Information Sciences 704 (2025) 121949

Contents lists available at ScienceDirect

Information Sciences

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

Enhancing efficiency and data utility in longitudinal data anonymization

Fatemeh Amiri [a,*], David Sánchez [b], Josep Domingo-Ferrer [b]

[a] Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran
[b] Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT Center for Cybersecurity Research of Catalonia, Av. Països Catalans 26, E-43007, Tarragona, Catalonia, Spain

ARTICLE INFO

Keywords:
Longitudinal data publishing
Sequence data publishing
Data anonymization
Data privacy

ABSTRACT

Longitudinal data consist of observations collected over time from a set of individuals. The accumulation of information on each individual over time makes longitudinal data particularly privacy-sensitive. However, existing anonymization methods are often inadequate for ensuring privacy-preserving publication of such data, as current privacy models assume unrealistic levels of attacker knowledge. To address this, we propose the $(k, \beta)^L$-privacy model, which assumes that an attacker's knowledge is limited to a subsequence of L quasi-identifiers. This provides a more realistic representation of the information an attacker might actually possess. Our model guarantees that every subsequence of L quasi-identifier values appears in either zero or at least k records within the longitudinal database. Additionally, it ensures that the confidence of any sensitive value within these k records is at most β times higher than its confidence in the entire dataset. This not only strengthens privacy protection but also enhances data utility.

Furthermore, we introduce FCLA, an anonymization algorithm designed to enforce our privacy model while prioritizing data utility. FCLA effectively mitigates identity and attribute disclosures, as well as skewness attacks in longitudinal data. It achieves this by partitioning sequences into groups and anonymizing them independently—a process that can be efficiently parallelized. Experimental results show that FCLA outperforms existing methods in preserving data utility while adhering to strict privacy constraints. Additionally, time complexity analysis and execution time measurements demonstrate that FCLA is more efficient and scalable than alternative approaches.

1. Introduction

Longitudinal data refer to information continuously collected from the same individuals over time. This type of data is valuable for understanding temporal changes and relationships between variables.

Ensuring the anonymization of longitudinal data before their release is crucial for safeguarding individuals' privacy and confidentiality while complying with privacy regulations and laws. However, the lack of anonymization methods in certain longitudinal surveys -such as the Scottish Longitudinal Survey [1], the Youth Survey Luxembourg [2], the Young People and Covid-19 (YAC) survey [3], the Framingham Heart Study (FHS) [4], the Child Asthma Management Program (CAMP) [5], and the HIVNET Informed Consent Substudy [6]- highlights the pressing challenges of privacy-preserving longitudinal data publication.

* Corresponding author.
E-mail address: f.amiri@hut.ac.ir (F. Amiri).

https://doi.org/10.1016/j.ins.2025.121949
Received 1 November 2023; Received in revised form 3 February 2025; Accepted 4 February 2025

Available online 6 February 2025
0020-0255/© 2025 Published by Elsevier Inc.

Table 1
Inpatient longitudinal data.

PID	VID	Y	Z	D	L	Disease
1	1	2018	41002	0	21	Infection
2	1	2018	41001	0	10	Cancer
2	2	2018	43003	30	1	Flu
3	1	2018	40012	0	30	Fever
3	2	2019	40012	0	35	Infection
3	3	2020	41001	0	3	Flu
4	1	2017	41001	0	10	Heart attack
4	2	2018	42003	0	10	Flu
4	3	2021	42003	0	35	Hepatitis
5	1	2018	42005	0	3	Fever
5	2	2018	42005	80	10	Cancer
6	1	2017	41001	0	5	Infection
6	2	2019	43002	0	30	Hepatitis
7	1	2018	42016	0	4	Fever
8	1	2020	43003	0	5	Fever
9	1	2019	42016	0	1	Flu
10	1	2018	42016	0	14	Heart attack

Table 1 presents a sample dataset that emulates longitudinal health records, resembling electronic health records (EHRs) and electronic medical records (EMRs). The longitudinal nature of this dataset is evident in the possibility of a patient undergoing multiple admissions within a certain time frame. These admissions are interconnected and pertain to the same individual. Within this dataset, each patient may have multiple records corresponding to different visits, which are systematically ordered by visit date, with each visit assigned a unique visit ID. The dataset includes six quasi-identifying (QI) attributes: PID (patient ID), VID (visit ID), Y (admission year), Z (zip code), D (number of days since the first claim in each year), and L (length of hospital stay). Additionally, it contains one sensitive attribute (SA): Disease. For example, in Table 1, patient PID $=$ 2 was hospitalized for ten days in 2018 and had another visit 30 days later in zip code 43003.

Privacy breaches can occur through individual visits or across multiple visits over time. If these data are released, adversaries can launch two types of attacks: *identity disclosure* and *attribute disclosure*. In identity disclosure, adversaries attempt to re-identify individuals from anonymized data. For example, if an adversary knows that Alice had a visit in 2017 and has resided in zip code 41003 since 2018, they can identify Alice's records (those with PID$=4$) and, consequently, learn all the diseases she has had.

Attribute disclosure, on the other hand, aims to infer sensitive information about individuals even without re-identifying their records. For instance, if Bob had two visits in 2018, an attacker might deduce that his condition is cancer, even without pinpointing his exact record.

The k-anonymity model [7] is widely used to protect relational data from privacy breaches by preventing identity disclosure. It ensures that each individual in a dataset is indistinguishable from at least $k-1$ others. However, it does not provide protection against attribute disclosure. To address this limitation, extensions of k-anonymity, such as β-likeness [8], t-closeness [9], and l-diversity [10], have been proposed. However, these models are not well-suited for longitudinal data due to their high dimensionality and sequential nature.

Protecting longitudinal data from re-identification and attribute disclosure is particularly challenging, as they contain multiple instances of each attribute at different time points. Some approaches, such as those proposed in [11], [12], [17], and [13], remain vulnerable to attribute disclosure attacks. Moreover, existing anonymization methods for longitudinal data often fail to generate datasets suitable for emerging clinical research. Effective anonymization requires accounting for temporal correlations and preventing privacy breaches through any combination of quasi-identifier (QI) values both within and across events.

Elemam et al. [11] and Tamersoy et al. [17] focus on protecting against identity disclosure, which may lead to overlooking sensitive values. In contrast, Sehatkar and Matwin [12] propose a partitioning method that considers the presence of highly sensitive items in a specified threshold percentage of records. However, such approaches may not adequately protect infrequent sensitive values, which are common in longitudinal data.

Additionally, privacy models such as k-anonymity and its extensions assume that attackers possess complete knowledge of all QI attributes for linkage attacks. When applied to longitudinal data, this assumption implies that adversaries have exhaustive knowledge of all events related to an individual -an unrealistic scenario that would require significant data distortion to prevent such attacks-. Some studies, including [14] and [15], explore the concept of limiting attackers' background knowledge.

1.1. Contributions and plan

We introduce the $(k,\beta)^L$-privacy model, which limits adversary's background knowledge to subsequences of L publicly known QI values. This model guarantees that each subsequence of L QI values appears in either zero or at least k records within the longitudinal database. Furthermore, it ensures that the confidence for any sensitive value within these k records is at most β times greater than its confidence in the entire dataset. Our model protects against identity disclosure, attribute disclosure, and attacks based on skewness and semantics [9].

It is unlikely that an adversary possesses complete knowledge of all QI attribute values for every event in an individual's record. Our model assumes a more realistic attack scenario, where adversaries have knowledge of only a subsequence of events. For instance, in medical longitudinal data, an attacker might know about a specific visit in 2018 or a two-day hospitalization in 2019 rather than an individual's entire medical history.

Our model further enhances privacy by allowing the definition of subsets of highly sensitive information that require stricter protection. By applying our privacy model to sequences containing values from these subsets, we achieve a balance between privacy and data utility.

To enforce our privacy model, we propose an anonymization algorithm that employs value generalization and suppression techniques. When releasing medical data for general research purposes without a predefined use case, it is crucial to ensure that each record in the published dataset accurately represents a real individual. For example, researchers may require real patient records to study the side effects of a specific drug. Unlike methods such as swapping, randomization, or synthetic data generation, generalization and suppression preserve the truthfulness of records while reducing the level of detail. This ensures that the released data remain meaningful and prevents the introduction of fabricated information during anonymization. Methods that perturb or synthesize original values can introduce semantic distortions, making the anonymized data unsuitable for certain analyses [16].

Our proposed algorithm specifically addresses the sequential nature of longitudinal data by operating on multidimensional sequences of events. To preserve data utility, it partitions the sequence data into clusters based on QI similarity, ensuring that each cluster contains at least k sequences. This clustering approach minimizes data distortion while maintaining data quality and satisfying privacy constraints. We employ a dynamic programming technique, inspired by the sequence alignment problem [17], to measure the distance between sequences. Within each cluster, anonymization techniques are applied independently, ensuring that every subsequence of at most L QI values satisfies our privacy model. Parallel processing further enhances computational efficiency. To detect privacy-violating sequences within each cluster, we utilize sequential pattern mining techniques [18].

We conducted experiments on three datasets -the CMS 2008-2010 DE-SynPUF dataset,[1] the Heritage Health Prize (HHP) claims dataset,[2] and the FDA Adverse Event Reporting System (FAERS) dataset[3]- to evaluate the effectiveness of our methods. By analyzing different parameters of our privacy model, we compared our algorithm's performance with the most closely related work, the HALT algorithm [12]. The results demonstrate that our approach outperforms HALT in terms of both data utility and privacy while achieving lower time complexity.

Section 2 discusses related work. Section 3 formally defines the problem. Section 4 presents our proposed privacy model. Section 5 introduces our anonymization algorithm. Section 6 provides implementation details and experimental results. Finally, Section 7 presents conclusions and outlines potential future research directions.

2. Related work

In this section, we review research on the anonymization of longitudinal data and related data types, including transaction data, trajectory data, and sequential data. Our focus is on evaluating key anonymization techniques to enhance privacy while ensuring their suitability for purely longitudinal data. Since many models used for these data types are derived from relational data models, we first examine research on relational data and discuss the limitations of these models.

