Framework to Reduce the Adversary Linkage Attacks in Data Publishing: MORAS

Nanna Babu Palla*, Dr A Vinaya Babu, K Govinda Raju****

* Department of Computer Science & Engineering, Research Scholar, JNTUHyderabad
** Department of Computer Science, Director , SCET(W)-Hyderabad, Associate Professor, Aditya Engineering College, Surampalm, India.

DOI: 10.29322/IJSRP.9.07.2019.p9122
http://dx.doi.org/10.29322/IJSRP.9.07.2019.p9122

Abstract- Collaborative data publishing from multiple data providers in recent is a common practice. Data owners publish these micro data for research and pattern analysis deploying data mining and high end intelligent retrieval systems. albeit, few attributes may necessitate sensitive data disclosure when nexus with external viable data sources resulting sensitive data leakage. MORAS algorithm , used in this work enlightens a framework to reduce the linkage attacks by adversary and deriving the population uniqueness to zero. Hence, the privacy infringements are eliminated in data publishing avenue.

Index Terms- data mining, equivalence class, privacy leakage, linkage attacks, Re-identification risk, uniqueness,

I. INTRODUCTION

Data collection and data publishing is a most preferred activity by repository systems such as health care. the published data may subsist person-centric information. few attributes published by various data holding agencies may be classified as i)sensitive ii)non-sensitive and non-identifiable iii) identifiable and iv)partially identifiable. the earlier studies rendered by researches draw out with insights adhering identity attacks and sensitive data disclosure menace. Hence, Privacy preserving data mining is evolved to protect individual's privacy while facading the users to share their data with trust and assurance. The earlier models imbibing privacy preservation in data publishing avenue are presented. *k*-Anonymity model which emphasize on constructing the clustering groups according to their similar features[1]. this suffers with homogeneity attacks , as the similar clusters were distributed in the published database, the adversary can figure out the similar classes and re-identity the records pertaining to an individual. hence, *l*-Diversity model proposed to guarantee the distribution of equivalence groups and clusters exhibiting the diversity of sensitive attributes, at each cluster level, so that the attacker is bewildered to identify the individual's record[2].

According to rules laid by HIPAA(Health Insurance Portability and Accountability Act,1996), while publishing the microdata of statistical databases and the privacy guidelines to be adapted while adhering NCBI suggestions. *t*-closeness offers reduced granularity in each cluster classes , such that they exhibit reduced distance at sensitive attributes[3]. The following scenario describes adversary attack using data linkage as presented in Fig.1.

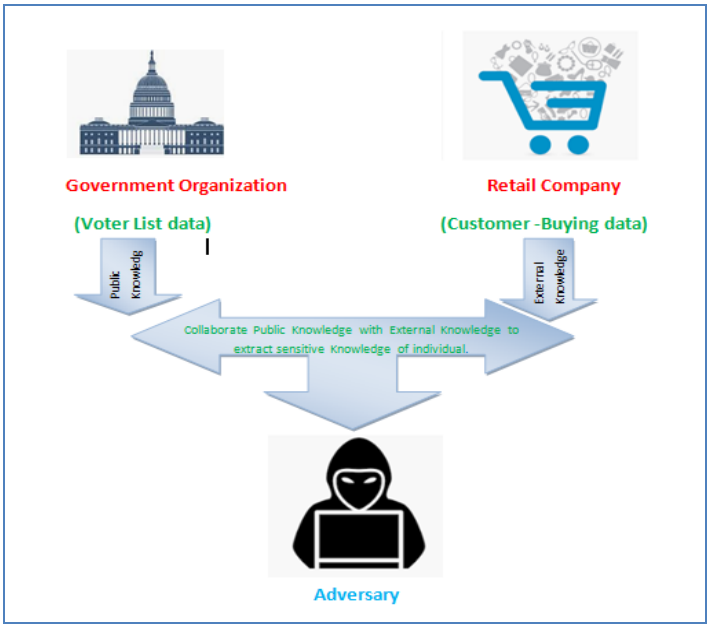

Fig.1 A Visualization of attacker model for linkage of data from public knowledge to private knowledge.

