A methodology for GDPR compliant data processing

Domenico Desiato and supervised by Genoveffa Tortora

Department of Computer Science, University of Salerno,
via Giovanni Paolo II n.132, 84084 Fisciano (SA), Italy
ddesiato@unisa.it

Abstract. Nowadays new laws and regulations to prevent the privacy of users have been proposed. For instance, the General Data Protection Regulation (GDPR) is taking effect in Europe, requiring organizations to define privacy policies complying with the preferences of their users. One way to abide by GDPR is to obscure sensitive data. However, in order not to limit the usage of data, it is vital to limit the amount of data to be obscured. To this end, we propose a methodology exploiting relaxed functional dependencies (RFDs) to automatically identify attributes from which sensitive values can be derived. The methodology prescribes to partially encrypt database values causing data privacy threats, identified through the automatically discovered RFDs.

Keywords: Data privacy, Anonymity, Data management

1 Introduction

When a user provides personal data to use a services on the web, s/he will no longer own them, rather they became property of the organization running the services. To this end, the European Community has issued the *General Data Protection Regulation (GDPR)*, in order to ensure the protection of personal user data while they are processed by organizations.

Standard privacy prevention techniques, such as cryptography and anonymity, could lead to the impossibility of using the data, even if part of them do not represent sensitive data. For this reason, it is necessary to detect the data to be considered sensitive, and those that would not affect user's privacy.

In this paper we present a new methodology that analyzes data correlations detected by means of relaxed functional dependencies RFDs [2], aiming to identify potentially sensitive data that could break privacy preservation. In particular, the proposed methodology aims to: (1) classify the data potentially yielding violations users' anonymity, and (2) enhance privacy prevention, by determining whether data declared as sensitive could be implied by identifying data that could imply the values of sensitive data.

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy. Copyright held by the author(s).

The paper is organized as follows. In Section 2 we provide a formalization of the privacy prevention problem, based on which the proposed methodology is described Section 3. In Section 4 we present the results of several experiments in order to validate the proposed methodology. Finally, conclusions and future research directions are discussed in Section 5.

2 Problem description

The two main concerns in data privacy are: *anonymity* and *information confidentiality*. Anonymity can be intended as non-identifiability. Thus, organizations must prevent the possibility to associate data to legitimate owners when letting third parts access them [4]. To formalize the concept of *anonymity*, we define the concept of *anonimity-violating attribute set*.

Anonimity-violating attribute set. Given a relation schema R containing user personal data, an attribute set $X = \{X_1, \ldots, X_k\}$, $X \subseteq attr(R)$, and a relation instance r of R, X represents an *anonimity-violating* attribute set if and only if it permits to *identify* data tuples in r (we denote this set by X_δ).

A relation R preserves the anonymity if and only if R does not contain an anonymity-violating attribute set X_δ. For instance, if a third-part knows an identifier value, then s/he is able to identify a user with a certainty degree of 100%. However, in order to limit third-part's power, we must also deny access to attributes enabling user's identification with a high certainty degree, even if less than 100%.

Information confidentiality is more a general concept. Here, the user would protect data s/he considers as *sensitive*. In this case, starting from a set of user specified sensitive data, we need to detect attributes from which it is possible to derive them. To formalize the concept of *information confidentiality*, we introduce the concept of *confidentiality-violating attribute set*.

Confidentiality-violating attribute set. Given a relation schema R containing user personal data, a relation instance r of R, and two attribute sets $X, Y \subseteq attr(R)$, where $Y = \{Y_1, \ldots, Y_h\}$ is the set of user specified sensitive attributes by user, then X represents a confidentiality-violating attribute set, if and only if it is not a key, but it determines at least one Y_i, one $Y_i \in Y$ (we denote this set by X_ς).

A relation R preserves the information confidentiality if and only if (i) it does not contain user specified sensitive attributes or (ii) they are obscured and R does not contain *confidentiality-violating* attribute sets X_ς. To this end, we use the concept of *functional determination* in order to exclude the possibility to derive values of attributes declared as sensitive. Thus, given a sensitive attribute A, if a third-part knows values of attributes determining those of A, then s/he could be able to discover values of A with a certainty degree of 100%, and with a maximum accuracy degree. However, it would be useful to limit third-part's power not only by decreasing the certainty degree (as defined for anonymity violations), but also by excluding the possibility to use values that are similar to those determining sensitive ones, i.e. by considering similarity-based matches.

For this reason, in this case we can identify a confidentiality violating attribute set X_ζ by using: (i) the above defined measure Ψ and a threshold ε, and (ii) a set of constraints Φ containing similarity-based matching predicates.

3 Methodology

We propose a methodology that guarantees user privacy for both *anonymity* and *information confidentiality* while permitting to continue use data. It exploits Relaxed Functional Dependencies (RFDs) [2], which enable us to detect sensitiveness of data, and to reduce the encryption processes only to them.

