FCA-based reasoning for privacy

GONZALO A. ARANDA-CORRAL[*], *Department of Information Technology, Universidad de Huelva, Avda. de las Fuerzas Armadas s/n. 21007 Huelva, Spain.*

JOAQUÍN BORREGO-DÍAZ[**], *Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, Avda. Reina Mercedes s.n. 41012-Sevilla, Spain.*

JUAN GALÁN-PÁEZ[†], *Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, Avda. Reina Mercedes s.n. 41012-Sevilla, Spain and Datrik Intelligence S.A., Spain.*

Abstract

Notwithstanding the potential danger to security and privacy, sharing and publishing data has become usual in Data Science. To preserve privacy, de-identification methodologies guided by risk estimation have been designed. Two issues associated with classical risk metrics are, on the one hand, the adequacy of the metric and, on the other hand, its static nature. In this paper, we present metrics for estimating risk based on the *emerging semantics* provided by Formal Concept Analysis. The metrics are designed to estimate the *a priori* risk of compromised data deletion. Furthermore, by applying specialized *variable forgetting* methods for association rules, it is shown how to reflect the effect of deleting attributes belonging to potentially dangerous quasi-identifier sets. Additionally, a study of the role of the risk metric in confidence-based reasoning for re-identification is presented.

Keywords: Privacy, re-identification, formal concept analysis, risk metrics, association rules.

1 Introduction

In Data Science, the problem of anonymization (or deidentification) consists of avoiding the individual identification using the information stored in the data set [14]. Deidentification has become a requirement for information sharing, mainly because anonymized data-sharing does not require the express consent of the individuals represented in the dataset. Data-sharing is crucial in sensitive fields such as Healthcare or Security. In general, it is potentially strategic in any area of Engineering. The dual problem, the reidentification, consists of manipulating databases to determine the identity of persons whose information appears in them.

To address both practices, it is necessary to consider the nature of the project in which the dataset will be used. Moreover, some considerations about the attacking agents that will attempt

[*]E-mail: gonzalo.aranda@dti.uhu.es
[**]E-mail: jborrego@us.es
[†]E-mail: juangalan@us.es

Vol. 32, No. 2, © The Author(s) 2024. Published by Oxford University Press.

Advance Access published 19 February 2024 https://doi.org/10.1093/jigpal/jzae011

to re-identify the dataset must be taken into account [14]. This kind of preliminary analysis allows the engineer to outline different criteria depending on the degree of sharing and the nature of the agents involved in the problem. Thus, the scenarios range from controlled use cases (such as e.g. the case where only a certain set of scientists have access to the dataset) to that of published open data (e.g. open data about the COVID-19 pandemic). In the latter case, the knowledge and skills of the potential attacking agents are unknown. Different scenarios bring to consider, in turn, several types of attacks against the unidentified dataset [15], as well as different methods of managing other related problems [25].

Historically, Privacy has been a major issue in Healthcare, due to the dilemma between pragmatic, scientific (reproducibility of experiments and data reusability) needs versus ethical and legal ones. The so-called Safe Harbor principles [23] provide a first approach to tackle the challenge, detailing which fields should be eliminated or anonymized (e.g. through pseudonyms). It is from this background that Data Science has inherited a series of best practices, guidelines and methodologies. Among them, there are issues that arise from the problem of reidentification attacks [15]. Moreover, different deidentification standards have been published (e.g. [10, 16, 19]) that complement and adjust traditional principles to the current technological and legal realms.

This paper concerns a particular case, ubiquitous in all privacy scenarios. We refer to the so-called *attribute disclosure*: when the attacker learns that an attribute or set of attributes provides information-rich enough to know who is one of the individuals in the de-identified dataset. For this problem, different scenario-based techniques are still needed. A *Closed World* scenario is where only a restricted set of agents have access, and there is little or no linkage of the data to other external data. In the *Open World* scenario, the data are linkable to external data, the set of agents with access is very large or the data are open to anyone. For the latter case, other techniques can—and should—improve on the traditional ones, such as Differential Privacy [12] that seeks to achieve the so-called privacy *ad omnia*.

From all of the above, it is clear that a major requirement for any Risk-minimization-Based Methodology (RBM) in Data Science is the availability of methods to quantify it; a metric for estimating the danger of re-identification in any stage [14]. When such metrics are available, the RBM will be concerned, on the one hand, with keeping the risk value small (such value is relative to the working scenario), considering robust thresholds. On the other hand, RBM will try to ensure that the techniques used to reduce risk do not cause an unacceptable loss of information that would make the dataset unusable.

Data scientists use privacy practices that work on variables or sets of variables that are *direct identifiers* (i.e. that allow the attacker to recognize an individual from the database). These practices also aim to analyse the impact of the elimination of records or variables on the knowledge implicit in the dataset. Examples of such techniques are the *masking* or the simple *variable deletion*. To carry them, it is necessary to identify variable sets for which, if some combination of values is known, then re-identification is very likely(*quasi-identifiers*). Finding such sets is a persistent activity in RBM.

1.1 Aim and structure of the paper

The thesis of the paper is that the so-called Formal Concept Analysis (FCA) can help to design metrics, semantic in nature, which consider some amount of *implicit knowledge* in the dataset. This paper (which extends a previously published work [7]) aims to enrich the classical risk analyses by employing new metrics that take into account the conceptual structure (in the sense of FCA [17]) extracted from the dataset. Also, the relationship among metrics and association rules mined from the dataset are analysed using FCA.

	Quasi-identifiers		Other Variables	
ID	Sex	Year of Birth	Lab Test	Lab Result
1	Male	1959	Albumin, Serum	4.8
2	Male	1969	Creatine Kinase	86
3	Female	1955	Alkaline Phosphatase	66
4	Male	1959	Bilirubin	Negative
5	Female	1942	BUN/Creatinine Ratio	17
6	Female	1975	Calcium, Serum	9.2
7	Female	1966	Free Thyroxine Index	2.7
8	Female	1987	Globulin, Total	3.5
9	Male	1959	B-type Natriuretic Peptide	134.1
10	Male	1967	Creatine Kinase	80
11	Male	1968	Alanine Aminotransferase	24
12	Female	1955	Cancer Antigen 125	86
13	Male	1967	Creatine Kinase	327
14	Male	1967	Creatine Kinase	82
15	Female	1966	Creatinine	0.78
16	Female	1955	Triglycerides	147
17	Male	1967	Creatine Kinase	73
18	Female	1956	Monocytes	12
19	Female	1956	HDL Cholesterol	68
20	Male	1978	Neutrophils	83
21	Female	1966	Prothrombin Time	16.9
22	Male	1967	Creatine Kinase	68
23	Male	1971	White Blood Cell Count	13.0
24	Female	1954	Hemoglobin	14.8
25	Female	1977	Lipase, Serum	37
26	Male	1944	Cholesterol, Total	147
27	Male	1965	Hematocrit	45.3

NOTE: BUN = blood urea nitrogen; HDL = high-density lipoprotein

	Quasi-identifiers		Probability of
ID	Sex	Decade of Birth	Re-identification
1	Male	1950-1959	0.33
2	Male	1960-1969	0.125
3	Female	1950-1959	0.167
4	Male	1950-1959	0.33
5	Female	1940-1949	1
6	Female	1970-1979	0.33
7	Female	1960-1969	0.33
8	Female	1980-1989	1
9	Male	1950-1959	0.33
10	Male	1960-1969	0.125
11	Male	1960-1969	0.125
12	Female	1950-1959	0.167
13	Male	1960-1969	0.125
14	Male	1960-1969	0.125
15	Female	1960-1969	0.33
16	Female	1950-1959	0.167
17	Male	1960-1969	0.125
18	Female	1950-1959	0.167
19	Female	1950-1959	0.167
20	Male	1970-1979	1
21	Female	1960-1969	0.33
22	Male	1960-1969	0.125
23	Male	1970-1979	0.33
24	Female	1950-1959	0.167
25	Female	1970-1979	0.33
26	Male	1940-1949	1
27	Male	1960-1969	0.125

FIGURE 1. Example from [20]. This paper focuses on some quasi-identifiers (right).

