Privacy Protection Using Generalization For Collaborative Data Publishing

Malgireddy Saidireddy,
Associate Professor
Dept of CSE
KG Reddy College of Engineering and Technology, Hyderabad

***Abstract:* -** Organizations share their data about customers for exploring potential business avenues. The sharing of data has posed several threats leading to individual identification. Owing to this, privacy preserving data publication has become an important research problem. The main goals of this problem are to preserve privacy of individuals while revealing useful information. An organization may implement and follow its privacy policy. But when two companies share information about a common set of individuals, and if their privacy policies differ, it is likely that there is privacy breach unless there is a common policy. One such solution was proposed for such a scenario, based on k-anonymity and cut-tree method for 2-party data. This paper suggests a simple solution for integrating n-party data using dynamic programming on subsets. The solution is based on thresholds for privacy and in formativeness based on k-anonymity.

Keywords: - Privacy preserving, data mining, k-anonymity, collaborative data publishing, dynamic programming.

I. INTRODUCTION

With numerous organizations collecting customer data, there exists a possibility of data sharing for exploring interesting data about behavior of customers [1]. This leads to identification of customers which can be treated as a privacy threat according to HIPAA[2] and EU directives[3]. These acts insist that anonymity should be guaranteed if the customers wish so.

A customer data normally contains attributes like SSN, name, age, postal code, date of birth and gender. This data enables identification of the individuals even though information like SSN and Name suppressed. This was first identified in [4]. The solution proposed k-anonymity property to be applied to the data before release. Subsequently several solutions were published. Most of them addressed issues related to preserving privacy of individuals related to a single organization [1, 5, 6]. This paper discusses an approach to protect privacy when anonymized data of two or more organizations is integrated.

1.1 Motivation

In real-life data publishing single organization often does not hold the complete data. Organizations need to share data for mutual benefits or for publishing to a third party. For example, banking sectors want to integrate their customer data for developing a system to provide better services for its customers. However, the banks do not want to indiscriminately disclose their data to each other for reasons such as privacy protection and business competitiveness. Figure 1 depicts this scenario, called collaborative data publishing, where several data publishers own different sets of attributes on the same set of records and want to publish the integrated data on all attributes. Say, publisher 1 owns {*Rec ID*, *Job*, *Sex*, *Age*}, and publisher 2owns {*RecID*, *Salary*, *Disease*}, where *Rec ID*, such as the *SSN*, is the record identifier shared by all data publishers. They want to publish an integrated *k*-anonymous table on all attributes. Also, no data publisher should learn more specific information, owned by the other data publishers, than the information that appears in the final integrated table.

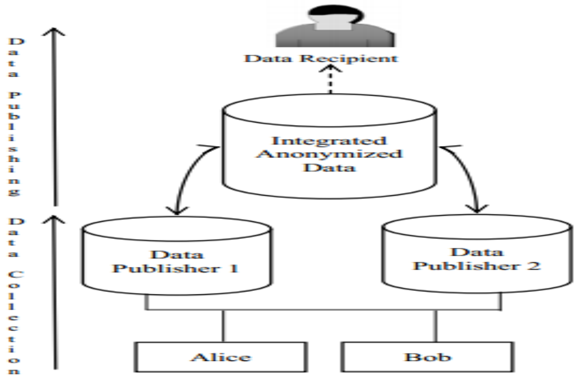

Figure1. Collaborative data publishing

The rest of the paper proceeds as follows. Section 2discourses the related work while section 3discusses the system architecture. Section 4 discusses the problem definition. Section 5 highlights secure data integration. Section 6 analyses our approach when compared to other published work. Section 7 concludes our work.

II. RELATEDWORK

The organizations share their data with many other research communities for various uses. Today technologies are providing easy way of information sharing. However sharing the data with outsiders should not reveal the individual identification of a person [7]. Care must be taken to provide the privacy for the person specific data at the time of publishing personal information for research purposes. The objective of privacy preserving mining is that this data, when published should not link back to the individual.

