Preserving Privacy in Multi Provider Systems with Sensitive Data Hiding: A Framework

Nanna Babu Palla [1], **Dr. A. Vinaya Babu** [2]

[1] Research Scholar, Dept .of.CSE, JNTU Hyderabad

[2] Professor in CSE, JNTU Hyderabad, Hyderabad, Telangana, India

Abstract: ***Data mining and information provider systems are widely used for knowledge retrieval and information excavation applications. Data sharing between organizations endorse in business analysis, data dissemination and strategy development. The gathered micro data may contain person-specific features. If the data published without keeping privacy protection, it's an infringement of human rights under law. Earlier works covers on protecting privacy in centralized and distributed systems with solo occurrences of entries related to individuals. The proposed approach considers multiple entries of individuals at different sources and ensures protecting the pricy by enforcing Collaborative Publishing Scheme.***

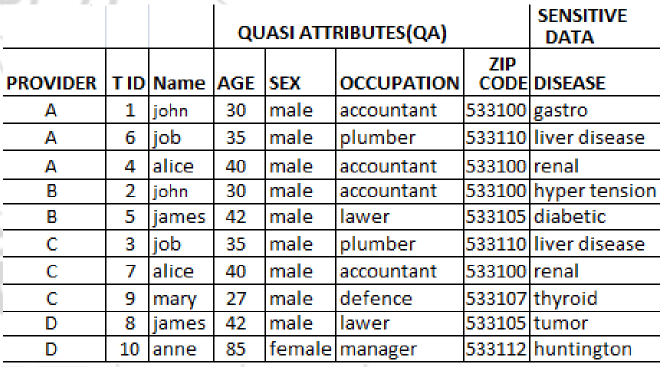
Keywords: Data mining, collaborative publishing scheme, multi provider systems, privacy, quasi-attribute

1. Introduction

Data collection and analysis is a vital activity practiced by government, business and other organizations. It's collected for specific and legitimate purpose. Data collection and processing of information related to persons is helpful for business, insurance, government and tax authorities. The analytical and knowledge extraction operations are performed using data mining techniques and intelligent techniques. The current data repository systems can accommodate unprecedented volume of data as storage technology expenses are plunge down. The repositories may contain information of individuals revealing ethnic, personal or social features. Privacy protection is an indispensable process in current socio-judicial systems and breaching privacy norms leads to infringement. The thrust of privacy preserving data mining is to thwart the identity leakage by fortifying confidentiality while sharing information of an individual in business and government organizations.

Data mining techniques can be performed without harming to a personal privacy data. In a demographic study, where an organization named Integrated Patient Information System(IPIS) collects patients details containing attributes like name, age ,gender, address ,disease and other miscellaneous fields in the form of database tables. The attributes which will biased for identifying an individual by associating them with viable social knowledge is known as ' Quasi Attributes(QA) ' or 'Quasi identifiers(QI)'. The Quasi Identifier set can be used for disclosing intended pattern with a heuristic approach. Earlier works focus on enforcing k-anonymity and ℓ-diversity approaches for protecting privacy in single source centralized systems. $\mathcal{K}$-*anonymity* approach protects an individual as the probability of identification is up to $1/k$ and in ℓ -diversity method, the number of sensitive attributes in a group is at least ℓ[2][3].

Table 1: Database maintained at IPIS



			QUASI ATTRIBUTES(QA)				SENSITIVE DATA
PROVIDER	T ID	Name	AGE	SEX	OCCUPATION	ZIP CODE	DISEASE
A	1	john	30	male	accountant	533100	gastro
A	6	job	35	male	plumber	533110	liver disease
A	4	alice	40	male	accountant	533100	renal
B	2	john	30	male	accountant	533100	hyper tension
B	5	james	42	male	lawer	533105	diabetic
C	3	job	35	male	plumber	533110	liver disease
C	7	alice	40	male	accountant	533100	renal
C	9	mary	27	male	defence	533107	thyroid
D	8	james	42	male	lawer	533105	tumor
D	10	anne	85	female	manager	533112	huntington

This paper enlightens on privacy preservation in Multi Provider Environment System (MPES), where data about an individual available at multiple sites if an individual avails medical services from one or more similar hospital by implementing data perturbation technique. It consider a case such as that mentioned earlier, Integrated Patient Information System IPIS) is one who collects, maintains and publishes data by collecting patient information from multiple hospital and healthcare agencies. If any adversary or intended user tries to publish data without enforcing anonymity, then it leads to identity leakage. . This paper focus on implementing hiding sensitive data using Collaborative Publishing Scheme (CPS) for multi provider environment system which safeguards against Background Knowledge (BK) attack and insider attacks so that the information provider systems can publish micro data without breaching privacy constrictions.

