# When Differential Privacy Implies Syntactic Privacy

Emelie Ekenstedt, Lawrence Ong, *Senior Member, IEEE*, Yucheng Liu, Sarah Johnson,
Phee Lep Yeoh, *Member, IEEE*, and Joerg Kliewer, *Senior Member, IEEE*

***Abstract*—Two main privacy models for sanitising datasets are *differential privacy* (DP) and *syntactic privacy*. The former restricts individual values' impact on the output based on the dataset while the latter restructures the dataset before publication to link any record to multiple sensitive data values. Besides both providing mechanisms to sanitise data, these models are often applied independently of each other and very little is known regarding how they relate. Knowing how privacy models are related can help us develop a deeper understanding of privacy and can inform how a single privacy mechanism can fulfil multiple privacy models. In this paper, we introduce a framework that determines if the privacy mechanisms of one privacy model can also guarantee privacy for another privacy model. We apply our framework to understand the relationship between DP and a form of syntactic privacy called $t$-closeness. We demonstrate, for the first time, how DP and $t$-closeness can be interpreted in terms of each other by introducing generalisations and extensions of both models to explain the transition from one model to the other. Finally, we show how applying one mechanism to guarantee multiple privacy models increases data utility compared to applying separate mechanisms for each privacy model.**

***Index Terms*—Privacy, differential privacy, t-closeness, syntactic privacy.**

## I. Introduction

DATASETS containing sensitive information are often modified before publishing to avoid breaching user privacy. These modifications, called *sanitation*, should allow a viewer of the published data to learn general trends and correlations in the data without being able to learn the exact attribute values of any individual record. Depending on how privacy is defined, one or more privacy mechanisms must therefore be applied to the dataset to sanitise it before publishing. Publishing a sanitised dataset is referred to as *privacy-preserving data publishing* [1] and is a non-interactive privacy approach. *Privacy-preserving data mining* [1], on the other hand, presents an interactive privacy approach to publishing a dataset, where users are only allowed to see parts of the information available in the dataset. In this latter approach, users can query the dataset but the response is sanitised before being sent back to the user.

Two common but semantically very different privacy models are *differential privacy* (DP) and *t-closeness*. DP [2] was introduced as a privacy-preserving data mining approach to prevent disclosure of individuals' confidential data or their presence in a dataset by limiting their impact on query results. Alternatively, $t$-closeness [3] was designed for the privacy-preserving data publishing approach and limits the information revealed about individuals' confidential attributes. It achieves this by restricting how far the confidential attributes bound to any individual can deviate in comparison to the overall distribution of these attributes in the dataset.

Despite initially being developed for a specific privacy model, a privacy mechanism may additionally protect a dataset under a different privacy model. Only a very limited amount of research has been conducted in this area, but initial findings reveal that there are some links between the models DP and *syntactic privacy*, a class of privacy models to which $t$-closeness belongs [4], [5]. Developing a better understanding of the link between these two privacy models can assist data publishers in selecting an appropriate level of privacy. It may also increase the utility (usefulness) of the sanitised data when multiple privacy levels under different models is required, by applying fewer privacy mechanisms.

This paper presents both previous and new findings on the links between the privacy models DP and $t$-closeness. We introduce a new framework to compare them through *privacy implications* which indicate if mechanisms for one can be used to guarantee the other. Using our privacy implication map, we extend and generalise the links between DP and $t$-closeness. Specifically, we

- propose a new privacy definition $(k, t)$-closeness, an extension of $t$-closeness;
- formally define $(k, \epsilon)$-DP, an extension of $\epsilon$-DP.

Here, $t$ and $\epsilon$ are privacy levels, and $k$ the number of records grouped within the dataset. These extensions provide crucial links between the models, and which collectively contribute to a more complete mapping of privacy implications. We show how our framework can be used to increase data utility and

Manuscript received September 13, 2021; revised February 13, 2022 and April 21, 2022; accepted May 11, 2022. Date of publication May 25, 2022; date of current version June 15, 2022. This work was supported in part by the Australian Research Council under Grant DP190100770 and in part by the U.S. National Science Foundation under Grant 1526547 and Grant 1813942. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. George Theodorakopoulos. *(Corresponding author: Emelie Ekenstedt.)*

Emelie Ekenstedt, Lawrence Ong, Yucheng Liu, and Sarah Johnson are with the School of Engineering, University of Newcastle, Callaghan, NSW 2308, Australia (e-mail: emelie.ekenstedt@newcastle.edu.au; lawrence.ong@newcastle.edu.au; yucheng.liu@newcastle.edu.au; sarah.johnson@newcastle.edu.au).

Phee Lep Yeoh is with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: phee.yeoh@sydney.edu.au).

Joerg Kliewer is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: jkliewer@njit.edu).

Digital Object Identifier 10.1109/TIFS.2022.3177953

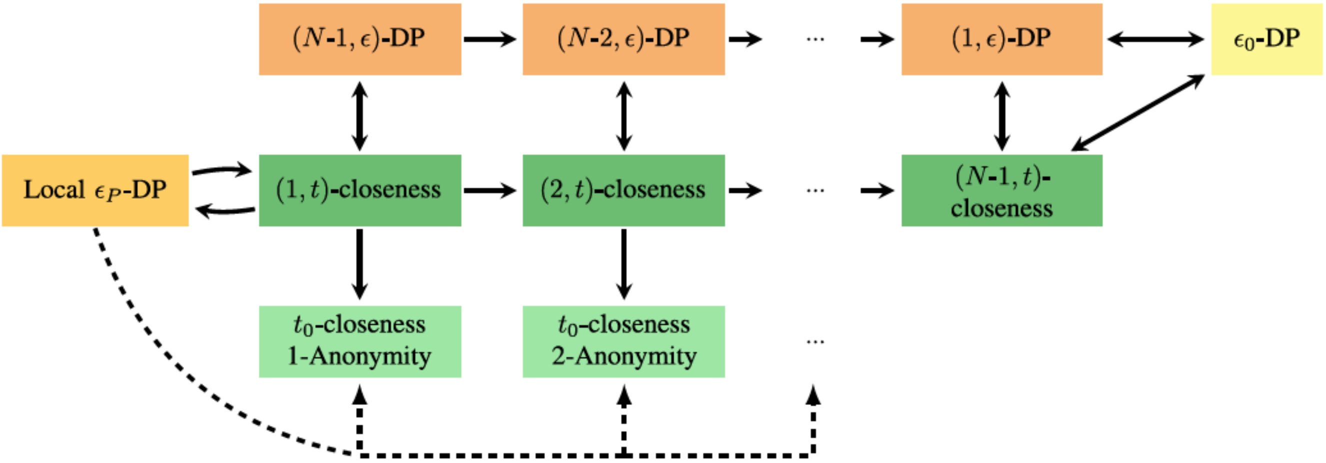

$(N\text{-}1, \epsilon)$-DP
$(N\text{-}2, \epsilon)$-DP
...
$(1, \epsilon)$-DP
$\epsilon_0$-DP
Local $\epsilon_P$-DP
$(1, t)$-closeness
$(2, t)$-closeness
...
$(N\text{-}1, t)$-closeness
$t_0$-closeness 1-Anonymity
$t_0$-closeness 2-Anonymity
...

Fig. 1. Our privacy implication map showing how $t$-closeness and DP relate. The definitions $(k, \epsilon)$-DP, $(k, t)$-closeness, and multiplicative $t$-closeness with $k$-anonymity are displayed for different values of the group size parameter $k$, where $N$ is the number of records in a dataset. Dashed lines represent findings by Domingo-Ferrer and Soria-Comas [5].

provide a numerical example with randomised response-based mechanisms [6] for both DP and $t$-closeness.

Our results are summarised in Figure 1 together with previous results by Domingo-Ferrer and Soria-Comas [5] which fits into our framework.

Figure 1 expands all definitions for different group size values $k$ and shows privacy implications both between different privacy definitions and within the same privacy definition. Additional implications can be inferred by tracing a directed path from a node to another through intermediate nodes. A missing path between two privacy definitions entails that no privacy implication exists or has yet to be found between them.

We discuss related work in Section II and explain the notation used in this paper, present our problem formulation and introduce our new framework in Section III. Section IV contains our new results on how and when the privacy guarantee translates between $t$-closeness and differential privacy while proofs for these claims are included in Appendix . Finally, the concept of utility and an example showing the potential utility gains available when using our framework are presented in Section V.

## II. Related Work

Syntactic privacy (where $t$-closeness belongs) and differential privacy were developed in the early 2000's as two separate approaches to achieving user privacy in large datasets.

### A. Syntactic Privacy

Syntactic privacy concerns methods of structuring a dataset to guarantee some form of user privacy. The dataset in general contains attributes that can be classified as either confidential, direct identifiers, or indirect identifiers. Indirect identifiers, also called *quasi-identifiers*, are attributes that can be used to identify a person when external information is available. Direct identifiers, for example names, are often removed in the first step when sanitising a dataset. Quasi-identifiers and confidential attributes are subsequently modified to prevent individuals from becoming distinguishable in the dataset and to make the link between any individual and its associated confidential attributes more uncertain.

$k$-anonymity [7] was the first syntactic privacy scheme to be introduced and it achieves user privacy by linking every individual record to at least $k$ confidential attributes. The link is established through partitioning the dataset into groups referred to as *equivalence classes* where, within each group, the quasi-identifiers are replaced by a single, generalised version of the previously present quasi-identifiers in the group. The problem of finding a partition and generalisation of attributes that satisfy a certain level of utility does not come with a straightforward solution and many different $k$-anonymity algorithms exist, such as Datafly [8], Incognito [9] and OLA [10]. One alternative solution that skips the generalisation step involves microaggregation [11].

Before long it became evident that $k$-anonymity does not put any restrictions on the diversity of the $k$ or more confidential attributes within each equivalence class, which implies that a record's confidential attributes can be revealed. In response, $l$-diversity was introduced to address this weakness by ensuring that there is a certain level of variation $l$ of the confidential attribute values in each equivalence class [12].

However, increasing the diversity of values within equivalence classes does not necessarily restrict confidential information from being released. The $l$-diverse values could, for example, have semantically similar meaning or represent extreme values entailing that the individual is much more likely to be linked to these types of values in comparison to the average. This motivated the introduction of $t$-closeness [3], which introduces further privacy guarantees restricting the distribution of sensitive attributes within each equivalence class such that their distribution closely follows the attributes' global distribution within the dataset. The Earth Mover Distance (EMD), was originally used to measure the distance $t$ between two distributions [3].

Methods for achieving $t$-closeness include modifying mechanisms for $k$-anonymity such as Incognito and microaggregation with $t$-closeness restrictions [3], [13].

It should also be mentioned that the classification of attributes as quasi-identifiers is often a difficult task. The task requires knowledge of which information will be available to a potential adversary. Consequently, confidential attributes are often also included as quasi-identifiers [14], [15], especially if the same dataset is sanitised and published multiple times [7]. However, increasing the number of quasi-identifiers increases the number of dimensions over which the syntactic privacy mechanism has to operate, and as a result the utility of the published data decreases [15].

### *B. Differential Privacy*

Differential privacy [16] was introduced as a privacy model that limits the impact of individual records on published results from statistical databases. More specifically, DP masks the evidence of an entity's presence in the dataset by restricting the relative difference between the results from any query applied to a dataset and the same query applied to a modified version of the dataset where one record has been removed [16] or changed [2]. The similarity in results in the privacy definition $\epsilon$-DP is restricted by the privacy level $\epsilon$ where a smaller $\epsilon$ results in higher privacy.

Dwork [16] argued that $\epsilon$-DP also holds for groups of any size $\kappa$ if the privacy level is adjusted to $\epsilon/\kappa$. An additive level $\delta$ was added to form a new definition of DP, $(\epsilon, \delta)$-DP, in order to ease the strict requirements of relative difference in cases where the actual difference in probabilities are small [17]. $\epsilon$-DP can be achieved by adding noise (Laplace [2], Gaussian [18], [19], or exponential [20]) of a carefully selected magnitude to each element in the query's result vector.

Additional reformulations of DP accounting for correlations between database records [21]–[25] have also been presented after DP was shown to lead to unintentional information leakage when these types of correlations are present [26].