The Adversary Attack Model(AAM) described , with linking the public vailable data(Voters data) with external knowledge such as Customer buying habits on an e-commerce platform. If customer purchases, few sales items, which are sensitive , then the attacker can mine such sensitive knowledge using high performant information extraction techniques and then, it leads to Privacy infringement. The Proposed work , MORAS(MOdel for Reduced Attacker Success) is instrumental in enhancing the privacy preservation and prudent in thwarting privacy breaches.

The privacy is jeopardize when linkage of published data by the repositories when compared with other private knowledge. This work focus on evaluating the Re-Identification Risk(RIR) , Data disclosure by quasi-identifiers and finally a model for reduced adversary attack with respect to Prosecutor Risk Model(PRM).

II. RELATED WORK

A. Key definitions

Quasi-identifiers : A set of attributes , {a_0-a_n} of a database D, which help to identify an individual from a released database(D).

Sensitive data : the data, which is an attribute A∈D, which an individual want non-disclosure to public domain.
Prosecutor Risk Model(PRM): the intruder extract the sensitive knowledge from released database, D using aprior knowledge about an individual from the given population P.

Uniqueness : the equivalence classes ,Eq € each Equivalence class is clustered such that the dankar population constant is set zero.

This work carried on de-facto dataset for privacy evaluation named US Adult dataset , provided from UCI machine learning repository, which is highly cited by research community. It's best known as Census dataset, which exhibits following properties. Table 1, presents sampling size of 32561 having QI= *{age,gender,occupation}* acting as QI for given data, D and *{race}* is a sensitive attribute. Table 2, shows the population size considered for two regions, namely USA and India having sampling fraction 0.0001 and 0.00003 respectively. the distinction and separation of quasi-identifiers were presented in the Table.3

Table 1. Adult dataset properties

Attributes type	Quasi-identifiers	Domain size	Cardinality
Multivariate	{age,gender,occupation}	32561	15

B. Evaluating the Privacy Disclosure- Risk(PDR) based on quasi-identifiers.

Table 2. Dataset Model

Region	Sampling fraction	Population size	Suppression Limit
USA	0.0001	317238626	0%
India	0.00003	1210569573	0%

.Table 3. Distinction and Separation of QI

Quasi-identifier	Distinction(%)	Separation(%)
age	0.2241	97.86
gender	0.0061	44.27
occupation	0.04	90.28
{age,gender,occupation}	4.89%	99.85

among the given quasi-identifiers, attribute *age* shows maximal distinction(.2241) and *occupation* having 0.04% and when *{age,gender,occupation}* then, distinction of records is 4.89%.

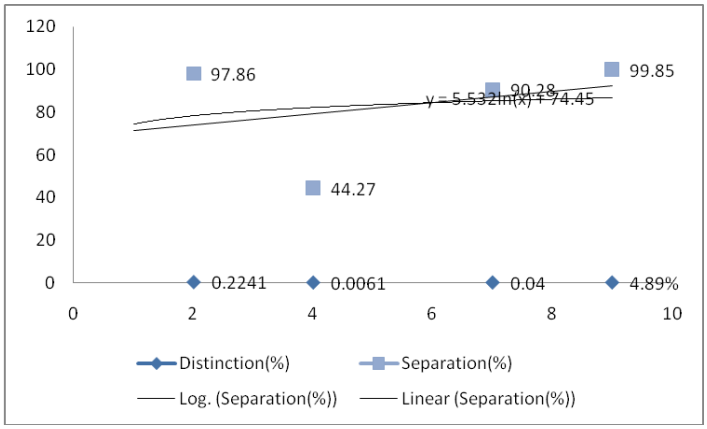

Fig.2 Graph showing the Distinction & Separation for attributes {age},{gender},{occupation} and {age,gender,occupation}.

