ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Secure Collaborative Data Publishing by Using Computation Trusted Third-Party Protocols

Karthikeyan.B[*1], R. Augustian Issac[2]
[*1,2] Assistant Professor, Computer Science Engineering, SRM university, Chennai, India
karthibalacse@gmail.com

Abstract

In this paper the collaborative data publishing issue for horizontally Partitioned data at different data providers was discussed. The attackers by colluding data from multiple providers, who use their own data to publish, were considered by implementing the two methods to improve the privacy constraints and secure in data publish. Initially, the Hidden Markov Models (HMM) is used to analysis the non-authenticated user who are trying to access the data is implemented to overcome the m-privacy notations which doesn't satisfies the privacy constraints against the group of colluding the different data providers in existing design , Second the Attribute based anonymity for preserving privacy is used for checking strategies and adaptive privacy of individuals in organisation's publishing of data was implemented.

Keywords: Data publishing, data providers, Hidden Markov Models

Introduction

The data sharing and data publishing in data mining technologies is rapidly developed nowadays, the personal information are shared in the distributed databases for research purposes. Consider this example in a healthcare domain the mail goal is to develop the nationwide health information network to share the data among the several hospitals and data providers. In data publishing the privacy of individual must preserved. The data provider (e.g., hospital) publishes a clean version of the data, at the same time providing utility for data users who are doing research, and privacy security for the individuals represented in the data (e.g., patients). When data are gathered from multiple data providers or data owners, two main settings are used for anonymization; one approach is for each provider to anonymize the data separately, which results in prospective loss of incorporated data utility. A more attractive approach is collaborative data publishing which anonymizes data from all providers as if they would come from one source using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols.

A. Problem setting

A typical scenario for data collection and publishing is described in Figure 1. In the data collection phase, the data publisher collects data from record owners (e.g., Alice and Bob). A typical scenario for data collection and publishing is described in Figure 1. In the data collection phase, the data publisher collects data from record owners (e.g., Alice and Bob). In the data publishing phase, the data publisher releases the collected data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data. In this survey, data mining has a broad sense, not necessarily restricted to pattern mining or model building. For example, a hospital collects data from patients and publishes the patient records to an external medical center. In this example, the hospital is the data publisher, patients are record owners, and the medical center is the data recipient. The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis. Publish data, not the data mining result. PPDP emphasizes publishing data records about individuals (i.e., micro data). Clearly, this requirement is more stringent than publishing data mining results, such as classifiers, association rules, or statistics about groups of individuals. For example, in the case of the Netflix data release, useful information may be some type of associations of movie ratings. However, Netflix decided to publish data records instead of such associations because the participants, with data records, have greater flexibility in performing the required analysis and data exploration, such as mining patterns in one partition but not in other

ISSN: 2277-9655
Impact Factor: 1.852

partitions; visualizing the transactions containing a specific pattern; trying different modelling methods and parameters, and so forth.

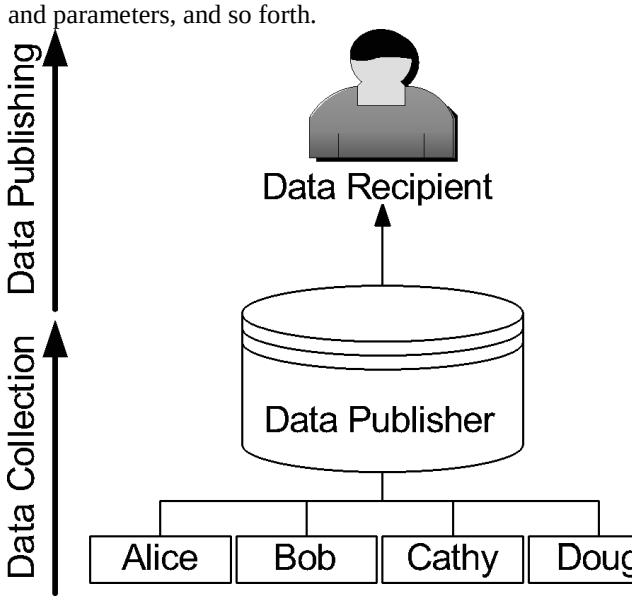


Fig: 1 Data collection and data publishing.

