# Working Paper: Short Summary of Syntactic Privacy

Niklas Zapatka[1], Joshua Stock[1], Hannes Federrath[1], and Jens Lindemann[1]

[1]Universität Hamburg

22nd December 2023

**Abstract**

Balancing the trade-off between personal data processing and the right to privacy requires a measure for privacy. Such measures are provided by formal models, which also enable proofs of privacy preservation. Syntactic privacy has been praised as an intuitive family of such models. This report reviews the literature on syntactic privacy models and their shortcomings. The results reveal their inherent complexity and fragility. A short review of open-source implementations supporting syntactic privacy is also included.

## 1 Introduction

In recent years, data-driven innovations such as machine learning have steadily increased the demand for data. If an application requires personal data, this demand conflicts with every individual's right to privacy. Resolving this trade-off requires a measure or definition of privacy. A formal approach is desirable because it allows to prove that privacy is accounted for.

Syntactic privacy models such as $k$-anonymity [59] are one attempt to address this. They define privacy as a property of a data set. Any data which complies with this criterion is considered privacy-preserving and can be processed or published. This promise, combined with their intuitiveness, have contributed to the popularity of syntactic privacy models.

This report reviews the literature on syntactic privacy. It focuses on the proposed models, algorithms and their shortcomings. The results demonstrate that the formal guarantees are fragile and dependent on a complex set of assumptions. This contrasts their apparent intuitiveness and motivates the use of a much more robust definition, differential privacy [26].

Preliminaries are established in Section 2. Section 3 outlines the historical development of syntactic privacy models. New models have been proposed in response to discovered attacks on their predecessors. The algorithms and implementations which put these models into practice are presented in Sections 4 and 5. Remaining shortcomings are reviewed in Section 3.4. Finally, Section 6 concludes and summarizes our results.

Table 1: A tabular data set.

| Name | Mail | Age | Profession | Diagnosis |
|---|---|---|---|---|
| Alice | alice@example.com | 27 | accountant | burn-out |
| Bob | abc@example.org | 35 | teacher | cancer |

# 2 Background

Before we review the research on syntactic privacy, we present its foundations. The field is rooted in statistical disclosure control, which aims at preventing privacy invasions from published statistics. This is demonstrated by the first publications [59, 60] on syntactic privacy, which view syntactic privacy as an improvement over existing statistical disclosure control approaches. Consequently, its underlying data model, privacy breach definitions and principal means to prevent such breaches are inherited from the statistics community.

## 2.1 Data Model

Syntactic privacy models [10, 43, 46, 48, 59, 67] assume that the data is a table. An example is shown in Table 1. There is exactly one row or *record* for each individual described by the data. These records consist of cells or *attributes* such as names or professions. Three types of attributes are distinguished.

- *(Direct) identifiers* such as names and mail addresses identify the individual described by a record directly.
- *Quasi-identifiers* [20] are known by an individual's surroundings or the general public. Examples include age and profession. An alternative definition [59] considers attributes whose values can appear in previously published data sets.
- *Sensitive* attributes [7] of an individual should remain private, for instance a diagnosis.

Syntactic privacy assumes that attributes are classified into these three groups without any overlaps and that the classification is the same for all records. Furthermore, it is commonly [43, 46, 66, 67] assumed that there is exactly one sensitive attribute.

Roughly speaking, the objective of syntactic privacy is preventing that information from a released data set is linked to the described individuals. This is achieved via *anonymization*. Identifiers must be removed to prevent trivial re-identification. However, unique quasi-identifier combinations may still identify individuals [20], which was confirmed experimentally: 87 % of the US population are uniquely identified by their sex, date of birth and postal code [60]. Hence, syntactic privacy anonymize data by modifying the quasi-identifiers. The sensitive attributes are left unaltered because they are viewed as unrelated to the re-identification risk. Consequently, misclassifying an attribute might enable privacy breaches. This caveat is elaborated in Section 3.4.2.

We formalize the resulting data model as follows. A data set $X = \{x_1, ..., x_n\}$ is a finite multiset of $n$ records. Every record $x_i = (q_1, ..., q_m, s)$ is a vector of $m$ quasi-identifier attribute values $q_j$ and a single sensitive value $s$. Note that finite attribute domains can be assumed because $X$ is finite.

## 2.2 Definitions of Disclosure

The previous section simplified the objective of syntactic privacy as preventing privacy breaches. Informally, a privacy breach occurs when a data set discloses some information it is not supposed to. This led to different definitions of *disclosures* in the literature, which are discussed in the following.

The first definition is *statistical disclosure* by Dalenius [19]. It defines disclosure as any event where an adversary can predict the sensitive attribute value of an individual more accurately with access to the data set than without it. This definition is close to above informal concept of privacy breaches. However, Dwork [26] provided a formal proof that statistical disclosure is impossible to prevent if the data set retains any utility.

Lambert [40] distinguishes *identity* and *attribute disclosure*. The former is equivalent to re-identification and occurs when the adversary can link a record to an individual. The latter describes events where the adversary links a sensitive value to an individual. Note that identity disclosure implies attribute disclosure as long as the sensitive attributes are not modified. However, the reverse is not true as illustrated by homogeneity attacks, which are discussed in Section 3.

Nergiz, Atzori and Clifton [48] added a third disclosure type to identity and attribute disclosure. Their definition describes the event where an adversary learns that an individual's data is contained within the data set. This knowledge may be combined with context information about the data set. For instance, inferring that a patient's record appears in a data set of cancer patients reveals their health status. While this disclosure was initially unnamed by For instance, knowing that someone's record is contained in a cancer research data set While Nergiz, Atzori and Clifton did not introduce a term for these disclosures, it is called *membership disclosure* in subsequent literature. To the best of our knowledge, this term was first introduced in [44].

These definitions differ in what kind of information is concerned. The following three distinctions have also been proposed in the literature. Dalenius [19] distinguishes *exact disclosure*, where the adversary's prediction is certain, and *approximate disclosure*, where some uncertainty remains. Lambert [40] observed that the adversary's prediction is either correct or not, leading to *true* and *false disclosures*. Finally, in the context of attribute disclosure, Machanavajjhala et al. [46] introduced the terms *positive disclosure*, where the adversary infers the attribute value, and *negative disclosure*, where they only eliminate a possible value.

Note the type of information and these three distinctions are all orthogonal to each other. This leads to a four-dimensional taxonomy of disclosures. For instance, if an adversary erroneously infers that an individual does not have the health status *healthy*, a false negative exact attribute disclosure occurred. Syntactic privacy models are designed to protect from certain disclosures within this taxonomy or some equivalent concepts.

## 2.3 Approaches to Sanitization

Preventing disclosures requires the removal of information. This, however, might harm legitimate data analysis. Sanitization approaches attempt to resolve this trade-off by leaving certain useful traits intact. An overview of methods can be found in [64]. This section summarizes only those on which the algorithms presented in Section 4 are based on.

Generalization [55] increases the ambiguity of quasi-identifier value combinations by replacing their values with less specific ones. For instance, the age 81 may be changed to *elder*. A related method is cell suppression [17], which erases values leaking too much information. It can be interpreted as maximal generalization [58]. Both methods are frequently used to enforce syntactic privacy models [9, 11, 30, 41, 42, 58]. A key advantage is that they are 'truthful' [58] unlike other approaches which introduce random noise to the data.

