IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 17, 2020, accepted November 4, 2020, date of publication November 9, 2020, date of current version November 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036916

GDPR Compliant Information Confidentiality Preservation in Big Data Processing

LOREDANA CARUCCIO, DOMENICO DESIATO, GIUSEPPE POLESE, (Member, IEEE), AND GENOVEFFA TORTORA, (Senior Member, IEEE)
Department of Computer Science, University of Salerno, 84084 Fisciano, Italy
Corresponding author: Loredana Caruccio (lcaruccio@unisa.it)

ABSTRACT Nowadays, new laws and regulations, such as the European General Data Protection Regulation (GDPR), require companies to define privacy policies complying with the preferences of their users. The regulation prescribes expensive penalties for those companies causing the disclosure of sensitive data of their users, even if this occurs accidentally. Thus, it is necessary to devise methods supporting companies in the identification of privacy threats during advanced data manipulation activities. To this end, in this paper, we propose a methodology exploiting relaxed functional dependencies (RFDs) to automatically identify data that could imply the values of sensitive ones, which permits to increase the confidentiality of a dataset while reducing the number of values to be obscured. An experimental evaluation demonstrates the effectiveness of the proposed methodology in increasing compliance to the GDPR data privacy, while reducing the set of values to be partially masked, hence enhancing data usage.

INDEX TERMS Data privacy, confidentiality, data dependencies.

I. INTRODUCTION

Nowadays, thanks to the digitalization of business processes and public administrations, many significant data collections are available. Users are direct suppliers of data when publishing content on social networks. However, when using a service on the web, users must often provide their data, which will become property of the company running the service. To this end, users are becoming aware of the privacy issues related to the management of their data, and governments are defining new laws and regulations to ensure the protection of users' personal data. At the same time, there exists the necessity not to limit the processing of data by companies and other public institutions.

The European Community has issued a new regulation, named *General Data Protection Regulation (GDPR)*, to guarantee greater control of users over the data they provide. To this end, companies have many difficulties in determining how they can use the data to avoid legal issues related to data privacy violations. Standard privacy preservation techniques, such as cryptography and obfuscation, could lead to the impossibility of using the data, even if some of them are not sensitive. Thus, it is necessary to distinguish between sensitive and non-sensitive data effectively. In general, the management of sensitive data is tackled into different application areas, such as IoT applications [1], Smart Grids [2], Social Networks [3], and so on. As an example, the necessity to manage sensitive data arises when hospitals adopt sensor networks to monitor patients and, in particular, disabled patients.

Guaranteeing privacy preservation becomes an even more complex problem in the presence of significant data processing operations like for instance data integration [4] and record linkage [5]. In fact, such processes could yield privacy violations when the compared data sources contain sensitive data. In fact, even if sensitive data are obscured to meet users' privacy requirements, such data processing operations could introduce new sensitive data, due to the generation of new identification patterns.

In this paper, we present a new methodology that analyses data correlations expressed in terms of relaxed functional dependencies (RFDs) [6], aiming to identify new potentially sensitive data upon significant big data operations, which could break privacy rules. In particular, the methodology exploits recent algorithms to automatically discover RFDs from data, which can also be applied to identify cross-correlations among data sources that could yield data privacy issues [6]. The main goal is to reduce the number of values

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

to be obscured while increasing the confidentiality preservation level of the dataset according to users' requirements, aiming to increase data usage. More specifically, the proposed methodology prescribes to partially encrypt sensitive data to increase the preservation of their confidentiality. This guarantees the possibility to use unobscured data within data analytics processes, without risking to jeopardise users' privacy. Concerning the GDPR, the proposed methodology represents a useful mean for guaranteeing the recital 71, regulating user profiling activities, as described in the GDPR [7]. The methodology has been experimentally validated on data derived from three public datasets, namely *CreditClient*,[1] *Health*,[2] and *London*.[3]

The paper is organized as follows. In Section II we discuss related works. Section III describes the *GDPR* and the basic concepts underlying RFDs, whereas Section IV illustrates a formalization of the privacy preservation problem. In Section V we describe the proposed methodology, whose general process is provided in Section VI. Section VII illustrates the results of several experiments that we performed to evaluate the effectiveness of the proposed methodology. Finally, conclusions and future research directions are discussed in Section VIII.

II. RELATED WORK

Nowadays, users are more aware of privacy preservation issues. To this end, GDPR has introduced a new way of thinking about privacy preservation, but it is essential to enforce such regulation without preventing the execution of important data analytics tasks. This has motivated several GDPR-related studies, most of which aim to analyse the impact of GDPR on specific application domains. In particular, a study on the impact of GDPR from the London Chamber of Commerce and Industry revealed that in 2018 companies were still not prepared to abide by the privacy preservation issues prescribed by GDPR [8]. Moreover, due to the necessity to motivate users to approach GDPR-based privacy issues, some recent works deal with the problem of making friendlier the management of such issues [9], [10]. In [9], authors analysed the state of the art of usability design for privacy notifications, by highlighting how approaches defined in the literature correlate to GDPR recitals, summarising them in terms of guidelines. Instead, in [10], a tool named privacyTracker is presented, which aims to support basic GDPR principles, including data traceability, allowing a user to get a cryptographically verifiable snapshot of his/her data trails.

We focus on how privacy preservation issues concerning Big Data context have been addressed in the literature [11]. In general, the main techniques used to manage sensitive information are: (i) *access control*, and (ii) *cryptography*. Access control requires the definition of data access policies by establishing whether the user has the authorisation to perform certain actions on the given data [12], [13]. Among the privacy preservation techniques exploiting access control, we find [14], which proposes a fine-grained approach for data stored in the cloud. It aims to simplify the users' workload in the encryption of their data before loading them on the cloud. Instead, in [15] biometry has been used to guarantee identification and authentication into the Naked environment. It aims to provide health services in a smart hospital environment, without using specific gadgets for accessing the services. Moreover, medical sensors embedded in the environment provide users with the required digital services by employing a biometric-based authentication schema. The last two approaches represent prominent examples to understand the usefulness of access control methodologies for privacy preservation in emerging and/or complex contexts, such as healthcare [15], and the cloud context in general [14].

Concerning cryptographic techniques useful to manage privacy requirements, several methodologies can be found in the literature, such as symmetric cryptography, hash functions, pseudo-anonymity, obscuration, and so on. One of the most recent proposals is described in [16], where a new approach named DBMask is proposed. It represents a novel solution supporting fine-grained, cryptographically enforced, access control policies, including column-, row-, and cell-level access control when evaluating queries on encrypted data.

Another way to categorise privacy preservation methodologies is to distinguish how they work to preserve data privacy, i.e. syntactically or semantically. In particular, there exist several approaches for guaranteeing syntactic data privacy. One of the most popular is the k-anonymity model [17], which defines a data record as k-anonymous whenever it is indistinguishable in its identifying information from at least k specific records or entities. Moreover, in [18] and [19] two extensions of k-anonymity have been defined, aiming to limit possible attribute disclosures. Furthermore, in [20] authors propose a two-party framework that can be easily employed to design a secure protocol, aiming to compute k-anonymous data from two vertically partitioned data sources.

Concerning privacy preservation approaches for semantic models, in [21] authors present p-sensitivity, aiming to mask user locations by taking into account query diversity and semantic information. The authors exploit a PE-Tree structure and search algorithms for implementing the p-sensitivity model, aiming to find the optimal p-sensitivity privatization in the tree.

Confidentiality issues have been studied in-depth in several application domains. In the data mining context, the Framework for Accuracy in Privacy-Preserving mining (FRAPP) represents a generalised model for random perturbation-based methods, operating on categorical data under strict privacy constraints [22]. Instead, guaranteeing privacy preservation in the context of big data processing tasks adds further complexity, since there might be several complex operations potentially yielding privacy threats. A typical example is

[1] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
[2] https://healthdata.gov/dataset/health-care-provider-credential-data
[3] https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset

represented by data integration, since data integrated from several data sources might partially or totally imply obscured data [4], [5]. An essential pioneering work in this context was carried out by Dalenius T. in 1986 [23]. The author shows how it is possible to identify unique records in a dataset by merely sorting them. This simple idea turns out to be highly effective because it permits to discover unique records. A more recent proposal is the general framework presented in [24], which concerns normalised measures to practically evaluate and compare privacy-preserving record linkage (PPRL) solutions. Furthermore, the authors aimed at comparing the state of the art PPRL techniques, by considering widely used numerical measures, such as scalability, linkage quality, and privacy. Instead, Schnell *et al.* propose the Bloom Filter-based protocol, which aims to guarantee privacy preserving record linkage [25]. It uses encrypted identifiers, and similarity computations of Bloom filters with Hash Message Authentication Codes (HMACs) on a q-grams similarity function [26]. Bit Vectors (BV) is another approach for representing numerical data values in privacy-preserving record linkage [27]. It represents an accurate distance preserving encoding schema for embedding numerical values into a privatization space, in a way that preserves the initial distances. In particular, in order to compute hash values, BV uses extremely simple and computationally cheap operations, instead of expensive cryptographic hash functions. Finally, a Privacy-Preserving Probabilistic Record Linkage (P3RL) methodology has been proposed in [28]. It facilitates the linkage of existing datasets in health-related research settings, and provides a different solution w.r.t. the classical cryptographic methods. Among all the analysed works in the context of privacy-preserving record linkage, the last three represent possible solutions that can be used to address the problem of record linkage. Instead, the first two can be analysed to deepen the record linkage problem in the data integration context.

With respect to all the proposals analysed before, our methodology preserves the information confidentiality by reducing the amount of data to be encrypted, hence increasing data usage. It exploits RFDs to identify data correlations that could potentially violate the privacy according to the user's requirements. As a consequence, the proposed methodology permits to partially encrypt data according to the different privacy preservation requirements that a user could specify.

III. BACKGROUND

Since the proposed methodology exploits RFDs to pursue GDPR compliant privacy preservation, in the following sections we provide details on GDPR and RFDs.

A. GENERAL DATA PROTECTION REGULATION

The General Data Protection Regulation (GDPR) prescribes how companies must process and manage private data [29] of their users, aiming to offer significant improvements to the regulatory environment of companies and institutions. In particular, GDPR establishes a uniform framework for data protection legislation across nations belonging to the European Community, without having to comply with the regulations of the single governments. This represents a significant advantage for companies operating across multiple countries of the European Community. Furthermore, even companies located outside the European Community must abide by the GDPR if they manage data of European users.

GDPR classifies as *personal data* any information related to individuals, without prescribing the usage of specific methodologies/technologies. Even if personal data are obscured and/or partially encrypted, the organization managing them still incurs violations if it is possible to disclose users' sensitive data upon some data processing activities, such as data integration, entity resolution, and so on. The central concept underlying GDPR concerns the ''user agreement'', i.e. the specification on how users' data should be processed through an explicit declaration, which should be freely given, specifically informed, and unambiguous.

GDPR prescribes the following two activities: (i) the adoption of a privacy preservation methodology, and (ii) the definition of default policies to preserve the privacy of any user. Thus, according to the first activity organizations need to employ a privacy preservation methodology from the design to the development of their services. Instead, the second activity prescribes the implementation of proper default methodologies/technologies to guarantee data processing in a trusted way. These prescriptions aim to provide a friendly privacy setting, by also providing the possibility to adopt default settings.