There are two models of data publishers. In the untrusted model, the data publisher is not trusted and may attempt to identify sensitive information from record owners. And statistical methods were proposed to collect records anonymously from their owners without revealing the owners' identity. In the trusted model, the data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher; however, the trust is not transitive to the data recipient. In this survey, we assume the trusted model of data publishers and consider privacy issues in the data publishing phase.

Data Publishing With Privacy-Preserving

In practice, every data publishing scenario has its own assumptions and requirements of the data publisher, the data recipients, and the data publishing purpose. The following are several desirable assumptions and properties in practical data publishing.

The data publisher is not required to have the knowledge to perform data mining on behalf of the data recipient. Any data mining activities have to be performed by the data recipient after receiving the data from the data publisher. Sometimes, the data publisher does not even know who the recipients are at the time of publication, or has no interest in data mining. For example, the hospitals in India publish patient records on the Web the hospitals do not know who the recipients are and how the recipients will use the data. The hospital publishes patient records because it is required by regulations or because it supports general medical research, not because the hospital needs the result of data mining. Therefore, it is not reasonable to expect the data publisher to do more than anonymize the data for publication in such a scenario. In other scenarios, the data publisher is interested in the data mining result, but lacks the in-house expertise to conduct the analysis, and hence outsources the data mining activities to some external data miners. In this case, the data mining task performed by the recipient is known in advance. In the effort to improve the quality of the data mining result, the data publisher could release a customized data set that preserves specific types of patterns for such a data mining task. Still, the actual data mining activities are performed by the data recipient, not by the data publisher.

The data recipient could be an attacker. In PPDP, one assumption is that the data recipient could also be an attacker. For example, the data recipient, say a drug research company, is a trustworthy entity; however, it is difficult to guarantee that all staff in the company is trustworthy as well. This assumption makes the PPDP problems and solutions very different from the encryption and cryptographic approaches, in which only authorized and trustworthy recipients are given the private key for accessing the clear text. A major challenge in PPDP is to simultaneously preserve both the privacy and information usefulness in the anonymous data.

The assumption for publishing data and not the data mining results, is also closely related to the assumption of a non-expert data publisher. For example, Netflix does not know in advance how the interested parties might analyse the data. In this case, some basic "information nuggets" should be retained in the published data, but the nuggets cannot replace the data.

In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life. Consider the example of patient records. The pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug. If a published record does not correspond to an existing patient in real life, it is difficult to deploy data mining results in the real world. Randomized and synthetic data do not meet this requirement. Although an encrypted record corresponds to a real life patient, the encryption hides the semantics required for acting on the patient represented.

Centralized Anonymization

The data publisher has a table of the form D(Explicit Identifier, Quasi Identifier, Sensitive

ISSN: 2277-9655
Impact Factor: 1.852

Attributes, Non-Sensitive Attributes),where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; Quasi Identifier (QID) is a set of attributes that could potentially identify record owners; Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories. The four sets of attributes are disjoint. Most works assume that each record in the table represents a distinct record owner.

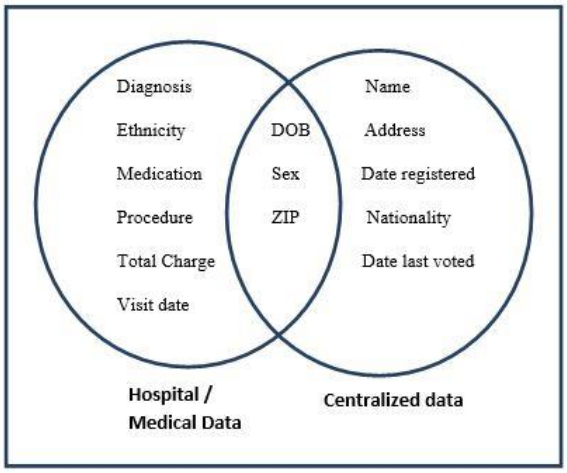


Fig: 2 linking to re identify record owner [Robert]

Anonymization refers to the PPDP approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed. Even with all explicit identifiers being removed, In Robert's example, an individual's name in a public voter list was linked with his record in a published medical database through the combination of zip code, date of birth, and sex, as shown in Figure 2. Each of these attributes does not uniquely identify a record owner, but their combination, called the quasiidentifier, often singles out a unique or a small number of record owners. According to Robert, the most of the people had reported characteristics that likely made them unique based on only such quasi-identifiers.

