IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Privacy Protection of Collaborative Data Using Slicing Technique

Sarita Kashid[1], Prof. Mrs. Manasi K. Kulkarni [2]

M.E. Scholar, Computer Engineering, PES Modern College of Engineering, Pune, India [1]

Asst Prof, Computer Engineering, PES Modern College of Engineering, Pune, India [2]

Abstract: There is an increasing need for sharing data that contains personal information from distributed database. The data is collected from multiple providers. The sharing of data can be used for research in different fields. While sharing a data first, personal information must be hidden from user. There are different techniques used for encryption of personal information. The process of removing personally identifiable information from data sets is called Anonymization. The techniques are k-anonymity, bucketization and slicing. Privacy for data and verification against privacy is very important and challenging task. Best utility must be preserved for collaborative data and loss of data must be avoided. Secure Multi Party (SMP) and Trusted Third Party (TTP) protocols create anonymized data. There may be a problem of insider and outsider attacker's for collaborative data. The proposed work will be the implementation of Slicing algorithm which gives more security to the database.

Keywords: M-Privacy; K-Anonymity; Suppression, Anonymization; Bucketization; Slicing.

I. INTRODUCTION

Anonymization is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personal information from data sets. There is need to publish collaborative data of any field for analysis purpose. This process either encrypting or removing personally identifiable information from data sets.

Figure 1.1: Privacy Preserving Data Model.

There are multiple providers to collaborate data. Integration of data, Privacy of data, verification against privacy is very important and challenging task. Individual information must be hiding from collaborative data. Data should be present in anonymized form. M-Privacy is a policy to protect the personal information. There are different techniques to provide privacy or anonymizing view of data. These techniques a r e K-anonymity [2], Bucketization [3] and Slicing. These techniques use different algorithms like k- Anonymity [6], l-Diversity [7] and t-Closeness [8].

There are different methods while proving anonymized view of data. In first method the data of individual provider gets anonymized and then aggregated [9]. In second method data get aggregated and then anonymized [6]. Database consists of following three types of attributes:

1. Identifiers (ID): These attributes are clearly identifying individual's information.

Examples: Employ ID and Name.

2. Quasi-identifiers (QI): It is a subset of attributes that can distinguish almost all tuples. Examples: Sex and Date of Birth.

3. Sensitive Attributes (SAs): These attributes are the researcher's need, so they are always released directly.

Examples: Salary and disease, It is responsibility of Trusted Third Party (TTP) to create anonymized data using Secure Multiparty Computation (SMC) protocol [5]. A data recipient may access some background knowledge, which represents any publicly available information about released data.

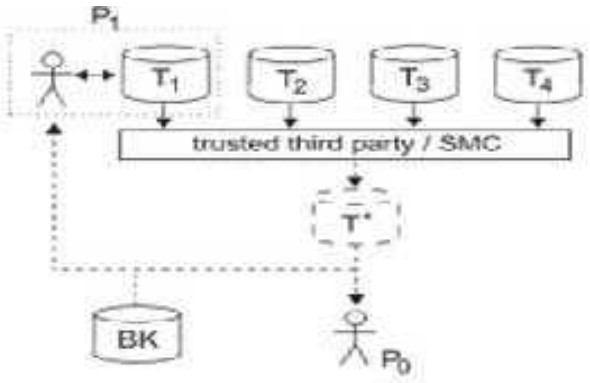

Figure1.2: General Architecture For Collaborative DAT A PubLish InG.

DOI 10.17148/IJARCCE.2016.5717

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Key Attribute	Quasi-Identifier			Sensitive Attribute
Name	DOB	Gender	Zip code	Disease
Andre	1/1/1999	Male	53457	Heart Disease
Beth	2/2/1986	Female	53715	Heart Disease
Dan	1/21/1976	Male	53703	Heart Disease

Figure 1.3: Classification Attributes.

II. RELATED WORK

Many Researchers and authors have contributed and put their efforts for creating anonymization view Web Search. Some of their researches and views have been presented.

A. K-anonymity

K-anonymity [2] is the process of replacing the information with semantically consistent value. It restore quasi identifier values with values that are less specific but semantically consistent. Due to the high dimensionality of the quasi identifier, with different possible items in thousands of order, that any k-anonymity method will cause high information loss and also rendering the data in useless.

In order to improve k-anonymity in efficient manner, records in the same bucket must be similar to each other so that during k- anonymity the records would not lose too much information.

