SW-SDF Based Personal Privacy with QIDB-Anonymization Method

Kiran P
Research Scholar
VTU,Belgaum
Karnataka, India

Dr Kavya N P
Prof & Head
Dept of MCA,RNSIT
Bangalore,Karnataka,India

***Abstract*— Personalized anonymization is a method in which a guarding node is used to indicate whether the record owner is ready to reveal its sensitivity based on which anonymization will be performed. Most of the sensitive values that are present in the private data base do not require privacy preservation since the record owner sensitivity is a general one. So there are only few records in the entire distribution that require privacy. For example a record owner having disease flu doesn't mind revealing his identity as compared to record owner having disease cancer. Even in this some of the record owners who have cancer are ready to reveal their identity, this is the motivation for *SW-SDF* based Personal Privacy. In this paper we propose a novel personalized privacy preserving technique that over comes the disadvantages of previous personalized privacy and other anonymization techniques. The core of this method can be divided in to two major components. The first component deals with additional attribute used in the table which is in the form of flags which can be used to divide sensitive attribute. Sensitive Disclosure Flag (SDF) determines whether record owner sensitive information is to be disclosed or whether privacy should be maintained. The second flag that we are using is Sensitive Weigh (SW) which indicates how much sensitive the attribute value is as compared with the rest. Second section deals with a novel representation called Frequency Distribution Block (FDB) and Quasi–Identifier Distribution Block(QIDB) which is used in anonymization. Experimental result show that it has lesser information loss and faster execution time as compared with existing methods.**

Keywords- Privacy Peserving Data Mining(PPDM);Privacy Preserving Data Publishing(PPDP); Personal Anonymization.

I. INTRODUCTION

Personal information present in different organizations can be used by research for understanding patterns there by achieving betterment of the community. For example a personal detail of the patient is present in different hospitals, this information can be used by researchers to understand the patterns for a particular disease and hence improve the identification of the diagnosis. The raw data present in hospitals contain detailed information regarding the patient like name, address, DOB, zip code, symptoms & disease. From this raw data, details regarding name and address which are considered personal are removed before it is given to Data Recipient and this information is also called Microdata. This microdata however contains details like zip, DOB that can be linked with other external publicly available data bases for re-identification of sensitive value.

This re-identification of the record by linking public data to Published data is called as linking attack. For example consider the details of the patient Published by the hospital in table 1, which does not contain details regarding name, address and other personal information. The attacker can use the publicly available external data base shown in table 2 and join these details with table 1 thereby personal details can be revealed. The query may look like

```
SELECT NAME, DISEASE
FROM VOTERS_TABLE AS V, PAIENT_TABLE AS P
WHERE V.ZIP=P.ZIPAND V.AGE=P.AGE;
```

The result of this query gives me entire details regarding sensitive information i.e. disease and the identity of the individual which is of great concern because the individuals are not ready to share their sensitive information. The join may give me a value <RAMA, Gastric ulcer > for zipcode 48677 & age 26 and is called Record Level Disclosure. The approaches used by researchers to mask sensitive data from Data Recipients come under a category called Privacy Preserving Data Publishing (PPDP). Attributes present in Published Patient Data that can be linked to external publicly available data bases like ZIP, DOB,… are called Quasi-Identifier (Q) attributes.

TABLE 1. PATIENT PUBLISHED DATA

ZIP Code	Age	Disease
48677	26	Gastric ulcer
48602	28	Stomach cancer
48678	32	Flu
48685	36	Flu
48905	42	Flu
48906	46	Flu
48909	43	Flu
48673	48	Heart Disease
48607	55	Heart Disease
48655	58	Stomach_cancer

