATION

6.1 Introduction

There are several types of individual data that can be published in a privacy – preserving way for fulfilling analysis needs, e.g. relational data, transaction data, sequence data, trajectory data, graph data… These data types differ in structure, properties and the information they contain about individuals. The dissemination of any specific type entails its own privacy risks and information preservation requirements, which should ideally be considered by the SDC approach selected to perform anonymization. Among these different types, longitudinal data are of particular interest in many areas, e.g. economics, medical research, sociology, finance, marketing... A dataset is longitudinal if it contains information on the same variables of interest about an individual at several points in time. For example, the information collected in clinical trials to evaluate the impact of treatments, or the dynamic of an individual's income, is longitudinal data. They are built from the pooling of observations on a cross-section of individuals over several time periods, achieved by surveying a number of individuals and following them over time.

However, despite the fact that the SDC literature offers a wide variety of tools suited to different contexts and data types [21], there have been very few attempts to deal with the challenges posed by longitudinal data. To the best of the author's knowledge, only one approach, formulated in the context of medical data and based on global suppression and generalization, has been proposed so far [49]. Hence, the objective of this Chapter, building on the permutation paradigm, is to contribute to filling this gap by proposing a general framework and some associated metrics of disclosure risk and information loss tailored to the specific challenges posed by longitudinal data anonymization. The contributions in this Chapter are currently under review.

6.2 Longitudinal data

Longitudinal data are repeated observations of the same respondents that are published at different points in time and are ubiquitous in a wide range of fields: medicine, public health, education, business, economics, psychology, biology, and more. Economists generally refer to it as panel data. They vary from cross-sectional data, i.e. where individuals are observed at a single point in time, and from time-series data, i.e. where one single entity is observed along a generally long time-span, in the sense that the defining feature of longitudinal data is that the multiple observations within several individuals can be ordered across time. Longitudinal surveys generally use calendar time, months or years, as the dimension separating observations on the same subject. Although the notion of time in longitudinal data can be quite intricate [53], in this Chapter we will focus on repeatedly measured attributes that can be ordered along a line to describe the sequence of measurement.

Compared to cross-sectional data, longitudinal data provide some clear advantages as they are generally more informative. Cross-sectional distributions that look relatively stable can in fact hide a multitude of changes that can only be captured if the same set of individuals is followed over time. For example, spells of unemployment, job turnover, residential and income mobility are better studied with longitudinal data. Longitudinal data are also well suited to study states durations, e.g. disease, unemployment and poverty, and if the time dimension is long enough, they can shed light on the speed of adjustments to medical treatments or policy changes. For instance, in measuring unemployment, cross-sectional data can estimate what proportion of the population is unemployed at a point in time. Repeated cross-sections can show how this proportion changes over time. But only longitudinal data can estimate what proportion of those who are unemployed in one period can remain unemployed in another period.

Longitudinal data has the potential to be plagued by several problems, the main one being attrition. While nonresponse from individuals is a standard issue in cross-sectional data, it is a more serious problem in longitudinal data because different periods of the data can be subject to varying rates of nonresponse from individuals. This issue generally leads to what is called an unbalanced longitudinal data set, i.e. not every individual is observed every year, while in the case of a balanced data set all individuals are observed at all periods. While the former case may seem more realistic, it remains barely considered in practice, and unbalanced data are generally made de facto balanced by not considering as relevant information individuals not observed across all periods. For example, econometric analysis techniques are much easier to implement and more developed on balanced than unbalanced data [55]. In this Chapter, we will assume that the longitudinal data set to be anonymized is balanced; anonymization on unbalanced data remains an avenue for future research.

Now, it is clear that the anonymization of longitudinal data poses some specific challenges. While it is beyond the scope of the Chapter to exhaustively investigate the possible forms of an attacker's background knowledge specific to longitudinal data, we can outline the main ones. Indeed, such knowledge may be thought of with its own characteristics compared to other types of data, and in particular cross-sectional data, and thus will carry specific privacy challenges. For example, an adversary may know that someone has transitioned from unemployment to employment between two time periods. Thus, while the employment status can be considered as a quasi-identifier in cross-sectional data, the change in employment status over time is also in itself a quasi-identifier in longitudinal data and can be used as additional background knowledge for the attacker.

Along the same lines, changes in confidential attributes, such as salary, can also be viewed as a quasi-identifier: an attacker may, for example, not know the salary of an individual at two periods, but may know that it has increased significantly between the two and can use that information to conduct the attack. Thus, the individual may consider as a privacy risk the fact that someone can learn about his salary variation, even if his salaries at the two time periods are not disclosed, e.g. the two salary values have been masked enough to avoid attribute disclosure but the masked values can still increase over time, providing the intruder with insights. Thus longitudinal data generally expand privacy threats.

Now, this widening is also a widening of information specific to longitudinal data. This is in fact what makes them specifically valuable in the first place, and must be preserved to a lesser or greater extent for the dissemination of longitudinal data to be useful. The trade-off between privacy and information is thus very direct in longitudinal data: the information on the dynamics of several variables at the individual level is valuable but is also problematic from a privacy perspective. The metrics developed later in this Chapter for the measures of disclosure risk and information loss in the context of longitudinal data will rely on this direct link.

6.3 A permutation-based approach to longitudinal data anonymization

6.3.1 Backward mapping of attributes in longitudinal data

We start with an observation regarding the relationship between two attributes that are followed over time and over the same set of individuals, i.e. the data are balanced, as assumed above. In fact, and while the context and the goal are different, it can be noted that one attribute observed during two periods t and t+1 can also always be reverse mapped in a way to express the attribute in t+1 as a function of itself in t. This approach, general in its scope, will lead to a simple characterization of the essential information and privacy risks specifically contained in longitudinal data.

By definition, to be followed over time, an attribute must keep the same form and definition, e.g. if it is categorical in t it must remain categorical in t+1 and track the same categories; if it is numerical in t it must remain numerical in t+1 and capture the same variable. Let denote by $X_{j,t} = (x_{1,j,t}, \dots, x_{n,j,t})$ the values taken by attribute j in t and $X_{j,t+1} = (x_{1,j,t+1}, \dots, x_{n,j,t+1})$ its values taken in t+1. As noted above, n is assumed to remain constant between t and t+1. Note that no assumption is made as to the nature of the attribute j, except that it can always be ranked: it can be numerical, categorical or nominal. The knowledge of $X_{j,t}$ and $X_{j,t+1}$ allows expressing the later as a function of the former by disentangling the nature of information in longitudinal data, using the following algorithm:

```
Algorithm: backward mapping of attributes in longitudinal data
Require: attribute in t $X_{j,t} = (x_{1,j,t}, \dots, x_{n,j,t})$
Require: attribute in t+1 $X_{j,t+1} = (x_{1,j,t+1}, \dots, x_{n,j,t+1})$
For i=1,…,n do
    Compute k=Rank($x_{i,j,t+1}$)
    Set $z_i$=Rank($x_{k,j,t}$) (where $x_{k,j,t}$ is the value of $X_{j,t}$ of rank k)
End for
Return $Z_{j,t} = (z_{1,j,t}, \dots, z_{n,j,t})$
```

The resulting backward mapped attribute $Z_{j,t}$ expresses $X_{j,t+1}$ as a permutation of $X_{j,t}$. Because the point values of the attribute may change over time, particularly in the case of a numerical attribute, one must also add $E_{j,t,t+1}$, the difference between $X_{j,t+1}$ and $Z_{j,t}$, to get an exact recomposition of $X_{j,t+1}$ as a function of $X_{j,t}$. Then, and because $Z_{j,t}$ is a permutation of $X_{j,t}$, it always hold that (with $P_{T,j}$ denoting a permutation matrix) :

$$X_{j,t+1} = P_{T,j}X_{j,t} + E_{j,t,t+1} \qquad (6.1)$$

It must be noted that the backward mapping procedure used here is analytically similar to the reverse mapping procedure developed in [39] (and outlined in Chapter 2), but serves a completely different purpose. It does not deal with anonymization but allows characterizing the two types of temporal information available in longitudinal data *by viewing time as an anonymization procedure*. Indeed, equation (6.1) disentangles the effect of time on an attribute, leading to two entities.

First, time modifies an attribute by changing the ranks of the individuals in a distribution. Because $Z_{j,t}$ is a permutation of $X_{j,t}$, the change of ranks through time can always be captured by the permutation matrix $P_{T,j}$. Note that, for convenience, we use here the compact notation for the permutation matrices, as in *Proposition 5.1*, but $P_{T,j}$ can also be decomposed following *Proposition 5.2*, to make explicit the key for temporal permutations. Equation (6.1) means that the main feature of longitudinal data can always be represented by the same entities used to express any anonymization method. As will be apparent below, this will turn out to be convenient for thinking about longitudinal data anonymization in a very general way.