A. Framework

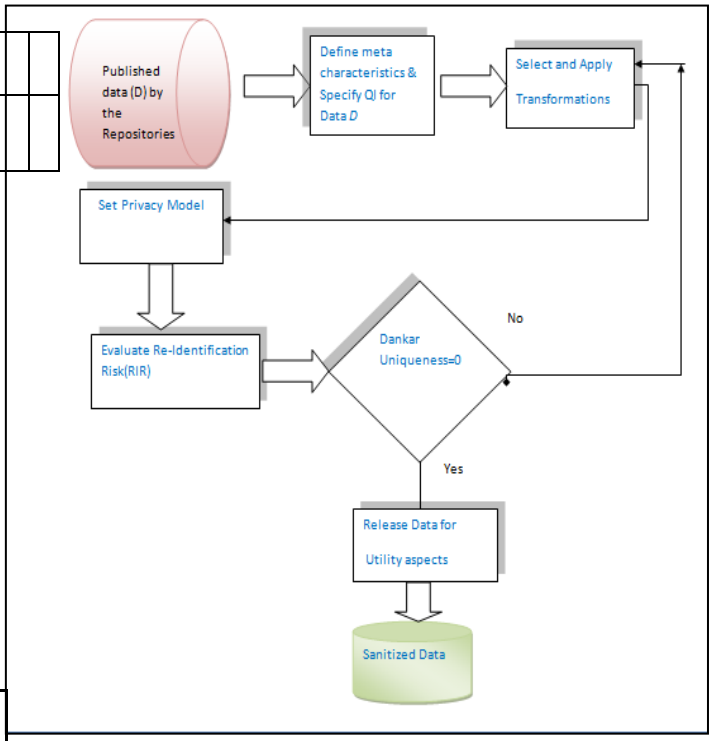

Fig.3 Framework deployed for the proposed work to be work on US-Adult data.

The transformations applied is shown with lattice representation as shown in Fig.4. The contingency between *'gender'* and *'occupation'* were analyzed for dataset showing maximal risk associated with *'managerial'* category with *'male'* gender as shown in Fig.5 demarked with an ellipse.

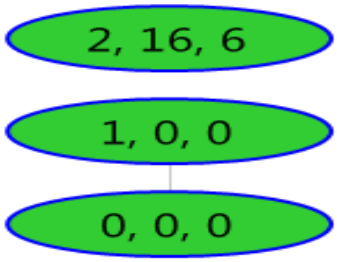

Fig.4 Lattice with transformations performed

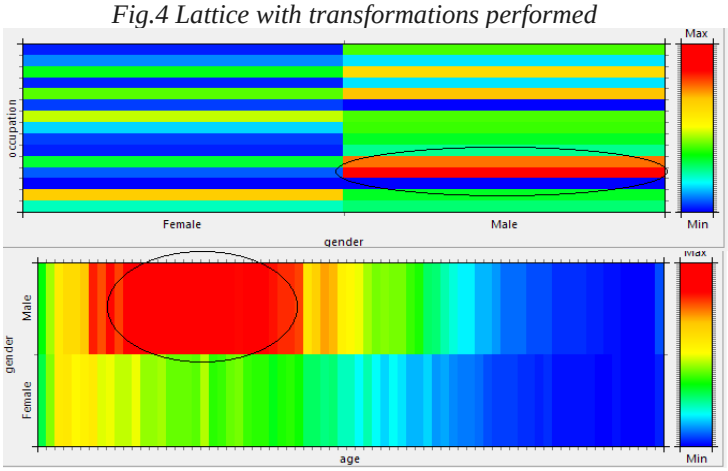

Fig. 5 Visualizing the Contingency of attributes 'occupation' and 'gender'