2. Problem Background

Table 1 illustrates a scenario where patient details are arranged in the chronological order of transaction, TID where patient John visited hospitals A and B, Job visited A, C and Alice availed health services from A and C with different health disorders . It shows that same individual availed well-ness services from two or more hospitals for multiple health diseases. If any provider namely A,B or C in this instance, not adhere to privacy protection, then the adversary can infer and misuse the data as the details about same individual are stored at multiple sites. This paper organized into 3 sections discussing Problem background in section 2, key literals and description in 4 and results in section 6.

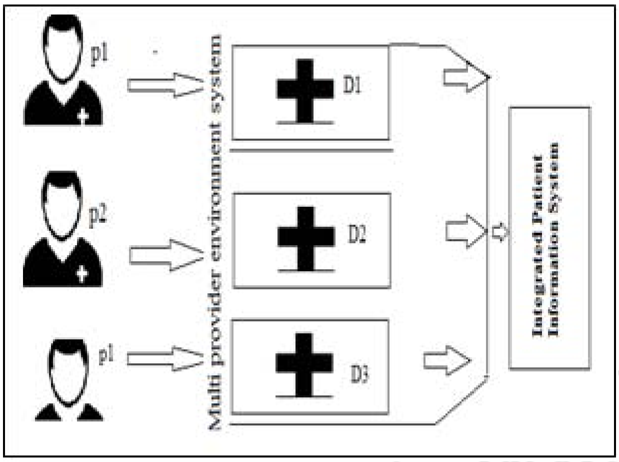

p1
p2
p1
Multi provider environment system
D1
D2
D3
Integrated Patient Information System

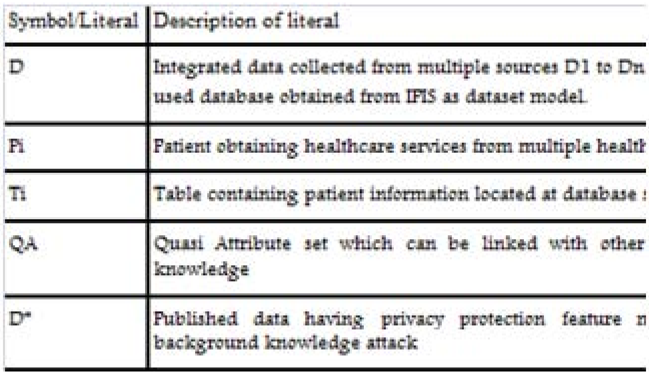
Figure 1: Diagram showing the scenario of availing multiple healthcare services by an individual

Fig.1 describes a scenario such that in Integrated Patient Information System (IPIS), the Quasi Attributes are (age, sex, occupation and zip code) or (age, sex, zipcode). They are denoted as QA=∑ (age, sex, occupation, zipcode). Quasi Attributes (QA) offers an inference based knowledge extraction by mapping them with other sources of public knowledge. Attribute 'disease' is a sensitive data pertaining to an individual. Direct identifiers like social security number, permanent account number and unique identification numbers reveals individual data without any alternative knowledge. In this paper, an Integrated Patient Information System(IPIS) which is obtained from multiple providers about an individual with their health status. Patient P1, P2, P3 is availing services from various health care centres.

Patient P1 is availing services from provider 1 and 3 respectively. If data providers are denoted by D, then D1, D2 and Dn are 'n' different provider systems. In this instance, patient p1 details are available at D1 and D3. If any provider, either D1 or D3 discloses and publishes data to an adversary, then the privacy of an individual is under menace. Integrated data collected from multiple provider systems are stored at single source site called as "Integrated Patient Information System. Database size depends on scalability and size metrics for each provider system. The tables maintained at various sites like D1 to Dn are represented as T1 to Tn.

3. Key Literals and Description

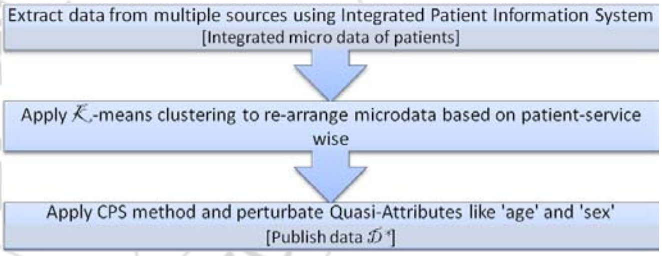
The following section provides notations for terminology used in the paper. Every hospital owner maintains patient health record $\mathcal{R}$ containing 'pid', 'age' , 'sex ', 'occupation', 'address ', 'disease' and other valid attributes. The database D assumes that every tuple and row contains one sensitive attribute 'disease'. The features which identify an individual without linkage to external knowledge are named as 'Unique Identifiers (UI)'. The used databases are cut with removing 'unique identifiers'. Data integration from multiple sources is performed, which is named as D.