The second type of information produced by time is what can be qualified as residual trajectories, i.e. changes in the attribute's values within two ranks, and is captured by $E_{j,t,t+1}$. Such information is contextual in nature. For a categorical attribute, $E_{j,t,t+1}$ will be by definition null. In the case of a numerical attribute, it will capture the effect of time on an attribute not due to rank changes. For example, if the salary of an individual moves from rank 4 to rank 7 in the salary distribution, then his residual trajectory will be such that his salary will still be contained between the values of ranks 6 and 8. By nature, this information is less relevant than the permutation patterns contained in $P_{T,j}$: the major effect of time is rank changes. However, it cannot be entirely discarded: if, for instance, the salaries in an economy grow at the same pace for everyone between two periods and no rank changes occur, this overall increase can only be expressed by $E_{j,t,t+1}$. Thus, $E_{j,t,t+1}$ will notably capture *how the entire distribution shifts through time*, while $P_{T,j}$ will always capture *how individuals move within the distribution over time*.

6.3.2 The effect of anonymization on temporal information

Now, using equation (6.1), the anonymized versions of $X_{j,t}$ and $X_{j,t+1}$, denoted respectively by $X_{j,t}^A$ and $X_{j,t+1}^A$, can always be written, whatever the anonymization methods considered for the two periods, as:

$$X_{j,t}^A = P_{j,t}X_{j,t} + E_{j,t} \qquad (6.2)$$

$$X_{j,t+1}^A = P_{j,t+1}X_{j,t+1} + E_{j,t+1} \qquad (6.3)$$

where $P_{j,t}$ and $P_{j,t+1}$ are, following the permutation paradigm, the matrices used to describe the core functioning of the anonymization method used for the attribute observed in t and t+1 respectively, and $E_{j,t}$ and $E_{j,t}$ are the eventual matrices of small noises. Here again, we use the notation of *Proposition 5.1* for the sake of convenience, while the decomposition of *Proposition 5.2* allows to extract the permutation keys from $P_{j,t}$ and $P_{j,t+1}$.

From an information perspective, it is clear that equation (6.1) has to remain exactly conserved for the specific temporal information conveyed by the longitudinal data to stay untouched. Now, by substituting (6.1) in (6.3), using the expression of $X_{j,t}$ in (6.2) as a function of its anonymized version and keeping in mind that the inverse of a permutation matrix is its transpose, one gets after rearrangements:

$$X_{j,t+1}^A = P_{j,t+1}P_{T,j}P'_{j,t}X_{j,t}^A + \left[P_{j,t+1}\left(E_{j,t,t+1} - P_{T,j}P'_{j,t}E_{j,t}\right) + E_{j,t+1}\right] \qquad (6.4)$$

As a result, if the two anonymization methods used in t and t+1 do not alter temporal information, it must hold, by comparison of (6.1) and (6.4), that:

$$P_{j,t+1}P_{T,j}P'_{j,t} = P_{T,j} \qquad (6.5)$$

$$P_{j,t+1}\left(E_{j,t,t+1} - P_{T,j}P'_{j,t}E_{j,t}\right) + E_{j,t+1} = E_{j,t,t+1} \qquad (6.6)$$

Equations (6.5) and (6.6) describe how the two anonymization methods in t and t+1 must be related to preserve the temporal information. First, the principal source of temporal information $P_{T,j}$ appears to be encased by the two permutation matrices of each method. Thus, for $P_{T,j}$ to remain unaltered in the anonymized version of the data set, we see by (6.5) that the product of the anonymizing permutation matrix used in t+1, the permutation matrix capturing the effect of time, and the transpose of the anonymizing permutation matrix used in t must be equal to the permutation matrix capturing the effect of time itself (note that because it is a product of matrices the terms cannot be rearranged conveniently).

Second, using the fact that small noises turn out to be irrelevant to describe the core functioning of an anonymization method, we can simplify equation (6.6) to:

$$P_{j,t+1}E_{j,t,t+1} = E_{j,t,t+1} \qquad (6.7)$$

Thus, for the residual trajectories to be preserved $P_{j,t+1}$ must be the identity matrix, i.e. no anonymization must take place at all on the attribute in period t+1. Therefore, for equation (6.5) to be verified, $P_{j,t}$ must also be the identity matrix, i.e. no anonymization at all must also take place in period t. This rather pointless and unsafe setting can be ignored given the fact that residual trajectories do not constitute the bulk of the relevant longitudinal information. In the remainder of this Chapter, we will thus focus on equation (6.5) and its implication for longitudinal data anonymization.

6.3.3 Universal measures of disclosure risk and information loss for longitudinal data anonymization

The preceding section outlined a general way to conceive longitudinal data anonymization when time is seen itself as an anonymization procedure. It can be applied to any kind of attribute and stipulates that, compared to cross-sectional data, longitudinal data offer an essential but specific feature, i.e. the permutation matrix $P_{T,j}$ describing the effect of time on one attribute. This matrix contains the main source of information that must be preserved somehow but which simultaneously entails some privacy risks. Thus, as stated above, the flip side of disclosure risk in longitudinal data is information. A data user will appreciate knowing how the attributes' values of some individuals change over time, but a data releaser may worry that such information could contribute to the knowledge of an intruder and that it may be operationalized for re-identification. As a result, any modification of $P_{T,j}$ will decrease disclosure risk but will also induce some information loss. The information/privacy trade-off is thus of a very direct nature in longitudinal data.

For data anonymization to take place, equation (6.5) can never hold in practice. The question is thus more about how $P_{j,t+1}P_{T,j}P'_{j,t}$ will depart from $P_{T,j}$. Bearing in mind that the result of the product of some permutation matrices is always a permutation matrix, this question can be assessed considering that the encasing of $P_{T,j}$ by $P_{j,t+1}$ and $P'_{j,t}$ will lead to a different pattern of rank changes over time.

For instance, assume that between t and t+1 an individual moved 4 ranks up in the distribution, i.e. in the rank displacement vectors derived from $P_{T,j}$ this individual is assigned +4 (see Chapter 4). Assume also that after anonymization of the attribute in t and t+1, the same individual is characterized by having moved 5 ranks up, i.e. in the rank displacement vectors derived from $P_{j,t+1}P_{T,j}P'_{j,t}$, this individual is assigned +5. Anonymization has altered information but in a minor way, as the individual is now characterized by a move between t and t+1 close to his ex-ante anonymization move. However, it implies that this individual is not equipped with sufficient protection against disclosure risk, because his move in the anonymized data is very close to his move in the original data, and such closeness can still lead to a privacy threat by enlarging, albeit now imperfectly, the background knowledge of an intruder.

Now, assume that the same individual is, after anonymization, characterized by having moved 100 ranks up. Here, anonymization has altered information in a major way as the individual is now characterized by a move between t and t+1 quite dissimilar to his real, ex-ante anonymization move. But it also implies that this individual is now equipped with sufficient protection against disclosure risk, as his move in the anonymized data is far from his move in the original data. Such dissimilarity can now only poorly enlarge the background knowledge of an intruder, if not fool him.

As a result, small differences between the rank shifting vectors derived from $P_{j,t+1}P_{T,j}P'_{j,t}$ and $P_{T,j}$ mean high disclosure risk and low information loss for the anonymization of longitudinal data, while large differences mean low disclosure risk and high information loss. Thus, the values in the vector of differences between the rank shifting vectors retrieved from $P_{j,t+1}P_{T,j}P'_{j,t}$ and $P_{T,j}$ will account both for disclosure risk and information loss. How to evaluate this vector of differences leads to the following proposition:

Proposition 6.1*: Denote by $r_{T,j}$ and $r_{A,j}$ the rank shifting vectors retrieved from $P_{T,j}$ and $P_{j,t+1}P_{T,j}P'_{j,t}$ respectively, and by $r_{T,A,j,i} = r_{T,j} - r_{A,j} = (r_{T,A,j,1}, \dots, r_{T,A,j,n})$ the vector of differences between $r_{T,j}$ and $r_{A,j}$ over the n individuals for which the attribute j is available in t and t+1. The following aggregative structure:*

$$J(\alpha) = \begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n}\left(abs(r_{T,A,j,i})\right)^{\alpha}\right)^{\frac{1}{\alpha}} & for\ \alpha \neq 0 \\ \prod_{i=1}^{n}\left(abs(r_{T,A,j,i})\right)^{\frac{1}{n}} & for\ \alpha = 0 \end{cases}$$

forms a class of both disclosure risk and information loss measures for the evaluation of longitudinal data anonymization.

$J(\alpha)$ aims at measuring the extent of dissimilarity that anonymization introduced on temporal information, with α capturing the different emphasis on the rank changes. It inherits the universal properties of the measures of disclosure risk and information loss developed in the context of cross-sectional data in Chapter 4, by making abstraction of the interplay between the distributional features of the data and the analytics of the methods. As a result, it can be applied to any kind of longitudinal data and for the ex-post evaluation of any anonymization methods applied to any attribute followed over time.

6.4 Experimental investigation

The objective of this section is to illustrate the use and effectiveness of the universal measures of disclosure risk and information loss developed above. The experimental data set used is one attribute of the Census data set, observed over 1080 individuals. The experiment is the following, assuming that the attribute from the original data is considered observed in period t:

I. Time scenario 1: Given that in period t the attribute is closely distributed as a normal law, we randomly generated the attribute for t+1 from a normal law with the same standard error as in t but with a mean of 2% more, assuming that overall the attribute's value has increased.

