A Probabilistic Approach to Mitigate Composition Attacks on Privacy in Non-Coordinated Environments

A.H.M. Sarowar Sattar[a,*], Jiuyong Li[a,*], Jixue Liu[a], Raymond Heatherly[b], Bradley Malin[b,c]

[a]*School of Information Technology and Mathematical Science, University of South Australia, Mawson Lakes, SA-5095, Australia*
[b]*Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, U.S.*
[c]*Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, U.S.*

Abstract

Organizations share data about individuals to drive business and comply with law and regulation. However, an adversary may expose confidential information by tracking an individual across disparate data publications using quasi-identifying attributes (e.g., age, geocode and sex) associated with the records. Various studies have shown that well-established privacy protection models (e.g., k-anonymity and its extensions) fail to protect an individual's privacy against this "composition attack". This type of attack can be thwarted when organizations coordinate prior to data publication, but such a practice is not always feasible. In this paper, we introduce a probabilistic model called (d, α)-linkable, which mitigates composition attack without coordination. The model ensures that d confidential values are associated with a quasi-identifying group with a likelihood of α. We realize this model through an efficient extension to k-anonymization and use extensive experiments to show our strategy significantly reduces the likelihood of a successful composition attack and can preserve more utility than alternative privacy models, such as differential privacy.

Keywords: Databases, Data publication, Privacy, Composition attack, Anonymization

[*] Corresponding author
Email addresses: `satay003@mymail.unisa.edu.au` (A.H.M. Sarowar Sattar), `Jiuyong.Li@unisa.edu.au` (Jiuyong Li)

[1] School of Information Technology and Mathematical Sciences Mawson Lakes, SA-5095. Mob. +61420863356

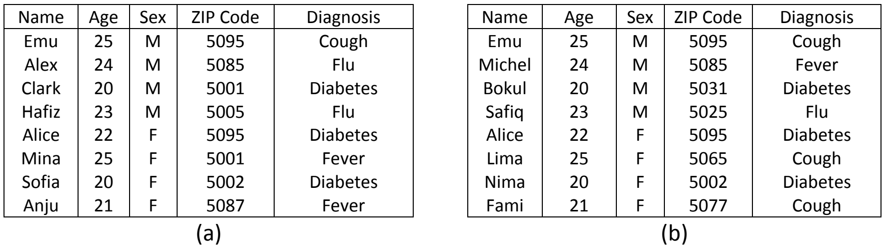
[2] D3-07, School of Information Technology and Mathematical Science, Mawson Lakes, SA-5095.

1. Introduction

The increasing collection of large quantities of person-specific information has created tremendous opportunities for knowledge-based decision making in a number of domains [24]. To fully maximize the knowledge that can be learned, the data needs to be made available beyond the organizations that performed the initial collection [13]. Data sharing, however, must be accomplished in a manner that respects the privacy of the individuals from which the data was gathered [31]. There are a wide variety of computational protection models that have been suggested [7], but, the majority fail to consider situations in which records on the same individual occur in multiple organizations' data sets. This is a concern because, in certain environments, an individual's information will be collected and published by disparate organizations [20]. And when such a situation arises, an adversary may invoke a *composition attack* [9] on the published data sets to compromise the privacy afforded by traditional protection models.

The composition attack will be formalized below, but here, we take a moment to illustrate how such a problem transpires to provide context. Imagine two healthcare organizations, Organization-A and Organization-B, collect demographics and confidential health information as shown in Tables 1(a) and 1(b), respectively. Notice that Alice, a 22 year-old female living in ZIP code 5095, was diagnosed with 'Diabetes' at both organizations. The organizations choose to publish versions of their data sets as depicted in Tables 2(a) and 2(b). These adhere to a traditional formal privacy model called k-anonymity, which classifies attributes as explicit identifiers (that is, information that allows for direct communication with an individual such as *name* and *Social Security Number*), quasi-identifiers (that is, in combination, can uniquely characterize an individual and be leveraged for identification purposes, such as *age*, *sex*, and *ZIP code* of residence) and confidential attributes (e.g., diagnoses). In a published data set, explicit identifiers are suppressed, quasi-identifiers are masked, and confidential attributes are retained in their original form. To mask quasi-identifiers, their values are often generalized to less specific concepts. Alice's information, for instance, has been generalized to an age range of 15-25 and a ZIP code of 50** in one table and 10-30 and 50** in the other table. Yet, when an adversary knows that Alice visited both institutions, they may learn her health status because there is only one common confidential value in the sets of records that could possibly correspond to Alice.

The *composition attack* can be thwarted when organizations coordinate during the k-anonymization process. Specifically, such coordination can take place by sharing their data sets in the clear (e.g., [18]) or computing over encrypted transformations (e.g., [11, 19]) prior to publication to discover and address potential violations. However, such coordination is not always possible and may even be



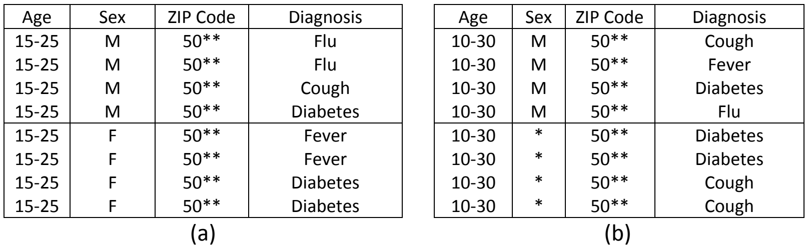
Name	Age	Sex	ZIP Code	Diagnosis
Emu	25	M	5095	Cough
Alex	24	M	5085	Flu
Clark	20	M	5001	Diabetes
Hafiz	23	M	5005	Flu
Alice	22	F	5095	Diabetes
Mina	25	F	5001	Fever
Sofia	20	F	5002	Diabetes
Anju	21	F	5087	Fever

(a)

Name	Age	Sex	ZIP Code	Diagnosis
Emu	25	M	5095	Cough
Michel	24	M	5085	Fever
Bokul	20	M	5031	Diabetes
Safiq	23	M	5025	Flu
Alice	22	F	5095	Diabetes
Lima	25	F	5065	Cough
Nima	20	F	5002	Diabetes
Fami	21	F	5077	Cough

(b)

Table 1: The data managed by (a) Organization-A (b) Organization-B in their private collections.



Age	Sex	ZIP Code	Diagnosis
15-25	M	50**	Flu
15-25	M	50**	Flu
15-25	M	50**	Cough
15-25	M	50**	Diabetes
15-25	F	50**	Fever
15-25	F	50**	Fever
15-25	F	50**	Diabetes
15-25	F	50**	Diabetes

(a)

Age	Sex	ZIP Code	Diagnosis
10-30	M	50**	Cough
10-30	M	50**	Fever
10-30	M	50**	Diabetes
10-30	M	50**	Flu
10-30	*	50**	Diabetes
10-30	*	50**	Diabetes
10-30	*	50**	Cough
10-30	*	50**	Cough

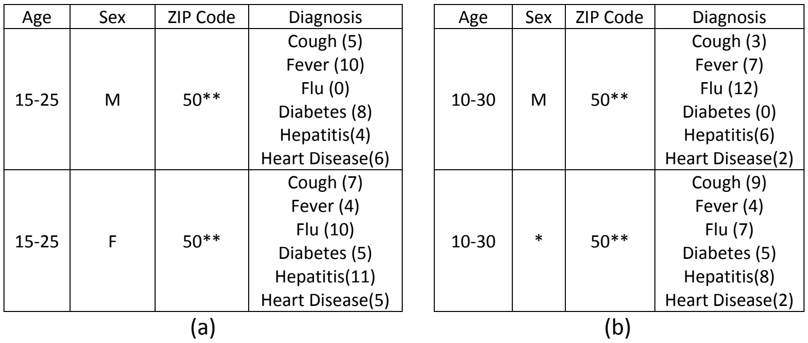
(b)

Table 2: Publications of data sets from (a) Organization-A (b) Organization-B after the application of k-anonymization.

prohibited by law [2]. Moreover, in some countries, such as the United States, healthcare is decentralized. As a result, it is not uncommon for a patient to be seen at multiple hospitals that do not coordinate with one another as discussed in [20]. We refer to this setting as a *non-coordinated environment*.