Symbol/Literal	Description of literal
D	Integrated data collected from multiple sources D1 to Dn used database obtained from IFIS as dataset model.
Pi	Patient obtaining healthcare services from multiple healt
Ti	Table containing patient information located at database :
QA	Quasi Attribute set which can be linked with other knowledge
D*	Published data having privacy protection feature n background knowledge attack

4. Framework & Approach

We consider integrated data obtained from Patient Information System (IPIS) using our developed framework for importing data from multiple sources. Each dataset contains 20000 rows which are extracted from 40 Multiple Provider Systems (MPS) at network hospitals circa 80000 tuple. If patient $\mathcal{P}_i$ is having services from $\mathcal{D}_i$ to $\mathcal{D}_n$ .K – means clustering algorithm applied on above dataset for re-arranging the database D by patient –record wise as shown in the Table 3. Figure 2 shows process used in hiding sensitive data using Collaborative Publishing Scheme (CPS).


Extract data from multiple sources using Integrated Patient Information System [Integrated micro data of patients]

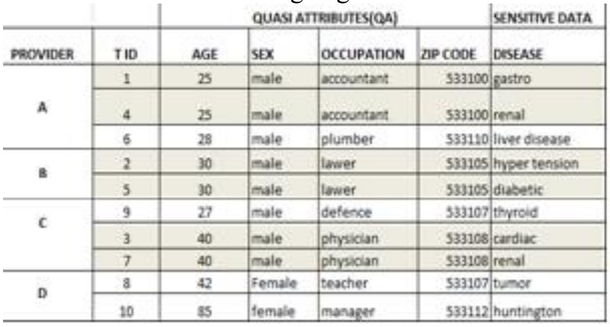
Apply $\mathcal{K}$-means clustering to re-arrange microdata based on patient-service wise

Apply CPS method and perturbate Quasi-Attributes like 'age' and 'sex' [Publish data $\mathcal{D}$']

Figure 2: Process flow used in data publishing

Quasi Attributes (QA) are perturbated with following schemata using 'Collaborative Publishing Scheme' [1]. Attribute 'age' and 'sex' is generalized using Taxonomy tree. 'zip code' is modified with aggregation. The taxonomy tree for 'age' and 'sex' are shown in fig.3 and in fig.4.

Table 3: Re-arranged Data after performing K-Means Clustering Algorithm



		QUASI ATTRIBUTES(QA)				SENSITIVE DATA
PROVIDER	T ID	AGE	SEX	OCCUPATION	ZIP CODE	DISEASE
A	1	25	male	accountant	533100	gastro
	4	25	male	accountant	533100	renal
	6	28	male	plumber	533110	liver disease
B	2	30	male	lawer	533105	hyper tension
	5	30	male	lawer	533105	diabetic
C	9	27	male	defence	533107	thyroid
	3	40	male	physician	533108	cardiac
	7	40	male	physician	533108	renal
D	8	42	Female	teacher	533107	tumor
	10	85	female	manager	533112	huntington

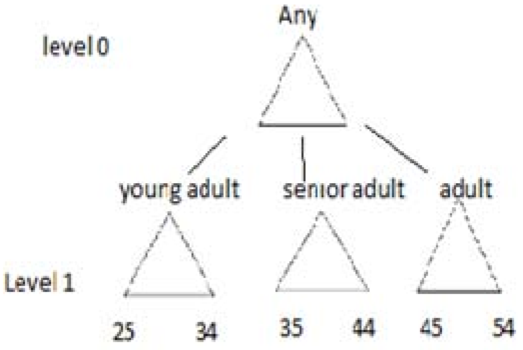


Figure 3: Taxonomy tree for attribute 'age' with level 2

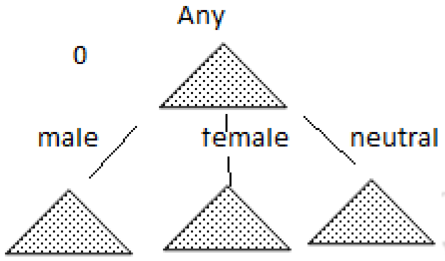


Figure 4: Taxonomy tree for 'sex' attribute

5. Results and Discussions

The extracted micro data is tested on Intel Pentium IV processor with 2.4GHz of clock speed. The micro data is selected from IPIS interface with 80000 records from 40 Provider systems. Number of perturbation operations performed on the above dataset while reading and writing operations on to the database are evaluated using Collaborated Publishing Scheme. The sanitized data can be published with less distortion to original data as shown in Fig. 6. The results shows comparative performance improvement in terms of access time and perturbation time for selected attributed. Fig. 5 shows linear time complexity considered on providers and access time.

