IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

# A Privacy Preserving Anonymization Approach using Scalable Top Down Specialization and Randomization for Big Data Security

**Athiramol S**

Department of Computer Science, CMS College Kottayam

*Corresponding Author: athiramol005@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i4.152156 | Available online at: www.ijcseonline.org

Accepted: 16/Apr/2019, Published: 30/Apr/2019

***Abstract*—** There are different analysis centers all around the world. These centers makes inferences that are beneficial to the society. From where did they get the data for doing such kind of analysis?. Is the data that is published for such kind of analysis is secure from privacy breach?. In America, different kind of records especially the medical records which contains disease information as a sensitive attribute are publishing publicly. This has to be taken as a serious issue. Taking into account these facts, different anonymization methods are developed. In this paper an approach that is different from all other approaches is proposed. The main concern is to manage background knowledge attack that most of the algorithms failed to heed. Although no algorithm is able to achieve 100 percent security without sacrificing information loss, a balance can be maintained between these two. This paper introduces an efficient method for handling the outlier tuples using a randomization algorithm.

***Keywords***— Background knowledge attack, IGPL metric , Quasi- identifier, K- anonymity, Randomization, Taxonomy tree, Top down specialization

## I. INTRODUCTION

Big data security [1] is a big matter of concern due to the introduction of various technologies like cloud. Various organizations like business organizations, medical centers, educational institutes etc. are publishing data publicly especially in America. Why they are publishing data? The purpose of such data publishing [2] is to get an inference from the data that can be used for their and society benefits. Is the data that we publish for analysis is free from illegitimate uses?. No, it can be misused. So we need to do some kind of anonymization on that record before publishing it publicly such that no privacy breach is possible. All the anonymization algorithms are having drawbacks especially the information loss, background knowledge attack, Homogeneity attack, etc.

The dataset used is an Adult dataset with 32,000 records and 10 attributes which is most commonly used for research works dealing with anonymization. By performing anonymization on the Quasi- identifiers (QIDs), it is possible to avoid linkage attacks. QIDs are the set of attributes when they are used in combination can reveal an individual record. In this proposed method, instead of just anonymizing only the QIDs, Sensitive attribute (Here Occupation) is also anonymized using TPTDS approach. The system is built on top of MapReduce framework such that the execution time gets, reduced. The Two phase top down specialization (TPTDS) [3] approach is one of the best methods in terms of information loss. It uses a taxonomy tree approach for doing the specialization.

The method of specialization starts from the root node advances towards the leaf node, replacing specific value to most general value. Specialization is done with a belief that general value is more secure than specific values. The selection of specialization value depends on the IGPL value which is an information metric used in anonymization. The components in this approach are partitioning, merging anonymization levels, specialization.

The TPTDS approach is performed only for categorical attributes like Education, Occupation, Relationship, etc. The numerical attributes like Age and Zipcode undergoes suppression which is termed as one of the powerful anonymization operators. Randomization is the process that is performed in-order to make use of the outlier tuples. So instead of just ignoring those records, do the randomization approach on that outliers.

Rest of the paper is organized as follows, Section I contains the introduction of the topic , Section II contain the related

 

works, Section III contain the methodology of anonymization with flowchart. Section IV contain the results and analysis of the method. section V is the conclusion of this paper.

## II. RELATED WORK

One commonly used technique for data security is Cryptography[4]. Studies states that cryptography cannot accomplish a balance on security and efficient data utility. So it is not widely used in anonymization process. L. Sweeney proposes K- Anonymity [5] method where the data that are intended for publishing are anonymized in such a way that there will be atleast k individuals with the same data entries especially the QIDs such that a particular individual will not get identified by a third party. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, proposes an approach named L- Diversity [6] which is a slight modification to Sweeny's K- anonymity. It ensures that the sensitive attribute takes diverse values within the anonymity groups. Another approach is the Semantic anonymization [7], which ensures the sensitive attribute values within an anonymity group is diverse semantically. For that some semantic rules are defined in-order to find a semantic relationship between two sensitive values. Another approach called Datafly algorithm [8] which is a practical application of K- Anonymity. It make use of full domain generalization. All the techniques are having their own advantages and disadvantages. The common drawbacks of many anonymization algorithms are Background knowledge attack, Information loss (as suppression is used), Privacy loss, Increased execution time. Although some of these limitations can be wiped out by these techniques, we need to find an algorithm that can optimize all of this problem by a single algorithm. The outliers should also be handled as it may hold some useful information.