II. Time scenario 2: We randomly generated some growth rates for each individual, constrained between -20% and 20%.

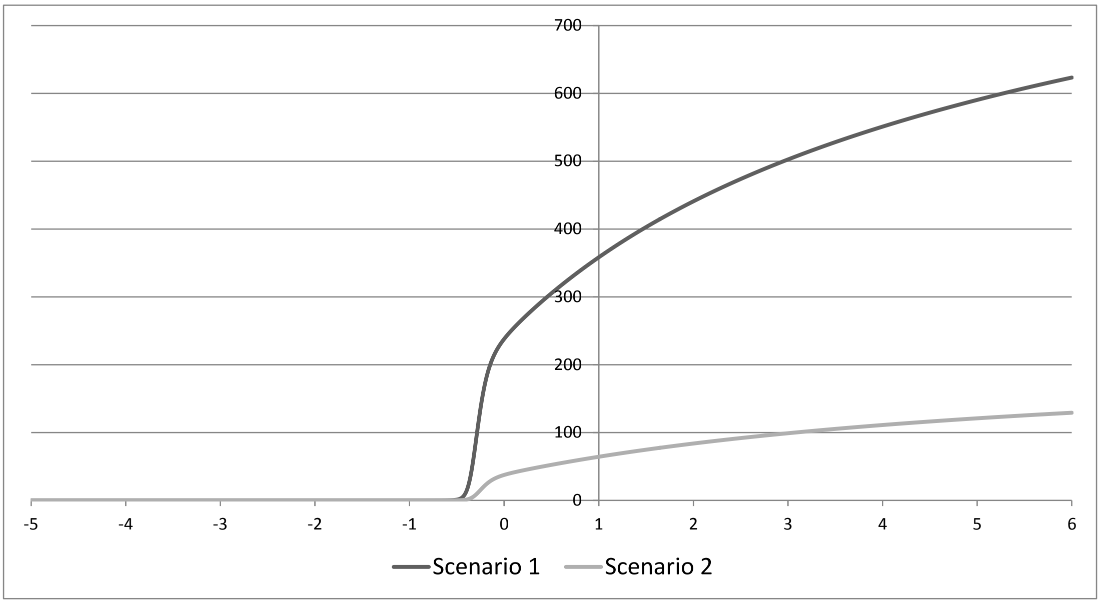
III. Anonymization methods: for each time scenario, the attribute in t has been anonymized using additive noise with a standard deviation equal to 50% of the standard error of the original values in t. For the attribute in t+1, we considered two versions: noise addition with half of the standard error in t+1 or the same standard error as in t+1.

IV. We then computed $r_{T,A,j,i}$, the values in the vector of differences between the rank shifting vectors derived from $P_{j,t+1}P_{T,j}P'_{j,t}$ and $P_{T,j}$, for each time scenario and anonymization procedures.

V. Finally, from these values we computed J(α) for a quasi-continuum of α parameters, that is by increments of 0.01. The results are displayed directly under the form of curves with the α parameters on the x-axis and the value of J(α) on the y-axis.

In this experiment, the purpose of having two time scenario aims at setting different longitudinal data configurations. In the first, the movements of individuals between t and t+1 are of larger magnitudes in terms of rank changes, while it is the reverse in the second. This can be seen in Figure 6.1, which shows the curves derived from applying power means under the same range of α to $abs(r_{T,j})$, i.e. the absolute values of the rank shifting vector derived from $P_{T,j}$. These curves demonstrate how time has moved individuals between t and t+1 and are a display of the essential time information contained in the longitudinal data, following the backward mapping procedure. In fact, for both curves a large chunk of individuals kept the same ranks between t and t+1, as both curves flat out at zero for α around -0.5. However, in the first time scenario the average level of rank changes (i.e. for α=1) is higher than for the second time scenario. When the focus is made on large rank changes (i.e. for α>1), scenario 1 also shows far greater magnitudes of rank changes.

Figure 6.1: Temporal information: time rank changes.



The effect of anonymization on longitudinal information can be seen in Figures 6.2 and 6.3. The curves displayed are the outcomes of anonymization on *both* disclosure risk and information. Indeed, individual trajectories through the attribute space represent the essential source of information brought by longitudinal data, but are also a specific source of disclosure risk. Thus a curve close to the x-axis means that anonymization didn’t alter time rank changes: disclosure risk is high but information loss is low. Conversely, a curve far above the x-axis means that time rank changes have been substantially distorted: disclosure risk is low but information loss is high.

One alternative way to consider this is viewing Figures 6.2 and 6.3 as two panels, taking α=1 as a dividing line. On the left, one is looking at disclosure risk first (by focusing on measures according relatively more weight to less altered time rank changes but with less information loss), while on the right one is looking at information first (by focusing on measures according relatively more weight to more altered time rank changes but with less risk of disclosure).

Figure 6.2: Disclosure risk and information loss: time scenario 1.

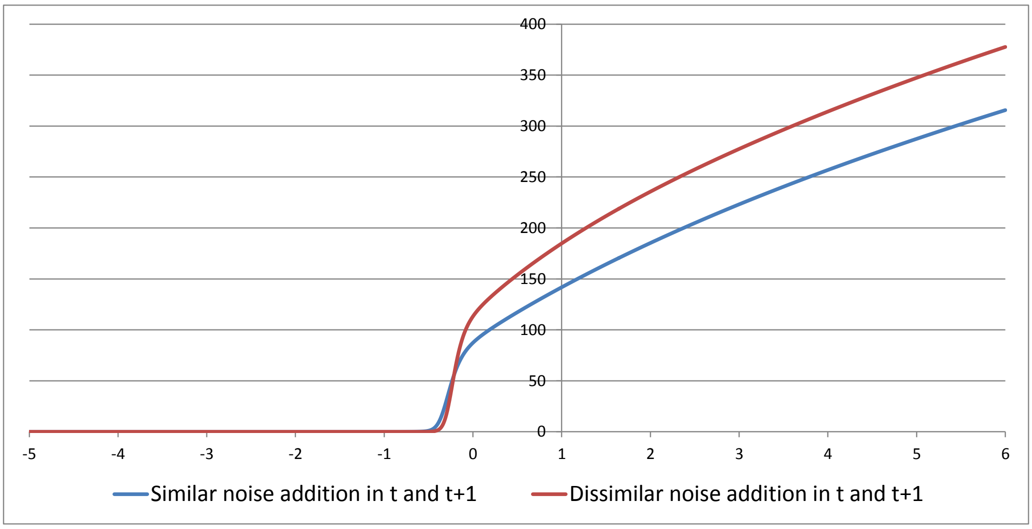
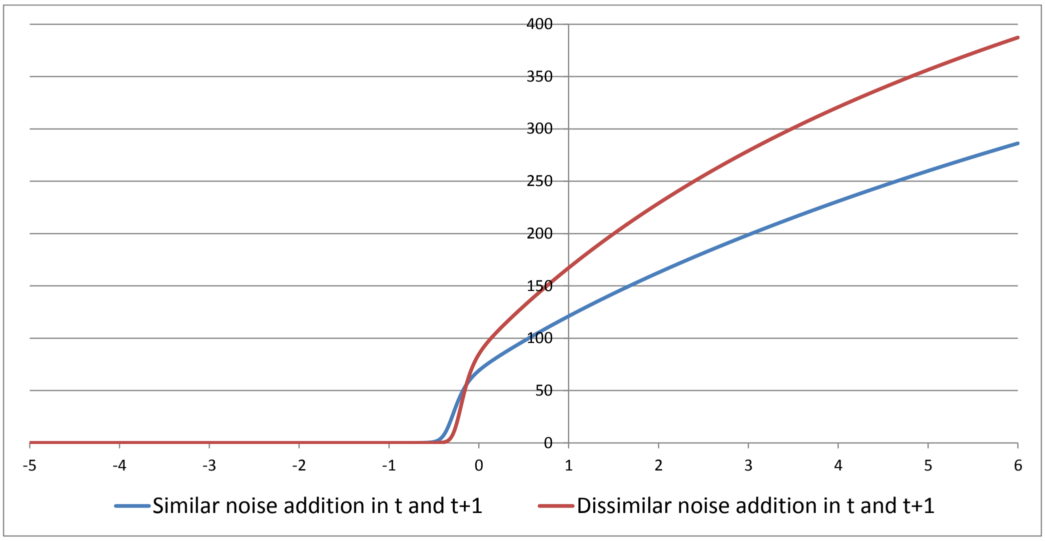


Figure 6.3: Disclosure risk and information loss: time scenario 2.



It seems that anonymization, when performed in a similar way between t and t+1, leads to less information loss and low protection against disclosure risk. This is a rather intuitive finding. When the attribute is anonymized with noise addition set as half of the standard error of the original data in t and t+1, the resulting curves are consistently lower than when the attribute in t+1 has been anonymized with the same standard error (Figures 6.2 and 6.3). It is thus clear that the dissimilarity in anonymization methods or parametrization through time will lead to better protection (but more information loss) of longitudinal data. However, and whatever the dissimilarity in methods, a large chunk of individuals is left with their time rank changes unmodified: across time scenario and anonymization methods, all curves are flat when crossing the geometric mean (i.e. for α=0) and below.