The k-anonymity categories are as follows:

- Global recoding-An attribute come from the same domain level in the hierarchy
- Regional recoding- It is also called multidimensional recoding which partitions the domain space into noninterest region and data points in the same region are represented by the region they are in. Regional allows different values of an attribute to be generalized to different levels.
- Local recoding- The same value to be generalized to different values in different records.

Limitation of K-anonymity

- Fails on high-dimensional data due to the curse of dimensionality
- Too much information loss due to uniform distribution.

B. Bucketization

Bucketization [3] separates the sensitive attribute from the non-sensitive attributes by randomly permuting the sensitive attribute values within each bucket.

Limitation of Bucketiztion

1. It does not prevent membership disclosure.
2. It requires a clear separation between QIs and SAs.
3. It breaks the attribute correlations between the QIs and the SAs by separating the SA from the QI attributes.

Bucketization partitions tuples in the table into buckets and then separate quasi identifiers with sensitive attribute by randomly permuting the sensitive attribute values in each bucket in the table. The anonymized data set consist of a set of buckets with permuted sensitive attribute values. Bucketization has been used for anonymizing high-dimensional data. The given approach assumes a clear separation between QIs and SAs.

There are two types of privacy preserving techniques used in bucketization.

C. k-anonymity

K-anonymity it is used to prevent identification of individual records in the given data set. The database is said to be k-anonymous where attributes are generalized until each row is identical with at least k-1 rows. It guarantees t h a t the data released is accurate.

The guarantee given by k-anonymity [6] is that no information can be linked to groups of less than k individuals. In K-anonymity for k-anonymity which losses considerable amount of information, for higher dimensionality data set.

– K-anonymity model for multiple sensitive attributes consist of three kinds of information disclosure:

Identity Disclosure: An individual who can link to a particular record in the published data set is known as identity disclosure.

Attribute Disclosure: When the sensitive information regarding particular individual revealed is known as attribute disclosure.

Membership Disclosure: Information regarding individual belongs from data set is presented or not revealed is a membership disclosure. Anonymity refers to a state where data does not show its identity. A dataset which satisfies k-anonymity if every record in the dataset is not distinguished from at least k-1 other records with respect to every set of quasi identifier attributes is known as k-anonymity dataset.

Limitations

It does not be able to hide whether a given individual is in the database.

1. It reveals individuals' sensitive attributes.

The attack based on background knowledge is not prevented.

It cannot be applied to high-dimensional data without data loss.

Different methods are required for a dataset which is anonymized and published more than once

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Attacks on k-anonymity are homogeneity attack and background knowledge attack.

Homogeneity Attack: It is sensitive information in the dataset may be revealed based on the known information, if the non sensitive information of an individual is revealed to an adversary. The method of information revealing is known as positive disclosure.

Background Knowledge Attack: If the user has some external information that can be linked to the released data which helps in neglecting some of the sensitive attributes.

D. l-diversity

l-diversity [7] prevents the association of an individual record with sensitive attribute value. l-diversity is a distribution of a sensitive attribute in each equivalence class which has at least l well represented values to protect against attribute disclosure. l-diversity over come limitations of k-anonymity, as l-diversity for privacy preserving provides that data publisher who does not know what kind of knowledge is possessed by an attacker. l- diversity is based on requirement is that values of the sensitive attributes are well-represented in each group.

l-diversity over come limitations of k-anonymity, as l-diversity for privacy preserving provides that data publisher who does not know what kind of knowledge is possessed by an attacker. l-diversity is based on requirement is that values of the sensitive attributes are well-represented in each group. [7]. the distribution of target values within a group is referred to be known as l-diversity.

This principle represents an important step beyond k anonymity in protecting against attribute disclosure.

Limitations of l-diversity

It may be difficult to achieve is insufficient to prevent the attribute disclosure.
It does not consider overall distribution of sensitive values.
Semantic meanings of sensitive values are not considered in l-diversity.

It is not be able to prevent probabilistic attack.
l diversity

1. Some of the attacks by which limitation occurs are: Skewness Attack: When the given overall distribution is skewed satisfying the l-diversity does not prevent attribute disclosure. Similarity Attack: When the sensitive values in a QI group are distinct but semantically similar, an adversary can able to learn information. T-closeness [8] is a further refinement of l-diversity group based anonymization. It is used to preserve privacy in data sets by reducing the granularity of a data representation. The t- closeness model extends the l-diversity model.