Relational data, typically structured in tabular format, share a fixed set of attributes across all records. Various privacy models have been introduced to prevent privacy breaches in such data [8–10,7]. The k-anonymity model [7] ensures that each record is indistinguishable from at least $k-1$ others based on its quasi-identifier (QI) values. However, while k-anonymity protects against identity disclosure, it does not safeguard confidential attributes from attribute inference attacks. To address this limitation, the l-diversity model [10] requires each equivalence class to contain at least l well-represented values for the sensitive attribute. However, this model assumes a uniform distribution of sensitive values, making it less effective for skewed data or multiple data releases. The t-closeness model [9] mitigates this issue by limiting disparities -up to a threshold t- between the frequency distributions of sensitive values in an equivalence class and the overall dataset. The β-likeness model [8] further refines this approach by controlling an adversary's confidence gain using a relative difference measure. However, it does not protect against identity disclosure.

Unlike relational data, transactional data are unstructured, high-dimensional, and non-sequential. Studies such as [19] and [20] propose anonymization methods for transactional data that do not rely on sequence information. However, simply applying transaction anonymization techniques to longitudinal data is inadequate, as adversaries can exploit the sequential nature of longitudinal records. Effective privacy protection for longitudinal data requires specialized techniques that account for their unique structure.

Trajectory data consist of spatio-temporal sequences that track the movement of objects such as GPS devices, cell phones, or RFID tags. Various anonymization approaches for this type of data are surveyed in [21]. Some works apply the l-diversity requirement to protect trajectory data [22,23]. Although trajectory data can be considered a subset of longitudinal data focused on spatial movement, existing anonymization methods primarily generalize or suppress location information [24]. However, when these techniques are applied to longitudinal data containing multiple attributes per time point, substantial information loss can occur. Moreover, trajectory-based distance measures may not accurately capture the complexity of longitudinal data, which typically include multiple quasi-identifiers per event.

[1] https://www.cms.gov/research-statistics-data-and-systems/downloadable-public-use-files/synpufs/de_syn_puf.
[2] https://foreverdata.org/1015/index.html.
[3] https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html.

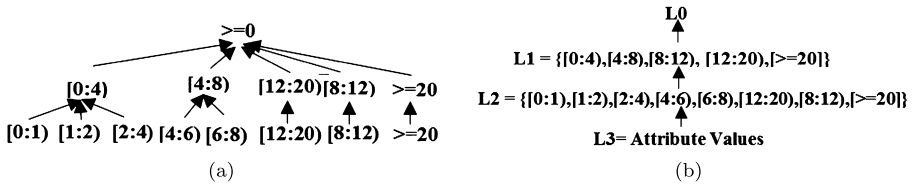


Fig. 1. Generalization hierarchies of L and D in terms of weeks: (a) value generalization VH_L, VH_D; (b) domain generalization DH_L, DH_D.

Longitudinal data comprise multidimensional sequences of events [25]. The anonymization techniques proposed by Elemam et al. [11] and Tamersoy et al. [17] enforce k-anonymity by modifying sequences to create indistinguishable groups of k events. However, these methods remain vulnerable to attribute disclosure attacks and fail to accurately model adversary knowledge. For example, in [17], records consist of sequences of (ICD, Age) pairs and DNA sequences. Existing approaches assume adversary knowledge as combinations of (ICD, Age) pairs but fail to consider the multidimensionality of longitudinal data, where each event may have multiple quasi-identifiers. Applying the method from [11] to our context would require restricting each event to two quasi-identifiers and assuming the adversary always knows both values per visit.

Sehatkar and Matwin [12] introduced the $(k,c)^L$-privacy model, which combines k-anonymity with a restriction that no more than c percent of records contain a particular sensitive value. However, their HALT algorithm examines all patterns with up to L QI values without considering QI similarity, leading to high computational costs, significant information loss, and susceptibility to skewness attacks.

Wang et al. [13] proposed the $MS(k,\theta^*)$-bounding model to protect published spontaneous reporting data. Their model assumes individuals may have multiple events, each with multiple sensitive values. However, their anonymization approach does not account for temporal characteristics. By generalizing all events for each individual into a single "super record" and clustering these records, this method results in larger information loss compared to [12], which anonymizes individual events while preserving temporal features.

Parameshwarappa et al. [26] developed multi-level clustering-based methods to enforce l-diversity for preventing attribute disclosure. Their framework anonymizes longitudinal activity data, treating quasi-identifiers as spanning all time points. However, quasi-identifiers themselves can sometimes reveal sensitive details. For example, knowledge that someone is on vacation for an extended period could help burglars identify vulnerable homes. Activity datasets generally have high dimensionality but few records. Consequently, [26] also examined risks associated with the duration of sensitive activities, noting that individuals absent from home for multiple days face higher risks than those away for just a few hours. Nevertheless, their approach relies on fixed-length sequences, whereas our research protects variable-length sequences unique to each individual. Furthermore, activity datasets typically assume a single data point per time interval, whereas longitudinal health records (as shown in Table 1) contain multiple admissions over specified periods.

Additionally, l-diversity is vulnerable to skewness attacks [26,9], where certain sensitive values become more easily identifiable within a subset of sequences than in the overall dataset. For example, consider a $(10, 0.1)$-privacy model [12] applied to a longitudinal health dataset T, where only 0.1% of individuals have HIV. Suppose a set of sequences Q in T contains 10 different sequences, with one individual diagnosed with HIV. In Q, the probability of linking to HIV is 10%, whereas in T, it is only 0.1%. This 100-fold increase in probability represents a significant privacy risk.

Our work addresses these challenges by proposing a novel, intuitive, distribution-based privacy model. We also introduce a specialized anonymization algorithm designed to enforce this model effectively.

3. Problem definition

We model data as consisting of quasi-identifying attributes denoted by $A = \{A_1, A_2, ..., A_n\}$ with their respective attribute domains represented as $R = \{R_1, R_2, ..., R_n\}$. Each QI attribute is associated with a domain generalization hierarchy DH_{A_i} and a value generalization hierarchy VH_{A_i}. A domain generalization is defined to be a set of domains that is totally ordered by direct generalization.

Value generalization is a value-level tree, in which edges are direct value generalizations. Additionally, there is a sensitive attribute (SA) with domain values $R_{SA} = \{s_1, s_2, ..., s_m\}$.

As an illustration, Figs. 1 to 3 visually depict the generalization hierarchies for the QIs listed in Table 1. The L and D attributes are generalized in terms of weeks.

The multidimensional sequence data set $T = \{T^1, T^2, ..., T^t\}$ consists of records of the form $T^i = (SID, SE)$, where each record represents a sequence of events $SE = \langle e_1, e_2, ...\rangle$. The events are ordered chronologically based on the temporal relation $<_t$, such that $e_1 <_t e_2 <_t ... <_t e_r$. Each event e_i contains values for n QIs and one SA, denoted as $(a_1, a_2, ..., a_n, s_j)$, where $a_i \in R_i$ and $s_j \in R_{SA}$. We also assume that attackers may have values for subsets of QIs, but the specific subset may vary for different attackers. The sensitive attribute values can be assumed private (if the attacker had them for a certain configuration of QIs, no attack would be needed).

More specifically, each event in a sequence can be represented as a collection of $(A_i, value)$ pairs, called items. The *value* corresponds to a node in VH_{A_i}. For example, to convert the longitudinal data in Table 1 into a multidimensional sequential data set, the visits of each patient are combined, sorted based on visit dates, and presented as sets of n attributes. Table 2 serves as an illustrative reference for analysis after this transformation. To compare sequences, we introduce the following concepts. For two items $x = (A_i, v)$ and

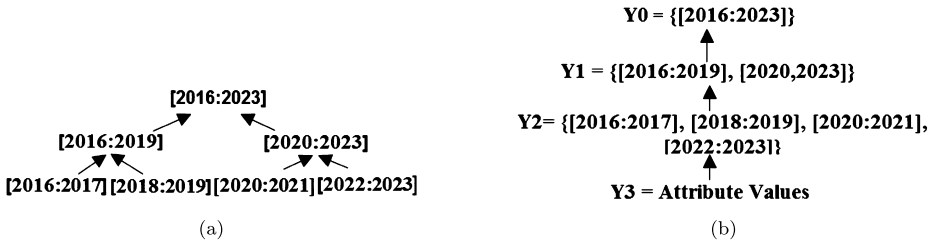


Fig. 2. Generalization hierarchy of Y: (a) value generalization VH_Y; (b) domain generalization DH_Y.

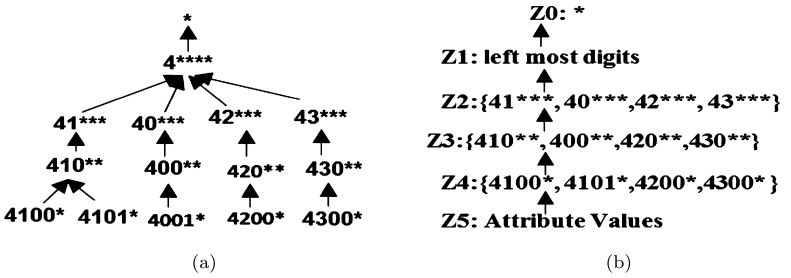


Fig. 3. Generalization hierarchies of Z: (a) value generalization VH_Z; (b) domain generalization DH_Z.

Table 2
Multidimensional sequential data.