### *C. The Intersection of t-closeness and DP*

Much attention has been focused on the strengths and weaknesses of syntactic privacy vs DP [15], [27]–[29] in areas where either one is applicable. However, in general one cannot fully replace one of the privacy models with the other since their areas of usage do not overlap fully [1].

Surprisingly little has been published on the intersection between the two, including how the privacy level of one can be expressed in terms of the other. Observing the lack of randomness offered by $k$-anonymity, Li *et al.* [4] added a random sampling step to $k$-anonymity and showed that a certain group of $k$-anonymity mechanisms (strongly-safe mechanisms) satisfy $(\epsilon, \delta)$-DP if preceded by random sampling.

Domingo-Ferrer and Soria-Comas [5] instead focused on making slight modifications to the definitions of $t$-closeness and $\epsilon$-DP to reveal underlying links between the privacy models. The distance metric for $t$-closeness was changed to be more similar to the one used in $\epsilon$-DP, while $\epsilon$-DP was modified to only apply to pairwise comparisons of records. Using these modified definitions, they were able to show that DP can satisfy $t$-closeness [5].

One major difference between the original definitions of $\epsilon$-DP and $t$-closeness lies in the interpretations of their privacy levels $t$ and $\epsilon$. The value of $t$ limits the amount of information revealed about any individual record in the results, while $\epsilon$ instead limits the individual record's effect on the results. The latter may therefore not in general be directly related to the former except in special cases such as limiting $\epsilon$-DP to queries that only consider a single individual record [1]. This fundamental difference between the two privacy definitions make it challenging to link them. Our solution introduces a form of group-wise DP in Definition 8 to bridge the gap.

## III. System Model

### *A. Notation*

In this paper, *privacy models* (e.g. DP or $t$-closeness) denote the main idea for how to achieve data privacy. A specific implementation of a privacy model, called a *privacy definition*, is defined by a mathematical expression limiting some form of information leakage. This information leakage is often bounded by a variable called the *privacy level*. The privacy level determines how private the resulting sanitised dataset will be according to the privacy definition.

Privacy levels can be compared in terms of their privacy strength.

*Definition 1 (Privacy Strength): Let A denote a privacy definition and $\alpha$ its associated privacy level.*

- *$\alpha$ is **stronger** than another privacy level $\alpha'$ for A, denoted $\alpha \succ \alpha'$, if any sanitised dataset satisfying $\alpha$ also satisfies $\alpha'$ but there exists at least one sanitised dataset satisfying $\alpha'$ that does not satisfy $\alpha$.*
- *$\alpha$ is the **optimal** privacy level for a given sanitised dataset iff it is the strongest level that the sanitised dataset satisfies. That is, denoting the range of all privacy levels for A that the sanitised dataset satisfies by S, we have for $\alpha \in S$*

$$\nexists \alpha' \in S : \alpha' \succ \alpha.$$

- *$\alpha'$ can be seen as a **lower limit** of $\alpha$ iff $\alpha \succeq \alpha'$.*

We further formally define privacy implication, or implication for short. We say that a privacy definition A implies another privacy definition B if we can find a function $f$ such that any mechanism that fulfils A at level $\alpha$ always fulfils B at level $f(\alpha)$. The function $f$ may also depend on some other non-record specific parameters $P_b$.

*Definition 2 (Implication): Consider two privacy definitions A and B with the privacy level $\alpha$ and $\beta$ respectively, shortened to $A(\alpha)$ and $B(\beta)$, and let $P_a$ and $P_b$ denote the associated parameters. We say that*

- *$A \Rightarrow B$, A **implies** B iff there exists a function $f_{P_b}(\alpha)$ such that, given any mechanism Z and any dataset D, if Z satisfies $A(\alpha)$ for D then Z also satisfies $B(f_{P_b}(\alpha))$ for D;*
- *$A \nRightarrow B$ iff for any function $f_{P_b}(\alpha)$ there exists at least one mechanism Z that when applied to a dataset D satisfies*

TABLE I

A SUMMARY OF SYMBOLS USED IN THIS PAPER

| | |
|---|---|
| $D$ | Database with $N$ records |
| $[1:N]$ | A set of all the record indexes in the dataset |
| $r_i$ | Database record $i, i \in [1:N]$ |
| $D_{-i}$ | $D$ with $r_i$ removed |
| $D_K$ | Database records with index in the index set $K \subseteq [1:N]$ |
| $Z$ | Privacy mechanism |
| $Q$ | Query |
| $Pr(.)$ | Probability |
| $E_i$ | An Equivalence class, $E_i \subseteq D$ |
| $B_l$ | A bucket, $B_l \subseteq \mathsf{range}(Z)$ |

*$A(\alpha)$ with an optimal privacy level $\beta^*$ under privacy definition B and where $f_{P_b}(\alpha) \succ \beta^*$.*

Proving A $\nRightarrow$ B therefore reduces to finding at least one dataset that satisfies A($a$) but where $\beta^* \prec f_{P_b}(\alpha)$ for any function $f_{P_b}(.)$. However, even if A $\nRightarrow$ B, the implication conditions may still hold for a limited range of $\alpha$ or for some mechanisms for A. Such partial implication is not considered in this paper and we have decided to include non-implications in our results to provide a word of caution - finding one mechanism for A that satisfies B does not entail that there exists an implication between A and B. A non-implication can also reveal that the information the privacy definitions are restricting can be independent of each other.

Implications can also be combined to provide further insights into the relationship between two privacy definitions.

*Definition 3 (Combined Implications): Let A and B be two privacy definitions with privacy levels $\alpha$ and $\beta$ respectively. We say that*

- *$A \Leftrightarrow B$, there exist **reciprocal implications** between A and B if $A \Rightarrow B$ for a function $f(\alpha)$ and $A \Leftarrow B$ for a function $g(\beta)$.*
- *$A \Lleftarrow\!\!\!\Rrightarrow B$, A and B are **equivalent** if $g(f(\alpha)) = (\alpha)$ and $f(g(\beta)) = (\beta)$ for all $\alpha$ and $\beta$.*
- *$A \overset{\Rightarrow}{\not\Leftarrow} B$, A is a **generalisation** of B if $A \Rightarrow B$ but $A \not\Leftarrow B$.*

### B. Preliminaries

Recall that an equivalence class, $E$, is a group of records (rows in the dataset) with identical quasi-identifiers. For $t$-closeness we want the distribution of the confidential attributes belonging to every equivalence class to be similar to that of the entire dataset. Also, a bucket defines a range of confidential attributes that are indistinguishable as far as privacy is concerned. Buckets were introduced to $t$-closeness by Soria-Comas and Domingo-Ferrer [30] to also enable datasets containing confidential attributes with rare values to fulfil $t$-closeness with $t < \infty$. Additional symbols and concepts used in this paper are summarised in Table I.

An alternative definition of $t$-closeness called stochastic $t$-closeness was introduced by Domingo-Ferrer and Soria-Comas [5] to enable the use of stochastic $t$-closeness mechanisms. They further focused on an alternative distance metric to the traditional EMD metric. We refer to the incorporation of this alternative distance metric into stochastic $t$-closeness as multiplicative $t$-closeness with $k$-anonymity, or *multiplicative $t$-closeness* for short.

*Definition 4 (Multiplicative t-closeness): Let $D$ be a dataset with $N$ records. Let $E_{\mathrm{m}} \subsetneq [1:N]$ for $m \in [1:M]$, be a set of $M \in [1:N]$ equivalence classes (in the sense that $E_{\mathrm{i}} \cap E_{\mathrm{j}} = \emptyset$ for all $i \neq j$, $\min_i |E_i| = k$, and $\sum_{i=1}^{M} E_{\mathrm{i}} = [1:N]$), $Z$ a privacy mechanism, $\hat{D}$ the sanitised dataset created by applying $Z$ to all records in $D$, $c_i$ the vector representation of record $i$'s confidential attributes in $D$, and $\{B_1, \dots B_p\}$ a set of buckets such that $\bigcup_{j=1}^{p} B_j = \mathsf{range}(Z)$ and $B_j \cap B_l = \emptyset, \forall j, l \in [1:p] : j \neq l$. The sanitised dataset $\hat{D}$ fulfils multiplicative $t$-closeness at a privacy level $t \in [1, \infty)$ iff for every equivalence class $E_m$, $m \in [1:M]$, and all $l \in [1:p]$,*

$$t^{-1} \leq \frac{\Pr_{\hat{D}}(B_l)}{\Pr_{E_m}(B_l)} \leq t$$

*where*

$$\Pr_{\hat{D}}(B_l) = \frac{1}{N} \sum_{j \in [1:N]} \Pr(Z(c_j) \in B_l)$$

*and*

$$\Pr_E(B_l) = \frac{1}{|E|} \sum_{j \in E_m} \Pr(Z(c_j) \in B_l)$$

*are the averaged probability of the event $Z(c_j) \in B_l$ of the records in $\hat{D}$ and $E_m$ respectively.*

Note that a random-worlds approach [31] is used to calculate the averaged probabilities $\Pr_{\hat{D}}(B_l)$ and $\Pr_{E_m}(B_l)$ where each confidential attribute in the equivalence class is considered equally likely of representing any of the records linked to the equivalence class.

In DP we want to restrict any individual's contribution to query replies, where the query can be any function $Q$ that extracts information from a dataset. $\epsilon$-DP is commonly defined as follows:

*Definition 5: ($\epsilon$-DP) Let $Q$ be a query, $D$ a dataset with $N$ records, $D_{-\mathrm{i}} = D \setminus \{r_{\mathrm{i}}\}$ a neighbouring dataset to $D$ with record $i$ removed,[1] and $\{B_1, \dots B_p\}$ a set of buckets such that $\bigcup_{j=1}^{p} B_j = \mathsf{range}(Z)$ and $B_j \cap B_l = \emptyset, \forall j, l \in [1:p] : j \neq l$.*

*A privacy mechanism $Z$ is said to fulfil $\epsilon$-DP at a privacy level $\epsilon \in [0, \infty)$ iff $\forall i, j : i \in [1:N], j \in [1:p]$*

$$e^{-\epsilon} \leq \frac{\Pr(Z(Q(D)) \in B_j)}{\Pr(Z(Q(D_{-\mathrm{i}})) \in B_j)} \leq e^{\epsilon}$$

For sequential application of DP on a dataset, the composition theorem [32] states that the resulting privacy level is the sum of the applied privacy levels. Note that $\epsilon$ is an exponential parameter and a multiplicative change in $\epsilon$ can have a huge impact on the privacy leakage. For example, selecting $\epsilon = 5$ and applying the mechanism for two different releases reduces the privacy level by a factor of 148. Special care must therefore be taken when selecting $\epsilon$ to avoid excessive privacy loss.

Domingo-Ferrer and Soria-Comas [5] used a modified definition of $\epsilon$-DP, which we identify as *local $\epsilon$-DP* ($\epsilon$-LDP) [33], defined as follows:

[1] A common alternative definition of $\epsilon$-DP changes the definition of $D_{-\mathrm{i}}$ to have one record *changed* instead of one record *removed*.

*Definition 6: ($\epsilon$-LDP) Let D be a dataset with N records, $c_i$ the vector representation of record i's confidential attributes in D, and $\{B_1, \dots B_p\}$ a set of buckets such that $\bigcup_{j=1}^{p} B_j =$ range$(Z)$ and $B_j \cap B_l = \emptyset, \forall j, l \in [1:p] : j \neq l$. A privacy mechanism Z fulfils $\epsilon$-LDP at a privacy level $\epsilon \in [0, \infty)$ iff $\forall i, j \in [1:N]$ and $\forall l \in [1:p]$,*

$$\frac{\Pr(Z(c_i) \in B_l)}{\Pr(Z(c_j) \in B_l)} \leq e^{\epsilon}.$$

Note that the definition of $\epsilon$-LDP does not require the removal or change of a record in the dataset.