Alternatives to generalization in syntactic privacy are microaggregation [21] and anatomization [66]. The former samples the data into subsets of equal size and replaces each subset with an aggregate statistic such as the mean. Thus, individual records cannot be observed directly while computations over subsets are still possible. Anatomization leaves the quasi-identifiers unaltered while introducing ambiguity in the mapping between quasi-identifiers and sensitive values. Section 4 presents it in more detail.

# 3 Syntactic Privacy

Syntactic privacy models define that a data set is anonymous if it satisfies some condition. These conditions depend only on the data set, its structure and symbols. An alternative is provided by differential privacy [26] which relies on random perturbation instead.

This section presents selected syntactic privacy models, their assumptions and attacks on them. Further models are found in survey articles such as [62, 71, 72].

## 3.1 Attacker Model

Syntactic privacy models assume a passive adversary who has obtained some released data sets and attempt a disclosure. They also have a list of targeted individuals and their quasi-identifier values.

The development of syntactic privacy was driven by refinements of attacker models. Where past restrictions and assumptions were deemed inappropriate, new disclosure types and background information was taken into account. This led to the discovery of novel attacks and the proposal of new privacy models.

The following subsections summarize these developments along two lines of research. One considers a scenario where a single data set is released and refines the disclosure types which are considered. This sequence focuses on increasing the attacker model's strength. The other line of research studies multiple releases of the same or overlapping data sets. Similarly to the first line, the assumptions are subsequently relaxed.

## 3.2 Single Release

The first syntactic privacy model was $k$-anonymity [59]. It assumes a single release of data and considers only identity disclosure. A data set is considered sanitized when every data point's quasi-identifier is indistinguishable from at least $k$ other records. Thus, the adversary cannot distinguish their target's record from at least $k-1$ other ones, preventing identity disclosure.

The notion of indistinguishable quasi-identifiers is reused in other models. We formalize it with the following relation:

**Definition 3.1** (Quasi-Identifier Equivalence)**.** *Let $X$ be a dataset and $x_i = (q_{i,1}, ..., q_{i,m}, s_i)$ for $i = a, b$ be two records. Define* ***quasi-identifier equivalence*** *as the relation $\approx$ such that*

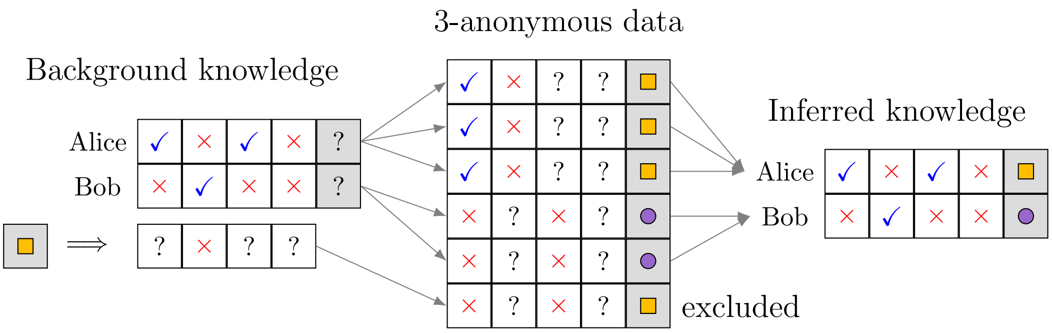


Figure 1: Homogeneity and background knowledge attacks.

$$x_a \approx x_b \iff \bigwedge_{i=1}^{m} q_{a,i} = q_{b,i}.$$

*We also define for some $x \in X$:*

$$[x] = \{x' \in X : x \approx x'\}$$

Observe that $\approx$ is an equivalence relation. Thus, $[x]$ denotes the equivalence class of a record $x$ with respect to $\approx$. This allows us give a formal definition for $k$-anonymity:

**Definition 3.2** ($k$-Anonymity [59])**.** *A dataset $X$ is **k-anonymous** if and only if $|[x]| \geq k$ for all $x \in X$.*

It is insufficient to protect from exact attribute disclosure as demonstrated by the homogeneity and background knowledge attacks [46]. The homogeneity attack occurs if all records in a target's equivalence class have the same sensitive value. While identifying the target's record is still impossible, inferring its sensitive value is trivial. The background knowledge attack assumes that the adversary possesses additional information. This might rule out enough records from the target's equivalence class to enable a homogeneity attack.

Figure 1 shows an example of these attacks. For simplicity, quasi-identifiers are depicted as binary attributes and sensitive values as coloured shapes. Some quasi-identifiers have been suppressed to achieve 3-anonymity. Yet, Alice's record is subject to a homogeneity attack. In Bob's case, the adversary knows that an amber square contradicts Bob's quasi-identifier.

To overcome these weaknesses, $l$-diversity [46] was proposed. It requires that each equivalence class contains at least $l$ 'well-represented' sensitive values. There are multiple definitions [46, 66] differing in how the term 'well-represented' is formalized. All of them prevent homogeneity attacks and render background knowledge attacks unlikely for a sufficiently large $l$.

**Definition 3.3** (Simple $l$-Diversity [16, 65, 66])**.** *Let $l \in \mathbb{N}, l > 1$ and $X$ be a dataset. Let $\phi(s, [x])$ be the relative frequency of a sensitive value $s$ in the equivalence class $[x]$. Let $\hat{s} = \arg\max_s \phi(s, [x])$ be the sensitive value with maximum relative frequency in $[x]$. An equivalence class $[x] \subseteq X$ is simple $l$-diverse if and only if $\hat{s} \leq \frac{1}{l}$. $X$ is **simple l-diverse** if and only if all contained equivalence classes are simple $l$-diverse.*

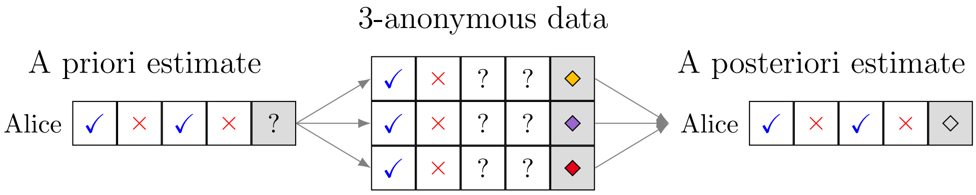


Figure 2: Similarity attack on 3-anonymous data.

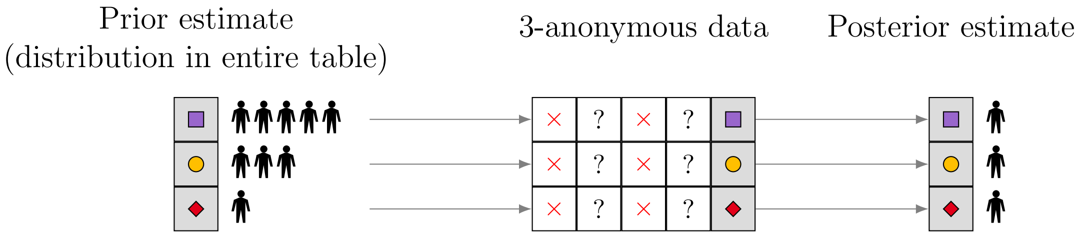


Figure 3: Skewness attack on 3-anonymous data.

Note that simple $l$-diversity implies $l$-anonymity: Any $\phi(s,[x])$ is bounded by $\frac{1}{l}$. With $\phi$ being a distribution, all $\phi(s,[x])$ must sum to 1. Thus, at least $l$ different sensitive values $s$ must appear in $[x]$. Similar proofs exist for other definitions of $l$-diversity [46].

