International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Privacy Preserving for Mobile Health Data

K. Komala Devi[1], S. Ram Prasad[2]

[1]PG student, Department of CSE, Vignan's Institute of Engineering for Women, Visakhapatnam, A.P
[2]Professor & HOD, Department of CSE, Vignan's Institute of Engineering for Women, Visakhapatnam, A.P

Abstract - *Confidential information of individuals is being collected by most of the enterprises which can be used by researchers for various purposes to perform analysis on the data. Individuals would not like to release their private information directly through any means. Even though the identifying attributes are suppressed, they can be still identified by linking it to the openly available data sources. The K-Anonymity is a protection model that forms the basis for many real-world privacy protection systems. A release provides K-Anonymity protection if the information for each individual contained in the release cannot be distinguished from at least K-1 individuals. This project is an attempt to develop a Structure for privacy protection using K-Anonymity under different scenarios by considering real world medical electronic record sets and to ensure that the principles provide maximum privacy and utility of the data. The project also re-identifies attacks that can be realized on releases that hold to K-Anonymity unless accompanying policies are respected.*

***Key Words*: Confidential, K Anonymity, Privacy, Patient data, Medical electronic records.**

1. INTRODUCTION

Data privacy is also called Information Privacy is the relationship between collection and dissimilation of data. It deals with the ability an organization or individual has to determine what data in a computer system can be shared with third parties [1] . Health privacy is the practice of keeping information about confidential. Sharing and disseminating electronic medical records while maintaining a commitment to patient confidentiality is one of the biggest challenges facing medical informatics and society at large. Society is experiencing exponential growth in the number and variety of data collections containing person-specific information as computer technology, network connectivity and disk storage space become increasingly affordable [2] . Data holders, operating autonomously and with limited knowledge, are left with the difficulty of releasing information that does not compromise privacy, confidentiality or national interests [3] . In many cases the survival of the database itself depends on the data holder's ability to produce anonymous data because not releasing such information at all may diminish the need for the data, while on the other hand, failing to provide proper protection within a release may create circumstances that harm the public or others.

The anonymized data is useful for research purposes. Large amount of person-specific data has been collected in recent years, both by governments and by private entities. Data and knowledge extracted by data mining techniques represent a key asset to the society analyzing trends and patterns. Formulating public policies laws and regulations require that some collected data must be made public for example, Census data. Public data conundrum as Health-care datasets are used in clinical studies, hospital discharge databases [3,4] . Genetic datasets are used for $1000 genome, Hap Map, decode. Demographic datasets are used as U.S. Census Bureau, sociology studies. Search logs, recommender systems, social networks, blogs are used for AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon etc. Privacy-preserving for both data mining and data publishing has become increasingly popular because it allows sharing of sensitive data for analysis purposes. K-Anonymity [1, 2, 3, 5, 6, and 7] model is an approach to preserve privacy. The term k in k-Anonymity implies the frequency of the similar tuples in the dataset.

The example above provides a demonstration of re-identification by directly linking (or "matching") on shared attributes. The work presented in this paper shows that altering the released information to map many possible people, thereby making the linking ambiguous, can thwart this kind of attack. The greater the number of candidates provided, the more ambiguous the linking, and therefore, the more anonymous the data.

Table -1.1: Patient Data

ID	Name	DOB	GENDER	Zip Code	Disease
202	Ramsha	29	M	53701	Heart Disease
347	Yadu	24	F	53712	Hypertension
101	Salima	28	M	53713	Bronchitis
901	Sunny	27	M	53714	Cancer

Table -1.2: Voter Registration Data

Name	DOB	Gender	Zip Code
Ramsha	29	Male	53701
Beth	40	Female	55410
Carol	25	Female	70210

Dan	31	Male	21174
Ellen	35	Female	12237

Ramsha has heart disease

Even though the identifying attributes are suppressed, the identity of the individual can be still obtained by linking it to the openly available data sources like Newspaper, Voter registration data etc. This results in a privacy threat to the individual. So privacy needs to be provided [4].

