A Look Ahead Approach to Secure Multi-party Protocols

Mehmet Ercan Nergiz · Ercüment Çiçek · Yücel Saygın

Received: date / Accepted: date

Abstract Secure multi-party protocols have been proposed to enable non-colluding parties to cooperate without a trusted server. Even though such protocols prevent information disclosure other than the objective function, they are quite costly in computation and communication. Therefore, the high overhead makes it necessary for parties to estimate the utility that can be achieved as a result of the protocol beforehand. In this paper, we propose a look ahead approach, specifically for secure multi-party protocols to achieve distributed k-anonymity, which helps parties to decide if the utility benefit from the protocol is within an acceptable range before initiating the protocol. Look ahead operation is highly localized and its accuracy depends on the amount of information the parties are willing to share. Experimental results show the effectiveness of the proposed methods.

Keywords Secure multi party computation · Distributed k-anonymity · Privacy · Security

1 Introduction

Secure multi party computation (SMC) protocols are one of the first techniques for privacy preserving data mining in distributed environment [19]. The idea behind these protocols is based on the theoretical proof that two or more parties, both having their own private data, can collaborate to calculate any function on the union of their data [7]. While doing so, the protocol does not reveal anything other than the output of the function and does not require a trusted third party. While this property is promising for privacy preserving applications, SMC may be prohibitively expensive. In fact, many SMC protocols for privacy preserving data mining suffer from high computation and communication costs. Furthermore, those that are closest to be practical are based on semi-honest model, which assumes that parties will not deviate from the protocol. Theoretically, it is possible to convert semi-honest models into malicious models. However, resulting protocols are even more costly.

The high overhead of SMC protocols raises the question whether the information gain (increase in utility) after the protocol is worth the cost. This is a valid argument for mining on horizontally or vertically partitioned data (but especially crucial for horizontally partitioned data where objective function is well defined on the partitions since they have the same schema.). More specifically, for private table $T_\sigma$ of party $P_\sigma$ and an objective function $O$; initiating the SMC protocol is meaningful only if the information gain from $O$: $|I_\sigma| = |I(O(T_\sigma))| - |I(O(T_{\sigma}))|$ where $T_{\sigma}$ is the union of all private tables, is more than a user defined threshold $c$. Of course $|I_\sigma|$ cannot be calculated without executing the protocol. However it may be possible to estimate it by knowing some prior (and non-sensitive) information about $T_{\sigma}$.

To the best of our knowledge, this is the first work that looks ahead of an SMC protocol and gives an estimate for $I_\sigma$. We state that an ideal look ahead satisfies the following:

1. Methodology is highly localized in computation, it is fast and requires little communication cost (at least asymptotically better than the SMC protocol).
2. Methodology relies on non-sensitive data, or better, data that would be implied from the output of the objective function.

We state that an ideal look ahead will benefit the parties in answering the following:

1. How likely the information gain $I_\sigma$ will be within an acceptable range?
2. Since efficiency of SMC depends heavily on data, what size of private data would be enough to get an acceptable $I_\sigma$?

Our focus is the SMC protocol for distributed k-anonymity previously studied in [31,11,10]. k-Anonymity is a well known privacy preservation technique proposed in [27,24] to prevent linking attacks on shared databases. A database is said to be k-anonymous if every tuple appears in the database at least $k$ times. k-Anonymization is the process of enforcing k-anonymity property on a given database
by using generalization and suppression of values. Works in [11, 10] assume that data is vertically partitioned among two parties and they share a common key making a join possible. Authors in [11] propose a semi-honest SMC solution to create a k-anonymity of the join without revealing anything else (The protocol takes around 2 weeks time to execute for k = 100 and 30162 tuples.). Work in [31] assumes horizontally partitioned data.

The motivation behind k-anonymity or distributed k-anonymity as a privacy notion has been studied extensively in the literature. Many extensions to k-anonymity has been proposed that address various weaknesses of the notion against different types of adversaries [8, 18, 20, 22, 29, 30, 21, 3]. Diversity [20] is one such extension that enforces constraints on the distribution of the sensitive values. We first focus on the k-anonymization process and show later how the proposed methodology can be extended for diversity. Our contribution can be summarized as follows:

1. We design a fast look ahead of distributed k-anonymization that bounds the probability that k-anonymity will be achieved at a certain utility. Utility is quantified by commonly used metrics from the anonymization literature.
2. Look ahead works for horizontally, vertically and arbitrarily partitioned data.
3. Look ahead exploits prior information such as total data size, attribute distributions, or attribute correlations, all of which require simple SMC operations. Look ahead returns tighter bounds as the security constraints allow more prior information.
4. We show how look ahead can be extended to enforce diversity on sensitive attributes as in [18, 20].
5. To the best of our knowledge, this work is the first attempt in making a probabilistic analysis of k-anonymity given only statistics on the private data.

2 Background

2.1 k-Anonymity and Table Generalizations

Given a dataset (table) $T$, $T[c][r]$ refers to the value of column $c$, row $r$ of $T$. $T[c]$ refers to the projection of column $c$ on $T$ and $T[,][r]$ refers to selection of row $r$ on $T$. We write $|t \in T|$ for the cardinality of tuple $t \in T$.

Although there are many ways to generalize a given data value, in this paper, we stick to generalizations according to domain generalization hierarchies (DGH) given in Figure 1 since they are widely used in the literature.

Definition 1 (i-Gen Function) For two data values $v'$ and $v$ from some attribute $A$, we write $v' = \Delta_i(v)$ if and only if $v'$ is the $i$th parent of $v$ in the DGH for $A$. Similarly for tuples $t, t', t'' = \Delta_{i_1 \ldots i_d}(t)$ iff $t''[c] = \Delta_{c_i}(t[c])$ for all columns $c$. Function $\Delta$ returns all possible generalizations of a value $v$. We also abuse notation and write $\Delta^{-1}(v')$ to indicate the leaf nodes of the subtree with root $v'$.

E.g., given DGH structures in Figure 1. $\Delta_1(\text{USA} \Rightarrow \text{AM})$, $\Delta_2(\text{Canada}) = \ast, \Delta_0, \Delta_1(\text{<M, USA>}) \Rightarrow \ast, \Delta(\text{USA}) = \{\text{USA}, \text{AM}, \ast\}$, $\Delta^{-1}(\text{AM}) = \{\text{USA}, \text{Canada}, \text{Peru}, \text{Brazil}\}$

Definition 2 (Single Dimensional Generalization) We say a table $T^*$ is a $\mu = [i_1, \ldots, i_n]$ single dimensional generalization of table $T$ with respect to set of attributes $Ql = \{A_{i_1}, \ldots, A_{i_n}\}$ if and only if $|T| = |T^*|$ and records in $T, T^*$ can be ordered in such a way that $T^*|Ql|[] = \Delta_{i_1 \ldots i_d}(T|Ql|[])$ for every row $r$. We say $\mu$ is a generalization mapping for $T$ and $T^*$; and write $T^* = \Delta_{\mu}(T)$.