PID	Sequence of visits
1	$\langle\{(Y,2018),(Z,41002),(D,0),(L,21),(Disease,Infection)\}\rangle$
2	$\langle\{(Y,2018),(Z,41001),(D,0),(L,10),(Disease,Cancer)\},$ $\{(Y,2018),(Z,43003),(D,30),(L,1),(Disease,Flu)\}\rangle$
3	$\langle\{(Y,2018),(Z,40012),(D,0),(L,30),(Disease,Fever)\},$ $\{(Y,2019),(Z,40012),(D,0),(L,35),(Disease,Infection)\}$ $\{(Y,2020),(Z,41001),(D,0),(L,3),(Disease,flu)\}\rangle$
4	$\langle\{(Y,2017),(Z,41001),(D,0),(L,10),(Disease,Heart\ attack)\},$ $\{(Y,2018),(Z,42003),(D,0),(L,10),((Disease,Flu)\},$ $\{(Y,2021),(Z,42003),(D,0),(L,35),(Disease,Hepatitis)\}\rangle$
5	$\langle\{(Y,2018),(Z,42005),(D,0),(L,3),(Disease,Fever)\},$ $\{(Y,2018),(Z,42005),(D,80),(L,10),((Disease,Cancer)\}\rangle$
6	$\langle\{(Y,2017),(Z,41001),(D,0),(L,5),(Disease,Infection)\},$ $\{(Y,2019),(Z,43002),(D,0),(L,30),(Disease,Hepatitis)\}\rangle$
7	$\langle\{(Y,2018),(Z,42016),(D,0),(L,4),(Disease,Fever)\}\rangle$
8	$\langle\{(Y,2020),(Z,43003),(D,0),(L,5),(Disease,Fever)\}\rangle$
9	$\langle\{(Y,2019),(Z,42016),(D,0),(L,1),(Disease,Flu)\}\rangle$
10	$\langle\{(Y,2018),(Z,42016),(D,0),(L,14),(Disease,Heart\ attack)\}\rangle$

$y=(A_i,v')$, we define y as being more specific than x (denoted as $y\leq_I x$) if either $v=v'$ or v is an ancestor of v' in VH_{A_i}. For example $(L,[0:1))\leq_I (L,[0:4))$.

Let $e=\{a_1,...,a_n\}$ and $e'=\{a'_1,...,a'_n\}$ represent two given events. We say that e is more specific than e' (denoted as $e\leq_E e'$) if, for all $1<i<n$, $a_i\leq_I a'_i$. For instance, $\{(Y,2016),(Z,4100*)\}$ is considered more specific than $\{(Y,[2016:2019]),(Z,4100*)\}$. Consider two sequences $X=\langle x_1,...,x_l\rangle$ and $X'=\langle x'_1,...,x'_{l'}\rangle$. We say that X is more specific than X' (represented as $X\leq_S X'$) if there exists a sequence of indices $1\leq i_1<i_2<...<i_{l'}\leq l$ where $x_j\leq_E x'_{i_j}$ holds for all $j=1,2,...,l'$, with $l'\leq l$. For example, $\langle\{(Y,2018),(L,2)\},\{(Y,2019),(Z,43003)\}\rangle\leq_S\langle\{(Y,[2016:2023]),(Z,4300*)\}\rangle$.

We also define the length of a sequence, denoted by $len(X)$, as the total number of QI items across all events in X. The items are numbered if their values are not the root of the value generalization hierarchy tree. For example, in the sequence $\langle\{(Y,[2016:2023]),(L,2)\},\{(Y,2019),(Z,43003)\}\rangle$, there are two events, and $len(X)=3$.

Definition 1. The support of a sequence X, denoted as $sup_s(X)$, is the set of sequences in T that are at least as specific as X ($sup_s(X)=\{S^i\in T\mid S^i\leq_S X\}$, and $|sup_s(X)|$ represents the count of sequences in $sup_s(X)$.

Definition 2. The support of an event e', denoted as $sup_e(e')$, is the set of sequences in T that contain an event at least as specific as e'.

4. Privacy model

We next present our privacy model, which assumes that the distribution of sensitive attribute (SA) values in a longitudinal data set T is publicly known. Furthermore, we assume that the attacker possesses a certain background knowledge regarding the longitudinal events of victim. The attacker's background knowledge about the sequence of events for the victim includes values of QIs as well as the order of these values. By leveraging publicly available hierarchies for QIs, the attacker can express her knowledge at varying levels of granularity. In real-life scenarios, it is in general not feasible for an attacker to have knowledge of all the QI values for every event. Therefore, it is more realistic to assume that the attacker has partial knowledge, which is limited to a maximum of P values of QIs. This bounded partial knowledge assumption aligns with the constraints and limitations typically encountered in practical situations. As an example, consider the case where the attacker knows that the victim had a visit between 2018 and 2019 and the victim was admitted to the hospital for 3 days in 2020. The attacker is aware of the hierarchy of the attributes shown in Figs. 1-3. In this scenario, she knows 3 QI values where two of them occur together in one visit. Such knowledge can be modeled as $X = \langle\{[Y, 2018 : 2019]\}, \{[Y, 2020], [L, 3]\}\rangle$. Equipped with this background knowledge, the attacker can initiate a privacy breach by identifying a matching sequence in the longitudinal data set that aligns with her background knowledge X. A sequence X' is a match to X if it contains X, *i.e.*, $X' \preceq_S X$. When examining the data in Table 2, the record with PID # 3 corresponds to sequence X. To define our proposed privacy model, we compute the confidence of a sensitive value as follows.

Definition 3. The confidence of a sensitive attribute value s_i in a set of sequences Q reflects the probability of inferring s_i for an individual in
Q. It is denoted by $conf\left(s_i|Q\right) = \frac{|sup_s(Q \cap s_i)|}{|sup_s(Q)|}$, where $Q \cap s_i$ represents the set of sequences containing both Q and s_i. In fact, $conf\left(s_i|Q\right)$ represents the percentage of sequences in $sup_s(Q)$ that contain s_i.

When the victim's record is in Q, the gain in confidence for the sensitive attribute value s_i is computed as $q_i - p_i$, where q_i represents the confidence of s_i in Q and p_i represents the confidence of s_i in the sequence data set T. A positive gain indicates increased correlation between the victim's record and s_i within Q, while a negative gain means reduced correlation, thereby enhancing privacy.[4] Therefore, our approach primarily focuses on positive confidence gains. To mitigate the risk of identity and attribute disclosure, we enforce two conditions in each set of sequences via our privacy model $(k, \beta)^L$-privacy, as follows.

Definition 4 *($(k, \beta)^L$-privacy).* Let $P_c = \{p_1, \ldots, p_m\}$ represent the confidences for all sensitive attribute (SA) values in a longitudinal dataset T. We define L as the maximum length of the adversary's background knowledge. Moreover, let Q be a set of sequences with SA confidence $Q_c = \{q_1, \ldots, q_n\}$. The relative distance between the confidences of SA values is determined by $D(p_i, q_i) = \frac{q_i - p_i}{p_i}$. A set of sequences Q satisfies $(k, \beta)^L$-privacy if and only if the following two conditions are met:

1. To protect against identity disclosure, it must hold that $|sup_s(Q)| \geq k$.
2. To protect against attribute disclosure, it must hold that the maximum relative distance between confidences of SA values is less than or equal to $\min\{\beta, -\ln p_i\}$, *i.e.* $\max\{D(p_i, q_i) | p_i \in P_c, p_i < q_i\} \leq \min\{\beta, -\ln p_i\}$, where β is a positive value.

By ensuring that each individual in the data set cannot be distinguished from at least $k - 1$ other individuals based on QI, Condition 1 in Definition 4 effectively mitigates the risk of identity disclosure. To preempt both attribute disclosure and skewness attacks, Condition 2 mandates that the difference in confidence of SA values within any group of sequences should not exceed a predefined threshold compared to the overall data set's confidence values. Note that using the absolute difference as a metric would fall short of treating all SA values and their relative differences equitably. As a result, we have chosen to adopt a relative difference distance approach. Additionally, we propose the use of a maximum threshold, rather than relying on a cumulative distance threshold. Our model is intentionally designed to accommodate SA values that may be absent from specific groups, granting more flexibility in the anonymization process and ultimately preserving higher information quality. We next prove the monotonicity of both Conditions 1 and 2.

Lemma 1 *(Monotonicity). Assume there are two disjoint groups of sequences Q^1 and Q^2 generated from sequences in the overall data set T. Let SA value $s_i \in S$ have confidence p_i in T, q_i^1 in Q^1, and q_i^2 in Q^2. Additionally, let q_i^3 be the confidence in $Q^3 = Q^1 \cup Q^2$. Both conditions of Definition 4 satisfy monotonicity.*

[4] However, it is important to note that reducing the gain for certain sensitive attribute values may inadvertently result in privacy violations. For example, if the victim's gain for the majority values of religion or sexual orientation is reduced, it means she is more likely to belong to a religion or sexual orientation minority. To address this issue, a potential solution, proposed in [8], is to transform the relative negative gain of a sensitive attribute value into positive gains for other sensitive attribute values.

Proof. The monotonicity property in Condition 1 requires $|sup_s(Q^3)| \geq \max(|sup_s(Q^1)|, |sup_s(Q^2)|)$. For all sequences S and S^0, if $S^0 \leq_S S$, then $|sup(S^0)| \leq |sup(S)|$ [27]. Therefore, $|sup_s(Q^1)| \leq |sup_s(Q^3)|$ and $|sup_s(Q^2)| \leq |sup_s(Q^3)|$ because $Q^1 \leq_S Q^3$ and $Q^2 \leq_S Q^3$. Thus,

$$|sup_s(Q^3)| \geq \max(|sup_s(Q^1)|, |sup_s(Q^2)|).$$

To prove the monotonicity of Condition 2, we need to demonstrate that $D(p_i, q_i^3) = \max\{D(p_i, q_i^1), D(p_i, q_i^2)\}$. Assume there are $n^1 = |sup_s(Q^1 \cap s_i)|$ and $n^2 = |sup_s(Q^2 \cap s_i)|$ sequences with the sensitive value s_i in Q^1 and Q^2, respectively. Then, we have $q_i^1 = \frac{n^1}{|sup_s(Q^1)|}$, $q_i^2 = \frac{n^2}{|sup_s(Q^2)|}$, and $q_i^3 = \frac{n^1+n^2}{|sup_s(Q^1)|+|sup_s(Q^2)|} = \frac{q_i^1 \cdot |sup_s(Q^1)| + q_i^2 \cdot |sup_s(Q^2)|}{|sup_s(Q^1)|+|sup_s(Q^2)|} \leq \max\{q_i^1, q_i^2\}$. Thus, $D(p_i, q_i^3) \geq \max\{D(p_i, q_i^1), D(p_i, q_i^2)\}$. □

Condition 2 limits the confidence level of sensitive values in set Q by requiring $q_i \leq (1 + \min\{\beta, -\ln p_i\}) \cdot p_i$. This constraint can be further understood by decomposing the upper bound defined by $\delta(p_i) = (1 + \min\{\beta, -\ln p_i\}) \cdot p_i$:

- For $0 < p_i \leq e^{-\beta}$, $\delta(p_i) = (1 + \beta) \cdot p_i$.
- For $e^{-\beta} \leq p_i \leq 1$, $\delta(p_i) = (1 - \ln p_i) \cdot p_i$.

