Information Sciences 337–338 (2016) 11–24

Contents lists available at ScienceDirect

Information Sciences

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

New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users

Josep Domingo-Ferrer[a,*], Krishnamurty Muralidhar[b]

[a] UNESCO Chair in Data Privacy, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia

[b] Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman, OK 73019-4007, USA

ARTICLE INFO

Article history:
Received 12 May 2015
Accepted 16 December 2015
Available online 24 December 2015

Keywords:
Data anonymization
Statistical disclosure control
Permutation paradigm
Subject-verifiability
Intruder model
Transparency to users

ABSTRACT

There are currently two approaches to anonymization: "utility first" (use an anonymization method with suitable utility features, then empirically evaluate the disclosure risk and, if necessary, reduce the risk by possibly sacrificing some utility) or "privacy first" (enforce a target privacy level via a privacy model, *e.g.*, *k*-anonymity or ε-differential privacy, without regard to utility). To get formal privacy guarantees, the second approach must be followed, but then data releases with no utility guarantees are obtained. Also, in general it is unclear how verifiable is anonymization by the data subject (how safely released is the record she has contributed?), what type of intruder is being considered (what does he know and want?) and how transparent is anonymization towards the data user (what is the user told about methods and parameters used?).

We show that, using a generally applicable reverse mapping transformation, any anonymization for microdata can be viewed as a permutation plus (perhaps) a small amount of noise; permutation is thus shown to be the essential principle underlying any anonymization of microdata, which allows giving simple utility and privacy metrics. From this permutation paradigm, a new privacy model naturally follows, which we call (**d**, **v**, *f*)-permuted privacy. The privacy ensured by this method can be verified by each subject contributing an original record (subject-verifiability) and also at the data set level by the data protector. We then proceed to define a maximum-knowledge intruder model, which we argue should be the one considered in anonymization. Finally, we make the case for anonymization transparent to the data user, that is, compliant with Kerckhoff's assumption (only the randomness used, if any, must stay secret).

1. Introduction

In the information society, public administrations and enterprises are increasingly collecting, exchanging and releasing large amounts of sensitive and heterogeneous information on individual subjects. Typically, a small fraction of these data is made available to the general public (open data) for the purposes of improving transparency, planning, business opportunities and general well-being. Other data sets are released only to scientists for research purposes, or exchanged among companies [4].

* Corresponding author. Tel.:+34977558109.
E-mail addresses: josep.domingo@urv.cat (J. Domingo-Ferrer), krishm@ou.edu (K. Muralidhar).

http://dx.doi.org/10.1016/j.ins.2015.12.014
0020-0255/

Privacy is a fundamental right included in Article 12 of the Universal Declaration of Human Rights. However, if privacy is understood as seclusion [30], it is hardly compatible with the information society and with current pervasive data collection. A more realistic notion of privacy in our time is informational self-determination. This right was mentioned for the first time in a German constitutional ruling dated 15 December 1983 as "the capacity of the individual to determine in principle the disclosure and use of his/her personal data" and it also underlies the classical privacy definition by Westin [31].

Privacy legislation in most developed countries forbids releasing and/or exchanging data that are linkable to individual subjects (re-identification disclosure) or allow inferences on individual subjects (attribute disclosure). Hence, in order to forestall any disclosure on individual subjects, data that are intended for release and/or exchange should first undergo a process of anonymization, also known as sanitization or statistical disclosure control (*e.g.*, see [14] for a reference work).

Statistical disclosure control (SDC) takes care of respondent/subject privacy by anonymizing three types of outputs: tabular data, interactive databases and microdata files. Microdata files consist of records each of which contains data about one individual subject (person, enterprise, etc.) and the other two types of output can be derived from microdata. Hence, we will focus on microdata. The usual setting in microdata SDC is for a data protector (often the same entity that owns and releases the data) to hold the original data set (with the original responses by the subjects) and modify it to reduce the disclosure risk. There are two approaches to control the disclosure risk in SDC:

- *Utility first*. An anonymization method with a heuristic parameter choice and with suitable utility preservation properties[1] is run on the microdata set and, after that, the risk of disclosure is measured. For instance, the risk of re-identification can be estimated empirically by attempting record linkage between the original and the anonymized data sets (see [29]), or analytically by using generic measures (*e.g.*, [15]) or measures tailored to a specific anonymization method (*e.g.*, [10] for sampling). If the extant risk is deemed too high, the anonymization method must be re-run with more privacy-stringent parameters and probably with more utility sacrifice.
- *Privacy first*. In this case, a privacy model is enforced with a parameter that guarantees an upper bound on the re-identification disclosure risk and perhaps also on the attribute disclosure risk. Model enforcement is achieved by using a model-specific anonymization method with parameters that derive from the model parameters. Well-known privacy models include ε-differential privacy [8], ε-indistinguishability [9], k-anonymity [24] and the extensions of the latter taking care of attribute disclosure, like l-diversity [18], t-closeness [16], (n, t)-closeness [17], crowd-blending privacy [13] and others. If the utility of the resulting anonymized data is too low, then the privacy model in use should be enforced with a less strict privacy parameter or even replaced by a different privacy model.

1.1. Diversity of anonymization principles

Anonymization methods for microdata rely on a diversity of principles, and this makes it difficult to analytically compare their utility and data protection properties [7]; this is why one usually resorts to empirical comparisons [5]. A first high-level distinction is between data masking and synthetic data generation. Masking generates a modified version **Y** of the original data microdata set **X**, and it can be perturbative masking (**Y** is a perturbed version of the original microdata set **X**) or non-perturbative masking (**Y** is obtained from **X** by partial suppressions or reduction of detail, yet the data in **Y** are still true). Synthetic data are artificial (*i*.e. simulated) data **Y** that preserve some preselected properties of the original data **X**. The vast majority of anonymization methods are *g*lobal methods, in that a data protector with access to the full original data set applies the method and obtains the anonymized data set. There exist, however, *l*ocal perturbation methods, in which the subjects do not need to trust anyone and can anonymize their own data (*e.g.*, [22,26]).

1.2. Shortcomings related to subjects, intruders and users

We argue that current anonymization practice does not take the informational self-determination of the subject into account. Since in most cases the data releaser is held legally responsible for the anonymization (for example, this happens in official statistics), the releaser favors global anonymization methods, where he can make all choices (methods, parameters, privacy and utility levels, etc.). When supplying their data, the subjects must hope there will be a data protector who will adequately protect their privacy in case of release. Whereas this hope may be reasonable for government surveys, it may be less so for private surveys (customer satisfaction surveys, loyalty program questionnaires, social network profiles, etc.). Indeed, a lot of privately collected data sets end up in the hands of data brokers [11], who trade with them with little or no anonymization. Hence, there is a fundamental mismatch between the kind of subject privacy (if any) offered by data releasers/protectors and privacy understood as informational self-determination.

The intruder model is also a thorny issue in anonymization. In the utility-first approach and in privacy models belonging to the k-anonymity family, restrictive assumptions are made on the amount of background knowledge available to the intruder for re-identification. Assuming that a generic intruder knows this but not that is often rather arbitrary. In the ε-differential privacy model, no restrictions are placed on the intruder's knowledge; the downside is that, to protect against

[1] It is very difficult, if not impossible, to assess utility preservation for all potential analyses that can be performed on the data. Hence, by utility preservation we mean preservation of some preselected target statistics (for example means, variances, correlations, classifications or even some model fitted to the original data that should be preserved by the anonymized data).