Concerning the possibility to share personal data, the GDPR is not limited to the European Economic Area (EEA),[4] since when data are transferred outside the EEA, all privacy preservation policies defined on data are transmitted along with the data themselves. Moreover, GDPR is composed of several recitals addressing the privacy preservation issues to specific activities, such as marketing, user profiling, data integration, and so on. Among all recitals defined in the GDPR, the recital 71 expresses, in a summarised way, the fact that a company performing analytical activities should use appropriate mathematical or statistical procedures, and implement technical and organizational measures necessary to ensure the privacy of data related to a physical person [7].

GDPR has become effective since May 25th, 2018. To this end, by offering the possibility to manage different policy requirements concerning single users, the proposed methodology turns out to particularly useful in pursuing GDPR compliant privacy preservation.

B. RELAXED FUNCTIONAL DEPENDENCIES

Let us recall some basic concepts of relational databases.

A relational database schema $\mathcal{R}$ is defined as a collection of relation schemas $(R_1, \ldots, R_n)$, where each R_i is defined over a set $attr(R_i)$ of attributes $(A_1, \ldots, A_m)$. Each attribute A_k has

[4]European Economic Area (EEA), includes all European Community countries, and Island, Liechtenstein, and Norway

associated a domain $dom(A_k)$, which can be finite or infinite. A relation instance (or simply a relation) r_i of R_i is a set of tuples such that for each attribute $A_k \in attr(R_i)$, $t[A_k] \in dom(A_k)$, $\forall\ t \in r_i$, where $t[A_k]$ denotes the projection of t onto A_k. A database instance r of $\mathcal{R}$ is a collection of relations $(r_1,\ldots,r_n)$, where r_i is a relation instance of R_i, for $i \in [1, n]$.

In the context of relational databases, data dependencies have been mainly used to define data integrity constraints, aiming to improve the quality of database schemas and to reduce manipulation anomalies. There are several types of data dependencies, including functional, multivalued, and join dependencies. Among these, functional dependencies (FDs) are the most commonly known, mainly due to their use in database normalization processes. Since RFDs extend FDs, let us recall the definition of FD.

Definition 1 (Functional Dependency): A functional dependency (FD) φ, denoted as $X \rightarrow Y$, between two sets of attributes $X, Y \subseteq attr(\mathcal{R})$, specifies a constraint on the possible tuples that can form a relation instance r of $\mathcal{R}$: $X \rightarrow Y$ iff for every pair of tuples (t_1, t_2) in r, if $t_1[X] = t_2[X]$, then $t_1[Y] = t_2[Y]$. The sets X and Y are also called Left-Hand-Side (LHS) and Right-Hand-Side (RHS), resp., of φ.

RFDs extend FDs by relaxing some constraints of their definition. In particular, they might relax on the *attribute comparison* method, and on the fact that the dependency must hold on the entire database.

Relaxing on the attribute comparison method means to adopt an approximate operator to compare tuples, instead of the ''equality'' operator. In order to define the type of attribute comparison method used within an RFD, we use the concept of *constraint* [30].

Definition 2 (Constraint): A constraint ϕ is a predicate evaluating whether the similarity/distance, or the order relation, between two values of an attribute A falls within a predefined interval.

Thus, a constraint depends on a similarity/distance function, or an order relation, defined on an attribute domain, plus one or more comparison operators with associated threshold values defining the feasible intervals of values.

A sample constraint ϕ defined on the attribute Address and the edit distance ED could be: $0 \leq ED(addr1, addr2) \leq \varepsilon$, where $addr1$ and $addr2$ are two address values, whereas 0 and ε are two given threshold value.

Definition 3 (Set of Constraints): Given a set of attributes $X = \{A_1, \ldots, A_k\}$, a set of constraints $\Phi = \{\phi_1, \ldots, \phi_k\}$ on them represents a collection of constraints that are applied to $\{A_1, \ldots, A_k\}$, respectively.

A functional dependency holding on ''almost'' all tuples or on a ''subset'' of them is said to relax on the extent [6]. In the case of ''almost'' all tuples, a *coverage measure* should be specified to quantify the degree of satisfiability of the RFD. Whereas, in the case of ''subset'' (*constrained domain* in the following), conditions on the attribute domains should be specified to define the subset of tuples satisfying the RFD.

Definition 4 (Coverage Measure): A coverage measure Ψ on φ, Ψ: $dom(X) \times dom(Y) \rightarrow \mathbb{R}^+$, quantifies the amount of tuple pairs in r satisfying φ.

As an example, the *confidence measure* introduced in [31] evaluates the cardinality of the greatest set of tuples $r_1 \subseteq r$ for which φ holds in r_1.

Several coverage measures can be used to define the satisfiability degree of an RFD, but usually they return a value normalized on the total number of tuples n, with n cardinality of r, so producing a value $v \in [0, 1]$. For the canonical FD, this measure evaluates to 1.

Definition 5 (Constrained Domain): Given a relation database schema $\mathcal{R}$ with attributes $\{A_1, \ldots, A_k\}$ defined on domains $\{D_1, \ldots, D_k\}$ respectively, $\mathbb{D}_1 \times \mathbb{D}_2 \times \cdots \times \mathbb{D}_k = dom(\mathcal{R})$, respectively, and let c_i be a condition on $\mathbb{D}_i$, $i = 1 \ldots k$, the constrained domain $\mathbb{D}_c$ is defined as follows

$$\mathbb{D}_c = \left\{ t \in dom(\mathcal{R}) \middle| \bigwedge_{i=1}^{k} c_i(t[A_i]) \right\}.$$

Constrained domains enable the definition of tuple ''subsets'' on which a functional dependency holds.

Then, a general definition of RFD can be given:

Definition 6 (Relaxed Functional Dependency): Let us consider a relational schema $\mathcal{R}$. An RFD ϱ on $\mathcal{R}$ is denoted by

$$\left[X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} Y_{\Phi_2} \right]_{\mathbb{D}_c} \tag{1}$$

where

- $\mathbb{D}_c$ is the constrained domain that filters the tuples on which ϱ applies;
- $X, Y \subseteq attr(\mathcal{R})$, with $X \cap Y = \emptyset$;
- Φ_1 and Φ_2 are sets of constraints on attribute sets X and Y, respectively;
- Ψ is a coverage measure defined on $\mathbb{D}_c$;
- ε is a threshold, with $0 \leq \varepsilon \leq 1$.

Given $r \subseteq \mathbb{D}_c$, a database instance r on $\mathcal{R}$ *satisfies* the RFD ϱ, denoted by $r \models \varrho$, if and only if: $\forall\ (t_1, t_2) \in r$, if Φ_1 is true for each constraint $\phi \in \Phi_1$, then *almost always* Φ_2 is true for each constraint $\phi' \in \Phi_2$. Here, *almost always* means that $\Psi(X, Y) \geq \varepsilon$.

In other words, 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.

Based on definition (1), the canonical FD $X \rightarrow Y$ can also be written as:

$$\left[X_{\mathrm{EQ}} \xrightarrow{\Psi_1} Y_{\mathrm{EQ}} \right]_{\mathbb{D}_{\mathrm{true}}} \tag{2}$$

where true is a sequence of tautologies, $\mathbb{D}_{\mathrm{true}} = dom(\mathcal{R})$, EQ is the equality constraint, and Ψ_1 represents the fact that the dependency must hold on all tuples of the instance r (i.e., $\Psi(X, Y) = 1$, and $\varepsilon = 1$).

In the following, we use RFDs having only one attribute on the RHS; this condition can always be reached employing the usual transformations of FDs.

TABLE 1. A database storing customers' information.

Name	Surname	SSN	Age	Street	Native-Country	Occupation	Sex	City	ZipCode
Katherine	Swavely	029-32-6730	35	ALOHA AVE	United-States	Employee	F	Pearl City	96782
Matthew	Costabile	475-96-3980	58	SARGENT ST	United-States	Unemployed	M	San Francisco City	94132
Jarrett	Albarado	214-20-7035	49	MARNE AVE	England	Worker	M	Newburgh	12550
Rowena	Hemeyer	481-98-9042	79	ESQUINA DR	United-States	Retired	F	San francisco	94134
Corina	Torris	490-03-6515	34	ALOHA AVE	United-States	Employee	F	Pearl City	96782
Carlotta	Bracker	659-05-8786	32	ALOHA AVE	United-States	Unemployed	F	Pearl City	96782
Zane	Bracker	678-14-8279	32	PALOS PL	United-States	Teacher	M	Illinois	60464
Joselyn	Bracker	004-03-6265	32	PALOS PL	United-States	Teacher	M	Illinois	60464
Sherry	Swavely	400-20-9834	80	ALOHA AVE	United-States	Retired	F	Pearl City	96782
Matthew	Costabile	255-73-7429	24	SARGENT ST	England	Unemployed	M	San Francisco City	94132

Example 1: Let us consider the database shown in Table 1, which represents a portion of *the census income dataset*, containing the following data of citizens: Name, Surname, SSN, Age, Address, Native-Country, Occupation, and Sex. According to this, it is likely to have the same Native-Country for costumers having the same Name and Surname thus, an FD Name, Surname → Native-Country might hold. However, the names, surnames, and countries might be stored by using different abbreviations/variations, and/or typos may have been introduced during tuple insertion operations. Thus, the following RFD might hold:

$$\left[\mathsf{Name}_{\approx}, \mathsf{Surname}_{\approx} \xrightarrow{\Psi_1} \mathsf{Native\text{-}Country}_{\approx}\right]_{\mathbb{D}_{\text{true}}}$$

where ≈ is the string similarity function. On the other hand, few cases of homonyms for the customers have to be considered. For this reason, the previous RFD should also admit exceptions. This can be modeled by introducing a different coverage measure to make the RFD relax on the extent:

$$\left[\mathsf{Name}_{\approx}, \mathsf{Surname}_{\approx} \xrightarrow{\Psi(X,Y)\geq 0.90} \mathsf{Native\text{-}Country}_{\approx}\right]_{\mathbb{D}_{\text{true}}}$$

IV. PROBLEM DESCRIPTION

According to the GDPR, companies and organizations can use sensitive data only for business application purposes, avoiding their exposure to third parties, or their transfer for commercial activities, such as user profiling. All activities affecting the confidentiality of data have to be considered as data privacy violations. To this end, we always need to pay attention to data representations that might refer to users.

Data privacy concerns several aspects, among which we focus on *Information Confidentiality* (IC). The latter is a general privacy preservation concept by which users request to preserve the confidentiality of their specific data, also referred to as *sensitive data*, aiming to protect them against unauthorized accesses [32].

In what follows, we formalize the concept of information confidentiality in the context of relational databases.

Definition 7 (Information Confidentiality): Given a relational database schema $\mathcal{R}$, defined on a set of attributes $attr(\mathcal{R} = \{A_1, \ldots A_n\})$, an instance r of it, where each tuple t over r represents a single user, and its projection $t[Y]$ onto $Y \subseteq attr(\mathcal{R})$ the data s/he defines as sensitive, then ensuring the information confidentiality for t requires that i) $t[Y]$ is masked, and ii) no subset of data $t'[Y']$ permits to disclose any value in $t[Y]$.

Starting from Definition 7, it is possible to derive the concept of *data usage* for our context, that is: a data can be used without jeopardising the privacy of any user if and only if i) it has not been declared as sensitive by its owner, and ii) it cannot be used to disclose any other data declared as sensitive.

In what follows, we formalize the *information confidentiality* problem issues in terms of attribute correlations expressed by RFDs, yielding the concept of *confidentiality-violating attribute set*.