The two segments intersect at $p_i = e^{-\beta}$. The function $\delta(p_i)$ monotonically increases within $(0, 1]$, ensuring confidence levels below 1, accounting for value frequency. For rare values ($0 < p_i \leq e^{-\beta}$), confidence is bounded by $(1+\beta)p_i$. For frequent values ($e^{-\beta} \leq p_i \leq 1$), confidence is restricted to $(1 - \ln p_i)p_i$, ensuring $q_i \leq (1 - \ln p_i)p_i \leq (1 + \beta)p_i$. This framework guarantees that confidences in Q are at most β times larger than those in T, maintaining privacy. The choice of $-\ln p_i$ is motivated by its properties (see [8]).

Condition 1 of our privacy model ensures that clusters contain no fewer than k elements, thereby protecting against identity disclosure. Condition 2 explicitly quantifies the relationship between the β threshold and positive information gain. Its maximum-distance threshold inherently safeguards against skewness and semantic attacks [9]. Furthermore, by clearly distinguishing between positive and negative information gain and allowing for the absence of sensitive attribute values within clusters, Condition 2 provides greater flexibility in anonymization. This results in higher information quality compared to the closely related $(k, c)^L$-privacy model [12]. As argued, a measure of relative difference serves our purpose by offering stronger protection for less frequent SA values. Moreover, Condition 2 defines only an upper bound on q_i and allows any q_i value less than p_i, but always disallows q_i values equal to 1 (its upper bound strictly less than 1). We assert that both of these choices are reasonable for preserving privacy compared to the $(k, c)^L$-privacy model.

Our model has several advantages over previous privacy models for longitudinal data. Firstly, it is more intuitive as it measures the risk based on the relative confidence gain of a sensitive value. This allows the data publisher to specify an acceptable maximum relative confidence threshold based on this intuitive measure. Secondly, unlike previous models proposed in [17], [12], our model explicitly quantifies the association between the threshold and the positive gain of information. Finally, our model inherently protects against identity and attribute disclosures, as well as skewness attacks, by enforcing a maximum-distance threshold [8].

Different sensitive attribute values may have varying privacy protection requirements. For instance, certain values like 'flu' for the disease attribute in medical data may be considered less sensitive and acceptable to disclose. To address this variability, we define a subset of sensitive attribute values denoted by $\pi \subseteq SA$, called highly sensitive values, that require stricter protection. In our privacy model, it is crucial to fulfill the second condition for sequences that contain at least one value from π in order to provide enhanced privacy for the specified subset. It's important to note that some works (such as [28], [29]) consider diverse privacy thresholds for sensitive values within anonymization frameworks.

5. The anonymization algorithm

In this section, we present FCLA –Fast Clustering-based Longitudinal data Anonymization–, which enforces the model above while striving to preserve data utility. FCLA microaggregates (that is, generates clusters of size at least k) similar sequences based on QIs and implements the privacy model for anonymizing each cluster. Sequence alignment techniques [17] are employed to measure sequence distances, aiding in effective cluster formation.

FCLA (Algorithm 1) uses the longitudinal data set T and privacy model parameters to generate anonymized sequences. Our methodology for the microaggregation component is based on the Maximum Distance to Average Vector (MDAV) algorithm [30], which is an efficient heuristic for creating homogeneous clusters of size at least k. The minimum cluster size k determines privacy, and cluster homogeneity favors utility preservation. First, FCLA defines a distance matrix to keep track of between-cluster distances, and it initializes the distance matrix with the distances between all singleton clusters using sequence alignment techniques. The algorithm begins by selecting a centroid sequence $\hat{c}$ and finding the most distant sequence r. It forms the first cluster G_1 with r and $k-1$ closest unclustered sequences. Additional sequences are added to G_1 until privacy condition 2 in Definition 4 is met for all SA values $s_i \in S$. The process continues by creating new clusters G_i with the farthest unclustered sequences until fewer than $2k$ unclustered sequences remain. If these sequences satisfy privacy condition 2, they form a final cluster; otherwise, they are inserted into other clusters without violating privacy condition 2. After cluster creation, each cluster within *Group* undergoes anonymization in lines 3 to 4 of Algorithm 1. Thus, FCLA identifies and removes sequences violating the privacy requirement to achieve $(k, \beta)^L$-privacy within each cluster G_i. We will provide descriptions of these functions and perform a time complexity analysis using FCLA in the following subsections.

Algorithm 1: The Fast Clustering-based Longitudinal Anonymization (FCLA) algorithm.

Input: T: data set of Sequences; k, β, L : privacy parameters
Output: Set of anonymized sequences satisfying privacy model

// Microaggregate the sequence data
1 $Group =$ Microaggregate the sequences based on QI values
// Generate the anonymized sequences
2 For each cluster of sequences $G_i \in Group$
3 Identify sequences that violate privacy
4 Remove the sequences that violate privacy

5.1. Identifying sequences that violate privacy

As we assume the attacker possesses a non-empty sequence of L different QI values within or across events, we need to identify the sequences that violate the privacy model. These sequences pose risks of identity and attribute disclosure. A sequence X in T, with a length between 1 and L, violates the privacy requirement if its $sup_s(X)$ violates $(k,\beta)^L$-privacy. We refer to these sequences as violating sequences. For instance, suppose we consider $k=2$, $\beta=1$, $L=2$, and $R_{SA}=\{Hepatitis, Cancer\}$ with $p_1 = conf(Hepatitis|T) = 0.2$ and $p_2 = conf(Cancer|T) = 0.2$, using Table 2. In this scenario, the sequence $X = \langle\{(Y,2017)\},\{(Y,2019)\}\rangle$ fails to meet the privacy requirement as the record $PID=6$ matches the sequence X and $|sup_s(X)|=1$, which is less than the specified threshold of k. Similarly, the sequence $X' = \langle\{(Y,2018)\},\{(Y,2018)\}\rangle$ also violates the privacy requirement as both records $PID=2$ and $PID=5$ match sequence X'. Both of these records have the sensitive value $Cancer$ in one of their visits. We have $q_1 = conf(Hepatitis|X') = 0$ and $q_2 = conf(Cancer|X') = 1$. Calculating $D(p_2,q_2)$ gives us $\frac{1-0.2}{0.2} = 4$, which exceeds $\min(\beta, -\ln(p_i)) = \min(1, -\ln(0.2)) = 1$.

One possible approach is to initially identify all potential violating sequences and then proceed with their removal. However, implementing this approach is impractical due to the excessively large number of violating sequences that would need to be enumerated. As a result, this approach becomes infeasible to execute effectively. Consider a violating sequence X with $|sup_s(X)| < k$. Any specific sequence of X', $X' <_S X$, with $|sup_s(X')| > 0$ in T is also a violating sequence because $|sup_s(X')| < |sup_s(X)| < k$. To efficiently identify privacy-violating sequences, we focus on a subset called minimal invalid sequences (MIS). An MIS is defined as a sequence that cannot be further generalized while still violating the privacy requirements.

Let us consider an example with $K=2$, $\beta=1$, $L=2$, and the SA set $R_s = \{Hepatitis, Cancer\}$. In this example, the sequence $S = \langle\{(Z,41001)\},\{(L,[4:6))\}\rangle$ violates $(k,\beta)^L$-privacy. This violation occurs because $D(p_i,q_i)=1.5$, which is greater than the specified threshold of $\min(\beta,-\ln(p_i))=1$ for *Hepatitis*. It is important to note that there is no more general sequence, such as $\langle\{(Z,4****)\}\rangle$, $\langle\{(Z,41***)\}\rangle$, $\langle\{(Z,410**)\}\rangle$, $\langle\{(Z,4100*)\}\rangle$, $\langle\{(L,[4:6))\}\rangle$ or $\langle\{(L,[4:8))\}\rangle$, that violates the privacy requirements. Hence, S qualifies as an MIS in this case.

Theorem 1. *A longitudinal database T satisfies $(k,\beta)^L$-privacy if and only if it does not contain any minimal invalid sequence (MIS).*

Proof. Let us consider a scenario where a database T does not meet the requirements for $(k,\beta)^L$-privacy, even if it does not contain an MIS. According to the definition of violating sequences, when a database violates the privacy requirements, it must contain violating sequences. However, as per the definition of MIS, a violating sequence either is an MIS itself or contains an MIS. This creates a contradiction with the initial assumption that the database T does not contain any MIS. Therefore, we can conclude that the database T must satisfy $(k,\beta)^L$-privacy because violating sequences cannot exist without the presence of an MIS. □

Algorithm 2 identifies MISs within cluster G_i. To clarify our proposed method, we introduce the concept of direct specializations (dsc). The direct specializations of a value g, denoted as $dsc(g)$, are defined as the set of values u in the domain D that are more specific than g ($u \leq_I g$), and there does not exist any value w in the domain such that $u \leq_I w$ and $w \leq_I g$. For example, dsc([2016:2023]) = {[2016:2019],[2020:2023]} regarding VH_Y in Fig. 2.

Similarly, the direct specializations of an event $e = \{g_1, g_2, ..., g_n\}$ are defined as $dsc_E(e) = \{\hat{e} = (\hat{g}_1, \hat{g}_2, ..., \hat{g}_n) |$ so that there is one value $i \in \{1, ..., n\}$ such that $\hat{g}_i \in dsc(g_i)$ and $\forall j \neq i, g_j = \hat{g}_j\}$.