Additionally, for supporting and refining the decision-making process, techniques from the so-called *Three-Way Decision* paradigm are applied. These may be successfully applied to analyse conceptual structures from FCA [8].

The structure of the paper is as follows. The next section recalls the fundamental elements of risk-based metrics and FCA. Section 3 is devoted to introducing FCA-based risk-based metrics. In Section 4 the dilemma of risk versus data preservation from the viewpoint of the Three-Way Decision paradigm is addressed. Section 5 tackles the relationship between the semantic risk and the association rule reasoning. Section 6 shows how to enrich RBM with the results of previous sections. The paper finishes with some considerations and ideas for future work.

2 Background

This section summarizes the main elements of both risk metrics and FCA. To illustrate the different proposals of the paper, it has been selected as running example a small dataset from [20] (Figure 1). The original dataset contained a direct identifier, the patient's name, has already been masked by the ID variable (it is supposed that the correspondence ID↔Patient is securely stored by the dataset sponsor).

There exist other interesting features which can be observed. For example, the ID set $\{10, 13, 14, 17, 22\}$ can be characterized as the individuals sharing the values $\{Male, 1967\}$ for the *quasi-identifier* $\{Sex, Year - of - Birth\}$. This type of individual set is referred to as *equivalence class* in [14]. This kind of class is characterized by variable values, so it is expected to be large enough to prevent the recognizing of a individual in the event that the attacker agent knows its attribute values, and they coincide with those of the quasi-identifier.

It can perform techniques that aim to augment the size of the equivalent class. For example, the variable $Year - of - Birth$ could be modified to lose precision, augmenting this way the

size of the new equivalent class. For instance, by taking decades as the time scale. For example, the equivalence class for the modified quasi-identifier $\{Male, 1960 - 1959\}$ increases, resulting in $\{2, 10, 11, 13, 14, 17, 22, 27\}$.

2.1 Risk-based metrics

Roughly speaking, metrics for re-identification risk could be classified in *exogenous* and *endogenous*. The former needs estimations of the probability of successful attacks, while the latter only uses features of the data set [14]. This paper is concerned with the second one. This way all the factors to consider come from the dataset itself.

Risk metrics aim to estimate risk when the adversary knows someone in the real world and is trying to find his entry in the dataset, or has chosen a record from the dataset and is trying to find out the identity of the individual [14, 20].

DEFINITION 2.1
Let D be a dataset and $Q = \{v_1, \ldots, v_k\}$ a variable subset (the quasi-identifier) of D.

- The **re-identification risk** of an entry (individual) e with respect to a set of variable values $v_1 = q_1, \ldots, v_k = q_k$ is defined as

$$r(e, q_1, \ldots, q_k) = \frac{1}{s(q_1, \ldots, q_k)}$$

  being $s(q_1, \ldots, q_k)$ the number of records sharing such values.
- An entry of D is a **unique record** with respect to Q if there exist values $v_1 = q_1, \ldots, v_k = q_k$ such that $r(e, q_1, \ldots, q_k) = 1$.

In the running example, it has $r(14, sex, Year - of - birth) = 0.2$. The case of unique records represents the maximum re-identification risk. If the attacker knew these variable values, then he/she could extract the remaining information. In the dataset of Figure 1, if the attacker knows the identity of the person born in 1965, then the attacker may also know his clinical information because $r(27, Male, 1965) = 1$.

DEFINITION 2.2

- The **maximum risk** of D with respect to $q_1, \ldots, q_k$ is

$$r_M(D, v_1, \ldots, v_k) := \max\{r(e, q_1, \ldots, q_k) \;\; e \in D, \; q_i \text{ value for } v_i, 1 \leq i \leq k\}$$

- The **average risk** of D is the expected value,

$$r_m(D, v_1, \ldots, v_k) := \mathbb{E}[r(., q_1, \ldots, q_k)]$$

For the example, $r_M(D, Sex, Year - of - Birth) = 1$, while the average risk would be $r_m(D, Sex, Year - of - Birth) = 0.59$. The selection of the risk metric depends again on the dataset usage scenario. It is common to select the maximum risk when *data sponsor* aims to share data within an open environment (Open Data); it is reasonable to think that the adversary will focus on the highest risk records. The opposite case is where data access is restricted, and re-identification could be forbidden by contract. In this case, the average risk could be more appropriate; the problem would be similar to inadvertent re-identification.

	River	Coast	Sea
Carp	×		
Escatofagus	×	×	
Bream		×	×
Sparus		×	×
Eel	×	×	×

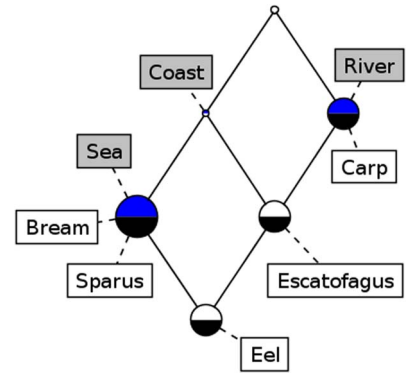


FIGURE 2. Formal context on fish and its associated concept lattice.

Anonymization practices consider generally acceptable to use 5 as the minimum size of a equivalence classes; $s(e, q_1, \ldots, q_k) \geq 5$, i.e. $r_M(D, v_1, \ldots, v_k) = 0.2$). However, for some studies higher risks are accepted, such as 0.33 (used by US CDC in HIV studies). In general terms, an average risk greater than 0.5 is also unacceptable [14]. Unique records deletion relieves both risks. For example, after applying the suppression of unique records in the example in Figure 1, the new database D' has $r_m(D') = 0.33$.

2.2 Formal Concept Analysis

The data structure in FCA is the object–attribute table called **formal context**. It is a three-element set $\mathbb{K} = (G, M, I)$, where G is a non-empty set of *objects* (events), M is a non-empty set of *attributes* and $I \subseteq G \times M$ is a binary relation. Figure 2 (left) shows a formal context that describes fish (objects) that live in different aquatic ecosystems (attributes).

The primary goal of FCA is the extraction and analysis of *concepts* from data. A concept is a unit of thought comprising its *extent* and its *intent*. The first covers all objects belonging to the concept, and the second comprises all attributes valid for all objects under consideration. On $\mathbb{K} = (G, M, I)$ two operators called *derivation operators* are considered. Given $A \subseteq G$ and $B \subseteq M$, these can be defined as:

$A' = \{m \in M \mid (g, m) \in I \text{ for all} g \in A\}$ and$B' = \{g \in G \mid (g, m) \in I \text{ for all } m \in B\}$

That is, the set of attributes shared by all objects in A (the *intent* of A), and the objects that have all the attributes of B (the *extent* of B), respectively. In terms of FCA—identifying the variable values with attributes—the extent of the quasi-identifier would correspond to its extent.

A **formal concept** is a pair $C = (A, B)$ object/attribute sets (extent $Ext(C) = A$ and intent $Int(C) = B$ of the concept, respectively) such that $A' = B$ and $B' = A$.

The relation *subconcept of*, $\leq$, between concepts is defined by: $(X_1, Y_1) \leq (X_2, Y_2)$ if $X_1 \subseteq X_2$. The derivation satisfies that, if $X_1 \subseteq X_2$, then $X_2' \subseteq X_1'$. It also satisfies $X \subseteq X''$ and $X' = X'''$. From these properties, the following characterization can be achieved:

- The smallest concept that contains an object set $X \subseteq G$ is (X'', X'),
- the largest concept that contains an attribute set $Y \subseteq M$ is (Y', Y'').
- the smallest concept containing $g \in G$; i.e. $Con(g) := (\{g\}'', \{g\}')$.

	Need Water	Aquatic	Mobility	Legs
Cat	×		×	×
Leech	×	×	×	
Frog	×	×	×	×
Corn	×			
Fish	×	×	×	

1 < 5 > { } ==> Need water;
2 < 3 > Need water Aquatic ==> Mobility;
3 < 2 > Need water Legs ==> Mobility;

FIGURE 3. Formal context on live beings (top) and Stem Basis (down).