Modification of data is done in such a way that the resultant table has duplicated records there by restricting the disclosure. Indirectly there must be more than one link to the

external data base and is done by using generalization [1, 2, 3, 4]. Once the table is generalized various methods were used to check the property of duplication and distribution. To measure this Samarati and Sweeney [6,7] introduced k-anonymity. 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; such a table is called a k-anonymous table. In other words each group of quasi identifier values must have at least k-1 records and can be cheeked by linking a record in released data to multiple records publicly available data base. Table 3 shows a 2-anonymus generalization for table 1. Let us assume that the attacker uses the publicly available data base and finds that Rama's zip code is 48677 and his age is 26 and wants to know the disease of Rama, the attacker observes the anonymized table 3 from which attacker understands that 48677 & 26 has been generalized to 486** & [20-30] which can be linked to two records of published table and hence the disease cannot be inferred. In this table <486**,[40-50],Heart Disease> has been suppressed and is not considered for publication. Similarly if the attacker tries to infer Sita's disease who is related to group 3 but since the entire group contains the same sensitive attribute the attacker infers that his disease is Flu. This leakage of sensitive value leads to Attribute Level Disclosure. This happens if all the diseases indicated in a group are related to the same disease. To overcome this *l*-diversity [8] was defied. An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have *l*-diversity if every equivalence class of the table has *l*-diversity. *l*-diversity also has the disadvantage that it suffers from skewness and similarity attack. To overcome this t-closeness was defined [9]. In this technique distribution of sensitive attribute must be equal to the anonymized block. This suffers from information loss

A. Motivation

Major disclosures that take place are record level and attribute level to avoid this various anonymity techniques have been proposed in literature. Among them most important are k-anonymity, l-diversity, t-closeness, but each of them have several drawbacks as indicated above it includes data utility loss and sensitivity disclosure. To overcome this author in [5] had indicated a method personalized privacy preservation which takes in to account record owners privacy requirement. In [5] the record owner can indicate his privacy by indicating in terms of a guarding node. The values of it are based on a sensitive hierarchy which is framed by Data Publisher. The core of this technique is to divide the sensitive value based on importance so that more privacy is given to those values and data utility is improved. The drawback of this method is that it may require several iterations based on the guarding node, sensitive attribute is also generalized which has larger information loss. The most important drawback is that distribution of sensitive attribute has not been taken in to account while anonymization.

B. Contribution and paper outline

In this paper we propose a novel privacy preserving technique that over comes the disadvantages of [5] and other anonymization techniques. The core of this method can be divided in to two major components. The first component deals with additional attribute used in the table which is in the form of flags. Sensitive Disclosure Flag (SDF) determines whether record owner sensitive information is to be disclosed or whether privacy should be maintained. The second flag that we are using is Sensitive Weigh (SW) which indicates how much sensitive the attribute is. SDF is dependent on SW.

TABLE 2. EXTERNAL VOTERS DATA BASE

Name	ZIP Code	Age
Rama	48677	26
Laxman	48677	35
Suresh	48602	28
Nagesh	48602	22
Anuma	48678	32
Sita	48905	42
Kushal	48909	43
Vihan	48906	46
	.	
	.	

TABLE 3. 2-ANONYMUS TABLE

ZIP Code	Age	Disease
486**	[20-30]	Gastric ulcer
486**	[20-30]	Stomach cancer
486**	[30-40]	Flu
486**	[30-40]	Flu
489**	[40-50]	Flu
489**	[40-50]	Flu
489**	[40-50]	Flu
486**	[40-50]	Heart Disease
486**	[50-60]	Heart Disease
486**	[50-60]	Stomach_cancer

SDF can be easily obtained from the individual when he/she is providing her data. SW can be based on the prior knowledge of sensitive attribute. General privacy methods provide the same level of security for all sensitive attributes which has been overcome in this method by the use of SDF and SW. The flag SDF=0 means that the record owner is not ready to disclose his sensitive attribute whereas SDF=1 doesn't mind revealing his sensitivity. SW is indicated by the publisher for those Sensitive attribute where privacy is at most important. For example record owner who has Flu or Gastritis doesn't mind revealing his identity as compared to a record owner who has Cancer. The value of SW=0 is used when the sensitive attribute is a common disease like Flu or Gastritis and SW=1 for sensitive attribute like Cancer which is not common. For SW=0 default value of SDF=1 & if SW=1 SDF values are accepted from record owner.