The notion of *k*-anonymity was proposed in [8], and generalization was used to achieve *k*-anonymity in Data fly system [9] and µ-Argus system [10]. All these works considered a single data source; therefore, data integration is not an issue. In the case of multiple private databases, joining all databases and applying a single table method would violate the privacy constraint private databases.

Information integration has been an active area of database research. This literature typically assumes that all information in each database can be freely shared [11]. Secure multiparty computation (SMC), on the other hand, allows sharing of the computed result, but completely prohibits sharing of data [12]. Liang et al. [13] and Agrawal et al. [11] proposed the notion of minimal information sharing for computing queries spanning private databases. They considered computing intersection, intersection size, equijoin and equijoin size. Their model still prohibits the sharing of databases themselves.

K.Wang et al [14] made two contributions. First, they defined the secure data integration problem. The goal is to allow data sharing in the presence of privacy concern. In comparison, classic data integration assumes that all information in private databases can be freely shared, whereas secure multiparty computation allows "result sharing" but completely prohibits data sharing. In many applications, being able to access the actual data not only leads to superior results, but also is a necessity. Second, they presented a solution to secure data integration where the two parties cooperate to generalize data by exchanging information not more specific than what they agree to share.

Jiang and Clifton [15, 16] addressed a similar problem by using a cryptographic approach. First, each data publisher determines a locally *k*-anonymous table. Then, the intersection of *RecID*s for the *qid* groups in the two locally *k*-anonymous tables is determined. If the intersection size of each pair of the *qid* group is at least *k*, then the algorithm returns the join of the two locally *k*-anonymous tables that is globally *k*-anonymous; otherwise, further generalization is performed on both tables and the *RecID* comparison procedure is repeated.

Pawel Jurczyk and Li Xiong [17] presented a distributed and decentralized anonymization approach for privacy-preserving data publishing for horizontally partitioned databases. This work addresses two important issues, namely, privacy of data subjects and privacy of data providers. They presented a new notion, l-site-diversity, to achieve anonymity for data providers in anonymized dataset.

Our solution is based on the problem and scenarios stated in [14].

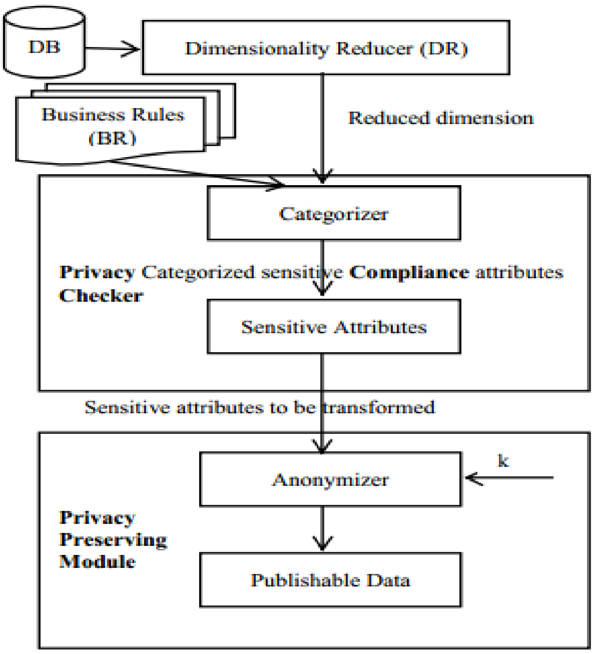
DB
Dimensionality Reducer (DR)
Business Rules (BR)
Reduced dimension
Categorizer
Privacy Categorized sensitive Compliance attributes
Checker
Sensitive Attributes
Sensitive attributes to be transformed
Anonymizer
k
Privacy Preserving Module
Publishable Data

Figure 2. System Architecture

III. SYSTEM ARCHITECTURE

This model (Figure 2) primarily has two objectives: preserving privacy while revealing useful information for sensitive attributes and to find an integrated table without loss of information. This involves the following steps:

1. Dimensionality reduction: Suppressing the unnecessary attributes.
2. Identifying sensitive attributes through business rules 3. Categorizing the attributes (Categorizer)
4. Use anonymizer for preserving linkage of individual with sensitive categorical attributes.
5. Release the anonymized data and announce the joint anonymity property.
6. Perform data integration without revealing any sensitive information and its associated individual.