The paper organized as follows: Section II discusses existing system and their drawbacks. Section III describes proposed work. The output results are presented in section IV. Finally, Section V concludes the paper.

2. EXISTING SYSTEM

There is increasing pressure to share health information with researchers and even make it publicly available in order to carry out the analysis. However, such disclosures of personal health information raise serious privacy concerns. To alleviate such concerns, it is possible to anonymize the data before disclosure. One popular anonymization approach is k-Anonymity [6]. There have been no evaluations of the actual re-identification probability of k-anonymized data sets. The information for each person contained in the released table cannot be distinguished from at least (k-1) individuals whose information also appears in the release [7].

Example: You try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender. Any quasi-identifier present in the released table must appear in at least k records. Personal information of individuals is collected by micro organizations that can be used for Business and Research purposes. Data sharing cannot be done directly as it includes person-specific information [8]. So organizations release micro data that is by removing explicit identifiers of an individual like name, address or phone number. But it contains data like DOB, Zip code; gender etc which when combined with other publicly released data like voter registration data can identify an individual. This joining attack can also obtain the sensitive information about an individual [9]. Thus, keeps the privacy of an individual in danger.

In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected patient specific data with nearly one hundred attributes per encounter along the lines of those shown in the leftmost circle of Figure 2.1 for approximately 135,000 state employees and their families. Because the data were believed to be anonymous, GIC gave a copy of the data to researchers and sold a copy to industry.

For twenty dollars Latanya Sweeney purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes. The rightmost circle in Figure 2.1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals.

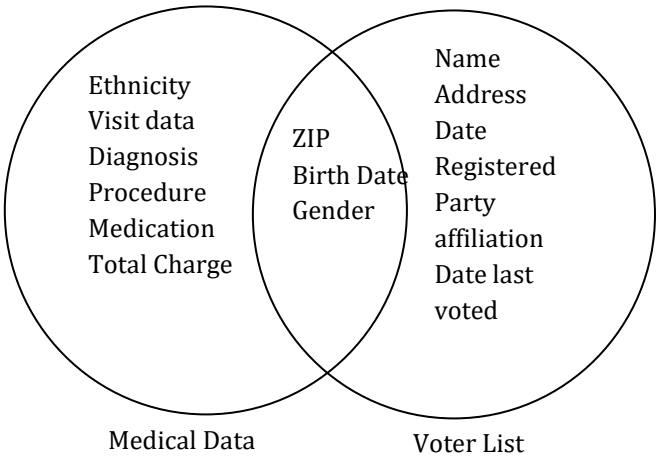

Fig -2.1: Linking to Re-identifying Data

For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

3. PROPOSED WORK

In our proposed work we are providing privacy for mobile health data. In this system we are using K – anonymity algorithm to provide security for the electronic health records. Electronics Health Records (EHR) is the assortment of patient and population electronically stored health information in digital format [10]. These include a range of data like demographics, medical history, medication & allergies, vaccination status, laboratory test results & personal statistics like age, weight & billing information etc. These information can be shared across diverse health care settings through network connected, enterprise wide information systems and through other information networks & exchanges. It eliminates the need to track down patients preceding paper medical records & assists in ensuring data is precise and readable. EHR can diminish risk of data imitation as there is only one

modifiable file, which means that the file is more likely up to date, and decreases risk of lost paper work. Owing to the digital information being searchable and in a single file, EHR's are more efficient when extracting medical data for the inspection of possible trends & long term changes in patient. Hence, we will apply K – anonymity for this mobile health data.