Second section deals with a novel representation called Frequency Distribution Block (FDB) and Quasi–Identifier Distribution Block (QIDB) used for measuring the distribution. FDB contains distribution of every disease with respect to original private data. For every record with SW=1 and SDF=0 QIDB is created. There will be multiple QIDB

blocks. These blocks are used to ensure that distribution of FDB is matched with individual QIDB.

In section II we have indicated Model and Notations used in our Personalized Privacy. Personalized Privacy Breach has been discussed in section III. Section IV gives the QIDB-Anonymization Algorithm. Experiment in section V has been analyzed. Related work has been discussed in section VI. Last section deals with conclusion and future work.

TABLE 4. SW FOR DISEASES

Disease	SW
Gastric ulcer	0
Stomach cancer	1
Flu	0
Heart Disease	1

TABLE 5. PATIENT PUBLISHED DATA WITH SW & SDF

ZIP Code	Age	Disease	SW	SDF
48677	26	Gastric ulcer	0	1
48602	28	Stomach cancer	1	0
48678	32	Flu	0	1
48685	36	Flu	0	1
48905	42	Flu	0	1
48906	46	Flu	0	1
48909	43	Flu	0	1
48673	48	Heart Disease	1	1
48607	55	Heart Disease	1	0
48655	58	Stomach_cancer	1	1

TABLE 6. FREQUENCY DISTRIBUTION BLOCK

Disease	Probability
Gastric ulcer	0.1
Stomach cancer	0.2
Flu	0.5
Heart Disease	0.2

II. MODEL AND NOTATION FOR PERSONALIZED PRIVACY

Let T be a relation containing private data about a set of individuals. there are four categories of attributes in T i) unique Identifiers UI_i which can be used for identification of a person and is removed from T ii) quasi identifiers Q_i whose values can be used for revealing the identity of a person by joining Q_i with publicly available data iii) sensitive attributes S_i which is confidential or sensitive to the record owner. iv)Non quasi identifiers NQ_i which do not belong to the previous three categories.

Objective of our approach is to find a generalized table T^* such that distribution of each *QIDB* is approximately equal to the diversity of the overall distribution which is there in *FDB*. For simplicity the entire quasi identifiers are represented as Q and their values as q. similarly we assume there is a single sensitive attribute S and its value is s. Relation T is made of n number of tuples $T=\{t_1,t_2,\ldots,t_n\}$. Record owner information can be retrieved by referring as $t_i.s$ to indicate sensitive value and $t_i.q$ for quasi identifier value $1 \leq i \leq n$.

A. Requirement for personal privacy

DEFINITION 1 (SENSITIVE WEIGHT) For each tuple $t \in T$, its *sensitive weight* is added. This value is taken from Relation $W(d,sw)$ where d disease and sw sensitive weight. W contains k records.

$$t_i.sw=\{ w_j.sw \text{ if } w_j.d=t_i.s \; 1 \leq j \leq k \} \; \forall \; 1 \leq i \leq n$$

For example table 4 shows the sw value for each disease. This distribution is taken from Table 1.

DEFINTION 2 (SENSITIVE DISCLOSURE FLAG) for each tuple $t \in T$, its *sensitive Disclosure Flag* is indicated as *t.sdf*.

$$t_i.sdf=\{1 \quad \text{if } t_i.sw=0$$

$$ud \quad t_i.sw=1 \; \}\forall \; 1 \leq i \leq n$$

ud represents *user defined* and the value is either *0* or *1*. $t_i.sdf=0$ then user is not ready to disclose his information and $t_i.sdf=1$ then user is ready to disclose his information. In table 5 value of sw and sdf are indicated assuming that sdf value is accepted from record owner for SW=1. We can also observe that if sw=0 its correspondent sdf is initialized to 1 indicating that the sensitivity of this record is not of much relevance.

