UBNB-PPDP: Utility-Boosting Negotiation-Based Privacy Preserving Data Publishing

M. H. Afifi, Ehab Zaghloul, Tongtong Li and Jian Ren

***Abstract*—In the era of big data and artificial intelligence, almost every aspect of research is driven by collecting data. However, privacy concerns substantially limit the usability of such data. This prevents a vast amount of possible advances in all branches of science. While privacy rules are inevitable, data owners will always seek data publishing models and techniques that can maximize data utility within the frame of the imposed privacy rules. In this paper we propose a negotiation-based data publishing model to jointly address the utility requirements of the Data User (DU) and the privacy and possibly the monetary requirements of the Data Owner (DO). We also re-define the data utility based on the DU's rather than the DO's perspective. Based on the proposed model, we present two data publishing scenarios that satisfy a given privacy constraint while achieving the DU's required data utility. The variation in a DO's flat or variable monetary rate objective motivates the data publishing scenarios. Our protocol fills the gap between the existing theoretical work and the ultimate goal of practicality.**

***Index Terms*—Data privacy, data utility, utility-privacy tradeoff, big data, healthcare systems, data publishing, data sharing.**

1. Introduction

Privacy concerns severely limit the information provided about certain sensitive attributes. Meanwhile, DOs sticking to privacy laws such as Health Insurance Portability and Accountability Act (HIPAA), favor individuals privacy over the public beneficiary. This results in a minimized utilization of the existing data. However, the main reason behind such miss-utilization is the lack of data publishing techniques that provide a satisfactory tradeoff between privacy and utility of the published data. Data utility inevitably conflicts with data privacy. From the data utility perspective, it is best to publish a dataset as is, while from the perspective of data privacy, it is best to publish a mostly generalized dataset.

In some scenarios, a data recipient is in crucial need for certain attributes of interest for some decision making problems. A data recipient is just interested in subset of the dataset that contains the attributes of interest. These attributes will help the data recipient in making the correct decision. Consider the example of preventive prophylactic surgeries that remove an organ or gland that shows no signs of cancer in an attempt to prevent high risk individuals from developing the disease. If a patient obtains accurate information of the infection risks, she would then be able to evaluate the risks and take suitable precautions.

Privacy related incidents have urged a demand for extensive research in privacy notions for data publishing and analysis, such as k-anonymity, ℓ-diversity and t-closeness, to name a few [1]–[3]. A table satisfies k-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of quasi-identifier attributes. However, it is insufficient to prevent attribute disclosure with side information. To deal with this issue, ℓ-diversity was introduced in [2], which requires that the sensitive attributes contain at least ℓ well-represented values in each equivalence class. Unfortunately, it is possible to acquire knowledge of a sensitive attribute from its generally available global distribution. t-closeness [3] requires the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the whole table. The differential privacy [4] learns whether or not an individual is included in the database. The strict definition of this model makes it impractical as a metric for data publishing [5]. These models only consider minimizing the amount of privacy loss without considering the utility-privacy tradeoff problem. The existing approaches of modeling the tradeoff and consequently the data publishing techniques are mostly either inapplicable or just intuitive rather than practical.

In this paper, based on our privacy characterization and quantification framework presented in [6], we propose a practical data publishing model that redefines data utility as a function of the DU's requirements, namely, attributes of interest. The model shown in Fig. 1 incorporates a negotiation process between the DO and the DU in order to reach a data publishing deal. The DU represents her requirements as utility patterns of the attributes of interest while the DO's requirements are represented as generalization policies. Based on this model we propose two Utility-Boosting Negotiation-Based Privacy Preserving Data Publishing (UBNB-PPDP) protocols that are guided by the DO's objective. The protocols provide a set of rules between the DO and the DU in order to reach a data publishing deal. The first protocol manages a negotiation process to publish any generalized dataset that matches the DU's utility requirements and meanwhile satisfies the DO's privacy constraints. In the second protocol the DO links the utility of the published dataset to a profit function. Our proposed framework can serve as an enabler to address the privacy issues in big data publishing [7], [8]. We contend that having a solid framework that manages the data publishing enables the big data owner to determine the trade-off between data utility and privacy loss.

