Preserving Privacy in Set Valued Data using Collaborative Publishing Scheme with Sensitive Data Hiding

Nanna Babu Palla
Research Scholar, JNTU Hyderabad
Telangana, India.

Dr A Vinaya Babu
Professor in Dept of CSE
JNTU Hyderabad, Telangana, India

***Abstract*— Data mining and warehousing systems usually deployed for information extraction, collection and maintenance respectively. It's a passion for many organizations to analyzing the data from general to intended applications. Data miners and researchers may apply sophisticated knowledge extraction algorithms on available data. The extracted information may contain sensitive data pertaining to an individual. Privacy Preserving Data Mining (PPDM) assert on protecting sensitive knowledge related to human being when side effects of mining occurs .it's an indispensable process to enforce sensitive data hiding techniques to mitigate the identity disclosure. This paper focus on multi provider environment where data sharing is permissible but prone to background knowledge attack as each provider conspire each other for identity leakage of an individual. We propose new approach named Collaborative Publishing Scheme (CPS) on set valued data a fragment from large data set. In this paper, we implemented anonymity CPS approach which shown resiliency to adversaries such as background knowledge attacks.**

Keywords*— Privacy, data mining, multi provider environment, sensitive data, data publishing, collaborative publishing scheme, background knowledge attack

I.INTRODUCTION

Modern information systems produce petabytes of data every day from various sources. They are grown from infant, centralized systems to the echelon of distributed and ubiquitous computational domains. Medicare data maintained by various organizations and hospitals may contain patient information like occupation, gender, age, treating disease, consulting doctor which contains sensitive attributes related to an individual. Sensitive data is defined as the information related to specific individual encompassing large range of information that includes ethnic,racial,health or personal life. Applying data mining techniques on Medicare data may reveal sensitive knowledge which should be protected .data mining helps in interpreting, retrieving and identifying relevant matching patterns. Data collection, publishing and sharing activities are common accession in government and private organizations for conjoint give-and-take policy. The acquired medical data might be exploited for patient population study systems, disease predictive systems, and public-health monitoring systems and consumed by government agencies, drug –manufacturing companies, insurance agencies, public –health safety institutions. In this paper, we identified a problem, where in Integrated Patient Information Systems (IPIS) which will be maintained by multiple organizations called multiple providers which contains patient information and their health profile along with donor details as mentioned in Figure 1.

When the data is shared by organizations and third party agencies (TPA) for mutual benefit, then the privacy of individual is under vulnerability. This issue raises remedial for privacy protection. The sensitive data must be safeguard against adversaries like background knowledge attack (BKA) and Insider attacks.

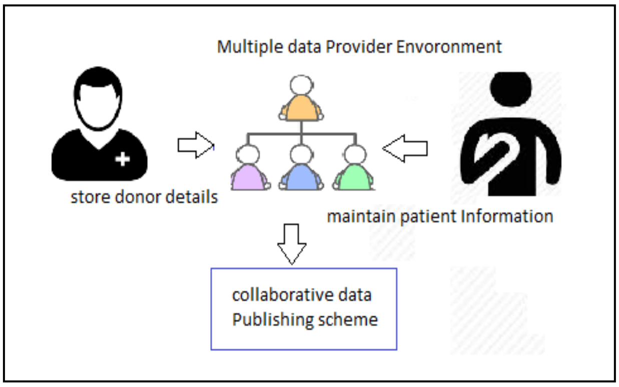
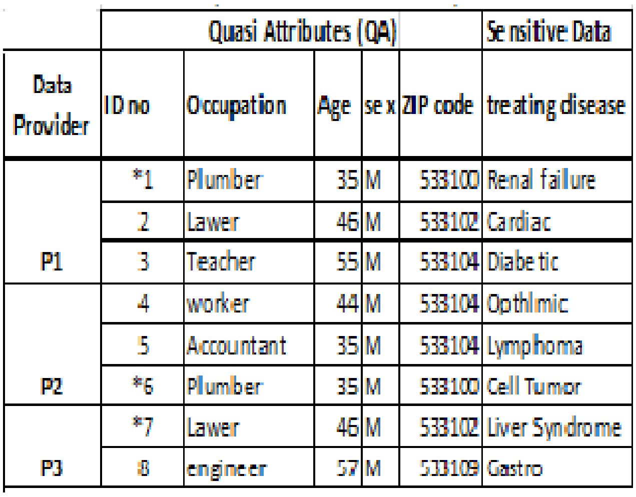

Multiple data Provider Envoronment
store donor details
maintain patient Information
collaborative data
Publishing scheme

Fig 1.Multiple data provider environment which maintains both donor and patient health profile.