B. Thresholds for Personalized Privacy

Threshold values are defined for various dimensions of personalized privacy to improve the overall performance of generalization, suppression and disclosure.

i)$TH\rho_n$ minimum number of records in T.

ii)$TH\rho_{iter}$ maximum number of iterations that must be performed . it indicates the amount of generalization & Height(VDH)

iii)$TH\rho_{suppr}$ minimum number of sensitive values for suppression.

iv)$TH\rho_{disc}$ minimum number of sensitive values for disclosure.

v) $TH\rho_{acct}$ minimum threshold that can be added or subtracted.

Since we are considering the distribution aspect we can indicate different threshold values. The first value indicates the minimum number of tuples that must be present for applying anonymization which was never defined in the previous representations. $TH\rho_{iter}$ based on the knowledge of the height of Value domain hierarchy. The larger the value of $TH\rho_{iter}$ higher the generalization and consequently information loss is more. $TH\rho_{suppr}$ indicates the minimum number of sensitive distribution that may be there in QIDB for removal of that block after $TH\rho_{iter}$. $TH\rho_{disc}$ indicates the threshold value that can be added or subtracted to each frequency distribution for each disease such that it is equivalent to the distribution FDB. The frequency of QIDB block and FDB will not be exactly same so while checking the distribution of each disease is checked whether the frequency in that $qidb.v.s \pm TH\rho_{acct}$ always $TH\rho_{disc} > TH\rho_{acct}$.

C. Additional Block Creations for personal privacy

DEFINITION 3 (FREQUENCY DISTRIBUTION BLOCK) Distribution of each $w_j.d$ with respect to the original distribution $t_i.s$ is saved in relation *FDB(d,p)* where *d* indicates disease and *p* indicates probability distribution of it. Each *p* for *d* is calculated by mapping each *d* in *T* (values of $t_i.s=fdb_u.d$) to the total number of tuples in *T* i.e. *n*, $\forall\ 1 \le u \le k$. let us assume there are *m* records in the relation.

DEFINITION 4 (Quasi–Identifier Distribution Block) for each $t_i.s$ where $t_i.sw=1$ & $t_i.sdf=0$ a new *QIDB* is created containing $t_i.s\ \forall\ 1 \le i \le n$. The relation *QIDB.V(q,s)* where $qidb.v_l.q=t_i.q$ & $qidb.v_l.s=t_i.s$. Let us assume there are *dn QIDB* blocks.

For example Table 6 shows the frequency distribution of each disease. This distribution shows that the disease flu is a common disease so its frequency is more, around 50% in the published data. The same distribution is maintained in each of the QIDB. In the first iteration two blocks of QIDB will be created for the qasi value <48602,28> and <48607,55> since its SW=1 & SDF=0 which is shown in table 7 & 8.

TABLE 7. QIDB.1 CONTENTS

ZIP Code	Age	Disease
48602	28	Stomach cancer

TABLE 8. QIDB.2 CONTENTS

ZIP Code	Age	Disease
48607	55	Heart Disease

D. Functions For Personal Privacy

DEFINITION 5 (GENERALIZATION) A general domain of an attribute *T.Q* is given by a generalization function. Given a value *t.q* in the original domain, function returns a generalized value within the domain.

For each $t \in T$ we use t^* to represent its generalized tuple in T^*.we denote it as *G(T)*

This is similar to earlier representations let us assume that Domain Generalization Hierarchy and Value Generalization Hierarchy are defined for each Quasi Identifiers. The distance vector of quasi attributes has also been generated. In figure 1 Value and Domain Generalization Hierarchy of zipcode has been indicated. Age is also generalized similarly. Distance vector is calculated which is shown in figure 2.