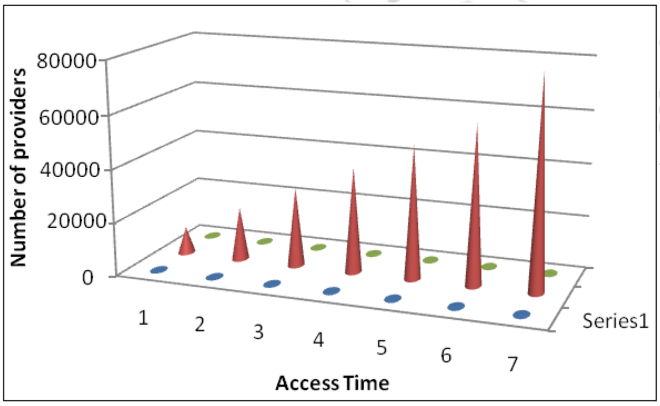
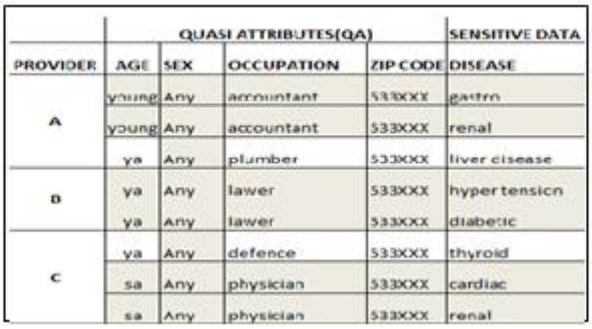


Figure 5: Graph depicting time variance between accee time vs number of providers.



	QUASI ATTRIBUTES(QA)				SENSITIVE DATA
PROVIDER	AGE	SEX	OCCUPATION	ZIP CODE	DISEASE
A	young	Any	accountant	533XXX	gastro
	young	Any	accountant	533XXX	renal
	ya	Any	plumber	533XXX	liver cisease
B	ya	Any	lawer	533XXX	hyper tensicn
	ya	Any	lawer	533XXX	diabetic
C	ya	Any	defence	533XXX	thyroid
	sa	Any	physician	533XXX	cardiac
	sa	Any	physician	533XXX	renal

Figure 6: Data with sanitization which can be published for miners and analysts

6. Conclusion

Multi provider systems offer data repository services having multiple instances of individuals. The proposed framework named 'Collaborative Publishing Scheme' for protecting privacy in data publishing arena is considered to be effective and efficient. This approach used for multi provider data systems where the source of information about an individual is available from different sites. We compared performance issues with respective to access time and data perturbation complexities. The study reveals that the performance can be further improved when homogeneous databases are employed.

7. Acknowledgement

Authors wish to thank Integrated Patient Information System(IPIS) data provider for granting the permission to use individuals data pertaining to healthcare in Rajahmundry division.

References

[1] P.Nanna babu, Dr A. Vinaya babu " Preserving privacy in set valued data using Collaborative Publishing scheme " in Proc. Of ICACC-2016, Page no.126-129.

[2] Manchanavajjhala " l-diversity : privacy beyond k-anonymity " IEEE Data engineering,2006.

[3] N. R. Adam and J. C. Wortmann." Security-control methods for statistical databases: A comparative study. " ACM Comput. Surv., 21(4):515-556, 1989.

[4] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k-anonymity: Algorithms and hardness. Technical report, Stanford University, 2004.

[5] Health Insurance portability and accountability act "HIPPA".

[6] 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.

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

[8] 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 Database systems, Santabarbara, California, USA, May2001.

[9] 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.

Author Profile

Nanna Babu Palla is a Research Scholar at JNTU Hyderabad, India. He did his graduation and post graduation from same university with a specialization in Computer Science & engineering. His research interests include privacy preserving data mining, distributed databases, design patterns and cloud computing. Currently, he is working as Associate Professor in department of computer science at Aditya Engineering College, Andhra Pradesh.

Dr A Vinaya Babu is a Professor in Computer Science and engineering with rich experience in teaching and administrative areas. He is an expert in various areas in advanced computing systems, parallel processors, algorithm design and analysis, compiler design, data mining and cloud computing.