Finally, the dissimilarity in anonymization methods delivers the same outcomes whatever the time scenario considered. In Figure 6.2, time rank changes are altered in similar ways whether half or the same standard error of the original data is used to generate noise in t+1. This is also the case in Figure 6.3, albeit the differences are larger for the second time scenario when one is putting relatively more weight on the largest disruption in time rank changes.

6.5 Conclusion

The objective of this Chapter has been to investigate longitudinal data anonymization. We first presented a backward mapping procedure that allows expressing any kind of attribute observed in t+1 as a function of its values in t. This procedure has nothing to do with anonymization per se but allows viewing the supplementary information contained in longitudinal data, in particular compared to cross-sectional data, mainly as a permutation matrix. Thus the backward mapping procedure appears to analytically align the specificities of longitudinal data with the overarching tool of data anonymization.

From this general view on longitudinal data, we then characterized the effect of anonymization on temporal information: anonymization of an attribute over two periods always appears to encase temporal information, leading to a specific alteration of time rank changes. This alteration can then be evaluated using a class of universal disclosure risk and information loss, two outcomes that are tightly linked in longitudinal data. This Chapter established such measures using a power-mean based aggregative structure, following Chapter 4, and provided some illustrations.

Intended to be very general in its scope, this framework for longitudinal data anonymization supports a research question that has so far been over-looked in the SDC literature.

7 Synthetic data as permutation

7.1 Introduction

While generally considered as part of the SDC literature, the publication of synthetic data is an appealing alternative to, but also a significant departure from, pure SDC methods. The idea is simple: instead of disseminating an anonymized version of a dataset, i.e. the original data altered by the application of a SDC method, some data are instead created by drawing from a model fitted to the original data. At first glance it is clear that, since all values are synthetic and none of the individuals in the original data are included, disclosure risk must be low if not zero. The original data are used to build the synthesizer, and thus the contribution of an individual to a data set is not pointless but is in fact used only as an informational basis. As a result, synthetic data seem to offer a clear and almost definitive advantage compared to other SDC methods: it would seem that synthetic data can be made as close as possible to the original data without any strong consideration for disclosure risk, while for non-synthetic SDC methods similarity to original data must be traded off against disclosure risk (and hence utility is necessarily limited).

However, further scrutiny appears to weaken the advantage offered by synthetic data. For the sake of illustration, assume a dystopian society in possession of a perfect synthesizer, i.e. one that is able to perfectly replicate the statistical information observed over its population. In this case, an intruder using the synthetic data to conduct his attack may be able to re-identify some individuals or acquire some sensitive information about them. From the point of view of the individuals, the fact that the information acquired by the intruder is synthetic does not much alter the situation: their right to privacy has been violated. While from a legal perspective this situation may not be unlawful [54], from an ethical perspective this can clearly be qualified as a negative outcome. Of course, in real life the perfect synthesizer does not exist. But the better the job done by the data releaser to create the synthetic data, the closer an attacker can be to obtaining valuable information about some respondents in the original data. Thus it can be reasonably argued that, ultimately, synthetic data are somehow subject to the same kind of risk/information trade-offs faced by non-synthetic SDC methods. This is the purpose of this Chapter. Its contributions are currently under review

7.2 Synthetic data

Synthetic data rely on a principle that is by nature similar to the imputation of missing values in a data set. The idea is to fit a model, called a synthesizer, to the original data; values are then drawn from the synthesizer to replace original data rather than merely imputing missing data. Three types of synthetic data can be distinguished [26]:

- *Fully synthetic data*: no original data are released and the values of all attributes across all records are synthetic.
- *Partially synthetic data*: across some if not all records, only sensitive attributes are synthesized while, for example, quasi-identifiers are original values.
- *Hybrid data*: original and fully synthetic data are combined, and the resulting data can be more or less similar to the original or fully synthetic data.

The above distinction will not have any consequences in what follows in this Chapter, so we will use the term synthetic data indistinctively to point to any of the three types. However, what is common to them is obviously the pivotal role of the synthesizer. Generating synthetic data worth disseminating is work-intensive, not least because creating a synthesizer that can replicate the intricate features of a micro data set necessitates some time and an involved level of expertise. It is beyond the scope of this Chapter to discuss the relative merits of the several approaches available to create a synthesizer, as well as the criteria that can be used to gauge it (see [17] for an extensive discussion), but a general principle is that the level of information offered by a synthetic data set can be only as good as the quality of the underlying synthesizer used to generate it. In what follows, we will simply assume that the data releaser did a good enough job so that the resulting synthetic data are worth disseminating and being analyzed by users.

Regarding the practical characteristics of synthetic data, let us emphasize that they do not always come under the same format as the original data. First of all, they do not have to be of the same size, although having the same number of synthetic records as the number of original records seems a natural choice. To the best of the author’s knowledge, no firm guideline exists in the literature on this criterion (see, however, [43] for an empirical discussion). Depending on the context, an argument can be made for releasing synthetic data smaller than, same size as, or larger than the original data. Given this, we will assume that the number of synthetic records is the same as the original data. However, we will not restrict an equal number of synthetic and original records to the case, as one of the features of synthetic data is that they can come under any size. Specifically, we will outline below a pre-sampling procedure that can be applied before undertaking the evaluation of the privacy guarantees of synthetic data; this will allow gauging synthetic data sets of any size.

A second difference with non-synthetic SDC methods is that synthetic data generally lead to the dissemination of several data sets, while for the former methods only one set is released. This practice is motivated by the goal of capturing the different designs of the original data [44]. Clearly, such a feature can quickly become cumbersome for the users (as well as for the releasers, who need to generate the sets under various design configurations) and thus has to balance cost and accuracy [43]. Moreover, in the case where the original data are numerical and approximately multivariate normal, a sufficiency-based perturbation approach will perform at least as well as synthetic data for the preservation of information, while at the same time necessitating the release of only a single data set, which eases the tasks of the users [37].

Here again, no firm guideline exists as to the right number of data sets to be released. The original proposal of releasing multiple data sets postulates as a rule-of-thumb a typical number between 3 and 10 [44], but this number is in fact context-dependent and may vary according to the analytical needs of the users and the properties of the employed synthesizer [42]. In this Chapter, we will assume that an arbitrary number M of synthetic data sets is released. As we will demonstrate, this number will turn out to be critical for the privacy guarantees of synthetic data.

Finally, in the introduction of this Chapter we briefly touched upon the fact that disclosure risk in fully synthetic data must always be by nature non-existent. Such a claim has been made on various occasions in the literature, e.g. [17,18,42,43], though it must be mentioned that this conclusion is less clear-cut for partially synthetic or hybrid data [17,18] (which by construction will contain some of the original data). In these last two cases however, it is again generally assumed that the risk is very low. Recent contributions to the SDC literature concerning the notion of intruders cast a new light on this crucial feature of synthetic data. In this Chapter, we will use the notion of a maximum knowledge intruder presented in Chapter 5. But using synthetic data does have some implications for such an intruder. For non-synthetic SDC methods, the releaser has the advantage over the maximum-knowledge attacker in knowing the mapping between the tracking numbers in X and Y. The releaser can use this knowledge, for example, to assess how an individual has been protected; even the individual herself can verify her protection, if she can identify her own record in the non-synthetic data set. But *for synthetic methods the mapping between original and synthetic records does not make much sense*: a synthetic record does not derive from any specific single original record. Thus, *the advantage of the releaser over the maximum-knowledge attacker vanishes: both possess the same level of knowledge*. The privacy risk in synthetic data is not tied to a mapping: rather, it is connected with knowing that synthetic records exist that are very close to some original records. In fact, real and synthetic individuals are linked by information. This can be assessed by a multivariate version of a rank-based record linkage procedure that is developed below.

7.3 Synthetic data from the maximum-knowledge attacker perspective

7.3.1 Multiple reverse mapping of synthetic data

We begin by observing that a synthetic data releaser can always transform the data such that each attribute in each synthetic data set can be expressed as a permutation of the original data. This procedure, called reverse mapping, has been recently proposed in the literature for non-synthetic SDC methods [12,39]. This is the first time that it is developed for synthetic data.

Assume that a releaser generates $m = 1, \dots, M$ synthetic data sets $Y^m = (Y_1^m, \dots, Y_p^m)$ based on an original data set $X = \left(X_1, \dots, X_p\right)$; denote by $X_j = (x_{1,j}, \dots, x_{n,j})$ and $Y_j^m = \left(y_{1,j}^m, \dots, y_{n_m,j}^m\right)$ the values of attribute *j=1,…,p* over n records in the original data and n_m records in the mth synthetic data set, respectively. No further assumptions are made, except that the values of an attribute can always be ranked, which is obvious in the case of numerical or categorical attributes, but also feasible in the case of nominal ones [14].