In the above example, the owner of a record is reidentified by linking his quasiidentifier to perform such linking attacks; the attacker needs two pieces of prior knowledge: the victim's record in the released data and the quasi-identifier of the victim. Such knowledge can be obtained by observation. For example, the attacker noticed that his boss was hospitalized, and therefore knew that his boss's medical record would appear in the released patient database. Also, it was not difficult for the attacker to obtain his boss's zip code, date of birth, and sex, which could serve as the quasi-identifier in linking attacks.

A. QID - Sensitive Attributes, Non-Sensitive Attributes, Individual Attribute based anonymity

QID is an anonymous version of the original QID obtained by applying anonymization operations to the attributes in QID in the original table D. Anonymization operations hide some detailed information so that several records become indistinguishable with respect to QID. Consequently, if a person is linked to a record through QID, that person is also linked to all other records that have the same value for QID, making the linking ambiguous. The anonymization problem is to produce an anonymous T that satisfies a given privacy requirement determined by the chosen privacy model and to retain as much data utility as possible. Information metric is used to measure the utility of an anonymous table. Note that the Non-Sensitive Attributes are published if they are important to the data mining task.

B. Attack and Privacy Models in Record linkage

Consider the integrated raw patient data in Table I (ignore Parties A, B, and C for now), where each record represents a surgery case with the patient-specific information. Job, Sex, and Age are quasi-identifying attributes. Hospitals want to release Table I to the BTS for the purpose of classification analysis on the class attribute, Transfuse, which has two values, Y and N, indicating whether or not the patient has received blood transfusion. Without a loss of generality, we assume that the only sensitive value in Surgery is Transgender. Hospitals express concern on two types of privacy threats.

Table I. Patient Data

		Quasi-identifier (QID)			Class	Sensitive
	ID	Job	Sex	Age	Transfuse	Surgery
Party A	1	Engineer	M	32	Y	Transgender
	2	Janitor	M	35	N	Plastic
	3	Doctor	F	28	N	Urology
Party B	4	Actor	M	23	Y	Vascular
	5	Electrician	M	34	N	Transgender
	6	Janitor	M	52	Y	Urology
Party C	7	Professor	F	33	Y	Urology
	8	Lawyer	F	26	Y	Vascular
	9	Carpenter	M	44	N	Plastic

—Identity linkage. If a record in the table is so specific that not many patients match it, releasing the data may lead to linking the patient's record and, therefore, her received surgery. Suppose that the adversary knows that the target patient is a Mover and his age is 34. Hence, record #5, together with his sensitive value (Transgender in this case), can be uniquely identified since he is the only Mover who is 34 years old in the raw data.

—Attribute linkage. If a sensitive value occurs frequently together with some QID attributes, then the sensitive information can be inferred from such

ISSN: 2277-9655
Impact Factor: 1.852

attributes even though the exact record of the patient cannot be identified. Suppose the adversary knows that the patient is a male of age 34. Even though there exist two such records (#1 and #5), the adversary can infer that the patient has received a Transgender surgery with 100% confidence since both the records contain Transgender.

High-Dimensionality - Many privacy models, such as K-anonymity and its extensions have been proposed to thwart privacy threats caused by identity and attribute linkages in the context of relational databases. The usual approach is to generalize the records into equivalence groups so that each group contains at least K records with respect to some QID attributes, and the sensitive values in each QID group are diversified enough to disorient confident inferences. However, has shown that when the number of QID attributes is large, that is, when the dimensionality of data is high, most of the data have to be suppressed in order to achieve K-anonymity, resulting in poor data quality for data analysis. Our experiments confirm this curse of high-dimensionality on K-anonymity [Aggarwal 2005]. In order to overcome this bottleneck, we exploit one of the limitations of an adversary.