In non-coordinated environments, privacy enhancing methods based on randomization can be applied to limit the detection of an individual in a data set. In particular, differential privacy (which is discussed in further detail in the following section), can prevent the composition attack [22]. However, the utility of such data sets may be too low for practical [22]. Consider Tables 2 and 3 depict examples of k-anonymized and differentially private data sets, respectively[3]. The composition attack is successful for the pair of k-anonymized data sets because an adversary can restrict their focus to only one confidential value, namely ‘Diabetes’, to link with Alice’s record. In contrast, for the pair of differentially private data sets, the adversary has all values from the confidential attribute’s domain to link with Alice’s

[3] The hypothesized tables in Table 3 have been created by following the differential privacy mechanism in [22]



Age	Sex	ZIP Code	Diagnosis
15-25	M	50**	Cough (5) Fever (10) Flu (0) Diabetes (8) Hepatitis(4) Heart Disease(6)
15-25	F	50**	Cough (7) Fever (4) Flu (10) Diabetes (5) Hepatitis(11) Heart Disease(5)

(a)

Age	Sex	ZIP Code	Diagnosis
10-30	M	50**	Cough (3) Fever (7) Flu (12) Diabetes (0) Hepatitis(6) Heart Disease(2)
10-30	*	50**	Cough (9) Fever (4) Flu (7) Diabetes (5) Hepatitis(8) Heart Disease(2)

(b)

Table 3: Published data sets from (a) Organization-A (b) Organization-B after the application of a differential privacy mechanism

record.[4]

Thus, the goal of our current work is to develop generalization-based strategies to protect individuals' privacy whose records are disclosed by disparate organizations when coordination is not permitted. We propose a protection model that is designed to increase the likelihood that an adversary will have multiple confidential values to link with an individual's record after combining disparate k-anonymized data sets. Specifically, the contributions of this paper are as follows.

- First, we propose a novel model to reduce the risk of the composition attack. Our model is applicable to each publisher's data set independently and without coordination. This model uses statistical information regarding the quasi-identifying and confidential attributes of the underlying population to simulate a k-anonymized data set published by another organization.

- Second, we design an efficient algorithm to achieve the proposed protection model. The algorithm is implemented as a post-processing method applied to partition-based k-anonymization [16, 17, 30] approaches[5]. Note that in the publications from different independent organizations, the privacy of the records is preserved by the k-anonymity model. Thus, when applying the post-processing method on top of k-anonymization, we retain this privacy

[4] In this example there are six confidential values {'Cough', 'Diabetes', 'Flu', 'Fever', 'Hepatitis', 'Heart Disease'} in the confidential attribute's domain.

[5] We acknowledge that even though there are a plethora of k-anonymity methods based on generalization, there are other methods like microaggregation, that replace equivalence classes by averaged values [5, 21, 4, 1]. However, in this paper, we only consider generalization because our algorithm is implemented as a post-processing step for partition-based anonymization.

guarantee.

- Third, we provide an extensive empirical evaluation of our method on publicly avaialble data sets from the U.S. Census Bureau. We compare our method with a strategy based upon differential privacy [6] and show that our method can preserve better utility, with a negligible effect on data quality.

The remainder of this paper is organized as follows. Section 2 provides a more detailed background on the composition attack and various models to protect data for publication. Section 3 formalizes the underlying concepts and the composition problem. Section 4 provides a theoretical foundation for privacy preservation and a model of protection for the situation in which two organizations publish data independently. Moreover, Section 4 presents an extension of our model to a more general case in which more than two organizations publish data. Section 5 introduces a computational method to achieve the model. Section 6 provides a series of empirical investigations to demonstrate the tradeoff in attack mitigation and data utility after applying the proposed approach and a differentially private publication strategy along with discussion on the limitations of our proposed model. Finally, Section 7 provides discussion on conclusions associated with this work.

2. Background

2.1. Composition attacks are challenging problems

In the composition attack, an adversary follows the confidential attribute to learn information about an individual's record (who we will refer to as the victim). The example offered in the introduction highlights this issue. This process is an extension to *linking*, or the identification of an individual's record in published data sets through the matching of their quasi-identifying attributes [30, 33]. This definition has traditionally been applied to single instance publications, which do not consider the composition attack. Consequentially, the principles in single instance publication settings are not applicable to multiple independent publications [9]. In this work, we call the property by which it is possible to infer an individual's confidential information by linking such attributes across multiple independent publications as *linkability*.

2.2. Existing solutions for composition attacks

Privacy-preserving data publication techniques can be broadly classified into two categories; *partition* and *randomization*. With a *partition* technique, the data values of some quasi-identifying attributes (e.g., age, sex and residential address) are generalized to form small groups, so that an individual cannot be identified and their confidential value(s) cannot be inferred with a high confidence. By contrast, in a *randomization* technique, the original values have noise added to them and, hence, it is difficult to pinpoint an individual in a published data set. While a substantial quantity of privacy-preserving data publication models and algorithms have been developed over the past decade [7], they do not appropriately address the non-coordinated composition problem. To provide context for our work and how it relates to the existing literature, we review several of the more relevant techniques and analyze why they are unsuitable to cope with the composition attack.

As mentioned earlier, a partition-based protection technique deals mainly with the single instance of publication [7]. Examples of such strategies include k-anonymity [30], l-diversity [17] and t-closeness [16]. However, these strategies do not appropriately address the composition attack [9]. This is because, when two publishers disclose data sets to satisfy a privacy criterion independently, there is no guarantee that the combination of the data sets continue to satisfy the criterion. Another technique, p-sensitive k-anonymity [29] proposes a way to unlink quasi-identifying and confidential attributes to protect the confidential information of individuals. The unlink-ability is only for one data set. This technique suffers for the same problem as other single instance of publications. When two or more data sets are published by the technique, the intersection can contain a unique confidential value and privacy is breached.

It should be recognized that some partition-based techniques deal with serial publication [8, 32, 34, 36] (i.e., multiple publications from the same data publisher). Yet, prior work assumes the published data sets are all controlled by a publisher who is aware of all versions of previously published data. Moreover, the data publisher can modify the data to ensure the combination with previous publications retain the privacy condition. In our setting, the data publisher does not have knowledge about other data sets that may be exploited for composition.

Alternatively, [22] showed that the randomization framework of ϵ-differential in a non-interactive setting (i.e., a one-time publication) can protect data from the composition attack. Yet, the utility of published information is often low when ϵ-differential privacy is used for data publication in a non-interactive manner. This is because differential privacy as presented in [22], is designed to mask both quasi-identifying and confidential attributes. Masking for quasi-identifying attributes is achieved by generalization, while masking for the confidential attribute is achieved by publishing noisy count of all confidential values for different equivalence classes (Definition 1) of people (an example in Table 3). Unfortunately, not all equivalence classes have records that cover all confidential values. As such, when the count of a confidential value within an equivalence is small in the original data set, say zero, it may be perturbed into any integer. And this is problematic when analytics over the published data require the complete absence of information to classify an individual, such as is the case in clinical phenotyping [23]. Moreover, this problem can occur for every equivalence class of individuals. The low utility of differential privacy in a non-interactive setting is also reported in [28]. This issue is further highlighted in our empirical analysis in Section 6.2.

Beyond the above discussed techniques, the composition attack can be resolved when organizations collaborate. For example, after k-anonymization is performed locally, the organizations can check which confidential values (e.g., diseases) are common between different equivalence classes of data sets and take actions to enforce a privacy requirement. Certain methods [11, 18, 19] solve this problem by suppressing the overlapping records from one (or more) of the data sets. Such methods do not apply in our setting because coordination is not permitted and the protection models invoked in such protocols do not explicitly address the linkability of the confidential values.

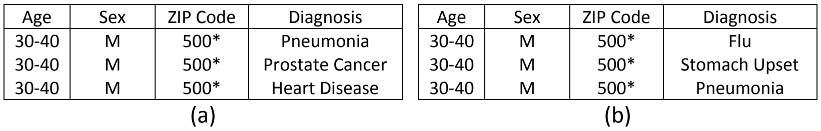
2.3. Fundamental causes and solution outline

Linkability enables attackers to narrow down the search space of a victim's confidential values (to be formally defined in Section 3). For example, independently of the two k-anonymized data sets in Tables 2(a) and 2(b), an attacker has two distinct confidential values to link with Alice's record. When combining both data sets, there is only one confidential value $\{$Diabetes$\}$ in common with the quasi-identifier attributes. Therefore, when the attacker knows that Alice visited both organizations, they can infer that Alice is suffering from $\{$Diabetes$\}$ with 100% confidence.

The number of confidential values shared across multiple k-anonymized data sets in an individual's equivalence classes determines the level of linkability. Let d represent the value of such common confidential values, then the linkability of those anonymous data set is d and one anonymous data set is d-linkable with another anonymous data set. In the previous example, the data set in Table 2(a) is 1-linkable with data set 2(b) and the privacy of an individual is clearly compromised in 1-linkable data sets. When the number of the shared confidential values increases, the risk of an individual's privacy being compromised decreases (and vice versa).

Without the knowledge of a linking data set, a data set from another publisher that can be used by an attacker to conduct composition attack, it is clearly difficult for one organization to k-anonymize its data set to eradicate the composition attack. Yet, based on the probability distributions of different confidential values, some combinations have a lower risk than others in the context of the composition attack.