Definition 3 ($\mu$-Cost) Given a generalization $T^*, \mu$-cost returns the generalization mapping of $T^*$: $\mu(T^*) = [i_1, \ldots, i_n]$ iff $T^* = \Delta_{i_1 \ldots i_n}(T)$

For example, Tables $T_0^*, T_1^*$ are $[0, 2]$ generalizations of $T_0$ and $T_1$ respectively w.r.t. attributes sex and nation. Similarly $T_{0,1}^* = \Delta_0(T_1), T_{0,1}^* = \Delta_0(T_2). \mu$-Cost of $T_{0,1}^*$ is $[0, 1]$

Definition 4 Given two generalization mappings $\mu^1 = [i_1^1, \ldots, i_n^1]$ and $\mu^2 = [i_1^2, \ldots, i_n^2]$, we say $\mu^1$ is a higher mapping than $\mu^2$ and write $\mu^1 \leq \mu^2$ iff $\mu^1 \neq \mu^2$ and $i_j^1 \geq i_j^2$ for all $j \in [1 \ldots n]$. We define $\mu^1 - \mu^2 = \sum_i i_j^1 - i_j^2$

E.g., $[0, 2]$ is a higher mapping than $[0, 1]$.

Corollary 1 Given mappings $\mu^1 \leq \mu^2$ and $T^*_1 = \Delta_{\mu^1}(T), T^*_2 = \Delta_{\mu^2}(T)$. $T^*_2$ is better utilized (contains more information) than $T^*_1$

The above corollary is true because $T^*_1$ can be constructed from $T^*_2$. E.g., $T_{0,1}^*$ is better utilized than $T_0^*$.

In this paper, without loss of generality, we use single dimensional generalizations. However, underlying ideas can also be applied to multi dimensional generalizations [16]. We now revisit briefly k-anonymity definitions.

While publishing person specific sensitive data, simply removing uniquely identifying information (SSN, name) from data is not sufficient to prevent identification because partially identifying information, quasi-identifiers, (age, sex, nation . . . ) can still be mapped to individuals (and possibly to their sensitive information such as salary) by using
external knowledge [26]. (Even though $T_\sigma$ of Table 1 does not contain info about names, releasing $T_\sigma$ is not safe when external information about QI attributes is present. If an adversary knows some person Alice is a British female; she can map Alice to tuple q1 thus to salary >40K.) The goal of privacy protection based on $k$-anonymity is to limit the linking of a record from a set of released records to a specific individual even when adversaries can link individuals via QI:

**Definition 5** ($k$-Anonymity [26]) A table $T^*$ is $k$-anonymous w.r.t. a set of quasi-identifier attributes $QI$ if each record in $T^*[QI]$ appears at least $k$ times.

For example, $T^*_2, T^*_1$ are 2-anonymous generalizations of $T_\sigma$ and $T_1$ respectively. Note that given $T^*_2$, the same adversary can at best link Alice to tuples q1 and q4.

**Definition 6** (Equivalence Class) The equivalence class of tuple $t$ in dataset $T^*$ is the set of all tuples in $T^*$ with identical quasi-identifier values to $t$.

For example, in dataset $T^*_1$, the equivalence class for tuple $q1$ is $\{q1,q4\}$.

There may be more than one $k$-anonymizations of a given dataset, and the one with the most information content is desirable. Previous literature has presented many metrics to measure the utility of a given anonymization [9, 23, 13, 4, 1]. We revisit Loss Metric (LM) defined in [9]. LM penalizes each generalization value $v'$ proportional to $|\Delta(v')|$ and returns an average penalty for the generalization. Let $a$ is the number of attributes, then:

$$LM(T^*) = \frac{1}{|T|} \cdot a \sum_{i \neq j} \frac{|\Delta(T[i][j])| - 1}{|\Delta(a)| - 1}$$

Since $k$-anonymity does not enforce constraints on the sensitive attributes, sensitive information disclosure is still possible in a $k$-anonymization. (e.g., in $T^*_1$, both tuples of equivalence class $\{q2,q3\}$ have the same sensitive value.) This problem has been addressed in [20, 18, 8] by enforcing diversity on sensitive attributes within a given equivalence class. We show in Section 6 how to extend the look ahead process to support diversity on sensitive attributes. For the sake of simplicity, from now on we assume datasets contain only QI attributes unless noted otherwise.

### 2.2 Distributed $k$-Anonymity

Even though $k$-anonymization of datasets by a single data owner has been studied extensively; in real world, databases may not reside in one source. Data might be horizontally or vertically partitioned over multiple parties all of which may be willing to participate to generate a $k$-anonymization of the union. The main purpose of the participation is using a larger dataset to create a better utilized $k$-anonymization.

Suppose in Table 1, two parties $P_\sigma$ and $P_1$ have $T_\sigma$ and $T_1$ as private datasets and agree to release a 2-anonymous union. Since data is horizontally partitioned, one solution is to 2-anonymize locally and take a union. $T^*_\sigma, T^*_1$ are optimal (with minimal distortion) 2-anonymous full-domain generalizations of $T_\sigma$ and $T^*_1$ respectively. However, optimal 2-anonymization of $T_\sigma \cup T_1; T^*_\sigma$ is better utilized than $T^*_\sigma \cup T^*_1$. So there is a clear benefit in working on the union of the datasets instead of working separately on each private dataset.

As mentioned above, in most cases, there is no trusted party to make a secure local anonymization on the union. So SMC protocols are developed in [11, 10, 31] among parties to securely compute the anonymization with semi-honest assumption.

We assume data is horizontally partitioned but we will state how to modify the methodology to work on vertically partitioned data. We assume we have $n+1$ parties $P_\sigma,P_1,\ldots,P_n$ with private tables $T_\sigma,T_1,\ldots,T_n$. The home party $P_\sigma$ is looking ahead of the SMC protocol and remote parties $P_1,\ldots,P_n$ are supplying statistical information on the union of their private tables, $\cup_i T_i$. We use the notation $T_i$ for the global union (e.g., $T_\sigma = T_\sigma \cup \cup_i T_i$). We use the superscript * in table notations to indicate anonymizations. We use the

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
<td>F</td>
<td>England</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q_2$</td>
<td>M</td>
<td>Canada</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q_3$</td>
<td>M</td>
<td>USA</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q_4$</td>
<td>F</td>
<td>Peru</td>
<td>≤40K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
<td>F</td>
<td>*</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q_2$</td>
<td>M</td>
<td>*</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q_3$</td>
<td>M</td>
<td>*</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q_4$</td>
<td>F</td>
<td>*</td>
<td>≤40K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_5$</td>
<td>M</td>
<td>AM</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q_6$</td>
<td>M</td>
<td>AM</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q_7$</td>
<td>F</td>
<td>AM</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q_8$</td>
<td>F</td>
<td>EU</td>
<td>&gt;40K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$T_\sigma$</th>
<th>$T^*_\sigma$</th>
<th>$T^*_1$</th>
<th>$T^<em>_\sigma \cup T^</em>_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>$T^*_1$</td>
<td>$T^*_1$</td>
<td>$T^<em>_\sigma \cup T^</em>_1$</td>
</tr>
</tbody>
</table>
equivalence class
gain to exceed some threshold
In this work, we assume that parties require the local info
decide if information gain is within acceptable range:
expectations. So it is imperative for collaborating parties to
such gain is nonnegative, but could be 0 or may not meet the
justify the information gain from the protocols. Surely,
Given the cost of most SMC protocols, there arises the need
anonymity have been proposed to deal
Many extensions to k-anonymity have been proposed to deal
expectations (information content) defined on the output domain of
 anonymity does not enforce diversity on the sensitive values
k[8,18,20,22,29,30,21,3]. Problems arise mostly because
k[5]. Let the or-
Le\uuml;r the data privacy
Definition 7 (Recursive (c, \ell)-diversity [20]) Let the or-
Table 1, \(T_i^*\) is (0.5, 2)-diverse since for all equiva-
From now on, without loss of generality, we assume we
have only two values in the sensitive attribute domain (\(m =
2, \ell = 2\)). In Table 1, \(T_i^*\) is (0.5, 2)-diverse since for all equiva-
ences classes, the frequencies of \(\leq 40K\) and \(>40K\) are the
same (i.e., \(r_1 = r_2\)). However \(T_i^*\) does not respect any di-
versity requirement (except when \(c = 0\)), since all tuples in
equivalence class \(\{q_2, q_3\}\), have salary \(\leq 40K\).