DEFINITION 6 (CHECK FREQUENCY) for any *QIDB*, we check *CF(QIDB.V)* wither *QIDB.V* frequency of distribution is equal to the frequency distribution in *FDB.* It is done as follows

Let c be the no of records in *QIDB.V*. for each *UNIQ(qidb.v_l.s)* find total no of mappings which match *qidb.v_l.s* to the no of records i.e. *c* in *QIDB.V*, thus CF will return true if

$\forall\ 1 \le u \le m$ such that $fdb_u.d=qidb.v_l.s$

$$fdb_u.p \approx \frac{UNIQ(qidb.v_l.s)}{c} \pm TH\rho_{acct}$$

this is checked in every iteration if a QIDB satisfies the frequency distribution then this block will not be considered for the next iteration.

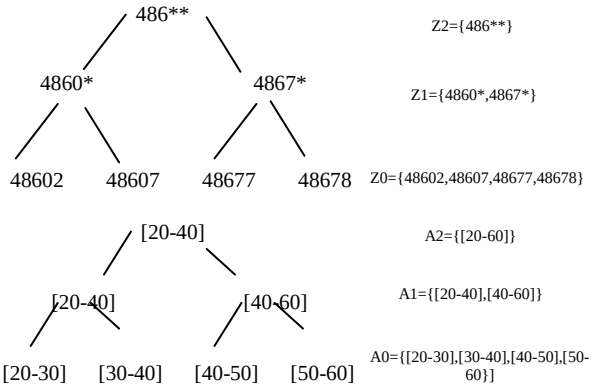

Figure 1. An example of Value and Domain generalization hierarchy for zipcode and Age

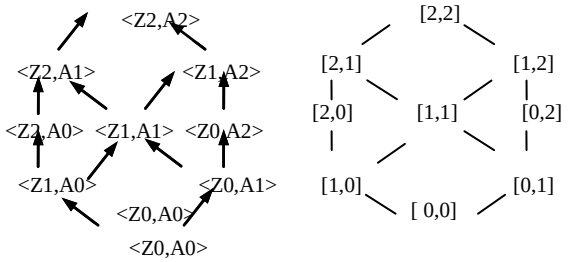

Figure 2. Hierarchy DGH<Z0,A0> and corresponding Hierarchy of distance vectors

DEFINITION 7 (SUPPRESSION) After $TH\rho_{iter}$ *iterations, SUPP(QIDB.v)* suppress the block if it satisfies the following condition

$\forall\ 1 \le u \le m$ such that for every $fdb_u.d=qidb.v_l.s \wedge fdb_u.d=w_j.d \wedge w_j.sw=1\ \forall$ j $1 \le j \le k$

$$Count(qidb.v_l.s) \le TH\rho_{suppr}$$

DEFINITION 8 (DISCLOSURE) After $TH\rho_{iter}$ *iterations, DIS(QIDB.v)* adds additional records if it satisfies the following condition

$\forall\ 1 \le l \le c$ such that for every $fdb_u.d=qidb.v_l.s \wedge fdb_u.d=w_j.d \wedge w_j.sw=1$ for some j $1 \le j \le k$

$$\frac{UNIQ(qidb.v_l.s)}{c} \approx TH\rho_{disc} \pm fdb_u.p$$

III. PERSONALIZED PRIVACY BREACH

Consider an attacker who attempts to infer the sensitive data of a record owner *x*. the worst case scenario assumes that the adversary knows *Q* of *X*, therefore the attacker observes only those tuples $t^* \in T^*$ whose *Q* value $t_i^*.q$ covers *x.q* for all *i* such that $1 \le i \le n$. These tuples form a *Q-group*. That is, if t_i^* and t_{ip}^* are two such tuples then $t_i^*.q=t_{ip}^*.q$ for all *i* such that $1 \le i \le n$.if this group is not formed the attacker cannot infer sensitive attribute of *x*.

DEFINITION 9 (REQUIRED Q-GROUP/ ACT(X)). Given an individual *x*, the Required Q-group *RG(X)* is the only Q-group in *t** covers *x.q*. let us assume *ACT(X)* refers to those records which are generalized to *RG(X)*.