II .PROBLEM DESCRIPTION

TABLE I SAMPLE DATABASE MAINTAINTED BY IPIS



	Quasi Attributes (QA)					Sensitive Data
Data Provider	ID no	Occupation	Age	sex	ZIP code	treating disease
P1	*1	Plumber	35	M	533100	Renal failure
	2	Lawer	46	M	533102	Cardiac
	3	Teacher	55	M	533104	Diabetic
P2	4	worker	44	M	533104	Opthlmic
	5	Accountant	35	M	533104	Lymphoma
	*6	Plumber	35	M	533100	Cell Tumor
P3	*7	Lawer	46	M	533102	Liver Syndrome
	8	engineer	57	M	533109	Gastro

In the Table 1, the common attributed maintained by IDIS are constructed with (*occupation, age, sex, ZIP and treating* *disease*). The *disease* is a sensitive data attribute. For the above table, the quasi –identifier set is QI=(occupation ,age,sex,ZIP). Linking this quasi –identifier attributes with other public domain available data, an individual identity can be easily disclosed. From the given set valued fragment obtained from large data set, 3 data providers named with P1,P2 and P3 offering both donor and patient information. The patient ID's with *1,*6,*7 are serving from multiple hospitals whose information available at multiple sources. This leads to the background knowledge attack if each provider intended for conspiracy of data leakage.

Patient Id with 1 is availing services for renal failure and cell tumor from two serving hospitals P1 and P2 with same patient profile as 2 and 7. If the data is published without implementing the anonymity for multi provider environment, then sensitive information will be misused by researchers and demographers. Data miners and demographers use them for health monitoring case studies, epidemic disease studies, public health surveys and insurance claiming approval process. Suppressing and hiding susceptible information by anonymity models, randomization techniques, data perturbation approaches is highly significant. Our work focus on Integrated Patient Information System (IPIS) which contain patient and donor information and employs collaborative publishing scheme (CPS) and analyzed the background knowledge attack w.r.t multi provider environment.

III. PREVIOUS WORK

This part focus on milestone in sensitive information hiding methods in early developments for privacy protection. Early techniques cover on sanitization of data on single source data providers. This paper focus on implementing privacy protection for multiple data providers systems. Multiple providers will maintain their organizational data at their own, but few occasions will motivate them for data exchange and sharing .this arises data leakage issues and background knowledge attack. Mohammed et.al described centralized and distributed anonymity methods for high volume health care data [1]. Fig 2. Describes anonymization in centralized systems, if any sensitive information is exists, then data perturbation, distortion and masking methods are to be implemented [2].

IV.PROPOSED METHODOLOGY AND ALGORITHM

It contains accessing medical data from Multiple providers named with P1 to Pn and collectively referred as T*, which is a set valued data from a large data set D, having A1, A2, and A3 … An as attributes. The quasi identifiers QI = Ai*. Integrated data is accessed from multiple heterogeneous sources with patient profile and then applies join () operation which performs data integration across multiple sources into one database T*.

The anonymization is then applied with Collaborative Publishing Scheme (CPS).

A. Algorithm for Collaborative Publishing Scheme

//Algorithm for collaborative publishing for sensitive data hiding in set valued data

```
Algorithm Collaborative Publish ( T*)
// reads integrated Medicare data collected from multiple
sources as input argument
//these records may contain occupation, age, sex and ZIP
code, disease information

// T* is integrated database of patients fetched from assorted
hospitals
Input: Health care data fetched from different data providers
Output: sanitize the data by hiding sensitive information
using collaborative scheme
{

Step 1: Create authentication credentials for publisher and
researcher

Step 2: Load data collected from hetero sources T* called as
set value data

Step 3: Set the sensitive data attributes for given database
A.  Step 4: Apply Combine( )function ,which now contains
    information of all individuals
Step 5: apply anonymity using generalization or masking
Publish anonymised data for miners and researchers using
collaborative publishing scheme
} // end of algorithm
```

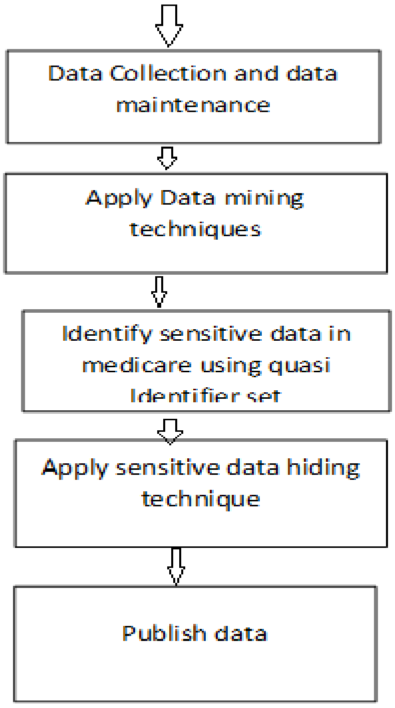

Data Collection and data maintenance
Apply Data mining techniques
Identify sensitive data in medicare using quasi Identifier set
Apply sensitive data hiding technique
Publish data