The rest of this paper is arranged as follows. Section

978-1-5386-4727-1/18/$31.00 ©2018 IEEE

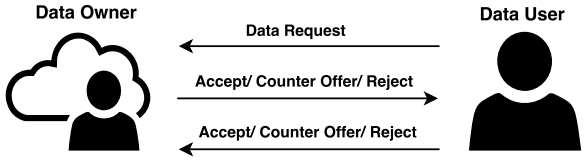


Figure 1: Negotiation-based data publishing model.

2 presents some preliminaries. Section 3 introduces the exploited utility and privacy metrics. Our UBNB-PPDP model is proposed in Section 4. In Section 5, two versions of the UBNB-PPDP protocol are formally presented. Section 6 provides the simulation results. We conclude in Section 7.

2. Preliminaries

System Model Data is usually released in the format of tables, where the rows are the records of the individuals and columns are their corresponding attributes. Some attributes can uniquely identify the individuals such as the social security or the driving license numbers. These attributes are referred to as *explicit-identifiers*. Some of the attributes are non-sensitive. These attributes are generally referred to as *quasi-identifiers*, which may include information such as Zip-Code, Age, and Gender. Sensitive attributes may include information such as disease and salary. When datasets are published, all explicit-identifiers are removed. In the proposed model, we assume no collusion attacks. Published data is protected by copyright. Data publishing to a user is not a grant of ownership.

Notations Let $U = \{u_n\}_{n=1}^{N}$ be N individuals participating in the data table T, $\mathcal{A} = \{A_l\}_{l=1}^{L}$ be the set of L attributes and $u_n[A_l]$ be the value of attribute A_l for individual u_n. We denote the set of *quasi-identifiers* as $\mathsf{QID} \subset \mathcal{A}$. We assume that any individual in a given table T only owns one record. Thus we, interchangeably, use the notation u_n to represent the record or the record owner. We also assume that any record is represented as a function of multi-variables $V = \{v_l\}_{l=1}^{L}$, where V corresponds to the set of attributes $\mathcal{A} = \{A_1, A_2, \cdots, A_L\}$ in the original dataset. The order of each variable v_l, denoted as $\mathrm{ord}(v_l) = |v_l|$, is the number of all possible attribute values.

Table Generalization To satisfy the privacy constraints, data publishing techniques apply some generalizations to the quasi-identifiers $\mathsf{QID}s$ to avoid linking individuals to records in the table. Any value in the original table is mapped to a generalized value in the published table following a certain mapping function. Records are generalized and represented as functions of multi-variables $V' = \{v'_l\}_{l=1}^{L}$, where V' is the generalization of V. The order of each generalized variable v'_l is defined as $\mathsf{ord}(v'_l) = |v'_l|$. After generalization, different combinations of v'_l's in the published Table T' naturally divide the table into a set $\mathcal{C} = \{[C_q]\}_{q=1}^{Q}$ of Q equivalence classes. Let $S = \{s_i\}_{i=1}^{m}$ be the set of all m attribute values of a sensitive attribute $S \in \mathcal{A}$. The estimated initial distribution of S for equivalence class $[C_q]$ is given as $a_{S,[C_q]} = (a_1, a_2, \cdots, a_m)$. The published distribution of S in an equivalence class $[C_q]$ is given as $x_{S,[C_q]} = (x_1, x_2, \cdots, x_m)$. Throughout the rest of this paper, we denote $a_{S,[C_q]}$ as a and $x_{S,[C_q]}$ as x.

3. The Utility and Privacy Metrics

3.1. The Utility Loss Metric

It is trivial that from the DU's perspective, an optimal published table is a table with the number of classes equal to the number of individuals in the original dataset. However, being subject to the privacy rules, the published dataset loses utility in the process of generalizing the original table. In order to quantify the utility loss resulting from data generalization, we introduce a utility loss metric that accumulates the loss of each individual in a class to find the utility loss for all classes in the published table.