With our notation, we can now state an existing privacy implication:

*Lemma 1 ( [5]): $\epsilon$-LDP $\Rightarrow$ multiplicative t-closeness*

In addition, with a few privacy definitions of DP and $t$-closeness now formally defined, we can now give more specific details of our notation presented in Section III-A. For $t$-closeness, the privacy strength requirement according to Definition 1 translates as follows: for two privacy levels $t_0, t_1 \in (1, \infty)$, if $t_0 < t_1$, then $t_0 \succ t_1$. Similarly, for two privacy levels $\epsilon_0, \epsilon_1 \in (0, \infty)$ for $\epsilon$-DP, if $\epsilon_0 < \epsilon_1$, then $\epsilon_0 \succ \epsilon_1$. We now explain the privacy parameters introduced in the privacy definitions (see Definition 2). The privacy parameters of $\epsilon$-DP and $\epsilon$-LDP that affect the privacy level are the buckets $B_{[p]}$ and query $Q$ (which returns some function of the dataset, e.g., the average). For $t$-closeness the privacy parameters are the equivalence classes $E_{[1:m]}$ and the buckets $B_{[1:p]}$. The information we focus on from the equivalence classes are the group size parameter $k$ and the number of records $N$.

### C. Problem Formulation

Not much has been published on how the privacy guarantees of DP and syntactic privacy, $t$-closeness in particular, are linked. Neither has much been published to demonstrate that no link exists. This entails that the mapping of privacy implications between the models is far from complete. Furthermore, to compare syntactic privacy with DP we need to focus on the privacy definitions that have theoretical privacy guarantees such as $t$-closeness. As discussed in Section II-A, $k$-anonymity alone does not have theoretical privacy guarantees.

Translating $\epsilon$-DP into $t$-closeness and vice versa is of interest since they are well known for having different desirable characteristics. DP is known for guaranteeing privacy for individuals, even when an adversary has extensive amounts of prior knowledge about the data, while $t$-closeness is appreciated for letting the users' work with a sanitised version of the dataset itself, not restricting the queries that can be applied to the dataset, and that the data can be generalised instead of being modified with additional noise [1]. Establishing a relationship between the two models can reveal how these desirable characteristics can be transferred between the two.

Furthermore, it is often difficult to choose an appropriate level of $\epsilon$ [29] for real-world applications, and it is not always clear how to interpret $\epsilon$ in terms of how much information it allows to be revealed about an individual record. The closeness level $t$ has a more obvious interpretation in that it reveals how far from the overall distribution any individual record's confidential attributes (the confidential attributes within its equivalence class) are allowed to deviate. Knowing how $t$ and $\epsilon$ are related can therefore help us to both better interpret and compare privacy levels.

Before going into detail of the relationship between $t$-closeness and DP, we will introduce some assumptions to simplify our analysis. In this paper, we only focus on the query that returns a record's confidential attributes.[2] We assume that the privacy models use the same set of buckets. Using different partitions for each privacy definition complicates the analyses and takes the focus off comparing the privacy definitions. With this in mind we purposefully shorten the notation for implications to omit the specific set of buckets $B_{[1:p]}$ and query $Q$ used.

### D. The Significance of Implications and Our Implication Mapping

An implication between two privacy definitions A and B requires that A and B base their computations of their privacy levels ($\alpha$ and $\beta$ respectively) on similar parts of the data. It also requires that the privacy level $\beta$ can be estimated based on $\alpha$ without using any other specific characteristics of the data. An advantage of this is that a mechanism designed to fulfil A will also fulfil B without any modifications to the mechanism. In particular, we would not have to apply a mechanism for A in addition to another mechanism for B to the same data in order to satisfy both A and B. On the other hand, the lack of an implication does not reveal if implications exist for certain values of the privacy levels. For example, it might be possible to estimate $\beta$ only for strict values of $\alpha$, which our results do not reveal.

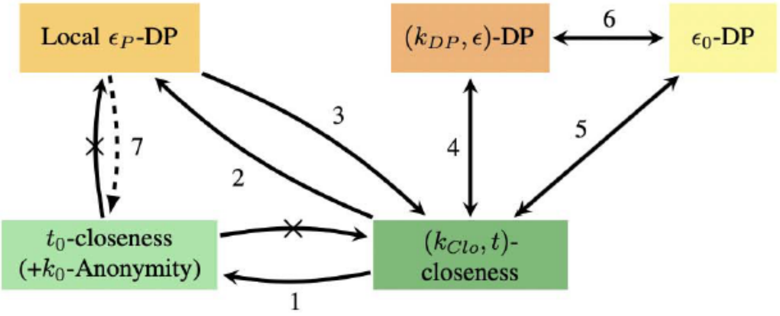
The implication paths of Figure 1 reveal that all nodes for $t$-closeness can be reached by a DP node but not vice versa. This result does not imply that $t$-closeness is a redundant model. Bear in mind that we are only considering application areas where $t$-closeness and DP are both applicable and that there are areas where $t$-closeness is more appropriate to use than DP [1].

## IV. Privacy Implication Mapping

This section presents and compares different privacy definitions of $t$-closeness and DP in order to make a broader comparison between the two models of privacy. A summary of these definitions is given in Table II and our implication mapping is depicted in Figure 2. Note that in this paper, the terminology $t$-closeness and DP refers to the privacy models, not specific privacy definitions.

The proofs of our results (lemmas and theorems) are presented in Appendix . Starting from Domingo-Ferrer and Soria-Comas' finding that $\epsilon$-LDP $\Rightarrow$ multiplicative $t$-closeness [5], we first extend their results by showing that there is no

[2] The results in this paper could be extended for other query functions by first bounding the range of the query function in terms of that of the confidential attributes.



1. $k_{Clo} \leq k_0, t = t_0$
2. $k_{Clo} = 1, e^{\epsilon_P} = t^2$
3. $t = f(k, N, \epsilon_P)$
4. $k_{DP} = N - k_{Clo}, e^{\epsilon} = t$
5. $k_{Clo} = N - 1, e^{\epsilon_0} = t$
6. $k_{DP} = 1, \epsilon_0 = \epsilon$
7. $t_0 = \max_E f(\epsilon, N, |E|)$ [5]

Fig. 2. Privacy implications between different definitions of the $t$-closeness and DP models. Dashed arrows denote implications found by [5] and crossed-out arrows between two nodes signify that no implication exist between these privacy definitions. The numbered expressions indicate when the implications are valid. Refer to Figure 1 for more details of implications between the definitions for different values of the group size parameter.

TABLE II

DESCRIPTIONS OF DIFFERENT PRIVACY DEFINITIONS FOR DP AND SYNTACTIC PRIVACY

| Name | Description |
|---|---|
| $k$-anonymity [7] | Merges records into groups of at least $k$ records with identical non-confidential attributes. |
| Multiplicative $t$-closeness [5] | Restricts the difference in distribution of confidential attributes between the overall distribution in the dataset and the distribution within groups of records with identical, non-confidential attributes, see Definition 4. |
| $(k, t)$-closeness | Multiplicative $t$-closeness for every possible subset of $k$ records in a dataset irrespective of their non-confidential attributes, see Definition 7. |
| Original $\epsilon$-DP [2] | Conceals the contribution of individual records to datasets, see Definition 5. |
| $\epsilon$-LDP [5] | Applies the $\epsilon$-DP constraints to pairwise comparisons of records, see Definition 6. |
| $(k, \epsilon)$-DP | Like the original $\epsilon$-DP but conceals the contribution from groups of $k$ records, see Definition 8. |

reciprocal implication between multiplicative $t$-closeness with $k$-anonymity and $\epsilon$-LDP.

*Theorem 1: Multiplicative $t$-closeness $\not\Rightarrow$ $\epsilon$-LDP*

The previously shown implication from $\epsilon$-LDP to multiplicative $t$-closeness [5] can be understood by observing that $\epsilon$-LDP limits the differences between individual records' confidential values, which in the extreme results in all records' confidential attributes being identical. This will in turn guarantee that any partition of the dataset will fulfil multiplicative $t$-closeness.

The lack of a reverse implication on the other hand follows from the fact that multiplicative $t$-closeness computes the average probability that a record in an equivalence class takes on certain confidential values. Relying on the average probability means that an individual record's probabilities are free to vary within each equivalence class and can even approach zero without making $t$ approach $\infty$. Since the presence of a zero probability for a confidential value would make $\epsilon = \infty$, Theorem 1 therefore must hold.

Note that an exception to Theorem 1 occurs if $k = 1$ and $|E_i| \leq 1$ for all equivalence classes $E_i$ in the dataset. In this special case multiplicative $t$-closeness implies $\ln t^2$-LDP which will become more evident in Theorem 3. However, setting $k = 1$ does not necessarily restrict the sizes of all equivalence classes in general which is why multiplicative $t$-closeness $\not\Rightarrow$ $\epsilon$-LDP.

The relationship between LDP and multiplicative $t$-closeness does not necessarily help us to intuitively understand the relationship between $\epsilon$-DP and multiplicative $t$-closeness since LDP is a conceptually very different privacy definition from $\epsilon$-DP. In order to extend the mapping between DP and $t$-closeness, we proceed by extending $t$-closeness and $\epsilon$-DP and use them as a bridge to better understand the relationship between $\epsilon$-DP and multiplicative $t$-closeness.

### *A. Extending t-closeness*

The foundation on which Theorem 1 is built on can be understood by considering that multiplicative $t$-closeness does not necessarily force all records' confidential attributes to converge onto similar values, which would have been necessary to have an implication to $\epsilon$-LDP. With this result in mind, we aim to bypass Theorem 1 by introducing a new extension of multiplicative $t$-closeness referred to as $(k, t)$-closeness. $(k, t)$-closeness applies restrictions similar to multiplicative $t$-closeness on all subsets of a given size of the dataset, not just the equivalence classes of one given partition.

*Definition 7 ($(k, t)$-closeness: Multiplicative $t$-closeness for all Fixed Size Subsets): Consider a mechanism $Z$ and a dataset $D$. Let $\hat{D}$ be the sanitised dataset resulting from applying $Z$ to $D$. Let $k \in [1, N-1]$ be a fixed parameter and $\{B_1, \ldots B_p\}$ be a set of buckets such that $\bigcup_{j=1}^{p} B_j = range(Z)$ and $B_j \cap B_l = \emptyset, \forall j, l \in [1:p] : j \neq l$. $\hat{D}$ fulfils $(k, t)$-closeness iff for every possible subset $K$ of size $k$ drawn from $\hat{D}$ and all $l \in [1:p]$*

$$t^{-1} \leq \frac{\Pr_{\hat{D}}(B_l)}{\Pr_K(B_l)} \leq t$$

*where*

$$\Pr_{\hat{D}}(B_l) = \frac{1}{N} \sum_{i \in [1:N]} \Pr(Z(c_i) \in B_l)$$

*and*

$$\Pr_K(B_l) = \frac{1}{k} \sum_{i \in K} \Pr(Z(c_i) \in B_l)$$

$(k, t)$-closeness extends multiplicative $t$-closeness by redefining the underlying equivalence classes. Where multiplicative $t$-closeness relies on the partition created by $k$-anonymity (recall from Definition 4 that $k$ is the size of the smallest equivalence class), $(k, t)$-closeness considers every possible partition of the data into groups of $k$ records and allows the 'equivalence' classes to overlap. Some additional properties of $(k, t)$-closeness and its relationship with $t$-closeness are presented in Lemma 2.

*Lemma 2 (Properties of $(k, t)$-closeness): $(k, t)$-closeness has the following properties*:

1) *$(k_0, t)$-closeness $\Rightarrow$ $(k_1, t)$-closeness if $k_1 \geq k_0$.*
2) *$(k_0, t)$-closeness $\nRightarrow$ $(k_1, t)$-closeness if $k_1 < k_0$.*
3) *$(k, t)$-closeness $\underset{\nLeftarrow}{\Rightarrow}$ Multiplicative $t$-closeness*
4) *$(k, t_0)$-closeness $\Rightarrow$ $(k, t_1)$-closeness if $t_0 \leq t_1$.*

The missing implications from higher to lower $k$ for $(k, t)$-closeness follow on from the inability to ensure that a group of size smaller than $k$ has a probability of zero for a bucket value when a $(k, t)$-Close dataset is given for $k < \infty$. Point 3 in Lemma 2 is a consequence of both point (1) in Lemma 2 and the fact that $(k, t)$-closeness always computes an equivalent or stricter $t$ than multiplicative $t$-closeness.

