IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 May 2023, accepted 28 May 2023, date of publication 12 June 2023, date of current version 26 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285407

RESEARCH ARTICLE

Enhancing Security and Privacy Preservation of Sensitive Information in e-Health Datasets Using FCA Approach

HEDI HAMDI[1], ZAKI BRAHMI[2], ALAA S. ALAERJAN[1], AND LOTFI MHAMDI[3], (Member, IEEE)

[1]Department of Computer Science, Jouf University, Sakkaka 72388, Saudi Arabia
[2]Computer and Information Sciences Department, College of Science and Arts at Al-Ola, Taibah University, Medinah 43522, Saudi Arabia
[3]School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT Leeds, U.K.

Corresponding author: Hedi Hamdi (hhamdi@ju.edu.sa)

This work was supported by the Deanship of Scientific Research at Jouf University Research under Grant DSR2020-04-2605.

ABSTRACT Advances in data collection, storage, and processing in e-Health systems have recently increased the importance and popularity of data mining in the health care field. However, the high sensitivity of the handled and shared data, brings a high risk of information disclosure and exposure. It is therefore important to hide sensitive relationships by modifying the shared data. This major information security threat has, therefore, mandated the requirement of hiding/securing sensitive relationships of shared data. As a large number of data mining activities that attempt to identify interesting patterns from databases depend on locating frequent item sets, further investigation of frequent item sets requires privacy-preserving techniques. To solve many difficult combinatorial problems, such as data distribution problem, exact and heuristic algorithms have been used. Exact algorithms are studied and considered optimal for such problems, however they suffer scalability bottleneck, as they are limited to medium-sized instances only. Heuristic algorithms, on the other hand, are scalable, however, they perform poor on security and privacy preservation. This paper proposes a novel heuristic approach based on Formal Concept Analysis (FCA) for enhancing security and privacy preservation of sensitive e-Health information using itemset hiding techniques. Our approach, named FACHS (FCA Hiding Sensitive-itemsets) uses constraints to minimise side effects and asymmetry between the original database and the clean database (minimal distortion on the database). Moreover, our approach does not require frequent itemset extraction before the masking process. This gives the proposed approach an advantage in terms of total availability. We tested our FCAHS heuristic on various reference datasets. Extensive experimental results showed the effectiveness of the proposed masking approach and the time efficiency of itemset extraction, making it very promising for e-Health sensitive data security and privacy.

INDEX TERMS Healthcare process data, security and privacy, sensitive itemsets, data anonymization and sanitization, formal concept analysis (FCA).

I. INTRODUCTION

In recent years, methods based on Knowledge Discovery in Database (KDD) have transformed multiple economic sectors such as manufacturing, transportation, and governance. Despite the fact that the field of health care has always been resistant to large-scale technological disruptions [1], these methods and techniques are now beginning to penetrate this field. Indeed, techniques such as Decision Tree, Random Forest, K-means Clustering, Support Vector Machine, Logistic Regression, Neural Network, Naive Bayes, and association rule mining, have recently shown promising results in versatile tasks such as diagnosing [2], prognosis [3], classification [4], constructing predictive models [5], and analyzing risk factors of various diseases [6].

As health systems become intelligent and ubiquitous, it is essential to protect their security and data confidentiality. However, although methods based on KDD perform well in

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

extracting and exploiting attribute associations in databases, sensitive or private information may still be exposed or inferred from related data as the exploration process moves forward [7], [8]. Indeed, in most applications of a health care system, the important data handled are usually contained in the electronic health record (EHR) and it is always considered as sensitive information that must be safely secured [9], [10]. In addition to sensitive patient or employee identity information that an EHR may contain, the combination of other attributes of the EHR with background knowledge of the process may also reveal other sensitive patient or employee information. For example, data from a blood test that is always performed by the same employee during a work shift, can reveal the identity of the concerned employee when combined with the execution time of this activity. Similarly, the combination of attributes such as time and date of admission, nature of diagnosis or treatment, age and language spoken could potentially identify a patient. Additionally, clinics and hospitals collaborate through a data sharing mechanism that is used to provide EHR data to access patient information. Similarly, various research institutes and hospitals use and share integrated data that is collectively constructed from individual information.

Existing EHR data sharing systems still face several challenges in e-Health systems [10], [11]. Indeed, privacy leaks and security threats may occur during the progress of data sharing; in particular, private or personal information could be disclosed if exchanged for money in an illegal market. The above issues lead us to ask a very trivial question which is: how do we sanitize medical records databases to secure information in healthcare systems? This important question has recently attracted much attention and has been the subject of several studies in the field of Privacy-Preserving Data Mining (PPDM). In health systems, the use of PPDM minimizes the disclosure of sensitive personal information and allows compliance with privacy constraints. Various techniques have been developed in the field of PPDM [12]. Among these techniques is sanitization in which confidential information regarding a patient's record is sanitized to distort the values of sensitive data by adding, subtracting or disturbing data with other means. In addition, hiding sensitive information often causes certain rules to be lost and artificial rules to appear as side effects of sanitization, mainly hiding failure, missing cost, and artificial cost. Many approaches have been proposed to sanitize an original database in order to hide sensitive information. However, selecting appropriate data for data sanitization with minimal side effects can be considered NP-hard optimization problems [12], [13].

In this paper, we propose a novel heuristic for sensitive itemset hiding using the Formal Concept Analysis paradigm [14], [15], [16]. Our proposed heuristic is termed FCAHS (FCA Hiding Sensitive-itemsets), that outperforms previous proposals in various ways. In particular, the main FCAHS's contributions are as follows:

- Hide all sensitive itemsets.
- Minimize side effects on non-sensitive itemsets.
- Keep the original database as much as possible. Unlike many other approaches that remove transactions, we keep transactions and hide only some sensitive items from sensitive transactions.
- To the best of our knowledge this is the first time the FCA concept is used in such application. As shown in this work, the FCA is key in selecting which sensitive items should be hidden from which transactions.

We have evaluated the performance for different hiding scenarios on different datasets observing various metrics, such as the effectiveness, the number of lost itemsets as a side effect, and runtime efficiency. As we shall see the experimental Section, our FCAHS approach has shown higher performance in successfully hiding sensitive information while preserving transaction semantics as well as better running time than previously proposed solutions.

This paper is organized as follows. Section II gives relevant works for Formal Concept Analysis applications, privacy-enhancing methods for securing healthcare data sharing environments and PPDM. Section III provides the most relevant concepts, notations, and definitions of hiding sensitive frequent itemsets while minimizing side effects, and then it states the problem that we are investigating in this work. Section V presents our FCA-Based sensitive itemsets hiding approach. In Section VI, the results of experiments for our approach are analyzed. Lastly, Section VII concludes the paper.

II. RELATED WORK

The relevant works for Formal Concept Analysis applications, privacy-enhancing methods for securing healthcare data sharing environment and PPDM are respectively reviewed and discussed in this section.

A. FORMAL CONCEPT ANALYSIS

Formal Concept Analysis (FCA) has been used in different domains such as healthcare [17], biology [18], chemistry [19], ontology engineering [20], functional magnetic resonance imaging(fMRI) scans [21], sentiments analysis [22], decision-making [23], elearning [24], criminal trajectories [25], terrorist threat [26], Breast cancer [32], eXplainable AI(XAI) [27], [28]. and others. The examples mentioned above and the list is far from exhaustive, show that FCA is a well-known approach in the literature. The set of selected examples range from simple problems to more complicated situations and show the main characteristics of the FCA-based approach. The diversity of the fields of application of FCA represents, in our opinion, the beginning of its applications in the field of privacy preservation. Thus, for FCA practitioners, privacy preservation represents a new example that enriches the FCA world. At the same time, from the perspective of the privacy preservation practitioner,

FCA could be the beginning of another effective technique to improve a better understanding of different issues. To the best of our knowledge, this is the first time FCA is used in the context of PPDM. Below we describe most relevant related work in the field.

B. PRIVACY-PRESERVING IN HEALTHCARE INFORMATICS

Despite the fact that health data offers enormous opportunities in various fields, maintaining the privacy of health data still poses several unresolved privacy and security challenges [29], [30], [31]. In the following, we present some well-established privacy models that are used to ensure privacy of health data. In particular, we focus on the two common PPDP technologies namely data anonymization and differential privacy and discuss their limitations and strengths.