Algorithm 2 iteratively generates MISs, starting with single events in lines 1 to 11, and expands to incorporate multiple events based on temporal relationships in lines 12 to 28. It begins by generating the most generalized event that covers the root of VH_{A_i}. For example with Table 2, the most generalized event is ⟨(Y,[2016:2023]), (Z,∗), (D,∗), (L,∗)⟩. Direct specializations of this event are stored in C_1. If an event is valid, it is added to the set V_1 for generating the next candidate set C_2. If an event is invalid, it is included in the set INV_1. The algorithm stores sequences containing i events in the sets C_i and V_i. C_2 is created by performing a self-join operation ($\bowtie$) on V_1 and subsequently eliminating more general sequences and those exceeding length L to reduce the search space for MIS identification. To generate candidate sequences in V_{i+1}, two sequences $X = \langle x_1, \ldots, x_i\rangle$ and $X' = \langle x'_1, \ldots, x'_i\rangle$ in V_i are joined if their first $i-1$ events are identical. The candidate sequence is created by appending the event x'_i to X based on specific conditions: if both x_i and x'_i have temporal information, x'_i can be added to X only if its temporal information is neither a descendant nor an ancestor of the temporal information in X; otherwise, x'_i can be added to X only if none of its items is a descendant or an ancestor of any of the items in X. For example, assume $L=4$, and we have sequences $X = \langle\{(Y,2018),(L,2)\},\{(Y,2019)\}\rangle$ and $X' = \langle\{(Y,2018),(L,2)\},\{(L,10)\}\rangle$. In this case, the candidate sequence $X'' = \langle\{(Y,2018),(L,2)\},\{(Y,2019)\},\{(L,10)\}\rangle$ is generated with three events and four items.

Algorithm 2: Identifying sequences that violate privacy.

Input: G_i: A set of sequences;
k, β, L : parameters of the privacy model
Output: Set of anonymized sequences satisfying the privacy model

r = The most general event based on $\{DH_{A_1}, ..., DH_{A_n}\}$
$C_1 = dsc_E(r)$ *// C_1 is a queue of events*
$INV_1 = \emptyset$ *// INV_i is a set of MISs*
$V_1 = \emptyset$ *// V_i = a set of non-violating sequences*
for each $e' \in C_1$ with $len(e) < L$
  delete e' from C_1
  If e' does not satisfy $(k, \beta)^L$-privacy in G_i
    $INV_1 = INV_1 \cup e'$
  else
    insert e' into V_1
    insert $dsc_E(e')$ into C_1
generate candidate set C_2 by $V_1 \bowtie V_1$
for each sequence $X \in C_2$ do
  If X is more general than any $v \in V_1$ or $len(X) > L$
    remove X from C_2
$i = 2$
while C_i is not empty
  compute $|sup_s(X)|$ and $conf(s_i|X)\ \forall\ X \in C_i,\ s_i \in R_s$
  for each sequence $X \in C_i$ with $|sup_s(X)| > 0$
    if X does not satisfy $(k, \beta)^L$-privacy in G_i
      $INV_i = INV_i \cup X$
    else
      insert into V_i
  i++
  generate candidate set C_i by $V_{i-1} \bowtie V_{i-1}$
  for each sequence $X \in C_i$ do
    if X is more general $\forall v \in V_{i-1}$ or $len(X) > L$
      remove X from C_i
$INV = INV_1 \cup ... \cup INV_{i-1}$

5.2. Removing the sequences that violate privacy

Algorithm 3 focuses on removing MISs from cluster G_i. In lines 1 to 8, a multi-domain generalization lattice is created to eliminate MISs within each cluster. Each node in the lattice, denoted as $N = [t_1, ..., t_n]$, corresponds to a vector of n QIs' domains. The values t_i represent the generalization level h for each QI, ranging from 0 to $height(DH_{A_i})$, where $l = 0$ corresponds to the root value and $l = height(DH_{A_i})$ corresponds to the leaf values of DH_{A_i}.

Algorithm 3: Removing the sequences that violate privacy.

Input: G_i: data set of sequences;
k, β, L : parameters of the privacy model
Output: Set of anonymized sequences satisfying the privacy model

// Second step: Removing the MISs
Node = a node with maximum domain generalization
while True
  children $= dsc_E$(*Node*)
  if $\nexists c \in$ *children* such that *InfLoss*(c) $\leq$ *InfLoss*(*Node*)
    break
  else if $\exists c \in$ *children* such that it removes all MIS in INV
    return c as solution of the sequences in G_i
  else *Node* =select $c \in$ *children* with maximum *Score*()
Update the set of INV based on *Node*
if $INV \neq \emptyset$
  supStrategy $= \emptyset$ *//This set stores the items to be suppressed*
  Items =store all items in INV
  Sort *Items* based on their *Score*()
  while $INV \neq \emptyset$
    insert x into *Items* with maximum *Score*() into *supStrategy*
    update INV according to x

Consider the data set in Table 2. A portion of the corresponding lattice is depicted in Fig. 4. Node $N = [Y1, Z1, D0, L0]$ in the lattice means that the values of Y and Z should be generalized to their level-1 descendants, and the values of D and L should be generalized to root level. Its direct generalization is the nodes $N1 = [Y2, Z1, D0, L0]$, $N2 = [Y1, Z2, D0, L0]$, $N3 = [Y1, Z1, D1, L0]$, and $N4 = [Y1, Z1, D0, L1]$.

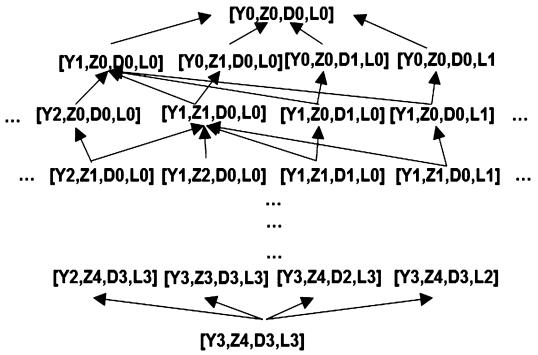


Fig. 4. Generalization lattice of Table 2.

The algorithm uses a top-down search to construct the lattice, starting from the maximum generalization node. Using direct generalization, child nodes $N_0 \in dsc_E(N)$ are generated and evaluated to eliminate extracted MISs. An MIS, X, can be eliminated by N_0 if at least one item's generalization level in X is higher than in N_0. For example, if $X = \langle\{(L, [0:1))\}\rangle$ is an MIS and the generalization level of L at N_0 is 1, X will be removed based on N_0's generalization. In line 8 of Algorithm 3, the child node with the maximum score is selected. For cluster G_i, the score of the sequence set Q under N_0 is given by $Score(N_0) = \frac{|MIS|}{infLoss}$, where $|MIS|$ represents the number of MIS eliminable by N_0, and *infLoss* denotes the normalized information loss incurred in Q due to anonymization operations on N_0.

A higher-scoring node is better for expansion as it removes more MISs with less information loss. Detailed descriptions of information loss measures can be found in Section 6.1. The search in the generalization lattice stops when the information loss of all child nodes surpasses that of the current node. At line 9 of Algorithm 3, the set INV is updated using the generalization strategy associated with *Node*.

If any MISs remain, a suppression method is employed (lines 10-16). To determine the optimal suppression strategy, all items in INV that cannot be eliminated through generalization are sorted in descending order based on their scores. The score of an item g indicates the number of MISs that contain g and can be eliminated by suppressing it at the cost of incurring information loss due to both the generalization strategy of *Node* and the suppression of items in $\{g\} \cup supStrategy$. The item scores take into account MISs that contain descendant items. In an iterative process, we add the item g with the highest score to the *supStrategy* set, we remove MISs containing g from the list, and we update the scores of the remaining items. This continues until no more MISs remain. Generalization and suppression strategies may differ across clusters in the *Group*.

5.3. Time complexity

The time complexity of the FCLA algorithm can be analyzed as follows: In line 1 of Algorithm 1, FCLA initializes the distance matrix with the distances of all singleton clusters using sequence alignment techniques, resulting in a time complexity $O(t^2 * w^2)$, where t is the number of sequences, and w is the maximum sequence length when dynamic programming is employed for sequence alignments. The algorithm generates clusters based on this distance matrix, resulting in a time complexity of $O(t^2)$ for line 1 of Algorithm 1.

This line produces several clusters, with a maximum number of $O(\frac{t}{k})$ clusters. Each cluster contains at least k sequences, satisfying the privacy model constraints. Lines 3 and 4 of the algorithm are executed for each cluster G_i. We enumerate all sequences of length 1 to L where L is the power of an adversary in $(K,\beta)^L$-privacy model. In the worst case, this enumeration incurs a computational complexity of $O(n * Height(DH_{A_i}))$, where A_i is a QI with the highest domain hierarchy, and n is the number of QIs. Removing MISs with a generalization lattice has a complexity of $O(|H| * |MIS|)$, where $|H|$ denotes the number of nodes in the generalization lattice, and $|MIS|$ represents the number of extracted MISs. The total cost complexity of FLCA is $O(t^2 + \frac{t}{k}\{n * Height(DH_{A_i}) + |H| * |MIS|\})$. While pruning strategies can significantly reduce this complexity in practice, removing MISs through suppression has a similar computational cost.

When comparing with the closest existing work, HALT [12], it must be noted that HALT searches for violating sequences throughout the entire data set with a time complexity $O(F^L)$, where F represents the number of non-minimal invalid sequences with length 1 in data set T and L is the power of an adversary in HALT's privacy model. In contrast, our method operates in parallel on clusters with sizes $k < |G_i| < t$. Thus, FCLA outperforms HALT in terms of time complexity. In the worst-case scenario where FCLA generates one cluster with a size $|t|$, the time complexity of our method is comparable to that of HALT.