In fact, $\epsilon$-LDP is closely related to $(1, t)$-closeness, as shown in the following theorem:

*Theorem 2:*

- *$(1, t)$-closeness $\Rightarrow$ $\epsilon$-LDP with $e^\epsilon = t^2$, and*
- *$\epsilon$-LDP $\Rightarrow$ $(1, t)$-closeness with $t = \frac{1}{N}(e^\epsilon(N-1)+1)$.*

Combining the above theorem and Lemma 2, we recover Lemma 1. In addition, we show the following:

*Theorem 3: For $k > 1$,*

- *$\epsilon$-LDP $\Rightarrow$ $(k, t)$-closeness where $t = \frac{k}{N}(1 + \frac{N-k-1}{k}e^\epsilon)$, but*
- *$(k, t)$-closeness $\nRightarrow$ $\epsilon$-LDP.*

### *B. Generalising Differential Privacy*

Having shown that privacy implications exist between multiplicative $t$-closeness and $\epsilon$-LDP, we now also want to include the original $\epsilon$-DP in the privacy mapping. First we present a generalisation of DP inspired by Dwork [16], which takes group privacy into consideration. We also draw inspiration from Liu *et al.* [22] which takes groups into consideration in $\epsilon$-DP due to record correlations. We refer to this generalisation of $\epsilon$-DP as $(k, \epsilon)$-differential privacy, which extends $\epsilon$-DP by allowing the impact of the information about a group of size $k$ to be restricted instead of only restricting the impact from any individual. Intuitively, as $t$-closeness depends on the behaviour of a group of $k$ records, the $\epsilon$-DP counterpart should also exhibit this behaviour.

*Definition 8 ($(k, \epsilon)$ Differential Privacy): Consider any dataset $D$. Let $D_{-\mathrm{K}}$ denote a subset of $D$ with a group $K \subsetneq [1:N]$ of $k$ records removed and $\{B_1, \dots B_p\}$ a set of buckets such that $\bigcup_{j=1}^{p} B_j = \mathrm{range}(Z)$ and $B_j \cap B_l = \emptyset, \forall j, l \in [1:p] : j \neq l$. A privacy mechanism $Z$ fulfils $(k, \epsilon)$-DP for $D$ if for all selections of $k$ record removals and all $l \in [1:p]$*

$$e^{-\epsilon} \leq \frac{\Pr(Z(Q(D)) \in B_l)}{\Pr(Z(Q(D_{-\mathrm{K}})) \in B_l)} \leq e^\epsilon.$$

This generalisation of DP allows the data publisher to restrict the effect of small groups on the result in order to prevent them from being identified. Some other properties of $(k, \epsilon)$-DP and how it relates to $\epsilon$-DP are presented in Lemma 3.

*Lemma 3 (Properties of $(k, \epsilon)$-DP): $(k, \epsilon)$-DP has the following properties*:

1) *$(k_0, \epsilon)$-DP $\Rightarrow$ $(k_1, \epsilon)$-DP if $k_0 \geq k_1$.*
2) *$(k_0, \epsilon)$-DP $\nRightarrow$ $(k_1, \epsilon)$-DP if $k_0 < k_1$.*
3) *$(1, \epsilon)$-DP $\Leftrightarrow$ $\epsilon$-DP.*
4) *$(k, \epsilon)$-DP $\underset{\nLeftarrow}{\Rightarrow}$ $\epsilon$-DP when $k > 1$.*
5) *$(k, \epsilon_0)$-DP $\Rightarrow$ $(k, \epsilon_1)$-DP if $\epsilon_0 \leq \epsilon_1$.*

### *C. New Implication Results*

All privacy implications presented so far have been based on the ability to restrict individual records' confidential attributes. With the introduction of $(k, \epsilon)$-DP we now shift focus and look at restricting the confidential attributes of groups. Comparing the definitions of $(k, t)$-closeness and $(k, \epsilon)$-DP suggests that there exists a connection between $t$-closeness and DP. In fact Theorem 4 states that $(k, \epsilon)$-DP $\Leftrightarrow$ $(N - k, e^\epsilon)$-closeness.

*Theorem 4:*

1) *For $k_{\mathrm{DP}} = N - k_{\mathrm{tClo}}$, $(k_{\mathrm{tClo}}, t)$-closeness $\Leftrightarrow$ $(k_{\mathrm{DP}}, \ln t)$-DP.*
2) *For $k_{\mathrm{DP}} > N - k_{\mathrm{tClo}}$, $(k_{\mathrm{tClo}}, t)$-closeness $\underset{\nLeftarrow}{\Rightarrow}$ $(k_{\mathrm{DP}}, \epsilon)$-DP.*
3) *For $k_{\mathrm{DP}} < N - k_{\mathrm{tClo}}$, $(k_{\mathrm{tClo}}, t)$-closeness $\overset{\nRightarrow}{\Leftarrow}$ $(k_{\mathrm{DP}}, \epsilon)$-DP.*

Theorem 4 establishes a strong reciprocal relationship between DP and syntactic privacy. It follows from point (1) in Theorem 4 that a dataset fulfilling the traditional $\epsilon$-DP (i.e. $(k, \epsilon)$-DP with $k = 1$) also satisfies $(N - 1, \epsilon)$-closeness. Considering that DP tries to hide individual values by making results similar for datasets differing in only one value, it makes sense that the same effect can be accomplished by ensuring that every subset of size $N - 1$ is similar to the overall dataset.

Before comparing multiplicative $t$-closeness and $(k, \epsilon)$-DP, a decision has to be made regarding which records in a dataset, that satisfies multiplicative $t$-closeness, should be considered when calculating the corresponding privacy level for $(k, \epsilon)$-DP. Since a dataset that fulfils multiplicative $t$-closeness has a structure that maps any record to at least $k$ other records we propose to calculate the $(k, \epsilon)$-DP privacy level within every equivalence class and not globally in the whole dataset. Based on this decision we can now show that multiplicative $t$-closeness $\nRightarrow$ $(k, \epsilon)$-DP.

*Theorem 5: Multiplicative $t$-closeness $\nRightarrow$ $(k, \epsilon)$-DP*

### *D. Interpretability*

A second look at Figure 1 reveals an interpretation of $t$-closeness in terms of DP and vice versa as illustrated in Figure 3. Note that grey arrows are used in Figure 3 to explain how one privacy definition can be turned into another one, not to be confused with implication arrows. The close relationship between $(k, t)$-closeness and $(k, \epsilon)$-DP enable us to establish a link between the privacy concepts of $t$-closeness and DP. In fact, we can interpret DP as multiplicative $t$-closeness

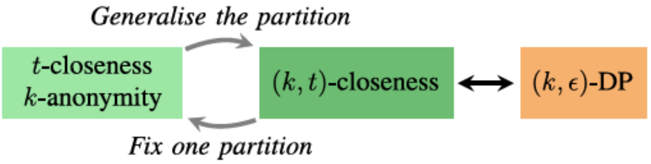


Fig. 3. $(k, t)$-closeness provides a link between the concept of $t$-closeness and DP. DP can be interpreted as $t$-closeness with a generalised partition while $t$-closeness can be seen as DP with a fixed partition.

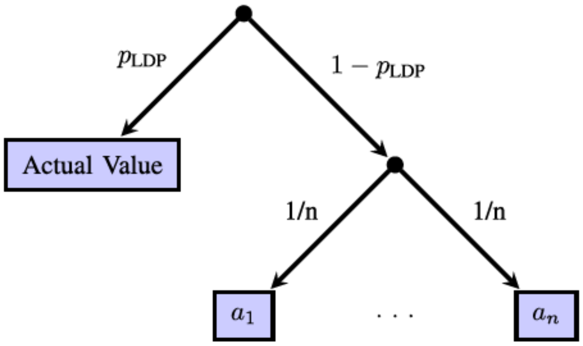


Fig. 4. A randomised response mechanism for LDP for a nominal variable where $p_{\text{LDP}}$ is selected to reach a desired value of $\epsilon$.

with a generalised partition, which describes the extension from multiplicative $t$-closeness to $(k, t)$-closeness. In the other direction we can interpret multiplicative $t$-closeness as DP where we only consider the privacy guarantee for one fixed partition of the dataset with groups of size $k$ or more.

## V. Evaluation

We will now show how the quality of sanitised data can be improved using our privacy implications. To this end, we will consider a range of datasets that need to satisfy two privacy definitions. Using our privacy implications, we then obtain better utility using just a $\epsilon$-LDP mechanism to satisfy both $\epsilon$-LDP and multiplicative $t$-closeness, instead of using two mechanisms, one for each privacy definition.

### A. Mechanism Selection

Randomised response [6] is a mechanism invented to allow deniability in reported survey responses. The mechanism specifies that every answer has a chance of being replaced with a value selected randomly from a fixed set of answers. This ensures that every participant in the survey can deny that the output of the mechanism is their true response. More recently, randomised response has been shown to satisfy DP [34] and LDP [35], and can be used in a modified version to ensure stronger privacy guarantees for data mining [36].

A randomised response algorithm can be represented by a directed stochastic decision tree of nodes connected by branches where every branch is associated with a probability of following that particular branch to a lower-level node; see Figure 4. The output of the algorithm is decided by the leaf node reached when starting from the root node and traversing the tree down through the branches.

To more easily compare and provide an explanation for the potential utility gain of using our proposed framework, we decided to focus on the randomised response mechanism for $\epsilon$-LDP represented in Figure 4. The mechanism is designed for a nominal variable and a parameter $p_{\text{LDP}}$ defines the probability that the output equals the input while there is a $1 - p_{\text{LDP}}$ chance that the output will be uniformly sampled from the set $\{a_1, \ldots, a_n\}$.[3] It can be shown that the parameter $p_{\text{LDP}}$ controls the strongest $\epsilon$ privacy level that the mechanism offers, as follows:

$$\epsilon = \ln\left(1 + \frac{p_{\text{LDP}}}{N(1 - p_{\text{LDP}})}\right). \tag{1}$$

Inspired by the use of randomised response in $\epsilon$-LDP we introduce a similar mechanism for a nominal variable for multiplicative t-closeness; see Figure 5. Similar to the mechanism for $\epsilon$-LDP, a parameter $p_t$ determines the probability that the output value equals the input value but $(1 - p_t)$ instead determines the chance that the output will be sampled from the empirical distribution of the attribute values in the entire dataset.[4] The randomised response mechanism is applied locally to every record individually after the dataset has been partitioned by a $k$-anonymity mechanism.

To find the dependence of the privacy level $t$ on $p_t$ we observe that independent of how the dataset is partitioned by a $k$-anonymity mechanism, an upper limit estimation of the value of $t$ can be provided by a group whose confidential attributes takes on the same value $a_i \in \{a_1, \ldots, a_n\}$ before applying the $t$-closeness mechanism. These extreme cases enable us to determine the value of $t$ as follows:

$$\begin{aligned} t &= \max_i \max\left\{\frac{\mu_i}{p_t/p_y(a_i) + (1 - p_t)}, \frac{\mu_i}{p_y(a_i)(1 - p_t)}\right\}, \\ \mu_i &= p_y(a_i)(p_t + (1 - p_t)p_y(a_i) + (1 - p_t)\sum_{j \neq i} p_y(a_j)). \end{aligned} \tag{2}$$

We will use these randomised response mechanisms to show benefits of privacy implications based on the gain in utility, defined in the next subsection.

### B. Utility Measure

Consider a dataset modelled by a random variable $X$ that contains some confidential and some non-confidential information. From $X$ we extract some information represented by the random variable $Y$ for data release. Before releasing the extracted information, a privacy mechanism $p_{Z|Y}(\cdot|\cdot)$ will be applied to $Y$, creating the random variable $Z$, to limit the amount of confidential information from $X$ that can be derived from observing $Z$. However, in addition to satisfying the privacy requirement we also need to consider the utility of $p_{Z|Y}(\cdot|\cdot)$, defined as how much information about $Y$ that can be extracted from $Z$. For this purpose we consider a user who based on observing $Z$ can use some post-processing to improve their guess $\hat{Y}$ of $Y$. A gain function $g(y, \hat{y})$ further

[3] This adheres to the spirits of common mechanisms for $\epsilon$-DP where independent noise of a specific distribution (independent of the dataset's statistics) is added [16].

[4] This choice ensures that the distribution of subsets of records gets closer to that of the dataset as $p_t$ decreases.

measures how useful it is to guess $\hat{y}$ when the actual value is $y$.

For the global utility $\mathcal{U}(Y, Z)$ of the privacy mechanism we use the definition presented by Alvim *et al.* [37]

$$\mathcal{U}(Y, Z) = \sum_{y} p_y(y) \sum_{\hat{y}} p(\hat{y}|y) g(y, \hat{y}) \tag{3}$$

A special case of the gain function is the binary gain function, where all guesses $\hat{y}$ but the correct one, $y$, are equally useless, or $g(y, \hat{y}) = \delta_{y,\hat{y}}$. This is applicable when no obvious distance metric exists to measure how the usefulness of a guess $\hat{y}$ changes based on the distance from the true answer $y$.