3 Information Gain

Given the cost of most SMC protocols, there arises the need
to justify the information gain from the protocols. Surely,
such gain is nonnegative, but could be 0 or may not meet the
expectations. So it is imperative for collaborating parties to
decide if information gain is within acceptable range:

Definition 8 (Info Gain) Let \(P_0, P_1, \ldots, P_n\) be \(n + 1\) par-
ties with private tables \(T_0, T_1, \ldots, T_n\). Let \(O\) be the objec-
tive function for the SMC protocol and \(I\) be the utility func-
tion (information content) defined on the output domain of
\(O\). Local info gain for a single party \(P_0\) is defined as \(|O_0| =
I(O(T_0)) - I(O(T_0^*))\) where \(T_0^* = T_0 \cup T_i\). Global info gain
for the protocol is \(|I| = \sum_i |I_i| + |O_0|\).

Each party involving in an SMC expects to gain from
SMC either locally or globally depending on the application.
In this work, we assume that parties require the local info
gain to exceed some threshold \(c\) before they proceed with
the SMC protocol. However, without total knowledge of all
private tables (\(T_i\)), parties can only have some confidence
that SMC will meet their expectations:

Definition 9 (c, p-sufficient SMC) For a party \(P_c\), an SMC
is \(c, p\)-sufficient with respect to some prior knowledge \(K\) on
\(\bigcup T_i\), if \(\mathcal{P}(|I_0| \geq c \mid K) \geq p\). We say SMC is \(c, p\)-sufficient
iff it is \(c, p\)-sufficient for all parties involved.

Our goal in a look ahead process will be to check if a
given SMC is \(c, p\)-sufficient for a user defined \(c\) and \(p\).
For distributed \(k\)-anonymity, the objective function \(O\) is
trivially the optimal \(k\)-anonymization which we name as \(O_k\).
Specifically, in this paper, we will make use of single dimen-
sional generalizations to achieve \(k\)-anonymity. This general-
ization technique has been used in many previous work
on anonymization [15, 20, 18, 22]. As mentioned above, our
work can be extended for multidimensional generalizations
[16, 22] as well.

Information gain \((I)\) is proportional to the quality of the
anonymization. It is challenging to come up with a standard
metric to measure the quality of an anonymization [23]. In
this work, we will be using the \(\mu\)-cost as the quality metric.
Recall that a higher mapping is less utilized than a lower
mapping, and ‘
\(\cdot\)’ operation has been defined over mappings
in Definition 4. \(\mu\)-cost can be used for horizontally parti-
tioned data.

Calculation of LM cost is possible if we know attribute
distributions (denoted with \(K_F\)) and the generalization map-
ning. So there is a direct translation between the \(\mu\)-cost and
LM cost for single dimensional generalizations given \(K_F\).
The advantage of translating \(\mu\)-cost to LM cost is that LM
cost can be used for arbitrarily partitioned data. For vertical
partitioning, each party has at least one missing attribute.
We assume a total suppression (\(\ast\)) for data entries from the
missing attributes when calculating LM cost.

We can now specialize \(c, p\)-sufficiency for distributed \(k\)-
anonymity problem:

Definition 10 (c, p-sufficient k-Anonymity) For a party \(P_c\),
a distributed \(k\)-Anonymity protocol is \(c, p\)-sufficient with re-
spect to some prior knowledge \(K\) on \(\bigcup T_i\), iff

\[\mathcal{P}(\mu(O_k(T_0)) - \mu(O_k(T_0^*)) \geq c \mid K) \geq p\]

We say SMC is \(c, p\)-sufficient iff it is \(c, p\)-sufficient for
all parties involved.

Informally, SMC is sufficient for an involving party if
the difference between the optimal generalization mapping
for the union and the optimal mapping for the local table is
more than \(c\) with \(p\) probability. Of course, the party can only
calculate such a probability if she has some knowledge on
the union denoted by \(K\). The amount of prior knowledge \(K\)
is crucial in successfully predicting the outcome of an SMC.
As mentioned before, prior knowledge \(K\) cannot be sensitive
information. Non-sensitive \(K\) can be derived in three ways:

1. Information that could also be learned from the anony-
mization such as the global dataset size.
2. Statistics about global data that are not considered as sensitive. In the case of $k$-anonymity, statistics that are not individually identifying such as attribute distributions are acceptable.

3. Based on the assumption that global joint distribution is similar with local distribution, information that can be gained from the local dataset. This type of prior knowledge is the most tricky one since over fitting to local distribution needs to be avoided. Such an information can be in terms of highly supported association rules in the local dataset.

We show, in later sections, how to check for sufficiency of distributed $k$-anonymity protocol given global attribute distributions which we denote with $K_F$.

**Definition 11 (Global attribute distribution $K_F$)** A distribution function $f_c^T$ for an attribute $c$ is defined over a dataset $T$ such that a given value $v^c$ returns the number of entities $t$ in $T$ with $v^c \in \Delta(t[c])$. Global attribute distribution $K_F$ sent to a home party $P_0$ contains all distribution function on $\bigcup_i T_i$.

In Table 1, $f^T_{\text{nation}}(\forall) = 3, f^T_{\text{nation}}(\forall') = 1$. For the parties $\{P_0, P_1\}$, $K_F = \{f^T_{\text{sex}}, f^T_{\text{nation}}\}$.

### 4 Problem Definition

Given Section 3, distributed $k$-anonymity protocol is $c, p$-sufficient for $P_0$ iff

$$P(\mu(O_k(T_o)) - \mu(O_k(T_g)) \geq c \mid K_F) \geq p$$

$$\mu^c = \mu(O_k(T_g))$$ requires local input and can be computed by $P_0$.

$$P(\mu(O_k(T_o)) - \mu^c \geq c \mid K_F) \geq p$$

Let $S_\mu = \{\mu_1^c, \cdot \cdot \cdot, \mu_m^c\}$ be the mappings that are exactly $c$ distance beyond $\mu^c$ and $\{\mu_1^c, \cdot \cdot \cdot, \mu_m^c\}$ be the mappings that are more than $c$ distance beyond $\mu^c$. Let also $A_\mu$ be the event that $\Delta_\mu(T_o)$ is $k$-anonymous. Then we have;

$$P(\mu(O_k(T_o)) - \mu^c \geq c \mid K_F)$$

$$= P((\bigcup_i A_{\mu_i^c}) \cup (\bigcup_i A_{\mu_{i+1}^c}) \mid K_F)$$

$$\geq \max_i P(A_{\mu_{i+1}^c} \mid K_F)$$

**Suppose in Table 1, $P_o$ needs to check for $(1,p)$-sufficiency.** Optimal 2-anonymization for $P_o$’s private table $T_o$ is $T_g$ with $\mu(T_o) = [0,2]$. There is only one mapping $[0,1]$ which is 1 away from $[0,2]$. So we need to check if $P(\Delta_{0,1}(T_o) < F, EU)$. When we refer to this bucket set, we will index the elements: $\{b_1, b_2, b_3, b_4\}$

### 5 $\mu$-Probability of a Protocol

**Definition 12 (Bucket Set)** A bucket set for a set of attributes $C$, and a mapping, is given by $B = \{\text{tuple } b \mid \exists t \in \text{domain of } C \text{ such that } b^c = \Delta_\mu(t)\}$

In Table 1, for the domain tables defined and the mapping $[0,1]$, the bucket set is given by $\{\forall, \forall', \forall''\}$. When we refer to this bucket set, we will index the elements: $\{b_1, b_2, b_3, b_4\}$

#### 5.1 Assumptions

Deriving the exact $\mu$-probability is a computationally costly operation. To overcome this challenge, we make the following assumptions in our probabilistic model:

**Attribute Independence:** Until Section 5.6, we assume that there is no correlation between attributes. This is a valid assumption if we only know $K_F$ about the unknown data. So from $P_o$’s point of view, for any foreign tuple $t \in T_i$;

$$P(t[i] = v_i) = P(t[i] = v_i \mid t[j] = v_i)$$

for all $i \neq j, v_i, v_j$.