B.Proposed Algorithm

Algorithm MORAS(Input Original Database-D)
// accepts original database which is to be sanitized.
Begin
step 1: Define Quasi-Identifiers ∀ attributes Ai in database -D
step2 : Apply Transformations ∃ QI are genenaralized
while (i<=domain size)
begin
set k value
//the k value is set for k-anonymity
step3 : Perform construction of equivalence class
step4: evaluate re-identity risk (RIR) and Uniqueness(UN)
end
if UN<=0 then
step5: publish database D
else
repeat step 3:
End// release sanitized data-D

C. Evaluating Re-Identification Risk(RIR)

The *RIR* for the dataset for the regions mentioned USA and India were evaluated using equivalence classes constructed with mentioned transformations.

$$if\ Database\ D\ having\ Domain_{size}\ DS$$
$$then\ \exists\ QI\ transformations\ T\ applied\ \in\ Database\ D$$
$$then\ equivalence\ classes\ (EQ) constructed\ using\ clustering\ model(N) \forall QI$$

$$Attacker\ Success\ Estimate(ASE) = \frac{Eq}{Ds} \quad \text{-------}$$ eq.(1)

$$where\ Eq = No.of.Equivalence\ class$$
$$and\ Ds = No.of.Records$$

ASE(Attacker Success Estimate) for USA population is evaluated using Eq.(1) and presented in Table 4 when local transformation scheme applied using 100 iterations.

Region	Parameters
USA	4.01
Maximal class size	151
Minimal Class Size	1
Equivalence Classes	1306
Total Records	32561
ASE(%)	4.01

Table 4 showing various parameters after transformations

III.RESULTS DISCUSSION

The experiment carried on US-Adult data ,which comprises 32561 rows with 15 attributes exhibiting multivariate features on 64-bit Intel i3-3220 ,3.30GHz Processor. The RIR for given dataset using MORAS algorithm analyzed, which results in minimal data disclosure to highest prosecutor risk=30%. the intruder can extract sensitive data of individuals from published database maximal up to 3.39%.

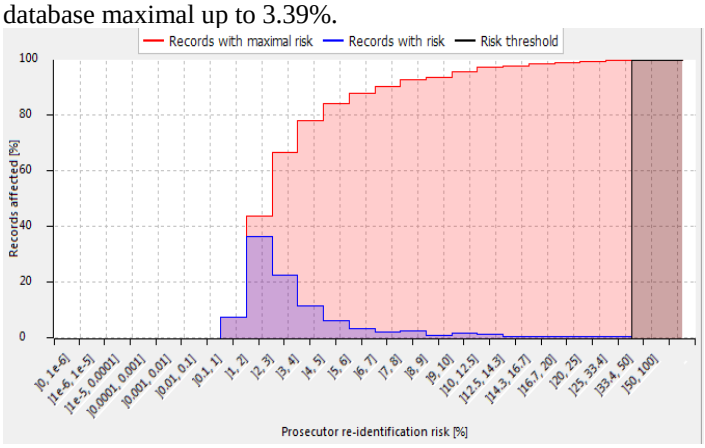

Fig.6 Visualization of Re-Identification Risks(RIR).

the population Uniqueness was evaluated using Dankar approach. this shows , the Uniqueness is derived to zero as represented in Fig.7.