The similarity and skewness attacks [43] prove that $l$-diversity does not prevent approximate attribute disclosure. The similarity attack is a conceptual cousin of the homogeneity attack: If all sensitive values in an equivalence class have a shared meaning, this common piece of information is revealed to the adversary. Figure 2 provides an example. While the attacker does not learn Alice's exact sensitive value, they infer that it is some coloured diamond.

The skewness attack exploits differences in the sensitive value distribution. It assumes that the adversary uses the entire data set to estimate probabilities of all sensitive values. If the sensitive value distribution in a group of indistinguishable records differs from the global distribution, an adversary can refine their estimates. Figure 3 shows an example. Imagine that a diamond corresponds to a lethal disease. While only one out of nine patients suffer from it in the overall table, the chances for the depicted equivalence class are one in three. Thus, if an adversary links their target to this equivalence class, their belief that the target has the lethal diagnosis triples. This constitutes a knowledge gain about the target.

A solution is provided by $t$-closeness [43]. It bounds the difference between these sensitive value distributions and thus the adversary's knowledge gain. This difference is measured through the Earth Mover's Distance [54]. Intuitively, it tells how much probability mass must be moved to make both distributions equal.

**Definition 3.4** ($t$-closeness [43]). *Let $t \in \mathbb{R}, 0 \le t \le 1$ and $X$ be a dataset. An equivalence class $[x] \subseteq X$ satisfies **t-closeness** if and only if*

$$d((\phi(s,[x]),\phi(s,X) \le t$$

*where $d$ is the Earth Mover's Distance and $\phi(s,X)$ the relative frequency of a sensitive value $s$ in the overall table $X$. $X$ is **t-close** if and only if all contained equivalence classes $[x] \subseteq X$ are $t$-close.*

Unlike $l$-diversity, $t$-closeness does not imply $k$-anonymity [43]. Therefore, both must be enforced in practical applications.

It has been observed by Cao and Karras [10] that $t$-closeness takes only the absolute knowledge gain of an adversary into account: The negligence of relative gains originates from the Earth Mover's Distance. In response, Cao and Karras proposed $\beta$-likeness [10]. Intuitively, it bounds the relative difference of empirical sensitive value distributions below a threshold $\beta$. However, if a value appears frequently in the overall table, this bound becomes meaningless. Thus, they enhanced their definition to use a logarithmic bound in these cases.

**Definition 3.5** (Enhanced $\beta$-Likeness [10])**.** *Let $\beta \in \mathbb{R}, \beta > 0$ and $X$ be a dataset. An equivalence class $[x] \subseteq X$ satisfies* ***$\beta$-likeness*** *if and only if for all sensitive values $s$*

$$\frac{\phi(s,[x]) - \phi(s,X)}{\phi(s,X)} \leq \min\{\beta, -\ln(\phi(s,X))\}$$

*$X$ satisfies $\beta$-likeness if and only if all contained equivalence classes are $\beta$-alike.*

Above models assume that the adversary knows that their target's data point is contained in the released data set. Thus, no protection from membership disclosure is provided. An extension of $k$-anonymity and $l$-diversity to incorporate this protection is $\delta$-presence [48]. It bounds the adversary's confidence whether an individual's record is contained in the data set. Both upper and lower bounds are provided to protect from positive and negative membership disclosure. We omit its definition here for brevity.

These attacks and improved models utilize different definitions of disclosure and assumptions on the adversary's background knowledge. While $k$-anonymity considers only identity disclosure, $l$-diversity, $t$-closeness and $\beta$-likeness protect from attribute disclosure. In contrast, $\delta$-presence takes also membership disclosure into account. The assumed disclosure definitions vary in whether the adversary's knowledge gain has must lead to certainty, an absolute increase above a threshold value or a relative one. Another difference in assumptions concerns the adversary's background knowledge: They may be able to rule out certain sensitive values or learn from the released data.

## 3.3 Multiple Data Releases

$k$-anonymity assumes that data is only released once. However, there might be multiple releases of overlapping data sets. Three scenarios are distinguished by [30]:

- the *multi-purpose publishing* scenario, where the same unaltered data set is released multiple times for different purposes,
- the *publishing* scenario, where subsequent releases contain new data points, and
- the *update* scenario, which extends publishing by also allowing deletions.

All three scenarios assume that the data set is maintained and released by the same entity. [32] introduces a fourth scenario, *independent releases*, where different parties release overlapping data sets independently of each other. Thus, not every previous release might be known when data is sanitized.

In all scenarios, there is a causal dependency between releases. It has been formalized as *correspondence* [30]: All released records belonging to the same individual must have

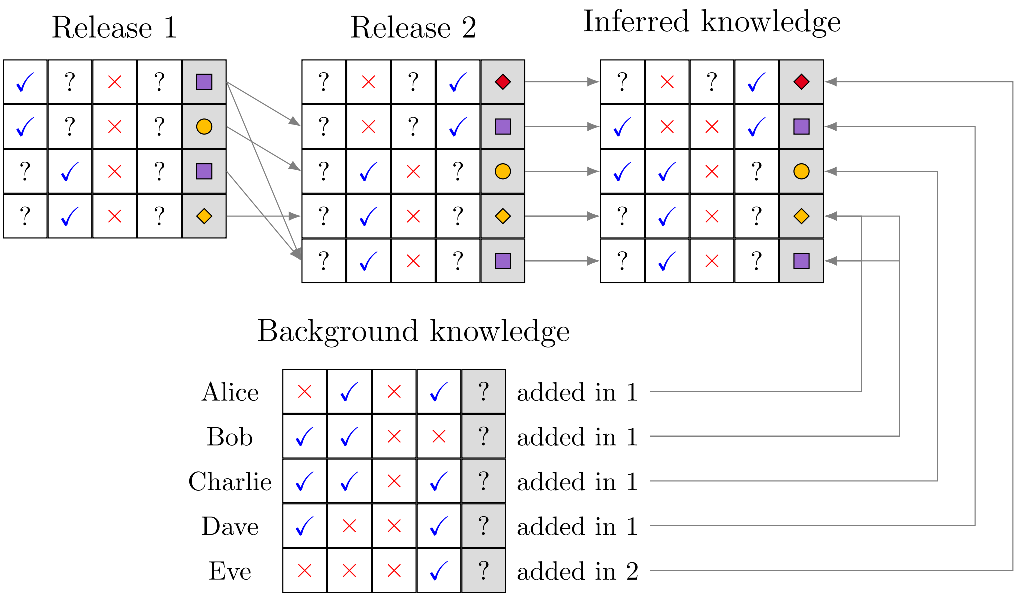


Figure 4: Composition attacks on two 2-diverse releases.

the same sensitive value and their quasi-identifiers are derived from the same raw data point. Furthermore, there must be exactly one sanitized record in every release containing a given raw data point.

While traditional syntactic privacy models such as $k$-anonymity or $l$-diversity ensure indistinguishability in single data sets, this dependency allows linking records over multiple releases. If the sanitization differs between releases or if data points are added or removed, candidates can be ruled out. This increases the adversary's confidence in the remaining options and can lead to a disclosure.

Figure 4 provides some examples. While both releases are 2-diverse, the correspondence knowledge enables inferences. Eve must have the red diamond because this record has no corresponding partner in the first release. Dave's quasi-identifier can be reconstructed entirely due to differences in sanitization. In Charlie's case, the differences reveal enough information to single-out her record. Only Bob's and Alice's records maintain ambiguity.