**Tuple Independence:** We assume foreign tuples are drawn from the same distribution but they are independent. Meaning for any two tuples $t_1, t_2 \in T_2$, $P(t_1[i] = v_i) = P(t_2[i] = v_i \mid t_2[j] = v_k)$ for all possible $i, v_i, v_j, v_k$. Such equality does not necessarily hold given $K_F$, but for large enough data, independence is a reasonable assumption. In Section 7, we experimentally show that tuple independence assumption does not introduce any deviation from the exact $\mu$-probability.

#### 5.2 Deriving $\mu$-Probability

Generalization of any table $T_o$ with a fixed mapping $\mu$ can only contain tuples drawn from the associated bucket set $B = \{b_1, \cdot \cdot \cdot, b_n\}$. Since we don’t know $T_o$, the cardinality of the buckets act as a random variable. However, $P_o$ can
extract the size of the $\bigcup_i T_i$ from $K_F$. Letting $X_i$ be the random variable for the cardinality of $b_i$, and assuming $\bigcup_i T_i$ has cardinality $N$, we have the constraint $\sum_i X_i = N$.

In Table 1, from $P_F$’s point of view $N = |T_1| = 4$. So for the four buckets above; $X_1 + X_2 + X_3 + X_4 = 4$.

The generalization $T^*_F$ satisfies $k$-anonymity if each bucket (generalized tuple) in $T^*_F$ has cardinality of either 0 or at least $k$. For horizontally partitioned data, party $P_F$ already knows his share on any bucket, so the buckets are initially non-empty. Let $X_i \geq 0$ denote the case when $X_i = 0$ or at least $k$ in the case of vertically partitioned data and $X_i + |b_i \in \Delta_k(T_0)| \geq 0$ in the case of horizontally partitioned data then $\mu$-probability takes the following form:

$$P(\bigcap_i X_i \geq 0 \mid \sum_i X_i = N, K_F)$$

If we have the knowledge of the distribution functions for the attributes $K_F = \bigcup_i f_i$, the probability that a random tuple $t \in T_i$ will be generalized to a bucket $b_i$ is given by

$$\ell_i = \prod_{c} \frac{f_c(b_i[c])}{N}$$

which we will name as the likelihood of bucket $b_i$.

For example, in Table 1, $P_F$ assumes to know the attribute distribution set $K_F = \{f_{\text{sex}}, f_{\text{nation}}\}$. (E.g., $f_{\text{sex}}^{(\text{Brazil})} = 1$). Thus the likelihood of bucket $b_1 (\{<\text{male}, \text{USA}\})$ is $\ell_1 = \frac{f_{\text{sex}}^{(\text{male})} f_{\text{nation}}^{(\text{USA})}}{N} = \frac{2}{8} \times \frac{3}{8} = \frac{3}{16}$. Similarly $\ell_2 = \frac{1}{8}, \ell_3 = \frac{3}{8}, \ell_4 = \frac{1}{8}$.

Without tuple independence assumption, each $X_i$ behaves like a hypergeometric random variable with parameters $(N, N\ell_i, N)$. However, hypergeometric distribution function is slow to compute. But with tuple independence, we can model $X_i$ as a binomial random variable $\mathcal{B}$ with parameters $(N, \ell_i)$. Such an assumption is reasonable for big $N$ and moderate $\ell$ values [14]. Figure 2 summarizes our probabilistic model. Each tuple is represented by a ball with a probability $\ell_i$ of going into a bucket $b_i$. Then the $\mu$-probability can be written as:

$$P_\mu = P(\bigcap_i X_i \geq 0 \mid \sum_i X_i = N, X_i \sim \mathcal{B}(N, \ell_i))$$

In Table 1, $|b_1 \in \Delta_k(T_0)| = 2$ similarly for $b_2, b_3, b_4$. initial bucket sizes are 0, 1, 1. So for $k = 2$, $P_\mu = P(X_1 \geq 0, X_2 \geq 0, X_3 \geq 1, X_4 \geq 1)$

5.3 Calculating exact $\mu$-Probability

$P_\mu$ can be calculated in two ways:

1. A recursive approach can be followed by conditioning on the last bucket:

$$P_{\mu^{n,\ell: \cdot}} = P(\bigcap_i (X_i \geq 0) \mid \sum_i X_i = N, X_i \sim \mathcal{B}(N, \ell_i))$$

$$= \sum_{x \geq 0} P(X_n = x) P(\bigcap_i (X_i \geq 0) \mid \sum_i X_i = N, X_i \sim \mathcal{B}(N, \ell_i), X_n = x)$$

$$= \sum_{x \geq 0} \mathcal{B}(x; N, \ell_n) \cdot P(\bigcap_i (X_i \geq 0) \mid \sum_i X_i = N - x, X_i \sim \mathcal{B}(N, \ell_i))$$

$$= \sum_{x \geq 0} \left(\frac{N}{N-x}\right)\ell_n^{x}(1-\ell_n)^{N-x} \cdot P_{\mu^{n-1,\ell: \cdot}}$$

where $\ell_i$ is the normalized likelihood $\ell_i = \frac{\ell_i}{\sum_j \ell_j}$.

2. Each tuple in $\bigcup_i T_i$ can be thought of an independent trial in a binomial process in which each trial results in exactly one of the $n$ possible outcomes (e.g., $b_1, \ldots, b_n$). In this case, the joint random variable $(X_1, \ldots, X_n)$ follows a multinomial distribution with the following density function:

$$P(X_1 = x_1 \ldots X_n = x_n) = \frac{N!}{x_1! \ldots x_n!} \ell_1^{x_1} \ldots \ell_n^{x_n}$$

$P_\mu$ can be calculated by summing up the probabilities of all assignments that respect $k$-anonymity:

$$P_\mu = \sum_{\sum_i N \land X_i \geq 0} \frac{N!}{x_1! \ldots x_n!} \ell_1^{x_1} \ldots \ell_n^{x_n}$$

In Table 1, following the example above, one assignment that satisfies 2-anonymity is $X_1 = 0, X_2 = 1, X_3 = 0, X_4 = \ldots$
3. The probability share of this assignment on \( \mathcal{P} \) can be calculated as
\[
\mathcal{P}(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 3) = \mathcal{P}(X_4 = 3) \mathcal{P}(X_3 = 0 | X_4 = 3) \\
\cdot \mathcal{P}(X_1 = 1 | X_2 = 0, X_4 = 3) \\
\cdot \mathcal{P}(X_1 = 0 | X_2 = 1, X_3 = 0, X_4 = 3) \\
= \frac{4!}{0! \cdot 1! \cdot 0! \cdot 3!} \cdot \frac{1}{3^4} = 0.026
\]

If we sum up all the probabilities for valid assignments, we get the \( \mu \) probability as 0.23.