1) DATA ANONYMIZATION

It is about modifying an original database by deploying generalization and deletion on its data, before sharing it as an anonymized database. The anonymized database could be studied instead of the original database. Some common data anonymization models to prevent privacy disclosure include:

- **k-anonymity [33]** k-anonymity has been developed with the aim of preventing identity disclosure. Indeed, in a table where a record has a certain Quasi-identifier (QID) value, there are at least $k-1$ other records in the same table that have the same QID value. Therefore, each record cannot be distinguished from at least $k-1$ other records with respect to the QID value in a k-anonymous table. In k-anonymity, therefore, no individual can be reidentified from published data with a probability greater than $1/k$. Although the k-anonymity model protects against identity disclosure, it remains vulnerable to attribute disclosure. Indeed, the deduction of the values of sensitive attributes from the published data remains possible and attacks such as the attack by homogeneity and the attack by background knowledge can always succeed. To remedy this and protect the value of the sensitive attribute, l-diversity and t-proximity have been proposed.
- **l-diversity** [34], this method depends on the range of sensitive attribute values. At least l distinct sensitive attribute values are required for each QID group. Some fictitious data may be added to achieve l-diversity if the number of distinct sensitive attribute values is less than the desired privacy parameter l. Adding fictitious data may further lead to excessive editing and may produce biased results in the statistical analysis. Moreover, when the global distribution of the sensitive attribute is skewed, the prevention of attribute disclosure is not guaranteed by l-diversity. Indeed, the attack by asymmetry and the attack by similarity are always possible.
- **t-proximity** [35], this method has been proposed in order to remedy the vulnerabilities mentioned above. It requires that the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the attribute in the global table be less than a threshold. This property prevents an attacker from performing an accurate estimation of sensitive attribute values and thus prevents their disclosure. However, this method only modifies the values of the sensitive attributes while all the QID values remain unchanged. Therefore, it makes identity disclosure possible. Moreover, in order to find the optimal solution, t-proximity deploys a brute-force approach to examine each possible partition of the table, and this takes enormous computation time and a complexity of $2^{O(n)O(m)}$.
- **δ-presence** [36] intended to address membership disclosure, δ-presence has been proposed to limit an attacker's confidence level to $\delta\%$ at most in inferring the existence of a targeted victim in published data.

2) DIFFERENTIAL PRIVACY

This model is characterized by its rigorous definition of privacy and its low computational load. For the past decade, it has been a de facto standard for a variety of data types in the privacy field. It ensures that adversaries cannot distinguish any pair of secrets by just observing the outputs and regardless of arbitrary background information. Its mode of operation is based on feeding individual data or database queries with well-calibrated noise. Several extensions from the original model [37] were developed and they resulted in many variants [38] for specific data scenarios, such as Metric DP [39], local DP [40], shuffled DP [41] and hybrid PD [42]. It has shown great effectiveness in protecting various sensitive information smartphone application usage [43], locations [44], surveys genome-wide association [45] and eye tracking data [46]. Differential privacy preserves the usefulness of low-sensitivity queries such as count queries, range queries, and predicate queries, since the presence or absence of a single record slightly changes the result. However, for very sensitive queries to a differentially private database could return extremely inaccurate results. Examples of high-sensitivity queries include calculating sum, maximum, minimum, averages, and correlation. Therefore, a differentially private database should provide highly biased results for more complex queries, such as calculating variance, skewness, and kurtosis.

3) PRIVACY-PRESERVING DATA MINING

In recent years, PPDM has been an important concern for data mining strategies. In effect, it can not only reveal important information but also hide sensitive information through the sanitization process. For healthcare systems, the EHR contains identifiable health data collected from patient information that is sensitive and confidential in nature and should not be disclosed. Typically, identifiable health data includes very sensitive attributes that can make up patient health reports, such as diagnoses and type of treatment undertaken. Therefore, preserving and securing personal or sensitive

information in medical data [47] remains an important topic especially in data sharing environment in cloud-assisted health systems [48] where the frequency of privacy leaks and security threats is relatively high. Indeed, as demonstrated in [51] the difficulties of retaining sensitive or private information, especially for medical datasets such as EHR in the shared environment are only increasing.

Authors of [49], introduced the distance measurements for the sanitized database. They took into account the number of updated items rather than the number of transactions. This distance is minimized by maximizing the occurrences of items of sensitive itemset. Constraints for maximizing item set occurrences and minimizing item modifications are defined using the positive and negative borders and the Apriori property. This work also proposes an approach for constraint reduction. When the constructed Constraint Satisfaction Problem (CSP) is not solvable, this approach removes a constraint and reconstructs the CSP iteratively until the CSP is solvable.

The authors in [50] implemented a cloud-based healthcare data sharing prototype using the number theory research unit to encrypt data collected from mobile and wearable devices. At the same time, and for the communication of the diagnosis of patients with similar diseases, they also presented a model of trust. A privacy preservation model for sharing medical records in cloud computing system has been presented in [55]. It combines both statistical analysis and cryptography, thus providing several paradigms of balance between the use of medical data and the protection of privacy. This model first uses the vertical partition to publish medical data. The authors of [56] present a scheme for sharing medical data that ensures the preservation of confidentiality in a given period of time by the possibility of grouping certain people in multimedia systems based on the cloud. A usable randomization algorithm for shared and published medical data has been developed in [57]. This algorithm can handle different types of datasets (i.e. categorical or numeric), and the published dataset will be independent of the adversary's background knowledge which helps reducing and neutralizing the risk of re-identification. The model proposed in [51] adds the system public parameters and moves the partial encryption computation to the offline tasks, thus eliminating the majority of the computational task and hence improving the computational efficiency for data processing. In [52], the authors proposes a multi-type approach to privacy-aware prediction of health data based on locality-aware hashing that can achieve a good trade-off between prediction accuracy and privacy preservation.

The work presented in [53], introduces the concept of multiple support thresholds for keeping sensitive information private In Cyber-Physical Systems (CPS), especially in human-in-the-loop situations (also known as HitLCPS). However, this approach which is based on a genetic algorithm (GA) does not give better performance than the conventional algorithms of the traditional Greedy PPDM approaches.

In [54], the authors proposed two techniques for frequently extracting sets of elements from horizontally partitioned datasets while preserving privacy: Protocol A - CCBR Dependent Computation and Protocol B - Data Owner Dependent Computation. However, the efficiency of the suggested homomorphic encryption scheme needs improvement.

The authors in [58] presented the state-of-the-art privacy preservation algorithms that are used in e-health clouds. This study showed that there is no relevant work to manage medical data in the shared environment using FCA-based model. Moreover, none of the existing algorithms considers the user-centered multi-threshold of attributes as the major consideration in PPDM, which will be the major contribution of the developed model.

III. PRELIMINARIES AND PROBLEM STATEMENT

This section, first presents the most relevant concepts, notations, and definitions of hiding sensitive frequent itemsets while minimizing side effects. Thereafter, it states the problem that we are investigating in this work. Let's start with an example of a transactions database.

A. CONCEPTS AND NOTATIONS

Consider $A = \{i_1, i_2, \cdots, i_m\}$ a finite set of r distinct items and $T = \{T_1, T_2, \cdots, T_n\}$ A database consisting of a set of transactions. A transaction T_q of T is a subset of A, with a unique identifier q, called Transaction IDentifer (TID).

TABLE 1. An example of transaction database.

TID	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Items	a, c	a, c,d,e	c,d	b, e	a, c, d, e	d, e	c	a, b	a, c	c, d

We assume that users or experts can manually set a minimum support threshold δ which can vary between 0% and 100%. The table 1, which contains 10 transactions will serve as an example for the following description. It is noted that each element is represented by a specific letter.

a: SUPPORT COUNT OF A FREQUENT ITEMSET

For an itemset i in a database D, the number of transactions made up of i defines the support count of this itemset. Consider that the product of the minimum support threshold and the number of transactions in the database, give the minimum support count. The support count of a set of frequent items f is greater than the minimum support count in the database and can be defined as:

$$sup(f) \geq \mid D \mid \times \delta \tag{1}$$

Example: Let's say the minimum support threshold δ is set at 30%. Thus, the minimum support count is calculated as $10 \times 0.3 = 3$. According to Table 1, the itemset $\{c, d\}$ appears in four transactions, so the support count for this itemset is $sup(\{c, d\}) = 4$. Therefore, for this database, $\{c, d\}$ is considered a frequent itemset because of the high support count.

b: SENSITIVE ITEMSETS

A set of itemsets
$SI_s = \{S_1, S_2, \cdots, S_p\}$ is said to be sensitive if and only if:

$$\begin{cases} SI_s \subseteq FI_s : \\ \forall s_i \in SI_s, s_i \text{ must be hidden in the database D.} \end{cases}$$

The main purpose of PPDM is to hide as much sensitive information as possible. That said, it is also used to minimize side effects of the pruning process, not just to prune sensitive information from a database. Besides, the major side effects of the pruning process can be categorized as follows:

- Fail To be Hidden(F-T-H): hiding failure or the inability to hide certain sensitive information.
- Not To be Hidden(N-T-H): missing cost or the hiding of important but non-sensitive information.
- Not To be Generated(N-T-G): artificial cost or the introduction of artificial information.