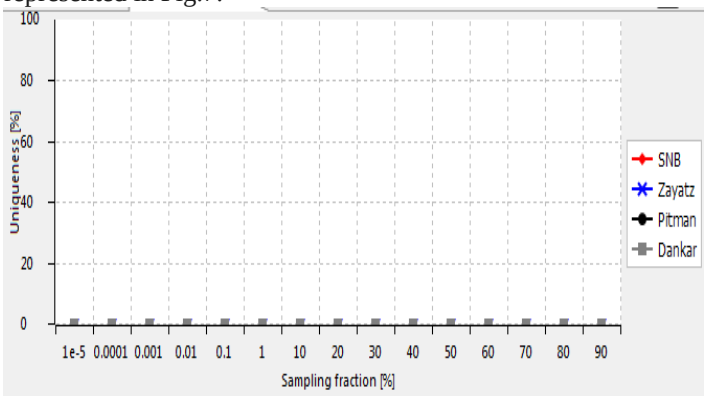

Fig.7 Population Uniqueness=0 using MORAS algorithm

IV. CONCLUSION

Data repositories publish collected data of individuals for diverse utilities. the public knowledge released to outer domain is innocuous, but due to external linkage from viable sources leads to sensitive data breach. To, thwart such linkage attacks by adversary, MORAS(Model for Reduced Attacker Success) is instrumental is reduction of Re-Identification Risks(RIR) by constructing equivalence classes with higher population spread at each bin. The Uniqueness of records is also reduced to zero, such that the attacker is minimal to data disclosure using intelligent extraction techniques. The individuals are open to share their data with trust and prudunce , so that

The proposed work enhances data utility by balancing data suppression=0 and minimal attack. Further, the work can be extended to use, MORAS for voluminous data in Statistical data Disclosure Control(SDC) applications.

ACKNOWLEDGEMENT

We wish to thank Aditya Engineering College(Surampalem) and Stanley College of Engineering & Technology(Hyderabad) for thy support in carry out this work into articulation.

REFERENCES

[1]L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; pp.557-570

[2]Ashwin M, Johannes *l*-diversity -a privacy beyond *k*-anonymity, ACM transactions on Knowledge discovery from data,Vol.1,article 1,2007,pp.1-48.

[3] AGRAWAL, R., EVFIMIEVSKI, A. V., AND SRIKANT, R. 2003. Information sharing across private databases. In Proceedings of the SIGMOD Conference. 86–97.

[4] BAYARDO, R. J. AND AGRAWAL, R. 2005. Data privacy through optimal k-anonymization. In Proceedings of the International Conference on Data Engineering (ICDE'05).

[5] CHIN, F. 1986. Security problems on inference control for sum, max, and min queries. J. ACM 33, 3, 451–464.

[6] http://counsel.cua.edu/fedlaw/hipaa.cfm

[7] https://www.ncbi.nlm.nih.gov/pubmed/citmatch

[8] T. Su and G. Ozsoyoglu. Controlling FD and MVD inference in multilevel relational database systems. IEEE Transactions on Knowledge and Data Engineering, 3:474-- 485, 1991

[9] M. Morgenstern. Security and Inference in multilevel database and knowledge based systems. Proc. of the ACM SIGMOD Conference, pages 357--373, 1987.

[10] D. Denning and T. Lunt. A multilevel relational data model. In Proc. of the IEEE Symposium on Research in Security and Privacy, pages 220-234, Oakland,

[11] D. Denning. Cryptography and Data Security. Addison-Wesley, 1982.

[12] L. Sweeney. Guaranteeing anonymity when sharing medical data, the Datafly system. Proceedings, Journal of the American Medical Informatics Association. Washington, DC: Hanley & Belfus, Inc., 1997

[13] T. Dalenius. Finding a needle in a haystack – or identifying anonymous census record. Journal of Official Statistics, 2(3):329-336, 1986.

AUTHORS

Nanna Babu Palla, working as Associate Professor at Department of Computer Science& Engineering having 16 years of academic experience at various levels. He is currently, a research scholar from JNTUHyderabad. His research interests include, data mining, cloud computing, algorithm analysis and privacy enhancement.

Dr A Vinaya babu, is former professor of CSE , JNTUHyderabad and present Director for Stanley College of Engineering-Hyderabad. He had wide research experience in computer architecture, data mining , automata design, algorithm analysis and natural language processing.

Kaladi Govinda Raju, working as Associate Professor at Department of CSE, Aditya Engineering College. He had rich academic experience in various subjects like, cryptographic techniques, net work security, data mining and data structures. His research interests include anonymity implementation , data publishing and image processing.

CORRESPONDING AUTHOR: Nanna Babu Palla, emial-id: nanibabup@rediffmail.com.