III. PROPOSED SYSTEM IMPLEMENTATION

A. System Architecture

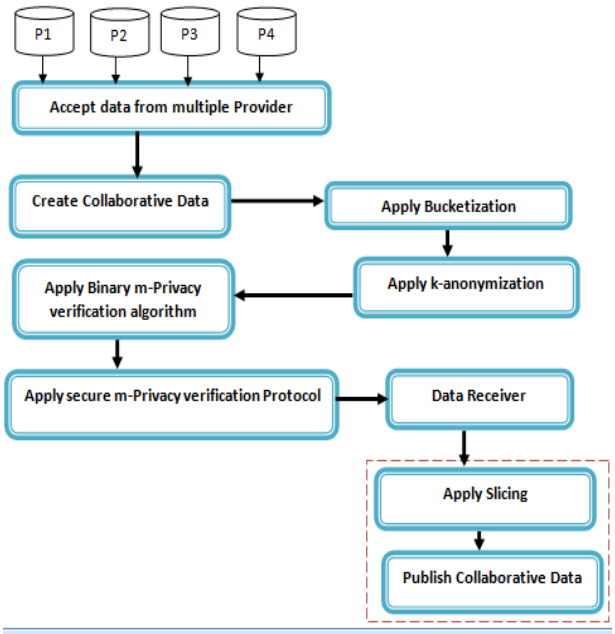

Figure 3.1: Proposed System Architecture

B. Proposed Modules

There are seven modules in the system:

- Accept input from multiple providers
- Create Collaborative d ata
- Apply Buckectizayion for creating anonymization view of data
- Apply k-Anonymity for creating anonymized view of
- Verify anonymized data by m-Privacy Verification protocol
- Publish Anonymized data
- Appliy Slicing for creating anonymization view of data

1. Accept input from multiple providers:

Multiple data provider of same field send their data for creating anonymized view of data. This data contains personal information. Trusted Third Party (TTP) accepts data of multiple providers for creating anonymized view.

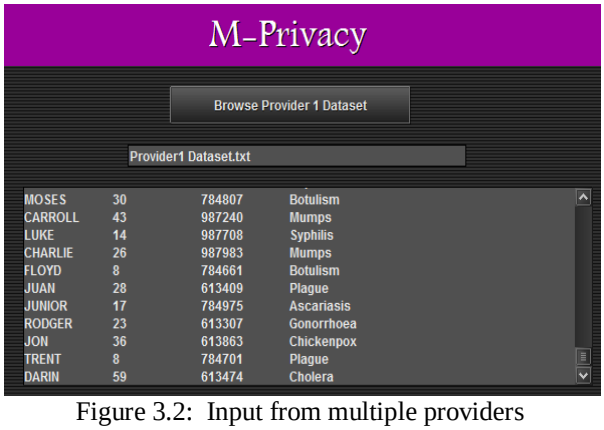

Figure 3.2: Input from multiple providers

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

2. Create Collaborative Data:
Data from multiple provider get collaborated into single format. Collaboration of data done by Trusted Third Party (TTP).