In what follows, we discuss the definitions, explanations and formal relationships between these three side effects. Let D' be a sanitized database, which was generated by deleting some transactions/itemsets from an original database D. FI_s is the set of frequent itemsets in dataset D, FI'_s is the set of frequent itemsets in the sanitized database D', SI_s is the set of sensitive itemsets that needs to be hidden and $\sim SI_s$ is the set of non-sensitive frequent itemsets in D.

c: FAIL TO BE HIDDEN(F-T-H)

this side effect is caused by the failure to hide some sensitive information, it is defined as the number of sensitive itemsets still appearing in the sanitized database D', and it is denoted by α, which is:

$$\alpha = |SI_s \cap FI'_s| \tag{2}$$

d: NOT TO BE HIDDEN (N-T-H)

This side effect is the number of non-sensitive itemsets hidden in the sanitized database D', it is denoted by β, which is:

$$\beta = | \sim SI_s - FI'_s| = |SI_s - FI_s - FI'_s| \tag{3}$$

e: NOT TO BE GENERATED(N-T-G)

This side effect is defined by the number of infrequent itemsets in the original database D, which were generated as frequent in the sanitized database D'. It is denoted by γ, which is:

$$\gamma = |FI'_s - FI_s| \tag{4}$$

B. PROBLEM STATEMENT

Given a database D with a set of sensitive itemsets $s_i \in SI_s$. The goal is to generate a sanitized database D' from D by hiding sensitive itemsets such that the support counts for all sensitive itemsets $s_i \in SI_s$ will be less than the minimum support count, namely:

$$sup(s_i) < \delta \times |D| \tag{5}$$

To assess the quality of our sanitization approach, we use as standard measures, the three aforementioned side effects as follows:

- A high F-T-H number means that too many sensitive patterns are still in the sanitized database.
- A high number of N-T-H indicates that some important information may be missing from the sanitized database.
- Finally, if the number of N-T-G is too high, it implies that a significant amount of artificial and meaningless information may have been generated by the sanitization process. It is quite possible that important non-sensitive information will be hidden by the sanitization process if the amount of sensitive information that needs to be hidden is very large.

Furthermore, the minimum support count will be reduced due to the reduction in the number of sensitive items following the hiding of certain items from the database. This causes several sets of infrequent items to become frequent as a result of the sanitization process. Thus, there is a trade-off relationship between F-T-H, N-T-H and N-T-G side effects. It is an NP-hard problem to find a solution that minimizes the three side effects.

For an efficient sensitive itemset hiding solution, the following goals should be achieved on the sanitized database:

1) Goal 1: Minimizing the modification of the original database D as possible. This can be expressed by (see eq. 6):

$$min(\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}); \tag{6}$$

$x_{ij} = 1$ if the items i is hidden from the transaction j, and 0 otherwise.

2) Goal 2: Hiding all sensitive itemsets. This goal can be achieved while keeping the support count of all sensitive itemsets less than the minimum support count in the sanitized database.
3) Goal 3: Hiding super-sets of sensitive itemsets. This goal is achieved by reaching the first goal according to the Apriori property.
4) Goal 4: keeping all non-sensitive frequent itemset in the sanitized database. We can achieve this goal by (Eq. 7):

$$\forall s_i \in \sim SI_s,\ sup(f) \geq | D | \times \delta \tag{7}$$

5) Goal 5: Neither new itemset is generated in the sanitized database. Nevertheless, approaches that act only on items, by removing items from the original database, and keeping all transactions, naturally accomplish this goal.

IV. FCA BASIC CONCEPTS

We remind the mathematical foundations of FCA-approach as they are basic for this work. We give the following definitions:

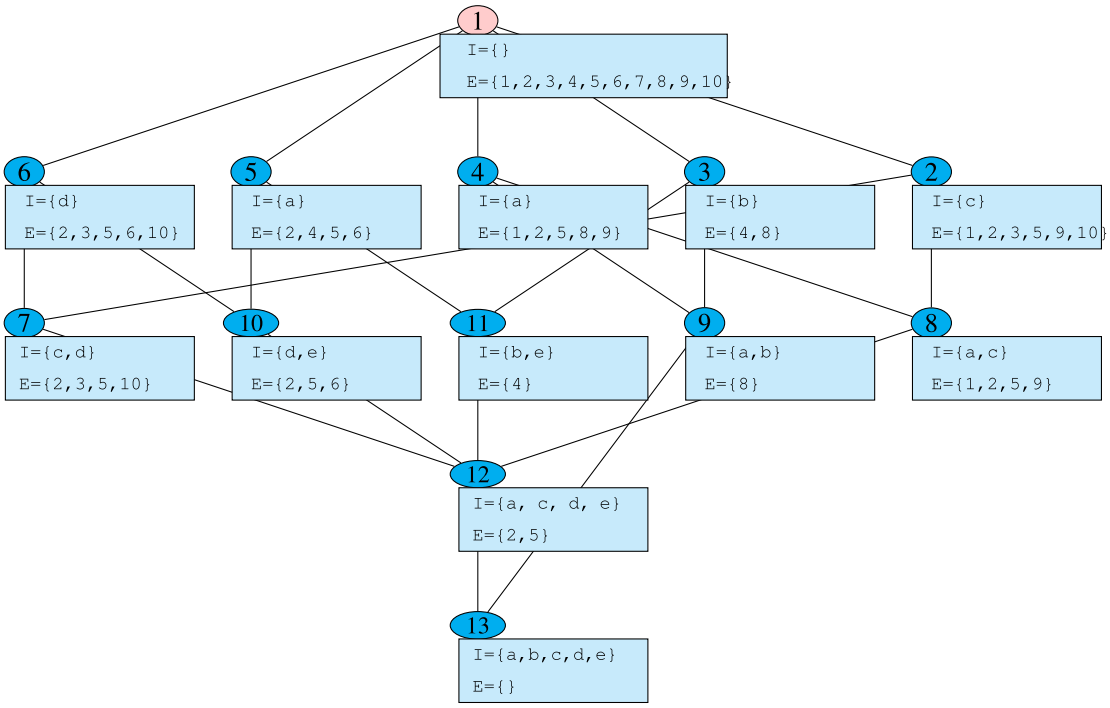

FIGURE 1. The Hasse diagram corresponding to the Galois lattice related to the formal context of Table 1.

TABLE 2. *FC* = (*A*, *T*, *I*).

	a	b	c	d	e
t_1	×		×		
t_2	×		×	×	×
t_3			×	×	
t_4		×		×	×
t_5	×		×	×	×
t_6				×	×
t_7					
t_8	×	×			
t_9	×		×		
t_{10}			×	×	

A. FORMAL CONTEXT

A formal context is a triplet (O, A, I) for which O is a set of objects, A is a set of attributes (or properties) and $I(P(O), P(A))$ a binary relation between O and A. R associates an object to a property: $(o, a) \in I$ when "o has the property a" or the property a is applied to the object o.

In our items hiding problem, objects are transactions and properties are items. The incidence relation indicates the transaction for which the item appears. Table 2 represents the binary relation of the formal context associated to the database presented in Table 1 where the set of transactions $T = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}\}$ and the set of items $A = \{a, b, c, d, e\}$.

B. GALOIS CONNECTION

Given a formal context $R(A, T, I)$, we define two functions f and g making it possible to express the correspondences between the subsets of objects $P(A)$ and the subsets of attributes $P(T)$ induced by relation R, as follows:

- f is the application which with any element $a \in A$ associates $(a) = \{t \in T \mid (a, at) \in I\}$,
- g is the application which with any element $t \in T$ associates $g(t) = \{a \in A | (a, t) \in I\}$. These two applications constitute the Galois correspondence of the context R.

C. CONCEPT

For our proposed method, a formal concept is like a transaction group. It connects a set of transactions (extent) to a set of items (intent). Indeed, a Formal Concept generalizes the notion of itemset, since it considers the itemset (as the intent) and the support (as the cardinality of the extent). Thereby, it's easy to determine the set of frequent itemsets. As an example, in Figure 1. if we consider 3 as the min-support count value, the itemset $\{a, c\}$ is frequent while $|\{1, 2, 5, 9\}| \geq$ min-support count.

D. GALOIS LATTICE (LATTICE OF CONCEPTS)

The set ł of all formal concepts, provided with order relation

$$\leq_l: (a2, t2) \leq_l (a1, t1), \Leftrightarrow t1 \subseteq t2$$

(or $a2 \subseteq a1$), is a complete lattice and is called Galois lattices (or formal concepts) of the context (A, T, I). The graphical representation of a Galois lattice is called the Hasse diagram. Figure 1 shows the Hasse diagram corresponding to the Galois lattice related to the formal context of Table 1.

According to the definition of order relation, we can say that the formal concept (a_1, t_1) is the super-concept of (a_2, t_2).

V. FCA-BASED SENSITIVE ITEMSETS HIDING APPROACH

Itemset hiding problem aims to generate a sanitized database by hiding sensitive frequent itemsets while minimizing side effects, non-sensitive frequent itemsets were preserved, ghost itemsets were not generated and dataset distortion is minimum. This problem is an NP-hard problem. In his paper, we propose a heuristic FCA-based approach, called FCAHS, to hide sensitive itemsets from a set of transactions. To keep the originality of the database as much as possible, unlike many other approaches that remove transactions, we keep transactions and hide only some sensitive items from sensitive transactions. The key idea behind the proposed approach is to use the FCA method's strength to select which sensitive items should be hidden from which transactions. The choice of FCA is motivated by the following advantages:

- The mathematical foundation of the FCA method makes it a robust approach to be used when resolving complex problems such as hiding sensitive frequent itemsets problems. Indeed, FCA is a solid mathematical framework to manage information based on logic and lattice theory.
- Lattice generated by FCA algorithm covers all possible itemsets from the set of items I. Indeed, a Formal Concept generalizes the notion of itemset, since it considers the itemset (as the intent) and the support (as the cardinality of the extent).
- From formal concepts it's easy to determine the support of any itemsets and afterward frequent itemsets.
- The hierarchical structure of Galois-lattice makes the navigation among formal concepts an easy process and then the navigation between itemsets and their supersets. This hierarchical navigation driven by the partial order operator among formal concepts can improve, considerably, the response time of the solution. For example, if a formal concept contains a sensitive itemset becomes non-frequent implies that all these super-sets are also non-frequent. This follows from the sub-concept/super-concept relationship.