Unfortunately, calculating \( \mu \)-probability with Equation 3 or 4 is computationally expensive. Number of assignments that satisfy \( X_1 + \cdots + X_6 = N \), thus number of binomials that needs to be calculated is in the order of \( \binom{N+n-1}{n} \). We next show how to bound and approximate \( \mu \)-probability.

5.4 Bounding \( \mu \)-Probability

Let \( E_i \) be the event that \( 0 \leq X_i \leq k \). Obviously,
\[
\mathcal{P}_\mu = \mathcal{P}(\bigcap E_i) = 1 - \mathcal{P}(\bigcup E_i)
\]

Bounding the probability of union of events is well studied in the literature. One of the most common bounds is given by Bonferroni [5]:
\[
\mathcal{P}_\mu = 1 - (z_1 - z_2 + z_3 - z_4 + \cdots (-1)^{n-1} z_n)
\]

where \( z_1 = \sum \mathcal{P}(E_i) \), \( z_2 = \sum_{i<j} \mathcal{P}(E_i, E_j) \),
\[
\mathcal{P}_\mu \leq 1 - z_1 + z_2, \quad \mathcal{P}_\mu \leq 1 - z_1 + z_2 - z_3 + z_4, \quad \cdots
\]

Following the example above, \( z_1 = \mathcal{P}(E_1) + \mathcal{P}(E_2) + \mathcal{P}(E_3) + \mathcal{P}(E_4) = 0 + \binom{4}{1} \ell_2 (1 - \ell_2)^3 + \binom{4}{0} (1 - \ell_3)^4 + \binom{4}{1} (1 - \ell_4)^4 = 1.074 \).
\[
z_2 = \mathcal{P}(E_1, E_2) + \mathcal{P}(E_1, E_3) + \mathcal{P}(E_1, E_4) + \mathcal{P}(E_2, E_3) + \mathcal{P}(E_2, E_4) + \mathcal{P}(E_3, E_4) = 0 + 0 + 0 + \binom{4}{0} \ell_2 (1 - \ell_2 - \ell_3)^3 + \binom{4}{0} \ell_3 (1 - \ell_2 - \ell_3)^3 + \binom{4}{1} \ell_4 (1 - \ell_2 - \ell_4)^3 + \binom{4}{3} (1 - \ell_3 - \ell_4)^4 = 0.336.
\]

Even though Bonferroni always holds, it does not guarantee tight bounds [25]. Besides that, calculation of high dimensional marginal distributions may still be infeasible for large data. In Section 7, we experimentally show the efficiency of the bounding algorithms.

5.5 Approximating \( \mu \)-Probability

In this section, we adapt the approximation of multinomial cumulative distribution given in [17] to \( \mu \)-probability. The resulting approximation is much faster to compute compared to bounding techniques. Even though the error of the approximation is unbounded, as we show in Section 7, the approximation is practically quite accurate.

Let \( A_i \) be the event that \( X_i \geq 0 \), then given \( X_i \sim \mathcal{B}(N, \ell_i) \) we have:
\[
\mathcal{P}_\mu = \mathcal{P}(\bigcap A_i) = \frac{\mathcal{P}(X_i = N)}{\mathcal{P}(X_i = N)}
\]

where \( Y_i \) is a truncated binomial; \( Y_i \sim (X_i|X_i \geq 0) \).

The second numerator term is a probability of independent binomials so it can easily be computed as:
\[
\mathcal{P}(\bigcap A_i) = \prod \mathcal{P}(X_i \geq 0)
\]

The first numerator term and the denominator however is the probability regarding sums of random variables which are independent but not identically distributed. However, since both \( X_i \) and \( Y_i \) are bounded, by Lindeberg theorem [5], the central limit theorem holds; distribution of the sums converges to a normal distribution \( \mathcal{N} \) as \( n \) goes to infinity. So given \( (X_i, Y_i) \) is the mean and \( (\sigma_X^2, \sigma_Y^2) \) is the variance of \( (X_i, Y_i) \) respectively, then
\[
\mathcal{P}_\mu \leq \frac{\mathcal{P}(|X_i - N| \leq 0.5)}{\mathcal{P}(|X_i - N| \leq 0.5)} \prod \mathcal{P}(X_i \geq 0)
\]

5.6 Handling Correlations

So far, we assumed that only the knowledge of the global attribute distributions is used to estimate \( \mu \)-probability. However, such information cannot successfully describe a global dataset with high attribute correlations. Sharing joint attribute distribution instead of single attribute distributions can be a...
solution. However, such sharing should be done carefully since supplying all joint probabilities with an SMC not only is inefficient due to the large domain of joint attributes but also might give out too much sensitive information.

Instead, parties can agree on summary structures that contain information on correlations. Bayesian networks (BN) are one such example. BNs are basically directed graphs in which each vertex corresponds to an attribute and each edge shows a dependency relation between the connected attributes (see Figure 3). Attribute pairs not connected by an edge are assumed to be conditionally independent. Joint probabilities on the connected attributes are supplied with the BN structures. The advantages of using BNs are 1. the level of dependency to be shared can be adjusted, thus information disclosure and the amount of communication traffic can be limited. 2. BNs are also proposed for query selectivity in database systems, thus might be readily available within the database management system [6].

In situations where local data of a party is assumed to follow a similar distribution with the global data, BNs can be constructed locally without requiring an SMC. Such an approach is however, only applicable to horizontally partitioned data.

Incorporating information from a BN structure does not complicate the computation of \( \mu \)-probability. BNs only affect the likelihoods of buckets given in Section 5.2. Thus only Equation 1 is rewritten in terms of joint distributions.

### 5.7 Communication Protocol

As we mentioned before, assuming we have \( n \) remote parties \( P_1, \ldots, P_n \), the knowledge \( K_F \) that \( P_0 \) gets should describe \( T_0^0 \rightarrow \bigcup_{i=1}^n T_i \). Having \( P_0 \) get separate distributions on \( T_i \) from each party \( P_i \) would disclose too much information. Our goal in this section is two folds:

- We limit the information disclosure by extracting information from the local anonymizations rather than private tables. This is useful if the privacy policies of the participating parties prevent disclosure of any information of finer granularity.
- We use SMC protocols to calculate the global \( K_F \). Thus for more than 2 parties, private shares inherent in the global \( K_F \) are indistinguishable. Fortunately, such a protocol is not costly for non-colluding parties.

**Algorithm 1 Secure distribution of \( K_F \)**

**Require:** Parties \( \{P_0, P_1, \ldots, P_n\} \). \( P_0 \) gets distribution functions on \( \bigcup_{i=1}^n T_i \).

1. for all dimensions \( d \) do
2. Let set of values \( \{v_1, \ldots, v_n\} \) be the domain of \( d \)
3. \( P_0 \) sends vector of random number \( R^n = \{r^n_1, \ldots, r^n_n\} \) to \( P_i \)
4. \( i = 1 \)
5. while \( i < n \) do
6. \( P_i \) calculates distorted distribution function
7. \( G_i = \text{getUniform} \ (d) \)
8. \( P_i \) sends \( R^n = R^{n-1} + \{G_i(v_1), \ldots, G_i(v_n)\} \) to \( P_{i+1} \)
9. \( i = i + 1 \)
10. \( P_0 \) calculates the distribution for \( d \); \( f_0(v_i) = r^n_i - r^n_i \)

**Algorithm 2 getUniform\( (d) \)**

**Require:** Party has the private table \( T \) and \( d \) is a dimension in \( T \). Set of values \( V = \{v_1, \ldots, v_n\} \) is the domain of dimension \( d \).