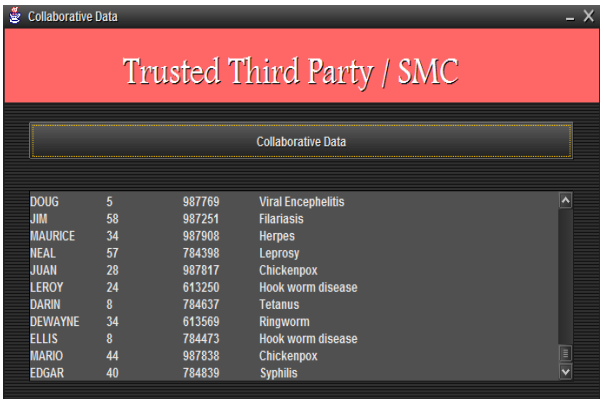

Figure 3.3: Create Collaborative Data

3. Applying Bucketization For Creating Anonymized View Of Data:
In this module, it create buckets on age ranges.

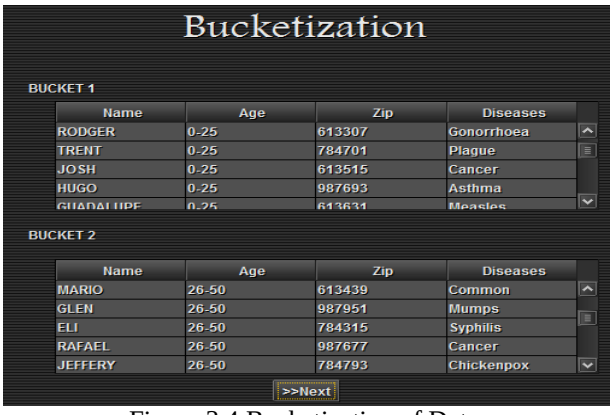

Figure 3.4 Bucketization of Data

4. Apply K-Anonymity For Creating Anonymized View Of Data:
In this step k-Anonymity convert collaborative data in anonymized view . It replaces some digit from value of attribute by * To calculate range for replacement of digits, it uses following formula of score.

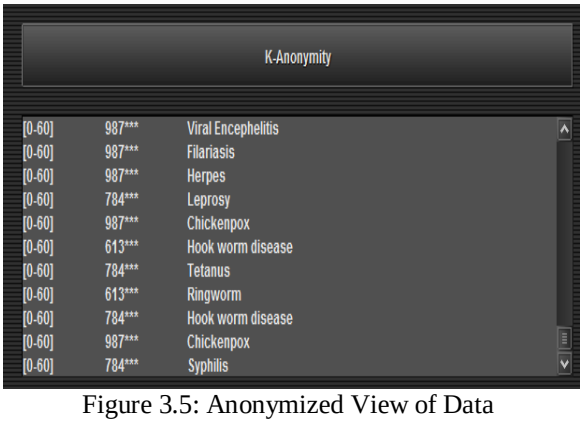

Figure 3.5: Anonymized View of Data

5. Verify Anonymized Data By M-Privacy Verification Protocol:
m-Privacy Verification protocol check correctness of anonymized data. Implementation of introduced algorithms can be run by a trusted third party (TTP).

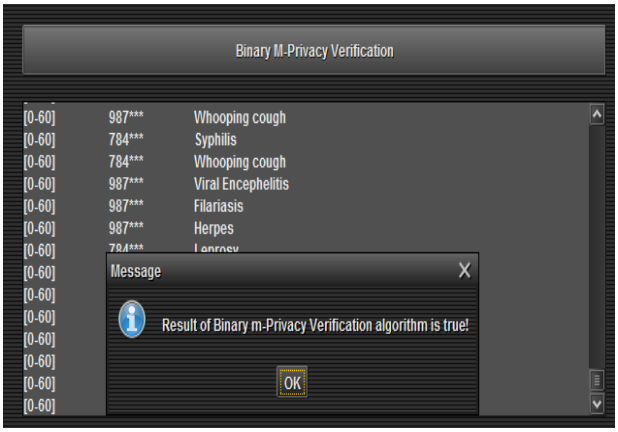

Figure 3.6: m-Privacy Verification of Data

6. Apply Slicing Algorithm And Publish Anonymized Data:
In this module slicing create anonymized view of data Anonymized data get published for receiver. In this data personal information is not present. This type of data provides privacy for personal information.

Figure 3.7: Publishing Data

III. ALGORITHMIC STRATEGY

A. Slicing Algorithm
Load Dataset
Attribute Partition and Column

An attribute partition consists of several subsets of A, such that each subset of attributes is termed a column. Precisely, let there be c columns c1, c2,cc where ci=1 to ci=A and for $1 \leq i1 \neq i2 \leq c$, $ci1 \cap ci2 = \infty$.

We consider only one sensitive attribute S. If the data contain multiple sensitive attributes, one can either consider them individually or consider their combined distribution. Exactly one of the c columns contains S. Without loss of generality, let the column that holds S be

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

the last column Cc. This column is called the sensitive column. All other columns contain only Q1 attributes.

Process tuples partition and buckets
A tuple partition comprises of several subsets of T, such that each tuple belongs to exactly one subset. Each subset of tuple belongs to exactly one subset. Each subset of tuples is called a bucket. Specifically, let there be b buckets b1, b2,……., bc then bi=1 to bi=T and for 1 ≤ i 1 ≠ i2 ≤ b, bi1∩ bi2 = ∅
Tuple partition (T,1)
Q = {T}, MSB = ∅; While Q is not empty
Remove first bucket B from Q; Q = Q – {B};
Similarity check (T,d)

Split B into two buckets B1, B2 as in Mondrian
If diversity-check (T,Q ∪ (B1, B2} ∪ SB ,1) Q = Q ∪ {B11 ∪ B21}
else
SB = SB ∪ {B11} ∪ {B21} Return SB
Slicing