Furthermore, to understand the proposed solution, we need first to mention the necessary used definitions.

A. DEFINITIONS

Top super-concept: A formal concept A is a Top sub-concept (*topConcept*) of concept B if the *intent* of B includes the *intent* of A and the number of items on *intent*(A) is the minimum over all others sub-concept of B. The set of Top sub-concepts of B can be defined as follows:

$$\begin{aligned} topConcept(B) = \{A, intent(A) \subseteq intent(B) \text{and} \\ |intent(A)| \leq |intent(X)|, \\ \forall X \in sub - concept(B)\} \end{aligned} \tag{8}$$

Sensitive concepts: a formal concept C is called sensitive if and only if the intent of C contains at least a sensitive Itemset:

$$iSsensitive(C) = \begin{cases} \text{True if } \exists\, a \in SI \subseteq intent(C) \\ \text{False otherwise.} \end{cases} \tag{9}$$

To be hidden: The number h of transactions form which items belong to a sensitive itemset s will be hidden is defined as follows:

$$h = support(s) - minsupport + 1 \tag{10}$$

We add the value one (+1) to h, in order ensure that the support of s after hiding is less than the minimum support count $| D | \times \delta$.

Sensitive transaction: a transaction t is called sensitive for the attribute $a \in SA$ ($t \in sensitive(a)$) if and only if hidden a from t generate the loss of a non-sensitive itemset. The transaction $t \in extent(C)$ means that the formal concept C is sensitive ($iSsensitive(C) = True$). A non-sensitive itemset is lost if its support change or becomes less than δ. In lattice, this non-sensitive itemset is a sub-concept of the formal concept containing the attribute a. This, in order to accomplish the second goal which is: *All non-sensitive frequent itemsets appear in the sanitized database with the same support or greater than the min-support count.*

For Example, suppose that $\{c, d\}$ is a sensitive Itemset in Figure 1. The formal concept 7 ($C7$) is sensitive because $\{c, d\}$ includes on $intent(C7)$ ($= \{c, d\}$). A sensitive transaction of c or d must belong to the extent of $C7(= \{2, 3, 5, 10\})$. Regarding this example, if δ is 3, the sensitive transaction of the item d is transaction 2 (or 5). In fact, removing item d from transaction 2 implies that the support count of $(\{d, e\}) = 2$ is less than the min-support count (= 3). Indeed; the set of sensitive transactions of all sensitive attributes will be defined as follows: $ST = \bigcup_{a \in SA} sensitive(a)$, where $SA = \{a \in A : a \in \bigcup_{s \in SI} S\}$.

B. PROPOSED FCA-BASED APPROACH

To deal with the NP-hardness of the PPDM problem with need, our FCA-based approach is based on a set of components aiming to reduce the execution time and generate a near-optimal solution. Figure 2 and Algorithm 1 present an architectural overview of the proposed solution.

1) PRE-PROCESSING

To reduce the complexity of the PPDM problem and enhance the execution time of the proposed solution, a pre-processing step is proposed which consists of two phases:

a: ORIGINAL DATABASE PRUNING

To avoid unnecessary processing which can be time consuming, we prune the original database by removing each transaction that doesn't contain an item belonging to the set SA. Thus, these transactions are not considered in our hiding process. Our main idea is to remove sensitive items from sensitive transactions.

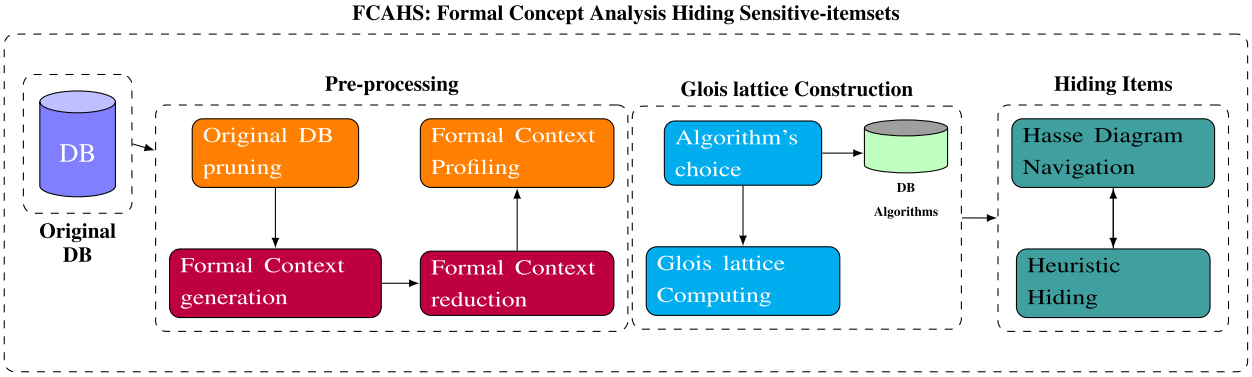

FIGURE 2. FCA-based architecture overview.

b: SMART GALOIS LATTICE CONSTRUCTION

The FCA-based approach is mainly based on Galois lattice manipulation. In fact, the computation time is mostly depending on Galois lattice generation which is, in general, expensive in time consumption. To overcome this issue, we opt for two ideas: Original formal context reducing and formal context profiling.

From the original database T, it's easy to generate the Formal Context $FC = (A, T, I)$. I is the binary relationship between A and T, it is set to 1 if attribute a_i appears in the transaction t_j for its execution, 0 otherwise.

Since the Formal context is the input of any Galois lattices construction algorithm, its size has a significant influence on the structure of formal concept lattice as well as time complexity when building the lattice, hence, it is a good idea to compact the original formal context. Many techniques were proposed for this purpose. In this paper, we opt for the method proposed by [59] which is based on the idea of lines and/or columns junction. If two objects/transactions *t1* and *t2* having the same set of attributes ($g(t1) = g(t2)$) then *t1* and *t2* can be merged to one single object. Dually, if for two attributes a and b appear on the same transactions/objects ($f(a) = f(b)$), then a and b can be replaced by one single attribute. Thereby, the number $|A|$ and $|T|$ can be reduced and therefore the formal context, which can significantly reduce the running time of Galois lattice algorithms.

The second step to enhance the response time of the proposed solution is to exploit the profile of the formal context when generating the Galois lattice structure. To identify the most convenient algorithm to be used for the generation of Galois-lattice, it's necessary to process by context formal profiling. Indeed, the profile of a context formal is mainly related to its density and its size. To be able to classify a context formal as dense or sparse, the number of 0s and 1s on the matrix is used. A sparse formal context is a matrix in which most of the elements are zero, and dense otherwise. The sparsity of the context formal F can be computed using eq. 11:

$$Sparsity(F) = \begin{cases} \text{Sparse } if \ \dfrac{number\ of\ 0}{|A| * |T|} < \beta \\ \text{Dense , } otherwise. \end{cases} \tag{11}$$

Based on the type of formal context, the lattice construction algorithms are recommended according to Table 3. For example, when the formal context is **small** and **sparse** Godin algorithm [59] is a good choice in this case.

2) GALOIS LATTICE GENERATION

This step aims at choosing the suitable Galois lattice algorithm to compute the Hasse diagram which describes the Galois lattice structure. Basically, we are motivated by the recommendation proposed in [60]. The efficiency of Galois lattice construction algorithms mainly depends on of density/sparseness of underlying formal contexts. We summarize the recommendation in Table 3. Once the suitable Galois lattice algorithm is selected, we proceed by computing all formal concepts and their relationships.