**Ensure:** Function \( G \) is the distorted distribution function for dimension \( d \).

1. Let \( T^* \) be the \( k \)-anonymization of \( T \) with mapping \( \mu \).
2. for all \( v_i \in V \) do
3. Get generalized value \( v^* = \Delta_\mu(v_i) \) in \( T^* \) and let \( p \) be the frequency of \( v^* \)
4. Let \( V^* = \{v_j \mid v_j \in V \wedge \exists v_j \in \Delta^* \mu(v^*)\} \) and let \( q \) be the size of \( V^* \)
5. Assuming each value is equally likely to appear, pick randomly a vector \( R = \{r_1, \ldots, r_q\} \) such that \( \sum_{v \in V} r_i = p \)
6. \( G(v_j) = r_j \) for \( v_j \in V^* \)
7. \( V = V - V^* \)
8. Return \( G \)

Algorithm 1 shows how parties can calculate global \( K_F \) securely. In line 3, \( P_0 \) supplies a random number for each domain value \( v_i \). In line 7, each party adds its private share (which we explain shortly) to the random sum and the last party sends the final sum back to \( P_0 \). \( P_0 \) finds the global distribution by subtracting the initial random number from the final sum.

The important point here is that private shares of parties do not contain the exact frequency of \( v_i \). Parties distort the frequency as given in Algorithm 2. The algorithm getUniform returns new distributions from the local \( k \)-anonymization other than the private table. The anonymized distribution (of values of possibly coarser granularity) is first extracted and a new distribution on atomic values (e.g., \( G \)) that respects the anonymized distribution is returned randomly. Randomization should enforce symmetry thus indistinguishability between each atomic value in the same
equivalence class \( \mathcal{P}(G(v_i) = x) = \mathcal{P}(G(v_j) = x) \) for all \( i, j, x \).

An example is to draw the frequencies from a multinomial distribution:

\[
P(r_1 = x_1, \ldots, r_q = x_q \mid r_1 + \cdots + r_q = p) = \begin{cases} 
0, & x_1 + \cdots + x_q \neq p; \\
\frac{p^q}{x_1! \cdots x_q! q!}, & \text{otherwise.}
\end{cases}
\]

Note that \( K_F \) from the remote parties is truthful only on a coarse granularity decided by their local anonymizations. Thus, information released by the remote parties is bounded by their local \( k \)-anonymizations.

Suppose this time we have the parties \( P_\sigma, P_1, \) and \( P_2 \) with private tables \( T_\sigma, T_1, \) and \( T_2 \) in Tables 1 and 2. We again assume \( P_\sigma \) initiates a look ahead with info \( K_F \). To get the frequency of a domain value, say \( \text{Male} \); first, \( P_\sigma \) picks a random number \( r \) and sends it to \( P_1, P_2 \) calculates \( G([t]) \) from \( T_1^* \). Since \( \text{Sex} \) column is not generalized in \( T_1^* \), \( G_2([t]) \) returns the exact frequency of the males: 2. Thus, \( P_2 \) sends \( r + 2 \) to \( P_2, P_2 \) calculates \( G_2([t]) \) from \( T_2^* \). * is the generalization of \( M \) in \( T_2^* \) with frequency 4 (e.g., \( p = 4 \)). There are two atomic values \( M \) and \( F \) under * (e.g., \( q = 2 \)). Thus the output of \( G_2 \) on \( M \) and \( F \) will respect the following multinomial distribution:

\[
P(G_2([t]) = x, G_2([t]) = y) = \begin{cases} 
\frac{1}{16}, & x = 0, y = 4 \\
\frac{2}{16}, & x = 1, y = 3 \\
\frac{6}{16}, & x = 2, y = 2 \\
\frac{4}{16}, & x = 3, y = 1 \\
\frac{1}{16}, & x = 4, y = 0
\end{cases}
\]

5.8 Information Disclosure

In this section, we discuss for a given party how much information on his/her private input is disclosed to other parties due to the look ahead. The amount of information disclosure depends on the outcome of the look ahead. So we evaluate the disclosure case by case. We start with disclosure by the remote parties:

1. Insufficient SMC If the look ahead concludes that SMC will not meet the expectations, we assume each party releases their local anonymizations. Note that Algorithm 1 only operates on the local anonymizations of the remote parties meaning any adversary can simulate the look ahead from the released anonymizations. Thus, with respect to remote parties, there is no information disclosure due to the look ahead process.

2. Sufficient SMC In this case parties initiate the SMC protocol. We again have two cases to consider depending on the output of the SMC protocol. Let \( T_0^* \) be the output and party \( P_i \) has private input \( T_i \) with local anonymization \( T_i^* \).

2.1. \( \mu(T_i^*) \subset \mu(T_0^*) \) If \( T_i^* \) is a higher level generalization than \( T_0^* \); the released data \( T_0^* \) contains more information on \( T_i \) than the local anonymization \( T_i^* \). Thus \( K_F \) on \( T_i^* \) does not give out any more than \( T_0^* \).

2.2. \( \mu(T_i^*) \not\subset \mu(T_0^*) \) We now try to upper bound the information disclosure in this case. Surely, \( P_i \) sending the exact local anonymization \( T_i^* \) to \( P_0 \) (as opposed to distributions on \( T_i^* \)) results in a higher information disclosure. In such a case, \( P_0 \) sees two different anonymizations of \( T_i \) enabling him/her conduct intersection attacks to recover some data cells in finer granularity. Table 3 shows an example where \( T_2^* \) is the local anonymization of \( T_2 \) and \( T_{0,2}^* \) is what \( P_0 \) sees at the end of the protocol. Seeing \( T_2^* \) and \( T_{0,2}^* \) together, \( P_0 \) can conclude that there is a tuple in \( T_2 \) with \( \text{sex} = \text{Male} \), \( \text{nation} = \text{EU} \), \( \text{salary} > 40K \). Note that such information cannot be extracted from \( T_{0,2}^* \) alone.

However, even though it is possible, it is quite unlikely to launch intersection attacks in our protocol. Because \( P_0 \) only sees the global \( K_F \) on all local anonymizations. It is not possible to distinguish distribution of one party from that of another. Even for the two party case, \( K_F \) is distorted and there is no way of telling the granularity of truth in \( K_F \) (e.g., generalization mapping). Besides even if \( P_0 \) knows the mapping, it is still unlikely to link values to an individual in \( T_i \). Consider in Table 3, \( P_0 \) knows \( T_{0,1}^* \) and distributions on \( T_i^* \). In addition to \( T_{0,1}^* \), \( P_0 \) will discover from the distributions that there are two nationalities of \( \text{EU} \) and \( \text{NAT} \) in \( T_1 \). However, he/she will not be able to link the attribute \( \text{nation} \) with \( \text{sex} \) or \( \text{age} \). In other words, what \( P_0 \) gets from \( K_F \) does not help to gain knowledge on any of the individuals.

Nevertheless, it is possible for parties to avoid case 2.2 by enforcing the output generalization mapping to be some descendant of all local mappings. This would negate any information disclosure at the cost of utility. Such an approach makes sense especially for the two party case in which disclosure risk is the highest. Also the utility loss for two party case is minimized since it becomes easier to find a common descendant mapping.