For example, two variations of k-anonymized data sets from Organization-C are shown in Tables 4(a) and 4(b). Intuitively, the data set in Table 4(a) has a



Age	Sex	ZIP Code	Diagnosis
30-40	M	500*	Pneumonia
30-40	M	500*	Prostate Cancer
30-40	M	500*	Heart Disease

(a)

Age	Sex	ZIP Code	Diagnosis
30-40	M	500*	Flu
30-40	M	500*	Stomach Upset
30-40	M	500*	Pneumonia

(b)

Table 4: Publications of data sets from Organization-C after the application of k-anonymization.

higher risk than the data set in Table 4(b). The chance for an adversary to see a pair of confidential values of Table 4(a) in another patient group (equivalence class) of another data set is higher than that of Table 4(b). In other words, the confidential values in Table 4(a) have a higher chance being unique when it is intersected with another group (equivalence class) of another data set than those in Table 4 (b). As such, if there is a confidential value in common between the data set in Table 4(a) and a linking data set, it is more likely caused by the victim who visited both organizations, than by two different records with the same quasi-identifying value and diagnosis.

For the composition attack, the risk is associated with the frequencies of confidential values. So, without knowledge of a linking data set, we should be able to reduce the probability of the attack. And, based on such an estimate, we can generalize a data set in a way that the chance for the data set in question and any linking data set will share two or more common confidential values is sufficiently high. Consider in the data set in Table 4(b), if the chance of any two data sets (i.e., the data set in Table 4(b) and a linking data set) sharing any pair of $\{$Flu, Stomach Upset$\}$, $\{$Flu, Pneumonia$\}$ or $\{$Stomach Upset, Pneumonia$\}$ confidential values is high, the probability of 1-linkable (i.e., two data sets sharing 1 confidential value) is low. As a consequence, the risk of privacy compromise by the composition attack is low as well.

Thus, in this paper, we propose a (d, α)-linkable privacy model. Informally, this model requires that the probability a k-anonymized data set and a linking data set share d confidential values in matching equivalence classes is α. And, it follows that the privacy risk of individuals with shared records is sufficiently low with high α.

3. Preliminaries

Here, we formalize the system. For reference, Table 5 summarizes the common notation used throughout this paper.

Let $D = \{t_1, t_2, \ldots, t_n\}$ be a multi-set of records, where each record t_i represents the information of an individual i. Each record is represented $t_i = \{id_i, q_i, s_i\}$, where $id_i \in ID$, $q_i \in QID$, and $s_i \in S$. ID represents the set of unique identifiers, which are used to uniquely identify records, such as personal name or medicare card number. QID is a set of quasi-identifying attributes that can potentially identify a person (e.g., age, ZIP code and sex) and S is a set of confidential values (e.g., disease). The quasi-identifying attributes $QID = \{q_1, q_2, \ldots, q_m\}$ consist of m attributes, each of which has its own domain that contains a set of possible values. Let $D^* = \{\hat{t_1}, \hat{t_2}, \ldots, \hat{t_n}\}$ be a published data set, where $\hat{t} = \{\hat{q_1}, \hat{q_2}, \ldots, \hat{q_m}, s\}$ and $\hat{q_i}$ is any value from the domain of q_i.

In a published data set D^*, the attribute ID has been removed, whereas the (modified) QID attributes and confidential attributes are kept in the published data sets. If the QID and the confidential values are published in their original state, an adversary may invoke record linkage [33] between QID attributes and external information to link an individual's identity to their confidential information. To avoid this disclosure, one frequently applied solution is to replace the QID values with more general values from their domains to ensure that the individuals in an equivalence class (Definition 1) are indistinguishable and their confidential values cannot be inferred with a high confidence [16, 17, 30, 35].

Definition 1 (Equivalence class). *For a k-anonymized data set, an equivalence class corresponds to the set of records in the data set with identical values over the combination of QID attributes.*

For example, records 1 to 4 in Table 2(a) form an equivalence class with respect to $\{$age, sex, ZIP code$\}$.

Let $D_1^*, D_2^*, \ldots D_n^*$ be the n independent k-anonymized data sets with minimum equivalence class size k. We use the notation E_i^j to represent the equivalence class of an individual i in a published data set D_j^*. Let $S(E_i^j)$ represent the set of (distinct) confidential values in an equivalence class.

In a published data set, the equivalence class size k represents the anonymity of an individual. This means an individual should be grouped with $(k-1)$ other individuals in the published data set. However, as pointed out in [17], the level of protection for an individual is equal to the number of distinct confidential values of the equivalence class in which the individual's record resides. Moreover, when there are multiple published data sets available (an individual may have record in the multiple published data sets) then her level of protection may change [9]. We refer to this as "composition anonymity" of an individual.

Therefore, the composition anonymity of an individual is defined by the Definition 2.

Definition 2 (Composition anonymity). *For an individual i, the composition anonymity in n independently anonymized data sets containing her record is equal to the number of distinct common confidential values in the equivalence classes where her*

Table 5: Common notation used in this paper.

Notation	Description
Ω	large population from which the records are collected
D, D_1, D_2	the original data sets
D^*, D_1^*, D_2^*	published data sets of D, D_1 and D_2, respectively
D_0	the hypothesized data set, $\|D_0\| = \|D_2^*\|$
$P(X)$	the probability that event X happens
t_i	record t of an individual i
$\hat{t_i}$	generalized record of record t
ID	the identifier attributes
QID	the quasi-identifying attributes
S	set of all confidential values
q_i	i^{th} QID attribute
$\hat{q_i}$	generalized value of q_i
s	the confidential attribute
S^d	the set of d different confidential values
E_i^j	the equivalence class of an individual i in a published data set D_j^*
$S(E_i^j)$	the set of (distinct) confidential values corresponding to E_i^j
o_i	composition anonymity of an individual i

record resides. Formally, the composition anonymity for i with respect to those anonymized data sets is

$$o_i = |\cap S(E_i^j)|, j = 1, \ldots, n$$

where E_i^j is an equivalence class in data set D_j^* containing quasi-identifier values of i.

The knowledge of an adversary A victim v is an individual in D_1 with $t = \{ID = v, QID, s\}$. The adversary knows the QID values of v and tries to infer s in the following scenario:

- v is also in another data set D_2. The adversary has access to D_1^* and D_2^*, the published data sets of D_1 and D_2, respectively.

The knowledge of a data publisher A publisher has no specific knowledge of another data set that may contain overlapping records with its published data set. However, a publisher anticipates that the values of QID and S of the another data set follow the same distribution and that both the data sets follow the same k-anonymization procedure.

When considering the above setting, the privacy breach is characterized by Definition 3.

Definition 3 (Privacy breach). *Given published data sets D_1^* and D_2^*, and the knowledge that a victim v is in both, a privacy breach occurs when the composition anonymity of an individual $i = v$ is less than value d (defined by the publishers). More specifically, the privacy breach occurs when $o_i < d$, where d represents a publisher's predefined protection parameter.*

For instance, imagine that $d = 2$. Then, Alice's privacy is breached in the data sets of Tables 2(a) and 2(b), where $o_{Alice} < 2$. Specifically, an adversary can identify that Alice is suffering from AIDS. This is because AIDS is the only common confidential value in the sets of records that could correspond to Alice. Therefore, the objective of the following model is to minimize the privacy breach that occurs for overlapping records.

We now formalize *linkability*, the main property of a published data set that makes the composition attack possible.

Definition 4 (Linkability). *Linkability is a property of a k-anonymized data set that offers an adversary the ability to identify the group of possible confidential values of an individual by following the confidential attribute without precisely identifying her record. Therefore, $linkability \propto o_i$, where o_i is the composition anonymity of an individual in n independent k-anonymized data sets.*

In Tables 2(a) and 2(b), the *linkability* reveals the confidential value which belongs to Alice without precisely identifying her record in those data sets.

In a partition-based protection technique, since the generalized values are faithful to their original values, it is possible to locate the equivalence class of a record; as such, *linkability* persists. However, we can control the confidence an adversary has in linking an individual's confidential information by increasing the number of distinct shared confidential values in that individual's equivalence class when combining different k-anonymized data sets. When a data set is k-anonymized in such a way that composition anonymity of an individual can have a pre-defined threshold d, we say that data set d-linkable (Definition 5). Therefore, the d-linkable property can be formalized by the following definition.

Definition 5 (d-linkable). *k-anonymized data sets D_1^* and D_2^* are called d-linkable if, for each individual $i = v$: $o_i \geq d$, where $d \geq 2$.*

4. Privacy model (d, α)-linkable

Our goal is to publish a data set that is d-linkable with any another data set. In this section, we provide a high-level characterization of the (d, α)-linkable model along with a statistical basis on how to achieve it.

Imagine that the original data sets D_1 and D_2 are samples from a large population Ω and that the intersection of the data sets is non-null. D_1^* and D_2^* are the corresponding k-anonymized versions of the original data sets. A data set D_0^* is a hypothesized data set of D_2^* (the generation process for which is explained below). We assume both data sets have the same attribute domain and size. Based on these assumptions, the (d, α)-linkable privacy model is defined as follows.