In particular, the synthetic data sets need not be of the same size as the original data set. However, in order to perform reverse mapping we need to compare sets of the same size. This issue can be fixed as follows: when the synthetic data have more (resp. less) records than the original data, synthetic data can be randomly sub-sampled (resp. super-sampled):

- When $n_m > n$, a subset Q^m of size *n* is randomly selected;
- When $n_m < n$, a superset Q^m of size n is created by randomly generating *n-n'* additional records from the original *n'* ones;
- When $n_m = n$, the synthetic data are not modified and $Q^m = Y^m$.

Such a preliminary sampling procedure is viable provided that the original data set is large enough for it to be analytically interesting and representative. In the remainder of this Chapter, we will assume that $n_m = n, \forall m = 1, \dots, M$, keeping in mind that the pre-sampling procedure can be eventually used to align the sizes of every synthetic data sets with the size of the original data. The multiple reverse mapping of synthetic data is then performed as follows:

```
Algorithm: multiple reverse mapping of synthetic data
Require: original data set X, with attributes $X_j = (x_{1,j}, \dots, x_{n,j})$, for j=1,...,p
Require: synthetic data sets $Y^m$, for m=1,..., M, where $Y^m$ has attributes $Y_j^m = \left(y_{1,j}^m, \dots, y_{n,j}^m\right)$, for j=1,...,p
For m=1, ..., M do
For j=1,...,p do
For i=1,...,n do
          Compute k=Rank($y_{i,j}^m$)
          Set $z_{i,j}^m$= $x_{(k,j)}$ (where $x_{(k,j)}$ is the value of $X_j$ of rank k)
          Next i
Let $Z_j^m = (z_{1,j}^m, \dots, z_{n,j}^m)$
```

```
Next j
    Let data set $Z^m = (Z_1^m, \dots, Z_p^m)$
Next m
Return data sets, $Z^1, \dots, Z^M$
```

The resulting reverse-mapped attribute j in the mth synthetic data set Z_j^m expresses Y_j^m as a permutation of X_j. Since the point values of a synthetic attribute are unlikely to be the same as the point values of the original data, particularly in the case of numerical attributes, one must also add E_j^m, the difference between Y_j^m and Z_j^m, to get an exact recomposition of Y_j^m as a function of X_j. Then, and since Z_j^m is a permutation of X_j, it always holds that (with P_j^m denoting a permutation matrix; see *Proposition 5.1*):

$$Y_j^m = \mathrm{P}_j^m X_j + E_j^m, \forall j = 1, \dots, p \text{ and } \forall m = 1, \dots, M \qquad (7.1)$$

Equation (7.1) shows that, conceptually, a synthetic data set is functionally equivalent to i) permuting the original data; ii) adding some noise to the permuted data. But, since the noise added has to be necessarily small, as it cannot by construction alter ranks, it does not offer protection of any sort against disclosure risk. In fact, it represents an information loss (as it modifies the marginal distributions of a data set) that is not matched by a decrease in disclosure risk: if, for example, an attacker learns from a synthetic data set that the income of an individual is 102 while in reality it is 100, privacy has been violated in the same way as if the intruder was able to retrieve the exact value. Thus, the imprecision due to the small noise is not relevant for privacy. But any anonymization method, synthetic or not, must intuitively comply with the basic principle that any information loss triggered by anonymization must have a counterpart in terms of improved protection. Clearly, the small noise addition does not comply with this principle and can thus be discarded. As a result, the synthetic version of a data set invariably has an underlying structure that exactly preserves the marginal distributions of the original data (as they are simply a permutation of the original ones), but alters the relative ranks across attributes (see [46] and Chapter 5). Stated otherwise, what ultimately brings protection (and also information loss), even in synthetic data, are the changes in relationships between attributes.

At first glance, viewing synthetic data as a rank permutation may seem counter-intuitive. After all, as mentioned above, there is no mapping between the synthetic records and the original records. However, the synthetic data set tries to mimic the information in the original data set. In turn, this mimicked information can be expressed as a function of the original data, but with a different rank structure. Thus, at a fundamental level of functioning, a synthesizer can be viewed as a generator of different permutation structures of the original data, or equivalently, as a way to generate some permutation matrices for the cipher of Chapter 5. The generation of M synthetic data sets is thus equivalent to the generation of M permutation keys. As any non-synthetic SDC method is also equivalent to the generation of specific permutation matrices, *the distinction between synthetic and non-synthetic approaches to anonymization does not seem a fundamental one. As a consequence, synthetic methods must undergo disclosure risk scrutiny just like their non-synthetic counterparts*.

The ramifications of the above conclusion can be articulated further by recalling the example of a perfect synthesizer. In that case, with a perfect mimic of the information, all multivariate relationships must be exactly preserved. As a result, the permutation matrix has to be the identity matrix (which is a particular case of a permutation matrix where no permutation takes place) and the synthetic data set is the same as the original data set. More realistically, *the better a synthesizer is, the closer to the identity matrix each of the underlying permutation patterns contained in the multiple synthetic data sets being generated will be.*

Finally, while the purpose of this Chapter is to investigate the privacy guarantee of synthetic data, it must be noted that the results developed above have broader implications. A releaser could, for example, decide to release only reverse-mapped synthetic data sets. This solution would not entail additional privacy risks as we saw, but will always offer superior information quality due to the exact preservation of the marginal distributions. Each synthetic data set will thus convey a different rank structure according to the targeted design feature of the original data. Such a possibility is an avenue for future research.

7.3.2 Multiple rank-based record linkage attack

The multiple reverse mapping procedure can be easily engineered by the data releaser because he has at his disposal both the original and the synthetic data sets, as in the case of non-synthetic SDC techniques [39]. But as we have argued, in the case of synthetic data, the releaser and the maximum-knowledge attacker are at the same level of knowledge. Thus the attacker, who tries to perform the equivalent of a known-plaintext attack in cryptography, can also reverse map each synthetic data set, eliminate the small noise addition and ultimately be confronted with a collection of data sets that contain only the original data but with different permutation structures. Here, a fundamental departure from non-synthetic anonymization is that the attacker is entitled to several attempts to perform his attack. For instance, if trying to learn, say, the level of income of an individual, the attacker will try on the *M* data sets to retrieve the value. Intuitively, one can see that the question of privacy in synthetic data may be trickier than previously thought: the attacker, by retrieving *M* values of income during his attack, could be confused (if the values are very different), comforted (if the values are close), or most likely be helped by narrowing the range of potential values. That is, it is in fact possible that synthetic data may entail a higher degree of privacy risk than non-synthetic anonymized data (in the latter type of data, only one anonymized data set is typically released).

To mount the attack against synthetic data, the recently developed procedure of rank-based record linkage [34] can be repeated *M* times. We privilege this specific linkage type ahead of other types, e.g. distance-based linkage or probabilistic linkage, because data anonymization can basically be described as rank perturbation. Thus, rank-based record linkage appears to be the overarching procedure for evaluating disclosure risk (see [46] and Chapter 5 for a detailed explanation).

Denote by $O = (o_{ij})$ and $S^m = (s_{lj}^m)$ the rank matrices of the original data set and of the mth synthetic data set, respectively. The procedure of multiple rank-based record linkage on synthetic data is as follows:

```
Algorithm: multiple rank-based record linkage
Require: rank matrix $O$ of the original data set
Require: rank matrices $S^1, \dots, S^M$ of the M synthetic data sets $Y^1, \dots, Y^M$
For m=1,..., M do
For i=1, …,n do
For l=1, …,n do
    Compute $d_{il}^m = Criterion[abs(o_{i1} - s_{l1}^m), \dots, abs\left(o_{ip} - s_{lp}^m\right)]$
Next l
  Linked index of i in $Y^m$ = $arg\,min_l(d_{il}^m)$
  Next i
Next m
Return linked indices of i in the M synthetic data sets
```

This procedure is the multi-data set version of the procedure outlined in [34]. It reports the *M* possible matches of an original record with the *M* synthetic data sets. Several criteria can be selected, such as the sum or the minimum of rank differences. To evaluate the privacy guarantees of non-synthetic methods, the criterion will generally depend on the method, e.g. the sum for noise addition or the maximum for data swapping [34]. In the context of synthetic data, this choice is less clear and several criteria should ideally be considered.

7.4 Empirical illustrations

We now illustrate the concepts of multiple reverse mapping of synthetic data and multiple rank-based record linkage. The experiment is based, without loss of generality, on a small data set of 20 observations and three attributes, and proceeds as follows:

- The assumed original data set is generated by sampling N(50,10^2), N(500,50^2) and N(2500,250^2) distributions, respectively. The correlation coefficient between the first and the second attribute is 0.56, 0.25 between the first and the third, and 0.16 between the second and the third.
- *M=3* synthetic data sets are generated using a similar sampling procedure. The synthetic data are directly generated with the same size as the original data, although one can use the pre-sampling procedure developed above to eventually align the sizes of the former with the size of the latter.
- For the sake of illustration, we consider three different levels of closeness to the original data. As stated previously, the goal of this Chapter is not to discuss the issue of how to generate a satisfying synthesizer. Rather, by using three different sets, we try to account for the difficulty in generating a satisfying synthesizer:
    - The first synthetic data set is very close to the original data (but does not replicate them perfectly). It was sampled from the same normal distributions from which the original data set was sampled. As a result, the joint relationships between the three attributes are slightly altered (the correlation coefficient between the first and the second synthetic attribute is 0.52, 0.18 between the first and the third and 0.21 between the second and the third).
    - The second synthetic data set also has the joint relationships between the three attributes slightly altered (the correlation coefficient between the first and the second synthetic attribute is 0.44, 0.25 between the first and

the third and 0.21 between the second and the third) but with the properties of the marginal distributions not exactly preserved, i.e. the attributes are sampled from N(45,8^2), N(450,40^2) and N(2200,200^2) distributions, respectively.