6. Experimental results

In this section, we present an experimental evaluation of our anonymization framework. Given that our study is the first to tackle the challenge of mitigating skewness attacks in privacy-preserving longitudinal data, direct comparisons with the techniques proposed by Terrovitis and Mamoulis [24] and Tamersoy et al. [17], primarily aimed at protecting against identity disclosure, are not pertinent. Nevertheless, we conducted experiments to compare our proposed anonymization framework with state-of-the-art privacy protection

Table 3
Privacy parameters used in the experiments.

k values	β values	L values
[5:45]	[1:7]	[2:16]

methods closely related to ours, such as HALT [12]. HALT's primary objective is to achieve k-anonymity while imposing constraints on the percentage C of records containing sensitive values, to prevent both identity and attribute disclosures. To ensure a fair and meaningful comparison with other methodologies, we adopted the same parameter values as those specified in [12]. Specifically, we varied the privacy parameters, including k, β, and L, within the following ranges: $[5:45]$ for k, $[1:7]$ for β, and $[2:16]$ for L (see Table 3). Experiments were conducted on a workstation with an Intel Core i5 processor and 8 GB RAM, running at 2.4 GHz. Our algorithms were implemented in Python.

The experiments were performed on three standard health data sets:

- CMS DE-SynPUF, provided by the Centers for Medicare and Medicaid Services (CMS), contains three years (2008 to 2010) of synthetic claims for hospital inpatient services. Each record represents a synthetic inpatient claim (hospital visit). Due to file size limitations, CMS released inpatient data in 20 separate samples. Each file contains 81 attributes, including sensitive information related to patient diagnoses and procedures. To protect privacy, we focused on one sensitive attribute, Diagnosis. For each patient visit, the diagnosis attribute is determined based on values from several related attributes. If any of these attributes contains highly sensitive data, the diagnosis attribute is set to that value; otherwise, a random value from these attributes is assigned to the diagnosis attribute. Some attributes such as CLM_FROM_DT and CLM_THRU_DT representing claim start and end dates, were almost identical to CLM_ADMSN_DT (admission date) and NCH_BENE_DSCHRG_DT (discharge date) in about 99.8% of records in the 20 sample files. As a result, we discarded CLM_FROM_DT and CLM_THRU_DT considering CLM_ADMSN_DT and NCH_BENE_DSCHRG_DT as quasi-identifiers. We also included CLM_UTLZTN_DAY_CNT (length of hospital stay) and CLM_PMT_AMT (claim payment amount) as potential quasi-identifiers, while excluding other payment-related attributes due to low variability or lack of quasi-identifier characteristics. The remaining attributes in the original DE-SynPUFs data are neither sensitive nor quasi-identifiers, and thus, they do not affect the anonymization process. In our study, we identified certain values such as HIV, abortion, abuse, psychosexual disorders, mental retardation, or plastic surgery as more sensitive. These specific values, referred to as the set π mentioned in Section 4, require protection to prevent attribute disclosure. Additionally, we introduced an extra quasi-identifier, DSFC, as suggested by [12], which tracks the number of days since the initial claim calculation for each patient in each calendar year. We generated five data sets using sample files from CMS DE-SynPUF. Our algorithm was applied to these data sets, and the results were averaged. Each data set comprised 66,253 visits, with a maximum of 12 visits per patient, an average of 8 visits per patient, and 8.12% of high sensitive values in a data set.
- The Heritage Health Prize (HHP) claims data set was part of a data mining competition aimed at predicting patients' hospitalization duration using historical claims data, which included various tables like claims and lab data. Our study focused on HHP claims data, encompassing patient claims over three years. In [11], the anonymization of HHP claims data to prevent identity disclosure attacks was examined. The focus was on attributes Specialty, Place of service (PlaceSvc), Length Of Stay in the hospital (LOS), Days Since the First Service that year (DSFS), (admission) Year, Diagnosis, and CPTCode. We chose DSFS and LOS as our quasi-identifiers. PlaceSvc and Specialty were discarded due to unavailable original values. Instead, we used patient zip codes and claim years as additional quasi-identifiers for each claim. As with the previous data set, we used Diagnosis as the sensitive attribute. The set of highly sensitive values, π, was chosen to be similar to the CMS DE-SynPUF dataset. To create a diverse set of synthetic data sets, we generated five variations, each one encompassing 1,000 patients. The average number of events (*i.e.*, claims) per sequence was set at 10, with a maximum of 16 events.
- FDA Adverse Event Reporting System (FAERS) datasets, which is a SRS system provided by U.S. Food and Drug Administration (FDA) and released quarterly. Each report in FAERS is uniquely identified by an attribute called ISR, and contains an attribute CaseID to identify distinct individuals, along with some demographic information such as Weight, Age, and Gender, drug name (Drug) and indication ($INDI_PT$), and reaction information (PT). We used Weight, Age, Gender as quasi-identifiers (QIDs) and CaseID as the individual identifier. We utilized the drug name and created a new feature named "disease". Disease values were constructed by referencing websites such as Drugs.com and wrongdiagnosis.com, which provide information on which drugs are effective for specific diseases. Datasets from 2004Q1 to 2011Q4 were selected to do experiments, where any record with QID containing missing values was discarded. To create a diverse set of synthetic data sets, we generated five variations, each one encompassing 1,000 patients. The average number of events (*i.e.*, ISR) per sequence was set at 4, with a maximum of 7 events.

Note that The value of β is determined by expert knowledge and context-dependent factors. In this work, we set β based on the percentage of more sensitive values within each dataset. For instance, in the CMS DE-SynPUF dataset, we identified several values, including those related to HIV, abortion, abuse, psychosexual disorders, mental retardation, and plastic surgery, as more sensitive, consistent with [12]. These more sensitive values constitute a maximum of 8.12% of the dataset. Setting $\beta = 1$ results in a frequency threshold of $e^{-\beta} \approx 37\%$, classifying all more sensitive values as "infrequent". This allows the frequency of any sensitive attribute value within any group to be at most $8.12\% * 2 = 16.24\%$. Consequently, $\beta = 1$ represents a relatively small value. For our experiments, we investigate β values within the range 1, 2, 3, 4, 7. In our privacy model, it is essential to satisfy the second condition for sequences

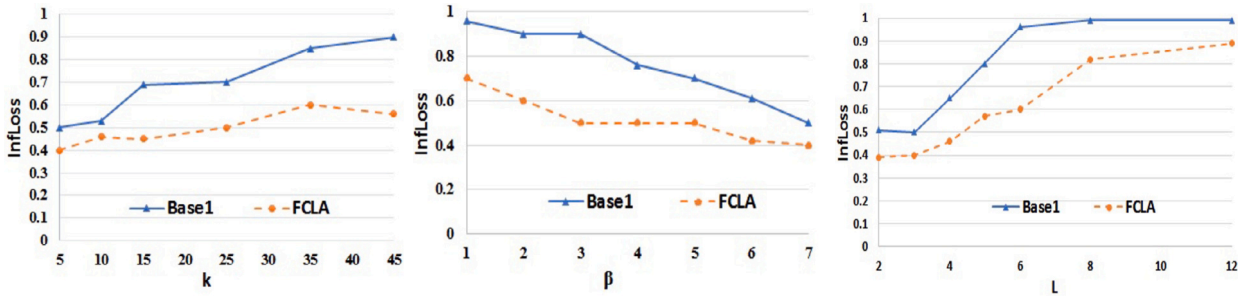


Fig. 5. Information loss of FCLA vs Base1 on the DE-SynPUFs data set when: (a) varying k values, $\beta = 6$, $L = 3$; (b) varying β values, $k = 5$, $L = 3$; (c) varying L values, $k = 5$, $\beta = 6$.

containing at least one value from π. This ensures enhanced privacy protection for the specified subset of data. Sequences containing no more sensitive values do not satisfy the second condition

6.1. Information loss measures

Different measures have been proposed to quantify information loss incurred by anonymization [31]. We use the normalized certainty penalty [32], denoted as $NCP(A_i, u)$, to measure the information loss of value u from the value generalization hierarchy of A_i. The information loss at node u is determined by the ratio of the subtree size rooted at u to the total number of leaves in the value generalization hierarchy of A_i. If u is a leaf node, the subtree size is considered to be 0. For example, for the value generalization hierarchy of attribute Y and the node [2016:2019], the information loss is calculated as $4/8 = 0.5$, which indicates that [2016:2019] covers four leaf values (2016, 2017, 2018, and 2019), while the total domain size of Y is 8 (2016 to 2023). In a set of sequences S, each sequence comprises multiple events, with each event containing an attribute value A_i. The overall information loss incurred by the generalization strategy *Node* on a set of sequences, Q, is defined as:

$$InfLoss(Node) = \Sigma_{X \in Q} \Sigma_{e \in X} \Sigma_{A_i \in A} NCP(A_i, u) \tag{1}$$

where X denotes a sequence in Q, $Q \subseteq T$, e represents an event in X, A refers to the set of QIs, and $NCP(A_i, u)$ represents the information loss at value u of attribute A_i.

We utilize a measure introduced by [12] to assess the information loss caused by suppression. When a leaf value u is suppressed, the information loss is 1. The total loss for suppressing all instances of u is given by $count(u)$. For non-leaf values in the value generalization hierarchy, the information loss for suppressing one instance of u is 1 - $NCP(A_i, u)$. Consequently, with a suppression strategy *supStrategy* that involves suppressing multiple items, the total information loss on a set of sequences S can be calculated as

$$InfLoss(supStrategy) = \Sigma_{u \in supStrategy} count(u) \cdot (1 - NCP(A_i, u)) \tag{2}$$

The normalized total information loss on a set of sequences S, considering the combined information loss from both the generalization and suppression strategies, is calculated for every possible anonymization solution by using the following expression:

$$InfLoss = \frac{InfLoss(Node) + InfLoss(supStrategy)}{|S| * |A|}. \tag{3}$$

6.2. Performance of our proposed method

We evaluate the utility preservation of FCLA compared to a baseline algorithm named Base1. Base1 is a clustering-based approach similar to FCLA but without dynamic programming for distance calculations. Instead, Base1 utilizes a list of representative event pairs to compute distances [30]. In the following, we assess the impact of different privacy parameters on the information loss of anonymized data.

Figs. 5(a), 6(a), and 7(a) show the information loss analysis of FCLA and Base1 algorithms on the CMS DE-SynPUF, HHP claims, and FAERS data sets. The evaluation was conducted for $\beta = 6$, $L = 3$, and values of k varying from 5 to 45. Increasing k resulted in higher information loss for both algorithms. However, it is noteworthy that FCLA occasionally exhibited a decrease in information loss for higher k values due to its greedy nature. All in all, FCLA's information loss was lower than that of Base1. This can be attributed to FCLA's utilization of dynamic programming for sequence alignment. This technique helps FCLA preserve data utility by maintaining intrinsic characteristics and patterns.