Generally, any privacy-preserving data publishing technique modifies the original dataset into a set of equivalence classes. The exploited utility loss metric computes the Euclidean distance between the published distribution x of the sensitive attribute at each class and the original conditional distributions b of the sensitive attribute for each individual given these classes. That is, it typically measures the deviation in the DU's belief about the sensitive attribute value of a certain individual between two cases. In the first case, access to the original dataset is granted. In the second case, the DU only has access to the published dataset which is the typical case. Thus, the original distribution b of an individual is a vector of zeros except for a one at the sensitive attribute value.

For an individual u_n belonging to an equivalence class $[C_q]$, the utility loss of attribute S given $[C_q]$ is defined as the Euclidean distance between the two distributions b and x. To measure the utility loss, we sum the loss of individuals in the class. The loss can be defined as follows.

Definition 1 (Utility Loss $\mathscr{L}_U(S, [C_q])$). The utility loss of an individual u_n in a certain class $[C_q]$ as a result of table generalization. Assuming that the number of individuals in a class $[C_q]$ is denoted as c_q. The utility loss is defined as

$$\mathscr{L}_U(S, [C_q]) = \frac{1}{c_q} \sum_{j=1}^{c_q} \sqrt{\sum_{i=1}^{m} (b_i - x_i)^2}.$$

The total utility loss is defined as

$$\mathscr{L}_U(S, T') = \frac{1}{N} \sum_{q=1}^{Q} \sum_{j=1}^{c_q} \sqrt{\sum_{i=1}^{m} (b_i - x_i)^2}.$$

Since this metric will be exploited in our model to quantify the utility loss, it is useful to have the following thresholding definition to express each entity's constraints.

Definition 2 (γ-Utility Loss). A published table T' has a γ-utility loss if $\mathscr{L}_U(S, \mathcal{C}) \leq \gamma$ for the set of all equivalence classes. That is, $\max(\mathscr{L}_U(S, [C_q])) \leq \gamma, \;\; q = 1, 2, \cdots, Q$.

3.2. The Privacy Loss Metric

Our approach to quantify privacy depends on quantifying the information loss between two adversary's states of knowledge. At the first state, based on public information of sensitive attribute's distribution, an adversary has some prior

belief about the sensitive attribute value of an individual. This prior belief a is based on the probability distributions of attributes and joint distributions of their combinations. After publishing, an adversary moves to the second state of knowledge that is, the posterior belief x, that is the conditional distribution of sensitive attribute given combinations of published attributes.

As a result of data publishing, the adversary gains more information about the individual from the posterior belief. This amount of information is the loss that we need to capture where it enables us to measure the extent to which this data publishing model leaks privacy. We are generally interested in $a_{S,[C_q]}$ and $x_{S,[C_q]}$, that are the prior and posterior beliefs of an adversary about a sensitive attribute S given a combination of generalized attributes represented in the specific class $[C_q]$ they form. We believe that matching the published distribution $x_{S,[C_q]}$ to the original estimated distribution $a_{S,[C_q]}$ would indeed achieve better privacy. Therefore we give the following definition.

Definition 3 (**Privacy Loss** $\mathscr{L}_P(S,[C_q])$)**.** For an individual u_n in an equivalence class $[C_q]$, the privacy loss of attribute S given an equivalence class $[C_q]$ is defined as the Euclidean distance between the two distributions a and x

$$\mathscr{L}_P(S,[C_q]) = \sqrt{\sum_{i=1}^{m}(a_i - x_i)^2}.$$

The privacy loss metric measures the overall divergence of attribute values distribution from one state to the other. When publishing a table T', it is optimum to maintain the same original distribution over the set of equivalence classes. That is, distribution loss $\mathscr{L}_P(S,\mathcal{C})$ is desired to be zero. However it is natural that the distance between distributions will change. This change contributes to the privacy loss. The DO's objective is to keep the privacy loss below a predetermined level.

Definition 4 (ε-**Privacy Loss**)**.** A published table T' has an ε-privacy loss if $\mathscr{L}_P(S,\mathcal{C}) \leq \varepsilon$ for the set of all equivalence classes. That is $\max(\mathscr{L}_P(S,[C_q])) \leq \varepsilon,\ \ q = 1,2,\cdots,Q$.