3.1 Dimensionality reduction

Dimensionality reduction and attribute selection aim at choosing a sub set of attributes sufficient to describe the data set. The goal of the methods designed for dimensionality reduction is to map *d*-dimensional objects into *k*-dimensional objects, where $k<d$.

Dimensionality reduction is beneficial only when the loss of information is not critical to the solution or the problem, or if more information is gained by the visualization of the problem than what is lost. Reduction from dimension *d* to *k* (*k*<*d*) reduces complexity, reduces communication cost and provides privacy since extra data given may help in re-identifying individuals or loss of sensitive information vulnerable for second use.

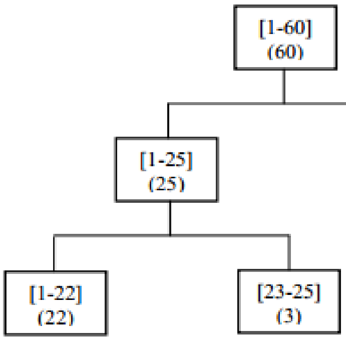

Figure 4. Taxonomy for Age

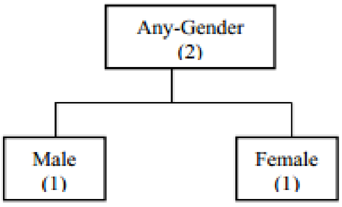

Figure 5. Taxonomy for Gender

3.2 Categorizing attributes

The attributes in the reduced table are classified as identifiers, sensitive attributes and quasi identifiers.

IV. PROBLEMDEFINITION

Consider the data in Table 1 and taxonomy trees in Figures3, 4 and 5. Party A and Party B own *TA* (*SSN; Gender;. . . . . . . ; Class*) and *TB*(*SSN; Education; Age; . . . . . . ; Class*) respectively. After joining the two tables on SSN, the “female, mechanical" on (*Gender; Education*) becomes unique, therefore, vulnerable to be linked to sensitive information such as *Age*. To protect against such linking, we can generalize *Civil* and Mechanical to Non-Computers so that this individual becomes one of many female professionals. However we preserve information as the basic classification is not changed.

Definition 1(*k*-anonymity): Consider *p* quasi-identifiers QID1, ..., QIDp on T. Let ri denote the number of records in T that share the value qidi on QIDi. The anonymity of QIDi, denoted ri, is the smallest ri for any value qidi on QIDi. A table T satisfies the anonymity requirement {<QID1; k1>, ..., <QIDp; kp> } if ri ≥ ki for 1 ≤ i ≤ p, where ki is the anonymity threshold on QIDi[17].If QIDj is a subset of QIDi, where i ≠ j, and if kj ≤ ki, then <QIDj; kj> is implied by <QIDi; ki>, therefore, can be removed.

Table 1. Compressed Table

Shared		Party A		Party B			
SSN	Class	Gender	…	Education	Age	…	# of Recs
1-3	0Y3N	M		Life Sc.	17		3
4-7	0Y4N	M		Physical Sc.	18		4
8-12	2Y3N	M		Fine Arts	19		5
13-16	3Y1N	F		Journalism	24		4
17-22	4Y2N	F		Computers	26		6
23-25	3Y0N	F		Computers	27		3
26-28	3Y0N	M		Civil	27		3
29-31	3Y0N	F		Civil	27		3
32-33	2Y0N	M		Mechanical	27		2
34	1Y0N	F		Mechanical	27		1
Total	21Y13N						34