ACT(X) is unknown to the attacker. To obtain *ACT(X)*, the attacker must find some external data base *EXT(X)* that must be covered in *RG(X)*.

DEFINITION 10(EXTERNAL DATA BASE EXT(X)) *EXT(X)* are set of individuals whose value is covered by *RG(X)*

In general ACT(X) $\subseteq$ *EXT(X)*

The attacker adopts a combinational approach to infer sensitive attribute of *x.* let us assume that *x.s* is present in one of t_i^* and the repetition of *x* is not present. The possible reconstruction of the *RG(X)* includes

r distinct record owners $x_1,x_2,x_3,\ldots,x_r$ who belong to *EXT(X)* are taken but there can be only *y* in *RG(X).* this can be understood by the probabilistic nature and can be indicated as perm(r,y). perm(r,y) is Possible Reconstruction(PR) that can be formed by using r owners and y mappings. Breach Probability (BP) indicates the probability of inferred knowledge. Let us assume *ACTN* indicates actual number of records with sensitive attribute that can be inferred to x.

$$BP=\frac{ACTN}{perm(r,y)}$$

BP will decide the privacy parameter, *BP* is 100% then x can be inferred if it is very low than the inference will be very much difficult for the attacker.

IV. *QIDB*-ANONYMIZATION ALGORITHM

In this algorithm we are using a sequential processing of quasi values since the assumption is that in each region usually the distribution of sensitivity is approximately same. The algorithm is as follows

Algorithm *QIDB*-Anonymization

Input: private data *T* with *SW-SDF*, threshold values $THp_n, THp_{iter}, THp_{suppr}, THp_{disc}, THp_{acct}$ and initialized *FDB(d,p)*

Output: publishable table T^*

1. if ($n < THp_n$) then return with *1*

2. for every $t_i.s$ where $t_i.sw=1$ & $t_i.sdf=0$ a new *QIDB* is created containing $t_i.s$ *and* $t_i.q$ $\forall\ 1 \le i \le n$.

3.ini_itr=0, accept_flag=0 and gen=first G(T)

4.while (ini_itr< $THp_{iter}$and accept_flag=0)

4.1.QIDB blocks are removed if CF() returns *true* then check the number of QIDB if it is equal to zero then accept_flag=1

4.2.itr=itr+1 and gen=next G(T)

5. if accept_flag=0 then invoke supp() & dis()

6. check number of QIDB if it is equal to zero accept_flag=1

7. publish T* if accept_flag=1

The resultant anonymization after applying Personal Anonymization of one of the QIDB with THp_{acct} =0.1 block is shown in Table 9.

TABLE 9. RESULTANT SW-SDF BASED QIDB-ANONYMIZATION WITH THp_{ACCT} =0.1

ZIP Code	Age	Disease
486**	[20-40]	Stomach cancer
486**	[20-40]	Gastric ulcer
486**	[20-40]	Heart Disease
486**	[20-40]	Flu
486**	[20-40]	Flu

V. EXPERIMENTS

In this section we try to evaluate the effectiveness of our technique as compared to k-anonymity and l-diversity. We have used a standard dataset used in the literature[7,8,9] for our experiment. We have considered Americal adult dataset of 400 records, with the following quasi attributes Age, Education , Maritial status & Occupation. The attribute age is numerical and the rest of the attributes are categorical. The sensitive attribute income has been converted to disease. Probability is used to find SDF value for SW=1.

We have defined and used generalization hierarchy for each qasi identifier and distance vector is generated which has been used in our algorithm. The maximum height of our generalization hierarchy is 10. Information loss parameter is shown in figure 3. Less the information loss better is the data quality. Minimal distortion (MD)is based on charging penalty for each value which is generalized or suppressed. Each hierarchy is assigned a penalty when it is generalized to the next level with in the domain generalization hierarchy. MD is shown in figure 4. In our experiment we have used a penalty of 10 for every generalization. This Discernibility Metric (DM) calculates the cost by charging a penalty to each tuple for being indistinguishable from other tuples which is shown in figure 5. Execution time is shown in figure 6. For our experiment the threshold values THp_n=400, THp_{iter}=10, THp_{suppr}=1, THp_{disc}=0.01 and THp_{acct}=0.1 was used. Experiment was conducted using Matlab 7 in which our algorithm out performs k-anonymity and l-diversity.