Attacks exploiting correspondence have been called *correspondence attacks* [30]. More generally, the term *composition attacks* [32] is used for attacks relying on multiple releases. Table 2 provides an overview of the composition attacks presented in this section. It maps them to the attacked privacy model and release scenario.

The threat of such links between releases was already noted when $k$-anonymity was introduced: Sweeney [59] observed that identity disclosure is possible if linking is unambiguous. She argued that subsequent releases must treat all attributes appearing in previous releases as quasi-identifiers. However, this increases information loss.

A refined solution for the multi-purpose publishing scenario was proposed in [63]. They extended $k$-anonymity and $l$-diversity to $(X, Y)$-privacy, which ensures an ambiguous join between different data views. It is achieved by generalizing attributes appearing in more than one release. Previously released tables do not have to be altered.

The publishing scenario was first studied by Byun et al. [9]. They identify several cases in which an adversary can eliminate records and thus threaten privacy. The proposed solution is to delay insertions until none of these cases occurs. However, this might delay insertions indefinitely. Furthermore, their solution requires that information about all

Table 2: Studies on the behaviour of syntactic privacy models in different release scenarios.

| Scenario | $k$-anonymity | $l$-diversity | both |
|---|---|---|---|
| multi-purpose | [63] | | |
| publishing | [59], [51], [30] | [9] | |
| update | | [67], [35] | |
| independent release | | | [32] |

previous releases is stored.

An alternative approach is investigated in [51]. The authors argue that linking all releases of the same record is important for data analysis. Consequently, their approach assigns a unique pseudonym to every record to make correspondence unambiguous. Using these assumptions, they identify privacy-breaching cases where differences in generalization reveal information about the raw quasi-identifier values. This threatens $k$-anonymity's promise of $k$ indistinguishable records. They argue that these precarious cases can be avoided when composing new releases. However, their solution assumes that the adversary does not know when each record has been added.

The shortcomings of [9] and [51] are overcome in [30]. Again, potentially privacy-breaching cases are identified and avoided. The assumptions in [51], unambiguous correspondence and the adversary's ignorance of insertion timestamps, are dropped. Their proposed solution, the syntactic privacy model $BCF$-anonymity, extends $k$-anonymity by ensuring ambiguous linking of releases. Unlike [9], its algorithm depends only on the last release and inserts new data points immediately.

[67] investigates the update scenario. The authors observe that deletions can leak information: If a candidate sensitive value disappears while the target's record is still known to be within the data, the adversary can rule out associated candidate records. To avoid this, they extended $l$-diversity to $m$-invariance:

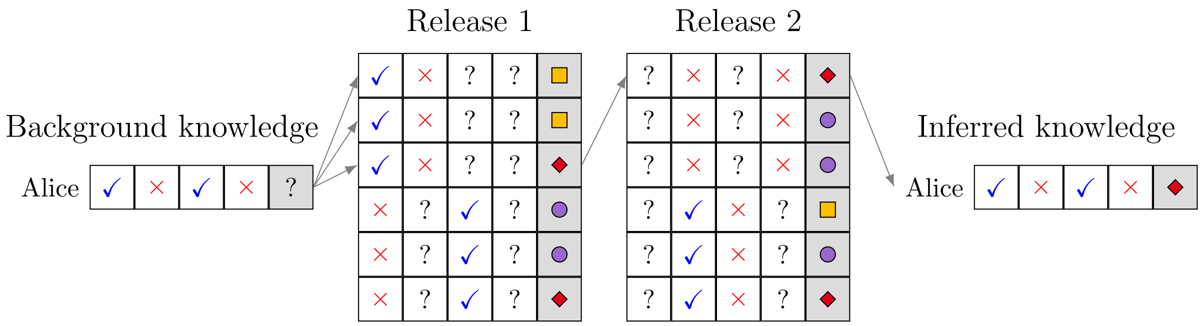
**Definition 3.6** ($m$-invariance [67])**.** *Let $X_1, ..., X_n$ be a series of datasets and $m \in \mathbb{N}\ \{0\}$. Define the **signature** $S([x])$ of an equivalence class $[x] \subseteq X_i$ as the set of sensitive values appearing in $[x]$:*

$$S([x]) = \{s \colon \exists (q_1, ..., q_p, s) \in [x]\}.$$

*$X_i$ is **m-unique** if and only if $|[x]| = |S([x])| \geq m$ for all equivalence classes $[x] \subseteq X_i$ The series $X_1, ..., X_n$ is **m-invariant** if and only if*

1. *$X_i$ is $m$-unique for all $1 \leq i \leq n$ and*
2. *$S([x_i]) = ... = S([x_j])$ for all data points $x_i \in X_i, ..., x_j \in X_j$ which describe the same individual.*

The definition implies two properties. Firstly, $m$-uniqueness requires that every equivalence class contains at least $m$ different sensitive values. Thus, $m$-invariance implies $m$-diversity. Secondly, the definition requires that all equivalence classes in which an individual's record appears over time contain the same sensitive values. This requires the replacement of deleted data points by either newly inserted ones or artificial records when no suitable insertion took place [67]. Consequently, $m$-invariance sacrifices syntactic privacy's tenet of *truthfulness*.



| | | | | | |
|---|---|---|---|---|---|
| Alice | ✓ | × | ✓ | × | ? |

| | | | | |
|---|---|---|---|---|
| ✓ | × | ? | ? | ■ |
| ✓ | × | ? | ? | ■ |
| ✓ | × | ? | ? | ◆ |
| × | ? | ✓ | ? | ● |
| × | ? | ✓ | ? | ● |
| × | ? | ✓ | ? | ◆ |

| | | | | |
|---|---|---|---|---|
| ? | × | ? | × | ◆ |
| ? | × | ? | × | ● |
| ? | × | ? | × | ● |
| ? | ✓ | × | ? | ■ |
| ? | ✓ | × | ? | ● |
| ? | ✓ | × | ? | ◆ |

| | | | | | |
|---|---|---|---|---|---|
| Alice | ✓ | × | ✓ | × | ◆ |

Figure 5: Intersection attack on 3-diverse releases.

Another disadvantage is that $m$-invariance is susceptible to *equivalence attacks* [35]. The adversary is assumed to know when each record was added and deleted. In some situations, this allows them to infer which new inserted data point replaced a deleted one. Consequently, both records must share the same sensitive value and a new causal dependency can be exploited.

Independent releases are studied by Ganta, Kasiviswanathan and Smith [32]. The aforementioned protective measures are inapplicable here because not every previous release is known at the time of sanitization. Ganta, Kasiviswanathan and Smith introduces the *intersection attack* to show how vulnerable syntactic privacy models are in this setting. It affects most syntactic privacy models including $k$-anonymity, $l$-diversity and $t$-closeness.

Figure 5 provides an example. An adversary can identify a group of records known to contain their target in every release using the known quasi-identifier. The target's sensitive value must appear in every group. Therefore, the set of candidate values can be reduces to the intersection of candidate values in every release. In Alice's case, only the red diamond satisfies this condition and thus her sensitive value is disclosed.

The intersection attack was demonstrated on a real-life data set by [14]. While a countermeasure was proposed in [5], it relies on cooperating data publishers and entails severe loss of information.

The release scenarios expand the assumptions required by syntactic privacy. The described attacks and countermeasures also assume different definitions of disclosure. For instance, intersection attacks only consider attribute disclosure, thus the protection mechanism proposed in [5] might still be susceptible to the identity disclosure attacks described in [30]. To the best of our knowledge, the applicability of all discovered composition attacks to syntactic privacy models has not been thoroughly evaluated. Finally, the attacks and proposed countermeasures also make different assumptions regarding the adversary's background knowledge.