Definition 6 ((d, α)-linkable). *A k-anonymized data set D_1^* is (d, α)-linkable with another k-anonymized data set D_2^* if d distinct confidential values appear in common for each individual's equivalence classes with α confidence.*

When D_1^* is (d, α)-linkable (see Definition 5) with another data set D_2^*, there are d distinct confidential values in common with each corresponding equivalence class (that is, the equivalence classes that should belongs to the same individual in both data sets) of D_1^* and D_2^*. This reduces the chance of a successful composition attack into an expected confidence bound (α).

We now provide a model for how to estimate the linkability of D_1^* with D_2^*. Imagine that the example data set D_0^* is a random sample of Ω with record probability (Definition 7 [27]) $P(\hat{t})$, where $\hat{t}$ is a record with confidential value s.

Definition 7 (Record probability). *We assume that the attribute values and the confidential value in a record are independent.*[6] *$P(q_i)$ and $P(s)$ are the frequencies of value q_i and confidential value s in the population. The probability of a record $\hat{t} = \{\hat{q}_1, \hat{q}_2 \ldots, \hat{q}_m, s\}$, denoted as $P(\hat{t})$, can then be assigned as follows.*

$$\begin{aligned} P(\hat{t}) &= P(\hat{q}_1) \times P(\hat{q}_2) \times \cdots \times P(\hat{q}_m) \times P(s) \\ &= (\prod_{i=1}^{m} P(\hat{q}_i)) \times P(s) \end{aligned} \tag{1}$$

For example, let us assume that $P(\text{age = [20-30]}) = 0.15$, $P(\text{gender = [female]}) = 0.5$, and $P(\text{disease = [diabetes]}) = 0.05$ are obtained from the patient population. Let $\hat{t} = \{20 - 30, \text{female}, \text{diabetes}\}$, then $P(\hat{t}) = 0.00375$.

[6] The independence assumption is invoked when additional knowledge is unavailable. When dependencies are known, their relationship can be modeled by other data mining frameworks. For example, the confidence of an association rule $(Age[40 - 60], M) \rightarrow Prostate\ Cancer$, can be used to model the probability of an equivalence class of people to a disease.

Linkability for D_1^* exists when D_0^* contains a record from the same individual with the same confidential value. Therefore, to check the linkability of D_1^* with D_0^*, we need to determine whether a record $\hat{t}$ is in the hypothesized data set D_0^* by chance alone.

Building on the definition of record probability (Definition 7), $P(\hat{t})$ and $1 - P(\hat{t})$ represent the probability of success and failure, respectively, of selecting a record $\hat{t}$ by a random draw. We can consider each draw as a Bernoulli trial and the process of generation of the data set D_0^* can be considered as n random draws with replacement that follow the binomial distribution: $f(\sigma, n, \rho)$, where σ is the exact number of successes for selecting a record $\hat{t}$ from a total of n random draws and ρ is the probability of a success of that event. Therefore, $\rho = P(\hat{t}) = (\prod_{i=1}^{m} P(q_i)) \times P(s)$.

Consequently, if $P(\hat{t}, s)$ represents the probability of having at least one $\hat{t}$ with confidential value s in D_0^*, then,

$$\begin{aligned} P(\hat{t}, s) &= 1 - f(0, n, \rho) \\ &= 1 - (1 - \rho)^n \end{aligned} \qquad (2)$$

In the above equation, since $\hat{t} = \{\hat{q_1}, \hat{q_2}, \ldots \hat{q_m}, s\}$ is a generalized version of t, its generalized QID can be linked with an equivalence class $E_i^1 = \{\hat{q_1}, \hat{q_2}, \ldots \hat{q_m}\}$ in D_1^*. Therefore, the probability $P(\hat{t}, s)$ also represents the probability of a confidential value s appearing in an equivalence class E_i^0 in D_0^*. Thus,

$$P(\hat{t}, s) = P(E_i^0, s) = 1 - (1 - \rho)^n \qquad (3)$$

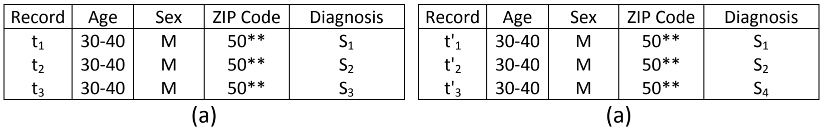
The above equation represents the probability that a confidential value s, already in E_i^1 of D_1^*, will be in E_i^0 of D_0^* by chance alone. As such, this probability also represents the chance that a confidential value s will be in common between equivalence classes (E_i^1 and E_i^0) with probability $P(E_i^0, s)$. Our goal is to determine such a probability for d distinct confidential values.

Now, let $P(E_i^0, S^d)$ represent the probability of d different confidential values appearing together in the equivalence class E_i^0 in D_0^* just by chance. Then,

$$\begin{aligned} P(E_i^0, S^d) &= P(E_i^0, s_1) \times P(E_i^0, s_2) \times \cdots \times P(E_i^0, s_d) \\ &= \prod_{r=1}^{d} P(E_i^0, s_r). \end{aligned} \qquad (4)$$

Since we have assumed that the equivalence class E_i^0 is already in D_1^*, here we only consider those s in S^d that are already in E_i^1. Therefore,

$$S^d = \{s_1, s_2, \ldots, s_d\} \subset S(E_i^1)$$



Record	Age	Sex	ZIP Code	Diagnosis
t_1	30-40	M	50**	S_1
t_2	30-40	M	50**	S_2
t_3	30-40	M	50**	S_3

(a)

Record	Age	Sex	ZIP Code	Diagnosis
t'_1	30-40	M	50**	S_1
t'_2	30-40	M	50**	S_2
t'_3	30-40	M	50**	S_4

(a)

Table 6: (d, α)-linkable data sets (a) D_1^* and (b) D_2^*

.

While $P(E_i^0, S^d)$ represents the probability of d distinct confidential values from E_i^1 appearing in E_i^0, it also represents the probability that d distinct confidential values will be in both equivalence classes E_i^0 and E_i^1 with confidence $P(E_i^0, S^d)$.

Therefore, after the publication of D_0^*, the composition anonymity of an individual i will be d with confidence $P(E_i^0, S^d)$. In other words, D_1^* is d-linkable with D_0^* with confidence $P(E_i^0, S^d)$. Hence, D_1^* is d-linkable with D_2^* with the same confidence. Thus,

$$o_i = |S(E_i^1) \cap S(E_i^0)| = |S(E_i^1) \cap S(E_i^2)| \geq d \tag{5}$$

Based on this framework, we can define the objective of our protection model. Given 1) a data set D_1^* that has already been k-anonymized and 2) the expected number of shared confidential value d, our objective is to generate a data set $\widehat{D_1^*}$, such that, for each individual i in $\widehat{D_1^*}$, the composition anonymity is d with hypothesized data set D_0^* for which the publisher's confidence is $\alpha = P(E_i^0, S^d)$.

4.1. How (d, α) linkable model protects privacy

In this section, we demonstrate how the model protects privacy in the discussed adversarial scenarios, that is the adversary knows that a victim visited two organizations.

Based on the proposed model and assume that $d = 2$ and $\alpha = 0.8$, we could publish data sets D_1^* and D_2^* as shown in Table 6(a) and Table 6(b), regardless of whether or not there is an individual common to both tables. Note that for the simplicity of discussion, we assume $d = 2$, however, to ensure sufficient confidence, we assume $\alpha = 0.8$.

An adversary will know that the victim's record is in both D_1^* and D_2^*. Since these data sets were generalized according to $(2, 0.8)$-linkable model, two confidential values will be common with 0.8 confidence. Say s_1 and s_2 are the common confidential values in the published data sets D_1^* and D_2^*. When both data sets are available to the adversary, after combing them she can identify the set of possible confidential values as $\{s_1,\, s_2\}$ that could possibly correspond to the victim.

Therefore, based on the published data sets, the adversary can infer that the confidential value of the victim is either s_1 or s_2. As such, the privacy of the overlapping individual is protected with 0.8 confidence.

4.2. n-independent k-anonymized data sets

In this section, we further extend the (d, α)-linkable model to the more general case when there are n organizations' data sets.

In this scenario the publisher needs to consider $n - 1$ other k-anonymized data sets. Furthermore, since the publisher does not have the original or k-anonymized data sets from other publishers, they must calculate the linkability with $n - 1$ hypothesized data sets, generated from the population Ω. Note that each hypothesized data set is a collection of random draws and is considered to be independent.

