Information Sciences 000 (2017) 1–14

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# 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

Nicolas Ruiz

*OECD, Rue André Pascal, 75016 Paris, France*

ARTICLE INFO

*Article history:*
Received 5 October 2016
Revised 2 October 2017
Accepted 25 October 2017
Available online xxx

*Keywords:*
Data anonymization
Statistical disclosure control
Permutation paradigm
Permutation matrices
Power means
Dominance

ABSTRACT

Recently, the permutation paradigm has been proposed in data anonymization to describe any micro data masking method as permutation, paving the way for performing meaningful analytical comparisons of methods, something that is difficult currently in statistical disclosure control research. This paper explores some consequences of this paradigm by establishing some class of universal measures of disclosure risk and information loss that can be used for the evaluation and comparison of any method on most data, under any parametrization and independently of the characteristics of the data to be anonymized. These measures lead to the introduction in data anonymization of the concepts of dominance in disclosure risk and information loss, which formalize the fact that different parties involved in micro data transaction can all have different sensitivities to privacy and information.

## 1. Introduction

Data on individual subjects are increasingly collected 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. In fact, such requirements shape the dissemination policy of micro 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 offering to users a reasonable richness of information. Moreover, 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. As a result, more parties, often very heterogeneous in their privacy and information requirements, are now involved in micro data transactions. This has opened a new range of questions and pressing needs 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 in 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

*E-mail address:* nicolas.ruiz@oecd.org

https://doi.org/10.1016/j.ins.2017.10.048
0020-0255/© 2017 Elsevier Inc. All rights reserved.

(see [8] for an authoritative survey). Over the years, it has burgeoned in many directions. In particular, techniques applicable to micro data, which are the focus of this paper, 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.

Streaming from the large variety of practical cases that can occur in micro data exchange is the diversity of techniques available for data anonymization. Such diversity is undoubtedly useful but has however one major drawback: a lack of agreement and clarity on the appropriate choice of tools in a given context, and as a consequence a lack of general view (or at best an incomplete one) across the relative performances of the techniques available. In fact, the cross-evaluation of current micro data masking methods is 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 [10]. 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 also a variety of definition for privacy and information loss, and picking some is often related to the context in which they are used and/or can result from an arbitrary choice. The fact that all evaluations can only be practical in nature and context-specific is clearly an issue, not least precluding 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.

A second reason is related to the variety of parties involved in micro data exchange. Indeed, it is natural to assert that across each party different sensitivities to privacy and information prevail. Some may place greater emphasis on the preservation of privacy, e.g. typically the data releasers, while others are relatively more concerned by the extent to which information is preserved, e.g. typically the researchers. Additionally, these sensitivities can differ also within groups, e.g. one researcher can 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.

A step toward the resolution of such limitations has been recently proposed [3,11], by establishing that any micro data masking method can be viewed as functionally equivalent to a permutation of the original data plus eventually a small noise addition. This insight, called the permutation paradigm, unambiguously establishes a common ground upon which any masking method can be gauged. It is independent of the underlying parameters of the masking mechanism and the characteristics of the data. Moreover, it presents the advantage of being meaningful and easy to grasp and implement, as the only necessary and sufficient information for the comparative evaluation of some methods, being under different parametrizations and/or different data sets, is a distribution of permutation distances. Thus, the permutation paradigm is also a tremendous simplifier for data anonymization.

While this paradigm is not considered by its author as a new anonymization method *per se* (a statement that can be reconsidered, see later), it offers the potential to re-interpret all the techniques available through the same lens. It remains however to develop a set of appropriate measures of disclosure risk and information loss based on permutation distances. This is the objective of this paper, which explores some 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 used for the evaluation of any anonymization method, independently of the context under which they operate. The construction of these measures allows introducing in data anonymization the notions of dominance in disclosure risk and information loss, which formalize the fact that different parties involved in micro data release can all have different sensitivity to privacy and information, and can inform about the methods that can reach a consensus among all parties involved. These two notions of dominance can in fact characterize which methods, under any tastes for privacy and information, always perform better than others.

This paper first starts in Section 2 with a brief reminder of the permutation paradigm and one of its first, simple consequence, which establishes permutation matrices as an encompassing tool in data anonymization. From permutation matrices, Section 3 derives a general class of disclosure risk measures and introduces the concept of dominance in disclosure risk. Section 4 then develops a general class of information loss measures as well as the related concept of dominance in information. Section 5 presents some empirical investigations on the two class measures introduced and notably how the concepts of dominance can shed a new light on the assessment of popular anonymization technique. Section 6 proceeds with possible extensions of the measures developed in this paper. Finally, conclusions and paths for future research are gathered in Section 7.

## 2. Centrality of permutation matrices in data anonymization

### 2.1. Restatement of the permutation paradigm

The current state of the literature on data anonymization offers a wide variety of techniques suited to different circumstances in terms of data, utility preservation and privacy requirement [8]. But as outlined above, this diversity in techniques also entails some difficulties in comparing the level of utility and privacy achieved through different methods on different data sets, as all of them are ultimately tied to the analytical framework selected, in particular their parameters which are data-dependent, and the underlying metrics used. This makes the comparison of different mechanisms such as e.g. additive vs. multiplicative perturbations, or the same mechanism applied on different data sets, an awkward task. However, a recent contribution in the literature (see [11] and its subsequent development in [3]) proposed a general functional equivalence to describe any data masking method. 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, the authors established that all masking methods can

Please cite this article as: N. 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 (2017), https://doi.org/10.1016/j.ins.2017.10.048