## 3.4 Remaining Shortcomings

Syntactic privacy is subject to shortcomings other than those discussed above. These problems are presented in the following.

### 3.4.1 Utility Problems

Several studies have reported problems concerning the analytical value of data which was sanitized according to syntactic privacy models. Firstly, the gain of privacy offered by $k$-anonymity, $l$-diversity and $t$-closeness comes with a high loss of data mining utility [8].

Secondly, the information loss is especially pronounced for $l$-diversity on skewed data sets [46]. Intuitively, the lack of diversity requires more generalization. Thirdly, syntactic privacy is affected by the curse of high-dimensionality [1]. Essentially, the number of data points required to maintain the same density in data space for an increasing number of attributes grows exponentially. In other words, the records which are made indistinguishable by the sanitization algorithm are increasingly dissimilar. The resulting information loss renders syntactic privacy unsuitable for high-dimensional data sets.

### 3.4.2 Limited Data Model

Syntactic privacy models assume a split between quasi-identifiers and sensitive records. This limits their application to tabular data. Unstructured data such as images or time-series data is not supported. Furthermore, only finite data sets are considered. Infinite data streams can be incorporated by splitting them into multiple releases as in the publishing scenario. However, this solution delays the publication of new data points until the next release, hence it is unsuitable for real-time applications where new data points have to be released immediately. Both unstructured data and real-time applications are becoming increasingly relevant due to developments such as ubiquitous computing and the internet of things.

Another problem concerns the classification of attributes into quasi-identifiers and sensitive attributes. It depends on the context, culture and individual [7]. For instance, a recurring example of a sensitive attribute in syntactic privacy research is salary [15, 37, 43, 66]. Yet, the salary can be considered public knowledge in Sweden and Norway. Both countries publish the tax records of their residents, which allows estimating an individual's salary [24].

### 3.4.3 Questionable Assumptions

Even when the separation between quasi-identifiers and sensitive attributes is clear, syntactic privacy relies on other questionable assumptions. They concern the adversary's background knowledge and how a disclosure is defined. For instance, $k$-anonymity only prevents identity disclosure in the single-release scenario if the adversary's knowledge is limited to quasi-identifier values. Violations of these assumptions enable for instance homogeneity, composition or skewness attacks as described above. Relaxations of these assumptions have led to new syntactic models such as $l$-diversity, $m$-invariance or $t$-closeness. However, the interactions between the assumptions are not well understood. For instance, no publication extending $t$-closeness, $\beta$-likeness or $\delta$-presence to a multiple release scenario has been found during our literature research. Similarly, $BCF$-anonymity, $m$-invariance and $(X, Y)$-privacy do not incorporate the refined approximate disclosure definitions embodied in $t$-closeness and $\beta$-likeness.

Another problematic assumption concerning the privacy promises was identified by Kifer [36]. Syntactic privacy models guarantee that the probability of a successful disclosure does not exceed a threshold value. However, calculating these probabilities relies on an assumed reasoning model. More precisely, the adversary's a priori distribution over potential raw data sets must be assumed. When the adversary observes the released data, they determine the set of records indistinguishable to their target. The resulting a posteriori distribution is calculated from these records and the a priori distribution.

Kifer observed that the majority of models, for instance $k$-anonymity, $l$-diversity, $t$-closeness or $m$-invariance, use the random worlds model [4], which assumes a uniform distribution. Thus, any record which might belong to a given target is assumed to be equally likely to be the target's real record. Kifer argues that this reasoning is not sound because an adversary can learn correlations from the released data to improve their estimates. For example, a 2-anonymous data set might link a known smoker to two records, which reveal that the person has either lung cancer or is healthy. Under the random worlds model, both options are considered equally likely. However, the adversary might learn a correlation between smoking and lung cancer from the entire published data set. This can shift the confidence of the adversary's predictions for individuals above the threshold promised by the syntactic privacy model, leading to approximate attribute disclosure. The attack has been called *DeFinetti attack* by Kifer. [36]

The applicability of these assumptions must be verified before syntactic privacy can be applied. Any mistake or oversight can result in privacy breaches. Consequently, syntactic privacy is a complex and fragile concept.

### 3.4.4 Attacks on Algorithms

There is also a line of research concerning attacks on sanitization algorithms. It was initiated by [65] and [70] independently of each other. Both research groups observed that minimizing information loss leaks information about the raw data. Intuitively, the adversary can determine that if an algorithm produced the observed sanitized table, any further partitioning of the data set would have violated the syntactic privacy model. This can be combined with the knowledge of raw quasi-identifier values to deduce possible raw data sets. If no option maps a sensitive value to a record, negative attribute disclosure occurs and the adversary's confidence in the remaining options increases. This attack was called *minimality attack* in [65]. A similar but less general description is provided in [70].

The minimality attack and its preconditions were formalized and further studied by [16]. They noticed that the attack relies on an asymmetric partitioning of the data and affects only the smaller subset. Therefore, the attack can be mitigated by restricting algorithms to roughly equally sized partitions. Another proposed countermeasures is generalizing the data more than necessary at the expense of data utility [65]. While $k$-anonymity is immune to minimality attacks because they rely on constraints on the sensitive value sets, an extension, the *downcoding attack* [14], removes this restriction.

### 3.4.5 Lack of Desired Properties

Kifer and Lin have proposed [39] and refined [38] two privacy axioms: *transformation invariance* and *convexity*, which is also called *privacy axiom of choice* in their first paper. While they provide formal definitions, we paraphrase them here informally for simplicity. Transformation invariance states that applying an arbitrary computation onto a sanitized data set must not compromise the privacy guarantees of the chosen privacy model. This ensures that anonymity cannot be breached by post-processing the data. Convexity requires that algorithms satisfying the privacy model can be selected for anonymization without a dependency on the input data. This ensures that anonymization algorithms are exchangeable. Kifer and Lin argue that syntactic privacy models, unlike differential privacy, do not satisfy these axioms.

A third axiom called *secure composition* was introduced by Ganta, Kasiviswanathan and Smith [32]. A privacy model composes securely when its privacy guarantee holds in an independent release scenario. The composition attacks presented in Section 3.3 show that syntactic privacy models do not satisfy this property.

# 4 Algorithms

This section presents selected algorithms achieving syntactic privacy. They have been chosen to outline different approaches and the historical development. More information on algorithms can be found in survey articles such as [13, 31, 69, 71].

Most algorithms achieving $k$-anonymity rely on generalizing the quasi-identifiers of data points. This makes records more similar to each other and thus increases indistinguishability. When observing the equivalence classes, generalization also increases the number of contained sensitive values [46] and moves their distribution towards the distribution in the overall table [43]. Therefore, the generalization algorithms for $k$-anonymity can be adapted to ensure $l$-diversity and $t$-closeness.

Note that generalization removes information from the data. Thus, algorithms which achieve a syntactic privacy model with minimal information loss are desired. A first proof-of-concept algorithm was proposed in [58]. It considers all possible generalizations of the input data set and outputs the optimal one. The set of eligible generalizations has to be specified by the user.

While this exhaustive approach is unsuitable for practice, improved algorithms have been proposed. One example is Incognito [42]. It exploits that once a $k$-anonymous sanitization is found, further generalizations only remove more information and thus do not have to be considered.