Based on this assumption, we transform the data set to ensure it is d-linkable with $n - 1$ independent hypothesized data sets with confidence α. Note that the common confidential values S^d are assumed to be the same for an individual's equivalence class of all data sets. Since the $n - 1$ hypothesized data sets are independent, the privacy condition becomes

$$\alpha = P^{(n-1)}(E_i^0, s^d) \quad = \quad \prod_{y=1}^{n-1} P^y(E_i^0, s^d) \tag{6}$$

5. An algorithm to achieve (d, α)-linkability

The algorithmic strategy to transform a data set to meet the the (d, α)-linkable requirement is summarized in Algorithm 1. This algorithm is called *dLink*. Informally, when an equivalence class fails to satisfy the (d, α)-linkable model, it is merged with its nearest equivalence class[7]. This process repeats until the model is satisfied.

The process of merging entails generalizing records to a common set of quasi-identifying values. Now, there are several possible ways by which this could be accomplished. First, we could increase the size of the equivalence classes and re-anonymize the original data set. However, in doing so, we would apply generalization to all equivalence classes, including those which have already satisfied the privacy criterion. Alternatively, we choose to merge the equivalence class that fail to satisfy the privacy criterion with another class that fails to meet the criterion as well.

[7] The equivalence class which has the minimum distance from the target equivalence class to all other equivalence classes is called the nearest equivalence class of the target equivalence class. Distance between two equivalence classes is measured using the distance defined in [15].

Algorithm 1 dLink

Input: D_1^*, a k-anonymized data set with m quasi-identifying attributes; T, a set of generalization taxonomies for each attribute; α and d, the linkability parameters .

Output: $\widehat{D_1^*}$, an (d, α)-linkable data set

```
1: Compute the equivalence classes $E_i^1$ and the set of confidential values $S(E_i^1)$ for each
   equivalence class. Let $|E_i^1|$ represent the number of equivalence classes.
2: while there is an equivalence class that does not satisfy $P(E_i^0, S^d) \geq \alpha$ do
3:     Calculate $P(E_i^0, S^d)$ for each observed combination of $d$.
4:     if $P(E_i^0, S^d) < \alpha$ then
5:         Merge the equivalence class $i$ with its nearest equivalence class
6:     end if
7: end while
8: output $\widehat{D_1^*}$
```

We now briefly elaborate on the key steps of our algorithm $dLink$.

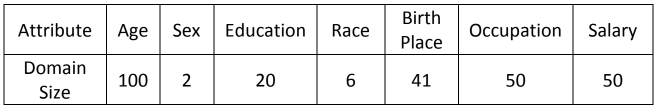
Initialization (Step 1) First, we count the number of equivalence classes. We store the equivalence class value (label) E_i^1 and the distinct confidential values for each equivalence class $S(E_i^1)$.

Checking the criteria (Step 2-3) Next, we calculate the probability that a group of confidential values (S^d) appear together in an equivalence class of the hypothesized data set D_0^*, where $S^d \subset S(E_i^1)$. Note that S^d can be any combination of d distinct confidential values from $S(E_i^1)$. For example, if $S(E_i^1) = \{s_1, s_2, s_3, s_4\}$ and $d = 3$, then there can be $\binom{4}{3} = 4$ possible groups, where each group has three distinct confidential values. Thus, $S^d = \{(s_1, s_2, s_3), (s_1, s_2, s_4), (s_1, s_3, s_4), (s_2, s_3, s_4)\}$. Generally speaking, the equivalence classes that do not have d distinct confidential values are also subject to the merge process.

Treating the equivalence class failing to satisfy the requirement (Step 5) When an equivalence class fails to satisfy the privacy criterion at Step 2 (i.e., the equivalence class is potentially subject to a composition attack), it is merged with its nearest equivalence class.

Output generalized table (Step 8) Finally, the algorithm outputs the generalized data set $\widehat{D_1^*}$.

We acknowledge that, since *dLink* applies generalization to achieve its objective, in a worst-case scenario, this may lead to suppression of the entire data set (i.e., when none of the equivalence class provides the privacy guarantee). However, our empirical analysis (in the following section) illustrates that, in practice, our method has negligible effect on data utility, while reducing the likelihood of a successful composition attack.



Attribute	Age	Sex	Education	Race	Birth Place	Occupation	Salary
Domain Size	100	2	20	6	41	50	50

Table 7: Attribute domain size

6. Experiments

We conduct our experiments in three stages. In the first stage, we compare published data sets with and without post-processing with respect to their strength against the composition attack (privacy) and, classification accuracy and query accuracy (utility). In the second stage, we compare the utility of differentially private data sets and the anonymized data sets with our method. Finally, we investigate the effects of the parameters α and d on the overall performance of $dLink$.

We perform experiments with real world data sets derived from the U.S. Census Bureau [8]. We split the data set into two independent data sets 1) Occupation and 2) Salary. Each data set consists of 600,000 records. The Occupation data set includes five QID attributes: age, sex, education, race, birth-place and one confidential attribute: occupation. The Salary data set contains the same QID attributes and replaces the confidential attribute with salary. All QID attributes consist of categorical values except age and education. The size of their domains are reported in Table 6.

We composed five disjoint data sets from each of Salary and Occupation via random draws of 100,000 records. The remaining 100,000 records are used as an overlapping pool. We made five copies of each data set in each group, and randomly inserted 1000, 2000, 3000, 4000 and 5000 records from the overlapping pool to the copies respectively yielding five sets of data of size 101,000, 102,000, 103,000, 104,000 and 105,000.

In these experiments, we apply *dLink* to data sets that have been k-anonymized via the Mondrian algorithm [14]. Composition attacks are conducted between all pairs of data sets with the same overlapping records. Note that in these experiments, we only consider the scenario that an individual visits both organizations and the adversary also has access to the anonymous data sets from those organizations. We apply Mondrian and (Mondrian + dLink) to the Occupation and Salary data sets to assess their risk and utility. The risk is measured as the accuracy of composition attack, which is defined as $\frac{\text{number of records with, } o_i{=}1 \text{ (Definition 2)}}{\text{Total number of overlapping records}} \times 100$. Unless noted otherwise, the parameters were set to $d = 4$, $\alpha = 0.8$ and $k = 10$.

[8] http://ipums.org

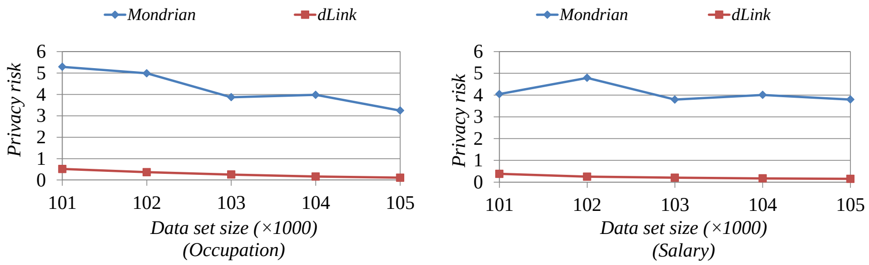


Figure 1: The average accuracy of the composition attack on 10-anonymized versions of the Salary and Occupation data sets.

We assess the utility of published data sets by the precision of classification accuracy and by the accuracy of answering range queries.

To evaluate the impact on classification accuracy, we divide the data into training and testing sets using a 10-fold cross-validation scheme based on stratified sampling. After applying the k-anonymization algorithm, the generalization level is determined solely by the training data set and then applied to the test data set. Moreover, values of the confidential attribute are discretized as binary values, class-I ($<$ 25) and class-II ($\geq$ 25), for classification purpose. For classification models, we use four classifiers $J48$ (an implementation of $C4.5$ classifier [25] in weka [10]), *naïve Bayes*, *logistic regression* and *support vector machines (SVM)*. For better visualization, we provide an additional measure *Baseline Accuracy (BA)*, which is the classification accuracy of the raw data without k-anonymization.

To evaluate the impact on query accuracy, we randomly generate 1000 queries using the following template.

SELECT COUNT (*) from D^* WHERE ($t[A_1] = x_1$ AND $t[A_2] = x_2$ AND $\ldots$ AND $t[A_m] = x_m$ AND $t[S] = s$)
where x_1, x_2, $\ldots$, x_m and s are randomly generated values.

For a query, we obtain its true result R_{act} from the original data set, and compute an estimated answer R_{est} from its anonymized data set. The relative error of a query is defined as $\frac{|R_{act}-R_{est}|}{R_{act}}$. We measure the workload error as the average relative error of all the queries of all data sets.