In Figs. 5(b), 6(b), and 7(b), the impact of the different values of β on information loss is shown. With $k = 5$ and $L = 3$, increasing β led to a decrease in information loss. Smaller values of β resulted in a higher number of MIS within the data set. Consequently, more data distortion was required to eliminate these MIS during anonymization. By increasing β, FCLA prioritized reducing information loss by minimizing data distortion. This trade-off between privacy protection and data utility is reflected in the decreased information loss.

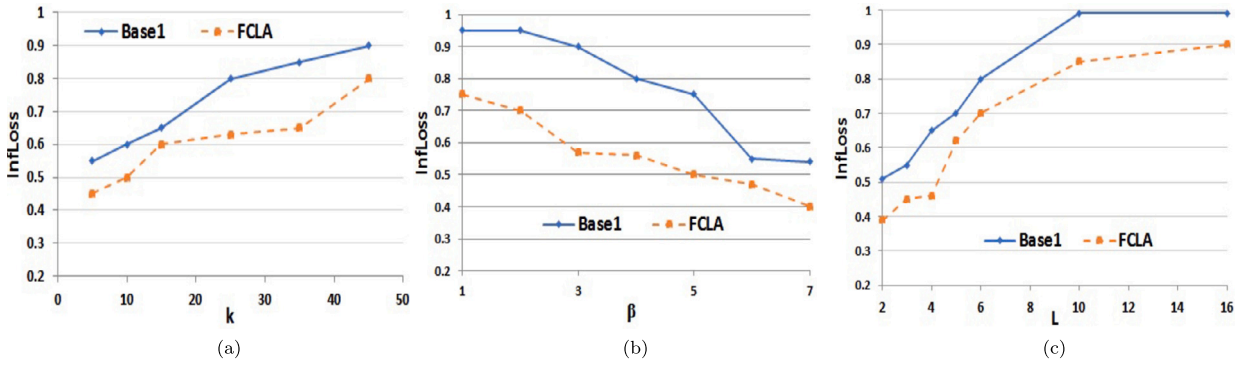


Fig. 6. Information loss of FCLA vs Base1 on the Heritage Health Prize claims data set when: (a) varying k values, $\beta = 6$, $L = 3$; (b) varying β values, $k = 5$, $L = 3$; (c) varying L values, $k = 5$, $\beta = 6$.

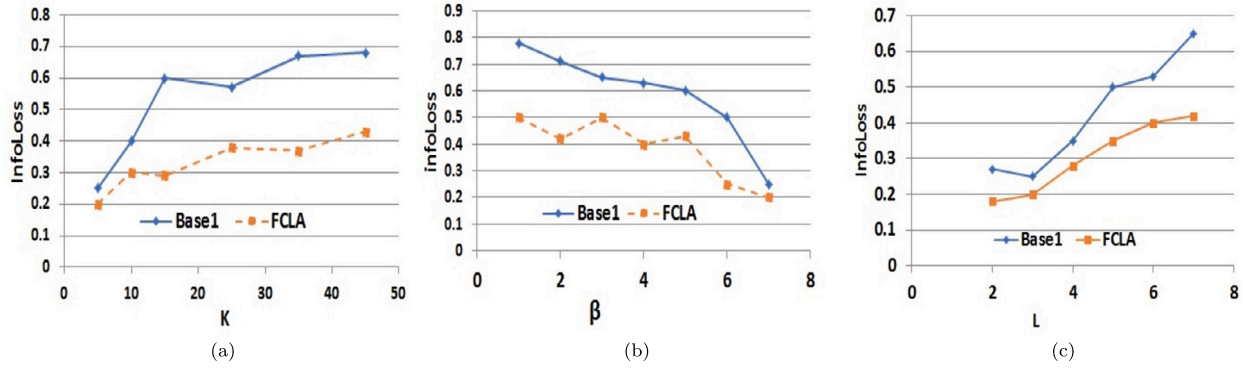


Fig. 7. Information loss of FCLA vs Base1 on the FAERS Dataset when: (a) varying k values, $\beta = 6$, $L = 3$; (b) varying β values, $k = 5$, $L = 3$; (c) varying L values, $k = 5$, $\beta = 6$.

In Figs. 5(c), 6(c), and 7(c), the impact of the adversary's background knowledge on data utility and information loss is shown. The length of the adversary's knowledge (L) was varied from 2 to the maximum length of the events in the corresponding data set, which were 12, 16, and 7, respectively, while k was fixed at 5 and β was fixed at 6. It is observed that for smaller values of L ($L = 2$ and $L = 3$), the information loss was relatively low. This indicates a lesser presence of MIS of size 2 and 3 in the data, resulting in lower information loss during anonymization. In this experiment, the maximum value of length referred to attackers with perfect knowledge, meaning that the length of the known sequences was set to the maximum length in the data set. The results show that the information loss of FCLA was less than that of Base1. However, as L increased, the number of extracted MIS grew, which required more extensive generalizations and suppressions to eliminate them, and hence led to larger information loss. Higher L values indicate an adversary with increased background knowledge. Consequently, more data distortion was required to safeguard privacy, thereby resulting in increased information loss.

Figs. 5(c), 6(c), and 7(c) highlight the importance of considering the adversary's background knowledge when determining the appropriate level of anonymization. Understanding the extent of the adversary's knowledge allows data controllers to make informed decisions about the necessary level of generalization and suppression to achieve a balance between privacy protection and data utility.

6.3. Comparison with other methods

We conducted a comparative analysis of our results with the closely related HALT method [12]. HALT is designed to anonymize longitudinal data while addressing identity and attribute disclosure concerns. However, a direct and fair comparison with HALT was challenging due to the utilization of different privacy models. Nevertheless, our framework offers a $(k, c)^L$-privacy guarantee, similar to that of

HALT. Condition 2 in our proposed model differs from Condition 2 in the $(k, c)^L$-privacy model. In $(k, c)^L$-privacy, Condition 2 ensures that the percentage of records in the set of sequences, Q, containing the sensitive attribute (SA) value s_i for any SA values, is below a specified threshold C. In FCLA, we address Condition 2 of the $(k, c)^L$-privacy model in line 1 of Algorithm 1. Moreover, line 7 of Algorithm 2 in FCLA also ensures adherence to the $(k, c)^L$-privacy model.

We took parameter values adjusted in accordance with the recommendations provided in [12]. Firstly, we compared the performance of the two methods in terms of information loss. Fig. 8 presents experimental results of HALT and FCLA on the evaluation data set with different parameter values. In Fig. 8(a), FCLA outperforms HALT in terms of information loss for varying values of k ($C = 0.7$, $L = 3$), thus achieving lower information loss and better data utility. Similarly, in Fig. 8(b), FCLA consistently demonstrates

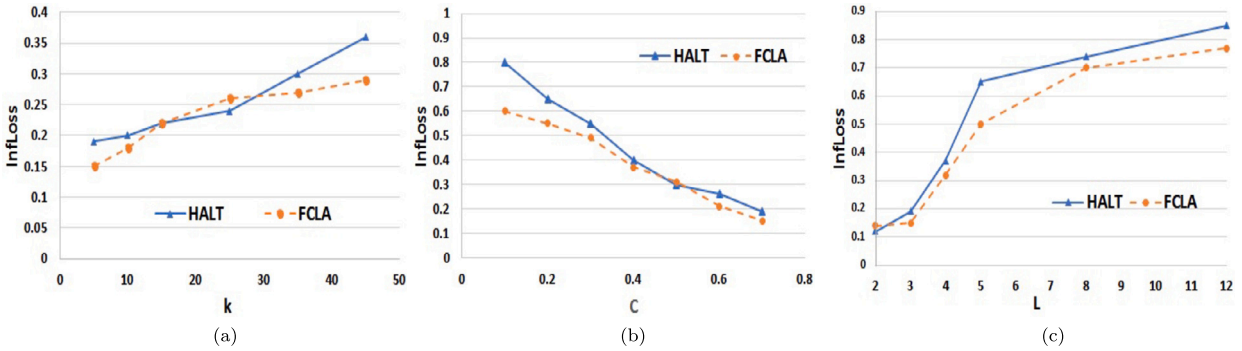


Fig. 8. Information loss of FCLA vs HALT on the DE-SynPUFs data set when: (a) varying k values, $C = 0.7$, $L = 3$; (b) varying C values, $k = 5$, $L = 3$; (c) varying L values, $k = 5$, $C = 0.7$.

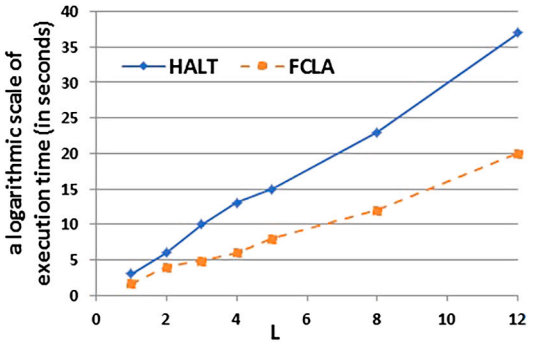


Fig. 9. Comparison of the execution times of FCLA and HALT on the DE-SynPUFs data set when varying L values, with $k = 5$ and $C = 0.7$.

lower information loss compared to HALT for different values of C (with $k = 5, L = 3$). Lastly, Fig. 8(c) shows that FCLA exhibits better performance with reduced information loss compared to HALT for different values of L ($k = 5, C = 0.7$). In this experiment, L was varied from 2 to 12, as the maximum length of sequences in the CMS DE-SynPUF data set was 12.