4. The Proposed UBNB-PPDP Model

The proposed negotiation-based data publishing model manages a negotiation process where a set of communication sessions is held between the DO and the DU in order to set a data publishing deal. More specifically, it redefines the data utility and adjusts the publishing rules accordingly. The ultimate goal of the proposed model is to provide the DU with the expected data utility while satisfying the privacy rules of the DO. Therefore, the proposed model reformulates the tasks of entities in order to express their needs.

The DU might give more priority to certain attributes than others. In order to boost the data utility, the DU seeks not just the highest possible data utility in general, but rather the highest possible data utility for the attributes of interest. To express these needs, the data utility is manifested as the DU's attributes of interest divided into levels with different priorities. These priorities can be represented in a requirements priority diagram namely, the utility pattern.

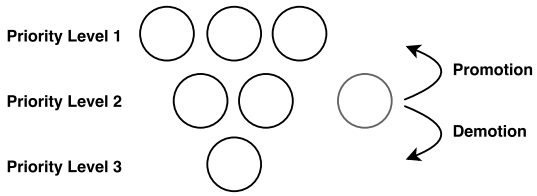


Figure 2: Utility pattern.

Definition 5 (**Utility Pattern** $\mathcal{U}$)**.** A layered hierarchy that ranks the DU's attributes of interest. The attributes with highest priority level are located at the top layers while the ones with the least priority level are located at the bottom.

The utility pattern can be modified by either ***promotion*** or ***demotion*** of different attributes according to their priority. As shown in Fig. 2, the promotion operation is done by the DU during the negotiation process in order to upgrade the priority level of an attribute of interest. On the other hand, the demotion operation is done by the DU in order to degrade the priority level of an attribute of interest.

For example, if the DU is a medical researcher that wants to evaluate the effect of some demographics such as Age, Gender, Zip-code, Race, Nationality and Occupation on a disease such as Breast Cancer. In the first negotiation attempt, the DU will have all the attributes of interest in the first priority level. If publishing the table with these attributes of interest as is satisfies the DO's privacy constraints, the deal is set. Otherwise, the DU will have to rearrange the priority level orders in the second negotiation attempt by the demotion of some attributes. For instance, in our example, the researcher might have an essential need to more specific data in terms of Age, Gender and Race as compared to other attributes. Therefore, other attributes can be demoted to the second priority level in the second negotiation attempt.

In response to a DU's requested utility pattern $\mathcal{U}$, the DO recommends a generalization policy that matches $\mathcal{U}$ in compliance with the privacy constraint ε. For all possible generalizations $\mathcal{G}$ satisfying ε, we define this generalization policy as follows.

Definition 6 (**Generalization Policy** $\mathcal{P}$)**.** A policy $\mathcal{P} \in \mathcal{G}$ proposed by the DO as a response to the utility pattern $\mathcal{U}$ requested by the DU. This policy comprises the generalization values of each attribute of interest in $\mathcal{U}$.

Consider a utility pattern $\mathcal{U}$ with d attributes of interest $\{v_1, v_2, \cdots, v_d\}$. The generalization policy is a mapping function $f: v \rightarrow v'$ that maps any attribute v to a generalized attribute v'. For example the Age in Table 1a consists of 12 values corresponding to the ages of 12 individuals. These values are mapped to 3 classes in Table 1b.

The generalization policy recommended by the DO will not essentially satisfy the DU's expected data utility. Therefore, using our proposed utility pattern, generalization policy, utility and privacy loss metrics, both the DO and the DU go through a negotiation process to set the data publishing deal. To decide whether an offer can be accepted or not, the DO computes the privacy and utility losses in order to check if the response $\mathcal{P}$ to a DU's offer satisfies the requirements and constraints (ε and γ).