The K-Anonymity is a defense model that forms the basis for many real-world seclusion protection systems. A discharge provides K-Anonymity protection if the information for each individual contained in the release cannot be distinguished from at least K-1 individuals. This paper is an endeavor to develop a Structure for privacy protection using K-Anonymity under diverse scenarios by considering mobile health data sets and to ensure that the principles give maximum privacy and utility of the data. This paper proposes and evaluates an optimization algorithm for the powerful de-identification procedure known as anonymization. An anonymized dataset has the property that each record is indistinguishable from at least others. Even simple restrictions of optimized -anonymity are NP-hard, leading to significant computational challenges. We present a new approach to exploring the space of possible anonymizations that tames the combinatory of the problem, and develop data-management strategies to reduce reliance on expensive operations such as sorting. Through experiments on real census data, we show the resulting algorithm can find optimal anonymizations under two representative cost measures and a wide range of. We also show that the algorithm can produce good anonymizations in circumstances where the input data or input parameters preclude finding an optimal solution in reasonable time. Finally, we use the algorithm to explore the effects of different coding approaches and problem variations on anonymization quality and performance. To our knowledge, this is the first result demonstrating optimal anonymization of a non-trivial dataset under a general model of the problem.

4. Anonymization algorithm

Input: Patient Table **PT;** quasi-identifier **QI**= (A1...., an),
Output: Anonymized table from **PT [QI]** with respect to k
Assumes: |PT|>=K

Methods:
1. Freq← a frequency list contains distinct sequences of values of **PT [QI].**
Along with the number of occurrences of each sequence.
2. While there exists sequences in **freq** occurring less than k times that account for more than k tuples **do**
2.1 let Aj be attribute in **freq** having the most number of distinct values
2.2freq<- generalize the values of Aj in **freq**
3. Freq<-suppress sequences in **freq** occurring less than k times.
4. Freq<-enforce k requirement on suppressed tuples in **freq.**
5. RETURN Anonymized table <- construct table from **freq**