3) HIDING ITEMS

The main phase of the proposed approach is the hiding of sensitive itemsets, which is, basically, achieved by navigation on the Galois lattice structure. In the following, we describe all steps of the hiding process:

1) Compute the set SC of Top sub-concepts of all sensitive itemsets:
   $SC = \bigcup_{s \in SI} topConcept(s)$
2) Sort SC by extent cardinality of formal concepts. The idea behind this is to start processing itemset $Y (= head(SC))$ with maximum support by choosing from what transaction items will be hidden. Starting by hidden Y that has maximum support can minimize losing non-sensitive Itemsets, because these items with a high probability compose, with others items, many non-frequent itemsets. Thereby, hidden Y form transaction that not contains any (or minimum) non-frequents itemset can minimize the loss of a non-frequents itemset. According to the hierarchy feature of the lattice, formal concepts with the minimum number of attributes (items) must have the maximum number of transactions (objects).
3) Choose random Item $a \in int(head(SC))$ and find the subset of non-sensitive transactions $Th \in T$ such that hidden a from these transactions doesn't lose any

Algorithm 1 Formal Concept Analysis Hiding Sensitive-Itemsets

Input:
- SI: set of sensitive itemsets
- D: original database
- A: set of items
- T: set of transaction
- δ: minimum support threshold

Output: D': sanitized database

```
/* Now this is an if else
   conditional loop                          */
1  begin
   /* Preprocessing                          */
   /* pruning the original database
      D by removing transaction that
      doesn't contain an item ∈ SI
      */
2  PD ← pruningDataBase(D, SI )
3  FC ← generatedFormalContext(A, T)
   /* reduce the formal context FC
      as described in section B.1.b
      */
4  FCR ← reduceFormalContext(FC)
   /* based on the density of FCR we
      select the appropriate
      algorithm as described in
      section B.1.b                          */
5  FCA_Algorithm ←
    FCA_Algorithm_Selection(FCR)
6  Lattice ←
    compute_Galois_Lattice(FCA_Algorithm, FCR)
   /* Hidding Items                          */
   /* the set of selected concept
      are in the level 1 of Lattice
      */
7  SC ← select_Top_Subconcepts(Lattice, SI)
8  sort(SC)
9  Y ← head(SC)
10 while SC ≠ Null & |intent(Y)| < |T| * δ do
11    a ← chooseRandomItem(intent(Y))
      /* according to step 3 of the
         hiding phase                        */
12    Th ← getTh(T, a)
13    for c ∈ sup_concept(Y) do
         /* remove/hiding from the
            concept c Th                     */
14       remove(c, Th)
15       remove(SC, Y)
16       Y ← head(SC)
```

(or minimum) itemset. Th is the set of transactions belonging only to extent of the formal concept containing a, and excludes transactions belonging to the super-concept of $head(SC)$ that their

$$|ext(c)| < minSupport$$

Thus, to avoid the loss of non-sensitive itemset according to goal 2.

$Th = ext(head(SC)) - E$;

$E = \bigcup_{c \in superConcept \text{ and } |ext(c)| < |D| \times \delta} |ext(c)|$;

Selecting a random item from $int(head(SC))$ is argued by navigation operators of the Galois lattice structure. Indeed, if two items a and b belong to the same formal concept C, according to partial order relation in Galois lattice, a and b share the same set of super-concepts. So, they have the same impact on hiding one of them. According to example 1:

$sub - concept(\{c, d\}) = \{C6, C2\}$. The formal concept $C6$. $Th = ext(C2)^{\smile} ext(C10), \{2, 3, 5, 10\}^{\smile} = \{2, 3, 5, 6, 10\}\{2, 5, 6\} = \{3, 10\}$. Indeed, we can hide D from 3 and/or 10. Given that the optimal number of transactions from these attributes that need to be hidden is equal to 2 ($sup(C, D) - \mid D \mid \times \delta = 4 - 3 + 1 = 2$), D will be hidden from transactions 3 and 10.

4) Update extent of all super-concept of $head(SC)$ by removing Th'.
5) If the first super-concept $\in SC$ contains a sensitive itemset it's $|intent| <\mid D \mid \times \delta$, stop algorithm. Indeed, according to the partial order feature of the Galois lattice, all other formal concepts containing sensitive itemsets become non-frequent.
6) Else $Y = head(SC)$, and go to 3.

C. COMPLEXITY ANALYSIS

In this section, we analyze the theoretical complexity of our sensitive itemsets hiding strategy, which depends on the steps described above: Galois lattice generation using an FCA algorithm, and the hiding of sensitive itemsets.

Regarding the first step, the complexity of the generation of Galois lattice is $\mathcal{O}(\Delta^2)$ where Δ denotes the maximum size of attributes list. In our context, attributes list is composed by the set of items $A : \Delta = |A|$. Indeed, the complexity of this step is analyzed as $\mathcal{O}(|A|)$. Although, this complexity can be minimized by the step of formal context reduction, in the worst case there no reduction.

The complexity of sensitive itemsets hiding phase depends, mainly, on the step of computing the set SC of Top sub-concepts of all sensitive itemsets, and the recursively navigation through the Galois lattice from each concept $c \in SC$ to the super-concept which has intent that verify: $|intent| < |D| \times \delta$. Computing SC can be made by selecting all concepts belongs to the first level of the Galois lattice, so the complexity is constant $\mathcal{O}(1)$. In the worst case the navigation of the lattice proceeded $|SC|$ time, while in the real word it's very hard to have this case. For each time the navigation consists on visiting, recursively, all super-concepts of the started concept $c \in SC$. This behavior is similar to the Depth-First Search Algorithm. The temporal complexity of DFS is $\mathcal{O}(V)$

TABLE 3. Galois lattice type and algorithms.

Context formal Type	Algorithms
Small and sparse	Godin [59]
Dense	**Norris** [61], **NextClosure** [62] and **Close by One** [63]
Average density	**Bordat** [64]

TABLE 4. Datasets properties.

Dataset	Transactions	Items	Minimum Support Count	Frequent Itemsets	Dataset type
T10I4D100K	100.000	870	500 (%0.5)	1073	Sparse
T40I4D100K	100.000	942	500 (%0.5)	1.286.037	Dense
Mushromm	8124	119	406(%5)	3.755.704	Dense
retail	88.162	16.470	440(%0.5)	581	Sparse
BMS1	59.602	497	60(%0.1)	3991	Sparse
BMS2	77.512	3.340	77 (%0.1)	24.143	Sparse

where V is the number of nodes. In our case the V is equal to the number of super-concepts of the concept c. Thereby, the temporal complexity of this step is evaluated to $\mathcal{O}(V * |SC|)$, in the worst case $|SC| = |A|$. So, temporal complexity of the proposal algorithm is analyzed as $\mathcal{O}(V * |A| + |A|)$.

VI. EXPERIMENTAL ANALYSIS

A. EVALUATION METRICS

As discussed in Section V, the aim of the sensitive itemset hiding problem is to transform the dataset in a way that sensitive itemsets were concealed, non-sensitive frequent itemsets were preserved, ghost itemsets were not generated and dataset distortion is minimum. These goals can be measured by the below metrics so-called side effects.

1) HIDING FAILURE (HF)

Hiding Failure is a side effect (failure to hide some sensitive patterns), which means sensitive itemsets remain frequent in the sanitized database. HF can be defined as the number of sensitive itemsets that appear in the sanitized dataset divided by the number that appeared in the original dataset, or simply the number of sensitive itemsets that exist at the same time on the original and sanitized dataset.

According to the algorithm of our approach, stop criteria is sensitive itemsets are hidden, it's clear to verify that our FCA-based solution FCAHS ensures that all sensitive itemsets are hidden. Therefore, HF = 0 for all scenarios.

2) NOT TO BE HIDDEN (NTH))

This side effect measures the apparition of new itemsets in the sanitized database that have not been in the original database. NTH can be computed by the ratio of itemsets that did not appear in the original dataset but appeared in the sanitized dataset to the itemsets that appear in both the original and the sanitized datasets.

Our FCA-based approach does not insert items on the original dataset given that the idea is only hidden items from transactions. Thus, ensure that NTH = 0 since it is not possible to produce new itemsets from the sanitized dataset. This is the same for the HISB algorithm [65], where the number of deleted items is identical.

3) DISSIMILARITY

As defined in [65], dissimilarity is a metric that measures the differences between the original and the sanitized dataset.

The main idea of our approach is to hide items from transactions, which gives dissimilarity between the original and sanitized dataset until the support of each sensitive itemset is less than the minimum support threshold. Based on the HISB algorithm and our algorithm, they delete almost the same number of items.

4) NOT TO BE GENERATED(NTG)

The side effect NTG is defined as the number of frequent itemsets in the sanitized database that was infrequent in the original database [66]. In another way, this measure is the number of lost itemsets. Given that it can be different from HISB algorithm, we conduct a deep comparison in the next Section.

B. DATASETS AND EXPERIMENTAL SETUP

To evaluate our algorithm, six different datasets obtained from[1] are used. Table 4 presents the characteristics of these datasets. Similar to HISB, our solution doesn't generate frequent itemsets discovered before the hiding process.

TABLE 5. Hiding scenarios.

Name Scenario	Number of sensitive itemsets	Size of the itemsets
$HS_{2.1}$	1	2
$HS_{2.2}$	2	2
$HS_{2.3}$	3	2
$HS_{3.1}$	1	3
$HS_{3.2}$	2	3
$HS_{4.1}$	1	4

To test the FCAHS and HISB algorithms, we conducted several experiments using the same scenarios proposed in [65]. These scenarios are presented in Table 5 try to hide 1-, 2-, 3- and 4- sensitive itemsets.

[1] http://fimi.uantwerpen.be/data/

TABLE 6. NTG side effect.