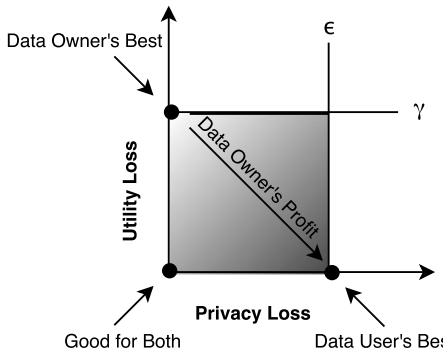


Figure 3: Diagram showing tradeoff and DO's profit.

Suppose that for a given privacy threshold ε, there exist g possible generalizations $\mathcal{G}$ that satisfy the privacy constraint. The ultimate goal is to find, out of these generalizations, the one that satisfies the DO's objective. After satisfying this objective, the DO now has a recommended generalization policy that is guided by the DU's utility pattern and meanwhile satisfies the objective. For a given ε, the DO has two possible objectives. The first is to publish the data with the best (or any) utility corresponding to the least (or any) possible generalization within the privacy threshold. The second is to control the utility loss of the published data for the purpose of linking it to a profit function.

The first proposed scenario is publishing the dataset that satisfies the privacy constraints disregarding how much utility loss it provides as long as it satisfies the DU's requested utility pattern $\mathcal{U}$. We name this the ***Flat Rate Objective***. Another proposed scenario is linking the data utility loss to a profit function. That is, for a requested DU's utility $\mathcal{U}$ and privacy constraint ε, there exist g possible generalizations $\mathcal{G}$ that satisfy the privacy constraint. Each of these generalizations provides a different level of data utility loss $\mathscr{L}_U(S,[C])$. As shown in Fig.3, the profit decreases with the increase in the data utility loss. We name this the ***Variable Rate Objective***. The profit is hence a function of the data utility loss,

$$\mathsf{Profit} = f(\mathscr{L}_U(S,[C])).$$

The profit function is typically determined by the DO depending on the data value in the market.

5. The Proposed UBNB-PPDP Protocol

In this section we describe two variants of the proposed UBNB-PPDP protocol. Based on the proposed model, the protocol relies on a negotiation process between the DO and the DU guided by both entities' needs and expectations.

Let $z = 1, \cdots, Z$ represent the z^{th} negotiation session between the DO and the DU where Z is the maximum number of negotiation sessions. Throughout the negotiation process, DU modifies the requested utility pattern $\mathcal{U}_z$ by either promoting or demoting different required attributes according to their priority level. Also the DO modifies the generalization policy $\mathcal{P}_z$ through modifying the mapping function by either increasing or decreasing the range by which each attribute is generalized. This continues until both entities set on a data utility level that matches the DU requirements on one hand and a data privacy level that satisfies the DO constraints on the other hand. We note that both entities can terminate the negotiation at any time.

Algorithm 1 Flat Rate UBNB-PPDP.

1) DU sends a request $\mathcal{U}_z$ to the DO.
2) DO responds with any $\mathcal{P}_z \in \mathcal{G}$ that matches the requested $\mathcal{U}_z$ and satisfies $\mathscr{L}_P(S,\mathcal{C}) \leq \varepsilon$.
3) If satisfied by $\mathcal{P}_z$, DU responds with an *offer accept*.
4) Otherwise, DO responds with another $\mathcal{P}_{z+1} \in \mathcal{G}$ as a counter offer.
5) If no $\mathcal{P}_{z+1}$ exists, DO responds with an *offer reject* or a *negotiation termination*.
6) If the response is an *offer reject*, if interested, the DU sends a new relaxed request $\mathcal{U}_{z+1}$. Otherwise, DU does a *negotiation termination*.
7) Repeat steps 2, 3, 4, 5, and 6.

5.1. The Flat Rate UBNB-PPDP

The DU submits an ***offer*** by sending a requested data utility pattern $\mathcal{U}_z$. If the DO accepts the requested $\mathcal{U}_z$ as is, the DO sends **any** generalization policy $\mathcal{P}_z$ that satisfies the privacy constraint $\mathscr{L}_P(S,\mathcal{C}) \leq \varepsilon$. The DU then reviews $\mathcal{P}_z$ and either responds with an ***offer accept*** to make a deal or an ***offer reject*** to refuse it and sends a modified utility pattern if interested. If no $\mathcal{P}_z$ that matches the ε is found for the requested $\mathcal{U}_z$, the DO refuses $\mathcal{U}_z$. The DU either modifies the utility pattern and presents a new offer or does a ***negotiation termination*** due to the failure of reaching a suitable deal. The flat rate UBNB-PPDP protocol is summarized in Algorithm 1.

