Data & Knowledge Engineering 100 (2015) 78–93

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Revisiting distance-based record linkage for privacy-preserving release of statistical datasets

Javier Herranz [a], Jordi Nin [b,*], Pablo Rodríguez [b], Tamir Tassa [c]

[a] Dept. de Matemàtica Aplicada IV, Universitat Politècnica de Catalunya, Barcelona, Spain
[b] Dept. d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona, Spain
[c] Department of Mathematics and Computer Science, The Open University, Ra'anana, Israel

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 5 July 2015
Accepted 7 July 2015
Available online 17 July 2015

Keywords:
Statistical Disclosure Control
Privacy measures
Distance-based record linkage

ABSTRACT

Statistical Disclosure Control (SDC, for short) studies the problem of privacy-preserving data publishing in cases where the data is expected to be used for statistical analysis. An original dataset T containing sensitive information is transformed into a sanitized version T' which is released to the public. Both utility and privacy aspects are very important in this setting. For utility, T' must allow data miners or statisticians to obtain similar results to those which would have been obtained from the original dataset T. For privacy, T' must significantly reduce the ability of an adversary to infer sensitive information on the data subjects in T.

One of the main a-posteriori measures that the SDC community has considered up to now when analyzing the privacy offered by a given protection method is the Distance-Based Record Linkage (DBRL) risk measure. In this work, we argue that the classical DBRL risk measure is insufficient. For this reason, we introduce the novel *Global Distance-Based Record Linkage* (GDBRL) risk measure. We claim that this new measure must be evaluated alongside the classical DBRL measure in order to better assess the risk in publishing T' instead of T. After that, we describe how this new measure can be computed by the data owner and discuss the scalability of those computations.

We conclude by extensive experimentation where we compare the risk assessments offered by our novel measure as well as by the classical one, using well-known SDC protection methods. Those experiments validate our hypothesis that the GDBRL risk measure issues, in many cases, higher risk assessments than the classical DBRL measure. In other words, relying solely on the classical DBRL measure for risk assessment might be misleading, as the true risk may be in fact higher. Hence, we strongly recommend that the SDC community considers the new GDBRL risk measure as an additional measure when analyzing the privacy offered by SDC protection algorithms.

1. Introduction

Nowadays, a huge amount of (digital) data is collected, processed, stored and eventually released to the public, in order to be used for different purposes. Sometimes, parts of the data contain sensitive information on individual users, and a careless dissemination of it could be inconsistent with current privacy laws. Therefore, data owners are required to protect the collected information before granting third parties access to it.

Studies in various areas of computer science have been dedicated to this problem in the last years. For instance, the *data mining* area mainly considers the interactive setting: in that setting, the data owner collects and stores the data T; external entities may

* Corresponding author. Tel.: +34 93 401 6995; fax: +34 93 401 7055.
E-mail addresses: jherranz@ma4.upc.edu (J. Herranz), nin@ac.upc.edu (J. Nin), pablor@ac.upc.edu (P. Rodríguez), tamir_tassa@yahoo.com (T. Tassa).

http://dx.doi.org/10.1016/j.datak.2015.07.009
0169-023X/

then submit queries $f(\cdot)$ on the data, to which the data owner replies with an approximate answer $y \approx f(T)$ that, on one hand, should be sufficiently close to the true answer $f(T)$, and, on the other hand, should not leak any significant information on the sensitive values contained in T. This problem led to the appearance of *privacy-preserving data mining* [3,37].

A different scenario is that of *data publishing* [17], where the data owner collects and stores some original (and sensitive) data T and then releases a modified or perturbed version T' of it. The release is made independently of the queries that could be submitted later on by external entities. The particular case in which the expected information on T that is of interest to external users is of statistical nature (means, averages, variances, correlations, etc.) has received a lot of attention and has led to the appearance of an independent area: *Statistical Disclosure Control* (SDC, for short) [13,58]. The goal of SDC is to design and analyze different methods to protect a dataset T in such a way that: (1) the released version T' allows external entities to obtain relatively accurate statistical information on T, and (2) the released dataset T' does not introduce privacy threats for the confidential information contained in T.

The SDC community proposed in the last decade many protection methods like noise addition, rank swapping, microaggregation and others. It also considered different ways to analyze both the utility level and the privacy level offered by such methods. Regarding utility, the current measures of (probabilistic) information loss all follow the same approach of measuring the difference between computing some statistical functions on the original data T and on the released data T'. Those measures are well accepted as they capture the utility of the published data for the purposes of statistical analysis. In this paper we focus on the second aspect by which SDC methods are evaluated, i.e. the privacy which they offer.