The set $\mathfrak{B}(\mathbb{K})$ of the concepts of $\mathbb{K}$ can be endowed with the mathematical structure of lattice throughout the subconcept relationship [17]. For example, the concept lattice of the formal context of Figure 2 left is shown at right. It actually represents the Hasse diagram of the lattice. This is a graph representing the relationship $C_1 \prec C_2$ from bottom to top defined by $C_1 \leq C_2$ *and there is not any intermediate concept*. In this representation, each node is a concept, and its intent (extent, resp.) is formed by the set of attributes (objects, resp.) included along the path to the top (bottom resp.) concept. For example, the bottom concept $(\{eel\}, \{Coast, Sea, River\})$ could be interpreted as a new concept, *euryhaline-fish*, which has been discovered.

2.2.1 Implication logic and association rules in FCA In the tradition of deductive databases, the logics on FCA are based on implications expressing attribute dependence. An **implication** of $\mathbb{K} = (G, M, I)$ is an expression $Y_1 \rightarrow Y_2$, where $Y_1, Y_2 \subseteq M$. The semantics are inherited from the propositional logic interpretation of implications, but relativized to the formal context. The set of implications will be denoted by $Imp(M)$.

Formally, $Y_1 \rightarrow Y_2$ is said to be **valid** for a set $T \subseteq M$ (or T is a *model* of the implication), written $T \models Y_1 \rightarrow Y_2$, if the following condition is satisfied: *if* $Y_1 \subseteq T$ *then* $Y_2 \subseteq T$. The implication $Y_1 \rightarrow Y_2$ is **valid in the context** $\mathbb{K} = (G, M, I)$, denoted by $\mathbb{K} \models Y_1 \rightarrow Y_2$, if $\{g\}' \models Y_1 \rightarrow Y_2$ for any object $g \in G$ (i.e. the set of attributes of any object models the implication).

For example, the implication $River, Sea \rightarrow Coast$ (*if a fish lives in both rivers and the sea then it also lives in the coast*) is valid in the context of Figure 2, whereas the $River \rightarrow Coast$ does not.

DEFINITION 2.3
Let $\mathbb{K} = (G, M, I)$ be a formal context and $\mathcal{L} \cup \{L\} \subseteq Imp(M)$. It is said that

1. L is **consequence of** $\mathcal{L}$ (denoted by $\mathcal{L} \models L$) if each model of $\mathcal{L}$ also models L.
2. $\mathcal{L}$ **entails** $\mathcal{L}'$, denoted by $\mathcal{L} \models \mathcal{L}'$, if every implication of $\mathcal{L}'$ is consequence of $\mathcal{L}$.
3. $\mathcal{L}$ is **complete** for $\mathbb{K}$ if for every implication L, if $\mathbb{K} \models L$ then $\mathcal{L} \models L$.
4. $\mathcal{L}$ is **non-redundant** if for all $L \in \mathcal{L}$, $\mathcal{L} \setminus \{L\} \not\models L$.
5. $\mathcal{L}$ is an **implication basis** for $\mathbb{K}$ if $\mathcal{L}$ is both complete for $\mathbb{K}$ and non-redundant.

A particular basis is the so-called *Duquenne-Guigues Basis* (also called *Stem Basis*, SB) [18], extracted from a type of attribute sets (pseudo-intents) [17]. An example is shown in Figure 3. Reasoning with implications can be performed using them as a production system [3, 4].

TABLE 1. Two Luxenburger bases (adding the Stem Basis) for Example of Figure 2

$\mathcal{L}(\mathbb{K}, 0.5, 2/5) =$

$\mathcal{L}(\mathbb{K}, 0.8, 1) =$

$Implication$	$Confidence$	$Support$
$Sea \rightarrow Coast$	1	1
$\{\} \rightarrow Coast$	$4/5$	1

$Coast \rightarrow Sea$	$3/4$	$4/5$
$River \rightarrow Coast$	$2/3$	$1/3$
$\{\} \rightarrow River$	$3/5$	$3/5$
$River, Coast \rightarrow Sea$	$1/2$	$2/5$

In FCA, association rules are implications—not necessarily valid—between attributes. Confidence and support are defined as usual in data mining:

- The **support of an implication** $L = Y_1 \rightarrow Y_2$ is $supp(L) = \frac{|(Y_1 \cup Y_2)'|}{|G|}$
- The **confidence of** $L = Y_1 \rightarrow Y_2$ is $conf(L) = \frac{|(Y_1 \cup Y_2)'|}{|Y_1'|}$

The analogous to the stem basis for association rules would be the so-called Luxenburger basis [22]. A set Y is **closed** if $Y'' = Y$. Given Y_1, Y_2 closed, it is denoted $Y_1 \prec Y_2$ when $Y_1 \subset Y_2$ and there is no Y closed such that $Y_1 \subset Y \subset Y_2$.

DEFINITION 2.4
Given $0 \leq \gamma, \delta \leq 1$, the **Luxenburger basis** of $\mathbb{K}$ with confidence γ and support δ is

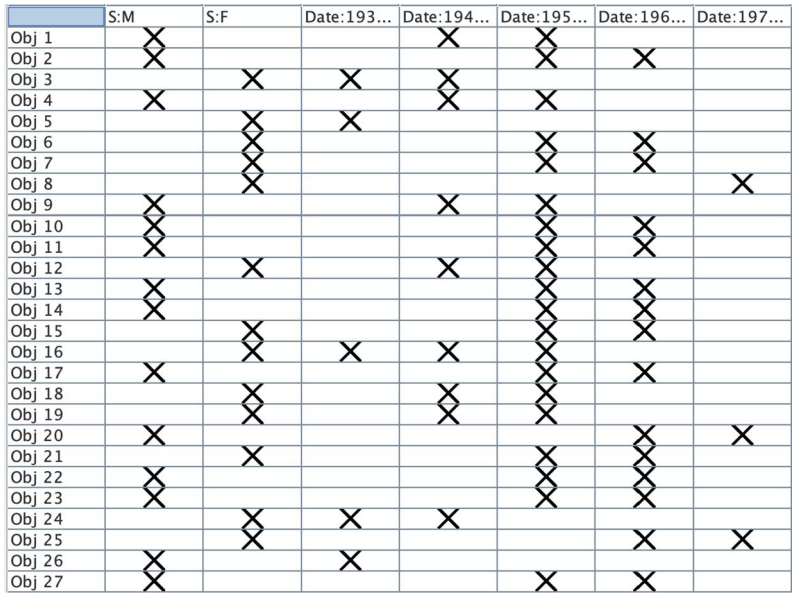
$$\mathcal{L}(\mathbb{K}, \gamma, \delta) := \{L : Y_1 \rightarrow Y_2 \mid Y_1, Y_2 \text{closed}, Y_1 \prec Y_2,\ conf(L) \geq \gamma,\ sup(L) \geq \delta\}$$

The reasoning system for Stem bases can be adapted for reasoning with Luxenburger bases [3]. To simplify the notation, it is supposed that stem basis (they are implications with confidence 1) is contained in Luxenburger basis. Implications of Luxenburger basis work as association rules from the classic data mining setting. Two Luxenburger bases for the example from Figure 2 are depicted in Table 1.

2.3 Evaluation-based three-way decision model

The so-called 3WD research paradigm is a relatively new proposal to study the decision-making processes in a general way (see e.g. [27, 28]). 3WD aims to unify different approaches to decision making under uncertainty. Techniques within 3WD have been developed in fields such as Data Science and Big Data [24], Machine Learning [1] or incremental concept learning [30], among others. This is not a paper on 3WD research, but we tackle here a decision problem, namely, to eliminate or preserve risk data. The theoretical solution will be 3WD-inspired.