- The third synthetic data set has its marginal distributions sampled from the same as the second one. However, no particular effort is made to preserve the joint relationships (the correlation coefficient between the first and the second synthetic attribute is 0.17, 0.12 between the first and the third and 0.09 between the second and the third).

Table 7.1 shows the multiple reverse-mapping procedure for the first attribute in the three synthetic data sets. It can be seen that each synthetic data set is expressed as a permutation of the original data. As outlined in the previous section, these versions do not entail more disclosure risk than the first generated synthetic data sets, but offer an improved level of information by exactly preserving marginal distributions.

Table 7.1: Example of multiple reverse mapping on synthetic data sets

Original data set			Synthetic data set 1				Synthetic data set 2				Synthetic data set 3			
ID	X1	Rank of X1	X1	Rank of X1	*Reverse-mapped X1*	Small noises	X1	Rank of X1	*Reverse-mapped X1*	Small noises	X1	Rank of X1	*Reverse-mapped X1*	Small noises
1	38	3	46	9	*51*	-5	33	2	*37*	-4	38	4	*39*	-1
2	66	19	36	1	*31*	5	54	19	*66*	-12	46	14	*57*	-11
3	56	12	43	5	*41*	2	50	16	*63*	-13	42	8	*50*	-8
4	53	11	59	14	*57*	2	37	6	*45*	-8	41	6	*45*	-4
5	31	1	41	4	*39*	2	43	13	*56*	-13	49	16	*63*	-14
6	63	16	61	16	*63*	-2	45	15	*61*	-16	49	17	*63*	-14
7	39	4	44	7	*49*	-5	33	3	*38*	-5	56	20	*70*	-14
8	63	17	56	13	*56*	0	41	11	*53*	-12	42	9	*51*	-9
9	51	9	76	20	*70*	6	40	9	*51*	-11	45	12	*56*	-11
10	56	13	49	10	*51*	-2	37	5	*41*	-4	53	19	*66*	-13
11	70	20	65	17	*63*	2	37	4	*39*	-2	42	7	*49*	-7
12	61	15	59	15	*61*	-2	43	12	*56*	-13	35	3	*38*	-3
13	41	5	40	3	*38*	2	32	1	*31*	1	44	11	*53*	-9
14	49	7	43	6	*45*	-2	51	17	*63*	-12	47	15	*61*	-14
15	51	10	53	12	*56*	-3	58	20	*70*	-12	28	1	*31*	-3
16	64	18	51	11	*53*	-2	39	8	*50*	-11	50	18	*64*	-14
17	45	6	66	18	*64*	2	45	14	*57*	-12	33	2	*37*	-4
18	57	14	44	8	*50*	-6	39	7	*49*	-10	42	10	*51*	-9
19	37	2	72	19	*66*	6	41	10	*51*	-10	40	5	*41*	-1
20	50	8	39	2	*37*	2	53	18	*64*	-11	46	13	*56*	-10

Now, a maximum-knowledge attacker can exactly perform reverse mapping for all attributes and can attempt to recreate the correct linkage. A releaser can also do the same to gauge the privacy of his synthetic data sets before release. Of course, identity disclosure may seem to be an odd notion for synthetic data but it is still conceivable: an attacker may try to identify which synthetic individuals are most similar to real individuals, i.e. trying to retrieve some clones. However, we believe that more interesting in the context of synthetic data is attribute disclosure, i.e. when confidential information contained in the synthetic data sets can be revealed and will closely or exactly correspond to the information of a real individual.

A maximum-knowledge attacker can conduct an attack on a specific attribute by ignoring his knowledge of this attribute in the original data; this is part of the flexibility offered by the maximum-knowledge attacker model [13]. The maximum-knowledge attacker can then use the multiple rank-based record linkage procedure to see how well he can recreate the ranks of the ignored attribute; that would simulate a partial-knowledge attacker who did not know the third original attribute and wanted to guess it. Table 7.2 shows the result of such an attack when knowledge of the third attribute of the original data set is ignored and the sum of rank differences criterion is used to perform multiple rank-based record linkage on the first and second attributes.

Table 7.2: Example of multiple rank-based record linkage: *third* attribute disclosure scenario

Original data set			Multiple rank-based record linkage: ranks identified by the intruder for X3		
ID	**X3**	**Rank of X3**	**Synthetic data set 1**	**Synthetic data set 2**	**Synthetic data set 3**
1	2228	2	11	9	8
2	2299	4	12	18	4
3	2534	10	1	8	12,17
4	2526	9	5	17	11
5	2336	5	16	13	2
6	2598	13	19	19	3
7	2736	16	2	9	8
8	2557	11	12	3	10.9
9	2704	15	17,4,5	16	12,2
10	2513	8	5	17	13
11	2942	19	17	3	10
12	2737	17	18	7	3
13	2559	12	2	2	8
14	2809	18	8	16	16
15	2195	1	4.5	16	11
16	2655	14	6,19	11	4
17	2963	20	15	5	15
18	2298	3	3	7	7
19	2382	6	11	9	8
20	2428	7	15	15	3,14

In this example, one can see that the outcome of an attack on synthetic data can either create confusion to a partial-knowledge attacker, or on the contrary help to narrow his knowledge of the attribute. Consider for example record no. 1 in the original data, with a value of rank 2 for the third attribute. What the attacker acquires information-wise is incorrect in each of the synthetic data sets, with a possible rank identified as ranging between 8 and 11. In fact, in that case, having multiple sets consistently orientates the partial-knowledge attacker in the wrong direction. The same is true for several records, e.g. nos. 7, 15, 17. For these individuals, it can be reasonably argued that synthetic data sets offer more privacy in that they fool the attacker consistently across all sets released.

Now consider records nos. 2 and 18. Respectively the third and first synthetic data sets perfectly disclose the attribute values of these records. But because the other sets point in another direction, the partial-knowledge attacker is again confused. As a result, synthetic data sets seem to provide better protection than non-synthetic approaches for these records. However, the partial-knowledge attacker can claim with reasonable confidence that the real value for record no. 2 is between ranks 4 and 18 of the original data and for record no. 18 between 3 and 7. That is, he can claim that the eighteenth individual has a value for the third attribute comprised between 2298 and 2428. Clearly, he has still gained some information from the synthetic data sets.

The information can also be narrowed for records where no exact attribute disclosure occurs across the three synthetic data sets in the first place. Consider, for example, records nos. 4 and 20. For the former, the attacker can claim that the real value is comprised between 2336 and 2737; for the latter, he can claim it is between 2298 and 2704.

Alternatively, assuming that the maximum-knowledge attacker now ignores his knowledge of the first attribute in the original data leads to the similar presence of edges in information (Table 7.3). For example, for records nos. 9 and 18 the knowledge of the first attribute is narrowed to a significant extent.

Table 7.3: Example of multiple rank-based record linkage: *first* attribute disclosure scenario

Original data set			Multiple rank-based record linkage: ranks identified by the intruder for X1		
ID	**X1**	**Rank of X1**	**Synthetic data set 1**	**Synthetic data set 2**	**Synthetic data set 3**
1	38	3	15,1	6	14
2	66	19	12	16,17	20
3	56	12	20	2,4	19,15
4	53	11	3,6	19	7,10
5	31	1	12,11	13	20
6	63	16	7	19,5	4,11
7	39	4	1,7,18	10	17
8	63	17	19	2,4	9
9	51	9	16	8	12
10	56	13	6	1,14	7,3
11	70	20	10	20	5
12	61	15	18	12,10	6
13	41	5	8,2	3	17,1
14	49	7	17	8	8
15	51	10	12	17	2
16	64	18	16	5	4
17	45	6	1	15	18
18	57	14	15	7	16
19	37	2	9	7	1,13
20	50	8	9,3,14	1,14	7,3,13

While these examples are meant to be illustrative, they tend to suggest that synthetic data do not come completely disclosure risk-free. Releasing multiple data sets can in fact be viewed as an additional privacy threat. Even if, by definition, no real individual is present in the synthetic data, some clones nonetheless are, and these clones can be re-identified to acquire some information about certain real individuals.

Originally, the proposal of releasing multiple data sets aimed at enhancing the quality of information offered by synthetic data. But, considering that such a practice can be cumbersome for users and that the quality of information can in some cases be made at least as well with a single data set [37], having multiple releases seems to entail some previously uncharacterized privacy risks that render this practice questionable.