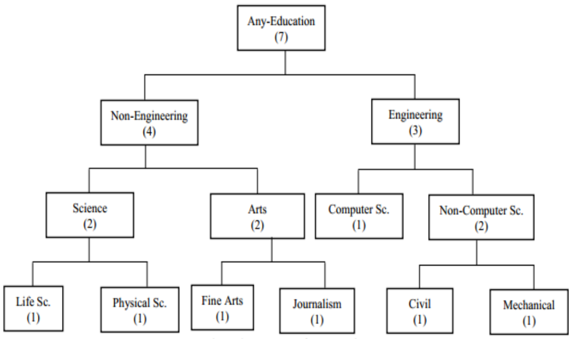

Figure 3. Taxonomy for Education

V. PRIVACY PRESERVING DATA INTEGRATION

The k-anonymity policies of an organization define information access threshold for its database. The threshold is the minimum amount of generalization required for giving information. Given two or more organizations who want to share their data without revealing information, the privacy preserving integration is to determine an optimal combination of attributes to disclose information while preserving privacy. Thus each organization may define privacy strength for each of their data sources and a joint anonymity requirement (Figure 6).
Generalization preserves the privacy whereas specialization makes it more informative.

Definition 2(Secure data integration [5]): Given two private tables TA and TB, a joint anonymity requirement {<QID1, K1>, ……, <QIDP, KP>} and a taxonomy tree for each attribute, the secure data integration is to produce a generalized integrated table T* such that it satisfies the joint anonymity requirement and retains as much information as possible.

Each node in the taxonomy is associated with a privacy strength value. The values vary for categorical data and continuous data. For continuous data the difference between ranges is taken as the privacy strength of the nodes. For categorical data the leaf nodes are given values „1‟ and for the internal nodes the privacy strength is the number of leaf nodes that each node has.

*Privacy strength (P)*of a given node is the number of leaves in the in the sub tree with this node as the root.

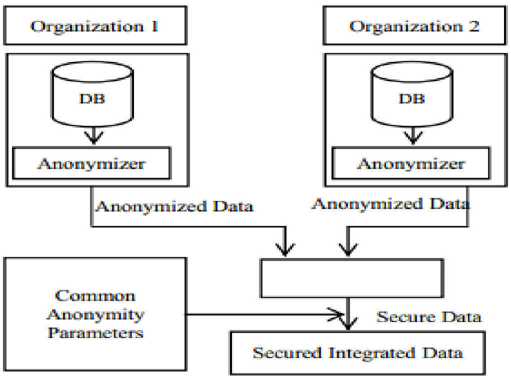

Figure 6. Secure Data Integration

VI. ANALYSIS

Dynamic Programming is a method for efficiently solving a broad range of search and optimization problems which exhibit the characteristics of overlapping sub-problems and optimal substructures.

The principle of optimality is the basic principle of dynamic programming, which was developed by Richard Bellman: that an optimal path has the property that whatever the initial conditions and control variables (choices) over some initial period, the control (or decision variables) chosen over the remaining period must be optimal for the remaining problem, with the state resulting from the early decisions taken to be the initial condition.

For our problem in this paper, we use the principle of knapsack model using dynamic programming.
The most common formulation of the problem is the 0-1 knapsack problem, which restricts the number *xi* of copies of each kind of item to zero or one. Mathematically the 0-1-knapsack problem can be formulated as: $subject$ $\leq W,$ $x_i \in \{0, 1\}$

In our problem, we have „*n'* kinds of possible subsets, 1through *n* as shown in Table 3. Each kind of possibility „*i'* in the subset has a privacy value *Pi* and an informativeness value *Ii*. The best possible subset that produces maximum privacy and maximum informativeness is chosen. Here the size of the knapsack is the cardinality of the data to be published. Here Si is the subset that satisfies the anonymity parameter.

$$maximize \sum_{i=1}^{n} P_i$$

$$maximize \sum_{i=1}^{n} I_i$$

$$subject\ to \sum_{i=1}^{n} S_i \geq k \qquad to \quad S_i \geq$$

The computation of privacy strengths for the taxonomies in figures 3, 4 and 5 are given below by indicating the privacy strength in the parenthesis.
Education=*{Any-Education(7), Engineering(3), Non-Engineering(4), Science(2), Arts(2), Computer Science(1), Non-Computer Science(2), Life(1), Physical(1), Fine Arts(1), Journalism(1), Civil(1), Mechanical(1)}*

Age=*{[1-60](60), [1-25](25), [26-60](35), [1-22](22), [22-25](3)}*

Gender=*{Any-Gender(2), Male(1), Female(1)}*

Suppose if the QID is {Gender, Age} then the possible and valid subsets would be as specified below in Set1.