For the binary gain function, (3) can be simplified to the following:

$$\mathcal{U}(Y, Z) = \sum_{z} \max_{y} p(y, z). \tag{4}$$

The expression in (4) represents the utility and gain functions that will be used in this paper. For our utility computations we further assume that $N$ is large so that the frequency of 1s and 0s in a dataset, $f_y$, closely resembles the underlying distribution $p_y$. That is: $p(y, z) = p_y p_{z|y} \approx p_{z|y} f_y$.

Expanding the utility function in (4) for the $\epsilon$-LDP mechanism gives

$$\mathcal{U}_{\text{LDP}}(Y; Z) = \sum_{z} \max \left\{ \max_{j \neq i}(p_y(a_j)) \frac{1 - p_{\text{LDP}}}{N}, p_y(a_i)(p_{\text{LDP}} + \frac{1 - p_{\text{LDP}}}{N}) \right\} \tag{5}$$

To similarly compute the utility for the combined multiplicative $t$-closeness and $\epsilon$-LDP mechanism we first define the functions $f_1$ and $f_2$ to simplify our notation.

$$\begin{aligned} f_1(a_i) &= p_y(a_i)(p_t[p_{\text{LDP}} + (1 - p_{\text{LDP}})/N] \\ &\quad + (1 - p_t)[p_y(a_i)p_{\text{LDP}} + (1 - p_{\text{LDP}})/N]) \\ f_2(a_i, a_j) &= p_y(a_i)(p_t(1 - p_{\text{LDP}})/N \\ &\quad + (1 - p_t)[p_y(a_j)(p_{\text{LDP}} + (1 - p_{\text{LDP}})/N)]) \end{aligned}$$

Using $f_1$ and $f_2$ from above we can now express the utility function as

$$\mathcal{U}_{\text{tClo+LDP}}(Y; Z) = \sum_{j} \max_{i} \max\{f_1(a_i), f_2(a_i, a_j)\} \tag{6}$$

In addition, we can also derive the utility function for the randomised response based multiplicative $t$-closeness mechanism, $\mathcal{U}_{\text{tClo}}(Y; Z)$, by setting $p_{\text{LDP}} = 1$ in (6). Note that this utility function also corresponds to $\mathcal{U}_{\text{tClo+}\infty\text{-LDP}}(Y; Z)$ which provides an upper limit for $\mathcal{U}_{\text{tClo+LDP}}(Y; Z)$ when varying $\epsilon_0$.

### C. Evaluation Setup

To evaluate the potential utility gains of using our proposed framework, we focus on a dataset structure with a nominal confidential attribute. We consider a situation where a data publisher aims to guarantee a certain privacy level $t_0$ of multiplicative $t$-closeness and at the same time also ensure a privacy level $\epsilon_0$ of $\epsilon$-LDP. For this purpose, the data publisher can choose to either apply both a $\epsilon$-LDP and a multiplicative $t$-closeness mechanism or use our framework to guarantee both privacy definitions with a $\epsilon$-LDP mechanism. For this reason, we focus on the following three mechanisms:

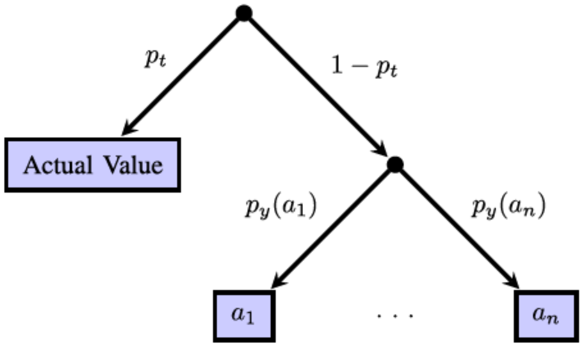


Fig. 5. A randomised response mechanism for $t$-closeness for a nominal variable where $p_t$ is selected to reach a desired $t$ and $p_y(a_i)$ is the frequency of the variable taking the value $a_i$, $i \in [1 : n]$.

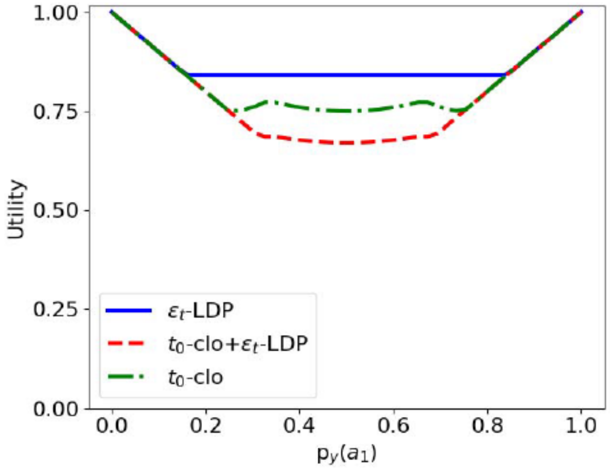


Fig. 6. Utility computed with (4) for the $\epsilon$-LDP mechanism (solid), combined multiplicative $t$-closeness and $\epsilon$-LDP mechanism (dashed), and $t$-closeness mechanism (dash dotted) for varying initial frequency of the attribute value $a_1$, $p_y(a_1)$, and $n = 2$. The utility of the $t$-closeness mechanism marks the upper limit for the utility of the combined mechanism for a fixed $t = t_0$. Thus, the utility for the $\epsilon$-LDP mechanism is always higher or equal to the utility for the combined mechanism.

1) The $\epsilon$-LDP mechanism in Figure 4
2) A sequential execution of the multiplicative $t$-closeness mechanism in Figure 5 and the $\epsilon$-LDP mechanism in Figure 4 where the root node of the $\epsilon$-LDP mechanism is attached to every leaf node in the multiplicative $t$-closeness mechanism
3) The $t$-closeness mechanism in Figure 5.

For mechanism 1 we use the relationship between $t$ and $\epsilon$, $N$ and $k$ for $\epsilon$-LDP [5] to compute the required privacy level of $\epsilon$-LDP, denoted as $\epsilon_t$, which guarantees both $\epsilon_0$ and $t_0$. We start by looking at examples where $\epsilon_t \succeq \epsilon_0$, but later argue for the result of letting $\epsilon_t \prec \epsilon_0$. Equation (5) is used to compute the utility for mechanism 1 where $p_{\text{LDP}}$ can be derived from (1) knowing $\epsilon_t$.

For mechanism 2 we choose $t = t_0$ and $\epsilon = \epsilon_0$ to satisfy our privacy requirement. To study the dependence of $\epsilon_0$ on the utility of mechanism 2 we evaluate $\mathcal{U}_{\text{tClo+LDP}}$ for two values of $\epsilon_0$: $\epsilon_t$ as computed for mechanism 1, and $\infty$. Selecting $\epsilon_0 = \infty$ gives an upper bound on $\mathcal{U}_{\text{tClo+LDP}}$ for a fixed value of $t_0$ and corresponds to only applying the multiplicative $t$-closeness mechanism, mechanism 3. Again, $p_{\text{LDP}}$ and $p_t$ can be derived from (1) and (2) respectively knowing $\epsilon_0$, $t_0$, and $p_y$.

To evaluate the difference between the utilities $\mathcal{U}_{\text{LDP}}$ and $\mathcal{U}_{\text{tClo+LDP}}$ we further vary the parameters $t_0$, the number of output options for the confidential attribute $n$, the frequency of $a_1$ in the initial dataset $p_y(a_1)$ with $n = 2$, the minimum group size parameter $k$ for multiplicative $t$-closeness, and the number of records $N$. As mentioned earlier, we first consider $\epsilon_t \succeq \epsilon_0$, which in numerical terms translates to $\epsilon_0 \geq \epsilon_t$. We will later complete the analysis for $\epsilon_0 < \epsilon_t$ with additional arguments. All other parameters can be derived from the above mentioned parameters.

Note that our assumption that the frequency of attribute values, $f_y$, closely resembles the underlying distribution $p_y$, enables us to use the utility functions $\mathcal{U}_{\text{LDP}}$ and $\mathcal{U}_{\text{tClo+LDP}}$ to compute the utility for our mechanisms without having to generate or use real-world datasets. Instead we represent a dataset with our parameters as follows: for a total of $N$ records, the number of records with attribute $a_i$ is $Np_y(a_i)$ for all $i \in [1:n]$. Varying $p_y(a_1)$ between 0 and 1 with $n = 2$ therefore enables us to compute the corresponding utility for any dataset of one binary variable with given privacy levels. Utilities for datasets with nominal attributes can similarly be computed by varying $n$ and the individual attribute values' frequencies.

### D. Results

To display the dependence of $p_y(a_1)$ on $\mathcal{U}_{\text{LDP}}$ and $\mathcal{U}_{\text{tClo+LDP}}$ we set $t_0 = 2$, $n = 2$, $N = 100$, and $k = 5$ which results in $\epsilon_t = 0.7$. The results are available in Figure 6. The shape of the graphs consist of splices of linear, constant, and polynomial functions where the splice occurs when there is a shift in which terms are the largest in the sums representing $\mathcal{U}_{\text{LDP}}$ and $\mathcal{U}_{\text{tClo+LDP}}$. From Figure 6 it is evident that choosing the same $\epsilon = \epsilon_t$ for mechanism 1 and 2 results in a value of $\mathcal{U}_{\text{LDP}}$ that is higher or equal to $\mathcal{U}_{\text{tClo+LDP}}$ for all values of $p_y(a_1)$. Furthermore, considering that $\mathcal{U}_{\text{tClo}}$ provides an upper limit for $\mathcal{U}_{\text{tClo+LDP}}$, corresponding to $\epsilon_0 = \infty$, and that $\mathcal{U}_{\text{tClo}}$ is lower or equal to $\mathcal{U}_{\text{LDP}}$ for all values of $p_y(a_1)$ we can conclude that the $\epsilon$-LDP mechanism has better utility than the combined mechanism for any $p_y(a_1)$ and $\epsilon_0 \preceq \epsilon_t$. If $\epsilon_0 \succ \epsilon_t$ we would have to select $\epsilon = \epsilon_0$ instead of $\epsilon_t$ in mechanism 1 to satisfy our required privacy levels of $t_0$ and $\epsilon_0$. Thus, both mechanism 1 and 2 are using the same privacy level $\epsilon = \epsilon_0$ for $\epsilon$-LDP and will have the same reduction in utility caused by applying the $\epsilon$-LDP part of the mechanism. However, mechanism 2 has an additional reduction in utility caused by applying the multiplicative $t$-closeness mechanism as seen in Figure 6 and 7 where the same $\epsilon$ is used for both mechanisms.

We also observed that decreasing the ratio of $N$ to $k$ causes the breakaway points between the utility curves of the $\epsilon$-LDP and combined mechanism in Figure 6 to move further away from each other. This increases the range of $p_y(a_1)$ values for which the $\epsilon$-LDP mechanism's utility is strictly larger than the combined mechanism's utility for $\epsilon = \epsilon_t$.

The values of $\mathcal{U}_{\text{LDP}}$ and $\mathcal{U}_{\text{tClo+LDP}}$ for $t_0 \in [1, 10]$, $p_y(a_i) = 1/n, i \in [1:n]$, $n = 2, 4, 6$, $N = 100$, and $k = 5$ are presented in Figure 7. Again, the utility for the multiplicative $t$-closeness mechanism provides an upper limit for the utility for the combined mechanism when varying $\epsilon_0$ for all tested values of $n$. Additionally, the difference between the highest and lowest utility of the different mechanisms seems to increase with increasing $n$. We can therefore conclude that mechanism 1 has higher or equal utility than mechanism 2 for all values of $t_0$ and $\epsilon_0$ for our setups. Selecting $p_y(a_1)$ close to either 0 or 1 does however cause $\mathcal{U}_{\text{LDP}}$ and $\mathcal{U}_{\text{tClo+LDP}}$ to coincide for smaller values of $t_0$, but we chose to display $p_y(a_i) = 1/n, i \in [1:n]$ since it gives the largest difference in utility.