RFDs extend Functional Dependencies (FDs) by relaxing some constraints of their definition. In particular, they might relax on the *attribute comparison* method, and or on the fact that the dependency must be satisfied by the entire database. Relaxing on the attribute comparison method means adopting an approximate tuple comparison operator, instead of the "equality" operator. In order to define the type of attribute comparison used within an RFD, we use the concept of *constraint.* Instead, a dependency holding for "almost" all tuples or for a "subset" of them is said to relax on the extent. In this case, a *coverage measure* or a *condition* is specified to quantify the subset of tuples on which the RFD holds.

More formally, the following RFD

$$X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} Y_{\Phi_2} \tag{1}$$

holds on a relation instance r of R iff: $\forall\ (t_1, t_2) \in r$, if $t_1[X]$ and $t_2[X]$ agree with the constraints specified by Φ_1, then $t_1[Y]$ and $t_2[Y]$ agree with the constraints specified by Φ_2 with a degree of certainty (measured by Ψ) greater than ε.

Our methodology exploits RFDs discovered from data to identify sensitive data, using block ciphers to encrypt a minimal set of attributes among those containing sensitive data. Then, ranking techniques are applied to discovered RFDs, in order to detect a minimal set of attributes to be encrypted to guarantee *anonymity* and *information confidentiality.*

Anonimity. Given a relation R, we need to identify all of its sets X_δ, defining a way to make them no longer accessible on R. Formally, to identify such set of attributes X_δ we map the concept of anonymity to that of *key dependency.* In particular, since a set X_δ identifies a user, it will also be the Left-Hand-Side (LHS) of a key dependency, to preserve the anonymity of R we need to identify the minimum attribute set $Z \subseteq attr(R)$, such that by obscuring all attributes in Z from R, then no anonymity violation can be found. To automatically obtain Z we must use a metric to rank RFDs discovered from data. We defined two simple and effective ranking metrics, but they do not always guarantee the minimality of the attribute set to be encrypted in order to satisfy privacy requirements, due to the fact that this problem can be reduced to the Minimum Feedback Vertex Set [3] that is NP-Complete.

Information confidentiality. Given a relation R, we need to identify all the confidentiality violating attribute sets X_ζ in R, and define a way through which

X_ζ is no longer accessible on R. Formally, to identify a set X_ζ, we map the concept of *information confidentiality* to RFDs relaxing on the extent by a coverage measure, and to attribute comparison by means of similarity constraints. The latter represents the required accuracy degree.

Given the set of all X_ζs in R, to preserve the confidentiality of R we need to identify a minimal set of attributes $Z \subseteq \mathit{attr}(R)$, such that by obscuring Z from R, no more confidentiality violation can be found.

The cryptographic technique that we use in the proposed methodology is *block cipher* [5]. In particular, given the set of "sensitive" attributes calculated for the i-th user (according to the user choices and the application of RFDs), denoted by $X^i = X^i_1, \ldots X^i_m$, we encrypt each X^i with a different key. The user's key permits to decrypt his/her set of sensitive data.

4 Evaluation

We validated the proposed methodology, by conducting experiments on six datasets derived from the real-world datasets *Customers*, *Cancer*, *Job*, *Votes*, *WholeSale* and *Echocardiogram*, available from the UCI machine learning repository, to which we added same personal data. They show that the number of attributes to be encrypted is in general small, and that the proposed methodology can effectively help to detect anonymity and/or confidentiality threats.

5 Conclusions and Future Work

We have proposed a new methodology to automatically identify and partially encrypt "sensitive" data in order to guarantee anonymity and information confidentiality. It can help organizations comply with the GDPR privacy prevention regulations. The identification procedure exploits automatically discovered RFDs [1], in order to derive the minimal set of data to be encrypted.

In the future, we would like to apply this methodology in the context of data manipulation processes that can potentially break data privacy, such as data integration [2], and schema evolution.

References

1. Caruccio, L., Deufemia, V., Polese, G.: On the discovery of relaxed functional dependencies. In: IDEAS. pp. 53–61 (2016)
2. Caruccio, L., Deufemia, V., Polese, G.: Relaxed functional dependencies - A survey of approaches. IEEE TKDE **28**(1), 147–165 (2016)
3. Fomin, F.V., Gaspers, S., Pyatkin, A.V., Razgon, I.: On the minimum feedback vertex set problem: Exact and enumeration algorithms. Algorithmica **52**(2), 293–307 (2008)
4. Mohammed, N., Fung, B., Hung, P.C., Lee, C.K.: Centralized and distributed anonymization for high-dimensional healthcare data. ACM TKDD **4**(4), 18 (2010)
5. Stallings, W.: The offset codebook (ocb) block cipher mode of operation for authenticated encryption. Cryptologia **42**(2), 135–145 (2018)

https://archive.ics.uci.edu/ml/index.php