The ground idea in 3WD is to work on the problem of the interpretation of the workspace (system inputs, beliefs, etc.) by studying three regions linked with decision-making procedures (cf. [29]): positive, negative and boundary (acceptance, rejection and non-commitment regions, respectively). It is the startpoint for tasks aimed to refine the boundary region (i.e. to solve non-commitment), a major concern within 3WD.



	S:M	S:F	Date:193...	Date:194...	Date:195...	Date:196...	Date:197...
Obj 1	×			×	×		
Obj 2	×				×	×	
Obj 3		×	×	×			
Obj 4	×			×	×		
Obj 5		×	×				
Obj 6		×			×	×	
Obj 7		×			×	×	
Obj 8		×					×
Obj 9	×			×	×		
Obj 10	×				×	×	
Obj 11	×				×	×	
Obj 12		×		×	×		
Obj 13	×				×	×	
Obj 14	×				×	×	
Obj 15		×			×	×	
Obj 16		×	×	×	×		
Obj 17	×				×	×	
Obj 18		×		×	×		
Obj 19		×		×	×		
Obj 20	×					×	×
Obj 21		×			×	×	
Obj 22	×				×	×	
Obj 23	×				×	×	
Obj 24		×	×	×			
Obj 25		×				×	×
Obj 26	×		×				
Obj 27	×				×	×	

FIGURE 4. Context obtained by scaling from the dataset of Figure 1.

There are multiple ways to formalize the idea that reflect different ways of understanding both the decision process and the ontological nature of such regions. One way is using an evaluation function to classify the inputs. The general formulation suggests to work with ranked evaluations in a poset. In the present paper, the range of values of the evaluation will be the (ordered) interval $[0, 1]$, identifying 0 with *false* and 1 with *true*. An **accept-reject evaluation** $\omega : U \to [0, 1]$ determines three decision regions,

$$POS_\omega = \omega^{-1}(1),\ NEG_\omega = \omega^{-1}(0) \text{ and} BND_\omega = \{u \in U\ \ 0 < \omega(u) < 1\}$$

No more elements of this approach will be detailed here (we refer the reader to [29]), since these will be introduced later, within specific approach of the paper. A natural way to refine the boundary region is to take new thresholds to expand positive and negative regions, e.g. $POS_{\omega,\beta} = \{u \in U\ \ \beta \leq \omega(u) \leq 1\}$.

3 Estimating risk using formal concepts

The standard technique to transform the dataset D into a formal context $\mathbb{K}$ is by employing *scales* [11, 17]. Roughly, these are designed by selecting attributes representing particular values or ranges of values of the dataset variables. This way a variable with multiple values is transformed into a (set of) attribute(s).

To transform the running example into one more illustrative for the paper, the so-called *nominal scale* has been chosen for the variable *Sex*. Also, for others that are (non-disjoint) time intervals for $Year - of - Birth$. In this way, the risk may increase since the value for an entry could belong to two different intervals; thus, more information is available. The result is shown in Figure 4. The object in

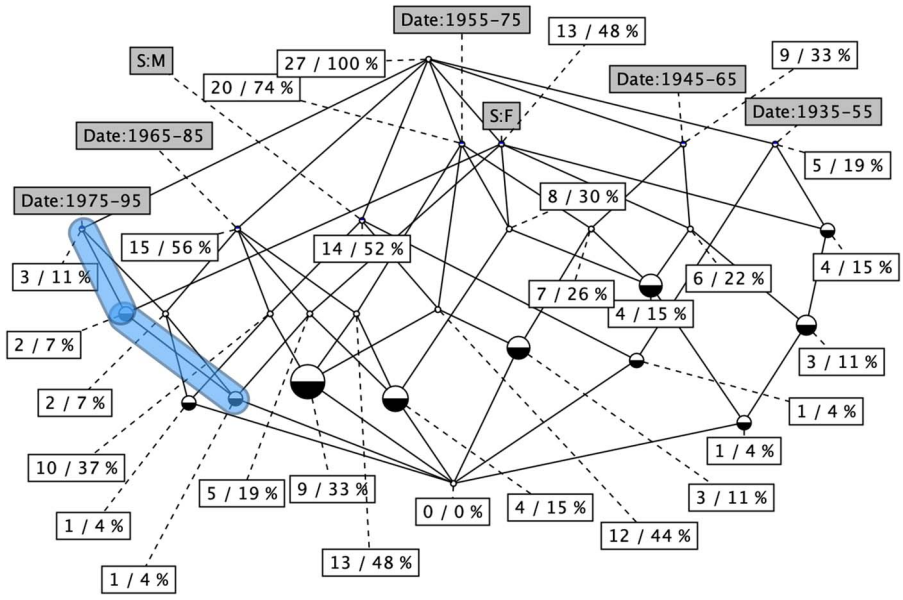


FIGURE 5. Concept lattice computed from Figure 4.

the formal context $\mathbb{K}$ corresponding to the record $e \in D$ will be denoted by $e^{\mathbb{K}}$. The concept lattice of the formal context obtained is shown in Figure 5. Throughout the paper, D will be the dataset to study and $\mathbb{K}$ the formal context associated with D.

Please note that, when working with the formal context, the set of individuals sharing a given set of variable values has been endowed with semantic meaning and is related to others, as shown in the concept lattice. Also, observe that $s(e^{\mathbb{K}}, q_1, \ldots, q_k)$ is the size of the large concept that contains attributes $\{q_1, \ldots, q_k\}$.

3.1 Semantic interpretation of risk metrics

In order to illustrate our proposal, is convenient to firstly analyse the particular case of unique records (i.e. $r_M(e^{\mathbb{K}}) = 1$). In the associated context, such a record corresponds to an object $e^{\mathbb{K}}$ with $|\{e^{\mathbb{K}}\}''| = 1$ (the extent of $Con(e^{\mathbb{K}})$ is a singleton). A feature of the proposed new metric—that the classic metric does not possess—is that it provides information on the effect of erasing an unique record.

DEFINITION 3.1
A $\prec$-chain

$$\mathcal{C} = C_1 \prec C_2 \prec \cdots \prec C_k$$

with$C_j = (X_j, Y_j) \in \mathfrak{B}(\mathbb{K})(1 \leq j \leq k)$ is called a **unique record identifier chain** if $|X_1| = 1$ and $|X_{j+1} \setminus X_j| = 1$ for any $j < k$.

The interest of detecting this type of chain lies in that it allows to predict the effect on the risk metric of deleting an unique record (it can create another one, the next concept in the chain). In

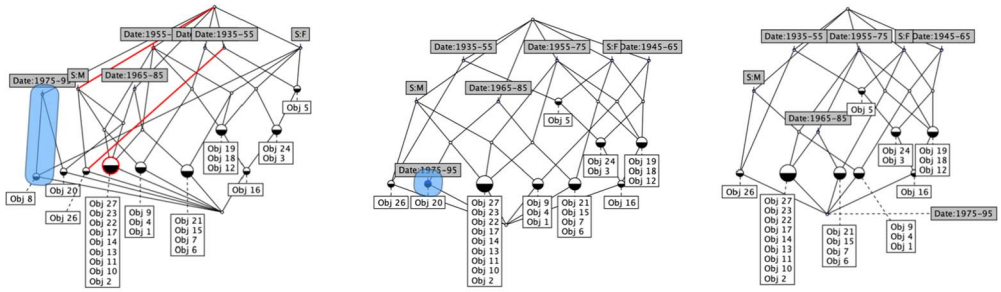


FIGURE 6. Effect of sequential record suppression from unique record identifier chain with extents $\{25\} \subseteq \{25, 28\} \subseteq \{25, 28, 20\}$.

Figure 1, right, a maximal identifier chain is highlighted:

$$(\{25\}, \{Date : 75 - 95, Date : 65 - 85, S : F\}\}) \prec (\{25, 28\}, \{Date : 75 - 95, S : F\}\}) \prec$$
$$\prec (\{25, 28, 20\}, \{Date : 75 - 95\}\})$$

The effect of the sequential suppression of unique records on the concept lattice, showing the apparition of new unique records, is shown in Figure 6. The formal contexts of Figure 6 has been obtained using Conexp,[1] and focuses on the chain discussed above. It has been recomputed the lattice after each record deletion.