## III. METHODOLOGY

This method treats numerical and categorical attributes in a different way. The taxonomy tree $T_i$ and attribute list *Alist* are provided to the system. Figure.1. depicts an architecture of the method. At first the entire record $R$ is partitioned according to random sampling method. For that, a partition parameter notated as $\alpha$ is given as input to the system.

The record is distributed to $\alpha$ partitions. Each partitions are read by the system and the MapReduce framework splits the tuples in each partitions and process accordingly. The categorical attributes in the records are generalized using TPTDS. The output from this module is taken for suppressing the numerical attribute values. The output of this module will be two text files one comprising the suppressed results and other an OutList which contains those tuples that are failed to get suppressed in-order to satisfy the anonymity value. Unlike other algorithms those neglects these outlier tuples, the new method make use of them by randomizing the values of the tuples. The randomization approach used here is reversible.

### *A. Two Phase Top Down Specialization with Randomization (TPTDSR)*

**Input:** Data record $R$, Partition parameter $\alpha$, Anonymization parameter $\beta$, Taxonomy Tree $TaT_i$, $1 \leq i \leq n$ where $n$ is the number of attributes, Attribute List *Alist*.

**Output:** Anonymized Record set

1. PARTITION($R$, $\alpha$)
2: **for** each partition $p_i$ **do**
3: Execute TPTDS ($p_i$, $\beta$, $AL^J$)
4: Find the minimum anonymization level $AL^{JJ}$ which is the overall anonymization level
5: Merge all the anonymized results for each partition into one, keeping the anonymity level
6: Execute TPTDS ($R$, $\beta$, $AL^{JJ}$) to achieve the desired anonymity
7: **end for**
8: **for** each numerical attributes **do**
9: SUPPRESS($R_{gen}$, $\beta$)
10: **end for**
11: **if** there exist Outliers **then**
12: RANDOMIZE(OutList)
13: **end if**

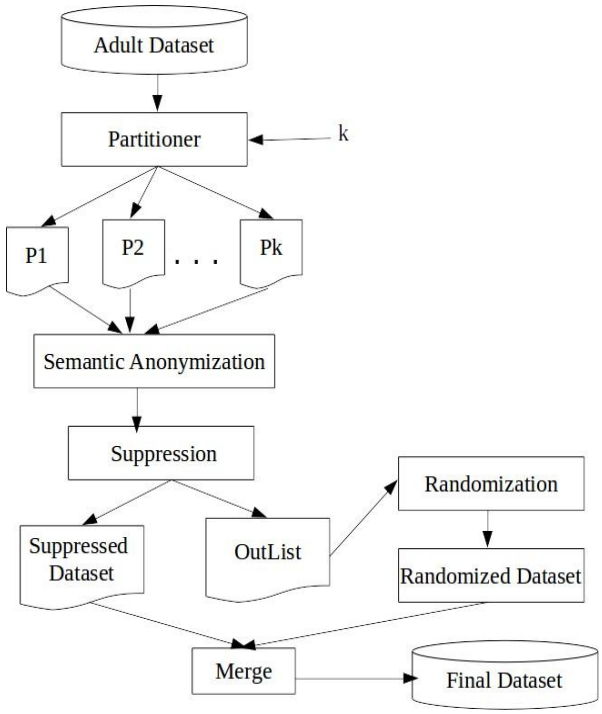


Figure. 1. TPTDSR Architecture

 

*1) Partition:* The Partition [3] is done using a random sampling approach. Figure. 2. shows an example of partitioning with $\alpha = 3$

PARTITION($R$, $\alpha$)

**Input:** Data record $R$, Partition parameter $\alpha$
**Output:** $p_i$, $1 \le i \le \alpha$

1: Generate a random number $\eta$, where $1 \le \eta \le \alpha$; $emit(\eta, r)$, where $r \in R$
2: **for** each $\eta$ **do**
3: $emit(null, list(r))$
4: **end for**