Table II. Anonymous Data (L = 2, K = 2, C = 0.5, S= {Transgender})

	Quasi-identifier (QID)			Class	Sensitive
ID	Job	Sex	Age	Transfuse	Surgery
1	Professional	M	[30-50]	Y	Transgender
2	Non Technician	M	[30-50]	N	Plastic
3	Professional	F	[10-30]	N	Urology
4	Professional	M	[10-30]	Y	Vascular
5	Technician	M	[30-50]	N	Transgender
6	Non Technician	M	[30-60]	Y	Urology
7	Professional	F	[30-60]	Y	Urology
8	Professional	F	[10-30]	Y	Vascular
9	Non Technician	M	[30-60]	N	Plastic

In real-life privacy attacks, it is very difficult for an adversary to acquire all the QID information of a target patient because it requires nontrivial effort to gather each piece of prior knowledge from so many possible values. Thus, it is reasonable to assume that the adversary's prior knowledge is bounded by at most L values of the QID attributes of the patient. Based on this assumption, we define a new privacy model called LKC-privacy for anonymizing high-dimensional data. The general intuition of LKC-privacy is to ensure that every combination of values in QIDj ⊆ QID with maximum length L in the data table T is shared by at least K records, and the confidence of inferring any sensitive values in S is not greater than C, where L, K, C are thresholds and S is a set of sensitive values specified by the data holder (the hospital). LKC-privacy bounds the probability of a successful identity linkage to be ≤ 1/K and the probability of a successful attribute linkage to be ≤ C, provided that the adversary's prior knowledge does not exceed L. Table II shows an example of an anonymous table that satisfies (2, 2, 50%)-privacy with S = {Transgender} by generalizing all the values from Table I according to the taxonomies in Figure 2 (Ignore the dashed curve for now). Every possible value of QIDj with maximum length 2 in Table II (namely, QID1, QID2, and QID3 in Figure 2) is shared by at least 2 records, and the confidence of inferring the sensitive value Transgender is not greater than 50%. In contrast, enforcing traditional 2-anonymity will require further generalization. For example, in order to make Professional, M, [30 −60)to satisfy traditional 2-anonymity, we may further generalize all instances of [1 − 30) and [30 − 60) to [1 − 60), resulting in much higher utility loss.

Table III. Distributed Anonymization (L = 2, K = 2, C = 0.5, S= {Transgender})

	Quasi-identifier (QID)			Class	Sensitive
ID	Job	Sex	Age	Transfuse	Surgery
1	ANY	ANY	[30-50]	Y	Transgender
2	ANY	ANY	[30-50]	N	Plastic
3	ANY	ANY	[10-30]	N	Urology
4	Professional	ANY	[10-30]	Y	Vascular
5	Technician	M	[30-50]	N	Transgender
6	Non Technician	ANY	[30-60]	Y	Urology
7	Professional	F	[30-60]	Y	Urology
8	Professional	F	[10-30]	Y	Vascular
9	Non Technician	M	[30-60]	N	Plastic

Attribute Linkage

In the attack of attribute linkage, the attacker may not precisely identify the record of the target victim, but could infer his/her sensitive values from the published data T,based on the set of sensitive values associated to the group that the victim belongs to. In case some sensitive values predominate in a group, a successful inference becomes relatively easy even if k-anonymity is satisfied. Clifton [2000] suggested eliminating attribute linkages by limiting the released data size. Limiting data size may not be desirable if data records such as HIV patient data, are valuable and are difficult to obtain. Several other approaches have been proposed to address this type of threat.The general idea is to diminish the correlation between QID attributes and sensitive attributes.

Example.1 from Table II, an attacker can infer that all female Doctor at age 30 have Urology, i.e., Doctor, Female,30→ Urology with 100% confidence. Applying this knowledge to Table III, the attacker can infer that female Doctor has Urology problem with 100% confidence provided that female Doctor from the same population in Table II.