Scenario	T10I4100K	T40I10D100K	Mushroom	retail	BMS1	BMS2
	HISB/FCAHS	HISB/FCAHS	HISB/FCAHS	HISB/FCAHS	HISB/FCAHS	HISB/FCAHS
$HS_{2.1}$	0/0	1/0	0/0	0/0	0/0	0/0
$HS_{2.2}$	0/0	1/1	0/0	0/0	0/0	0/0
$HS_{2.3}$	0/0	1/0	0/0	0/0	1/0	1/0
$HS_{3.1}$	0/0	0/0	0/0	0/0	0/0	0/0
$HS_{3.2}$	0/0	0/0	0/0	0/0	1/1	1/1
$HS_{4.1}$	0/0	1/2	1/1	1/0	2/1	2/1

C. RESULT ANALYSIS

The performance evaluation of our proposed algorithm against the HISB algorithm with respect to the side effects as the number of lost itemsets (NTG), and running time in seconds is presented in this Section. The algorithms were implemented in the java language. All experiments were conducted on a PC running MS Windows 10 with an Intel i7-6500U CPU and 8 GB of RAM.

Table 6 gives NTG side effect evaluation regarding all datasets presented in Table 4. The value of each cell c_i is a comparison between HISB and FCAHS algorithms (HISB/FCAHS) regarding NTG.

As indicated in Table 6, both algorithms perform well regarding the number of lost items expressed by NTG side effects. In the majority of cases, HISB and FCAHS are similar except, in some cases, there is a slight difference. For example, in the scenario, $HS_{4.1}$ HISB performed better regarding the dataset T40I10D100K, but FCAHS is better regarding BMS1 and BMS2.

What can be concluded from the evaluation expressed in the Table 6:

- The sparsity of a dataset does not affect the number of lost items.
- In the majority of scenarios and datasets, both HISB and FCAHS algorithms perform well. This is due to the importance of the FCA-based approach and sibling itemset formal concept to hide sensitive itemsets.
- There is no significant difference in the performance of HISB and FCAHS algorithms regarding NTG side effects.
- For the majority of datasets, our algorithm FCAHS performs well when the size of the itemset is small. This is explained by the fact that the FCAHS algorithm is mainly based on formal concepts. So, when the size of items is small (e.g., example 1) implied that items of the concerned itemsets, mostly exist on one formal concept which is easy to proceed without losing any items.

To analyze the performance of algorithms regarding runtime, FCAHS and HISB algorithms are compared in detail. Results are shown in Figure 3 to 7 where both algorithms are compared regarding different scenarios and different datasets. FCAHS performs well compared to HISB algorithms for the majority of cases except for the dataset Mushroom (Figure 5, they are mostly similar. This is because the pre-processing phase will reduce significantly the running time of FCAHS algorithm.

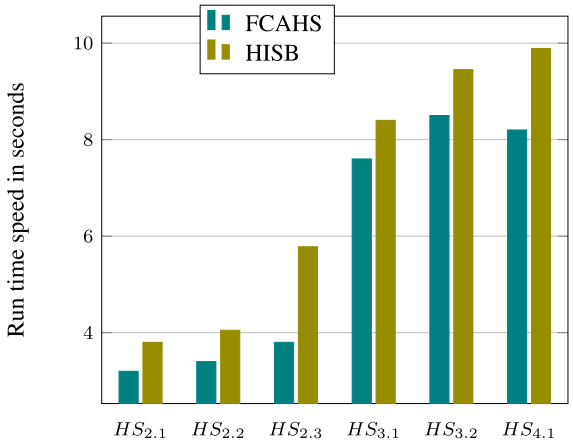

FIGURE 3. Comparison of the performance of FCAHS and FHISB algorithms in terms of execution time for different scenarios and Dataset T10I4100K.

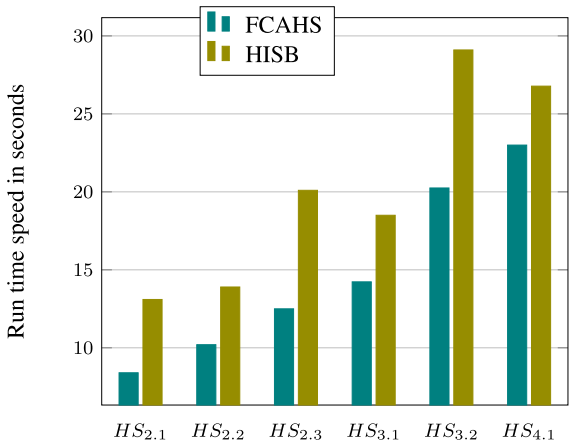

FIGURE 4. Comparison of the performance of FCAHS and FHISB algorithms in terms of execution time for different scenarios and Dataset T40I10100K.

D. DISCUSSION

First of all, we have to mention that our main goal was to provide a solution that hides sets of sensitive elements without suppressing transactions like most other approaches do. Naturally, we sought to compare our work to works that share the same objective while addressing the problem with other approaches. The most closely relevant existing works important to us and meet our comparison criterion are [49] and [65]. Since the authors of [65] compared their results

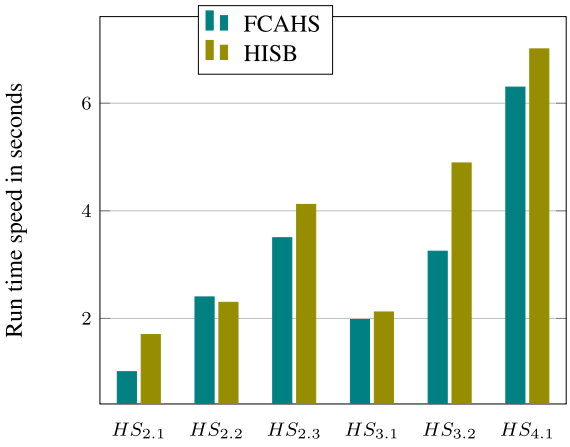

FIGURE 5. Comparison of the performance of FCAHS and FHISB algorithms in terms of execution time for different scenarios and Dataset Mushroom.

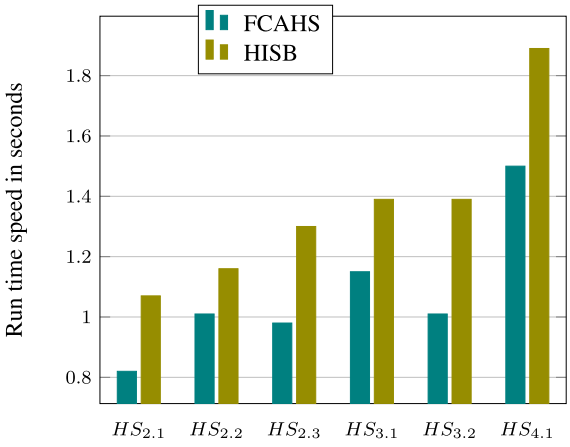

FIGURE 6. Comparison of the performance of FCAHS and FHISB algorithms in terms of execution time for different scenarios and Dataset BMS1.

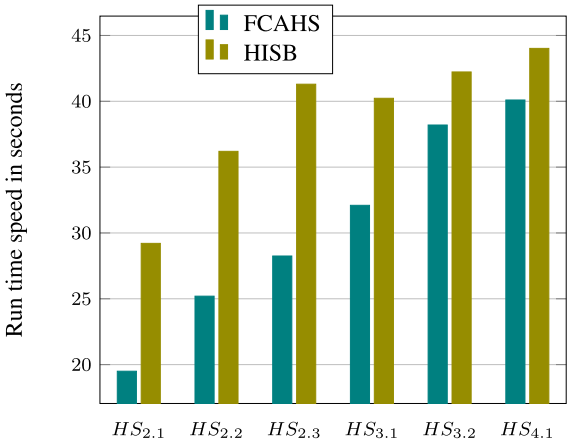

FIGURE 7. Comparison of the performance of FCAHS and FHISB algorithms in terms of execution time for different scenarios and Dataset retail.

to the results presented in [49] and they showed that their approach is better in terms of runtime where side effects such as sets of lost items are similar. It was wise to compare our work to [65] which, in addition to its superior results, it is the most recent. The analysis of the experimental results presented above, show that our approach to preserving privacy in the extraction of frequent items is more time-efficient than heuristic approaches while minimizing side effects on non-sensitive itemsets similar to that of exact approaches. Our approach uses the notion of formal concept and Galois lattice to obtain a good solution with a minimum of side effects such as the inability to hide certain sensitive itemsets and the generation of new itemsets in the sanitized database. Compared to a the benchmark algorithm [65], experiments revealed that pre-suppression of frequent item set extraction on the original database, combined with formal concepts and Galois lattice, is time efficient while side effects are minimized. This is due to the pre-processing phase which significantly minimizes the response time of the algorithm by reducing the initial database and the formal context. Since the time efficiency of FCAHS is driven by the algorithm used to generate the Galois lattice, formal context profiling shows its efficiency using the appropriate Galois lattice generation algorithm. Moreover, the experimental results showed that the proposed FCAHS algorithm can effectively hide all sensitive information and achieves good performance regarding HF, NTH, NTG and dissimilarity.