VI. RELATED WORK

Different methods of PPDM exist, among them the most important are Randomization Method [13], Data Swapping [14], Cryptographic Approach [15] and Data Anonymization. Data Anonymization is considered as one of the most important anonymization technique since it has lesser information loss and higher data utility. There are different anonymization algorithms has been proposed in literature [1, 3, 4, 6, 10, 11, 12]. Initial anonymization algorithm was called k-anonymity [6] but the drawback of this approach is that it is prone to record level disclosure. To overcome this

disadvantage *l*-diversity[8] was proposed. Disadvantage is that it is prone to Skewness and Back ground Knowledge Attack. *t*-closeness[9] is used to overcome the disadvantages of l-diversity but it has larger information loss. Personalized Privacy[5] was added on to anonymization which gave lesser information loss. This is the motivation of our approach.

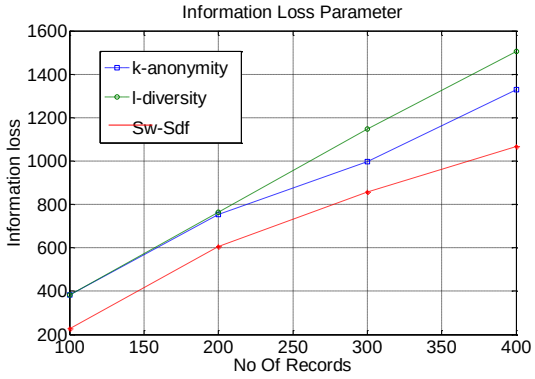

Figure 3. Information Loss Of *SW-SDF* Personal Anonymization As Compared With K-Anonymity & L-Diversiy

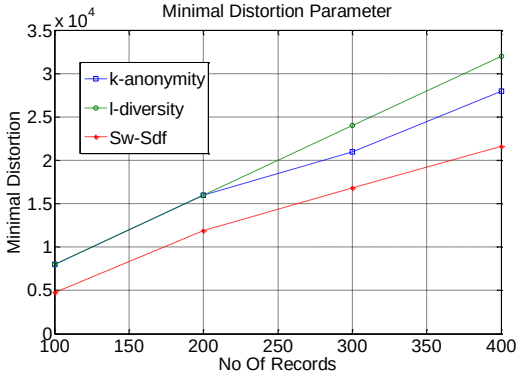

Figure 4. Minimal Distortion Parameter Of SW-SDF Personal Anonymization As Compared With K-Anonymity & L-Diversiy

VII. Conclusions and FutureWork

Personalized privacy is an important research direction in PPDP since its data quality and execution time is less. Usage of *SW* not only improves the indication of sensitivity as the entire records do not require privacy but also improves the data utility. *SDF* is an additional flag which once again improves data utility with in *SW* record since some of the record owners are ready to reveal their identity. Thus the combination of *SW-SDF* is a better option for personalized privacy as compared to just using a guarding node.

QIDB based anonymization allows different quasi group to be generalized independently. In this approach each quidb block is checked for the frequency distribution of sensitive value approximately equal to the frequency distribution of the sensitive value in original contents thereby improving privacy.

It also overcomes record linkage, attribute linkage and even probabilistic attack. This approach works well when the frequency distribution of a particular sensitivity is concentrated within a region of individual pattern.