Fig 2. Process flow diagram used in anonymization approach

The proposed framework contains n number of data providers, each providers identified with P1, P2 and Pn. T* is an Integrated data collected from heterogeneous data sources with patient health profile. The researcher R applies data mining techniques like clustering, association, classification for analyzing the data which is a sanitized data with hiding the sensitive attributes. The background knowledge attack can be mitigated by reducing the data leakage of each provider at the earliest

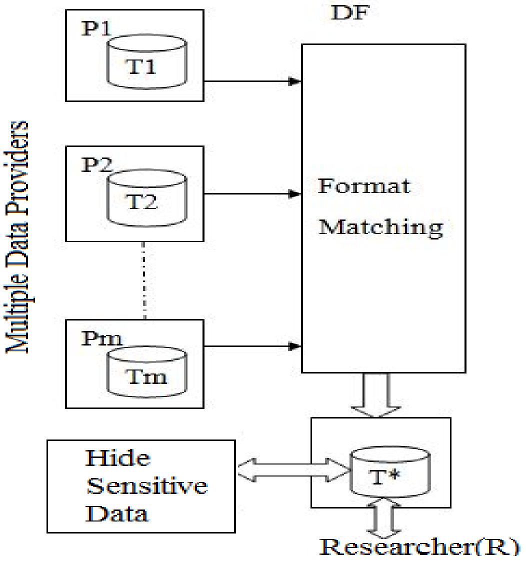
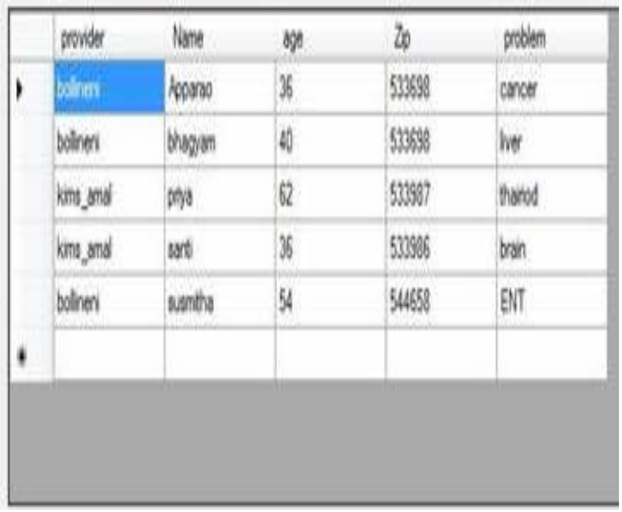


Fig 3.Architectural diagram showing multi data providers (P1 to Pm) offering patient details and researcher may use data mining technique to extract matching, relevant interested patterns.sensitve data hiding process using collaborative method applied on T*.

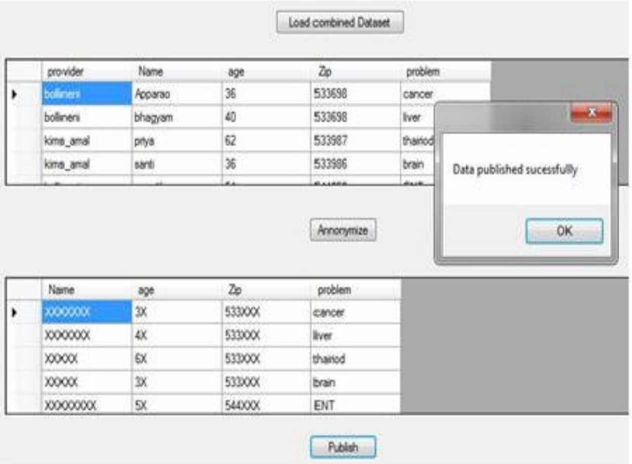
B. Experimental Results and Discussion

The interface accepts various individual data who are availing medical services, then data integration from multi provider data environment system is performed. The proposed Collaborative Publishing Scheme (CPS) is applied on the integrated data implementing anonymity with perturbation technique.



provider	Name	age	Zip	problem
bollineni	Apparao	36	533698	cancer
bollineni	bhagyam	40	533698	liver
kims_amal	priya	62	533987	thariod
kims_amal	santi	36	533986	brain
bollineni	susmitha	54	544658	ENT

Fig 4 .Interface which accepts patient health profile with name, age, sex, ZIP and disease details.