5. OUTPUT SCREENS

Fig -5.1: Introduction to privacy

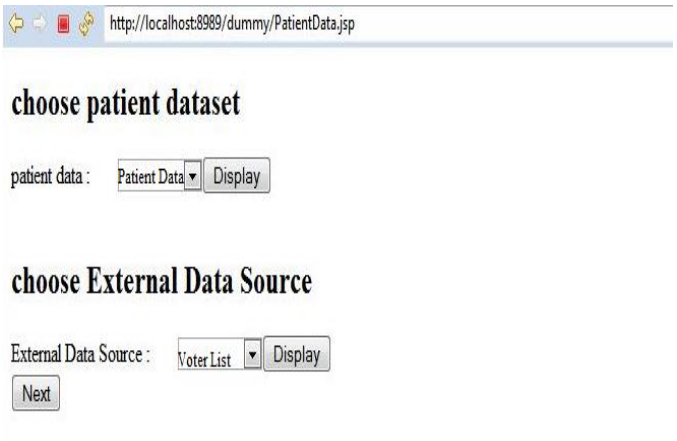

Fig -5.2: Choosing Patient &Voter list dataset

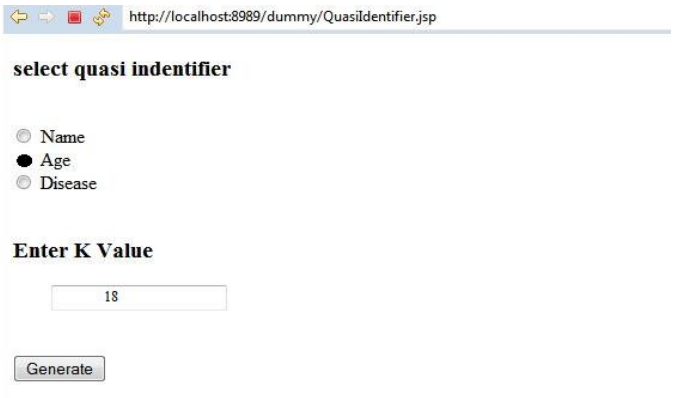

Fig -3: Choosing Quasi-identifiers & applying anonymization algorithm

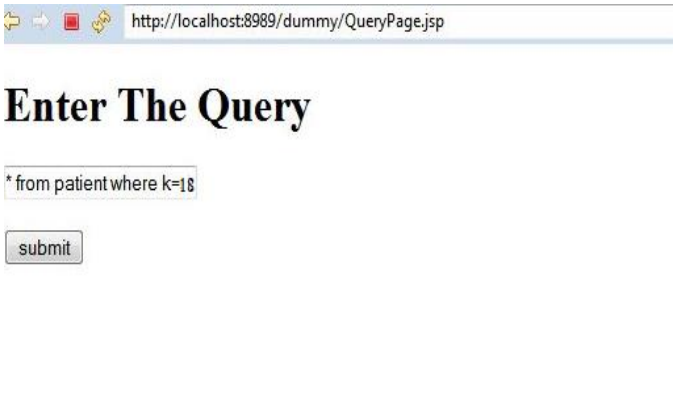

Fig -5.4: To share required data in anonymized format

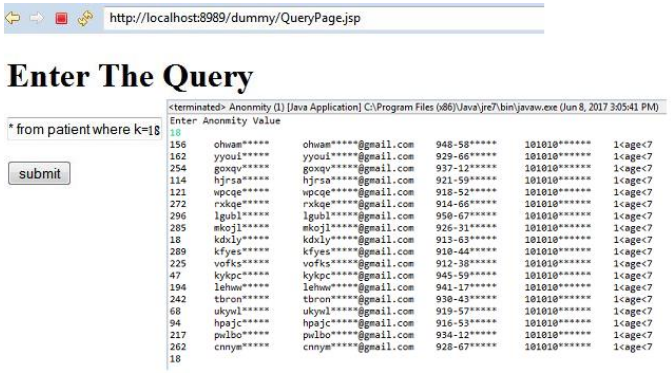

Fig -5.5: Anonymized Output

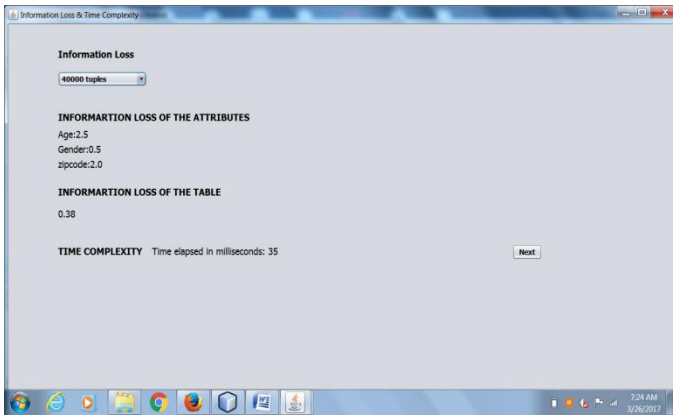

Fig -5.6: Information loss & Time complexity

6. RESULTS & ANALYSIS

The final result is the anonymized format of data where the k-Anonymization algorithm is applied on the original dataset. The anonymized format of data is generated to share with the third party. Finally the Information Loss and Time Complexity is calculated and displayed.

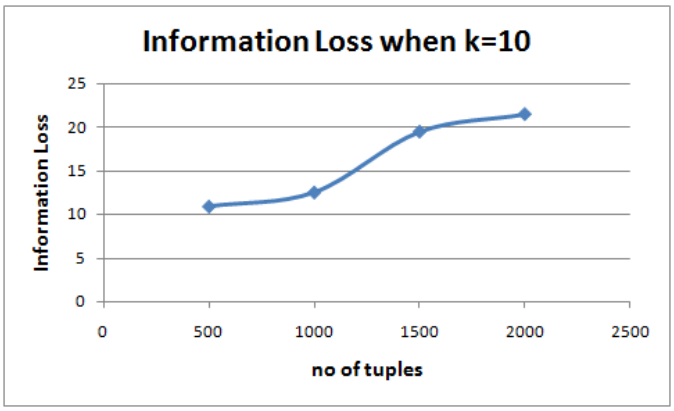

Chart -6.1: Information Loss when K = 10

When k value is constant (here k=10), the information loss is increases with the increase in number of tuples.