Set1 = { (Any-Gender, [1-60]), (Any-Gender, [1-25]), (Any-Gender, [26-60]), (Any-Gender, [1-22]), (Any-Gender, [23-25]),(Male, [1-60]), (Male, [1-25]), (Male, [26-60]), (Male, [1-22]), (Male , [23-25],(Female, [1-60]), (Female, [1-25]), (Female, [26-60]), (Female, [1-22]), (Female , [23-25] }

$$\text{maximize} \quad \sum_{i=1}^{n} v_i x_i$$

For each subset, k-anonymityrequirement is verified. If the required threshold is satisfied, the Privacy strength(P), Informativeness(I) are computed. These values are recorded along with the respective subset and the threshold value. After considering all the possible subset, the subsets that have been recorded are taken into consideration. From these subsets, the supersets if any, are identified and discarded from the set. From the remaining subsets, the subset that produces the optimal solution is taken and the data is published satisfying the required threshold.

$$\text{to} \sum_{i=1}^{n} w_i x_i$$

Algorithm: Collaborative Data Publishing

Input: Integrated table that contains data of both the parties.
Output: Optimal anonym zed table.
Step1: identify all the QID sets. { QID1, QID2, ………, QIDn }
Step2: for each QIDi that belongs to QID set, generate the power set.
Step3: for each qidj combination in the power set of QIDi
Step4: generate the corresponding equivalence classes
Step5: compute the relative Privacy strength(P), Informativeness(I)
Step6: end for
Step7: end for
Step8: discard the equivalence classes that do not satisfy the threshold.
Step8: find the equivalence classes that provide optimal solution by considering Privacy strength and Informativeness.
Step9: Publish the anonymized data.
For instance see table 2, let us consider one combination and observe the resultant dataset for a threshold value k=3. Here the qid is <Any-Gender, Any-Education>.

$$r_i$$

$$\text{subject} \sum_{i=1}^{n} k$$

From table 2, we observe that the privacy strength, P = {2, 7}; the informativeness, I = {0, 0}; the number of equivalence classes = 6 and the threshold value, k = 3.

The combinations that satisfy the threshold value would only be considered for analysis. The same is repeated for remaining combinations. From these sets, a combination value is selected such that it produces an optimal value where it provides maximum privacy and more information.

Table 2. Anonymized table

Class (Shared	Gender	...	Education	Age	…	# of Recs
0Y3N	Any-Gende		Any-Educatio	17		3
0Y4N	Any-Gende		Any-Educatio	18		4
2Y3N	Any-Gende		Any-Educatio	19		5
3Y1N	Any-Gende		Any-Educatio	24		4
4Y2N	Any-Gende		Any-Educatio	26		6
12Y0N	Any-Gende		Any-Educatio	27		12

The dataset for the combination <Gender, Education>is given in table 3.

Table 3. Subsets satisfying the threshold (ki)