5.2. The Variable Rate UBNB-PPDP

In the variable rate UBNB-PPDP scenario, the DU submits an ***offer*** by sending a requested $\mathcal{U}_z$. The DO finds the set $\mathcal{G}$ of the generalization policies that matches the requested $\mathcal{U}_z$ and meanwhile satisfies $\mathscr{L}_P(S,\mathcal{C}) \leq \varepsilon$. The DO can either accept the requested $\mathcal{U}_z$ as is, or refuse it. In the first case, if she accepts, the DO computes the utility loss $\mathscr{L}_U(S,\mathcal{C})$ and the $Profit$. The DO then sends an **optimized** $\mathcal{P}_z$ that satisfies the privacy constraint $\mathscr{L}_P(S,\mathcal{C}) \leq \varepsilon$ and matches the expected profit. The DU then reviews $\mathcal{P}_z$ and either responds with either ***offer accept*** to make a deal or ***offer reject*** to refuse it and sends a modified utility pattern if interested. In the second case, if the DO refuses due to the non-existence of a generalization policy that matches the privacy constraint for the requested $\mathcal{U}_z$, the DU either modifies the utility pattern and presents a new offer or does a ***negotiation termination***. The variable rate UBNB-PPDP protocol is summarized in Algorithm 2.

We note that a flat rate objective saves computations at the DO's side where the DO is not required to optimize the generalization process to be linked to the profit function. However, this does not have any guarantees about neither achieving the best possible data utility for the DU nor the best possible profit for the DO.

Algorithm 2 Variable Rate UBNB-PPDP.

1) DU sends a request $\mathcal{U}_z$ to the DO.
2) DO finds the set $\mathcal{G}$, computes $\mathscr{L}_U(S,\mathcal{C})$ and the Profit, and responds with an optimized $\mathcal{P}_z$.
3) DU reviews $\mathcal{P}_z$ and either responds with an *offer accept* or *offer reject* and sends a modified $\mathcal{U}_{z+1}$ if interested.
4) If no $\mathcal{P}_z \in \mathcal{G}$ exists, DO responds with an *offer reject* or a *negotiation termination*.
5) If the response is an *offer reject*, if interested, the DU sends a new relaxed $\mathcal{U}_{z+1}$. Otherwise, DU does a *negotiation termination*.
6) Repeat steps 2, 3, 4 and 5.

TABLE 1: Impatient Micro-data

(a) Original Dataset

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Virus Infection
4	13053	23	American	Virus Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Virus Infection
8	14850	49	American	Virus Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

(b) 4-anonymous Impatient Micro-data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Virus Infection
4	130**	<30	*	Virus Infection
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart Disease
7	1485*	≥40	*	Virus Infection
8	1485*	≥40	*	Virus Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

6. Empirical Analysis and Simulation Results

6.1. Empirical Analysis

In this subsection we introduce an empirical example to help understand the capabilities of the utility and privacy metrics in evaluating the utility and privacy losses of published classes.

Example 1. In the example from [2], the original impatient dataset is given in Table 1a and the 4-anonymous impatient dataset is given in Table 1b. The probability distribution for the three diseases is $\left(\frac{3}{12}, \frac{4}{12}, \frac{5}{12}\right)$. In this case, the privacy loss $\mathscr{L}_P(S,\mathcal{C})$ for individuals within the first, second and third equivalence classes is $[0.5137, 0.2357, 0.7619]$, while the utility loss $\mathscr{L}_U(S,\mathcal{C})$ is $[0.707, 0.77, 0]$ respectively.