VII. CONCLUSION

This paper presents an FCA-based approach for hiding sensitive itemsets in transactional datasets. Our main goal is to hide sensitive itemsets without removing transactions like other approaches. We benefit from the notion of formal concept and Gallois lattice to obtain a good solution with minimal side effects such as failure to hide some sensitive patterns and apparition of new itemsets in the sanitized database. Experimentation showed that the FCA approach is an efficient solution. Compared with a reference algorithm, experiments revealed that removing prior mining of frequent itemsets on the original database combined with formal concepts and Galois lattice is time-efficient while side effects are minimized. This is due to pre-processing phase which significantly minimizes the response time of the algorithm by reducing the initial database and the formal context. Given that the FCAHS's time-efficiency is guided by the algorithm used to generate the Galois lattice, formal context profiling shows its effectiveness to use the suitable Galois lattice generation algorithm. Also, experimental results showed that the proposed algorithm FCAHS can effectively hide all sensitive information and obtains good performance regarding HF, NTH, NTG and, Dissimilarity. For future work, since the response time is the major issue in stream applications and health databases are usually very large in the real world, distributed version of the Galois lattice generation algorithm on cloud computing infrastructure can further improve its performance, making it highly appealing in such environments.

REFERENCES

[1] M. Tavakoli, J. Carriere, and A. Torabi, "Robotics, smart wearable technologies, and autonomous intelligent systems for healthcare during the COVID-19 pandemic: An analysis of the state of the art and future vision," *Adv. Intell. Syst.*, vol. 2, no. 7, Jul. 2020, Art. no. 2000071.

[2] C. W. Song, H. Jung, and K. Chung, "Development of a medical big-data mining process using topic modeling," *Cluster Comput.*, vol. 22, no. 1, pp. 1949–1958, 2019.

[3] G.-P. Diller, A. Kempny, S. V. Babu-Narayan, M. Henrichs, M. Brida, A. Uebing, A. E. Lammers, H. Baumgartner, W. Li, S. J. Wort, K. Dimopoulos, and M. A. Gatzoulis, "Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients," *Eur. Heart J.*, vol. 40, no. 13, pp. 1069–1077, Apr. 2019.

[4] U. R. Acharya, W. L. Ng, K. Rahmat, V. K. Sudarshan, J. E. W. Koh, J. H. Tan, Y. Hagiwara, C. H. Yeong, and K. H. Ng, "Data mining framework for breast lesion classification in shear wave ultrasound: A hybrid feature paradigm," *Biomed. Signal Process. Control*, vol. 33, pp. 400–410, Mar. 2017.

[5] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease," in *Information and Communication Technology for Competitive Strategies*. Singapore: Springer, 2019, pp. 447–454.

[6] M. G. Ahamad, M. F. Ahmed, and M. Y. Uddin, "Clustering as data mining technique in risk factors analysis of diabetes, hypertension and obesity," *Eur. J. Eng. Technol. Res.*, vol. 1, no. 6, pp. 88–93, Jul. 2018.

[7] J.-S. Lee and S.-P. Jun, "Privacy-preserving data mining for open government data from heterogeneous sources," *Government Inf. Quart.*, vol. 38, no. 1, Jan. 2021, Art. no. 101544.

[8] M. Sheikhalishahi, A. Saracino, F. Martinelli, and A. L. Marra, "Privacy preserving data sharing and analysis for edge-based architectures," *Int. J. Inf. Secur.*, vol. 21, no. 1, pp. 79–101, Feb. 2022.

[9] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges," *Egyptian Informat. J.*, vol. 22, no. 2, pp. 177–183, Jul. 2021.

[10] A. V. Deorankar and K. T. Khobragade, "A review on various data sharing strategies for privacy of cloud storage," in *Proc. 4th Int. Conf. Comput. Methodolog. Commun. (ICCMC)*, Mar. 2020, pp. 98–101.

[11] H. Jin, Y. Luo, P. Li, and J. Mathew, "A review of secure and privacy-preserving medical data sharing," *IEEE Access*, vol. 7, pp. 61656–61669, 2019.

[12] S. Sharma and S. Ahuja, "Privacy preserving data mining: A review of the state of the art," in *Harmony Search and Nature Inspired Optimization Algorithms*. Singapore: Springer, 2019, pp. 1–15.

[13] J. M.-T. Wu, G. Srivastava, A. Jolfaei, P. Fournier-Viger, and J. C.-W. Lin, "Hiding sensitive information in eHealth datasets," *Future Gener. Comput. Syst.*, vol. 117, pp. 169–180, Apr. 2021.

[14] R. Missaoui, L. Kwuida, and T. Abdessalem, Eds., *Complex Data Analytics With Formal Concept Analysis*. Cham, Switzerland: Springer, 2022.

[15] S. Ferre et al., "Formal concept analysis: From knowledge discovery to knowledge processing," in *A Guided Tour of Artificial Intelligence Research: AI Algorithms*, vol. 2. Cham, Switzerland: Springer, 2020, pp. 411–445.

[16] Y. Eslami, S. Ashouri, C. Franciosi, and M. Lezoche, "Knowledge extraction in cyber-physical systems meta-models: A formal concept analysis application," in *Proc. 3rd Int. Conf. Innov. Intell. Ind. Prod. Logistics*, 2022, pp. 129–136.

[17] I. Coyne, I. Holmström, and M. Söderbäck, "Centeredness in healthcare: A concept synthesis of family-centered care, person-centered care and child-centered care," *J. Pediatric Nursing*, vol. 42, pp. 45–56, Sep. 2018.

[18] F. Roberts, Ed., *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*. Cham, Switzerland: Springer, 2012.

[19] N. Quintero and G. Restrepo, "Formal concept analysis applications in chemistry: From radionuclides and molecular structure to toxicity and diagnosis," in *Partial Order Concepts in Applied Sciences*. Cham, Switzerland: Springer, 2017, pp. 207–217.

[20] R. Jindal, K. R. Seeja, and S. Jain, "Construction of domain ontology utilizing formal concept analysis and social media analytics," *Int. J. Cognit. Comput. Eng.*, vol. 1, pp. 62–69, Jun. 2020.

[21] D. Endres, R. Adam, M. A. Giese, and U. Noppeney, "Understanding the semantic structure of human fMRI brain recordings with formal concept analysis," in *Proc. Int. Conf. Formal Concept Anal.* Berlin, Germany: Springer, 2012, pp. 96–111.

[22] S. T. Li and F. C. Tsai, "A fuzzy conceptualization model for text mining with application in opinion polarity classification," *Knowl.-Based Syst.*, vol. 39, pp. 23–33, Feb. 2013.

[23] B. Long, W. Xu, and X. Zhang, "Double threshold construction method for attribute-induced three-way concept lattice in incomplete fuzzy formal context," *J. Eng.*, vol. 2020, no. 13, pp. 549–554, Jul. 2020.

[24] F. Pérez-Gámez, M. Ojeda-Hernández, Á. M. Bonilla, D. López-Rodríguez, and N. Madrid, "Using formal concept analysis to explore hidden knowledge in the assessment of a math course," in *Proc. Int. Conf. e-Learn.*, 2020, pp. 1–8.

[25] J. M. Rodriguez-Jimenez, "Analyzing criminal networks using formal concept analysis with negative attributes," in *Proc. Int. Conf. Comput. Math. Methods Sci. Eng.*, 2016, pp. 1–11.

[26] P. Elzinga, J. Poelmans, S. Viaene, G. Dedene, and S. Morsing, "Terrorist threat assessment with formal concept analysis," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, May 2010, pp. 77–82.

[27] C. D. Maio, G. Fenza, M. Gallo, V. Loia, and C. Stanzione, "Toward reliable machine learning with congruity: A quality measure based on formal concept analysis," *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1899–1913, Jan. 2023.

[28] A. Sangroya, "Using formal concept analysis to explain black box deep learning classification models," in *Proc. FCA4AI IJCAI*, 2019, pp. 19–26.

[29] V. S. Naresh and M. Thamarai, "Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions," *WIREs Data Mining Knowl. Discovery*, vol. 13, no. 2, Mar. 2023, Art. no. e1490.

[30] J. C. Lin, P. Fournier-Viger, L. Wu, W. Gan, Y. Djenouri, and J. Zhang, "PPSF: An open-source privacy-preserving and security mining framework," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 1459–1463.

[31] U. H. W. A. Hewage, R. Sinha, and M. A. Naeem, "Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: A systematic literature review," *Artif. Intell. Rev.*, pp. 1–38, Feb. 2023.

[32] I. I. Amin, S. K. Kassim, A. E. Hassanien, and H. A. Hefny, "Formal concept analysis for mining hypermethylated genes in breast cancer tumor subtypes," in *Proc. 12th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, Nov. 2012, pp. 764–769.

[33] J. Wang, Z. Cai, and J. Yu, "Achieving personalized k-anonymity-based content privacy for autonomous vehicles in CPS," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4242–4251, Jun. 2020.