It was proven that the problem of optimal generalization is NP-hard [47]. Some algorithms such as Incognito [42] address this by increasing the granularity of generalizations, which decreases the number of steps the algorithm must execute, however, they still have an exponential worst-case run-time. Another approach is taken by approximate algorithms. They relax the guarantee of minimal information loss to improve performance.

One example of an approximate algorithm is Mondrian [41]. Unlike the aforementioned algorithms, which generalize the data first and then derive the partitioning of data into indistinguishable groups, Mondrian partitions the data first and then derives the required generalization. This removes the dependency to user-specified generalization descriptions. Mondrian achieves this by splitting the data recursively at the median until no subset can be split without violating the chosen syntactic privacy model. The quasi-identifier of every subset is then replaced by the multi-dimensional interval defined by the previous splits.

While Incognito and Mondrian can be modified to support $l$-diversity and $t$-closeness, dedicated algorithms for these models have also been proposed. An example for $l$-diversity is SABRE [11]. Its experimental evaluation shows that it offers both increased performance and decreased information loss when compared to a Mondrian adaption. Most variations and extensions of $k$-anonymity, $l$-diversity and $t$-closeness which have been proposed in the literature are accompanied by a dedicated algorithm, for instance [10, 30, 67].

While most algorithms rely on generalization, alternatives have also been explored. A clustering-based approach is described in [3]. Microaggregation has been applied by [2] and [23]. Similar to Mondrian, these algorithms create a partitioning of the input data. The quasi-identifiers are then replaced by a summary statistic such as the mean or the minimum and maximum values.

A unique interpretation of $l$-diversity is found in Anatomy [66]. It leaves the quasi-identifiers unaltered. Thus, no protection from identity disclosure is offered. Instead, the sensitive values are separated from the quasi-identifiers. Groups of records are created such that all records from the same group have an ambiguous mapping between their quasi-identifiers and their sensitive values. Essentially, the sensitive value is replaced by the set of all sensitive values appearing in a record's group. Anatomy protects from attribute disclosure by ensuring that this sensitive value set is $l$-diverse.

# 5 Implementations

Our objective was to identify existing implementations of syntactic privacy algorithms which can be integrated into a benchmarking platform for data privacy solutions. Implementations have been collected from a survey article [34] and websites: A list of tools related to the implementation ARX[1] and software recommendations by the Johns Hopkins University[2]. The results were supplemented by a web search on anonymization tools and software packages. The following selection criteria have been applied:

- The software must be available as a library or command-line tool so that it can be executed fully automatically. UI-based tools require manual input and hence are at most semi-automatic.
- It must be licensed using an open source licence and the source code must be available. Otherwise, adaptions and extension of the implementation are not possible.
- At least the three traditional syntactic privacy models $k$-anonymity, $l$-diversity and $t$-closeness have to be supported. They do not have to be implemented as long as the available algorithms can be adapted (e.g. Incognito or Mondrian).
- The software must be relatively mature. Some encountered implementations were only research prototypes and not designed for actual use.

The excluded results are shown in Tables 3. Four implementations are UI-based tools requiring manual input from a user. The source code or binary distributions could not be located for six solutions. In these cases, neither the links provided in the aforementioned surveys nor a supplementary web search yielded a working download. Two implementations support only $k$-anonymity. The remaining three are immature research prototypes. Two of them are also missing a licence.

Three implementations remained after these exclusions. They are presented in Table 4. All of them are open source tools. Anonypy [29] and Mondrian_py [33] are Python implementations of Mondrian. The former has superior maintainability and already implements $k$-anonymity, $l$-diversity and $t$-closeness. Unlike the latter, it uses well-tested Python packages such as *pandas* instead of providing own implementations.

An alternative is ARX [53]. It is a mature and well-tested Java application, which can be used as both a UI-based tool and as a Java library. Like Anonypy, $k$-anonymity, $l$-diversity and $t$-closeness are offered. While Java cannot be integrated directly into the Python-based platform under development, the library could be used to implement a back-end service invoked by Python scripts. This trades development effort and performance for maturity.

[1] https://arx.deidentifier.org/overview/related-software/ (visited on 2023-08-14)

[2] https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data (visited on 2023-08-14)

[3] The link https://cs.utdallas.edu/dspl/cgi-bin/toolbox/ is broken as of August 2023.

[4] The link http://www.privacyanalytics.ca/software/parat/ is broken as of August 2023.

Table 3: Excluded implementations

| Implementation | Reason for Exclusion |
|---|---|
| $\mu$-Argus [49] | UI-based tool |
| $\tau$-Argus [50] | UI-based tool |
| Cornell Anonymization Toolkit [68] | UI-based tool |
| ANON [12] | UI-based tool |
| UTD Anonymization Toolbox[3] | unavailable |
| PARAT[4] | unavailable |
| PPSF [45] | unavailable |
| TIAMAT [18] | unavailable |
| SECRETA [52] | unavailable |
| Anon-Tool [25] | unavailable |
| Amnesia [22] | only $k$-anonymity |
| prioprivacy [6] | only $k$-anonymity, research prototype |
| $\mu$-ANT [56] | no licence, research prototype |
| OpenAnonymizer [57] | no licence, research prototype |
| k-Anon-Tool [61] | student semester project |

Table 4: Remaining implementations

| Implementation | Language | License | Notes |
|---|---|---|---|
| ARX [53] | Java | Apache-2.0 | mature |
| Anonypy [29] | Python | MIT | Mondrian |
| Mondrian_py [33] | Python | MIT | Mondrian, bad code quality |

# 6 Conclusion

This report has reviewed the research on syntactic privacy models. While these models may appear simple and intuitive at first glance, they rely on inherent assumptions which must be checked before their use. If any of these assumptions is violated, privacy breaches may be possible. However, several assumptions cannot be controlled by users of syntactic privacy, for instance which background knowledge an attacker possesses or whether independent parties will release an overlapping data set in the future. Hence, syntactic privacy offers highly fragile privacy guarantees at best. Furthermore, selecting an appropriate metric remains a complex task. Other problems of syntactic privacy concern their restriction to tabular data, lack of analytical value and the violation of privacy axioms.

In contrast, differential privacy [26] provides desirable guarantees and properties. While syntactic privacy mostly ignores membership disclosure, differential privacy bounds the probability of both positive and negative membership disclosure [26]. This uncertainty extends to identity and attribute disclosures. Differential privacy also offers transformation invariance and complexity [39]. While it does not offer secure composition according to the definition in [32], it bounds the privacy decrease linearly [28].

However, differential privacy also has its shortcomings. Firstly, it requires the introduction of random noise [27], which impacts data utility. Secondly, differential privacy is a property of algorithms instead of data sets [27]. This requires that a trusted party holds the data to execute these algorithms. Thirdly, differential privacy must be proven for every new algorithm [26].

To summarize, the evolution of syntactic privacy was driven by identifying attacks and shortcomings, followed by solution proposals and fixes. However, no robust mathematical foundation has been developed. Furthermore, the assumptions were not investigated thoroughly, resulting in fragile privacy definitions. Differential privacy takes a different approach: Starting from a provable privacy guarantee with known assumptions, its implications, properties and possible use cases are derived.

# References

[1] Charu C Aggarwal. ‘On K-Anonymity and the Curse of Dimensionality’. In: *VLDB*. Vol. 5. 2005, pp. 901–909.

[2] Charu C Aggarwal and Philip S Yu. ‘A Condensation Approach to Privacy Preserving Data Mining’. In: *Advances in Database Technology-EDBT 2004: 9th International Conference on Extending Database Technology, Heraklion, Crete, Greece, March 14-18, 2004 9*. Springer, 2004, pp. 183–199.