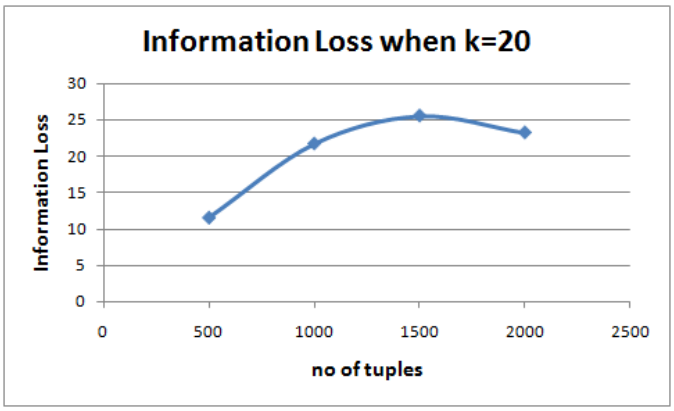

Chart -6.2: Information Loss when K = 20

When k value is constant (here k=20), the information loss is increases with the increase in number of tuples.

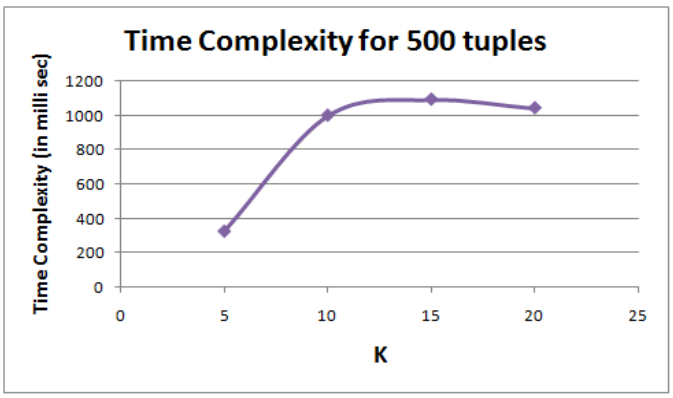

Chart -6.3: Time complexity for 500 tuples

The time complexity increases with increase in k-values keeping the number of tuples constant (Here number of tuples=500).

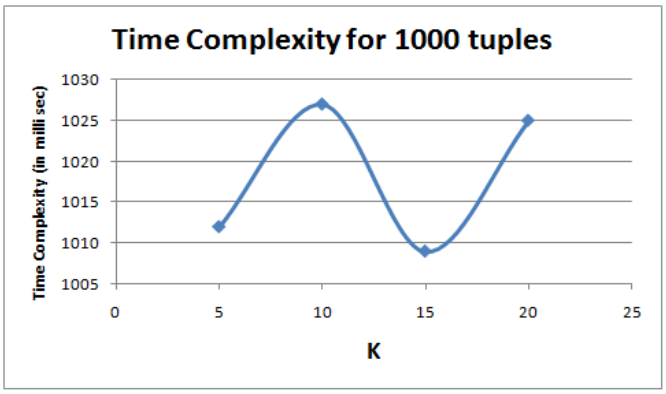

Chart -6.4: Time complexity for 1000 tuples

The time complexity increases with increase in k-values keeping the number of tuples constant (Here number of tuples=1000).

The Implementation and results is the phase where the patient and newspaper datasets should be selected and mapped to identify the records. The k value is set so that the datasets are generalized and re-identification is done on the data. The query entered should be valid and anonymization algorithm is applied on the original data. The anonymized format of data is obtained. Information Loss and Time Complexity is calculated and displayed.

7. CONCLUSIONS

Researchers request person specific data in order to carry out the analysis, the data sources should upload the data. Here, privacy policy is the issue to be considered mostly in these days. So many techniques have been proposed to provide privacy to person specific data, where as this project discusses another privacy protection model, the k-Anonymity, to provide better privacy protection to the data compared to previous models. The important key feature of k-Anonymity algorithm is that, the frequency of the data is calculated and generates the anonymized format of data. It increases the privacy protection. The results proven that the k-Anonymity gives a better performance compared to the any other algorithms in terms of Information Loss and Time Complexity.