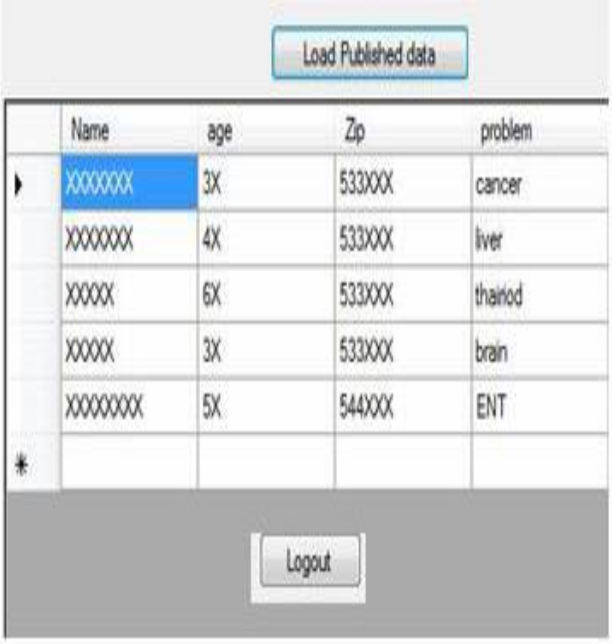


provider	Name	age	Zip	problem
bollineni	Apparao	36	533698	cancer
bollineni	bhagyam	40	533698	liver
kims_amal	prya	62	533987	thariod
kims_amal	santi	36	533986	brain

Data published sucessfully

Name	age	Zip	problem
XXXXXXX	3X	533XXX	cancer
XXXXXXX	4X	533XXX	liver
XXXXX	6X	533XXX	thariod
XXXXX	3X	533XXX	brain
XXXXXXXX	5X	544XXX	ENT

Fig 5 .data integration performed with multiple data providers



Name	age	Zip	problem
XXXXXX	3X	533XXX	cancer
XXXXXX	4X	533XXX	liver
XXXXX	6X	533XXX	thariod
XXXXX	3X	533XXX	brain
XXXXXXX	5X	544XXX	ENT

Fig 6. Collaborative Publishing Scheme (CPS) performed on T* to hide the sensitive data

C. Performance Issues and Evaluation

The proposed algorithm for Multi Provider Data Environment Systems (MPDES) as in IPIS, collaborative publishing scheme applies integration of data across multiple sources and implementing anonymization on the combined data which is ready for data publishing for the purpose of researchers and demographers. They can utilize this published data which is sanitized outcome pruning sensitive information from the raw data. The performance of Collaborative publishing scheme depends on number of data providers and their data storage formats. If similar database patterns are maintained, then the performance and access time are observer with high values.

Case 1: if number of providers is more with data incompatible, then the performance will be aggravated.

Case 2: if number of providers is more with homogeneous storage compatibility, performance is optimal.

Case 3: if optimal number of providers with homogeneous storage compatibility, then performance is high.

V. CONCLUSION

Preserving privacy with sensitive data hiding is an indispensable phenomenon in data publishing arena. With the advent of internet and novel computing methods, the identity disclosure is easier. The stored data can be published with high utility without losing its significance for research and analysis by data miners and analysts. Our approach called Collaborative Publishing Scheme (CPS) is an effective approach for publishing sensitive data across multi provider data environment systems. It safeguards individual sensitive data with purging pre knowledge attacks. Hence data owner can release the private information by guarantying confidentiality and trust to an individual. This approach offers an immune to Background Knowledge (BK) by improving trust and privacy preservation while sharing and publishing healthcare data by an individual. This work can be extensively measured with data utility factor keeping an optimal data distortion for future data analysis.

ACKNOWLEDGEMENT

The authors wish to thank Integrated Patient Information Systems (IPIS) supported from various network hospitals in and around rajahmundry,Andhrapradesh for providing such valuable information.

REFERENCES

[1] 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.

[2] Sweeney, L. "*K-anonymity*: A model for protecting privacy" *International journal on uncertainty and . Fuzz. Knowledge .Based Systems, 2002.*

[3] 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.

[4] D. Agarwal and C.C.Aggarwal, " On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th Symposium on Principles of Database systems,Santabarbara,California,USA, May2001

[5] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., and Theodoridis,Y. 2004. "State-of-the-art in privacy preserving data mining". *ACM SIGMOD Record 33*, 50-57.

[6] P Kamakshi, A Vinaya Babu "preserving privacy and sharing data in distributed environment using cryptographic technique on perturbated data" Journal Of Computing, Volume 2, Issue 4, April 2010,pp.115-119.