Our findings for Table 1b reveal that patients under 30 have Heart-Disease or Virus-Infection with equal probability. The scheme provides $\mathscr{L}_P(S,[C_1]) = 0.51$ and $\mathscr{L}_U(S,[C_1]) = 0.707$. For patients over 40, $\frac{1}{4}$ have Cancer, $\frac{1}{4}$ have Heart-Disease and $\frac{1}{2}$ have Virus-Infection. The scheme provides $\mathscr{L}_P(S,[C_2]) = 0.23$ and $\mathscr{L}_U(S,[C_2]) = 0.77$. Finally, patients in their 30s, all have Cancer. The scheme returns $\mathscr{L}_P(S,[C_3]) = 0.76$ and $\mathscr{L}_U(S,[C_3]) = 0$.

TABLE 2: Description of Adults Database

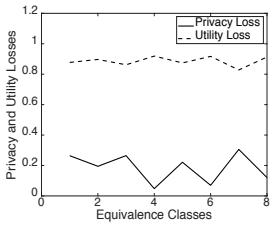
	Attribute	Type	Domain Size	Height
1	Age	Numeric	74	4
2	Work Class	Categorical	7	2
3	Education	Categorical	16	3
4	Country	Categorical	41	2
5	Marital Status	Categorical	7	2
6	Race	Categorical	5	1
7	Gender	Categorical	2	1
8	Occupation	Sensitive	14	

6.2. Simulation Results

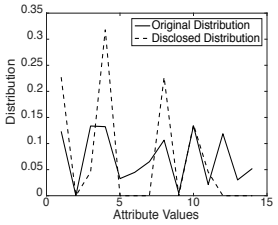
Simulation results give us an insightful understanding of utility and privacy losses and how the negotiation can be handled in our proposed UBNB-PPDP protocol. Specifically, the DO is able to analyze the utility and privacy loss for different combinations of QIDs and interpret the losses in each class to determine which classes leak more privacy or provide more utility. Simulations are done on a sample of the US census dataset from the UC Irvine machine learning repository [9]. After eliminating records with missing values, we have a total of 30,162 records. Following the work in [2], as shown in Table 2, we utilize only 8 attributes, 7 of which form the set of possible quasi-identifiers while Occupation is the sensitive attribute. We adopt the Incognito algorithm [10] for generating the generalized tables. The number of quasi-identifiers QIDs is represented by the variable l that takes values from 1 to 7 with the same order in Table 2.

We start by considering a published table satisfying 0.5-closeness, 6-diversity, and 6-anonymity at $l = 2$. Quasi-identifiers are chosen to be Age and WorkClass where QID $= (1, 2)$. From the results shown in Fig. 4a, an observed instance has a considerably high privacy loss for individuals in $[C_7]$. To further understand the reason behind this loss, we refer to Fig. 4b showing the distribution of the Occupation in the original table versus the distribution at this specific class after publishing. It is obvious that $[C_7]$ has some unrepresented attribute values. Hence, an observer can eliminate these values and thus gain an increased confidence about the Occupation of the individual of interest. Also as shown in Fig. 4a, this class has the least utility loss. This is also justifiable by the fact that the DU gains a high level of certainty about the sensitive attribute values of individuals in this class where only 6 out of 14 sensitive attribute values are represented. Thus, the DO can use the privacy and utility metrics to manage the negotiation process. In particular, the DO can control the privacy and utility losses of different classes based on the threshold values and the expected profit.

For the requested DU's utility patterns $\mathcal{U}$s, the DO can analyze the utility and privacy losses of any dataset



(a) Privacy and Utility Losses



(b) Original Vs. published distribution of class 7

Figure 4: Evaluation of a table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $l = 2$

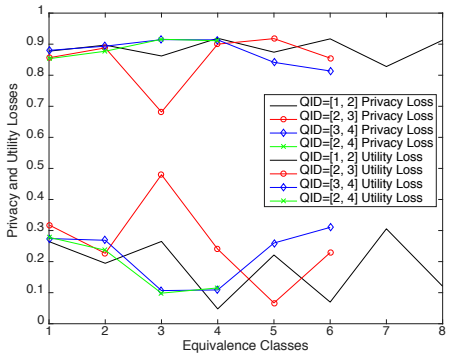


Figure 5: Privacy and utility losses at different sets of QIDs

generalization before publishing. For example, in Fig. 5, we compare privacy and utility losses of four published tables at $l = 2$ while choosing a different combination of quasi-identifiers for each table. Quasi-identifiers are chosen to be $\mathsf{QID} = [(1,2),(2,3),(2,4),(3,4)]$.