Definition 8 (Confidentiality-Violating Attribute Set): Given a relational database schema $\mathcal{R}$, an instance r of it, and two attribute sets $X, Y \subseteq attr(\mathcal{R})$, where $Y = \{Y_1, \ldots, Y_h\}$ is the set of data defined as sensitive, then X is a confidentiality-violating attribute set if and only if it is not a key, and there exists $Y_i \in Y$ that is the RHS of an RFD holding on r and having X as LHS.

According to Definition 8, a relational database schema $\mathcal{R}$ preserves the information confidentiality if and only if: (i) $\mathcal{R}$ contains all the user-specified sensitive attributes in a masked form, and (ii) $\mathcal{R}$ does not contain confidentiality-violating attribute sets. In other words, if the user specifies a set of sensitive attributes, other than obscuring them, we also need to prevent the possibility to derive their values from other attribute values. For instance, a sensitive attribute might be derived by the LHS of an RFD φ in which it appears as RHS. In this case, we say that the LHS of φ determines the RHS. Thus, given a sensitive attribute A, knowing the values of attributes determining A, a third party could infer the values of A with high certainty and accuracy degrees according to the thresholds of φ. As a consequence, we need to identify all the confidentiality violating attribute sets, that is, all the attribute sets functionally determining sensitive attributes.

V. METHODOLOGY

From the discussion above, it is clear that the GDPR might be a serious burden, especially for big companies managing huge volumes of data concerning their customers. By referring to the scenario shown in Table 1, a solution could be to obscure all data, by means of cryptographic techniques. However, in this way a company could never use such data, even those that are not sensitive, and would have to deal

with computationally expensive encryption processes, e.g. not all the data showed in Table 1 can be considered as sensitive. Moreover, by manually specifying both sensitive data and those from which they can be derived could require a huge effort when managing big data collections. To this end, we propose a new methodology that reduces the number of attributes to be encrypted while pursuing information confidentiality, hence maximizing data usage. The methodology exploits attribute correlations, expressed in terms of Relaxed Functional Dependencies (RFDs) [33], to identify attribute sets from which sensitive data can be derived.

More specifically, the proposed methodology exploits algorithms to automatically discover RFDs from data [6], [34], together with ranking techniques to decide their application order, aiming to derive a minimal set of attributes to encrypt for pursuing information confidentiality.

Given a relational database schema $\mathcal{R}$, and an instance r of it, we need to identify the set $X_\Xi = \{X_{\zeta_1}, \ldots, X_{\zeta_n}\}$ of all confidentiality violating attribute sets X_{ζ_i} within $\mathcal{R}$, and define a way to make each of them not accessible. To this end, we consider the following types of RFDs holding on r:

$$\left[X_{\Phi_1} \xrightarrow{\Psi \geq \varepsilon} A_{\Phi_2}\right]_{\mathbb{D}_{\text{true}}} \tag{3}$$

where $A \in attr(\mathcal{R})$, and search for the LHSs of RFDs having a sensitive attribute on the RHS.

More formally, in order to preserve the confidentiality of $\mathcal{R}$ we need to identify the minimal set of attributes $Z \subseteq attr(\mathcal{R})$ such that there exists no valid RFD $X_{\zeta_i} \backslash Z \rightarrow A$, with A sensitive attribute of $\mathcal{R}$. In other words, it is necessary to invalidate all the RFDs having a user-specified sensitive attribute as RHS. The set of user-specified sensitive attributes is also named *IC-attribute set*.

In order to automatically derive the minimal set Z of attributes to be removed, we must use a heuristic. This is due to the fact that this problem is NP-complete, since the Minimum Feedback Vertex Set [35], which is the problem of finding the smallest set of vertices to be removed from an undirected cyclic graph to make it acyclic, can be reduced to it. In particular, each $X_{\zeta_i} \in X_\Xi$ can be modeled as a cycle in an undirected graph, where the vertices of the cycle are the attributes in X_{ζ_i}. Thus, given an undirected graph G with one or more cycles, the vertices of a cycle can be seen as a confidentiality-violating attribute set X_{ζ_i}. Thus, solving the problem of finding the minimal set Z defined above would also solve the MFVS one.

A. HEURISTICS

We defined three heuristics: (i) the *counting* heuristic, scoring the number of X_{ζ_i} containing a given attribute; (ii) the *weighted counting*, similar to the counting heuristic, but instead of adding a 1 for each X_{ζ_i} in which an attribute appears, it adds $1/|X_{\zeta_i}|$, which represents the weight of the attribute over X_{ζ_i}; and (iii) the *MFVS* heuristic, derived from an approximate solution for the MFVS problem, which is based on the Depth First Search (DFS) visit to approximately evaluate the number of times a node is involved in a cycle.

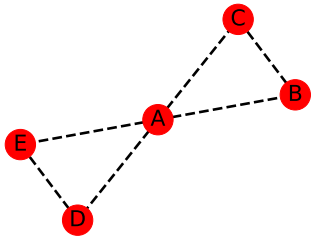


FIGURE 1. MFVS problem associated to (4) and (5).

In particular, the first two heuristics associate scores to the attributes belonging to the confidentiality-violating attribute sets in X_Ξ, eliminating them in descendant order of their score, until all the RFDs associated to X_Ξ are invalidated. Instead, as mentioned above, with the third heuristic an undirected graph is produced. In particular, the heuristic scores each node with the number of backward edges encountered during a DFS visit. More specifically, we adapted the well-known DFS visit to count backward edges. Then, the heuristic removes nodes in descendant order of their score, until no more cycles exist in the graph. For all three heuristics, a basic case is represented by RFDs with one attribute on the LHS, since we can remove it without considering any score.

As an example, let us consider the following two RFDs:

$$A\ B\ C \rightarrow F \tag{4}$$

$$A\ E\ D \rightarrow F \tag{5}$$

where F is a confidential attribute. If they are the only RFDs with F on the RHS, we must encrypt some of the attributes on their LHSs together with F, in order to guarantee the information confidentiality of F. By applying the counting heuristic defined above, attribute A has a score of 2, whereas each of the remaining attributes has a score of 1. Thus, we first remove A, which already invalidates both RFDs (4) and (5). A similar action is decided upon applying the weighted counting heuristic, since attribute A has a score of $2/3$, whereas the remaining ones have a score of $1/3$ each. Finally, the MFVS heuristic could be used to solve the MFVS problem of the graph in Figure 1, yielding the deletion of the vertex A only, since it is the node with the maximum number of backward edges derived from the DFS visit. Moreover, upon removing A, the resulting graph is acyclic. Thus, in all three cases, A will be the only attribute to be encrypted together with F.

Example 2: Let us consider the database of customers shown in Table 1, and let us suppose that a user wants to "obscure" the Occupation attribute in order to preserve his/her privacy. In this case, there are three attribute sets determining Occupation, i.e. Name, {Age, Sex}, and {Age, Street}, since they are the LHSs of all the RFDs holding on the considered relation, and having attribute Occupation as RHS. To guarantee information confidentiality, besides "obscuring" the attribute Occupation we should also "obscure" the attribute Name, since it determines

TABLE 2. A privacy-preserving database of customers' information.

Name	Surname	SSN	Age	Street	Native-Country	Occupation	Sex	City	ZipCode
Katherine	Swavely	029-32-6730	35	ALOHA AVE	United-States	Employee	F	Pearl City	96782
Matthew	Costabile	475-96-3980	58	SARGENT ST	United-States	Unemployed	M	San Francisco City	94132
Jarrett	Albarado	214-20-7035	49	MARNE AVE	England	Worker	M	Newburgh	12550
Rowena	Hemeyer	481-98-9042	79	ESQUINA DR	United-States	Retired	F	San francisco	94134
Corina	Torris	490-03-6515	34	ALOHA AVE	United-States	Employee	F	Pearl City	96782
*****	Bracker	659-05-8786	*****	ALOHA AVE	United-States	*****	F	Pearl City	96782
*****	Bracker	678-14-8279	*****	PALOS PL	United-States	*****	M	Illinois	60464
*****	Bracker	004-03-6265	*****	PALOS PL	United-States	*****	M	Illinois	60464
*****	Swavely	400-20-9834	*****	ALOHA AVE	United-States	*****	F	Pearl City	96782
*****	Costabile	255-73-7429	****	SARGENT ST	England	*****	M	San Francisco City	94132

the attributes Occupation and Age, based on the defined heuristics.

B. PARTIAL ENCRYPTION

The cryptographic technique used in the proposed methodology is *block cipher* [36]. The latter is a method using a secret key to encrypt text (to produce cipher-text). In particular, it applies encryption to blocks of data (e.g., 64 contiguous bits) rather than to one bit at a time. Formally, a block cipher is a permutation with a key that can be efficiently computed, i.e. $\mathcal{F} : \{0,1\}^n \times \{0,1\}^l \rightarrow \{0,1\}^l$ such that, given a key k and a block of data to be encrypted x

$$F_k(x) \stackrel{\text{def}}{=} F(k,x) \text{ is a permutation} \tag{6}$$

where n is the length of k, l the length of x, and $\mathcal{F}_k$, $\mathcal{F}_k{}^{-1}$ must be efficiently computed.

In particular, given the set $X^i = X^i_1, \ldots X^i_m$ of user-specified "sensitive" attributes, together with those derived through RFDs, we apply the block cipher to all X^i. It is worth to notice that each X^i is encrypted with a different secret parameter k, which is the user's secret parameter that permits to decrypt his/her sensitive data. This implies that we can have a database containing both visible and encrypted data, that is still privacy-preserving. The block cipher guarantees security with respect to the Chosen Plaintext Attacks (CPA-security) [37].

Example 3: Let us consider the database of customers shown in Table 1, and let us suppose that the last five of them required attribute Occupation to be confidential. As shown in Table 2, by applying the proposed methodology we obtain partial encryption, where the values denoted as "*****" are encrypted as explained in the previous examples.

C. OVERVIEW OF THIRD PARTIES

In what follows, we analyze the robustness of the proposed methodology by considering the power of third parties in disclosing values of attributes specified as confidential. In particular, we prove how RFDs can help identify confidentiality threats by also analyzing several critical scenarios.

One of the main properties of RFDs is *minimality* [6], which concerns both the number of attributes on their LHS and the associated similarity thresholds. For the critical scenarios analyzed below, we are only concerned with how the minimality property is related to the LHS attributes. Let r be an instance of a relational database schema $\mathcal{R}$, and φ: $X \rightarrow Y$ a minimal RFD holding on r, then for each $A \in X$, φ': $X \backslash A \rightarrow Y$ does not correspond to a RFD holding on r. In general, RFD discovery algorithms aim at finding the set of all minimal RFDs holding on a given dataset.

Before detailing the sample scenario, we first introduce the preliminaries of the third parties that we consider for our threat model. Let us suppose that a third party can access:

- the dataset structure together with metadata concerning the value distribution of each attribute;
- the set of all minimal RFDs holding on the dataset;
- the dataset partially encrypted according to the proposed methodology.

Moreover, we assume that the third party can ask an oracle all the information defined above by simply providing the name of the dataset. In particular, the value distributions enable the third party to know all possible values that an attribute can assume, whereas the set of minimal RFDs holding on the unencrypted dataset enables the third party to catch the data validating possible RFDs. Notice that, we use the term dataset also referring to the ones obtained as a result of data integration, data augmentation, or any other big data processing task.

By considering the characteristics of this threat model, we can reduce the likelihood of success for a third party to the safest scenario, i.e. a totally encrypted dataset. Thus, the likelihood of success for a third party in disclosing a target value can be reduced to a random guess on the value distribution it belongs to. In other words, even when a dataset is completely encrypted, the third party can try to disclose the target value by only choosing one of the values of its distribution. To this end, in what follows, we show that our target is to reduce the likelihood of success of the third party to a random guess on the value distribution of each IC attribute.