A. Distributed Anonymization

The centralized anonymization method can be viewed as "integrate-then generalize" approach, where the central government health agency first integrates the data from different hospitals then performs generalization. In real-life information

ISSN: 2277-9655
Impact Factor: 1.852

sharing, a trustworthy central authority may not always exist. Sometimes, it is more flexible for the data recipient to make requests to the data holders, and the data holders directly send the requested data to the recipient. For example, in some special occasions and events, BTS has to directly collect data from the hospitals without going through the government health agency. In this distributed scenario, each hospital owns a set of raw patient data records. The data can be viewed as horizontally partitioned among the data holders over the same set of attributes. Consider the raw patient data in Table I, where records 1–3 are from Party A, records 4–7 are from Party B, and records 8–11 are from Party C. To achieve distributed anonymization, the approach is to anonymize the patient data independently by the hospitals and then integrate as shown in Table III. However, such a distributed "generalize-then-integrate" approach suffers significant utility loss compared to the centralized "integrate-then-generalize" approach as shown in Table II. The distributed anonymization problem has two major challenges in addition to high dimensionality. First, the data utility of the anonymous integrated data should be as good as the data quality produced by the centralized anonymization algorithm. Second, in the process of anonymization, the algorithm should not reveal more specific information than the final anonymous integrated table.

Problem Definition

We first describe the privacy and information requirements, followed by the problem statement.

A. Privacy Measure

Suppose a data holder (e.g., the central government health agency) wants to publish a health data table T(ID, D1, . . . , Dm, Class, Sens) (e.g., Table I) to some recipient (e.g., the Red Cross BTS) for data analysis. ID is an explicit identifier, such as SSN, and it should be removed before publication. We keep the ID in our examples for discussion purpose only. Each Di is either a categorical or a numerical attribute. Sens is a sensitive attribute. A record has the form v1, . . . , vm, cls, s, where vi is a domain value of Di, cls is a class value of Class, and s is a sensitive value of Sens. The data holder wants to protect against linking an individual to a record or some sensitive value in T through some subset of attributes called a quasi-identifier or QID, where QID ⊆ {D1, . . . , Dm}. One recipient, who is an adversary, seeks to identify the record or sensitive values of some target victim patient V in T. As explained in Section 1, weassume that the adversary knows at most L values of QID attributes of the victim patient. We use qid to denote such prior known values, where |qid| ≤ L. Based on the prior knowledge qid, the adversary could identify a group of records, denoted by T[qid], that contains qid. |T[qid]| denotes the number of records in T[qid]. For example, T[Janitor,M] = {ID#1, 6} and |T[qid]| = 2. Then, the adversary could launch two types of privacy attacks:

(1) Identity linkage. Given prior knowledge qid, T[qid] is a set of candidate records that contains the victim patient V's record. If the group size ofT[qid], denoted by |T[qid]|, is small, then the adversary may identify V's record from T[qid] and, therefore, V's sensitive value. For example, if qid Mover, 34in Table I, T[qid] = {ID#5}. Thus, the adversary can easily infer that V has received a Transgender surgery.

(2) Attribute linkage. Given prior knowledge qid, the adversary can identify T[qid] and infer that V has sensitive value s with confidence P(s|qid) =|T[qidΛs]||T[qid]| , where T[qidΛ s] denotes the set of records containing both qid and s. P(s|qid) is the percentage of the records in T[qid] containing s. The privacy of V is at risk if P(s|qid) is high. For example, given qid = M, 34 in Table I, T[qid ∧ Transgender] = {ID#1, 5} and T[qid] = {ID#1, 5}, hence P(Transgender|qid) = 2/2 = 100%. To thwart the identity and attribute linkages on any patient in the tableT, we require every qid with a maximum length L in the anonymous table to be shared by at least a certain number of records, and the ratio of sensitive value(s) in every group cannot be too high. Our privacy model, LKC-privacy, reflects this intuition.