6.1. *Comparison with a generalization method*

We first assess the effectiveness of our method in the reduction of privacy risk of the composition attack. Figures 1 and 2 show that our method reduces the success rate of the composition attack greatly. The reduction increases with the record overlap ratio, as well as the size of the equivalence class. In some cases, the privacy risk is lower than one-tenth of the privacy risk without our method. Moreover, from Figure 2 it can be observed that the privacy risk of Mondrian and

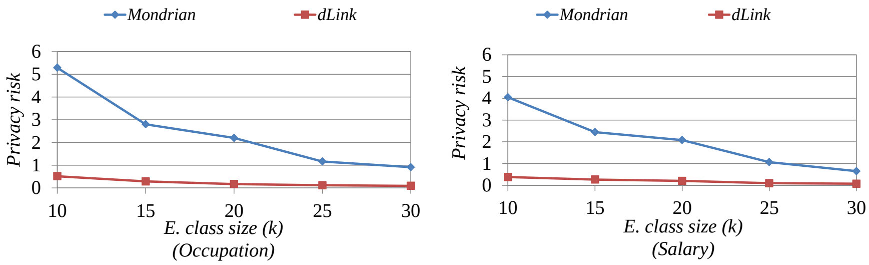


Figure 2: The average accuracy of the composition attack on k-anonymized versions of the 101K Salary and Occupation data sets.

$dLink$ approaches zero when k is large. This effect also observed in [9]. The reason for this effect is that when two equivalence classes are large, the likelihood that there are common confidential values in two equivalence classes increases significantly. Therefore, the risk of the composition attack is reduced. However, this does not mean that is unnecessary to utilize $dLink$. Let us take a moment to articulate several reasons for this claim. First, if we increase k to reduce the risk of the composition attack, we unnecessarily reduce the utility of data set. Second, a large k does not directly bound the risk of the composition attack to a certain.

Figures 3 and 4 compare the classification accuracy of k-anonymized data with and without *dLink* with increasing data set size and equivalence class size, respectively. The results show that the relative difference in accuracy is at most 0.8% before and after the application of *dLink*. The accuracies do not change much with different data size and k. Firstly, the data sets are large and the differences between them are small. Secondly, classification models mainly make use of aggregated information. k-anonymity has aggregated attribute values and such aggregation does not lose information for classification.

Figures 5 and 6 list the average query errors of the k-anonymized data sets with and without $dLink$ as a function of the data set size and equivalence class size, respectively. Each average value is obtained from 1000 random queries. The results show that relative difference in query error is at most 5% before and after the application of $dLink$. Moreover, there is no fixed pattern of the query errors with respect to data set size and k values. This is because we use different sets of queries for different pairs (Mondrian and $dLink$) of data sets. Thus, each reported pair of query errors is independent from the others.

The above results support our claim that the post-processing strategy *dLink* has negligible effect on classification accuracy and a small effect on query accuracy. It can be observed that it is possible to achieve more privacy (in the magnitude of 10) by sacrificing a little (5% or less errors in the query processing) by using the $dLink$.

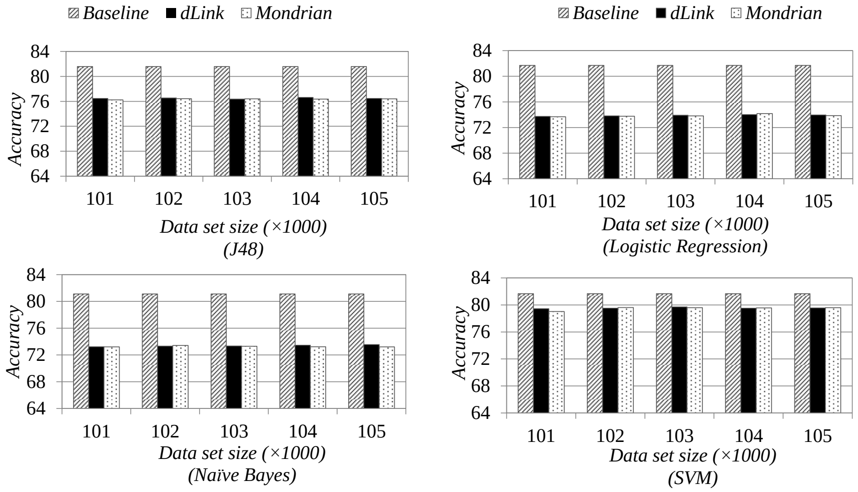


Figure 3: The average classification accuracy of the 10-anonymized Salary data sets as a function of their sizes.

6.2. Comparison with differential privacy

As discussed earlier, ϵ-differential privacy is capable of protecting data from the composition attack, but the data utility may be low. We use the following set of experiments to demonstrate this point empirically. To do so, it is crucial to choose an appropriate value for ϵ (i.e., the privacy budget). We followed the technique described in [27] to choose ϵ. We run 100,000 random queries on our data sets, and accumulate the number of unchanged responses. In Figure 7, it can be seen that when $\epsilon > 0.1$, more than 30% of the query results remain unchanged. Thus, we set the upper bound of ϵ to 0.1.

Figures 8 compares the results of the query errors of the 10-anonymous data sets after the application of $dLink$ with the differentially private data sets. The differentially private data sets are obtained by using the implementation in [22]. It can be observed that in all cases we have at least 7%, 22% and 44% less error than differentially private data sets when ϵ is set to 0.1, 0.05 and 0.01, respectively.

In addition to the above experiments, we also used Kullback-Leibler distance [12] and city block distance to measure the difference between two histograms of confidential values in the original and differentially private counts. The smaller the distance, the better the preservation of the original distribution. From Figure 9, it can be seen that the distributional distances using $dLink$ are at least 21.03%, 56.51% and 76.54% smaller than those using differential privacy when ϵ is set to 0.1, 0.05 and 0.01, respectively.

Both the results support our claim that $dlink$ can retain more data utility than

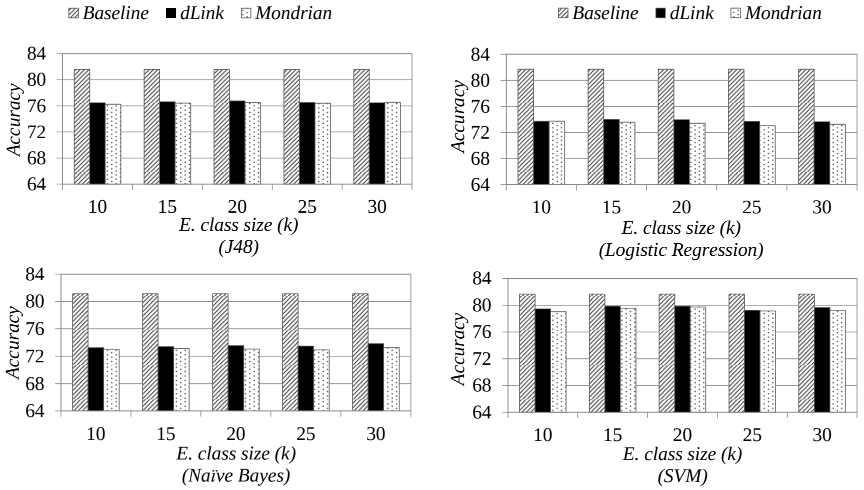


Figure 4: The average classification accuracy of the 101K Salary set as a function of the k-anonymization level.

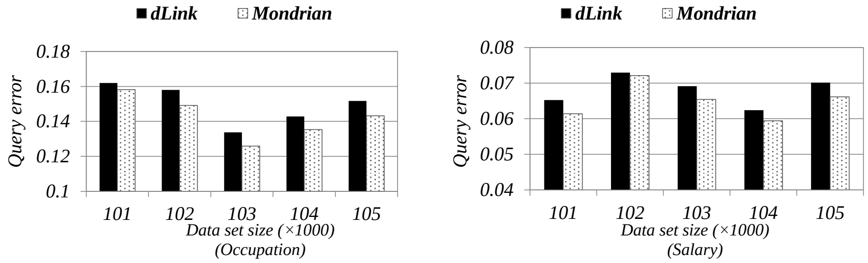


Figure 5: The average query errors of the 10-anonymized versions of the Salary and Occupation data sets as a function of their sizes.

in differentially private settings in our context.

6.3. *Effect of different critical parameters*

In the following experiments, we show the effects of the parameters d and α on the overall performance of $dLink$.

The parameter d has significant effects on the privacy risk and on the data utility. Figure 10 shows that for small d ($\leq$ 3 in our data sets), the privacy is as high as 30 times larger than the privacy risk for large d ($\geq$ 6 in our data sets). This is because with large d ($d \geq 6$), $dLink$ causes additional merging, which leads to an increase in common confidential values between corresponding equivalence classes across different data sets and a reduction in the privacy risk. However, large values of d leads to additional merging and, thus, a reduction in data utility.

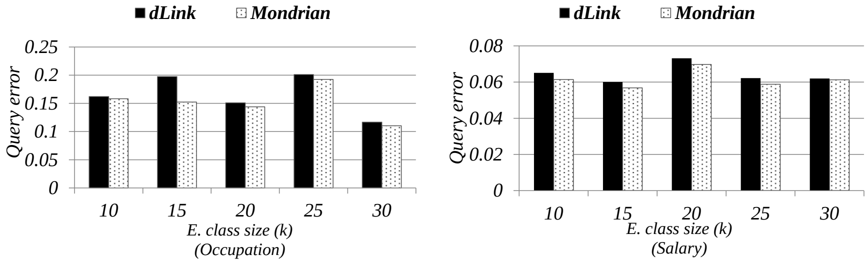


Figure 6: The average query errors of the 101K Salary and Occupation data sets as a function of the k-anonymization level.