3.1.1 Identifier chains Along the rest of the paper, a threshold λ for the risk metric is fixed. The generalization of the concept of unique record identifier chain is as follows, which represents the fact that the elimination of records belonging to a high-risk concept could cause an increase in risk to the next concept of the chain.

DEFINITION 3.2
Let $\mathbb{K}$ be the formal context associated with the database D.

- A chain $\mathcal{C} = C_1 \prec C_2 \prec \cdots \prec C_k$ with $C_j = (X_j, Y_j) \in \mathfrak{B}(\mathbb{K})$ $(1 \leq j \leq k)$ is called a **chain with identification risk** if $|X_1| \leq \lambda$, and $|X_{j+1} \setminus X_j| \leq \lambda$ $(j \leq k)$
- The **risk associated with the chain** $\mathcal{C}$, $r_{\mathcal{C}} : \bigcup X_i \to \mathbb{R}$ is defined as

$$r_{\mathcal{C}}(e^{\mathbb{K}}) = \begin{cases} \frac{1}{|X_1|} & \text{if } e^{\mathbb{K}} \in X_1 \\ \frac{1}{|X_i \setminus X_{i-1}|} & \text{if } e^{\mathbb{K}} \in X_i \setminus X_{i-1} \end{cases}$$

The idea is generalized considering all the $\prec$-predecessors in the lattice, by defining a FCA-based risk by recursion on $\prec$ (without dependence on a particular chain). To simplify the notation, the quasi-identifiers $\{q_1, \ldots q_n\}$ will be omitted in the following definitions. The so-called *basic semantic risk* estimates the risk when there has been a deletion of records at risk on some subconcept of $Con(e^{\mathbb{K}})$.

[1] http://conexp.sourceforge.net/. Updated version in https://github.com/keinstein/Conexp/releases

DEFINITION 3.3
Let e be a record of D. The **basic semantic risk** of e is defined by

$$r_{BS}(e,\mathbb{K}) = \max\{\frac{1}{|\{e^{\mathbb{K}}\}'' \setminus ext(C)|} \;\; C \in \mathfrak{B}(\mathbb{K}),\; C \prec Con(e^{\mathbb{K}})\}$$

The next metric to be introduced formalizes the idea of not assuming the suppression of the elements of a subconcept, but using the probability of occurrence of that event (considering an uniform distribution). It is also well defined on the partial order $\prec$. The definition exploits the fact that the basic semantic risk actually depends on $Con(e^{\mathbb{K}})$: if $\{e_1^{\mathbb{K}}\}'' = \{e_2^{\mathbb{K}}\}''$, then $r_{BS}(e_1^{\mathbb{K}},\mathbb{K}) = r_{BS}(e_2^{\mathbb{K}},\mathbb{K})$.

DEFINITION 3.4
The **semantic risk** of e is $r_S(e,K) := r_L(Con(e^{\mathbb{K}}))$ where $r_L(.) : \mathfrak{B}(\mathbb{K}) \to [0,1]$ is defined by

$$r_L(C,\mathbb{K}) = \begin{cases} \frac{1}{|ext(C)|} & \textit{if } C \textit{ has not non-empty subconcepts} \\ \max\{\frac{1}{|ext(C)\setminus ext(D)|} \cdot r_L(D,\mathbb{K}) \;\; D \prec C\} & \textit{in other case} \end{cases}$$

It is easy to see that $r_L(e,\mathbb{K}) = 1$ iff C belongs to a unique record identifier chain.

DEFINITION 3.5
The **maximum semantic risk** and the **medium semantic risk** are, respectively:

$$r_{BS,M}(\mathbb{K}) = \max\{r_{BS}(e^{\mathbb{K}})\; e \in D\},\; \text{and} r_{BS,m}(\mathbb{K}) = \mathbb{E}[r_{BS}(.)]$$

In Figure 7 the basic semantic risk and the semantic risk for the running example is shown, with threshold $\lambda = 0.35$. Records belonging to highlighted concepts are at risk. Note that such a elimination of some of them do not affect the risks of superconcepts (with $r_{BS} > \lambda$). However, $r_S < \lambda$ indicates that it is very unlikely that the records located in that concept and not in the subconcepts are reidentifiable.

4 A 3WD framework based on the basic semantic risk

In a 3WD evaluation-based decision scenario, two thresholds could be taken, one for discarding $\beta \leq \lambda$ and a safety threshold for preserving records, $\alpha < \lambda$. In such a situation, the 3WD-inspired regions POS_{BS} (decision to be discarded), NEG_{BS} (decision not to be discarded) and the boundary BND_{BS} are defined by

- $POS_{r_{BS}} := r_{BS}^{-1}([\beta,1],\mathbb{K})$
- $NEG_{r_{BS}} := r_{BS}^{-1}([0,\alpha],\mathbb{K})$
- $BND_{r_{BS}} := r_{BS}^{-1}((\alpha,\beta),\mathbb{K})$

However, the analysis should be refined to obtain a dataset with acceptable risk. For example, to try maximizing the number of records preserved (i.e. extend $NEG_{r_{BS}}$) without compromising privacy.

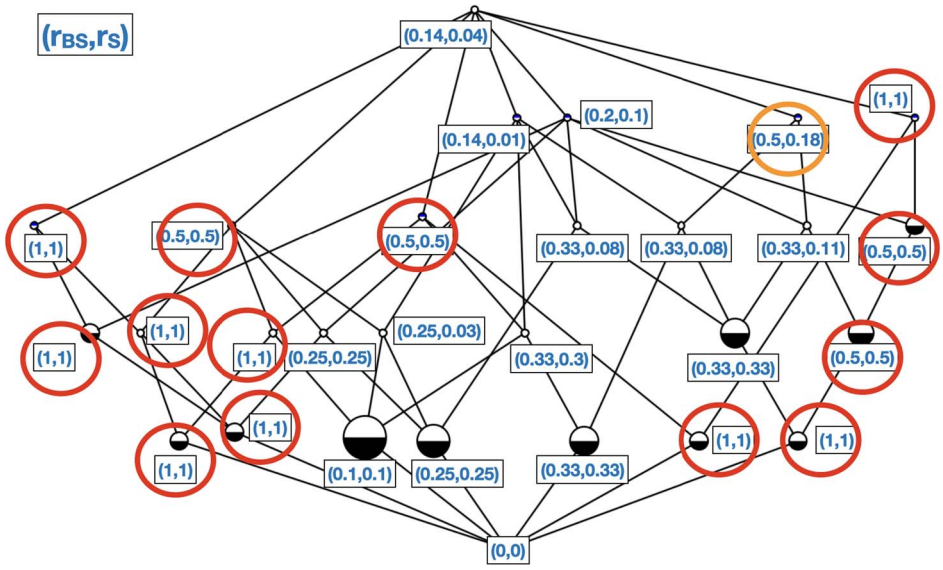


FIGURE 7. Identification of concepts with higher risk.

That is, to seek the decision function $d : D \to \{0, 1\}$ verifying that the formal context induced by $d^{-1}(\{1\})$ has acceptable risk. That is, if

$$\mathbb{K}_{3WD} := (d^{-1}(\{1\}), M, I \cap (d^{-1}(\{1\}) \times M))$$

then $r_{BS,M}(\mathbb{K}_{3WD}) \leq \lambda$.