The data holder specifies the thresholds L, K, and C. The maximum length L reflects the assumption of the adversary's power. LKC-privacy guarantees that the probability of a successful identity linkage to be ≤ 1/K and the probability of a successful attribute linkage to be ≤ C. LKC-privacy has severalnice properties that make it suitable for anonymizing high-dimensional data. First, it only requires a subset of QID attributes to be shared by at least records. This is a major relaxation from traditional K-anonymity, based on a very reasonable assumption that the adversary has limited power. Second,

LKC-privacy generalizes several traditional privacy models. K-anonymity [Samarati 2001; Sweeney 2002] is a special case of LKC-privacy with L =|QID| and C = 100%, where |QID| is the number of QID attributes in the data table. Confidence bounding [Wang et al. 2007] is also a special case of LKC-privacy with L = |QID| and K = 1. (α, k)-anonymity [Wong et al. 2006]is also a special case of LKC-privacy with L = |QID|, K = k, and C = α. Thus,the data holder can still achieve the traditional models, if needed.

ISSN: 2277-9655
Impact Factor: 1.852

B. Problem Statement

We generalize the problems faced by BTS to the problems of centralized anonymization and distributed anonymization. The problem of centralized anonymization models the scenario of the central government health agency that anonymizes the integrated data before transferring it to BTS. The problem of distributed anonymization results from the scenario of the hospitals that distributively anonymize the data without the need of the central governmenthealth agency

Algorithm: Centralized Anonymization Algorithm
1: Initialize every value in T to the topmost value;
2: Initialize Cuti to include the topmost value;
3: while some x ∈ ∪Cuti is valid do
4: Find the Best specialization from ∪Cuti;
5: Perform Best on T and update ∪Cuti;
6: Update Score(x) and validity for x ∈ ∪Cuti;
7: end while
8: Output T and ∪Cuti;

C. Analysis

The distributed anonymization algorithm produces the same anonymous integrated table as the centralized anonymization algorithm. This claim follows from the fact that Algorithms 2 and 3 perform exactly the same sequence of specializations as the centralized anonymization algorithm in a distributed manner where Ti is kept locally at each party. For the privacy requirement, the only information revealed to the leader is content found in the global count statistics of Information message. The count statistics are needed for the calculation of Score and validity of the candidates. However, such information can also be determined from the final integrated table because a specialization should take place as long as it is valid. The disclosure of the score part does not breach privacy because it contains only the frequency of the class labels for the candidates. These values only indicate how good a candidate is for classification analysis, and does not provide any information for a particular record. Moreover, the Score is computed by the leader over the global count statistics without the knowledge of the individual local counts.

Experimental Evaluation

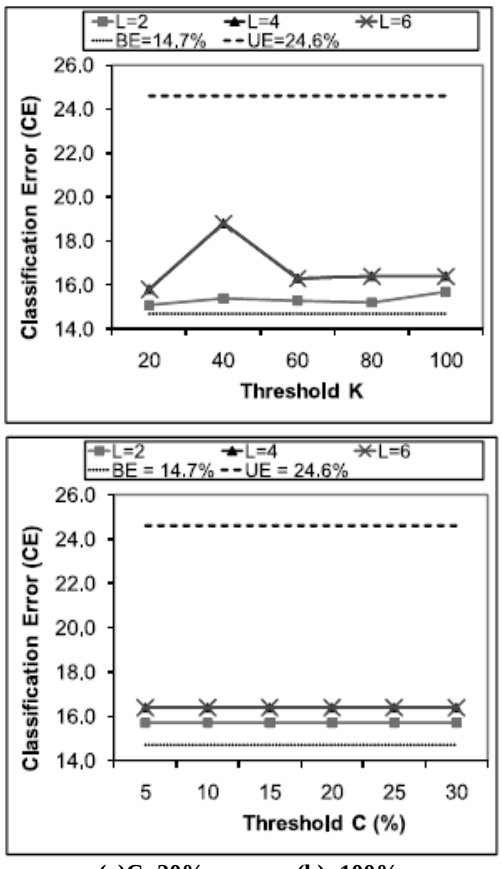
In this section, our objectives are to study the impact of enforcing various LKC privacy requirements on the data quality in terms of classification error and discernibility cost, and to evaluate the efficiency and scalability of our proposed centralized and distributed anonymization methods by varying the thresholds of maximum adversary's knowledge L, minimum anonymity K, and maximum confidence C. Blood Group represents the Class attribute with 8 possible values. Diagnosis Codes, which has 15 possible values representing 15 categories of diagnosis, is considered to be the sensitive attribute. The remaining attributes are neither quasi-identifiers nor sensitive. Blood contains 10,000 blood transfusion records in 2008. Each record represents one incident of blood transfusion. The publicly available Adult dataset [Newman et al. 1998] is a de facto benchmark for testing anonymization algorithms [Bayardo and Agrawal 2005; Fung et al. 2007; Iyengar 2002; Machana vajjhala et al. 2007;Wang et al. 2007]. Adult has 45,222 census records on 6 numerical attributes, 8 categorical attributes, and a binary Class column representing two income levels, ≤50K or >50K. See Fung et al. [2007] for the description of attributes. We consider Divorced and Separated in the attribute Marital-status as sensitive, and the remaining 13 attributes QID. All experiments were conducted on an Intel Core2 Duo 2.4GHz PC with 2GB RAM.