To put our results into context, consider a large dataset with 2 binary attributes that describes if a hospital patient has tested positive to HIV and whether or not they have skin cancer. If our dataset has a uniform distribution over the combination of the attributes and we select $t = 2$ in the setup described in this section, then Figure 7 tells us that using our framework almost doubles the utility of the sanitised data compared to applying the LDP mechanism and multiplicative $t$-closeness mechanism sequentially. With better utility we can, for example, expect better machine learning models being developed based on our sanitised data.

### E. Additional Remarks

The mechanisms introduced in Figure 4 and 5 can also be used in settings with multiple nominal confidential attributes. Let $a^1_{i_1}, a^2_{i_2}, \ldots, a^m_{i_m}$ where $i_j \in [1:n_j]$ for $j \in [1:m]$ represent the possible values for $m$ confidential attributes. Instead of releasing the attributes separately by applying the privacy mechanism independently to each attribute, we can instead release them in one vector. We must then consider the distribution of all combinations of values for the confidential attributes and the output for the mechanism will be of the format $[a^1_{i_1}, a^2_{i_2}, \ldots, a^m_{i_m}]$. With this in mind we can now argue that the results in Figure 7 also apply to multiple nominal attributes. For example, $n = 4$ not only represent one nominal attribute with 4 possible values but also corresponds to 2 binary attributes. Hence, our results show that increasing the amount of attributes increases the utility difference between the mechanisms.

In addition, we note that the computation of achieved $t$ in (2) from using mechanism 3 can also be used to compute the achieved $t$ for $(k, t)$-closeness. The computation in (2) relies on the worst-case scenarios of all confidential values being identical in an equivalence class, which corresponds to the same computation needed to compute $t$ for $(k, t)$-closeness. By applying Theorem 4 we can now also argue that mechanism 3 can be used to satisfy $(k, \epsilon)$-DP. With this in mind, we can provide another example for when our framework provides increased utility. Suppose we need to achieve both $(k_0, t_0)$-closeness and $(N - k_1, \epsilon_0)$-DP where $k_0 \geq k_1$. Selecting mechanism 3 to be used for both

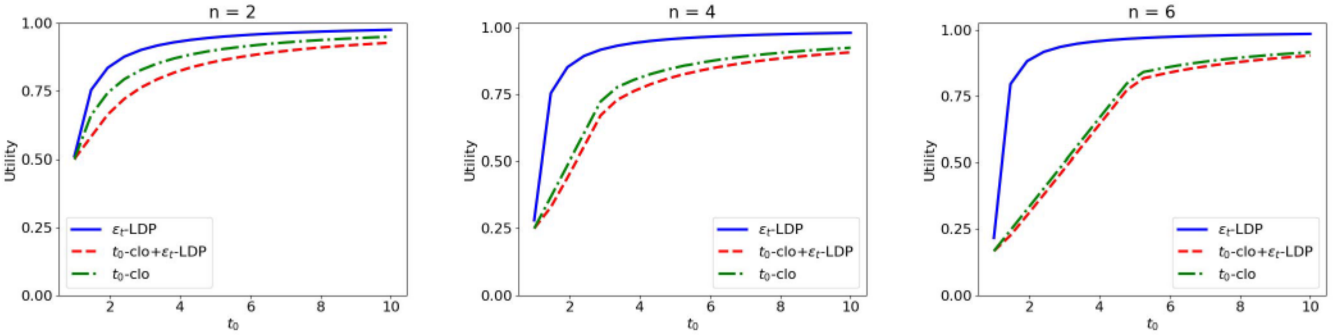


Fig. 7. Utility computed with (4) for varying $t_0$ for the $\epsilon$-LDP mechanism (solid), combined $t$-closeness and $\epsilon$-LDP mechanism (dashed), and $t$-closeness mechanism (dash dotted) when the number of output variables $n$ is 2,4, and 6. The utility of the $t$-closeness mechanism marks the upper limit for the utility from the combined mechanism for a fixed $t = t_0$ and shows that the utility for the $\epsilon$-LDP mechanism is always higher or equal to the utility for the combined mechanism. The utility difference is also seen to increase with increasing $n$.

$(k, t)$-closeness and $(k, \epsilon)$-DP, we will now compare the use of mechanism 3 once to achieve both $(k_0, t_0)$-closeness and $(k_0, \epsilon_0)$-DP, Scheme 1, to using mechanism 3 twice as a concatenation of the mechanisms for $(k_0, t_0)$-closeness and $(k_0, \epsilon_0)$-DP, Scheme 2. For Scheme 1, Theorem 4 gives $\epsilon_t = \epsilon_0$, which implies that we need to apply the $(k_0, \epsilon_0)$-DP mechanism once to satisfy both privacy definitions. Since Scheme 2 uses this exact mechanism plus another mechanism we can conclude that Scheme 1 provides better utility than Scheme 2 due to additional noise being applied in Scheme 2. Thus, our framework improves the utility in this case too.

As a final remark, it should be mentioned that the utility gain from using the proposed framework is dependent on existing mechanisms for the privacy definitions between which an implication exists. The framework reveals new mechanisms for privacy definitions by allowing mechanisms from other privacy definitions to be used. If a utility gain is achieved or not therefore depends on properties of these mechanisms as well as and properties of the found implication function. For example, we would expect to see a utility gain in the case A $\Rightarrow$ B where the mechanism for A has better or similar utility than the mechanism for B, and the implication function $f(\alpha)$ does not require very strict privacy levels $\alpha$ for A to achieve any privacy level $\beta$ for B.

## VI. Conclusion

We have shown that DP and a form of syntactic privacy called $t$-closeness are closely linked in both directions in terms of how the privacy levels can be translated between the models. An extension and respectively a generalisation to the models, $(k, t)$-closeness and $(k, \epsilon)$-DP, provide the missing link that enables us to more fully connect $\epsilon$-LDP, $k$-anonymity with multiplicative $t$-closeness, and $\epsilon$-DP. This extended knowledge of the models' connection can assist data publishers when deciding on an appropriate privacy definition for their application and can increase the utility of the data in cases where both $t$-closeness and DP are desired. We showed how using a $\epsilon$-LDP mechanism to guarantee both $\epsilon$-LDP and multiplicative $t$-closeness for certain desired privacy levels, $t_0$ and $\epsilon_0$, provides better data utility than sequentially applying a multiplicative $t$-closeness mechanism and a $\epsilon$-LDP mechanism to obtain the same privacy levels.

Future research on this topic could reveal functions that give a more precise estimate of the optimal privacy levels in the implications than what we have shown in this paper. It may also be possible to show more general results regarding the utility gains of utilising our framework. In addition, our privacy implication mapping can be extended to include additional privacy definitions under the privacy models $t$-closeness and differential privacy.

## Appendix

This section contains proofs of Lemmas and Theorems from the main part of the paper. We consider the tables presented in the proofs as being the result of having applied a privacy mechanism $Z$ to a dataset. The tables have two or three of the following attributes: *Id* which is the row number of the record, *QI* which is a quasi-identifier attribute for the records, and *attribute probability* which presents the probability that a record with Id $i \in [1 : N]$, where N is the number of records, has a confidential attribute value $c_i = a_1$. Here we only consider the case where the confidential attribute is limited to two values, $a_1$ and $a_2$, as it captures the ideas required for proving the results in general. We use the notation $q_i$, $i \in [1 : N]$, to represent the quasi-identifier for the record with Id $i$ and $x_0 \in range(\text{QI})$ is a value used to sort the records into equivalence classes based on their quasi-identifiers.

For our proofs we use the set of buckets $\{B_1, \ldots, B_p\}$ as defined in Definition 7. We have also adopted a shortened notation for $\Pr(Z(c_i))$, where $\Pr(Z(c_i)) = \Pr(Z(c_i) \in B_l)$, $l \in [1 : p]$, in order to declutter some equations. In these cases, buckets play a minor role in the proof and it is implied that the equations are to hold for all buckets $B_l$, $l \in [1 : p]$.

### A. Proof of Theorem 1

*Proof:* (Theorem 1) Consider the two equivalence classes in Table III. Using Definition 4, we get

$$\Pr_{\hat{D}}(\{a_1\}) = \Pr_{E_1}(\{a_1\}) = \Pr_{E_2}(\{a_1\}) = \mu,$$
$$\Pr_{\hat{D}}(\{a_2\}) = \Pr_{E_1}(\{a_2\}) = \Pr_{E_2}(\{a_2\}) = 1 - \mu.$$

The anonymised dataset in Table III thereby satisfies 1-closeness independently of the value of $\sigma$. To calculate $\epsilon$ in local $\epsilon$-DP for the dataset we note that from Table III we

TABLE III

RECORDS 1-2 AND 3-4 CONSTITUTE TWO EQUIVALENCE CLASSES $E_1$ AND $E_2$. THIS DATASET FULFILS MULTIPLICATIVE 1-CLOSENESS AND LOCAL $\ln{(2\mu-\sigma)/\sigma}$-DP

| | Id | QI | Attribute probability |
|---|---|---|---|
| $E_1$ | 1 | $qi_1 \leq x_0$ | $\Pr(c_1 = a_1) = \mu$ |
| | 2 | $qi_2 \leq x_0$ | $\Pr(c_2 = a_1) = \mu$ |
| $E_2$ | 3 | $qi_3 > x_0$ | $\Pr(c_3 = a_1) = \sigma$ |
| | 4 | $qi_4 > x_0$ | $\Pr(c_4 = a_1) = 2\mu - \sigma$ |

have $\Pr(Z(c_1) \in \{a_1\}) = \mu$ and $\Pr(Z(c_3) \in \{a_1\}) = \sigma$. From Definition 6, the dataset fulfils local $\epsilon$-DP only if $\epsilon \geq \ln(\mu/\sigma)$.

Now suppose that there exists a function $f$ such that multiplicative $t$-closeness $\Rightarrow f(t)$-LDP. We know that $\epsilon = \ln \mu/\sigma$ and that $\mu$ is independent of $\sigma$, which imply that letting $\sigma \to 0$ causes $\epsilon \to \infty$. We also know that the mechanism that has been applied to Table III guarantees multiplicative $t$-closeness for $t = 1$, that is, the $t$ value is independent of $\sigma$. The value of $\sigma$ can therefore be chosen arbitrarily close to 0 without affecting the $t$ value. In addition, if we let each unique value of $\sigma$ represent a different mechanism for multiplicative $t$-closeness, then we can always choose a sufficiently small $\sigma$ such that the resulting optimal privacy level $\epsilon^* = \ln(\mu/\sigma)$ fulfils $f(t) \succ \epsilon^*$. Hence, there exists at least one mechanism for multiplicative $t$-closeness that cannot guarantee $\epsilon$-LDP for a fixed $\epsilon$ which leads to a contradiction. Thus, multiplicative $t$-closeness $\not\Rightarrow$ local $\epsilon$-DP. See Section A for additional comments. ■

*1) Additional Comments for the Proof of Theorem 1:* We can further show that the privacy parameters for multiplicative $t$-closeness, $k$ and $N$, where $k$ is the size of the smallest equivalence class in $\hat{D}$ and $N$ is the number of records, also do not impact the lack of implication between multiplicative $t$-closeness and local $\epsilon$-DP. Suppose that there exists another function $f_{k,N}$ such that multiplicative $t$-closeness $\Rightarrow$ local $f_{k,N}(t)$-DP. Such a function would only be able to provide a lower limit for $\epsilon$ if $k, N$, and $t$ affect the value of $\epsilon$. We can however show that $\epsilon$ can vary independently of $k$, $N$, and $t$. For example, record 1 in Table III can be replicated and added to the first equivalence class without affecting $t$ or $\epsilon$ since the mean probability is preserved. Hence $t$ and $N$ cannot solely determine $\epsilon$. Furthermore, $k$ can be increased by adding records to the first equivalence class as described and changing the second equivalence class to contain $k$ records where $k-1$ records contain attribute $a_1$ with probability $\sigma/(k-1)$ and the remaining record contains $a_1$ with probability $k\mu - (k-1)\sigma$. This new table has the same value of $t$ and $\epsilon$ can still vary independently of $t, k$ and $N$.