Example 4: Let us consider the sample dataset 1 shown in Table 3(a), for which we assume that there is only one IC attribute for the tuple t_1, e.g. attribute D. This means that the owner of t_1 requires confidentiality for the value $t_1[D]$. According to the proposed methodology, we consider RFDs implying attribute D. Thus, the only RFD to be considered among those holding on the given dataset is φ: $AB \rightarrow D$. Then, if we obscure only the value $t_1[D]$, it could still be derived from the correlation expressed by φ. In fact, by looking at tuple t_2, a third party could infer the value on $t_1[D]$

TABLE 3. A sample scenario.

(a) Sample dataset 1

	A	B	C	D
t_1	1	2	7	True
t_2	1	2	8	True
t_3	3	2	6	False
t_4	1	3	9	True

(b) Masked dataset 1

	A	B	C	D
t_1	***	2	7	***
t_2	1	2	8	True
t_3	3	2	6	False
t_4	1	3	9	True

(c) Sample dataset 2

	A	B	C	D
t_1	1	2	7	True
t_2	3	2	8	False
t_3	3	2	6	False
t_4	1	4	9	False

(d) Masked dataset 2

	A	B	C	D
t_1	***	2	7	***
t_2	3	2	8	False
t_3	3	2	6	False
t_4	1	4	9	False

through the similarity between t_1 and t_2 on the combination of values for attributes A and B.

Example 4 shows how minimal RFDs help us solve some issues concerning the third parties' derivation process. In fact, since φ: $AB \rightarrow D$ is minimal on the dataset shown in Table 3(a), then by masking $t_1[A]$ or $t_1[B]$ would guarantee that a third party could not derive the value $t_1[D]$, since both φ': $A \rightarrow D$ and φ'': $B \rightarrow D$ do not hold on the considered dataset. For this reason, by simply observing the values of A (or B), a third party could not derive the value $t_1[D]$, since φ' (φ'') is not valid.

Example 5: Starting from the scenario described in example 4, let us assume that our methodology prescribes to mask attribute A to break the attribute correlation expressed by φ: $AB \rightarrow D$, as shown in Table 3(b). To this end, the third party can only observe the free values of attribute B and consider the tuples that are similar to $t_1[B] = 2$, which means all tuples. In particular, the dataset contains the value $'True'$ for $t_2[D]$ and the value $'False'$ for $t_3[D]$. Thus, the third party can only try a random guess, which in this case is equivalent to a coin toss. Similar considerations also apply when the value $t_1[B]$ is masked and $t_1[A]$ is not.

In what follows, we analyze a borderline case of the aforesaid scenario, which is the only one jeopardising the proposed methodology to the risk of value disclosures.

Example 6: Let us consider the sample dataset shown in Table 3(c), and suppose that there is only one IC attribute for the tuple t_1, e.g. attribute D. Consequently, the only RFD to be considered, among those holding on the given dataset, is φ: $AB \rightarrow D$. Let us now suppose that the methodology prescribes to mask attribute A in order to break the attribute correlation expressed by φ, as shown in Table 3(d). If a third party knew that φ is a minimal RFD holding on the dataset, then s/he would be aware that φ': $B \rightarrow D$ did not hold on the dataset. Furthermore, since the value distribution of attribute D is limited to $\{True, False\}$, and the tuples similar to t_1 on B are t_2 and t_3, which have value *False*, a third party could exactly infer the value of $t_1[D]$, since the only violation invalidating the RFD φ' can be generated from the value $'True'$.

In general, this case can occur only when the RFD violation is caused by the attribute value declared as confidential, which has been obviously masked. However, although this borderline case occurs rarely, we need to undertake additional actions in order to guarantee the requested confidentiality.

More formally, let A be an attribute, and t be a tuple for which $t[A]$ is declared as confidential, then a third party can infer the masked value $t[A]$ with higher likelihood than a random guess, if and only if:

1) A third party knows the minimal RFDs holding on the unencrypted dataset, hence s/he can also infer the non-holding RFDs by looking at the partially encrypted dataset;
2) There exists a set of attributes X such that φ': $X \rightarrow A$ does not hold on r, but it holds on $r \setminus t$, and there exists a non-empty set of tuples s whose projection on X is similar to $t[X]$, and all tuples in $r \setminus t$ share the same value of A.

In fact, in this case, the reason why φ' does not hold on r can only be that the masked value $t[A]$ is different from that of the tuples in s, hence the third party can discard that value from his/her guesses. To tackle this borderline case, we encrypt the value of a further attribute on the LHS of the minimal RFD.

Example 7: Let us consider the scenario described in the example 6, where for the RFD φ: $AB \rightarrow D$ we highlighted a borderline case (see Table 3(d)). According to the proposed methodology, also attribute B is encrypted. In this way, also the violation induced by tuple t_1 is masked so that a third party could only give a random guess on the value distribution for attribute D.

VI. THE GENERAL PROCESS

In this section, we describe the general process of the proposed methodology and provide a sample scenario.

Figure 2 shows how the proposed methodology can be applied to a generic scenario. The process starts by considering a given dataset, the set of RFDs holding on it, and a file containing several IC attribute sets, i.e. users' policies concerning attributes specified as confidential. The first step (RFD parsing) aims to filter out only non-key RFDs from the set of RFDs holding on the given dataset, since key RFDs cannot permit to determine any values, since all tuples differ on the attributes of their LHS, hence they are not a threat to confidentiality. Moreover, users are grouped according to the specified policies, through an aggregator module. Then, for each specified policy, RFDs are filtered out by selecting those having one of the confidential attributes on their RHS (filter by IC attributes). All of their LHSs will represent the collection of confidentiality-violating attribute sets for the specific policy. Thus, one of the three heuristics defined above can be applied to retrieve the minimal set of attributes to be encrypted. Moreover, to verify whether the borderline case described in Section V occurs, we must check whether its two conditions are satisfied. To this end, we need to compute the set of non-holding RFDs, which is accomplished by removing attributes to be masked from the LHSs of the RFDs in which they are involved. If a resulting RFD reveals a borderline case, then its LHS will be added as confidentiality-violating

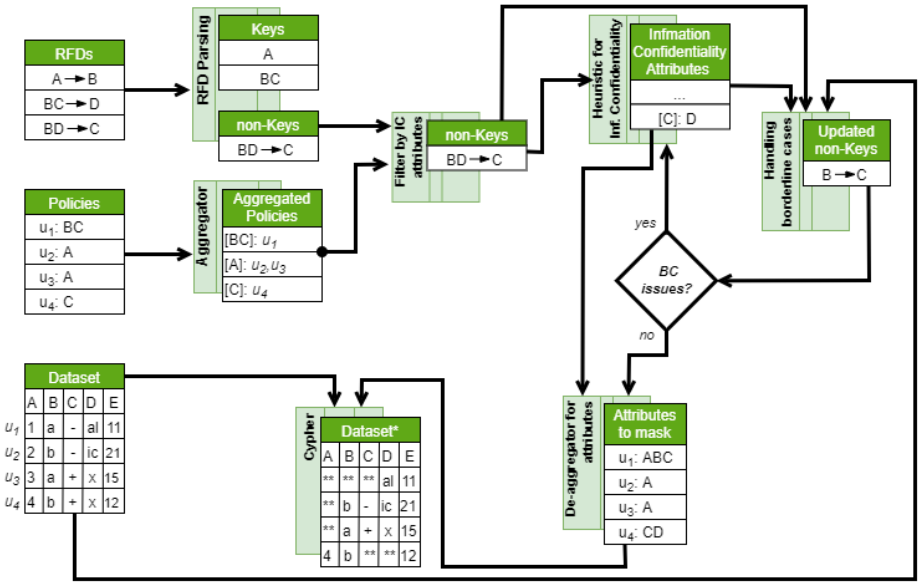


FIGURE 2. The general process for masking data according to users' policies.

attribute set. Once all borderline cases have been detected, the process iterates the application of a heuristic to derive the additional attributes to be encrypted. At the end of this process, a de-aggregator module permits to obtain the attributes to be encrypted for each user, according to his/her specified policy. Finally, the prescribed masking is applied to the entire dataset.

The pseudo-code of the proposed methodology is provided in Algorithm 1. It takes as input the dataset D to be masked, the set Σ of RFDs holding on D, and the set Λ of users specified policies, each defined in terms of IC attribute sets, and it returns as output the masked dataset D^*. At Lines $1-2$, the algorithm invokes the functions `REMOVE_KEY_RFDs` and `POLICY_AGGREGATOR` to remove key RFDs and group equal policies specified by different users. Then, for each specified policy, the algorithm performs the following steps: (i) it computes its associated *confidentiality-violating attribute set* (Line 4), (ii) it applies one of the three proposed heuristics to identify the additional attributes to be encrypted (Line5); (iii) it verifies whether there exist *borderline cases*, by updating the confidentiality-violating attribute sets (Line 6), and iteratively repeating the application of heuristics until no more borderline cases exist (Lines $7-10$); and (iv) it runs the `DEAGGREGATOR_FOR_POLICY` function for mapping attributes to be encrypted to the data of users (Line 11). Finally, the encryption step is performed (Line 13).

Figure 3 shows a masked dataset resulting from the application of Algorithm 1 to the *CreditClient* dataset. In particular, aiming to simulate the definition of users' policies, we implemented a module that randomly assigned confidential attributes to each user tuple. It is possible to notice that at the end of the application of the proposed methodology, only few values are encrypted, whereas many others remain free. Moreover, it is worth to notice that many differences among the encrypted values are obtained. More specifically, the number of values encrypted for each tuple depends on (i) its associated policy, and (ii) the RFDs holding on the considered dataset. In this way, data declared as confidential can never be derived from *free* data. Thus, this strategy permits to limit the number of values to be encrypted in order to preserve information confidentiality, by increasing the possibilities to perform data analytic processes.

Algorithm 1 The Main Algorithm

INPUT: A dataset D, a set of rfds Σ, a set of policies Ψ
OUTPUT: A dataset partially encrypted D^*

1: $\Sigma' \leftarrow$`REMOVE_KEY_RFDs`(Σ)
2: $\Lambda \leftarrow$`POLICY_AGGREGATOR`(Σ', Ψ)
3: **for each** $p_i \in \Lambda$ **do**
4: $X_\zeta \leftarrow$ `FILTER_BY_IC_ATTRIBUTES`(Σ', p_i)
5: $Z \leftarrow$`GET_IC_ATTRIBUTES`(X_ζ)
6: $X_\zeta \leftarrow$ `UPDATE_X_SET`(Z)
7: **while** `BORDERLINE_CASE`(X_ζ, D) **do**
8: $Z \leftarrow$ `ADD_IC_ATTRIBUTES`(X_ζ)
9: $X_\zeta \leftarrow$ `UPDATE_X_SET`(Z)
10: **end while**
11: $\Psi' \leftarrow$`DEAGGREGATOR_FOR_POLICY`(Λ, Z, p_i)
12: **end for**
13: $D^* \leftarrow$`DATASET_ENCRYPTION`(D, Ψ')

Proof of correctness. In the following, we prove the correctness of the proposed methodology.

Theorem 1: Each attribute value $t[A_i]$ defined as sensitive by user t is preserved after the application of the proposed methodology.