QID	PS(P)	Info(I)	Anony (k_i)	# Recs
Any-Gender, Any-Education	(2,7)	(0,0)	3	34
Any-Gender, Non-Engineering	(2,3)	(0,0.33)	3	16
Any-Gender, Engineering	(2,4)	(0,0.33)	6	18
Any-Gender, Science	(2,2)	(0,0.66)	3	07
Any-Gender, Arts	(2,2)	(0,0.66)	3	09
Any-Gender, Computers	(2,1)	(0,1)	3	09
Any-Gender, Non-Computers	(2,2)	(0,0.66)	9	09
Any-Gender, Life Sc	(2,1)	(0,1)	3	03
Any-Gender, Physical Sc	(2,1)	(0,1)	4	04
Any-Gender, Fine Arts	(2,1)	(0,1)	5	05
Any-Gender, Journalism	(2,1)	(0,1)	4	04
Any-Gender, Civil	(2,1)	(0,1)	6	06
Any-Gender, Mechanical	(2,1)	(0,1)	3	03
Male, Any-Education	**(1,7)**	**(1,0)**	**3**	**17**
Male, Non-Engineering	**(1,3)**	**(1,0.33)**	**3**	**12**
Male, Engineering	**(1,4)**	**(1,0.33)**	**5**	**05**
Male, Science	(1,2)	(1,0.66)	3	07
Male, Arts	(1,2)	(1,0.66)	5	05
Male, Computers	(1,2)	(1,0.66)	5	05
Male, Life Sc	(1,1)	(1,1)	3	03
Male, Physical Sc	(1,1)	(1,1)	4	04
Male, Fine Arts	(1,1)	(1,1)	5	05
Male, Civil	(1,1)	(1,1)	3	03
Female, Any-Education	(1,7)	(1,0)	4	17
Female, Non-Engineering	(1,3)	(1,0.33)	4	04
Female, Engineering	(1,4)	(1,0.33)	6	13
Female, Arts	(1,2)	(1,0.66)	4	04
Female, Manager	(1,1)	(1,1)	3	09
Female, Computers	(1,2)	(1,0.66)	4	04
Female, Journalism	(1,1)	(1,1)	4	04
Female, Civil	(1,1)	(1,1)	3	03

From the generated subsets, we select the subsets for which we get the maximum privacy strength value and maximum informativeness value. On applying the dynamic programming principle, the highlighted rows in Table 3, give the subsetcombinations that provide the optimal result and are to be selected. The correlation between Privacy strength and Informativeness is shown in figure 7.

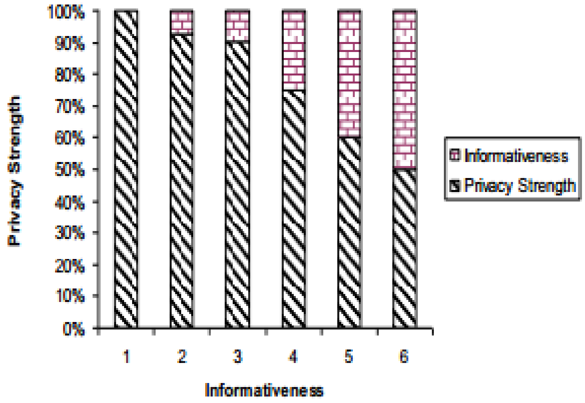

Figure 7. Privacy strength Vs. Informativeness

VII. CONCLUSION

The paper proposed a solution for achieving anonymity when data from two organizations with common privacy policy are integrated. The solution is a simple and effective method as it uses cost effective algorithms for achieving anonymity. The solution proposed in [14] is based on tree data structure called TIPS. We base our solution on subset generation and selecting the most relevant subset. We are currently examining the feasibility of this approach for achieving anonymity on the fly in dynamically growing databases.

REFERENCES

[1] M-Privacy For Collaborative Data Publishing, IEEE Transactions On Knowledge And Data Engineering Vol:Pp No:99 Year 2013.
[2] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.
[3] Council Directive (EC) 2001/29/EC of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society.
[4] L. Sweeny. : K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,2002.
[5] K. Wang, P. S. Yu, and S. Chakraborty. : Bottom-up generalization: A data miningSolution to privacy protection. In *ICDM*, 2004.
[6] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *IEEE ICDM*, November 2005.
[7] R. Agrawal and R. Srikant. : Privacy-Preserving Data Mining. In Proceedingsof the ACM SIGMOD International Conference on Management of Data, Dallas, Texas, May 2000.
[8] Dalenius, T.: Finding a needle in a haystack - or identifying anonymous census record. Journal of Official Statistics, 1986.
[9] Sweeney, L.: Achieving k-anonymity privacy protection using generalization andsuppression. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 2002.
[10] Hundepool, A., Willenborg, L.: μ- and τ-argus: Software for statistical disclosurecontrol. In: Third International Seminar on Statistical Confidentiality, Bled, 1996.
[11] Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across privatedatabases. In: Proceedings of theACM SIGMOD International Conference on Management of Data, San Diego, California, 2003.