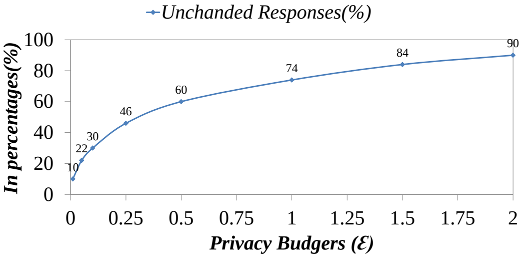


Figure 7: Unchanged responses through differentially private mechanism

Figures 13 and 12 show that for a large d, the classification accuracies decrease by 5% and the query errors are increase by 40% in comparison to the results for a small d.

The confidence parameter α influence the privacy risk and the data utility as expected. Specifically, when α is large (≥ 0.9 in our data set) we achieve more (as high as 18 times larger) privacy than a low α (≤ 0.6 in our data sets). However, it can be observed from Figures 13 and 12 that high α also leads to a decrease in the data utility. In particular, when $\alpha \geq 0.9$ the classification accuracies decrease by an average of 3.5% and the query errors increase by an average of 70% in comparison to when $\alpha \leq 0.6$. This is because to achieve high confidence, $dLink$ merges more equivalence classes.

6.4. *Efficiency*

Figure 14 shows the execution time of our post-processing $dLink$ in comparison with the anonymization process by Mondrian with increasing data set size. It can be observed that the $dLink$ incurs very small increase in process time in comparison to the anonymization. In addition, to observe the effect of the parameters d and α in Figure 15, we present the runtime of $dLink$ with different d and α along with the runtime of k-anonymization of the salary data set of 101000 records. The

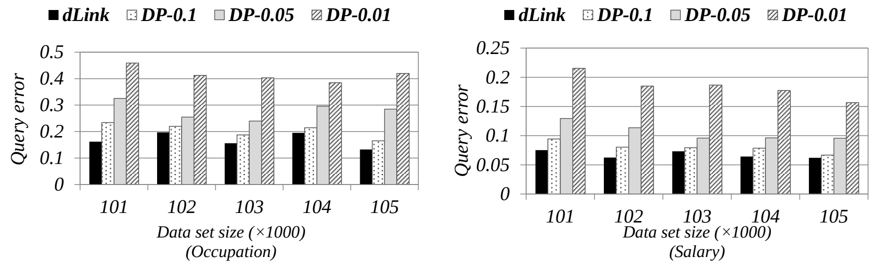


Figure 8: The average query errors of the DP [22] and 10-anonymized with $dLink$ versions of the Salary and Occupation data sets.

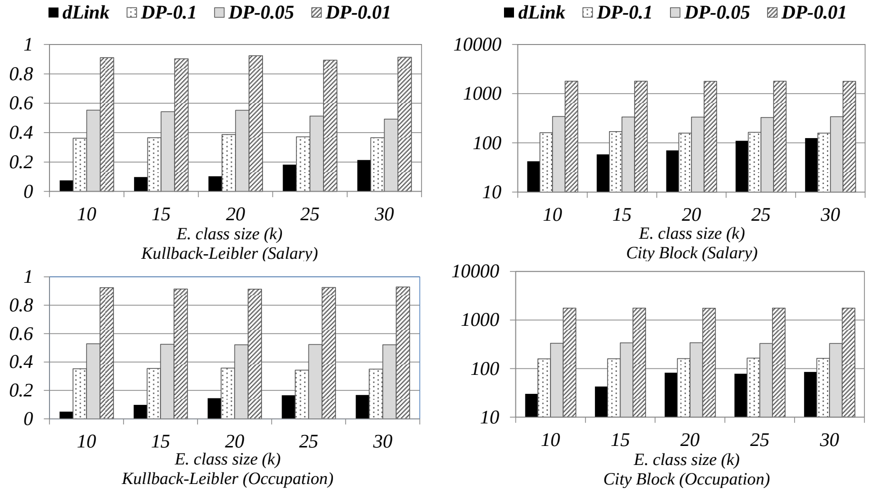


Figure 9: Distance between the original data set, the output of *dLink*, and several privacy budgets of differential privacy ($\epsilon = 0.01, 0.05, 0.1$)

X-axis represents the values of d. It can be seen that thought the runtime increases with d, the runtime of $dLink$ also incurs small additional time in comparison to the time for anonymization process.

6.5. Discussion and limitations

The composition problem transpires when an adversary combines multiple data sets to reveal the confidential information of an individual. The proposed (d, α)-model requires that a data set be published such that each equivalence class has d confidential values with a certain likelihood. Though organizations cannot coordinate in this setting, the experiments illustrate that the likelihood two data sets contain a d confidential values in common for corresponding equivalence classes can be estimated with relatively strong accuracy. Moreover, the experiments show

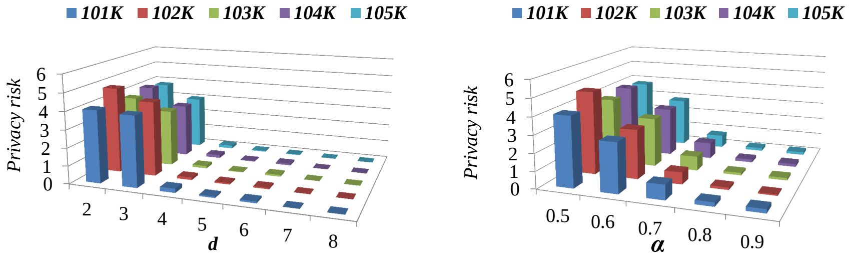


Figure 10: The average accuracy of the composition attack on 10-anonymized versions with $dLink$ of different sizes Salary data sets as functions of the parameters d and α.

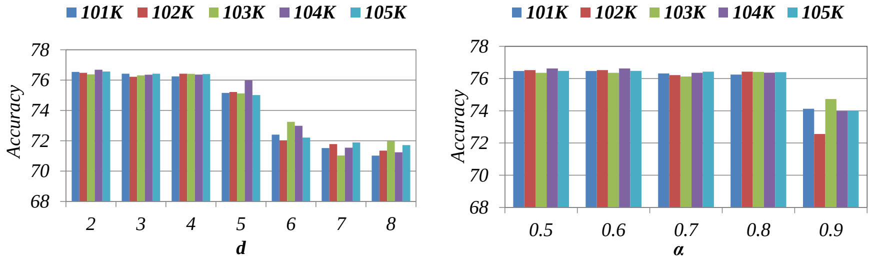


Figure 11: The average classification accuracy of the 10-anonymized versions with $dLink$ of different sizes Salary data sets as functions of the parameters d and α with classifier J48

that the proposed *dLink* algorithm can significantly reduce the likelihood of a composition attack and can retain information that is significantly closer to the original data set than differentially private publications.

There are, however, several ways in which this work could be extended and thus improved. First, our model is based on the assumption that all attributes of a record are independent. Yet, this is not necessarily the case when considering the relationship between quasi-identifying and confidential attributes. For example, in a healthcare setting, a female patient certainly has a higher chance of being diagnosed with breast cancer than a male patient. Modeling and accounting for such correlations is a complex challenge; however, we believe that more accurate statistical modeling of such relationships (i.e., used in [26, 3]) could help minimize loss in data utility. Second, our model was evaluated only in the setting of composition over two data sets, whereas the composition attack may be executed in a more distributed setting. Nonetheless, the literature suggests that in real environments, such as healthcare, the number of locations visited by a patient may be small. For instance, in [20] it was shown that patients visited a median of two hospitals in the state of Illinois over an eight year period.

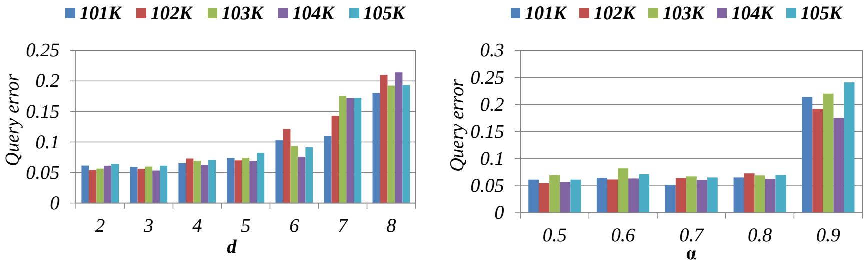


Figure 12: The average query errors of the 10-anonymized versions with $dLink$ of different sizes Salary data sets as functions of the parameters d and α.