A. Data Utility

To evaluate the impact on classification quality (Case 1 in Section 3.2.1), we use all records for generalization, build a classifier on 2/3 of the generalized records as the training set, and measure the classification error (CE) on 1/3 of the generalized records as the testing set. Baseline Error (BE) is the error measured on the raw data without generalization. BE − CE represents the cost in terms of classification quality for achieving a given LKC-privacy requirement. A naıve method to avoid identity and attributes linkages is to simply remove all QID attributes. Thus, we also measure upper bound error (UE), which is the error on the raw data with all QID attributes removed. UE− CE represents the benefit of our method over the approach. To evaluate the impact on general analysis quality we use all records for generalization and measure the discernibility ratio (DR) on the final anonymous data. DR = qid|T[qid]|2|T|2 . DR is the normalized discernibility cost, with 0 ≤ DR ≤ 1. Lower DR means higher data quality. Centralized Anonymization depicts the classification error CE with adversary's knowledge L = 2, 4, 6, anonymity threshold 20 ≤K ≤ 100, and confidence threshold C = 20% on the Blood dataset. This setting allows us to measure the performance of the centralized algorithm against identity linkages for a fixed C. CE generally increases as K or L increases. the increase is not monotonic. Generalization has removed some noise from the data, resulting in a better classification structure in a more general state. For the same reason, some test cases on L = 2 and L = 4 have CE < BE, implying that generalization not only achieves the given LKC-privacy requirement but

ISSN: 2277-9655
Impact Factor: 1.852

sometimes may also improve the classification quality. BE = 22.1% and UE = 44.1%. For L = 2 and L = 4, CE− BE spans from -2.9% to 5.2% and UE−CE spans from 16.8% to 24.9%, suggesting that the cost for achieving LKC-privacy is small, but the benefit is large when L is not large. However, as L increases to 6, CE quickly increases to about 40%, the cost increases to about 17%, and the benefit decreases to 5%. For a greater value of L, the difference between LKC-privacy and K-anonymity is very small in terms of classification error since more generalized data does not necessarily worse classification error. This result confirms that the assumption of an adversary's prior knowledge has a significant impact on the classification quality. It also indirectly confirms the curse of high dimensionality [Aggarwal 2005].