7.5 Conclusion

It has frequently been claimed in the literature that disclosure risk in synthetic data must always be very low, if not zero. This Chapter challenges such statements. Despite the fact that no real individuals are included in a data release, at least as far as fully synthetic data are concerned, synthetic and real individuals remain linked by the information they convey. If an attacker is able to retrieve some information on real individuals that happens to be correct, it ultimately does not matter that this information is based on simulated data. Even if such a disclosure does not fall under the purview of any legislation on privacy, it can still be viewed as unethical insofar as it affects real individuals.

The objective of this Chapter was thus to investigate the privacy guarantee of synthetic data. Using recent advances in the literature on the definition of an attacker in data anonymization, we confronted synthetic data to an attack by a maximum-knowledge intruder. While conservative in its stance, this model has the ability to establish a common benchmark to gauge the privacy guarantees of non-synthetic anonymization methods. It thus seems plausible to consider synthetic data in the same context. Actually, the maximum-knowledge attacker is the counterpart of the popular and widely used notion of known-plaintext attack in cryptography.

We first presented an extension of a reverse-mapping procedure that can be performed both by an attacker and a synthetic data releaser. Under a reasonable assumption as to the size of the synthetic data sets to be released, this procedure shows that any synthetic data set can invariably be expressed as a permutation of the original data, in a way similar to non-synthetic SDC techniques. This result offers applications beyond disclosure risk assessment. For one thing, it is always possible to release synthetic data sets with the same privacy properties but with an improved level of information, because the marginal distributions can always be preserved without increasing risk. On the privacy front, reverse mapping leads to the consequence that the distinction made in the literature between non-synthetic and synthetic data is not so clear-cut. Both approaches must thus be evaluated against the same privacy challenges.

Next, we proposed an extension of the rank-based record linkage procedure that can also be performed both by the attacker and the synthetic data releaser. In particular, the latter can use it to assess the privacy guarantee of its synthetic data before release. This procedure shows that the practice of releasing several synthetic data sets for a single original data set gives rise to privacy issues that do not arise in non-synthetic anonymization (where typically only one anonymized data set is released). Indeed, the multiple releases can lead to better privacy guarantees, by confusing the attacker, or can facilitate attribute disclosure by helping the attacker narrow the range of the possible values that he is attempting to retrieve. An empirical investigation in the previous section illustrated these issues.

8 CONCLUSIONS

8.1 Summary of contributions

This thesis has dealt with privacy-preserving data publishing at a general level of functioning. Practitioners in this field currently benefit from a wide variety of methods and concepts available in the literature to foster dissemination and unleash new sources of information for the benefits of society at large. But such variety does not come without difficulties. Over the years, this literature has developed dynamically in numerous directions but with no overarching framework emerging. As a result, the current diversity of concepts, models and tools available makes complicated the task of selecting the optimal analytical environment in which to conduct anonymization and to evaluate privacy and information outcomes, due to the multitude of available choices.

Relying on recent contributions from the literature which established permutations as the core functioning of data anonymization, our main contributions are as follows.

- We have derived two general classes of disclosure risk and information loss measures, which we argued are easy to compute for most methods and data sets and can be used for the comparisons of any methods on any data sets. These two classes are based on the aggregative structure of p-norms, i.e. power means, and the degrees of these norms can be harnessed with an interpretation in terms of aversion. In the case of disclosure risk, the aversion translates to a different emphasis on the lowest permutation distances achieved among records for one attribute. For information loss, the aversion translates to a different emphasis on the highest relative permutation distances among records between two attributes. In addition, these measures can derive unanimity of judgments following the concepts of dominance introduced. Finally, some graphical representations of disclosure risk and information loss can be derived from these measures, which we believe can ease communication around privacy and information's outcomes.

- We then brought data anonymization closer to cryptography. Borrowing from the latter, we developed a general cipher for data anonymization which leads to a new approach to data anonymization. This cipher is able to replicate the outcome of any method, while some of its properties outline general lessons for data anonymization. In particular, at a general level of functioning, anonymization can always be performed independently of the data to be anonymized. As a result, beyond being a universal mimicker, the cipher is a tool in itself that can be used via the exploration of permutation structures. We then provided some guidance as to how to explore these structures, notably by proposing to calibrate permutation keys using the power means-based measures developed in Chapter 4, for which we also suggested a new theoretical justification. We believe that this allows for a new and more efficient, ex-ante approach to data anonymization.
- We then derived a general view on longitudinal data anonymization based on permutations. By noting that time can be conceived as an anonymization method, we presented a backward mapping procedure that allows expressing any kind of attribute observed in t+1 as a function of its values in t. This procedure allows: i) viewing the supplementary information contained in longitudinal data,

which has to be preserved but can also be a source of privacy risk, mainly as a permutation matrix; ii) analytically aligning the specificities of longitudinal data with the cipher developed in Chapter 5. From this approach, we then characterized the effect of anonymization on temporal information: anonymization of an attribute over two periods always appears to encase temporal information, leading to a specific alteration of time rank changes. This alteration can then be evaluated using the class of measures developed in Chapter 4.

- Finally, we reconsidered the privacy guarantees of synthetic data. Despite the fact that no real individuals are included in a synthetic data release, synthetic and real individuals remain linked by the information they convey: if an attacker is able to retrieve some information about real individuals that happens to be correct, it ultimately does not matter that this information is based on simulated data. Thus through a permutation-based approach, we first demonstrated that the distinction made in the literature between non-synthetic and synthetic data is not so clear-cut and, as a result, both approaches must be evaluated against the same privacy challenges. We then proposed an extension of a recently developed rank-based record linkage procedure that can be used to assess the privacy guarantee of synthetic data. This procedure shows that the practice of releasing several synthetic data sets for a single original data set entails privacy issues that do not arise in non-synthetic anonymization. Indeed, the multiple releases can lead to better privacy guarantees, by confusing the attacker, or facilitate attribute disclosure, by helping the attacker narrow the range of the possible values that he is attempting to retrieve.

8.2 Publications

The publications supporting this thesis are:

- *ISI JCR Journals*:
  - Nicolas Ruiz, "On some consequences of the permutation paradigm for data anonymization: Centrality of permutation matrices, universal measures of disclosure risk and information loss, evaluation by dominance", *Information Sciences*, Vol. 430–431, pp. 620-633, March 2018. Impact factor: 4.832. (Contributions presented in Chapter 4)
  - Nicolas Ruiz, "A general cipher for individual data anonymization", *under review for Information Sciences*, 2018. Impact factor: 4.832. Arxiv link: https://arxiv.org/abs/1712.02557 (Contributions presented in Chapter 5)
  - Nicolas Ruiz, "A multiplicative masking method for preserving the skewness of the original micro-records", *Journal of Official Statistics*, Vol. 28, No.1, pp. 107–120, 2012. Impact factor: 0.411. (Contributions presented in Chapter 3)
- *Lecture Notes in Computer Science*:
  - Nicolas Ruiz, Krishnamurty Muralidhar and Josep Domingo-Ferrer, "On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective", *Lecture Notes in Computer Science*, (Privacy in Statistical Databases - PSD2018), 2018. *Submitted*. (Contributions presented in Chapter 7)
  - Nicolas Ruiz, "A general framework and metrics for longitudinal data anonymization", *Lecture Notes in Computer Science*, (Privacy in Statistical Databases - PSD2018), 2018. *Submitted*. (Contributions presented in Chapter 6)

- *Other papers*:
    - Nicolas Ruiz, "Universal measures of disclosure risk and information loss for individual data anonymization", *Proceedings of the 4th URV Doctoral Workshop in Computer Science and Mathematics*, 2017.

8.3 Future work

From the contributions of this thesis, several avenues for new research can be pursued:

- An inventory of popular SDC methods under different parametrizations and data contexts should be established, using the class of measures developed in Chapter 4, in particular for benchmarking the values of these measures into existing practices. This could allow for characterizing the existing methods that are dominant in terms of disclosure risk and information loss, and in particular if some methods can be dominant in both, which could provide a strong rational for their use.
- As data anonymization relies on the single principle of permutation, which could be phrased as a general principle as "to be protected, become someone else", an intuitive privacy guarantee and thus a new privacy model around the cipher developed in Chapter 5 should be formulated.
- Exploring the composition of an approach by permutation is warranted, i.e. when merging two data sets with certain permutation patterns, the result of the merge with its subsequent privacy and information guarantees should be identified.
- The assessment of the notion of disclosure risk in longitudinal data anonymization should be deepened. In particular, how disclosure risk from time-variant attributes relates and combines with disclosure risk stemming from time-

invariant attributes, as generally longitudinal data sets contain both, should be examined.

- The possibility of considering synthesizers as tools to generate different permutation patterns, which could offer some insights for non-synthetic anonymization techniques, should be assessed.

9 REFERENCES

[1] Aczél J., *Lectures Notes on Functional Equations and their Applications*, Dover, 2006.

[2] Aczél J. and C. Alsina, "Synthesizing judgments: a functional equations approach", *Mathematical Modelling*, Vol. 9, pp. 311-320, 1987.