Given $X \subseteq G$, consider $\mathbb{K}_X := (X, M, I \cap (X \times M))$. An issue with the design of d is that it must be incremental. Please note that the evaluation risk is not necessarily non-decreasing; $X_1 \subset X_2$ does not imply that $r_{BS,M}(\mathbb{K}_{X_2}) \leq r_{BS,M}(\mathbb{K}_{X_1})$ (this fact is true if the classical risk r_M is considered). This fact could be not true for $r_{BS,M}$ since it depends on the lattice structure of $\mathbb{K}_X$, and it is expected that some concepts have smaller size (increasing the risk). The decision on a threshold depends on the self selection of preserved records. Therefore, it is necessary to to formalize the process of eliminating records with higher risk described in Figure 6. Let $\mathbb{K}_n = (G_n, M, I \cap (G_n \times M))$ be defined by

- $G_0 := r_{BS}^{-1}([0, \lambda], \mathbb{K})$, and $\mathbb{K}_0 := (G_0, M, I \cap (G_0 \times M))$ (the context induced by the records with initial acceptable risk).
- Let $G_{n+1} := r_{BS}^{-1}([0, \lambda], \mathbb{K}_n)$, and $\mathbb{K}_{n+1} := (G_{n+1}, M, I \cap (G_{n+1} \times M))$.

Then $\mathbb{K}_{n+1} \subseteq \mathbb{K}_n$. Let $G^* = \bigcap_n G_n$. Then there are two possibilities:

- $G^* = \emptyset$. It is not possible to select a sub-dataset $X \subseteq G$ with $r_{BS,M}(\mathbb{K}_X) \leq \lambda$.
- $G^* \neq \emptyset$. Then $\mathbb{K}^* := (G^*, M, I \cap (G^* \times M))$ has to be studied.

THEOREM 4.1

$$r_{BS,M}(\mathbb{K}^*) \leq \lambda$$

PROOF. If $\bigcap_n G_n \neq \emptyset$ then $G_{j+1} = G_j = G^*$ for some j (since G is finite). Thus $\mathbb{K}^* = \mathbb{K}_{j+1}$. Let e with $Con(e) = C = (X, Y) \in \mathfrak{B}(\mathbb{K}^*)$ and $D \prec C$. Then

$$r_{BS,M}(e, \mathbb{K}^*) = r_{BS,M}(e, \mathbb{K}_{j+1}) \overset{\mathbb{K}_j = \mathbb{K}_j}{=} r_{BS,M}(e, \mathbb{K}_j) \overset{e \in \mathbb{K}_{j+1}}{\leq} \lambda$$

Thus

$$r_{BS,M}(\mathbb{K}^*) = \max\{r_{BS,M}(e, \mathbb{K}^*) \;\; e \in G^*\} \leq \lambda$$

□

Therefore, the decision preserving acceptable risk would defined by: $d(e) = 1$ iff $e \in G^*$.

5 Effect of risk on implication basis and association rules

This section analyses the usefulness of the basic semantic risk metric in a scenario in which the attacker uses some sort of rule mining. The attacker will choose a set of attributes and apply on them some kind of learning algorithm (rule mining, in our case) to obtain certain patterns about those attributes. Such patterns will be association rules. From these, the attacker will try to deduce new attributes for the record. Working with *arguments* formed by attribute and implication sets has already been achieved in the FCA context [9]. Specifically, two types of analysis are proposed.

Firstly, the implication basis is used to estimate how knowledge changes when removing dangerous attributes (Sect. 5.1). One way to estimate how much knowledge is lost would be to relate the stem basis $\mathcal{L}$ of the original dataset to a complete implication set $\mathcal{L}'$ for the formal context obtained by eliminating such attributes. This $\mathcal{L}'$ can be obtained by *conservative retraction* [2]. Secondly, association rules are considered to investigate the relationship between them and a sound risk metric. Two principles will be stated in this regard (Sect. 5.2).

5.1 Effect of attribute elimination on the implication basis

The case of eliminating dangerous information represented by an attribute set Q, is considered to attempt to reduce the re-identification risk. The impact of the elimination could be studied in the lattice (as in Figure 6). However, there is a more efficient way of analysing knowledge loss working with a basis $\mathcal{L}$ of $\mathbb{K}$.

For the context resulting from eliminating these attributes, a complete implication set can be obtained from a basis using the so-called *variable forgetting* [2, 5, 13]. By forgetting all the attributes of Q, the so-called *conservative retraction* [2] is obtained. This is an implication set $\mathcal{L}_0$ such that $\mathcal{L} \models \mathcal{L}_0$ and verifying that any valid implication in the language of $\mathcal{L}_0$ valid in the original context is also a consequence of $\mathcal{L}_0$;

$$\forall L \in Imp(M \setminus Q)[\mathcal{L} \models L \Longrightarrow \mathcal{L}_0 \models L]$$

With these notions, it is possible to specify the case in which it is not feasible to discard the set of attributes: the set Q will be non-disposable if it renders the elimination result to be useless: $\mathcal{L}_0 \equiv \top$ (i.e. is a tautology set). That is, the reduced context does not provide real information. Due to the lack of space, other not so extreme case, for which the information that is lost with the new base must be estimated, will not be addressed. For this case, it is possible to resort to 3WD-inspired metrics for *Knowledge Harnessing* [6].

```
1 < 3 > S:M Date:1945-65 ==> Date:1955-75;
2 < 3 > Date:1935-55 Date:1945-65 ==> S:F;
3 < 1 > Date:1935-55 Date:1955-75 ==> S:F Date:1945-65;
4 < 1 > S:M Date:1975-95 ==> Date:1965-85;
5 < 0 > Date:1935-55 Date:1975-95 ==> S:M S:F Date:1945-65 Date:1955-75 Date:1965-85;
6 < 0 > Date:1945-65 Date:1975-95 ==> S:M S:F Date:1935-55 Date:1955-75 Date:1965-85;
7 < 0 > Date:1955-75 Date:1975-95 ==> S:M S:F Date:1935-55 Date:1945-65 Date:1965-85;
8 < 0 > S:M S:F ==> Date:1935-55 Date:1945-65 Date:1955-75 Date:1965-85 Date:1975-95;
9 < 0 > Date:1935-55 Date:1965-85 ==> S:M S:F Date:1945-65 Date:1955-75 Date:1975-95;
10 < 0 > Date:1945-65 Date:1965-85 ==> S:M S:F Date:1935-55 Date:1955-75 Date:1975-95;
```

```
[Date:1935-55, Date:1955-75, S:M] => [Date:1945-65]
[Date:1935-55, Date:1945-65, S:M] => [Date:1955-75]
[Date:1935-55, Date:1955-75] => [Date:1945-65]
[S:M, Date:1945-65] => [Date:1955-75]
```

FIGURE 8. Stem basis (top) and the result of forgetting {`Date:75-95, Date:65-85,S:F`} (bottom).

Stem basis $\mathcal{L}$ of the running example is shown in Figure 8. If one aims to remove the attributes that appear in the identifier chain of Sect. 3.1,

$$Q = \{Date : 75 - 95, Date : 65 - 85, S : F\}$$

attribute forgetting can be applied (e.g. using the algorithm from [5]) the implication set $\mathcal{L}_0$, shown in Figure 8 (right), is obtained. Thus $\mathcal{L}_0 \not\equiv \top$, so the elimination of Q does not make the dataset useless.

5.2 Risk versus confidence

Figure 9 shows a particular Luxenburger basis for the running example. Association rules can also be used to obtain information about a record (with some probability). Let us now assume that the attacker attempts the re-identification by mining and reasoning with association rules. The elimination of a record guided by the risk metric affects the confidence of such rules (possibly increasing it). It is therefore necessary to study the relationship between both factors. We claim two general principles which can be considered as minimal requisites to be required to any metric:

- *Principle of Diminution of Confidence* (PDC): *If the risk is low, then the confidence of dangerous rules should be small.* The attacker will not entail, from the properties he or she already knows about the individual, new information with relevant confidence.
- *Principle of Lack of Confidence* (PLC): *If the risk is high, upper bounds for rule confidence obtained using the risk metric would be high.* Data owner will not ensure that the attacker can not entail sensitive information with confidence.