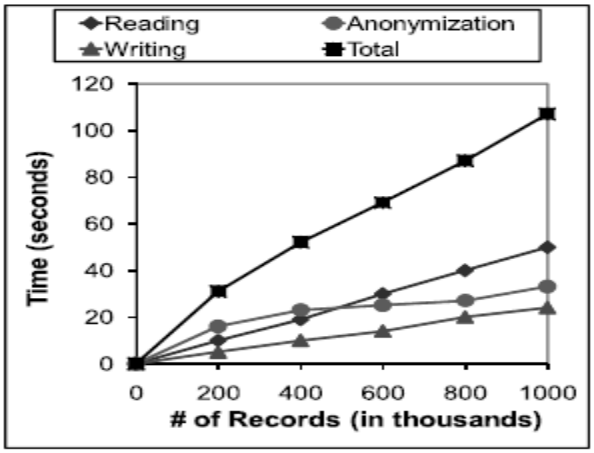


(a)C=20% **(b)=100%**

These results suggest that the cost for achieving LKC-privacy is small, while the benefit of our method over the naıve method is large. Figure b depicts the CE with adversary's knowledge L = 2, 4, 6, confidence threshold 5% ≤ C ≤ 30%, and anonymity threshold K = 100. This setting allows us to measure the performance of the algorithm against attribute linkages for a fixed K. The result suggests that CE is insensitive to the change of confidence threshold C. CE slightly increases as the adversary's knowledge L increases.

C. Efficiency and Scalability

One major contribution of our work is the development of an efficient and scalable algorithm for achieving LKC-privacy on high-dimensional healthcare data. Every previous test case can finish the entire anonymization process within 30 seconds. First, we combined the training and testing sets, giving 45,222 records. For each original record r in the combined set, we created $\alpha - 1$ "variations" of r, where $\alpha > 1$ is the blowup scale. Together with all original records, the enlarged dataset has $\alpha \times 45,222$ records. Figure 9 depicts the runtime of the centralized anonymization algorithm from 200,000 to 1 million records for L = 4, K = 20, C = 100%. The total runtime for anonymizing 1 million records is 107s, where 50s are spent on reading raw data, 33s are spent on anonymizing, and 24s are spent on writing the anonymous data.



Conclusion

The proposed two anonymization algorithms to address the centralized and distributed anonymization problems for healthcare institutes with the objective of supporting data mining. Motivated by the BTS's privacy and information requirements, we have formulated the LKC-privacy model for high-dimensional relational data. Moreover, our developed algorithms can accommodate two different information requirements according to the BTS' information need. Our proposed solutions are different from privacy-preserving data mining (PPDM) due to the fact. This is an essential requirement for the BTS since they require the flexibility to perform various data analysis tasks. We believe that our proposed solutions could serve as a model for data sharing in the healthcare sector. Finally, we would like to share our

ISSN: 2277-9655
Impact Factor: 1.852

collaborative experience with the healthcare sector. Health data is complex, often a combination of relational data, ACM Transactions on Knowledge Discovery from Data, Vol. 4, No. 4, Article 18, Pub. date: October 2010. Anonymization for High-Dimensional Healthcare Data. Thus, the project focuses only on the relational data, but notice that some recent works [Gardner and Xiong 2009;Ghinita et al. 2008; Terrovitis et al. 2008; Xu et al. 2008], are applicable to solve the privacy problem on transaction and textual data in the BTS case. Besides the technical issue, it is equally important to educate health institute management and medical practitioners about the latest privacy-preserving technology.

When management encounters the problemof privacy-aware information sharing as presented in this paper, their initial response is often to set up a traditional role-based secure access control model. In fact, alternative techniques,such as privacy-preserving data mining and data publishing [Aggarwal and Yu 2008; Fung et al. 2010], are available provided that the data mining quality does not significantly degrade.

References

[1] *S. Goryczka, L. Xiong, and B. C. M. Fung, "m-Privacy for collaborative data publishing," in Proc. of the 7th Intl. Conf. on Collaborative Computing: Networking, Applications and Worksharing, 2011*

[2] *C. Dwork, "Differential privacy: a survey of results," in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1–19.*

[3] *B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy- preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, pp. 14:1–14:53, June 2010.*

[4] *C. Dwork, "A firm foundation for private data analysis," Communication ACM, vol. 54s January 2011.*

[5] *N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowl. Discovery from Data, vol. 4, no. 4, pp. 18:1–18:33, October 2010.*

[6] *W. Jiang and C. Clifton, "Privacy-preserving distributed k- anonymity," in DBSec, vol. 3654, 2005.*

[7] *W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 2006.*

[8] *O. Goldreich, Foundations of Cryptography: Volume 2, 2004.*