As for the disclosure by the home party, at the end of the look ahead, the remote parties will know the decision of \( P_\sigma \) on inputs \( T_\sigma, K_F \). This information cannot be simulated from \( T_\sigma^* \) alone, thus there is a non-zero information disclosure. This situation could have been avoided by enforcing the home party to use \( T_{\sigma}^* \) instead of \( T_\sigma \) in the look ahead. However, we advocate that the risk of disclosure is very small to take such a precaution at the cost of utility. It is unlikely for the remote parties to infer anything from the decision since no party knows the exact global \( K_F \). Besides, decision does not disclose the exact \( \mu \)-probability making any inference on \( T_\sigma \) difficult if not impossible.
Table 2 Tables for party $P_2$

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q^9$</td>
<td>30</td>
<td>F</td>
<td>England</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q^{10}$</td>
<td>17</td>
<td>M</td>
<td>Canada</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{11}$</td>
<td>24</td>
<td>M</td>
<td>USA</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{12}$</td>
<td>25</td>
<td>F</td>
<td>Italy</td>
<td>≤40K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q^9$</td>
<td>10-20</td>
<td>*</td>
<td>EU</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q^{10}$</td>
<td>20-30</td>
<td>*</td>
<td>AM</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{11}$</td>
<td>30</td>
<td>*</td>
<td>AM</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{12}$</td>
<td>20-30</td>
<td>*</td>
<td>EU</td>
<td>≤40K</td>
</tr>
</tbody>
</table>

Table 3 Intersection Attack

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q^9$</td>
<td>30</td>
<td>F</td>
<td>England</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q^{10}$</td>
<td>40</td>
<td>M</td>
<td>Canada</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{11}$</td>
<td>24</td>
<td>M</td>
<td>USA</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{12}$</td>
<td>25</td>
<td>F</td>
<td>Italy</td>
<td>≤40K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q^9$</td>
<td>10-20</td>
<td>*</td>
<td>EU</td>
<td>&gt;40K</td>
</tr>
<tr>
<td>$q^{10}$</td>
<td>10-20</td>
<td>F</td>
<td>*</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{11}$</td>
<td>20-30</td>
<td>M</td>
<td>*</td>
<td>≤40K</td>
</tr>
<tr>
<td>$q^{12}$</td>
<td>20-30</td>
<td>F</td>
<td>*</td>
<td>≤40K</td>
</tr>
</tbody>
</table>


6 Look Ahead for Distributed Recursive Diversity

6.1 Problem Definition

In this section, we show how to modify our methodology to work with distributed $\ell$-diversity assuming we have different distribution functions describing tuples with different sensitive values. However, to the best of our knowledge, there is no proposed protocol for distributed $\ell$-diversity problem. Thus, we choose to leave the practical evaluation of the theory as a future work.

While $k$-anonymity constraints the size of the equivalence classes, recursive diversity constraints on the distribution of sensitive attributes. In this section, we try to propose an extension for recursive diversity when we have only two sensitive values. We first revise our problem definition for distributed diversity.

**Definition 13** Let $A_\mu$ be the event that $\Delta_\mu(T_o)$ is recursive $(c, 2)$ diverse. Assuming we have only two values $s^1, s^2$ in the sensitive attribute domain, the problem of sufficiency in this case is to prove that, for at least one $\mu \in S_\mu$:

$$\mathbb{P}(A_\mu \mid K_{c^1, c^2}) \geq \rho$$

where $K_{c^i}$ is the distribution function for the set tuples with sensitive value $s^i$.

Note that we assume in this case, the home party collects separate distributions for each class of tuples. However by doing so, we assume independence between QI attributes with the sensitive attribute. A better approach would be to consider correlations through summary structures such as bayesian networks. We leave this challenge as a future work.

6.2 Deriving $\mu$ Probability for Recursive Diversity

The probabilistic model we construct in this section is similar to the one presented in Section 5.2. The difference is that there are now two separate distributions (one for each sensitive value $s^i$) of random tuples to buckets. Let $N_i^j$ be the number of $s^i$ tuples in $T_o$, then we have two sets of likelihood probabilities

$$N_1 = \sum_i N_i^1$$

and cardinality random variables $\sum_i X_i^1$, $\sum_i X_i^2$, each associated with one sensitive value, $\mu$ probability for recursive $(c, \ell)$ diversity can be written as

Let $Z_i$ be the event that $\max(X_i^1, X_i^2) \leq c \cdot \min(X_i^1, X_i^2)$, then

$$\mathbb{P}(A_\mu \mid K_{c^1, c^2}) \geq \rho$$

where $K_{c^i}$ is the distribution function for the set tuples with sensitive value $s^i$.

Note that we assume in this case, the home party collects separate distributions for each class of tuples. However by doing so, we assume independence between QI attributes with the sensitive attribute. A better approach would be to consider correlations through summary structures such as bayesian networks. We leave this challenge as a future work.

Since we assume QI attributes and sensitive attribute are independent, $c^1, c^2$ becomes independent variables. Thus,

$$\mathbb{P}(A_\mu \mid K_{c^1, c^2}) \geq \rho$$

where $A_\mu$ is the event that $\Delta_\mu(T_o)$ is recursive $(c, 2)$ diverse. Assuming we have only two values $s^1, s^2$ in the sensitive attribute domain, the problem of sufficiency in this case is to prove that, for at least one $\mu \in S_\mu$:

$$\mathbb{P}(A_\mu \mid K_{c^1, c^2}) \geq \rho$$

where $K_{c^i}$ is the distribution function for the set tuples with sensitive value $s^i$.

Note that we assume in this case, the home party collects separate distributions for each class of tuples. However by doing so, we assume independence between QI attributes with the sensitive attribute. A better approach would be to consider correlations through summary structures such as bayesian networks. We leave this challenge as a future work.

Since we assume QI attributes and sensitive attribute are independent, $c^1, c^2$ becomes independent variables. Thus,

$$\mathbb{P}(A_\mu \mid K_{c^1, c^2}) \geq \rho$$

where $A_\mu$ is the event that $\Delta_\mu(T_o)$ is recursive $(c, 2)$ diverse. Assuming we have only two values $s^1, s^2$ in the sensitive attribute domain, the problem of sufficiency in this case is to prove that, for at least one $\mu \in S_\mu$:

$$\mathbb{P}(A_\mu \mid K_{c^1, c^2}) \geq \rho$$

where $K_{c^i}$ is the distribution function for the set tuples with sensitive value $s^i$.

Note that we assume in this case, the home party collects separate distributions for each class of tuples. However by doing so, we assume independence between QI attributes with the sensitive attribute. A better approach would be to consider correlations through summary structures such as bayesian networks. We leave this challenge as a future work.
7 Experiments

We evaluate our approaches in three different contexts based on the data source. In Section 7.1, we work on a synthetic uniform dataset. In this set of experiments, we vary the data size and the number of equivalence classes while uniformly distributing tuples to equivalence classes. The second and third class of experiments in Sections 7.2 and 7.3 use the famous UCI Adult dataset [2]. For the second set of experiments, we use a shuffled version of the dataset to achieve attribute independence. To be more precise, keeping the attribute distributions fixed, we shuffled the values of each attribute independently. By doing so, we set our expectation on the joint probability of a given tuple to be the product of the distributions of its values. The third set of experiments is run on the original Adult dataset with correlations between the attributes.

The experiments aim to convey the accuracy of the probability approximation and the time performance with respect to different variables such as data size and the value of $k$.