A	B	C	D	E	F	G	H	I	J
LastName	FirstName	StreetAddress	CredentialType	Status	BirthYear	CEDueDate	FirstIssueDate	LastIssueDate	ExpirationDate
*****	Vasanti	39111 Janiya Harbors	*****	SUPERSEDED	1969	*****	20110927	20110927	20130927
*****	Kinda	93379 Rocky Knolls	*****	EXPIRED	1968	?	*****	20161109	*****
*****	Wassan	7661 Wolff Motorway	Nursing Assistant Registration	EXPIRED	1984	?	20091008	*****	20100826
*****	India	29040 Champlin Cape	Counselor Registration	EXPIRED	1957	?	20090224	20090224	*****
*****	*****	05672 Tyrese Turnpike	Nursing Assistant Registration	*****	1990	?	*****	20081110	20090413
*****	Stacey	723 Magali Ways	Licensed Practical Nurse	*****	*****	?	19931013	19961202	19961202
Roy	Casey	05663 Jarred Pine	Medical Assistant Registration	ACTIVE	1981	?	20180501	20190410	20210524
EVERETT	*****	7160 Tyreek Stream	*****	EXPIRED	1970	?	19930323	19930901	19930901
*****	Meaghan	46333 Brice Village	Registered Nurse License	*****	*****	*****	?	*****	*****
*****	Lauren	790 Billy Terrace	Social Worker Independent Clinical License	PENDING	1981	?	?	*****	?
*****	Melissa	*****	Counselor Agency Affiliated Registration	EXPIRED	*****	*****	20170913	20180608	20190706
*****	ROSALIND	023 Heaney Mission	Nursing Assistant Registration	EXPIRED	1947	?	*****	*****	20010317
*****	Debra	51018 Blaze Ways	Nursing Assistant Certification	ACTIVE	1953	?	20100209	*****	20200521
OLSON	JEAN	856 Rudy Knolls	Nursing Assistant Registration	EXPIRED	1951	?	20070327	20100127	20110127
BORST	AMY	*****	Health Care Assistant Certification	EXPIRED	1980	?	20021010	20041010	20041010
*****	Miriam	3281 Bart Creek	*****	EXPIRED	*****	?	20140818	20140818	20150910
Mathew	*****	*****	*****	EXPIRED	1965	?	19900227	19920514	19920514
*****	LILLIE	8421 Cayla Summit	Nursing Assistant Registration	EXPIRED	*****	*****	19911002	19950406	19950406
*****	JUDY	97813 Garret Pine	Health Care Assistant Certification	EXPIRED	1935	*****	19920501	19990304	*****
*****	Kimberly	337 Effertz Mountain	Nursing Assistant Registration	EXPIRED	1990	*****	20160823	20170912	*****
*****	Alyssa	61764 Shawn Spur	Medical Assistant Registration	ACTIVE	1997	*****	20171201	20181227	*****
Patel	Ankit	89136 Baron Parks	Pharmacist Intern Registration	EXPIRED	1991	?	*****	20180531	20190706
*****	Irina	*****	*****	ACTIVE	*****	?	*****	20190706	*****
*****	Kathryn	93355 Sylvester Ville	Counselor Certified Certification	ACTIVE	1977	*****	*****	*****	*****
Numata	Jaclyn	4762 Bogan Alley	Emergency Medical Technician Certification	INOPERABLE	1984	?	20120820	20151210	20191130
*****	ELLEN	*****	Emergency Medical Technician Certification	EXPIRED	1978	?	*****	20040202	20071031
Hatfield	*****	53117 Wiza Rue	Registered Nurse Temporary Practice Permit	EXPIRED	1974	?	20090731	20090731	20100131
*****	Patricia	*****	*****	EXPIRED	*****	*****	*****	19890810	19890810
*****	*****	1416 Crona Locks	Medical Assistant Certification	EXPIRED	*****	?	20130701	20140326	20160228
*****	Flavia	6368 Rhett Prairie	Registered Nurse Temporary Practice Permit	*****	*****	*****	20110729	20110729	20120129
STALLARD	IDABELLE	452 Brown Inlet	Registered Nurse License	EXPIRED	1926	?	19500925	19931031	19931031
*****	*****	*****	*****	EXPIRED	1959	*****	*****	19890328	19890328
WAHYUNI	RETNO	*****	Nursing Assistant Registration	*****	1958	?	20011029	20031002	20031002
*****	JUDY	427 Williamson Dam	Nursing Assistant Certification	EXPIRED	1969	?	19940307	20000824	*****

FIGURE 3. Masked dataset after the application of the proposed methodologies.

Proof: We proceed by contradiction. Let us assume that the user t defined a sensitive value on the attribute A_i and a third party is able to disclose $t[A_i]$ after the application of the proposed methodology. To this end, according to the threat model described in Section V.*c*, the third party can access i) the value distribution of each attribute $d(A_i)$, ii) the set Σ of all minimal RFDs holding on the dataset D, and iii) the partially encrypted dataset D^* resulting upon the application of the proposed methodology. The latter says that $t[A_i]$ is encrypted, together with other values on t. Thus, since $t[A_i]$ is encrypted on D^*, the third party has been able to disclose $t[A_i]$ by only using some free values on D^* and some of the RFDs in Σ. This could occur if and only if at least one of the two following cases occurs:

1) there exists a tuple t' containing a combination of free values such that $t'[X]$ is similar to $t[X]$, and there exists an RFD φ: $X \rightarrow A_i$. This means that whenever two tuples (t, t') are similar on X, then almost always they are similar on A_i, yielding the possibility of determining $t[A_i]$ by looking at $t'[A_i]$;
2) there exists a combination of free values on all the tuples of D^*, such that $t[Z]$ is similar to any $t'[Z]$ on D^*, then all tuples in $D^* \setminus t$ are free and have a similar value on A_i, and it does not exist an RFD φ': $Z \rightarrow A_i$ in Σ, but there exists at least one RFD φ: $X \rightarrow A_i$ in Σ such that Z is a direct subset of X, i.e. $ZB = X$ for at least an attribute $B \notin Z$ and $B \neq A_i$. This means that, since all tuples of D^* are similar on Z, and all tuples of $D^* \setminus t$ are similar on A_i, then only the tuple t represents a violation making φ': $Z \rightarrow A_i$ not holding on D, yielding the possibility of determining $t[A_i]$ with a higher likelihood than a random guess, by looking at the value distribution of $d(A_i)$ and by excluding all the values similar to at least one $t'[A_i]$ on $D^* \setminus t$. This becomes a certainty for $|A_i| = 2$.

However, the third party is unable to exploit case 1), since the proposed methodology considers the LHS of each RFD in Σ that determines A_i as a confidentiality-violating attribute set, and encrypts at least one attribute for each of them. Thus, no combination of free values can satisfy the LHS of any RFD in Σ that determines A_i. Moreover, the third party is unable to exploit case 2), since the proposed methodology considers it as *borderline case* and forces the encryption of at least another attribute on the confidentiality-violating attribute set representing the LHS of an RFD revealing such borderline case. This implies that neither case 1) nor case 2) occur, and a third party cannot exploit attribute correlations and free values to disclose values declared as sensitive, contradicting the original assumption. ■

VII. EVALUATION

This section presents the experiments we performed for evaluating the proposed methodology on several public datasets. Our goal is to evaluate the performances of the three defined heuristics on different real-world datasets. This is due to the fact that they represent approximate solutions to the problem of finding the minimum number of attributes to be encrypted. For this reason, we expected that heuristics produce different results in terms of the number of attributes. As detailed in the following, we also tried to use different settings for RFD relaxation criteria.

A. IMPLEMENTATION DETAILS

We implemented several tools to support our methodology by using the Java language. In particular, to discover RFDs holding on a given dataset, we used the discovery algorithm

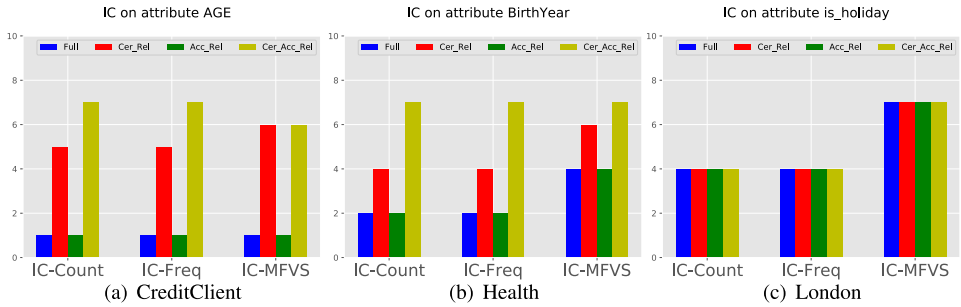


FIGURE 4. Evaluation results of the proposed methodology for Information Confidentiality.

TABLE 4. Statistics on the datasets used in the evaluation.

Datasets	# Columns	# Rows	# FD	Size [KB]
CreditClient	10	30000	152	1730
Health	10	45250	72	4650
London	13	17414	514	1500

defined in [6], and analyzed output RFDs through the algorithm described in Section VI. The latter also implements the three heuristics *counting*, *weighted counting*, and *MFVS*, in order to select values to be partially encrypted. In particular, the values of selected attributes are encrypted with AES in Cipher-Block-Chaining (CBC) mode [37].

B. HARDWARE

The experiments were performed on a machine with an Intel Core i5-4210U CPU with 2.4 GHz, 8 GB of memory, running Windows 10 operating system, and a 64-Bit Java environment.

C. DATASETS

We used three public datasets [38], which were augmented by artificially introducing some confidential data. In particular, we added attributes Name, Surname, and StreetAddress, randomly selecting their values for all tuples. Statistics on the characteristics of the considered datasets are reported in Table 4.

D. EVALUATION SESSION ON INDIVIDUAL DATASETS

We defined a privacy preservation scenario, in which we supposed that each user specified one attribute to be confidential. Moreover, for each of them, we used the proposed methodology to derive the minimum *number of attributes* to be encrypted for guaranteeing users' privacy. This scenario has been evaluated through four experimental sessions, in which we considered different sets of RFDs, according to several threshold settings. In particular, we tried canonical FDs, RFDs relaxing on the extent only (total accuracy degree), on the attribute comparison method only (total certainty degree), and on both.

In the first session, we considered total certainty and total accuracy degree. In the second session, we reduced the certainty degree by also considering RFDs relaxing on the extent only, admitting a $g3$-error of 10%. In the third session, we reduced only the accuracy degree, by considering RFDs relaxing on the attribute comparison method only, setting a distance threshold equal to 1 for each attribute in the dataset. Finally, in the last session, we considered RFDs relaxing on both criteria.

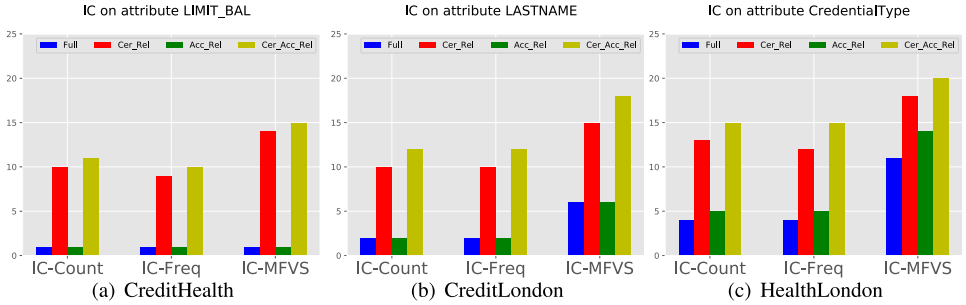
Figure 4 shows evaluation results for each considered dataset, grouping bars according to the used heuristics: counting (IC-Count), weighted counting (IC-Feq), and MFVS (IC-MFVS). More specifically, we show the number of attributes to be encrypted for each used heuristic, and each of the sessions specified above. We use *Full* to denote no relaxation, *Cer_Rel* to denote relaxation on the extent only, *Acc_Rel* to denote relaxation on the attribute comparison method only, and *Cer_Acc_Rel* to denote relaxation on both.