For each tuple t ε T, L[t] = ∅

For each bucket B in T*

Record f(v) for each column value v in bucket B. For each tuple t ε T
Calculate p(t,B) and find D(t,B) L[t] = L[t] ∪ {hp (t,B), D(t,B)i} For each tuple t ε T
Calculate p(t,s) for each s based on L[t] If p(t,s) ≥1/l, return false
Return true

IV. EXPERIMENTAL SETUP

The UPS framework is implemented on a PC with a Processor Intel core i3, Speed 3.00 GHz, RAM – 2.00GB running Microsoft Windows XP/7/8. All the algorithms are implemented using Java. The algorithms are compared based on their response time.

Name of Dataset-
Medical Clinic Activity Data set
Link-https://www.gov.uk/guidance/genitourinary-medicine -clinic-activi ty-dataset-gumcadv2VI.
Size- 16KB
It contains 4 attributes (Name, Age, Zipcode, Disease)

V. RESULT ANALYSIS

When we create anonymized view of data, we found that the following results. There are four performance parameters.

1 .Required Time
2. Information Loss
3. Privacy Gain
4. Utility Loss

A. Required Time
As experimental results shows, Slicing required less time than Bucketization and k-anonymity.

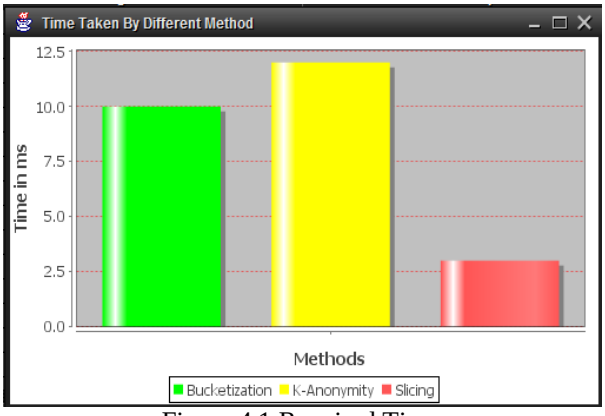

Figure 4.1 Required Time

B. Information Loss

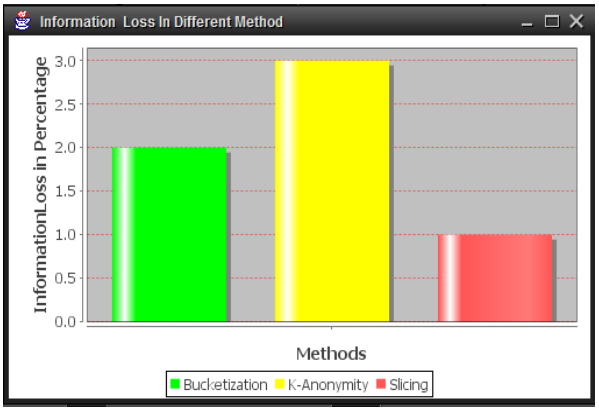

Figure 4.2 Information Loss

Above graph shows Slicing having less information loss than Bucketization and k-anonymity.

C. Privacy Gain

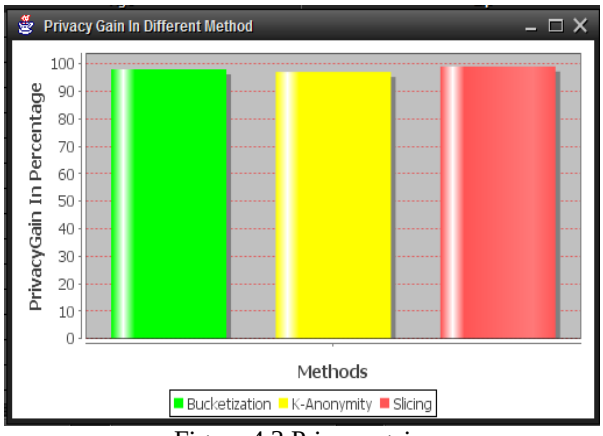

Figure 4.3 Privacy gain

Privacy gain of slicing is more than bucketization and k-anonymity.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

D. Utility Loss

Utility Loss of Slicing is less than bucketization k-anonymity.