[3] Gagan Aggarwal et al. ‘Achieving Anonymity via Clustering’. In: *ACM Transactions on Algorithms* 6.3 (July 2010). ISSN: 1549-6325. DOI: 10.1145/1798596.1798602.

[4] Fahiem Bacchus et al. ‘From Statistical Knowledge Bases to Degrees of Belief’. In: *Artificial Intelligence* 87.1 (1996), pp. 75–143. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(96)00003-3.

[5] Muzammil M Baig et al. ‘Data Privacy against Composition Attack’. In: *International Conference on Database Systems for Advanced Applications*. Springer, 2012, pp. 320–334.

[6] Alexandros Bampoulidis, Ioannis Markopoulos and Mihai Lupu. 'PrioPrivacy: A Local Recoding k-Anonymity Tool for Prioritised Quasi-Identifiers'. In: *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume.* 2019, pp. 314–317.

[7] Jelke G Bethlehem, Wouter J Keller and Jeroen Pannekoek. 'Disclosure Control of Microdata'. In: *Journal of the American Statistical Association* 85.409 (1990), pp. 38–45.

[8] Justin Brickell and Vitaly Shmatikov. 'The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing'. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 70–78. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401904.

[9] Ji-Won Byun et al. 'Secure Anonymization for Incremental Datasets'. In: *Secure Data Management: Third VLDB Workshop, SDM 2006, Seoul, Korea, September 10-11, 2006. Proceedings 3.* Springer, 2006, pp. 48–63.

[10] Jianneng Cao and Panagiotis Karras. *Publishing Microdata with a Robust Privacy Guarantee.* 2012. arXiv: 1208.0220. preprint.

[11] Jianneng Cao et al. 'SABRE: A Sensitive Attribute Bucketization and REdistribution Framework for t-Closeness'. In: *The VLDB Journal* 20 (2011), pp. 59–81.

[12] Margareta Ciglic, Johann Eder and Christian Koncilia. 'ANON—a Flexible Tool for Achieving Optimal k-Anonymous and l-Diverse Tables'. In: *Klagenfurt, AUT: University of Klagenfurt* (2014), pp. 1–23.

[13] Valentina Ciriani et al. 'K-Anonymous Data Mining: A Survey'. In: *Privacy-Preserving Data Mining: Models and Algorithms* (2008), pp. 105–136.

[14] Aloni Cohen. 'Attacks on Deidentification's Defenses'. In: *31st USENIX Security Symposium (USENIX Security 22).* 2022, pp. 1469–1486.

[15] Graham Cormode. 'Personal Privacy vs Population Privacy: Learning to Attack Anonymization'. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 1253–1261. ISBN: 978-1-4503-0813-7. DOI: 10.1145/2020408.2020598.

[16] Graham Cormode et al. 'Minimizing Minimality and Maximizing Utility: Analyzing Method-Based Attacks on Anonymized Data'. In: *Proc. VLDB Endow.* 3.1–2 (Sept. 2010), pp. 1045–1056. ISSN: 2150-8097. DOI: 10.14778/1920841.1920972.

[17] Lawrence H Cox. 'Suppression Methodology and Statistical Disclosure Control'. In: *Journal of the American Statistical Association* 75.370 (1980), pp. 377–385.

[18] Chenyun Dai et al. 'TIAMAT: A Tool for Interactive Analysis of Microdata Anonymization Techniques'. In: *Proc. VLDB Endow.* 2.2 (Aug. 2009), pp. 1618–1621. ISSN: 2150-8097. DOI: 10.14778/1687553.1687607.

[19] T. Dalenius. 'Towards a Methodology for Statistical Disclosure Control'. In: *Statistik Tidskrift* 15 (1977), pp. 429–444.

[20] Tore Dalenius. 'Finding a Needle in a Haystack or Identifying Anonymous Census Records'. In: *Journal of official statistics* 2.3 (1986), p. 329.

[21] D Defays and Ph Nanopoulos. ‘Panels of Enterprises and Confidentiality: The Small Aggregates Method’. In: *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*. 1993, pp. 195–204.

[22] Manolis Terrovitis Dimakopoulos Dimitris Tsitsigkos and Nikolaos. *Amnesia Anonymization Tool - Data Anonymization Made Easy*. Amnesia. 2022. URL: https://amnesia.openaire.eu/index.html (visited on 28/08/2023).

[23] Josep Domingo-Ferrer and Vicenç Torra. ‘Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation’. In: *Data Mining and Knowledge Discovery* 11 (2005), pp. 195–212.

[24] Alister Doyle and Alistair Scrutton. ‘Privacy, What Privacy? Many Nordic Tax Records Are a Phone Call Away’. In: *Reuters* (Apr. 2016).

[25] Johannes Drepper. *V086-01 Anon-Tool*. 6th Mar. 2019. URL: https://www.tmf-ev.de/Themen/Projekte/V08601_AnonTool.aspx (visited on 28/08/2023).

[26] Cynthia Dwork. ‘Differential Privacy’. In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 978-3-540-35908-1.

[27] Cynthia Dwork et al. ‘Calibrating Noise to Sensitivity in Private Data Analysis’. In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Ed. by Shai Halevi and Tal Rabin. Berlin, Heidelberg: Springer Berlin Heidelberg / Springer, 2006, pp. 265–284. ISBN: 978-3-540-32732-5.

[28] Cynthia Dwork et al. ‘Our Data, Ourselves: Privacy via Distributed Noise Generation’. In: *Advances in Cryptology - EUROCRYPT 2006*. Ed. by Serge Vaudenay. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 486–503. ISBN: 978-3-540-34547-3.

[29] Taisuke Fujita. *AnonyPy*. Github. 17th July 2023. URL: https://github.com/glassonion1/anonypy (visited on 28/08/2023).

[30] Benjamin C. M. Fung et al. ‘Anonymity for Continuous Data Publishing’. In: *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*. EDBT ’08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 264–275. ISBN: 978-1-59593-926-5. DOI: 10.1145/1353343.1353378.

[31] Benjamin C. M. Fung et al. ‘Privacy-Preserving Data Publishing: A Survey of Recent Developments’. In: *Acm Computing Surveys* 42.4 (June 2010). ISSN: 0360-0300. DOI: 10.1145/1749603.1749605.

[32] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan and Adam Smith. ‘Composition Attacks and Auxiliary Information in Data Privacy’. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008, pp. 265–273.

[33] Qiyuan Gong and Katharina M. *Mondrian*. Github. 3rd Sept. 2021. URL: https://github.com/KatharinaMoel/Mondrian_py3 (visited on 28/08/2023).

[34] Haber, Anna C and Sax, Ulrich and Prasser, Fabian. ‘Open Tools for Quantitative Anonymization of Tabular Phenotype Data: Literature Review.’ In: *Briefings in bioinformatics* 23.6 (Nov. 2022).

[35] Yeye He, Siddharth Barman and Jeffrey F. Naughton. 'Preventing Equivalence Attacks in Updated, Anonymized Data'. In: *2011 IEEE 27th International Conference on Data Engineering*. 2011, pp. 529–540. DOI: 10.1109/ICDE.2011.5767924.

[36] Daniel Kifer. 'Attacks on Privacy and DeFinetti's Theorem'. In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 127–138. ISBN: 978-1-60558-551-2. DOI: 10.1145/1559845.1559861.