## B. Proof of Lemma 2

*Proof:* (Lemma 2) The proof of points (1)-(3) in Lemma 2 can be simplified by observing that the value of $t$ in $(k,t)$-closeness is solely determined by the group of $k$ records, $E^k_{\max}$ and $E^k_{\min}$, that has the highest respectively smallest probability of taking on any of the possible bucket values. For the three properties we can then state the following

TABLE IV

THIS DATASET SATISFIES $(k,t)$-CLOSENESS WHERE $t = N - k + 1 < \infty$ WHILE $t = \infty$ FOR $(k_0, t)$-CLOSENESS WITH $k_0 < k$

| Id | Attribute probability |
|---|---|
| 1 | $\Pr(c_1 = a_1) = 0$ |
| ... | ... |
| $k-1$ | $\Pr(c_{k-1} = a_1) = 0$ |
| $k$ | $\Pr(c_k = a_1) = \mu$ |
| ... | ... |
| $N$ | $\Pr(c_N = a_1) = \mu$ |

1) Consider the two record groups $E^{k_0}_{\max}, E^{k_0}_{\min}$ of size $k_0$ that have the highest respectively smallest probability of taking on any bucket value $B_l$, $l \in [1:p]$, in a $(k_0, t_0)$-closeness dataset. More formally, $E^{k_0}_{max} = \{E_i : E_i = \{r_{j_1}, r_{j_2}, \ldots, r_{j_{k_0}}\}, E_j = \{r_{j'_1}, r_{j'_2}, \ldots, r_{j'_{k_0}}\}$ such that $\Pr_{E_i}(B_l) \geq \Pr_{E_j}(B_l)$ and $j_q \neq j_{q^1}, j'_{q'} \neq j'_{q'^1}$, if $q \neq q^1, q' \neq q'^1$ for $j_q, j_{q'} \in [1:N], l \in [1:p]$, and $q, q^1, q', q'^1 \in [1:k_0]\}$, where $r_j$ is referring to a record in the dataset. We can similarly define $E^{k_0}_{min}$ with the inequality turned so that $\Pr_{E_i}(B_l) \leq \Pr_{E_j}(B_l)$. For the corresponding record groups $E^{k_1}_{\max}, E^{k_1}_{\min}$ of size $k_1 \geq k_0$ we have two cases.
   a) An extreme valued record set $E^{k_1}_x$, $x \in \{\text{max,min}\}$ occurs for the *same* bucket value $B_l$ as $E^{k_0}_x$. Then $E^{k_0}_x \subseteq E^{k_1}_x$ and any additional records in $E^{k_1}_x$ will have probabilities closer to the mean value of the whole dataset or equal to the $k_0$th highest or lowest probability. That is, $\Pr_{E^{k_1}_{\max}}(B_l) \leq \Pr_{E^{k_0}_{\max}}(B_l)$ or $\Pr_{E^{k_1}_{\min}}(B_l) \geq \Pr_{E^{k_0}_{\min}}(B_l)$ which results in $t_1 \leq t_0$ according to Definition 7.
   b) An extreme valued record set $E^{k_1}_x$, $x \in \{\text{max,min}\}$ occurs for a *different* bucket value $B_j$, $j \in [1:p]: j \neq l$. $\Pr_{E^{k_0}_{\text{x}}}(B_j)$ must then be closer or equally close to the mean value compared to $\Pr_{E^{k_0}_{\text{x}}}(B_l)$. Using the result in (a) we therefore get
   $\Pr_{E^{k_1}_{\max}}(B_j) \leq \Pr_{E^{k_0}_{\max}}(B_j) \leq \Pr_{E^{k_0}_{\max}}(B_l)$ or $\Pr_{E^{k_1}_{\max}}(B_j) \geq \Pr_{E^{k_0}_{\max}}(B_j) \geq \Pr_{E^{k_0}_{\max}}(B_l)$ which results in $t_1 \leq t_0$.

   Thus, we can choose the function $f(t) = t$ which fulfils $(k_0, t)$-closeness $\Rightarrow (k_1, f(t))$-closeness for $k_1 \geq k_0$.
2) Consider the sanitised dataset in Table IV where $\mu > 0$.

   This dataset satisfies $(k,t)$-closeness where $t = {(N-k+1)\mu}/{\mu} = N - k + 1 < \infty$ while $(k_0, t_0)$-closeness for $k_0 < k$ results in $t_0 = \infty$. Thus, there exist no function $f_k(t)$ that based on the values of $k$ and $t$ for a dataset satisfying $(k,t)$-closeness can compute a lower limit of $t_0$ for $(k_0, t)$-closeness with $k_0 < k$, that is, $(k,t)$-closeness $\not\Rightarrow (k_0, t_0)$-closeness for $k_0 < k$.
3) To show that $(k,t)$-closeness $\Rightarrow$ multiplicative $t_0$-closeness, consider a dataset that satisfies multiplicative $t$-closeness with a fixed partition of equivalence classes $\{E_1, \ldots, E_n\}$. For any equivalence class $E_j \in \{E_1, \ldots, E_n\} : |E_j| = k$ and any bucket $B_l$, $l \in [1:p]$, we have

$$\Pr_{E_j}(B_l) \leq \Pr_{E^k_{\max}}(B_l) \tag{7}$$

$$\Pr_{E_j}(B_l) \geq \Pr_{E^k_{\min}}(B_l) \tag{8}$$

TABLE V

THIS DATASET SATISFIES 1-CLOSENESS INDEPENDENTLY OF $\sigma$ WHILE THE VALUE OF $t$ FOR $(k, t)$-CLOSENESS IS DEPENDENT ON $\sigma$

| | Id | QI | Attribute probability |
|---|---|---|---|
| $E_1$ | 1 | $qi_1 \leq x_0$ | $\Pr(c_1 = a_1) = 2\mu - \sigma$ |
| | 2 | $qi_2 \leq x_0$ | $\Pr(c_2 = a_1) = \sigma$ |
| $E_2$ | 3 | $qi_3 > x_0$ | $\Pr(c_3 = a_1) = 2\mu - \sigma$ |
| | 4 | $qi_4 > x_0$ | $\Pr(c_4 = a_1) = \sigma$ |

due to the definitions of $E^k_{\max}$ and $E^k_{\min}$. For equivalence classes $E_j \in \{E_1, \ldots, E_n\} : |E_j| = k' \geq k$ we can use this result and the result in item (1a) in this proof to show that $\Pr_{E_j}(B) \leq \Pr_{E^{k'}_{\max}}(B) \leq \Pr_{E^k_{\max}}(B)$ and $\Pr_{E_j}(B) \geq \Pr_{E^{k'}_{\min}}(B) \geq \Pr_{E^k_{\min}}(B)$. According to Definition 7 we therefore must have $t_0 \leq t$. Thus, we can choose the function $f(t) = t$ which fulfils $(k, t)$-closeness $\Rightarrow$ multiplicative $f(t)$-closeness.

To show that multiplicative $t$-closeness $\not\Rightarrow$ $(k_0, t_0)$-closeness, first observe the dataset fulfilling multiplicative 1-closeness in Table V.

We select $k_0 = 2$ and assume that $\mu > \sigma$. Table V then satisfies $(k_0, t_0)$-closeness for $t_0 = \mu/\sigma$ but $\sigma$ does not affect the value of $t$ for multiplicative $t$-closeness. Similar to the proof of Theorem 1 we can show that there exist no function $f_{k_0}(t)$ that for a dataset satisfying multiplicative $t$-closeness can compute a lower limit of $t_0$ for $(k_0, t_0)$-closeness.

4) The property follows directly from Definition 7. ■

### C. Proof of Theorem 2

*Proof:* (Theorem 2) Let $\mu = \frac{1}{N}\sum_{i\in[1:N]} \Pr(Z(c_i))$, for a bucket $B_l, l \in [1 : p]$. Starting with $k = 1$ in Definition 7 we can write the privacy guarantee for $(1, t)$-closeness as

$$\frac{\mu}{t} \leq \Pr(Z(c_i)) \leq t\mu \tag{9}$$

for all $i \in [1 : N]$. This gives

$$\Pr(Z(c_i)) \leq t\mu = t^2\frac{\mu}{t} \leq t^2 \Pr(Z(c_j))$$

for any $i, j \in [1 : N]$. The same can be shown to hold for all $l \in [1 : p]$. Thus, $(1, t)$-closeness $\Rightarrow$ local $\ln(t^2)$-DP since the function $f(t) = \ln(t^2)$ approximates a lower estimate of $\epsilon$ in local $\epsilon$-DP for a dataset that satisfies $(1, t)$-closeness.

Starting instead with Definition 6 for local $\epsilon$-DP we have $\Pr(Z(c_i)) \leq e^\epsilon \Pr(Z(c_j))$ for all $i, j \in [1 : N]$. Thus

$$\mu = \frac{1}{N}\sum_{i\in[1:N]} \Pr(Z(c_i)) \leq \frac{1}{N}(e^\epsilon(N-1)\Pr(Z(c_j)) + \Pr(Z(c_j))) = \frac{\Pr(Z(c_j))}{N}(e^\epsilon(N-1)+1) \tag{10}$$

for any $j \in [1 : N]$. Similarly, by shuffling the local $\epsilon$-DP requirement we get $e^{-\epsilon}\Pr(Z(c_i)) \leq \Pr(Z(c_j))$ and

$$\mu = \frac{1}{N}\sum_{i\in[1:N]} \Pr(Z(c_i)) \geq \frac{1}{N}(e^{-\epsilon}(N-1)\Pr(Z(c_j)) + \Pr(Z(c_j))) = \frac{\Pr(Z(c_j))}{N}(e^{-\epsilon}(N-1)+1) \tag{11}$$

TABLE VI

THIS DATASET SATISFIES $(2, 2)$-CLOSENESS INDEPENDENTLY OF $\sigma$ WHILE THE VALUE OF $\epsilon$ IN LOCAL $\epsilon$-DP IS DEPENDENT ON $\sigma$

| Id | Attribute probability |
|---|---|
| 1 | $\Pr(c_1 = a_1) = \mu$ |
| 2 | $\Pr(c_2 = a_1) = 2\mu$ |
| 3 | $\Pr(c_3 = a_1) = \mu - \sigma$ |
| 4 | $\Pr(c_4 = a_1) = \sigma$ |

for any $j \in [1 : N]$. Equations (9), (10), and (11) now reveal that we can choose

$$f(\epsilon) = \max\left(\frac{1}{N}(e^\epsilon(N-1)+1), \frac{N}{(e^{-\epsilon}(N-1)+1)}\right) = \frac{1}{N}(e^\epsilon(N-1)+1)$$

since the first term is always greater than the second, and hence, local $\epsilon$-DP$\Rightarrow$ $(1, f(\epsilon))$-closeness. We have thereby shown that local $\epsilon$-DP $\Leftrightarrow$ $(1, t)$-closeness. ■

### D. Proof of Theorem 3

*Proof:* (Theorem 3) To prove that local $\epsilon$-DP $\Rightarrow$ $(k, t)$-closeness for $k > 1$ we can use all but the last step in the proof for local $\epsilon$-DP $\Rightarrow$ multiplicative $t$-closeness by Domingo-Ferrer and Soria-Comas [5]. Instead of assuming that there exist multiple equivalence classes in the last step we ensure that the closeness of the distributions ensured for multiplicative $t$-closeness has to hold for all subsets of size $k$, concluding that local $\epsilon$-DP $\Rightarrow$ $(k, t)$-closeness where

$$t = \frac{k}{N}\left(1 + \frac{N-k-1}{k}e^\epsilon\right).$$

On the other hand, we can use the example in Table VI to show that $(k, t)$-closeness $\not\Rightarrow$ local $\epsilon$-DP.

Assuming that $\sigma < \mu$ and selecting $k = 2$ we can compute that Table VI satisfies $(2, 2)$-closeness since $t = \frac{(\mu+2\mu+\mu-\sigma+\sigma)/4}{(\mu-\sigma+\sigma)/2} = 2$ independently of $\sigma$ and $\mu$. However, noting that $\epsilon$ in local $\epsilon$-DP is determined by the multiplicative difference between the largest and smallest attribute probability we get $\epsilon = \ln(2\mu/\sigma)$, that is, $\epsilon$ is dependent of $\epsilon$ and $\mu$. With similar arguments to the ones in Section A we can show that there exist no function $f(t)$ that can estimate a lower bound for $\epsilon$ in local $\epsilon$-DP and hence $(k, t)$-closeness $\not\Rightarrow$ local $\epsilon$-DP for $k > 1$. ■

### E. Proof of Lemma 3

*Proof:* (Lemma 3) We prove each point listed in Lemma 3 separately.