In Figure 4(a) it is possible to notice that although the number of attributes to be encrypted for the *CreditClient* dataset is quite different across several configurations, it is quite similar across the three heuristics on the same configuration. Among the three heuristics, IC-MFVS is the best-performing one with the *Cer_Acc_Rel* configuration. On the contrary, with the *Cer_Rel* configuration IC-Count and IC-Freq heuristics achieve better performances than IC-MFVS. In Figure 4(b) we notice that on the *Health* dataset IC-Count and IC-Freq achieve better performances than IC-MFVS in all configuration settings, except for the *Cer_Acc_Rel* configuration, where the number of attributes to be encrypted is the same as the other two heuristics. Similar considerations apply for the *London* dataset (Figure 4(c)), where IC-MFVS results are worse than those of IC-Count and IC-Freq. Moreover, for this dataset, we notice that no variability is encountered across several configuration settings.

We can conclude that relaxation settings can affect the number of attributes to be encrypted. As expected, RFD relaxation usually increases the number of attributes to be encrypted, since more attribute correlations are generated, but this also yields stronger confidentiality preservation. In particular, results highlight the trade-off between the amount of encryption and the degree of confidentiality preservation that could be achieved with the proposed methodology.

E. GENERAL EVALUATION SESSION ON INTEGRATED DATASETS

Since the proposed methodology aims to highlight the information confidentiality risks arising during several big



(a) CreditHealth (b) CreditLondon (c) HealthLondon

FIGURE 5. Evaluation results for Information Confidentiality on the integrated datasets.

TABLE 5. Statistics of the integrated datasets.

Datasets	# Columns	# Rows	# FD	Size [KB]
CreditHealth	16	30000	1518	368000
CreditLondon	20	17414	11364	193000
HealthLondon	20	17414	6816	272000

data processing activities, like for instance data integration, we performed an evaluation session on datasets derived through data integration processes. In particular, we repeated the general evaluation session defined above for each of the following integrated datasets:

(i) *CreditHealth*, integrating *CreditClient* and *Health*
(ii) *CreditLondon*, integrating *CreditClient* and *London*
(iii) *HealthLondon*, integrating *Health* and *London*

Statistics on the characteristics of the integrated datasets are reported in Table 5. The data integration process was accomplished based on the following attributes common to the three datasets: Name, Surname, and StreetAddress.

Results are shown in Figure 5. In detail, in Figure 5(a) it is possible to notice that for the *CreditHealth* dataset the IC-Freq heuristic performs better than the other two. In particular, this arose in both *Cer_Rel* and *Cer_Acc_Rel* settings. Figure 5(b) shows that the IC-Count and IC-Freq heuristics perform better than the IC-MFVS on the *CreditLondon* dataset, in all configuration settings. A similar behaviour occurred for the *HealthLondon* dataset (Figure 5(c)).

Although we expected that the number of attributes to be encrypted would increase when integrating datasets with respect to the single datasets, by comparing results in Figure 4 and Figure 5, we notice that this happens only when RFD relaxation is introduced. This might be due to the fact that RFD relaxation potentially increases the possibility to catch inter-schema relationships between attributes.

F. IC VARIABILITY EVALUATION SESSION ON INDIVIDUAL DATASETS

As a further experiment, we performed IC variability evaluation, by monitoring the number of attributes to be encrypted as the number of IC attributes grows.

Table 6 shows the attributes used for the IC variability evaluation on the dataset *CreditClient*, Table 7 those used for the *Health* dataset, and Table 8 those used for the *London* dataset. We varied the IC attributes in the range [1, 5], by adding an attribute concerning personal or non-personal user's data at a time. Furthermore, in order to compare results concerning all three defined heuristics, for each of them we have analysed how the number of attributes to be encrypted changes as the number of IC attributes increases.

TABLE 6. Attributes selected for evaluating IC variability on the CreditClient dataset.

Attributes	Length	Description
AGE	1	User's age.
PAY_0	2	User's amount paid added.
FIRSTNAME	3	User's first name added.
BILL_AMT1	4	User's bill in September 2005 added.
STREETADDRESS	5	User's street address added.

TABLE 7. Attributes selected for evaluating IC variability on the Health dataset.

Attributes	Length	Description
BirthYear	1	User's year of birth.
CEDueDate	2	User's CE due date posted added.
FirstName	3	User's first name added.
Status	4	User's status added.
CredentialType	5	User's credential type added.

TABLE 8. Attributes selected for evaluating IC variability on the London dataset.

Attributes	Length	Description
is_holiday	1	User's boolean value.
is_weekend	2	User's boolean value added.
firstname	3	User's first name added.
lastname	4	User's last name added.
streetAddress	5	User's street address added.

Figure 6 shows the results achieved on the *CreditClient* dataset. In particular, the x-axis represents the number of IC attributes, whereas the y-axis represents the number of attributes to be encrypted in order to guarantee requirements on IC attributes for each defined heuristic. More specifically, for the *CreditClient* dataset, all the defined heuristics show a linear growth of the number of attributes to be encrypted w.r.t the number of IC attributes for both *Full* and *Acc_Rel* configurations, and a sub-linear growth for both *Cer_Rel* and *Cer_Acc_Rel* configurations. However, although we can notice an increasing trend for all the three heuristics, sometimes the number of attributes to be encrypted decreases

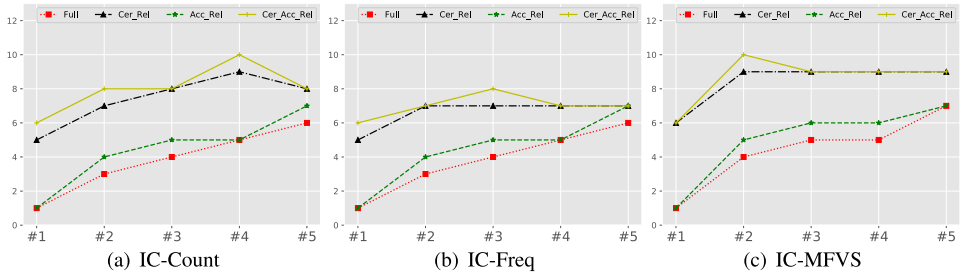


FIGURE 6. Evaluation results of the proposed methodology on IC variation for the CreditClient dataset.

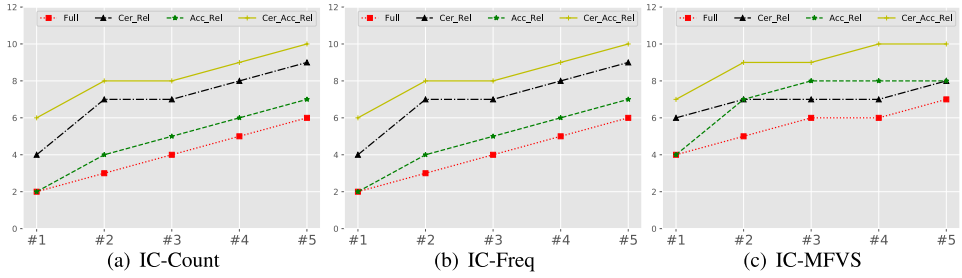


FIGURE 7. Evaluation results of the proposed methodology on IC variation for the Health dataset.

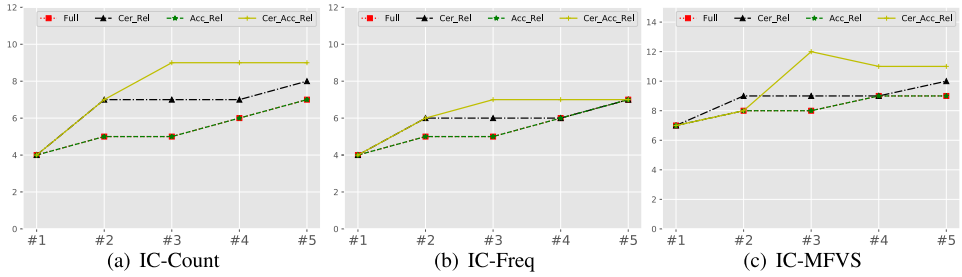


FIGURE 8. Evaluation results of the proposed methodology on IC variation for the London dataset.

as the number of IC attributes increases (*Cerr_Acc_Rel* configuration).

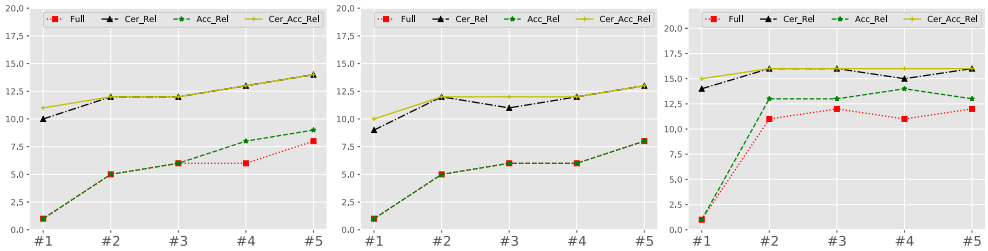
Figure 7 shows the results achieved on the *Health* dataset. In particular, IC-Count in Figure 7(a) and IC-Freq in Figure 7(b) exhibit a linear growth for all the configurations. Instead, IC-MFVS in Figure 7(c) shows more variability in the growing trend for all configurations. In particular, for the *Cer_Rel* configuration, results exhibit a strong growth in the range $[1-2]$, and a constant trend in the range $[2-4]$. Similarly, for the *Acc_Rel* configuration a strong growth is registered in the range $[1-3]$ and a constant trend in the range $[3-5]$.

Figure 8 shows results achieved on the *London* dataset. In particular, IC-Count and IC-Freq heuristics (Figure 8(a)-Figure 8(b)) follow a similar trend for each considered configuration, i.e. a linear growth for both *Full* and *Acc_Rel* configurations, and a sub-linear growth for both *Cer_Rel* and *Cer_Acc_Rel* configurations. However, the number of attributes to be encrypted is greater for IC-Count than for IC-Freq. For IC-MFVS (Figure 8(c)) the trends are similar to those described above for *Full*, *Cer_Rel*, and *Acc_Rel* configurations, but not for *Cer_Acc_Rel*, due to the strong growth registered in the range $[2-3]$. Moreover, it also registered a decrease in the range $[3-4]$. Generally, results of IC-MFVS are worse in terms of the number of attributes to be encrypted w.r.t. the other two heuristics.

In general, it is not obvious that the number of attributes to be encrypted decreases when the number of confidential attributes increases. However, when a new confidential attribute is added, the process typically considers many more RFDs. Thus, the incidence of each attribute w.r.t. the selection criteria of a heuristic could change. Consequently, a heuristic could converge towards a more optimal solution, i.e. fewer attributes to be encrypted.