In particular, this means that for $L = 12$ we conducted the experiment with an attacker who had perfect knowledge. By doing so, we effectively demonstrated the superior utility benefits of our methods compared to HALT in scenarios where the attacker possesses complete knowledge. These results highlight the effectiveness of our framework at preserving data utility while meeting the privacy requirements of HALT. Under the same privacy model, HALT identifies patterns that violate the privacy model across the entire data set and generalizes the data without considering the similarity of QI values and data utility. In contrast, our method partitions sequences into groups with similar QI values, thereby preserving utility and handling violating sequences independently within each group. In this way, our method is able to generalize groups in parallel, resulting in reduced time costs. Notably, our approach differs from HALT in terms of identifying and removing violating sequences.

Finally, we compared the execution times of our algorithm and HALT on the CMS DE-SynPUF data set.

To do that, we set $k = 5$ and $C = 0.7$. The L values were selected within the range $[1 : 12]$, as the maximum length of a sequence in the CMS DE-SynPUF data set was taken to be 12. Fig. 9 displays the execution times of both FCLA and HALT for various values of L. It is important to emphasize that the Y-axis is represented on a logarithmic scale. Our findings indicate a substantial performance advantage for FCLA over HALT. We next give an explanation for this performance difference. HALT identifies violating sequences by extending a given sequence S through two distinct procedures: one appends a new event consisting of a single item to the end of S, and the other adds a single item to the last event of S. On the other hand, HALT considers all items in the entire data set during both of these procedures. In contrast, our method focuses exclusively on items within a specific cluster, denoted as G_i, where $k \leq |G_i| \leq t$. Furthermore, FCLA employs a search strategy that involves finding violating sequences by joining violating sequences. This results in a smaller search space compared to HALT. It is worth noting that Fig. 9 provides empirical evidence that confirms the time complexity analysis presented in Section 5.3.

7. Conclusions and future work

We have proposed an anonymization framework for longitudinal data, specifically designed for scenarios where an adversary's knowledge is limited to a certain number of events. Our framework is adaptable, allowing for different adversarial models with varying degrees of background knowledge, thereby accommodating diverse privacy requirements. Unlike existing approaches, our method effectively mitigates identity disclosure, attribute disclosure, and skewness attacks in an integrated manner. Additionally, by partitioning data into sequence groups, our approach enables efficient parallel execution, enhancing scalability.

Empirical evaluations demonstrate that our framework outperforms the most similar state-of-the-art method in preserving both privacy and data utility while also achieving superior computational efficiency.

As future work, we plan to explore data republishing strategies for longitudinal data, addressing a critical gap in existing dynamic data publishing research (e.g., [33], [34], and [13]). In practice, data sources are continuously evolving, leading to regular updates and republication of datasets. Current approaches do not adequately handle the unique challenges of anonymizing longitudinal data in such dynamic settings. As a result, privacy breaches may arise not only from individual data releases but also from adversaries aggregating information across multiple releases.

CRediT authorship contribution statement

Fatemeh Amiri: Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Data curation, Conceptualization. **David Sánchez:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Josep Domingo-Ferrer:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Partial support for this work has been received from the European Commission (project H2020-871042 “SoBigData++”), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez, and grant 2021SGR-00115), MCIN/AEI/10.13039/501100011033 and “ERDF A way of making Europe” under grant PID2021-123637NB-I00 “CURLING”, and the EU’s NextGenerationEU/PRTR via INCIBE (project “HERMES” and INCIBE-URV cybersecurity chair).

Data availability

Data will be made available on request.

References

[1] Scottish Longitudinal Study, https://calls.ac.uk/ls-units/scottish-longitudinal-study-scotland/.
[2] M.E. Sozio, A. Procopio, R. Samuel, Youth Survey Luxembourg. Technical report 2019, Ministry of National Education, Children and Youth (MENJE) and University of Luxembourg (UL), Esch-sur-Alzette, 2020.
[3] C. Residori, M.E. Sozio, L. Schomaker, R. Samuel, YAC Young people and Covid-19. Preliminary results of a representative survey of adolescents and young adults in Luxembourg, University of Luxembourg (UL) and Ministère de l’Éducation nationale, de l’Énfance et de la Jeunesse (MENJE), Esch-sur-Alzette, 2020.
[4] Framingham Heart Study (FHS), https://www.framinghamheartstudy.org/.
[5] S. Szefler, S. Weiss, A. Tonascia, N. Adkinson, B.R.C. Bender, M. Donithan, H. Kelly, J. Reisman, G. Shapiro, G. Sternberg, R. Strunk, V. Taggart, M. VanNatta, R. Wise, M. Wu, R. Zeiger, Long-term effects of budesonide or nedocrimil in children with asthma, N. Engl. J. Med. 343 (15) (2000) 1054–1063.
[6] A.S. Coletti, P.J. Heagerty, A.R. Sheon, M. Gross, B.A. Koblin, D.S. Metzger, G.R. Seage, Randomized, controlled evaluation of a prototype informed consent process for HIV vaccine efficacy trials, J. Acquir. Immune Defic. Syndr. 32 (2003) 161–169.
[7] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression, Technical Report. SRI-CSL-98-04, 1998.
[8] J. Cao, P. Karras, Publishing microdata with a robust privacy guarantee, Proc. VLDB Endow. 5 (11) (2012) 1388–1399.
[9] N. Li, T. Li, S. Venkatasubramanian, T-closeness: privacy beyond k-anonymity and l-diversity, in: 2007 IEEE 23rd International Conference on Data Engineering, 2006, pp. 106–115.
[10] P. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, L-diversity: privacy beyond k-anonymity, ACM Trans. Knowl. Discov. Data 1 (1) (2007) 3.
[11] K. El Emam, L. Arbuckle, G. Koru, L. Gaudette, E. Neri, S. Rose, J. Howard, J. Gluck, De-identification methods for open health data: the case of the Heritage Health Prize Claims data set, J. Med. Internet Res. 14 (1) (2012), https://doi.org/10.2196/jmir.2001.
[12] M. Sehatkar, S. Matwin, HALT: hybrid anonymization of longitudinal transactions, in: 2013 Eleventh Annual Conference on Privacy, Security and Trust, 2013, pp. 127–134.
[13] J.T. Wang, W.Y. Lin, Privacy-preserving anonymity for periodical releases of spontaneous adverse drug event reporting data: algorithm development and validation, JMIR Med. Inform. 9 (10) (2021) e28752.
[14] M. Terrovitis, N. Mamoulis, P. Kalnis, Privacy-preserving anonymization of set-valued data, Proc. VLDB Endow. 1 (1) (2008) 115–125.
[15] Y. Xu, K. Wang, A.W-C. Fu, P.S. Yu, Anonymizing transaction databases for publication, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 767–775.
[16] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: a survey on recent developments, ACM Comput. Surv. 42 (2010) 4.
[17] A. Tamersoy, G. Loukides, M.E. Nergiz, Y. Saygin, B. Malin, Anonymization of longitudinal electronic medical records, IEEE Trans. Inf. Technol. Biomed. 16 (2012) 413–423.
[18] P. Fournier-Viger, W. Gan, Y. Wu, M. Nouioua, W. Song, T. Truong, H. Duong, Pattern mining: current challenges and opportunities, in: International Conference on Database Systems for Advanced Applications, 2022, pp. 34–49.
[19] A. Gkoulalas-Divanis, G. Loukides, Utility-guided clustering-based transaction data anonymization, Trans. Data Priv. 5 (2012) 223–251.
[20] T. Tassa, A. Mazza, A. Gionis, k-Concealment: an alternative model of k-type anonymity, Trans. Data Priv. 5 (2012) 189–222.
[21] F. Jin, W. Hua, M. Francia, P. Chao, M.E. Orlowska, X. Zhou, A survey and experimental study on privacy-preserving trajectory data publishing, IEEE Trans. Knowl. Data Eng. 35 (2023) 5577–5596.

[22] X. Liu, L. Wang, Y. Zhu, SLAT: sub-trajectory linkage attack tolerance framework for privacy-preserving trajectory publishing, in: 2018 International Conference on Networking and Network Applications (NaNA), 2018, pp. 298–303.
[23] L. Yao, X. Wang, X. Wang, H. Hu, G. Wu, Publishing sensitive trajectory data under enhanced L-diversity model, in: 20th IEEE International Conference on Mobile Data Management, 2019, pp. 160–169.
[24] M. Terrovitis, N. Mamoulis, Privacy preservation in the publication of trajectories, in: 9th International Conference on Mobile Data Management, 2008, pp. 65–72.
[25] N. Ruiz, A general framework and metrics for longitudinal data anonymization, in: Privacy in Statistical Databases, in: Lecture Notes in Computer Science, vol. 11126, 2018, pp. 215–230.
[26] P. Parameshwarappa, Z. Chen, G. Koru, Anonymization of daily activity data by using l-diversity privacy model, ACM Trans. Manag. Inf. Syst. 12 (2021) 1–21.
[27] D. Kunkle, D. Zhang, G. Cooperman, Mining frequent generalized itemsets and generalized association rules without redundancy, J. Comput. Sci. Technol. 23 (1) (2008) 77–102.
[28] W.Y. Lin, D.C. Yang, J.T. Wang, Privacy preserving data anonymization of spontaneous ADE reporting system dataset, in: Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics, 2015, pp. 2–16.
[29] K. Wang, B.C.M. Fung, P.S. Yu, Handicapping attacker's confidence: an alternative to kanonymization, Knowl. Inf. Syst. 2006 (2006).
[30] J. Domingo-Ferrer, S. Martínez, D. Sánchez, Decentralized k-anonymization of trajectories via privacy-preserving tit-for-tat, Comput. Commun. 190 (2022) 57–68.
[31] F. Amiri, N. Yazdani, A. Shakery, A.H. Chinae, Hierarchical anonymization algorithms against background knowledge attack in data releasing, Knowl.-Based Syst. 101 (2016) 71–89.
[32] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. Fu, Utility-based anonymization using local recoding, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 785–790.
[33] F. Amiri, N. Yazdani, A. Shakery, S.S. Ho, Bayesian-based anonymization framework against background knowledge attack in continuous data publishing, Trans. Data Priv. 12 (3) (2019) 197–225.
[34] S. Li, H. Tian, H. Shen, Y. Sang, Privacy-preserving trajectory data publishing by dynamic anonymization with bounded distortion, ISPRS Int. J. Geo-Inf. 10 (2) (2021) 78.