1) This lemma follows from point (1) in Lemma 2 by first observing that $(k_0, t)$-DP $\Leftrightarrow$ $(N - k_0, t)$-closeness and $(k_1, \epsilon)$-DP $\Leftrightarrow$ $(N - k_1, t)$-closeness and secondly that $k_1 \geq k_0$ entails $N - k_1 \leq N - k_0$.
2) Analogously to point (1), this lemma follows from point (2) in Lemma 2 by first observing that $(k_0, t)$-DP $\Leftrightarrow$ $(N - k_0, t)$-closeness and $(k_1, \epsilon)$-DP $\Leftrightarrow$ $(N - k_1, t)$-closeness and secondly that $k_1 < k_0$ entails $N - k_1 > N - k_0$.

TABLE VII
THIS DATASET SATISFIES $\epsilon$-DP WHERE $\epsilon = \max(N/N-1\,,\,(N-1)k/(k-1)N)$ INDEPENDENTLY OF THE VALUE OF $\alpha$ WHILE IT SATISFIES $(k, \epsilon_k)$-DP WITH $\epsilon_k = (N-k)\mu k/N\alpha$ WHICH IS DEPENDENT OF $\alpha$

| Id | Attribute probability |
|---|---|
| 1 | $\Pr(c_1 = a_1) = \mu$ |
| ... | ... |
| k-1 | $\Pr(c_{k-1} = a_1) = \mu$ |
| k | $\Pr(c_k = a_1) = \mu - \alpha$ |
| k+1 | $\Pr(c_{k+1} = a_1) = \alpha$ |
| k+2 | $\Pr(c_{k+2} = a_1) = 0$ |
| ... | ... |
| N | $\Pr(c_N = a_1) = 0$ |

3) This is a trivial case since setting $k = 1$ in Definition 8 for $(k, \epsilon_k)$-DP gives Definition 5 for $\epsilon$-DP. Thus the function $f(\epsilon) = \epsilon$ can estimate a lower limit in both $(1, \epsilon_k)$-DP $\Rightarrow$ $f(\epsilon_k)$-DP and $\epsilon$-DP $\Rightarrow$ $(1, f(\epsilon))$-DP. In addition, $f(f(\epsilon)) = \epsilon$ and thus $(1, \epsilon)$-DP $\Longleftrightarrow$ $\epsilon$-DP.

4) $(k, \epsilon)$-DP $\Rightarrow$ $\epsilon$-DP follows as a special case of point (1) in Lemma 3.

To show that $\epsilon$-DP $\not\Rightarrow$ $(k, \epsilon_k)$-DP, consider the sanitised dataset in Table VII where $0 < \alpha \leq \mu$.

The dataset in Table VII satisfies $\epsilon$-DP and $(k, \epsilon_k)$-DP where

$$\epsilon = \max\left(\frac{N}{N-1}, \frac{(N-1)k}{(k-1)N}\right)$$
$$\epsilon_k = \frac{(N-k)\mu k}{N\alpha}$$

Since $\epsilon_k$ is inversely proportional to $\alpha$ while $\epsilon, k, N$ are independent of $\alpha$, knowing $\epsilon, k, N$ is not enough to put an lower bound on the value of $\epsilon_k$. Thus, there exists no function $f_k(\epsilon)$ that can compute a lower limit for the privacy level of $(k, \epsilon_k)$-DP given a dataset satisfying $\epsilon$-DP since we can always choose a value of $\sigma$ so that $f_k(\epsilon) \succ \epsilon_k$

5) The property follows directly from Definition 8. ∎

### F. Proof of Theorem 4

*Proof:* (Theorem 4) To show that $(k, t)$-closeness $\Longleftrightarrow$ $(N-k, \epsilon)$-DP we start by using more similar notation for the two privacy definitions. In Definition 8, we can fix $Q(D)$ to be a query that returns the distribution of the confidential attributes $c_i$ for the dataset $D$ so that $\Pr(Z(Q(D)) \in B_l) = \Pr_D(B_l)$, for all $l \in [1 : p]$, where the right hand side is defined in Definition 7.

We can now present the two privacy definitions as

$$e^{-\epsilon} \leq \frac{\Pr_D(B_l)}{\Pr_{D_{N-k_{\text{DP}}}}(B_l)} \leq e^{\epsilon}$$
$$t^{-1} \leq \frac{\Pr_D(B_l)}{\Pr_{D_{k_{\text{tClo}}}}(B_l)} \leq t$$

where $k_{\text{DP}}$ and $k_{\text{tClo}}$ are the separate $k$ parameters for the two privacy definitions.

The numerator in both expressions is independent of the choice of $k$ and we can select $k_{\text{DP}} = N - k_{\text{tClo}}$ to make the definitions identical for $t = e^{\epsilon}$, irrespective of the choice of $l \in [1 : p]$. We can thereby identify two functions $f(t) = \ln t$ and $g(\epsilon) = e^{\epsilon}$ that for all $t \in [1, \infty)$ and $\epsilon \in [0, \infty)$ fulfil $(k, t)$-closeness $\Rightarrow$ $(N-k, f(t))$-DP and $(k, \epsilon)$-DP $\Rightarrow$ $(N-k, g(\epsilon))$-closeness. In addition, $f$ and $g$ satisfy $f(g(\epsilon)) = \epsilon$ and $g(f(t)) = t$. Hence, $(k, t)$-closeness $\Longleftrightarrow$ $(k, \epsilon)$-DP.

To show that $(k_{\text{tClo}}, t)$-closeness $\not\Longleftrightarrow$ $(k_{\text{DP}}, \epsilon)$-DP when $k_{\text{DP}} > N - k_{\text{tClo}}$ we can use point (1) in Lemma 2 and point (1) in Theorem 4. Observing that $(k_{tClo}, t)$-closeness $\not\Longleftrightarrow$ $(N - k_{DP}, t)$-closeness $\Leftrightarrow$ $(k_{DP}, \ln t)$-DP proves the implication in point (2) in Theorem 4 since the function $f(t) = \ln(t)$ is able to give a lower bound estimation to $\epsilon$ for $(k_{\text{DP}}, \epsilon)$-DP given a dataset satisfying $(k_{\text{tClo}}, t)$-closeness.

The lack of implication in point (2) follows by reversing the previous observation and applying Lemma 2 again since $(k_{DP}, \epsilon)$-DP $\Leftrightarrow$ $(N - k_{DP}, e^{\epsilon})$-closeness $\not\Rightarrow$ $(k_{tClo}, e^{\epsilon})$-closeness.

Analogously, a proof for point (3) in Theorem 4 can be constructed using point (1) in Lemma 2 and point (1) in Theorem 4. ∎

### G. Proof of Theorem 5

*Proof:* (Theorem 5) Consider the two equivalence classes in Table III. Assume that individuals' quasi-identifiers are known so that any individual can be linked to an equivalence class. By using Definition 8 for $(k_0, \epsilon)$-DP, applying it to each equivalence class with $k_0 = 1$, and noting that $\Pr(Z(c_3) \in \{a_1\}) = \sigma$ and $\Pr(Z([c_3, c_4]) \in \{a_1\}) = \mu$ we compute that the dataset in Table III satisfies $(1, \mu/\sigma)$-DP. The dataset in Table III also satisfies multiplicative 1-closeness.

With the same arguments as in the proof of Theorem 1, Section A, we can show that there exists no function $f_{k_0}(t)$ such that multiplicative $t$-closeness $\Rightarrow$ $(k_0, f_{k_0}(t))$-DP. Therefore multiplicative $t$-closeness $\not\Rightarrow$ $(k, \epsilon)$-DP. ∎

## REFERENCES

[1] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *Proc. IEEE 29th Int. Conf. Data Eng. Workshops (ICDEW)*. Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2013, pp. 88–93.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Conf. Theory Cryptography*. Berlin, Germany: Springer-Verlag, 2006, pp. 265–284.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. 23rd Int. Conf. Data Eng. (ICDE)*. Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2007, pp. 106–115.

[4] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," in *Proc. 7th ACM Symp. Inf., Comput. Commun. Secur.*, New York, NY, USA: ACM, 2012, pp. 32–33.

[5] J. Domingo-Ferrer and J. Soria-Comas, "From t-closeness to differential privacy and vice versa in data anonymization," *Knowl.-Based Syst.*, vol. 74, pp. 151–158, Jan. 2015.

[6] S. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Stat. Assoc.*, vol. 60, no. 309, pp. 6–63, 1965.

[7] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[8] L. Sweeney, "Datafly: A system for providing anonymity in medical data," in *Database Security XI: Status and Prospects*, T. Y. Lin and S. Qian, Eds. Boston, MA, USA: Springer, 1998, pp. 356–381.

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA: ACM, 2005, pp. 49–60.

[10] K. E. Emam *et al.*, "A globally optimal k-anonymity method for the de-identification of health data," *J. Amer. Med. Informat. Assoc.*, vol. 16, pp. 82–670, Sep. 2009.

[11] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," *Data Mining Knowl. Discovery*, vol. 11, no. 2, pp. 195–212, Sep. 2005.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *22nd Int. Conf. Data Eng. (ICDE)*. Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2006, p. 24.

[13] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2016, pp. 1464–1465.

[14] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125.

[15] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proc. 31st Int. Conf. Very Large Data Bases*, Trondheim, Norway. New York, NY, USA: ACM, Aug. 2005, pp. 901–909.

[16] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Conf. Automata, Lang. Program.* Berlin, Germany: Springer-Verlag, 2006, pp. 1–12.

[17] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. 24th Annu. Int. Conf. Theory Appl. Cryptograph. Techn.* Berlin, Germany: Springer-Verlag, 2006, pp. 486–503.

[18] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2018, pp. 394–403.

[19] F. Liu, "Generalized Gaussian mechanism for differential privacy," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 747–756, Apr. 2016.

[20] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2007, pp. 94–103.

[21] J. Zhao, J. Zhang, and H. V. Poor, "Dependent differential privacy for correlated data," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Singapore. New York, NY, USA: Institute of Electrical and Electronics Engineers, Dec. 2017, pp. 1–7.

[22] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *Proc. Netw. Distrib. Syst. Secur. Symp.* Reston, VA, USA: The Internet Society, Feb. 2016, pp. 21–24.

[23] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA: ACM, May 2015, pp. 747–762.

[24] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA: ACM, Jun. 2014, pp. 1447–1458.

[25] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *Proc. 31st Symp. Princ. Database Syst. (PODS)*. New York, NY, USA: ACM, 2012, pp. 77–88.

[26] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. Int. Conf. Manage. Data (SIGMOD)*. New York, NY, USA: ACM, 2011, pp. 193–204.

[27] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," in *Proc. 3rd Int. Conf. Availability, Rel. Secur.* Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2008, pp. 990–993.

[28] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, 2008, pp. 70–78.

[29] J. Lee and C. Clifton, "How much is enough? Choosing $\epsilon$ for differential privacy," in *Proc. 14th Int. Conf. Inf. Secur.* Berlin, Germany: Springer-Verlag, 2011, pp. 325–340.

[30] J. Soria-Comas and J. Domingo-Ferrert, "Differential privacy via t-closeness in data publishing," in *Proc. IEEE 11th Annu. Conf. Privacy, Secur. Trust*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2013, pp. 27–35.

[31] F. Bacchus, A. J. Grove, D. Koller, and J. Y. Halpern, "From statistics to beliefs," in *Proc. 10th Nat. Conf. Artif. Intell.*, W. R. Swartout, Ed. San Jose, CA, USA: MIT Press, Jul. 1992, pp. 602–608.

[32] Y. Li, X. Cao, Y. Yuan, and G. Wang, "PrivSem: Protecting location privacy using semantic and differential privacy," *World Wide Web*, vol. 22, no. 6, pp. 2407–2436, 2019.

[33] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Monticello, IL, USA, Oct. 2013, p. 1592.

[34] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection," in *Proc. EDBT/ICDT Workshops*, 2016, pp. 1–12.

[35] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 492–542, 2014.

[36] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," 2014, *arXiv:1407.6981*.

[37] M. S. Alvim, M. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "Differential privacy: On the trade-off between utility and information leakage," in *Proc. Formal Aspects Secur. Trust*, 2011, pp. 39–54.