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.

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.

The Adversary Attack Model(AAM) described, with linking the public available 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- an\} of a database D, which help to identify an individual from a released database(D).
Sensitive data: the data, which is an attribute \( A \in D \), which an individual want non-disclosure to public domain.

Prosecutor Risk Model (PRM): the intruder extract the sensitive knowledge from released database, \( D \) using aprprior knowledge about an individual from the given population \( P \).

Uniqueness: the equivalence classes \( E_q \) 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

<table>
<thead>
<tr>
<th>Attributes type</th>
<th>Quasi-identifiers</th>
<th>Domain size</th>
<th>Cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate</td>
<td>{ age, gender, occupation }</td>
<td>32561</td>
<td>15</td>
</tr>
</tbody>
</table>

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

Table 2. Dataset Model

<table>
<thead>
<tr>
<th>Region</th>
<th>Sampling fraction</th>
<th>Population size</th>
<th>Suppression Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>0.0001</td>
<td>317238626</td>
<td>0%</td>
</tr>
<tr>
<td>India</td>
<td>0.00003</td>
<td>1210569573</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3. Distinction and Separation of QI

<table>
<thead>
<tr>
<th>Quasi-identifier</th>
<th>Distinction(%)</th>
<th>Separation(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.2241</td>
<td>97.86</td>
</tr>
<tr>
<td>gender</td>
<td>0.0061</td>
<td>44.27</td>
</tr>
<tr>
<td>occupation</td>
<td>0.04</td>
<td>90.28</td>
</tr>
<tr>
<td>{age,gender,occupation}</td>
<td>4.89%</td>
<td>99.85</td>
</tr>
</tbody>
</table>

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.
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 ∃ QI transformations $T$ applied ∈ Database D
then equivalence classes (EQ) constructed using clustering model(N)$\forall QI$

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

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

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.

<table>
<thead>
<tr>
<th>Region</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>4.01</td>
</tr>
<tr>
<td>Maximal class size</td>
<td>151</td>
</tr>
<tr>
<td>Minimal Class Size</td>
<td>1</td>
</tr>
<tr>
<td>Equivalence Classes</td>
<td>1306</td>
</tr>
<tr>
<td>Total Records</td>
<td>32561</td>
</tr>
<tr>
<td>ASE(%)</td>
<td>4.01</td>
</tr>
</tbody>
</table>

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

![Image](image1.png)

**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 in 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 prudence, 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


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, email-id: nanibabup@rediffmail.com.