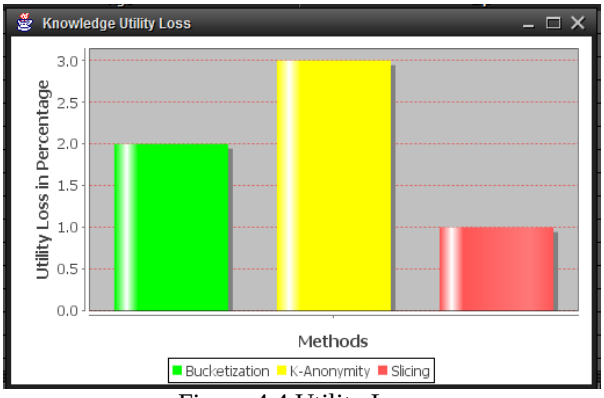

Figure 4.4 Utility Loss

VI. CONCLUSION AND FUTURE WORK

Two anonymization algorithms, namely k-anonymity and Bucketization have been implemented. Experimental result shows that Slicing performs better than k-anonymity and Bucketization in terms of response time, Privacy gain and utility Loss. data utility while protecting against privacy threats. Combination of these methods will give best privacy.

Experimental result shows that Bucketization performs better than K-anonymity in terms of. The proposed work will be the implementation of an Slicing algorithm, which will give more security to the database.

REFERENCES

[1] C. Dwork, "A firm foundation for private data analysis, " in Commun. ACM, Vol. 54, No.1, pp. 8695, 2011.
[2] K. Wang, P. S. Yu and S. Chakrab orty,"Bottom-up k-anonymity: A data mining solution to privacy protection," In Proc. of the 4th IEEE Int. Conf. on Data Mining (ICDM), 2004.
[3] Younho Lee, "Secure Ordered Bucketization," IEEE Transactions on Dependable and secure computing, Vol.11, No. 3, 2014.
[4] S. Ram Prasad Reddy, KVSVN Ra ju, and V. Valli Kumari, "A Dynamic Programming Approach for Privacy Preserving Collaborative Data Publishing," Int. J. Comp. Appli. (0975 8887), Vol. 22 No.4, 2011.
[5] S. Goryczka, Li Xiong, and Benjamin C. M. Fung, "m-Privacy for Collaborative Data Publishing," IEEE Trans. Knowl. Data En g., Vol.26, No. 10, 2014.
[6] L. Sweeney, "k-Anonymity: A model for protecting privacy," Int. J. Uncertain. Fuzz. Knowl. Based Syst., Vol. 10, No. 5, pp. 557570, 2002.
[7] A. Machanava jjhala, J. Gehrke and D. Kifer, "l-Diversity: Privacy Beyond k- Anonymity, " IEEE Proceedings of the 22nd Int. Conf. Data Eng., 2006.
[8] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," IEEE Trans. Knowl. Data Eng, 1-4244-0803-2/07/2007.
[9] C. duanzhi, "Program Slicing ", Int. Forum on Infor. Technology. and Application.IEEE, DOI 10.1109/IFITA. 58,2010.
[10] G. Aggarwal, N. Mishra, and B. Pinkas,"Secure computation of the kth-ranked element," in Proc. EUROCRYPT, Interlaken, pp.4055,Switzerland, 2004.
[11] M. Burkhart and X. A. Dimitropoulos,"Fast privacy-preserving top-k queries using secret sharing," in Proc. 19th ICCCN, Zurich, Switzerland, pp. 17,2010.
[12] S. Pohlig and M. Hellman, "An improved algorithm for computing logarithms over GF(p) and its cryptographic significance (Corresp.)," IEEE Trans. Inf. Theory, vol.24, no. 1, pp. 106110, Jan. 2006.
[13] Y. Tao, X. Xiao, J. Li and D. Zhang, "On anti-corruption privacy preserving publication, "in Proc. IEEE 24th ICDE, Cancun, Mexico, pp. 725734, 2008.
[14] W. Jiang and C. Clifton, "A secure distributed framework for achieving k- anonymity, "VLDB J., vol. 15, no. 4, pp. 316333, 2006.
[15] N. Mohammed, B. C. M. Fung, K. Wang and P. C. K. Hung, "Privacy-preserving data mashup," in Proc. 12th Int. Conf. EDBT, Biopolis, Singapore, pp. 228239, 2009.
[16] S. Zhong, Z. Yang, and R. N. Wright, "Privacy-enhancing k- anonymization of customer data," in Proc. 24th ACM SIGMOD- SIGACT-SIGART Symp. PODS, Baltimore, MD, USA, pp. 139147 2005