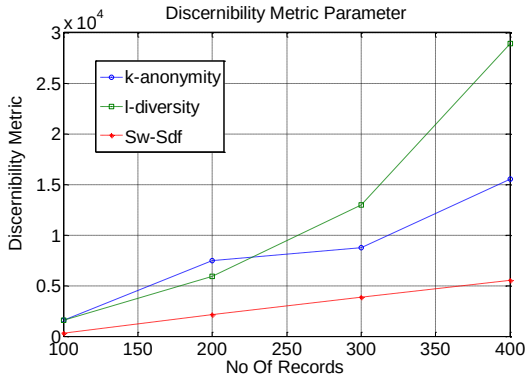

Figure 5. Discernibility Metric Parameter of SW-SDF personal anonymization as compared with k-anonymity & l-diversiy

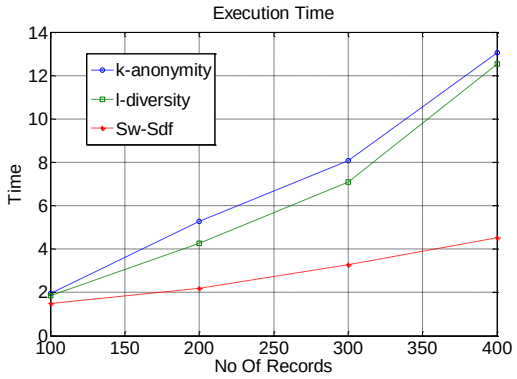

Figure 6. Execution Time of SW-SDF personal anonymization as compared with k-anonymity & l-diversiy

There are several future research directions along the way of analyzing *SW-SDF* personal privacy with *QIDB* anonymization. First we haven't considered the effect of Sequential Release and Multiple Release of published data. Research on giving different Weight on sensitivity can be considered. In this approach we have used sequential processing of records to check the generalized record matches with *QIDB* generalized value if they are same then it would be included in the block. Instead of sequential processing alternative methods can be looked in to. This method can be extended to unstructured schema and multi-dimensional data.

References

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

[2] Fung B. C. M., Wang K. and Yu P. S., "Anonymizing classification data for privacy preservation". IEEE Trans. Knowl. Data Engin, pp. 711–725,2007.

[3] Fung B. C. M.,Wang K. and Yu P. S., " Top-down specialization for information and privacy preservation". In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE). 205–216, 2005.

[4] Iyengar V. S., "Transforming data to satisfy privacy constraints". In Proceedings of the 8th ACMSIGKDD. ACM, New York, pp. 279–288,2002.

[5] Xiao X. and Tao Y., "Personalized privacy preservation". In Proceedings of the ACM SIGMOD Conference.ACM, New York, 2006.

[6] P. Samarati, "Protecting Respondent's Privacy in Microdata Release". IEEE Trans. on Knowledge and Data Engineering (TKDE), vol. 13, no.6, pp. 1010-1027, 2001.

[7] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy". Int'l J.Uncertain. Fuzz., vol. 10, no. 5, pp. 557-570, 2002.

[8] Machanavajjhala A, Gehrke J, Kifer D and Venkitasubramaniam M, "l-diversity: Privacy beyond k-anonymity". In Proceedings of the 22nd IEEE International Conference on Data Engineering(ICDE), 2006.

[9] Ninghui Li , Tiancheng Li , Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l–Diversity". ICDE Conference, 2007.

[10] R. Bayardo and R. Agrawal. "Data privacy through optimal k-anonymization". In ICDE, pp. 217–228, 2005.

[11] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian multidimensional k-anonymity". In ICDE, 2006.

[12] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection". In *ICDM*, pp. 249–256,2004.

[13] Agrawal D. Aggarwal C. C., "On the Design and Quantification of Privacy- Preserving Data Mining Algorithms". ACM PODS Conference, 2002.

[14] Fienberg S.,McIntyre J., "Data Swapping: Variations on a Theme by Dalenius and Reiss". Technical Report, National Institute of Statistical Sciences, 2003.

[15] Pinkas B.,"Cryptographic Techniques for Privacy-Preserving Data Mining". ACM SIGKDD Explorations, 2002.