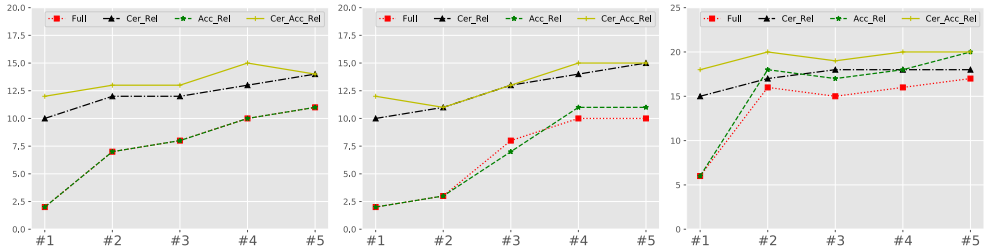
G. IC VARIABILITY EVALUATION SESSION ON INTEGRATED DATASETS

We have accomplished a further IC variability evaluation session on the previously described integrated datasets (Table 5). In particular, also in this case we started by specifying one



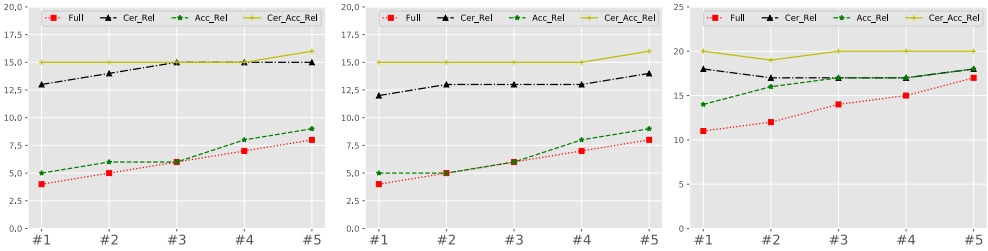
(a) IC-Count (b) IC-Freq (c) IC-MFVS

FIGURE 9. Evaluation results of the proposed methodology on IC variation for the CreditHealth dataset.



(a) IC-Count (b) IC-Freq (c) IC-MFVS

FIGURE 10. Evaluation results of the proposed methodology on IC variation for the CreditLondon dataset.



(a) IC-Count (b) IC-Freq (c) IC-MFVS

FIGURE 11. Evaluation results of the proposed methodology on IC variation for the HealthLondon dataset.

TABLE 9. Attributes selected for evaluating IC variability on the CreditHealth dataset.

Attributes	Length	Description
LIMIT_BAL	1	User's balance.
ExpirationDate	2	User's expiration date added.
AGE	3	User's age added.
Status	4	User's status added.
EDUCATION	5	User's education added.

TABLE 10. Attributes selected for evaluating IC variability on the CreditLondon dataset.

Attributes	Length	Description
LASTNAME	1	User's last name.
t1	2	User's temperature added.
AGE	3	User's age added.
is_holiday	4	User's boolean value added.
MARRIAGE	5	User's boolean added.

confidential attribute, and adding new ones up to 5. More precisely, Table 9 shows the attributes used for IC variability experiments on the *CreditHealth* dataset, Table 10 those for the *CreditLondon* dataset, and Table 11 those for the *HealthLondon* dataset.

TABLE 11. Attributes selected for evaluating IC variability on the HealthLondon dataset.

Attributes	Length	Description
CredentialType	1	User's credential type.
hum	2	User's humidity added.
BirthYear	3	User's birth year added.
weather_code	4	User's weather code added.
StreetAddress	5	User's street address added.

Figure 9 shows the results achieved on the *CreditHealth* dataset. In general, all three heuristics mainly show an increasing trend for all considered configurations. More specifically, the trend is exactly the same for *Full* and *Acc_rel* configurations with IC-Freq. Moreover, in these two configurations, the number of attributes to be encrypted remains sufficiently low. Instead, a remarkable growth occurs with IC-MFVS for *Full* and *Acc_Rel* configurations in the variability range $[1-2]$.

Figure 10 shows the results achieved on the *CreditLondon* dataset. In particular, also for this dataset IC-MFVS exhibited a similar behaviour for *Full* and *Acc_Rel* configurations, and

for *Cer_Rel* and *Cer_Acc_Rel* configurations. Better performances are achieved with IC-Count and IC-Freq, where the latter follows a non-monotonic trend for the *Cerr_Acc_Rel* configuration.

Figure 11 shows the results obtained on the *HealthLondon* dataset. In particular, IC-Count and IC-Freq heuristics (Figure 11(a)-Figure 11(b)) do not require to encrypt many attributes for *Full* and *Acc_Rel* configurations. This does not occur with IC-MFVS for the same configurations (Figure 11(c)). Moreover, we notice that although the *Cerr_Acc_Rel* configuration requires the maximum number of attributes to be encrypted, it follows a quasi-constant trend with all three defined heuristics.

By comparing results achieved in this evaluation w.r.t. the previous one, we can notice that there are no relationships between the trends on the integrated datasets and those on the single datasets from which they are derived. Often, *Full* and *Acc_Rel* configurations required less attributes to be encrypted than *Cer_Rel* and *Cer_Acc_Rel* configurations.

H. INFORMATION GAIN EVALUATION SESSION

In this section, we describe another evaluation session aiming to analyze the effectiveness of the proposed methodology. In particular, we considered a classification scenario in which it is important to guarantee the quality of data even if the privacy preservation must be ensured. Thus, we measured the data quality in terms of information gain on the unencrypted dataset, and compared it to the partially encrypted one obtained by the application of the proposed methodology. More specifically, we aimed to understand the dispersion of the data in terms of information gain [39], which exploits the concept of entropy. The latter is defined in equation (7), and characterizes the purity of an arbitrary collection of examples.

$$\text{Entropy} \;=\; H(X) = -\sum p(X)\log p(X) \qquad (7)$$

where

- X is an attribute of the dataset;
- $H(X)$ is the entropy of X;
- $p(X)$ is the probability of getting a value of X when randomly selecting one from the set.

Instead, the Information Gain is the expected reduction in entropy caused by partitioning the examples according to a given attribute. The formal definition of the information gain is expressed in (8).

$$\text{Information Gain} = \; I(X, Y) = H(X) - H(X|Y) \qquad (8)$$

where

- X and Y are attributes of the dataset;
- $I(X, Y)$ is the information gain on the attribute Y;
- $H(X)$ is the entropy on X;
- $H(X|Y)$ is the entropy of X given Y.

To perform our analysis, we evaluated the variation of information gain for each attribute in the Health dataset, and used Status as the target attribute of the classification scenario. Thus, we evaluated the information gain of each attribute w.r.t. Status. Moreover, in the computation of information gain, we considered every encrypted value as belonging to the same class, like in the case of null values.

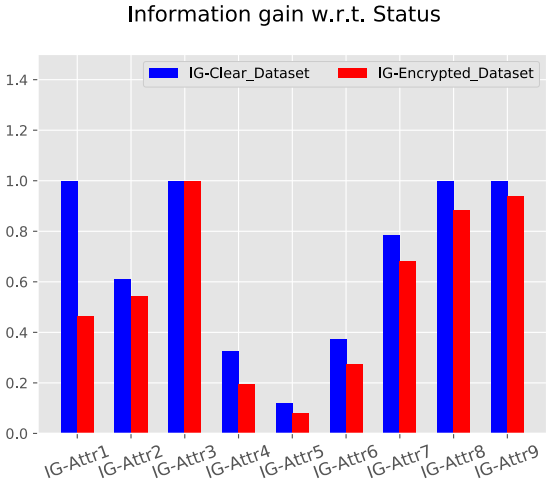


FIGURE 12. Evaluation results of the information gain on the Health dataset.

Figure 12 shows the obtained results, where the blue bar is related to the information gain computed on the attributes without encryption (e.g. exposed to privacy threats), and the red one is related to the information gain computed after the application of the proposed methodology, i.e. the partially encrypted dataset. According to Figure 12, it is possible to notice that the variation of information gain is almost always small. This highlights the fact that the proposed methodology is a useful means to guarantee privacy preservation without affecting too much the quality of data. More specifically, some exceptions have been encountered. A slightly worse behavior is obtained for *IG-Attr1*, i.e. LastName, due to the many encryptions on an attribute whose distribution contains many values. Instead, for *IG-Attr3* the information gain remains unchanged.

This evaluation represents a specific analysis scenario, which allowed us to verify how in a real-world scenario it is possible to work with partially encrypted data, aiming to ensure both privacy preservation and data usage.

VIII. CONCLUSION

We proposed a methodology to automatically identify and partially encrypt ''sensitive'' data, in order to detect several threats to information confidentiality. The methodology provides a contribution to organizations in the effort to comply with new regulations concerning privacy preservation, like GDPR. The identification procedure exploits the semantic correlations among data, represented through automatically discovered RFDs [6], [33], to derive the minimal set of data to be encrypted. The methodology exploits RFDs holding between attributes belonging to different sources since an attacker might inquire them, or they might be put together as a result of big data processing operations, like for instance data integration, entity resolution, and so on, because data

belonging to different data sources could imply the values of previously specified confidential ones. To this end, the proposed methodology supports the detection of such threats, providing organizations with a useful means to accomplish complex data processing operations in a safe way, contributing to reduce the risk of incurring expensive penalties, as prescribed by the *GDPR* regulation.

The proposed methodology exploits partial encryption methods, aiming to maximize the amount of free data on which organizations can perform analysis. In particular, we have defined a threat model and analyzed how the proposed methodology reduces the possibility of breaking information confidentiality. To this end, one of the limitations of the proposed methodology is that a third party could identify unique records and attempt to disclose confidential information by exploiting, for instance, record linkage techniques. Nevertheless, in our application scenario, we exclude attributes typically used for identifying tuples referring to a single user (e.g. SSN, student ID, and so on). Moreover, although not impossible, an attacker should have much detailed external information to link user data to the specific target ones. Finally, another limitation of the proposed methodology is that users are not always aware of all of their sensitive data. To this end, even if organizations could support users by defining some default sensitive attribute sets, in our opinion this problem should be further investigated in the context of GDPR compliant privacy preservation. Experimental results demonstrated that the proposed methodology can help to detect many confidentiality threats while requiring to encrypt a reduced number of attributes to prevent them.

To tackle the above-mentioned problems, in the future we plan to extend the proposed methodology in several directions. In particular, beyond the possibility to use the proposed methodology in combination with some anonymity preserving strategies (e.g. k-anonymity, and so on), we would like to leverage on recent data profiling tools [40] to gain additional metadata other than RFDs, which could be useful to detect this and other potential confidentiality threats, highlighting further privacy preservation threats and the corresponding actions for neutralizing them. In particular, these tools can provide metadata concerning many different types of data dependencies, unique values, foreign keys, and so on, based on which we can empower the proposed methodology to prevent several additional potential attacks. Moreover, we would like to define new heuristics from the characteristics of RFDs, aiming to further minimize the number of attributes to be encrypted. Finally, we aim to embed the proposed methodology within self-service data preparation tools, especially those targeted to end-users and data stewards.

REFERENCES

[1] X. Lin, R. Lu, and X. S. Shen, "MDPA: Multidimensional privacy-preserving aggregation scheme for wireless sensor networks," *Wireless Commun. Mobile Comput.*, vol. 10, no. 6, pp. 843–856, 2010, doi: 10.1002/wcm.796.

[2] F. Li, B. Luo, and P. Liu, "Secure and privacy-preserving information aggregation for smart grids," *Int. J. Security Netw.*, vol. 6, no. 1, pp. 28–39, Apr. 2011, doi: 10.1504/IJSN.2011.039631.

[3] A. C. Squicciarini, M. Shehab, and F. Paci, "Collective privacy management in social networks," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, 2009, pp. 521–530, doi: 10.1145/1526709.1526780.

[4] V. M. Shelake and N. Shekokar, "A survey of privacy preserving data integration," in *Proc. Int. Conf. Electr., Electron., Commun., Comput., Optim. Techn. (ICEECCOT)*, Dec. 2017, pp. 59–70.