The aim of the section is to show that the basic semantic risk satisfies both principles. To this end, risk-based upper bounds for confidence will be established. Firstly, it is necessary to state such a bound for rules of Luxenburger basis.

```
1 < 15 > Date:1965-85 =[87 %]=> < 13 > Date:1955-75;
2 < 14 > S:M =[86 %]=> < 12 > Date:1955-75;
3 < 10 > S:M Date:1965-85 =[90 %]=> < 9 > Date:1955-75;
4 < 12 > S:M Date:1955-75 =[75 %]=> < 9 > Date:1965-85;
5 < 9 > Date:1945-65 =[78 %]=> < 7 > Date:1955-75;
6 < 5 > S:F Date:1965-85 =[80 %]=> < 4 > Date:1955-75;
7 < 5 > Date:1935-55 =[80 %]=> < 4 > S:F;
8 < 3 > S:M Date:1945-65 =[100 %]=> < 3 > Date:1955-75;
9 < 3 > Date:1935-55 Date:1945-65 =[100 %]=> < 3 > S:F;
10 < 4 > S:F Date:1935-55 =[75 %]=> < 3 > Date:1945-65;
11 < 1 > S:M Date:1975-95 =[100 %]=> < 1 > Date:1965-85;
12 < 1 > Date:1935-55 Date:1955-75 =[100 %]=> < 1 > S:F Date:1945-65;
13 < 0 > S:M S:F =[100 %]=> < 0 > Date:1935-55 Date:1945-65 Date:1955-75 Date:1965-85 Date:1975-95;
14 < 0 > Date:1935-55 Date:1965-85 =[100 %]=> < 0 > S:M S:F Date:1945-65 Date:1955-75 Date:1975-95;
15 < 0 > Date:1935-55 Date:1975-95 =[100 %]=> < 0 > S:M S:F Date:1945-65 Date:1955-75 Date:1965-85;
16 < 0 > Date:1945-65 Date:1965-85 =[100 %]=> < 0 > S:M S:F Date:1935-55 Date:1955-75 Date:1975-95;
17 < 0 > Date:1945-65 Date:1975-95 =[100 %]=> < 0 > S:M S:F Date:1935-55 Date:1955-75 Date:1965-85;
18 < 0 > Date:1955-75 Date:1975-95 =[100 %]=> < 0 > S:M S:F Date:1935-55 Date:1945-65 Date:1965-85;
```

FIGURE 9. Luxenburger basis with confidence ≥ 0.75.

PROPOSITION 5.1
Suppose $Con(e^{\mathbb{K}}) = (X, Y) \in \mathfrak{B}(\mathbb{K})$, and let $Y \to Y_1$ belonging to $\mathcal{L}(\mathbb{K}, \gamma, \delta)$. Then

$$conf(Y \to Y_1) \leq \frac{r_{BS}(e^{\mathbb{K}}) \cdot |\{e^{\mathbb{K}}\}''| - 1}{r_{BS}(e^{\mathbb{K}}) \cdot |\{e^{\mathbb{K}}\}''|}$$

PROOF. Since $Con(e^{\mathbb{K}}) = (X, Y)$, it has $\{e^{\mathbb{K}}\}' = Y$ and $\{e^{\mathbb{K}}\}'' = X$. Suppose $r_{BS}(e) = u$, and let $C_0(X_0, Y_0) \prec C$ with which the maximum risk is achieved,

$$u = \frac{1}{|Y' \setminus Y_0'|}.$$

Since $Y_0' \subseteq Y'$, then $|Y' \setminus Y_0'| = |Y'| - |Y_0'|$. Therefore, $|Y_0'| = |Y'| - 1/u$.
Suppose that $Y \to Y_1$ is other Luxenburger basis rule. Since $u = r_{BS}(e, \mathbb{K})$, it has

$$\frac{1}{|Y'| - |Y_1'|} \leq \frac{1}{|Y'| - |Y_0'|}$$

Thus $|Y_1'| \leq |Y_0'|$, hence

$$conf(Y \to Y_1) = \frac{|Y_1'|}{|Y'|} \leq \frac{|Y_0'|}{|Y'|} = \frac{|Y'| - 1/u}{|Y'|} = \frac{u|Y'| - 1}{u|Y'|}$$

□

The above result can be generalized using the maximum basic semantic risk:

COROLLARY 5.2
If $Y \to Y_1 \in \mathcal{L}(\mathbb{K}, \gamma, \delta)$ then

$$conf(Y \to Y_1) \leq \frac{r_{BS,M}(\mathbb{K})|Y'| - 1}{r_{BS,M}(\mathbb{K})|Y'|}$$

PROOF. By definition of the Luxenburger basis, it has $(Y', Y), (Y'_1, Y_1) \in \mathfrak{B}(\mathbb{K})$ and $Y \prec Y_1$. Let $e \in Y'$ such that $Con(e) = (Y', Y)$. From Prop. 5.1 it follows that

$$conf(Y \to Y_1) \leq \frac{|Y'| - 1/r_{BS}(e^{\mathbb{K}})}{|Y'|} \leq \frac{|Y'| - 1/r_{BS,M}(\mathbb{K})}{|Y'|} = \frac{r_{BS,M}(\mathbb{K})|Y| - 1}{r_{BS,M}(\mathbb{K})|Y'|}$$

□

Finally, it can be justified that our proposed maximum risk, $r_{BS,M}$, satisfies the above suggested principles. It will be a consequence of the following result:

THEOREM 5.3

$$conf(Y_1 \to Y_2) \leq \frac{r_{BS,M}(\mathbb{K})|Y'_1| - 1}{r_{BS,M}(\mathbb{K})|Y'_1|}$$

PROOF. The confidence has three properties that will be used (see e.g. [21]):

1. $conf(Y_1 \to Y_2) = conf(Y''_1 \to Y''_2)$
2. $conf(Y_1 \to Y_2) = conf(Y_1 \to Y_1 \cup Y_2)$
3. $conf(Y_1 \to Y_2) \cdot conf(Y_2 \to Y_3) = conf(Y_1 \to Y_3)$

By using (1), (2) it has

$$conf(Y_1 \to Y_2) = conf(Y''_1 \to Y''_1 \cup Y''_2)$$

Please note that $Y''_1 \cup Y''_2$ is closed:

$$(Y''_1 \cup Y''_2)'' = (Y'''_1 \cap Y'''_2)' = (Y'_1 \cap Y'_2)' = Y''_1 \cup Y''_2$$

It is therefore the intent of a concept

$$((Y''_1 \cup Y''_2)', Y''_1 \cup Y''_2) \leq (Y'_1, Y''_1)$$

Let us consider a descending path within the Hasse diagram from (Y'_1, Y_1) to that subconcept. In this path there exists $(X_0, Y_0) \in \mathfrak{B}(\mathbb{K})$ such that

$$((Y''_1 \cup Y''_2)', Y''_1 \cup Y''_2) \leq (X_0, Y_0) \prec (Y'_1, Y''_1)$$

(see Figure 10). By (3),

$$conf(Y''_1 \to Y''_1 \cup Y''_2) = conf(Y''_1 \to Y_0) \cdot conf(Y''_0 \to Y''_1 \cup Y''_2)$$

Confidence is less than 1, so it can bound by substituting the second factor with 1 and by applying Corollary 5.2 to the first one (as it belongs to the Luxenburger base):

$$conf(Y''_1 \to Y''_1 \cup Y''_2) \leq \frac{r_{BS,M}(\mathbb{K})|Y'''_1| - 1}{r_{BS,M}(\mathbb{K})|Y'''_1|} = \frac{r_{BS,M}(\mathbb{K})|Y'_1| - 1}{r_{BS,M}(\mathbb{K})|Y'_1|}$$

□

The confidence bound shown in the previous result allows us to argue that the basic semantic risk satisfies the required principles. Roughly speaking:

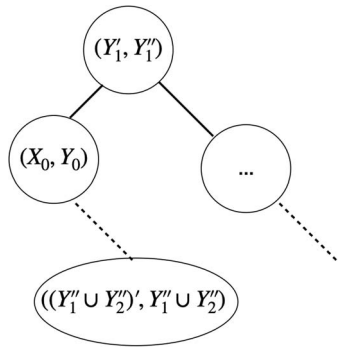


FIGURE 10. Sketch considered in the proof of Thm. 5.3.