[34] P. Parameshwarappa, Z. Chen, and G. Koru, "Anonymization of daily activity data by using *ell*-diversity privacy model," *ACM Trans. Manage. Inf. Syst.*, vol. 12, no. 3, pp. 1–21, Sep. 2021.

[35] R. Wang, Y. Zhu, T.-S. Chen, and C.-C. Chang, "Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness," *J. Comput. Sci. Technol.*, vol. 33, no. 6, pp. 1231–1242, Nov. 2018.

[36] M. E. Nergiz, M. Atzori, and C. W. Clifton, "δ-presence," in *Encyclopedia of Cryptography, Security and Privacy*, S. Jajodia, P. Samarati, and M. Yung, Eds. Berlin, Germany: Springer, 2019, pp. 1–5.

[37] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *J. Privacy Confidentiality*, vol. 7, no. 3, pp. 17–51, May 2017.

[38] M. C. Tschantz, S. Sen, and A. Datta, "SoK: Differential privacy as a causal property," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 354–371.

[39] K. A. M. E. Chatzikokolakis, "Broadening the scope of differential privacy using metrics," in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.* Berlin, Germany: Springer, 2013, pp. 82–102.

[40] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2013, pp. 429–438.

[41] B. Balle, J. Bell, A. Gascón, and K. Nissim, "Private summation in the multi-message shuffle model," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 657–676.

[42] B. Avent, A. Korolova, and D. Zeber, "BLENDER: Enabling local search with a hybrid differential privacy model," in *Proc. 26th USENIX Secur. Symp. (USENIX Secur.)*, 2017, pp. 747–764.

[43] W. Jung, S. Kwon, and K. Shim, "TIDY: Publishing a time interval dataset with differential privacy," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 2280–2294, May 2021.

[44] S. Takagi, Y. Cao, and Y. Asano, "Geo-graph-indistinguishability: Protecting location privacy for LBS over road networks," in *Proc. IFIP Annu. Conf. Data Appl. Secur. Privacy*. Cham, Switzerland: Springer, 2019, pp. 143–163.

[45] F. Tramèr, Z. Huang, J.-P. Hubaux, and E. Ayday, "Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1286–1297.
[46] J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling, "Privacy-aware eye tracking using differential privacy," in *Proc. 11th ACM Symp. Eye Tracking Res. Appl.*, Jun. 2019, pp. 1–9.
[47] J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, and A. Toval, "Security and privacy in electronic health records: A systematic literature review," *J. Biomed. Informat.*, vol. 46, no. 3, pp. 541–562, Jun. 2013.
[48] Y. Harel, I. B. Gal, and Y. Elovici, "Cyber security and the role of intelligent systems in addressing its challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 4, pp. 1–12, Jul. 2017.
[49] A. Gkoulalas-Divanis and V. S. Verykios, "An integer programming approach for frequent itemset hiding," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2006, pp. 748–757.
[50] M. Chen, Y. Qian, J. Chen, K. Hwang, S. Mao, and L. Hu, "Privacy protection and intrusion avoidance for cloudlet-based medical data sharing," *IEEE Trans. Cloud Comput.*, vol. 8, no. 4, pp. 1274–1283, Oct. 2020.
[51] J. Li, Y. Zhang, X. Chen, and Y. Xiang, "Secure attribute-based data sharing for resource-limited users in cloud computing," *Comput. Secur.*, vol. 72, pp. 1–12, Jan. 2018.
[52] L. Kong, L. Wang, W. Gong, C. Yan, Y. Duan, and L. Qi, "LSH-aware multitype health data prediction with privacy preservation in edge environment," *World Wide Web*, vol. 25, no. 5, pp. 1793–1808, Sep. 2022.
[53] J. M.-T. Wu, G. Srivastava, A. Jolfaei, M. Pirouz, and J. C.-W. Lin, "Security and privacy in shared HitLCPS using a GA-based multiple-threshold sanitization model," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 1, pp. 16–25, Feb. 2022.
[54] D. Dhinakaran and P. M. J. Prathap, "Preserving data confidentiality in association rule mining using data share allocator algorithm," 2023, *arXiv:2304.14605*.
[55] J.-J. Yang, J.-Q. Li, and Y. Niu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment," *Future Gener. Comput. Syst.*, vols. 43–44, pp. 74–86, Feb. 2015.
[56] K. Yang, Z. Liu, X. Jia, and X. S. Shen, "Time-domain attribute-based access control for cloud-based video content sharing: A cryptographic approach," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 940–950, May 2016.
[57] A. N. K. Zaman, C. Obimbo, and R. A. Dara, "An improved data sanitization algorithm for privacy preserving medical data publishing," in *Proc. Can. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2017, pp. 64–70.
[58] A. Abbas and S. U. Khan, "A review on the state-of-the-art privacy-preserving approaches in the e-health clouds," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1431–1441, Jul. 2014.
[59] Y. Lin, J. Li, and H. Wang, "Granular matrix method of attribute reduction in formal contexts," *Soft Comput.*, vol. 24, no. 21, pp. 16303–16314, Nov. 2020.
[60] S. O. Kuznetsov and S. A. Obiedkov, "Comparing performance of algorithms for generating concept lattices," *J. Experim. Theor. Artif. Intell.*, vol. 14, nos. 2–3, pp. 189–216, Apr. 2002.
[61] E. M. Norris, "An algorithm for computing the maximal rectangles in a binary relation," *Revue Roumaine de Mathematiques Pures et Appliquees*, vol. 23, no. 2, pp. 243–250, 1978.
[62] B. Ganter, "Two basic algorithms in concept analysis," in *Proc. Int. Conf. Formal Concept Anal.* Berlin, Germany: Springer, 2010, pp. 312–340.
[63] S. Kuznetsov, "A fast algorithm for computing all intersections of objects from an arbitrary semilattice," *Nauchno-Tekhnicheskaya Informatsiya Seriya Informatsionnye Protsessy I Sistemy*, vol. 1, no. 1, pp. 17–20, Jan. 1993.
[64] J.-P. Bordat, "Calcul pratique du treillis de Galois d'une correspondance," *Mathématiques et Sci. Humaines*, vol. 96, pp. 31–47, Jan. 1986.
[65] B. Yildiz, A. Kut, and R. Yilmaz, "Hiding sensitive itemsets using sibling itemset constraints," *Symmetry*, vol. 14, no. 7, p. 1453, Jul. 2022.
[66] J. M. Wu, J. Zhan, and J. C. Lin, "Ant colony system sanitization approach to hiding sensitive itemsets," *IEEE Access*, vol. 5, pp. 10024–10039, 2017.

HEDI HAMDI received the Ph.D. degree in computer sciences from the University of Bordeaux, France, in December 2009. Currently, he is an Assistant Professor with the Computer Science Department, College of Computer and Information Science, Jouf University, where he is conducting research activities in the areas of cloud computing, information security, software defined network, and network functions virtualization. He is a member of the RIADI-GLD Laboratory, ENSI, Manouba University.

ZAKI BRAHMI received the Ph.D. degree in computer sciences from the Faculty of Mathematical, Physical and Natural Sciences, University of Tunis Manar, Tunisia, in December 2010. He is currently an Associate Professor with the Computer Science Department, College of Science and Arts at Al-Ola, Taibah University. He is a member of the RIADI-GLD Laboratory, ENSI, Manouba University. His research interests include web services composition, intensive workflow scheduling in cloud computing and data stream mining for outlier, and workload detection in cloud computing.

ALAA S. ALAERJAN received the B.S. degree in computer and information sciences from Jouf University, Saudi Arabia, in 2009, the M.S. degree in computer science from Ball State University, in 2013, and the Ph.D. degree in computer science and informatics from Oakland University, in 2019. He is currently an Assistant Professor with the Department of Computer Science, College of Computer and Information Sciences, Jouf University. His research interests include distributed systems, software engineering, the IoT, information security, and smart grids.

LOTFI MHAMDI (Member, IEEE) received the Master of Philosophy (M.Phil.) degree in computer science from The Hong Kong University of Science and Technology (HKUST), in 2002, and the Ph.D. degree in computer engineering from the Delft University of Technology (TU Delft), The Netherlands, in 2007. He continued his work at TU Delft as a Postdoctoral Researcher, working on high-performance networking tropics within various European Union funded research projects. Since July 2011, he has been a Lecturer with the School of Electronic and Electrical Engineering, University of Leeds, U.K. His research interests include high-performance networks, including the architecture, design, analysis, scheduling, and management of high-performance switches, and internet routers. He is/was a Technical Program Committee Member of various conferences, including the IEEE International Conference on Communications (ICC), the IEEE GLOBECOM, the IEEE Workshop on High Performance Switching and Routing (HPSR), and the ACM/IEEE International Symposium on Networks-on-Chip (NoCS). He is/was the TPC Co-Chair of the Green Computing, Networking, and Communications Symposium (GCNC 2020), and the TPC Co-Chair of GLOBECOM (NGNI Symposium), in 2020. He is currently serving as the Vice-Chair for the IEEE ComSoc Technical Committee on Communication Switching and Routing (CSR-TC).