Defining good privacy measures for SDC methods is not as simple as defining utility measures, due to their dependence on the adversarial model. The first thing that needs to be done is to define both the *goals* and the *resources* of the attacker who is trying to break the privacy barrier. Regarding the goals, the SDC community has considered two possibilities [13]: first, an *interval disclosure* attacker may try to find a good approximation for some sensitive value in T; second, a *link disclosure* attacker may try to link a perturbed record of T' with an original (non-perturbed) record obtained from another source. An interval disclosure attacker is not assumed to have additional resources (other than T'). However, when considering link disclosure attackers, one has to define what external resources are available to them. As it happens in cryptography, the most recommendable option (in order to ensure privacy even in the worst case) is to assume that the attacker has obtained some information on *all* original records in T, and then he uses this information in order to infer links between protected records in T' and original records in T. (See for example [20,35,50,51,56,61,62].) The goal of the attacker is then to infer correct links between the records in T (which typically include non-sensitive information, called quasi-identifiers, such as age, location, profession etc.) and the records in T' (that may include additional sensitive information such as medical or financial data) in order to reveal sensitive information on the data subjects. A link disclosure measure can be defined as the percentage of correct links that the attacker may infer between original and protected records.

But then a new problem emerges: what is the best strategy for an attacker to find correct links? Different attackers could use very different strategies, and it is impossible to take all such strategies into account when defining a link disclosure measure. One of the linkage strategies for the attacker which has been widely adopted by the SDC community up to now is the so-called *distance-based record linkage*. A distance-based record linkage strategy finds, for every protected record $\mathbf{v}'_j \in T'$, an original record $\mathbf{v}_i \in T$ which minimizes the distance to $\mathbf{v}'_j$, for some pre-specified distance function (for instance, the Euclidean distance if all attributes in T are numerical). The pair $(\mathbf{v}_i, \mathbf{v}'_j)$ is then added to the list of links, and then the number of correct links, divided by $|T|$, gives a distance-based record linkage (DBRL for short) measure of disclosure risk. This is the only distance-based link disclosure measure that is typically considered in the SDC area when analyzing and comparing SDC protection methods.

1.1. Our contributions

The main contribution of this work is in observing that the classical DBRL disclosure measure is insufficient in order to analyze the privacy offered by SDC protection methods. The reason is that, when computing the DBRL measure, different protected records in T' may be linked to the same original record in T. For example, it is possible that some record $\mathbf{v}$ in T will be, at the same time, the closest original record to both $\mathbf{v}'_{i_1}$ and $\mathbf{v}'_{i_2}$ in T'. However, the record $\mathbf{v} \in T$ has in T' only one true perturbed image (which could be $\mathbf{v}'_{i_1}$ or $\mathbf{v}'_{i_2}$ or even another record in T'). Namely, the true global linkage is a bijection (or a perfect matching) between the records of T and those of T': each original record in T has exactly one protected image in T'. Hence, a clever attacker may try more accurate strategies to find links between records of T and records of T'. For instance, if an attacker runs the classical DBRL process and observes that more than one protected record is linked to the same original record, he would choose different candidates for the correct links. Therefore, the classical DBRL-based privacy definition for SDC methods has to be revisited.

We would like to point out that the fact that an attacker may use his background knowledge that the correct record linkage is a bijection was already observed in [35] and was implemented in RELAIS, a software for record linkage that was developed at the Italian National Statistical Office. It was also used in the context of record linkage; in that context it is sometimes known as the exclusivity constraint [18]. However, and maybe even surprisingly, this type of constraint has never been taken into account by the SDC community for the purpose of risk assessment.

We make a first and important step in this direction by introducing the *Global Distance-Based Record Linkage* (GDBRL) risk measure. Namely, we propose a new privacy measure that offers a more careful assessment of the risk of link disclosure for existing (or future) SDC protection methods. The new measure is also based on distances between original and protected records, but it takes into account the fact that the true linkage between the records in T and the records in T' is a perfect matching. After formally defining the new measure and describing algorithms to compute it, we ran experiments with datasets of different sizes, which are protected with different parameterizations of several SDC methods, in order to compare the values obtained by the classical DBRL measure and the new measure. The obtained results show that the new linkage strategy (which could be launched by attackers that have a global