[37] Daniel Kifer and Johannes Gehrke. 'Injecting Utility into Anonymized Datasets'. In: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. SIGMOD '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 217–228. ISBN: 1-59593-434-0. DOI: 10.1145/1142473.1142499.

[38] Daniel Kifer and Bing-Rong Lin. 'An Axiomatic View of Statistical Privacy and Utility'. In: *Journal of Privacy and Confidentiality* 4.1 (July 2012). DOI: 10.29012/jpc.v4i1.610.

[39] Daniel Kifer and Bing-Rong Lin. 'Towards an Axiomatization of Statistical Privacy and Utility'. In: *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 147–158. ISBN: 978-1-4503-0033-9. DOI: 10.1145/1807085.1807106.

[40] Diane Lambert. 'Measures of Disclosure Risk and Harm'. In: *Journal of Official Statistics-Stockholm-* 9 (1993), pp. 313–313.

[41] K. LeFevre, D.J. DeWitt and R. Ramakrishnan. 'Mondrian Multidimensional K-anonymity'. In: *22nd International Conference on Data Engineering (ICDE'06)*. 2006, pp. 25–25. DOI: 10.1109/ICDE.2006.101.

[42] Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan. 'Incognito: Efficient Full-Domain K-anonymity'. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. SIGMOD '05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 49–60. ISBN: 1-59593-060-4. DOI: 10.1145/1066157.1066164.

[43] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian. 'T-Closeness: Privacy beyond k-Anonymity and l-Diversity'. In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2006, pp. 106–115.

[44] Tiancheng Li et al. 'Slicing: A New Approach for Privacy Preserving Data Publishing'. In: *IEEE Transactions on Knowledge and Data Engineering* 24.3 (2012), pp. 561–574. DOI: 10.1109/TKDE.2010.236.

[45] Jerry Chun-Wei Lin et al. 'PPSF: An Open-Source Privacy-Preserving and Security Mining Framework'. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018, pp. 1459–1463. ISBN: 1-5386-9288-0.

[46] Ashwin Machanavajjhala et al. 'L-Diversity: Privacy beyond k-Anonymity'. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 3–es.

[47] Adam Meyerson and Ryan Williams. 'On the Complexity of Optimal K-anonymity'. In: *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '04. New York, NY, USA: Association for Computing Machinery, 2004, pp. 223–228. ISBN: 1-58113-858-X. DOI: 10.1145/1055558.1055591.

[48] Mehmet Ercan Nergiz, Maurizio Atzori and Chris Clifton. ‘Hiding the Presence of Individuals from Shared Databases’. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 665–676. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247554.

[49] Statistics Netherlands. *Mu-Argus Open Source*. Github. 24th Mar. 2023. URL: https://github.com/sdcTools/muargus (visited on 28/08/2023).

[50] Statistics Netherlands. *Tau-Argus Open Source*. Github. 25th Aug. 2023. URL: https://github.com/sdcTools/tauargus (visited on 28/08/2023).

[51] Jian Pei et al. ‘Maintaining K-anonymity against Incremental Updates’. In: *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*. 2007, pp. 5–5. DOI: 10.1109/SSDBM.2007.16.

[52] Giorgos Poulis et al. ‘SECRETA: A System for Evaluating and Comparing Relational and Transaction Anonymization Algorithms’. In: (2014).

[53] Fabian Prasser and Florian Kohlmayer. ‘Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool’. In: *Medical Data Privacy Handbook*. Ed. by Aris Gkoulalas-Divanis and Grigorios Loukides. Cham: Springer International Publishing, 2015, pp. 111–148. ISBN: 978-3-319-23633-9. DOI: 10.1007/978-3-319-23633-9_6.

[54] Yossi Rubner, Carlo Tomasi and Leonidas J Guibas. ‘The Earth Mover’s Distance as a Metric for Image Retrieval’. In: *International journal of computer vision* 40.2 (2000), p. 99.

[55] Pierangela Samarati and Latanya Sweeney. ‘Generalizing Data to Provide Anonymity When Disclosing Information’. In: *PODS*. Vol. 98. 188. 1998, pp. 10–1145.

[56] David Sánchez et al. ‘$\mu$-ANT: Semantic Microaggregation-Based Anonymization Tool’. In: *Bioinformatics (Oxford, England)* 36.5 (2020), pp. 1652–1653.

[57] Konrad Stark. ‘Scientific Workflows, Data Provenance Management and Data Anonymization in Context of the Genome Austria Tissue Bank’. PhD thesis. Universität Wien, 2013.

[58] Latanya Sweeney. ‘Achieving K-Anonymity Privacy Protection Using Generalization and Suppression’. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 571–588.

[59] Latanya Sweeney. ‘K-Anonymity: A Model for Protecting Privacy’. In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002), pp. 557–570.

[60] Latanya Sweeney. ‘Simple Demographics Often Identify People Uniquely’. In: *Health (San Francisco)* 671.2000 (2000), pp. 1–34.

[61] Louis Tochon. *K-Anonymization Tool for Databases*. Github. 9th June 2022. URL: https://github.com/Ltochon/k-anon_tool (visited on 28/08/2023).

[62] Isabel Wagner and David Eckhoff. ‘Technical Privacy Metrics: A Systematic Survey’. In: *Acm Computing Surveys* 51.3 (June 2018). ISSN: 0360-0300. DOI: 10.1145/3168389.

[63] Ke Wang and Benjamin C. M. Fung. ‘Anonymizing Sequential Releases’. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 414–423. ISBN: 1-59593-339-5. DOI: 10.1145/1150402.1150449.

[64] Leon Willenborg and Ton De Waal. *Statistical Disclosure Control in Practice*. Vol. 111. Springer Science & Business Media, 1996.

[65] Raymond Chi-Wing Wong et al. ‘Minimality Attack in Privacy Preserving Data Publishing’. In: *Proceedings of the 33rd International Conference on Very Large Data Bases*. 2007, pp. 543–554.

[66] Xiaokui Xiao and Yufei Tao. ‘Anatomy: Simple and Effective Privacy Preservation’. In: *Proceedings of the 32nd International Conference on Very Large Data Bases*. 2006, pp. 139–150.

[67] Xiaokui Xiao and Yufei Tao. ‘M-Invariance: Towards Privacy Preserving Re-Publication of Dynamic Datasets’. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 689–700. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247556.

[68] Xiaokui Xiao, Guozhang Wang and Johannes Gehrke. ‘Interactive Anonymization of Sensitive Data’. In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 1051–1054. ISBN: 978-1-60558-551-2. DOI: 10.1145/1559845.1559979.

[69] Yang Xu et al. ‘A Survey of Privacy Preserving Data Publishing Using Generalization and Suppression’. In: *Applied Mathematics & Information Sciences* 8.3 (2014), p. 1103.

[70] Lei Zhang, Sushil Jajodia and Alexander Brodsky. ‘Information Disclosure under Realistic Assumptions: Privacy versus Optimality’. In: *Proceedings of the 14th ACM Conference on Computer and Communications Security*. CCS ’07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 573–583. ISBN: 978-1-59593-703-2. DOI: 10.1145/1315245.1315316.

[71] Yan Zhao et al. ‘A Survey on Privacy Preserving Approaches in Data Publishing’. In: *2009 First International Workshop on Database Technology and Applications*. 2009, pp. 128–131. DOI: 10.1109/DBTA.2009.149.

[72] Athanasios Zigomitros et al. ‘A Survey on Privacy Properties for Data Publishing of Relational Data’. In: *IEEE Access* 8 (2020), pp. 51071–51099. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2980235.