- If the risk is high, $r_{BS,M}(\mathbb{K}) \approx 1$, the upper bound of Thm. 5.3 is close to 1. Thus dataset owner has not a good confidence estimation (PLC).
- Let e, $Con(e) = (X, Y) \in \mathfrak{B}(\mathbb{K})$ being Y the target of attack. It has that $r_{BS}(e, \mathbb{K}) \geq 1/|Y'|$. If the risk decreases toward this, $r_{BS}(e, \mathbb{K}) \approx 1/|Y'|$, the confidence is approaching 0, which makes the rule useless for the attacker (PDC).

6 Semantic risk management

In [7] a procedure to maintain low semantic risk was sketched. The study carried out in this paper can enrich the procedure, which would be as follows:

1. By exploring the data, the attributes that require a risk analysis are found. This is made by *conceptual selection* (something similar to what is exposed in [9] to design explanations in formal AI models).
2. The risk associated with a particular record is computed if one wants to study specific individuals. The concepts containing such record(s) at risk are located.
3. With the results of this paper, the two possibilities of anonymization based on the elimination of information can be addressed:
   a. Record (object) elimination: it is analysed by simulating the effect on the lattice. Possible strategies for dealing with the records involved are also studied:
      - If it is not desired to remove some individuals at risk, noise could be introduced in the dataset variable values (e.g. Differential Privacy), and go back to step 1.
      - It is possible to analyse which maximal object set can be maintained while preserving low risk by using the techniques of Sect. 4.
   b. Attribute elimination: It is analysed whether it is worthwhile (even if there is some re-identification risk) by studying the conservative retraction (Sect. 5.1)

7 Conclusions and future work

A revision of classical metrics under the prism of FCA has been proposed. This revision is based on the consideration of formal concepts representing quasi-identifiers. The new metric exploits the lattice structure of the induced concepts; data at risk can be detected by exploring the lattice. Other

approaches to the problem are based on statistical modeling (such as the one based on copulas [26]), while ours is semantic in nature, working with concepts. We have shown that our proposal satisfies some basic principles in relation to data (rule) mining. (PDC, PLC of Sect. 5.2).

The impact of the elimination of attributes and concepts can be analysed considering implications and association rules. For this task, variable forgetting techniques on implications have been used [2, 5]. In addition, the upper bounds for confidence are set depending on the semantic risk. Future work will be directed to the analysis of the effect of scales on our formalization, and to the design of metrics to estimate the loss of knowledge that the elimination of attributes would produce, following the ideas of [6].

Acknowledgements

This work was supported by *Agencia Estatal de Investigación* project PID2019-109152GB-I00/AEI/10.13039/501100011033 and Universidad de Huelva project UHU-1266216.

References

[1] M. K. Afridi, N. Azam, J. T. Yao and E. Alanazi. A three-way clustering approach for handling missing data using GTRS. *International Journal of Approximate Reasoning*, **98**, 11–24, 2018.

[2] J. A. Alonso-Jiménez, G. A. Aranda-Corral, M. Joaquín Borrego-Díaz, M. Fernández-Lebrón and M.-J. Hidalgo-Doblado. A logic-algebraic tool for reasoning with knowledge-based systems. *J. Log. Algebr. Meth. Program*, **101**, 88–109, 2018.

[3] G. A. Aranda-Corral, J. Borrego-Díaz and J. Galán-Páez. Confidence-based reasoning with local temporal formal contexts. In *Proc. 11th Int. Conf. Artificial Neural Networks on Adv. Comput. Intell., IWANN'11*, pp. 461–468. Springer, Berlin, Heidelberg, 2011.

[4] G. A. Aranda-Corral, J. Borrego-Díaz and J. Galán-Páez. Complex concept lattices for simulating human prediction in sport. *Journal of Systems Science and Complexity*, **26**, 117–136, 2013.

[5] G. A. Aranda-Corral, J. Borrego-Díaz, J. Galán-Páez and A. T. Caballero. *On Experimental Efficiency for Retraction Operator to Stem Basis, Chapter 8*, pp. 73–79. Springer, 2019.

[6] G. A. Aranda-Corral, J. Borrego-Díaz and J. G. Páez. A model of three-way decisions for knowledge harnessing. *International Journal of Approximate Reasoning*, **120**, 184–202, 2020.

[7] G. A. Aranda-Corral, J. Borrego-Díaz and J. G. Páez. Estimating re-identification risk by means of formal conceptualization. In *14th Int. Conf. Computational Intelligence in Security for Information Systems and 12th Int. Conf. European Transnational Educational (CISIS-ICEUTE 2021)*. Advances in Intelligent Systems and Computing, vol. 1400, pp. 13–22. Springer, 2022.

[8] G. A. Aranda-Corral, J. Borrego-Díaz and J. G. Páez. Concept learning consistency under three-way decision paradigm. *International Journal of Machine Learning and Cybernetics*, **13**, 2977–2999, 2022.

[9] J. Borrego-Díaz and J. G. Páez. Knowledge representation for explainable artificial intelligence. *Complex & Intelligent Systems*, **8**, 1579–1601, 2022.

[10] Canadian Institute for Health Information. Best practice guidelines for managing the disclosure of de-identified health information. *Technical Report*. Canadian Institute for Health Information, 2011.

[11] R. Cole, P. Eklund and D. Walker. Constructing conceptual scales in formal concept analysis. In *Research and Development in Knowledge Discovery and Data Mining*, pp. 378–379. Springer, Berlin Heidelberg, 1998.

[12] C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, **54**, 86–95, 2011.
[13] T. Eiter and G. Kern-Isberner. A brief survey on forgetting from a knowledge representation and reasoning perspective. *KI*, **33**, 9–33, 2019.
[14] K. El Emam and L. Arbuckle. *Anonymizing Health Data: Case Studies and Methods to Get You Started*, 1st edn. O'Reilly Media, Inc., 2013.
[15] K. El Emam, E. Jonker, L. Arbuckle and B. Malin. A systematic review of re-identification attacks on health data. *PLoS One*, **6**, e28071–e28071, 2011.
[16] Federal Committee on Statistical Methodology Report on statistical disclosure limitation methodology. *Technical Report 12*. Federal Committee on Statistical Methodology, 2005.
[17] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*, 1st edn. Springer New York, Inc., Secaucus, NJ, USA, 1997.
[18] J. L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, **95**, 5–18, 1986.
[19] ICO Anonymisation: managing data protection risk code of practice. *Technical Report*. UK's Information Commissioner's Office, 2012.
[20] Institute of Medicine. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. The National Academies Press, Washington, DC, 2015.
[21] M. Kryszkiewicz. Concise representations of association rules. In *Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK, September 16–19, 2002, Proceedings*. Lecture Notes in Computer Science, vol. 2447, pp. 92–109. Springer, 2002.
[22] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, **29**, 1991.
[23] Office Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. *Technical Report*. U.S. Dept. Health & Human Services, Washington DC, 2012.
[24] Y. Qian et al. Local rough set: a solution to rough data analysis in big data. *International Journal of Approximate Reasoning*, **97**, 38–63, 2018.
[25] S. Ribeiro-Navarrete, J. R. Saura and D. Palacios-Marqués. Towards a new era of mass data collection: assessing pandemic surveillance technologies to preserve user privacy. *Technological Forecasting and Social Change*, **167**, 120681, 2021.
[26] L. Rocher, J. Hendrickx and Y.-A. de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, **10**, 2019.
[27] Y. Yao. The superiority of three-way decisions in probabilistic rough set models. *Information Sciences*, **181**, 1080–1096, 2011.
[28] Y. Yao. Three-way decisions with probabilistic rough sets. *Information Sciences*, **180**, 341–353, 2010.
[29] Y. Yao. An outline of a theory of three-way decisions. In *Rough Sets and Current Trends in Computing*, pp. 1–17. Springer, Berlin Heidelberg, 2012.
[30] T. Zhang, H.-h. Li, M.-q. Liu and M. Rong. Incremental concept-cognitive learning based on attribute topology. *International Journal of Approximate Reasoning*, **118**, 173–189, 2020.

Received 20 May 2022