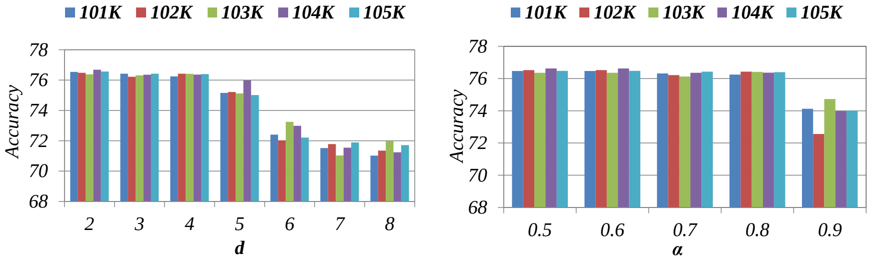


Figure 13: The average classification accuracy of the 10-anonymized versions with $dLink$ of different sizes Salary data sets as functions of the parameters d and α with classifier J48

7. Conclusion

This paper presented a k-anonymization model to limit the likelihood an adversary can successfully complete a composition attack when organizations are unable to coordinate prior to data publication. In doing so, we provided a theoretical foundation for reducing the risk of the composition attack. We further provided an effective method to achieve the privacy principle using a probabilistic approximation. We experimentally showed that the k-anonymized data adequately protects privacy and yet supports effective data analysis in a manner that is more effective than a noise-based technique (i.e., differential privacy).

Nonetheless, we believe there is room for improvement and, in particular, believe the approach would benefit from incorporating dependency models regarding the relationship between quasi-identifying and confidential values. Moreover, applying k-anonymity to both quasi-identifiers and confidential attributes could be an interesting extension of this work. In addition, incorporating a practical implementation of t-closeness to deal with data sets that lack sufficient diversity may avoid the suppression of the entire data set (i.e., the worst case scenario).

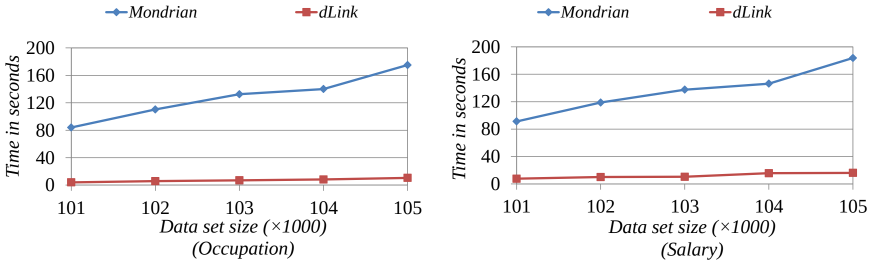


Figure 14: The execution time of the 10-anonymized versions of the Salary and Occupation data sets as a function of their sizes.

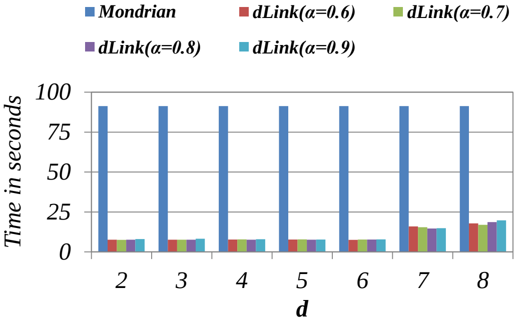


Figure 15: Execution time of $dLink$ along with the execution time of 10-anonymized version of the 101K Salary data sets as a function of d.

8. Acknowledgements

This work has been supported by Australian Research Council (ARC) discovery grant DP110103142, National Science Foundation grant CCF-0424422, and National Institutes of Health grants R01LM009989, U01HG006378 and U01HG006385.

References

[1] Charu C. Aggarwal and Philip S. Yu. A condensation approach to privacy preserving data mining. In *Proceedings of the 9th International Conference on Extending Database Technology*, pages 183–199, Heraklion, Crete, Greece, 2004.

[2] Randall D. Cebul, James B. Rebitzer, Lowell J. Taylor, and Mark Votruba. Organizational fragmentation and care quality in the U.S. health care system. Working Paper 14212, National Bureau of Economic Research, August 2008.

[3] Richard Chow, Philippe Golle, and Jessica Staddon. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 893–901, Las Vegas, Nevada, U.S., 2008.

[4] Josep Domingo-Ferrer and Vicenç Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[5] Josep Domingo-Ferrer and Úrsula González-Nicolás. Hybrid microdata using microaggregation. *Information Sciences*, 180(15):2834–2844, 2010.

[6] Cynthia Dwork. Differential privacy. In *Proceedings of the 5th International Colloquium on Automata, Languages and Programming*, pages 1–12, Venice, Italy, 2006.

[7] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53, 2010.

[8] Benjamin C. M. Fung, Ke Wang, Ada W. Fu, and Jian Pei. Anonymity for continuous data publishing. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pages 264–275, Nantes, France, 2008.

[9] Srivatsava R. Ganta, Shiva P. 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*, pages 265–273, Las Vegas, Nevada, U.S., 2008.

[10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[11] Wei Jiang and Chris Clifton. A secure distributed framework for achieving k-anonymity. *The VLDB Journal*, 15(4):316–333, 2006.

[12] Solomon Kullback and Richard Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[13] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S. Hopkins, and Nina Kruschwitz. Big data, analytics, and the path from insights to value. *MIT Sloan Management Review*, 52:21–31, 2011.

[14] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering*, pages 25–25, Atlanta, Georgia, U.S., 2006.

[15] Jiuyong Li, Raymond C. Wong, Ada W. Fu, and Jian Pei. Achieving k-anonymity by clustering in attribute hierarchical structures. In *Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery*, pages 405–416, Krakow, Poland, 2006.

[16] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering*, pages 106–115, Istanbul, Turkey, 2007.

[17] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[18] Bradley Malin. k-unlinkability: A privacy protection model for distributed data. *Data & Knowledge Engineering*, 64(1):294–311, 2008.

[19] Bradley Malin. Secure construction of k-unlinkable patient records from distributed providers. *Artificial Intelligence in Medicine*, 48(1):29–41, 2010.

[20] Bradley Malin and Latanya Sweeney. How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37(3):179–192, 2004.

[21] Sergio Martínez, David Sánchez, and Aida Valls. Semantic adaptive microaggregation of categorical microdata. *Computers & Security*, 31(5):653–672, 2012.

[22] Noman Mohammed, Rui Chen, Benjamin C.M. Fung, and Philip S. Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493–501, San Diego, California, U.S., 2011.

[23] Katherine M. Newton, Peggy L. Peissig, Abel N. Kho, Suzette J. Bielinski, Richard L. Berg, Vidhu Choudhary, Melissa Basford, Christopher G. Chute, Iftikhar J. Kullo, Rongling Li, Jennifer A. Pacheco, Luke V. Rasmussen, Leslie Spangler, and Joshua C. Denny. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the

eMERGE network. *Journal of the American Medical Informatics Association*, 20(e1):e147–e154, 2013.

[24] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, 2013.

[25] John R. Quinlan. *C4.5: programs for machine learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, San Francisco, CA, U.S., 1993.

[26] David Sánchez, Montserrat Batet, and Alexandre Viejo. Minimizing the disclosure risk of semantic correlations in document sanitization. *Information Sciences*, 249:110–123, 2013.

[27] A.H.M. Sarowar Sattar, Jiuyong Li, Xiaofeng Ding, Jixue Liu, and Millist Vincent. A general framework for privacy preserving data publishing. *Knowledge-Based Systems*, 54:276–287, 2013.

[28] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and Sergio Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, pages 1–24, 2014.

[29] Xiaoxun Sun, Hua Wang, Jiuyong Li, and David Ross. Achieving p-sensitive k-anonymity via anatomy. In *IEEE International Conference on e-Business Engineering*, pages 199–205, Macau, China, 2009.

[30] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(5):557–570, 2002.

[31] Omer Tene and Jules Polonetsky. Privacy in the age of big data: a time for big decisions. *Stanford Law Review*, 64:63–69, 2012.

[32] 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*, pages 414–423, Philadelphia, PA, U.S., 2006.

[33] William E. Winkler. Advanced methods for record linkage. In *Proceedings of the Selection on Survey Research Methods, American Statistical Society*, pages 467–472, 1994.

[34] Raymond C. Wong, Ada W. Fu, Jia Liu, Ke Wang, and Yabo Xu. Global privacy guarantee in serial data publishing. In *Proceedings of 26th IEEE*

International Conference on Data Engineering, pages 956–959, Long Beach, California, U.S., 2010.

[35] Raymond C. Wong, Jiuyong Li, Ada W. Fu, and Ke Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 754–759, Philadelphia, PA, U.S., 2006.

[36] Xiaokui Xiao and Yufei Tao. M-invariance: towards privacy preserving republication of dynamic data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 689–700, Beijing, China, 2007.