[5] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, "Privacy-preserving record linkage for big data: Current approaches and research challenges," in *Handbook of Big Data Technologies*, A. Y. Zomaya and S. Sakr, Eds. Springer, 2017, pp. 851–895, doi: 10.1007/978-3-319-49340-4_25.

[6] L. Caruccio, V. Deufemia, and G. Polese, "Mining relaxed functional dependencies from data," *Data Mining Knowl. Discovery*, vol. 34, pp. 443–477, Dec. 2019, doi: 10.1007/s10618-019-00667-7.

[7] European Commission. (2018). *Recital 71 of Reform of EU Data Protection Rules 2018*. [Online]. Available: https://www.privacy-regulation.eu/en/r71.htm

[8] J. Garber, "GDPR–compliance nightmare or business opportunity?" *Comput. Fraud Secur.*, vol. 2018, no. 6, pp. 14–15, Jun. 2018.

[9] K. Renaud and L. A. Shepherd, "How to make privacy policies both GDPR-compliant and usable," in *Proc. Int. Conf. Cyber Situational Awareness, Data Analytics Assessment (Cyber SA)*, Jun. 2018, pp. 1–8, doi: 10.1109/CyberSA.2018.8551442.

[10] K. H. Gjermundrød, I. Dionysiou, and K. Costa, "PrivacyTracker: A privacy-by-design GDPR-compliant framework with verifiable data traceability controls," in *Proc. Int. Workshops Current Trends Web Eng. (ICWE)*, 2016, pp. 3–15, doi: 10.1007/978-3-319-46963-8_1.

[11] E. Bertino, "Big data–security and privacy," in *Proc. IEEE Int. Congr. Big Data*, New York City, NY, USA, Jun. 2015, pp. 757–761, doi: 10.1109/BigDataCongress.2015.126.

[12] E. Bertino, G. Ghinita, and A. Kamra, "Access control for databases: Concepts and systems," *Found. Trends Databases*, vol. 3, nos. 1–2, pp. 1–148, 2010, doi: 10.1561/1900000014.

[13] L. Caruccio, V. Deufemia, C. D'Souza, A. Ginige, and G. Polese, "A tool supporting end-user development of access control in Web applications," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 25, no. 2, pp. 307–331, Mar. 2015, doi: 10.1142/S0218194015400112.

[14] M. Nabeel and E. Bertino, "Privacy preserving delegated access control in public clouds," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2268–2280, Sep. 2014, doi: 10.1109/TKDE.2013.68.

[15] T. Kumar, A. Braeken, M. Liyanage, and M. Ylianttila, "Identity privacy preserving biometric based authentication scheme for naked healthcare environment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7, doi: 10.1109/ICC.2017.7996966.

[16] M. I. Sarfraz, M. Nabeel, J. Cao, and E. Bertino, "DBMask: Fine-grained access control on encrypted relational databases," in *Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, 2015, pp. 1–11, doi: 10.1145/2699026.2699101.

[17] L. Sweeney, "K-ANONYMITY: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002, doi: 10.1142/S0218488502001648.

[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond K-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[19] J. Domingo-Ferrer and J. Soria-Comas, "From t-closeness to differential privacy and vice versa in data anonymization," *Knowl.-Based Syst.*, vol. 74, pp. 151–158, Jan. 2015, doi: 10.1016/j.knosys.2014.11.011.

[20] W. Jiang and C. Clifton, "A secure distributed framework for achieving K-anonymity," *VLDB J.*, vol. 15, no. 4, pp. 316–333, Nov. 2006, doi: 10.1007/s00778-006-0008-z.

[21] Z. Xiao, X. Meng, and J. Xu, "Quality aware privacy protection for location-based services," in *Proc. 12th Int. Conf. Database Syst. Adv. Appl. (DASFAA)*, 2007, pp. 434–446, doi: 10.1007/978-3-540-71703-4_38.

[22] S. Agrawal, J. R. Haritsa, and B. A. Prakash, "FRAPP: A framework for high-accuracy privacy-preserving mining," *Data Mining Knowl. Discovery*, vol. 18, no. 1, pp. 101–139, Feb. 2009, doi: 10.1007/s10618-008-0119-9.

[23] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *J. Financial Statist.*, vol. 2, no. 3, p. 329, 1986.

[24] D. Vatsalan, P. Christen, C. M. O'Keefe, and V. S. Verykios, "An evaluation framework for privacy-preserving record linkage," *J. Privacy Confidentiality*, vol. 6, no. 1, p. 42, Jun. 2014, doi: 10.29012/jpc.v6i1.636.

[25] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," *BMC Med. Informat. Decis. Making*, vol. 9, no. 1, p. 41, Dec. 2009, doi: 10.1186/1472-6947-9-41.

[26] T. Churches and P. Christen, "Blind data linkage using n-gram similarity comparisons," in *Proc. 8th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, 2004, pp. 121–126, doi: 10.1007/978-3-540-24775-3_15.
[27] D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, "Distance-aware encoding of numerical values for privacy-preserving record linkage," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 135–138, doi: 10.1109/ICDE.2017.58.
[28] K. Schmidlin, K. M. Clough-Gorr, and A. Spoerri, "Privacy preserving probabilistic record linkage (P3RL): A novel method for linking existing health-related data and maintaining participant confidentiality," *BMC Med. Res. Methodol.*, vol. 15, no. 1, p. 46, Dec. 2015.
[29] European Commission. (2016). *General Data Protection Regulation—Final Version of the Regulation*. Accessed: Apr. 6, 2016. [Online]. Available: http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf
[30] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proc. Workshop Inf. Integr. Web*, 2003, pp. 73–78. [Online]. Available: http://www.isi.edu/info-agents/workshops/ijcai03/papers/Cohen-p.pdf
[31] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, "TANE: An efficient algorithm for discovering functional and approximate dependencies," *Comput. J.*, vol. 42, no. 2, pp. 100–111, Feb. 1999, doi: 10.1093/comjnl/42.2.100.
[32] R. W. Proctor, E. E. Schultz, and K.-P. L. Vu, "Human factors in information security and privacy," in *Handbook of Research on Information Security and Assurance*. Hershey, PA, USA: IGI Global, 2009, pp. 402–414.
[33] L. Caruccio, V. Deufemia, and G. Polese, "On the discovery of relaxed functional dependencies," in *Proc. 20th Int. Database Eng. Appl. Symp. (IDEAS)*, 2016, pp. 53–61, doi: 10.1145/2938503.2938519.
[34] L. Caruccio, V. Deufemia, F. Naumann, and G. Polese, "Discovering relaxed functional dependencies based on multi-attribute dominance," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 17, 2020, doi: 10.1109/TKDE.2020.2967722.
[35] F. V. Fomin, S. Gaspers, A. V. Pyatkin, and I. Razgon, "On the minimum feedback vertex set problem: Exact and enumeration algorithms," *Algorithmica*, vol. 52, no. 2, pp. 293–307, Oct. 2008, doi: 10.1007/s00453-007-9152-0.
[36] W. Stallings, "The offset codebook (OCB) block cipher mode of operation for authenticated encryption," *Cryptologia*, vol. 42, no. 2, pp. 135–145, Mar. 2018, doi: 10.1080/01611194.2017.1422048.
[37] O. Reparaz and B. Gierlichs, "A first-order chosen-plaintext DPA attack on the third round of DES," in *Proc. 16th Int. Conf. Smart Card Res. Adv. Appl. (CARDIS)*, 2017, pp. 42–50, doi: 10.1007/978-3-319-75208-2_3.
[38] C. L. Blake and C. J. Merz. *UCI Repository of Machine Learning Databases*. Accessed: Mar. 11, 2018. [Online]. Available: http://archive.ics.uci.edu/ml/index.php
[39] J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Mach. Learn.*, vol. 3, no. 4, pp. 319–342, Mar. 1989, doi: 10.1007/BF00116837.
[40] F. Naumann, "Data profiling revisited," *ACM SIGMOD Rec.*, vol. 42, no. 4, pp. 40–49, Feb. 2014, doi: 10.1145/2590989.2590995.

LOREDANA CARUCCIO received the B.Sc. *(cum laude)* and M.Sc. *(cum laude)* in computer science from the University of Salerno, in 2009 and 2012, respectively, and the Ph.D. degree in management and information technology, in 2018. In 2017, she has been a visiting student with the Hasso Plattner Institute, University of Potsdam. In 2020, she has been an Adjunct Professor with the Department of Computer Science, Université Claude Bernard Lyon 1. She is currently an Adjunct Professor and Postdoctoral Researcher with the Department of Computer Science, University of Salerno. Her research interests include data science, artificial intelligence, and web engineering. She is the Co-Chair of the 27th International DMS Conference on Visualization and Visual Languages (DMSVIVA). She regularly serves as a Reviewer in conferences and journals, such as the *Data Science and Engineering* (Springer), *Information Sciences* (Elsevier), *Fuzzy Sets and Systems* (Elsevier), and IEEE Access.

DOMENICO DESIATO received the B.Sc. and M.Sc. *(cum laude)* degrees in computer science from the University of Salerno, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree in computer science and information & communication technology. His research interests include data privacy, big data, and machine learning.

GIUSEPPE POLESE (Member, IEEE) received the Laurea degree *(cum laude)* in computer science from the University of Salerno, the M.Sc. degree in computer science from the University of Pittsburgh, PA, USA, and the Ph.D. degree in applied mathematics and computer science from the jointed Universities of Salerno, Naples, and Catania. He was the Project Manager of the Italian Airspace Company, Alenia. He is currently an Associate Professor with the University of Salerno, the Director of the Data Science and Technologies Laboratory, and the Coordinator of the master track in data science and machine learning of the master's degree in computer science. His research interests include data science, artificial intelligence, information visualization, and web engineering. He is a member of the ACM. He is also a member of the editorial board of the following international journals: *Information Systems (Elsevier)*, *ACM Journal on Data and Information Quality*, *Distributed and Parallel Databases* (Springer), *Multimedia Tools and Applications* (Springer), *International Journal of Software and Knowledge Engineering* (Kluwer), and *Translational Medicine@UniSa*. He regularly serves as a Reviewer in journals and conferences in the fields of data science, big data, and information visualization. He has been a Visiting Professor with the Computer Science Department, University of Pittsburgh, Pittsburgh, PA, USA, and the Institute of Computational Science, University of Italian Switzerland.

GENOVEFFA TORTORA (Senior Member, IEEE) has been a Full Professor of Computer Science, since 1990. She was the Dean of the Faculty of Mathematical, Natural and Physical Sciences, University of Salerno, from 2000 to 2008. She is the Scientific Director of the Laboratory on Context-Aware Intelligent Systems, Department of Computer Science. Her interests include software engineering, visual languages and human-machine interaction, image processing and biometric systems, big data, data warehouses, data mining, and geographic information systems. She is the author and coauthor of more than 290 papers published in scientific journals or proceedings of refereed conferences and is co-editor of three books. She is also the editorial board member of high-quality international journals; a steering committee member of the International Working Conference on Advanced Visual Interfaces, held in cooperation with the ACM; the Program Chair and Program Committee Member of several relevant international conferences; a Senior Member of the IEEE Computer Society; a member of the ACM; the European Association of Theoretical Computer Science (EATCS) Member; the International Association of Pattern Recognition (IAPR) Reviewer for several international scientific journals; an Evaluator of research projects for Italian Ministries, Regions, Universities; and European Commission Member of the Board of Examiners for several researchers, associate professor, and full professor positions both at a national and at an international level.