7.1 Synthetic Dataset

For the synthetic dataset, we set $k$ to 100 and the number of equivalence classes to 4, which means each equivalence class will contain 25% of the total number of tuples due to the uniform distribution assumption. Figure 4 plots the results of Bonferroni Bounds (Lower and Upper) and the Approximation algorithm against the actual probability of being $k$-anonymous, when the size of the data varies. Bonferroni Bounds get more precise as the data size increases. Approximation tends to overestimate with a margin of at most 0.1, but seems to be independent of the data size.

The question ‘Why don’t we calculate the exact probability?’ is eliminated with Figure 5. The time it takes to calculate the exact probability grows exponentially with respect to the data size, whereas efficiency of Bonferroni Bounds and the Approximation seems to be independent of the data size.

Although the Bonferroni Bounds seem to yield more accurate results in less time, when we increase the number of equivalence classes, the time performance of the bounds decrease drastically as shown in Figure 6. The exact probability behaves similarly whereas the time requirement of the Approximation is independent of the number of equivalence classes.

7.2 Shuffled Adult Dataset

To test our approach, we have generated an SMC scenario (Scenario 1) similar to the one in Section 2.2. We have two parties ‘home site’ and ‘remote site’ that are willing to initiate an SMC to create a global $k$-anonymization. The ‘home site’ employs the Look Ahead of the SMC protocol with the ‘remote site’, thus approximates the local info gain (Section 3). Info gain is calculated with respect to the $\mu$-cost metric and we search for a $(1, p)$-sufficient SMC protocol. In other words, we try to look ahead to calculate the probability that
Repeating the latter step for 100 times we had 100 different
of the remote site, data itself is invisible to the home site.
The remote site which has a size equal to the home site. Al-
the data of the home site. Then, using the remaining 85% of
marital \( - \) status attribute.

To create the datasets for each site, we have partitioned
the dataset as follows: First, we have selected and removed
15% of the data (which corresponds to 4524 tuples) to form
the data of the home site. Then, using the remaining 85%
of the data, by random sampling, we have formed the data
of the remote site which has a size equal to the home site.
Although home site knows the general distribution of the data
of the remote site, data itself is invisible to the home site.
Repeating the latter step for 100 times we had 100 different
remote sites that are going to be subject to Look Ahead based
on the data of the home site. Conducting several experiments
on randomized data gives us an idea on the algorithm behavior
at the mean. To show the effect of the data size, we have
repeated the above mentioned procedure by increasing the
size of the data used for both home site and remote sites to
25% of the data (7540 tuples) and 35% of the data (10556
tuples).

Figure 7 shows the accuracy of Naive Approximation
and Randomized Naive Approximation for each data size
mentioned above and for different \( k \) values. We name the
approximation 'naive' as it does not take attribute correlations
into account, thus Naive Approximation calculates the prob-
ability on the information gain given only attribute distribu-
tions. We use the adjective 'Randomized' when the remote
site shares a distorted version of the atomic frequencies in-
stead of the actual distribution, to bound the information re-
leased by its local \( k \)-anonymization (Section 5.7). Note that
in this set of experiments, we have shuffled the data and bro-
en the correlations between the attributes.

We have applied the algorithm against every remote site
taking its corresponding home site as a base and then av-
eraged the results to be the representative result against the
actual probability. We calculate the actual probability as fol-
lows: We first combine all the local datasets to get the uni-
versal dataset (e.g., \( T_1 \)). Then we check if \( k \) anonymity
is reached with the selected mapping as described above, and
take the average of the 100 runs. In other words, we try to
find out the percentage of a positive information gain out
of 100 cases and compare it with the probability value that
the Naive Approximation returns. Experiments show that
when there is a low correlation between the attributes, our
results are spot on. For instance, for the 25% Data Case and
for \( k = 30 \), our Naive Approximation algorithm finds the
\( \mu \)-probability as 0.65 and we found that out of 100 uni-
versal datasets, 65 of them were \( k \)-anonymous with respect to
the selected mapping. This shows our probabilistic model is
very successful on predicting the probability of \( k \)-anonymity
when there is little correlation between the attributes. Al-
though it has a larger error margin in some cases, Random-
ized Naive Approximation is almost as successful as Naive
Approximation. This difference obviously stems from the
distortion in the information provided by the remote site. It
is a trade off between accuracy and privacy but results show
that we do not lose that much by disclosing less.

Figure 8 shows the time performance of the Naive Ap-
proximation for different \( k \) values. Time requirement for the
Naive Approximation is identical to the Randomized ver-
sion as the only difference between the two is the frequen-
cies of the atomic values. Thus the time requirement of the
Randomized Naive Approximation is not shown. 25% and
35% Data Cases follow similar patterns but there is no total
domination between any pair of the lines. This is because
the time taken by the algorithm does not depend directly on
data size or \( k \), instead it depends on the mapping used and
the number of equivalence classes (e.g., buckets in Section
5) in the resulting generalization.
7.3 Adult Dataset

In this set of experiments, we again use the Scenario 1, but this time we use the original Adult dataset, which has attribute correlations. We compare the results of the Naive Approximation and Correlation Aware Approximation (which we name as the CA Approximation from now on) that uses the Bayesian Network Structure to consider the effect of the correlations between attributes. To create data for the parties, we again sampled 15%, 25% and 35% of the Adult Dataset in the same manner as in Section 7.2, but this time we did not shuffle the data values.

Figure 9 shows the absolute error for each data case and for different $k$ values. It can again be inferred that there is no direct relation between neither $k$ and the absolute error nor the data size and the absolute error. We rather expect a relation between the mapping and the absolute error. We see clearly that Naive Approximation fails to capture the correlations and yields results that are off the target. For instance, in the 15% Data Case, for $k = 30$, Naive Approximation gives a $\mu$-probability of 1, but out of 100 unions of the home site with the remote sites, only 4 of them are $k$-anonymous with respect to the given mapping $\mu$. On the other hand, the CA Approximation dominates the Naive Approximation in all cases. It yields a $\mu$-probability of 0.12 for the above mentioned case and has an absolute error of 0.08,
than the data size and the $k$ values. We see that time taken depends on $k$ for each data case and for different $k$ values. We attempt to approximate the probability of being $k$ anonymous looking at a single sample from the original dataset for a fixed mapping. Our aim is to work with a larger dataset and factor out the effect of the generalization mapping on anonymity by approximating the probability that the output of the SMC will be more utilized than their local anonymizations. Experiments on real data showed that the look ahead process is perfectly accurate given non-identifying statistics on the global union.

Figure 11 shows the absolute error of CA Approximation for each data case and for different $k$ values. We see that the error rate is at most 0.17. Although both columns have similar shapes, 80% Data Case seems to be shifted to the right. That is because there is no direct relation between the $k$ value and the error rate, rather there is a relation between the number and size of equivalence classes in data and the error rate.

Finally, Figure 12 shows the time performance of CA approximation. We again see that time taken depends on the equivalence classes in the resulting generalization rather than the data size and the $k$ value, as 60% and 70% Data Cases have similar structures and 80% Data Case requires more time. Note that the time required for a CA approximation is no more than 6 seconds. Compared to the running time of a semi honest SMC protocol where execution time is generally expressed in days or weeks, the look ahead approach is quite cheap and effective.

8 Conclusion and Future Work

Most SMC protocols are expensive in both communication and computation. We introduced a look ahead approach for SMC protocols that helps involved parties to decide whether the protocol will meet the expectations before initiating it. We presented a look ahead specifically for the distributed $k$-anonymity approach. As for the look ahead process on distributed anonymization protocols, definitions of $k$-anonymity definitions can be revisited, more efficient techniques can be developed and experimentally evaluated.

References