[3] Atkinson A. B., T. Piketty, and E. Saez, "Top incomes in the long run of history", *Journal of Economic Literature*, American Economic Association, Vol. 49(1), pages 3-71, March 2010.

[4] Bhatia R., *Matrix Analysis*, Springer-Verlag, 1997.

[5] Blackorby C. , D.Primont and R. Russel, *Duality, Separability, and Functional Structure*, North-Holland, 1978.

[6] Brand R., "Microdata protection through noise addition", *Inference Control In Statistical Databases*, LNCS 2316, 97-116, Springer-Verlag Berlin Heidelberg, 2002.

[7] Brand R., J. Domingo-Ferrer and J. M. Mateo-Sanz, "Reference data sets to test and compare SDC methods for the protection of numerical microdata", *Deliverable of the EU IST-2000-25069 "CASC" Project*, 2003.

[8] Burridge J., “Information preserving statistical obfuscation”, *Statistics and Computing*, Vol. 13, 321-327, 2003.

[9] Clauset A., C. R. Shazili, and M. E. J. Newman, “Power-law distributions in empirical data”, *SIAM Review*, Vol. 51, 661-703, 2009.

[10] Dalenius T., and R. Steven, “Data-swapping: A technique for disclosure control (extended abstract)”, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, Washington, DC, pp. 191–194, 1978.

[11] Domingo-Ferrer J., O. Farràs, S. Martínez, D. Sànchez and J. Soria-Comas, "Self-enforcing protocols via co-utile reputation management", *Information Sciences*, Vol. 367, pp. 159-175, Nov 2016.

[12] Domingo-Ferrer J. and K. Muralidhar, "New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users", *Information Sciences*, Vol. 337, pp. 11-24, Apr 2016.

[13] Domingo-Ferrer J., S. Ricci and J. Soria-Comas, "Disclosure risk assessment via record linkage by a maximum-knowledge attacker", *13th Annual International Conference on Privacy, Security and Trust-PST 2015, Izmir, Turkey*, Sep 2015.

[14] Domingo-Ferrer J., D. Sánchez and G. Rufian-Torrell, "Anonymization of nominal data based on semantic marginality", *Information Sciences*, Vol. 242, pp. 35-48, May 2013.

[15] Domingo-Ferrer J. and V. Torra, "Disclosure risk assessment in statistical disclosure control of microdata via advanced record linkage." *Statistics and Computing*, Vol. 13(4), pp. 343-354, 2003.

[16] Domingo-Ferrer J. and V. Torra, “A quantitative comparison of disclosure control methods for microdata" *In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, North-Holland, pp. 111-134, 2001.

[17] Drechsler J., *Synthetic Datasets for Statistical Disclosure Control*, Springer, 2011.

[18] Drechsler J., S. Bender and S. Rässler, "Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel", *Transactions on Data privacy*, Vol. 1, pp. 105-130, 2008.

[19] Eichorn W., *Functional Equations in Economics*, Addison-Wesley, 1978.

[20] Fuller W. A., "Masking procedures for micro-data disclosure limitation", *Journal of Official Statistics*, Vol. 9, 383-406, 1993.

[21] Fung B. C. M., K. Wang, R. Chen and P. S. Yu, "Privacy-preserving data publishing: a survey of recent developments", *ACM Computing Surveys (CSUR)*, Vol. 42, pp. 14:1-14:53, June 2010.

[22] Gao X., H. Xu, and D. Ye, "Asymptotic behaviour of tail density for sum of correlated lognormal variables", *International Journal of Mathematics and Mathematical Sciences*, Volume 2009.

[23] Grafakos L., *Classical Fourier analysis*, Springer-Verlag, 2008.

[24] Greenberg B., "Rank swapping for masking ordinal microdata", *US Census Bureau, (unpublished manuscript)*, 1987.

[25] Hardy G. H., J. E. Littlewood and G. Polya, *Inequalities*, Cambridge University Press, 1988.

[26] Hundepool A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.-P. de Wolf, *Statistical Disclosure Control*, Wiley, 2012.

[27] Kaliski B. S., R. L. Rivest and A. T. Sherman, "Is the data encryption standard a group?", *Journal of Cryptology*, Vol. 1, pp. 3-36, Jan 1988.

[28] Kleber C., and S. Kotz, *Statistical size distributions in economics and actuarial sciences*, Wiley Series in Probability and Statistics, Wiley-Interscience, 2003.

[29] Krishnamoorthy K., *Handbook of statistical distributions with applications*, Statistics: Textbooks and Monographs, Chapman & Hall/CRC, 2006.

[30] Laeven R. J. A., M. J. Goovaerts, and T. Hoedemakers, “Some asymptotic results for sums of dependent variables, with actuarial applications”, *Insurance: Mathematics and Economics*, Vol. 37, 154-172, 2005.

[31] Lydall H. F., *The structure of earnings*, Clarendon Press, Oxford, 1966.

[32] Marshall A. W., I. Olkin and B. C. Arnold, *Inequalities: theory of majorization and its applications*, Springer-Verlag, 2004.

[33] Muralidhar K., D. Batra, and P. J. Kirs, “Accessibility, security, and accuracy in statistical databases: the case for the multiplicative fixed perturbation approach”, *Management Science*, Vol. 41, 1549-1564, 1995.

[34] Muralidhar K. and J. Domingo-Ferrer, "Rank-based record linkage for re-identification risk assessment", *Lecture Notes in Computer Science, Vol. 9867 (Privacy in Statistical Databases - PSD2016)*, pp. 225-236, Sep 2016.

[35] Muralidhar K. and J. Domingo-Ferrer, “Microdata Masking as Permutation,” *UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, Helsinki, Finland*, October 2015.

[36] Muralidhar K. and R. Sarathy, “Generating sufficiency-based non-synthetic perturbed data”, *Transactions on Data Privacy*, Vol. 1, 17-33, 2008.

[37] Muralidhar K. and R. Sarathy, " A Comparison of Multiple Imputation and Data perturbation for Masking Numerical Variables ", *Journal of Official Statistics*, Vol. 22, pp. 507-524, 2006.

[38] Muralidhar K. and R. Sarathy, “An enhanced data perturbation approach for small data sets”, *Decision Sciences*, Vol. 36, 513-529, 2005.

[39] Muralidhar K., R. Sarathy and J. Domingo-Ferrer, "Reverse mapping to preserve the marginal distributions of attributes in masked microdata", *Lecture Notes in Computer Science, Vol. 8744 (Privacy in Statistical Databases - PSD 2014)*, pp. 105-116, Sep 2014.

[40] OECD, *Report to the MacArthur Foundation: feasibility study of a harmonised access to labour force and migration statistics (Phase I),* OECD Publishing, 2010.

[41] Piketty T. and E. Saez, “Income inequality in the United States 1913-1998”, *The Quarterly Journal of Economics*, Vol. CXVIII, 1-39, 2003.

[42] Reiter J. P., "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study", *Journal of The Royal Statistical Society Series A*, Vol. 168, pp. 185-205, 2005.

[43] Reiter J. P., "Satisfying disclosure restrictions with synthetic data sets", *Journal of Official Statistics*, Vol. 18, pp. 531-544, 2002.

[44] Rubin D. B., "Discussion: statistical disclosure control limitation", *Journal of Official Statistics,* Vol. 9, pp. 462-468, 1993.

[45] Ruiz N., "On some consequences of the permutation paradigm for data anonymization: Centrality of permutation matrices, universal measures of disclosure risk and information loss, evaluation by dominance", *Information Sciences*, Vol. 430–431, pp. 620-633, March 2018.

[46] Ruiz N., "A general cipher for individual data anonymization", *under review for Information Sciences* (https://arxiv.org/abs/1712.02557), 2017.

[47] Ruiz N., “A multiplicative masking method for preserving the skewness of the original micro-records”, *Journal of Official Statistics*, Vol. 28, No.1, pp. 107–120, 2012.

[48] Sánchez D., J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation", *Information Fusion*, Vol. 30, pp. 1-14, Jan 2016.

[49] Sehatkar M. and S. Matwin, “HALT: Hybrid anonymization of longitudinal transactions” *In Eleventh Conference on Privacy, Security, Trust (PST) (2013)*, pp. 127–134.

[50] Shannon C., "Communication theory of secrecy systems", *Bell System Technical Journal*, Vol. 28(4), pp. 656–715, 1949.

[51] Stinson D. R., *Cryptography: Theory and Practice*, *Third Edition*, Chapman and Hall/CRC, 2005.

[52] Soria-Comas J. and J. Domingo-Ferrer, “A non-parametric model for accurate and provably private synthetic data sets”, *Proc. of International Conference on Availability, Reliability and Security-ARES 2017*, art. no. 3. ACM, 2017.

[53] Weiss R. E., *Modelling Longitudinal Data*, Springer, 2005.

[54] Willenborg L. and T. De Waal, *Elements of Statistical Disclosure Control*, Springer, 2001.

[55] Wooldridge J. M., *Econometric Analysis of Cross Section and Panel Data*, *Second Edition*, The MIT Press, 2010.

UNIVERSITAT
ROVIRA i VIRGILI