To satisfy the privacy constraints, for different requested utility patterns with different attributes of interest, the generalization would decrease the number of classes in the published table and hence, the data utility decreases. Fig. 5 also illustrates the number of classes Q at each chosen combination of QIDs and different levels of privacy and utility loss for each class. The generalization ended up with 8 classes at $\mathsf{QID} = (1,2)$, 6 classes at $\mathsf{QID} = [(2,3),(3,4)]$, and 4 classes at $\mathsf{QID} = (2,4)$. Depending on the sensitivity of different classes formed by different combinations of QIDs, the DO can select the generalization that achieves the desirable privacy level for a requested $\mathcal{U}$. For instance, if a DO is more concerned about privacy of certain users that fall into $[C_3]$, then choosing $\mathsf{QID} = [2,3]$ would leak considerable private information. Also for these specific QIDs the class $[C_3]$ has a low utility loss level. Hence, according to the metrics, a DO can not only design the suitable data publishing technique for all individuals in a dataset, but also for a subset of them.

Let us also consider 3 versions of a published table T', with different privacy loss levels, at $l = 3$, 5 and 7. As shown in Fig. 6, QIDs are chosen to be $\mathsf{QID} = [(1,2,3),(1,2,3,4,6),(1,2,3,4,6,7,8)]$. This is useful where we can see the tradeoff between the utility

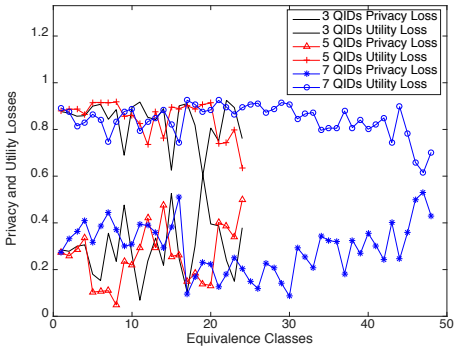


Figure 6: Utility-privacy tradeoff for different attributes of interest at $l = 3, 5$ and 7

and privacy and how different utility patterns can affect the published data utility levels for a given privacy constraint.

7. Conclusion

In this paper, we introduced two data utility and privacy loss metrics. Using these metrics, we were able to practically address the utility-privacy tradeoff problem. We then propose a utility-boosting privacy-preserving data publishing model that redefines the data utility based on the DU's perspective. Based on this model we incorporate our utility and privacy metrics to propose two versions of a privacy-preserving data publishing protocol. The protocol sets rules for the negotiation between the DO and the DU in order to set a data publishing deal. The proposed protocol inherently boosts the data utility from the DU's perspective with the satisfaction of the DO's privacy constraint and monetary objectives.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, Mar. 2007.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, pp. 106–115, 2007.

[4] C. Dwork., "Differential privacy," *ICALP*, 2006.

[5] G. Cormode, "Individual privacy vs population privacy: Learning to attack anonymization," *CoRR*, vol. abs/1011.2511, 2010.

[6] M. H. Afifi, K. Zhou, and J. Ren, "Privacy characterization and quantification in data publishing," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018.

[7] I. S. Rubinstein, "Big data: The end of privacy or a new beginning?," *N.Y.U. Public Law and Legal Theory Working Papers*, 2012.

[8] I. Kerr and J. Earle, "Prediction, preemption, presumption: How big data threatens big picture privacy," *66 Stan. L. Rev. Online 65*, 2013.

[9] A. Asuncion and D. Newman, "Uci machine learning repository, http://www.ics.uci.edu/ mlearn/ml-repository.html, 2007,"

[10] K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," *In Proceedings of ACM SIGMOD*, pp. 49–60, 2005.