Further work need to be carried out for this project is to be applicable for various types of datasets.

REFERENCES

[1] Data fly: A System for Providing Anonymity in Medical Data by Latanya Sweeney 1998

[2] k-Anonymity: A MODEL FOR PROTECTING PRIVACY by Latanya Sweeney.

[3] L. Sweeney. "K-Anonymity": A model for protecting privacy". Intl. Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5):557 (570), 2002

[4] L. Sweeney. "Achieving k-Anonymity privacy protection using generalization and Suppression". Intl. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

[5] Roberto J. Bayard, RakeshAgrawal, "Data Privacy Through optimal k-Anonymization"2005.

[6] M.ErcanNergiz, Chris Clifton, "Thoughts on k-Anonymization", Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW06).

[7] V Ciran, S De Captianidi Vimercati, S Foresti, and P, Samarthi k anonymity by, 28 sep2012.

[8] Khaled EI Emam and Fida Kamal Dankar Protecting Privacy Using k-anonymity 2014.

[9] Q. Long, "Privacy preservation based on K-anonymity," Science & Technology association forum, vol. 3, no. 5, pp.41-43, 2010.

[10] Q. Long, "A K-anonymity Study of the Student-Score Publishing", Journal of Yunan University of Nationalities(Natural Sciences Edition), vol.3, NO. 2, pp. 144-148,March, 2011.

[11] P. Lü, N. Chen, and W. Dong, "Study of Data Mining Technique in Presence of Privacy Preserving", Computer Technology and Development, 2006, 16(7).

[12] T. Cen, J. Han, J. Wang and X. Li, "Survey of Kanonymity research on privacy preservation," Computer Engineering and Applications, vol. 44, no. 4, pp. 130-134,2008.

[13] K. Yin, Z. Xiong and J. Wu. "Survey of Privacy Preserving in Personalization Service," Application Research of Computers, vol. 25, NO. 7, pp. 123-140, 2008.

[14] L. Sweeney, "Achieving K-anonymity privacy preservation using generalization and suppression, "International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 54, no. 5, pp. 571-588,2002.

[15] A. Machanavajjhala, J. Gehrke and D. Kifer, C, "l-Diversity: Privacy beyond K-Anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1.No. 1, pp: 24-35, 2007.

[16] J. Li, G. Liu , J. Xi, and Y. Lu, "An anonymity approach satisfying demand of maximum privacy disclosure rate";Journal of YanShan University, 2010, 34(3).

[17] X. Qin, A. Men and Y. Zou, "Privacy preservation based on K-anonymity algorithms," Journal of ChiFeng university , vol. 26, no. 5, pp. 14-16, 2010.

[18] H. Jin, Z. Zhang, S. Liu, S. Ju, (αi, K)-anonymity Privacy Preservation Based on Sensitivity Grading, Computer Engineering, Vol.37 No.14, pp.12-17.

[19] R. Wong, J. Li, A. Fu, et al "(α, K)-anonymity: An enhanced K-anonymity model for privacy preserving data publishing," Inter-national Conference on knowledge Discovery and Data Mining, 2006: 754-759.

[20] Y. Kan and T. Cao, "Enhanced privacy preserving K anonymity model: (α, L)-diversity K- anonymity," Computer Engineering and Applications, vol. 46, no. 21,pp. 148-151, 2010.

[21] Y. Tong, Y. Tao, S. Tang and B. Yang, "Identity-Reserved anonymity in privacy preserving data publishes," Journal of Software, 2010, 21(4):771-781.
[22] T. Ma, M. Tang, "Data Mining Based on Privacy Preserving," Computer Engineering, 2008, 34(9).
[23] S. Zhou, F. Li, Y. Tao, X. Xiao, "Privacy Preservation in Database Applications: A Survey", Chinese Journal of Computers, 2009, 32(5).