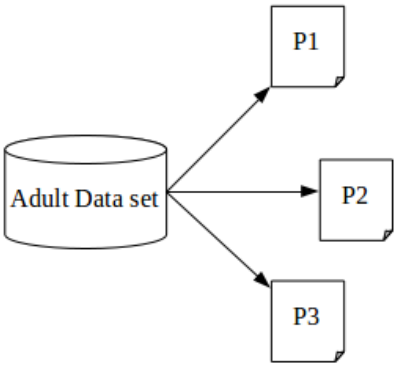


Figure 2:Data Patitioning for α=3

*2) Two Phase Top Down Specialization (TPTDS):* The TPTDS approach is performed on categorical attributes in the dataset. The values of each attributes are read out and find cut set represented as *Cuts*. There will be exactly one cut for each attributes. So the number of elements in the cutset will be equal to the number of attributes in the list. These cuts are used for the specialization operation. Whenever the number of tuples in an equivalence class is less than the prescribed anonymity level $\beta$, the specialization is performed again and it iterates until the desired level has reached. The goodness of specialization is determined by an information metric called IGPL which is a product combination of Information Gain (IG) and Privacy Loss (PL).

$$IttPL(Spec_i) = Itt(Spec_i)/(PL(Spec_i) + 1) \quad (1)$$

$$Itt(Spec_i) = |Des(Spec_{i-1})| - |Des(Spec_i)| \quad (2)$$

$$PL(Spec) = A_{pt}(Spec) - A_{cd}(Spec) \quad (3)$$

Where $Itt(Spec_i)$ is the Information Gain when doing $i^{th}$ specialisation, $Des(Spec_{i-1})$ and $Des(Spec_i)$ are the descendants of the value $Spec_{i-1}$ and $Spec_i$ in the taxonomy tree. $PL(Spec)$ is the Privacy Loss occured upon doing the Specialization $Spec$. $A_{pt}(Spec)$ is the Anonymity before performing $Spec$ and $A_{cd}(Spec)$ is the anonymity after performing $Spec$.

This specialization approach is performed using a Taxonomy tree. This algorithm is performed on each partition. For each of the attribute in the attribute list, a cut is made in such a way that, only one cut should be present for each attribute. Advancing in this way we will get a cut set with cardinality equal to the number of attributes in the list. An example is depicted in Figure.3.

If there are tuples with attribute values Engineer, Lawyer, Artist and the anonymity is not satisfied, then they are generalized to Any. This kind of cuts are formed for every attributes. The choice of good specialization depends on the Equations (1),(2), and (3). The output of this module will be the generalized record $R_{gen}$

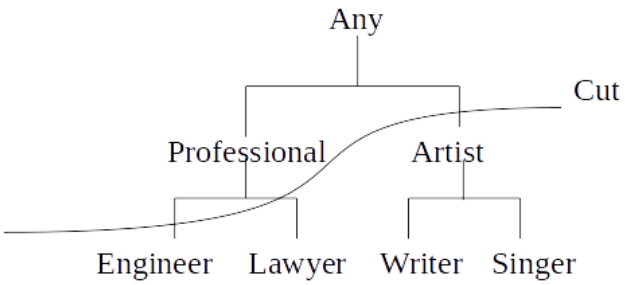


Figure. 3. An Example for Cut in Job Taxonomy Tree

Table 1: An Example For Suppressed Tuples

| Age | Gender | Zip code | Disease |
|---|---|---|---|
| 3* | * | 736** | Headache |
| 3* | * | 736** | Headache |
| 4* | F | 639** | Headache |
| 4* | F | 639** | Cough |

*3) Suppression:* Table 1 shows suppressed zipcode, age and gender values so as to satisfy anonymity requirement

SUPPRESS($R_{gen}$, $\beta$)

**Input:** Data record $R_{gen}$, Anonymity parameter $\beta$
**Output:** $R_{sup}$, OutList

1: Find $\nu$ , where $\nu$ is the frequency of each QID set.
2: **while** $\nu < \beta$ **do**
3: **for** each $QID_i$, $1 \le i \le n$ **do**
Replace LSD of the values to *
4: **end for**
5: **end while**
6: **if** there exist $\psi$, where $\psi$ is the tuple that cannot be suppressed
**then**
7: OutList $\leftarrow \psi$
8: **end if**

 

*4)Randomization:* In the proposed method, a randomization [9] approach that can be reversed is used. The values in the probability matrix ranges from 0.1 to 0.9. Figure. 4. shows an example for probability matrix.

RANDOMIZE(OutList)

**Input:** OutList
**Output:** Conversed OutList

1: **for** each QID **do**
2: Generate probability matrix $PM_i$ randomly with size $j \times j$, where $j$ is the number of tuples in OutList and $1 \le i \le n$, where $n$ is the number of QIDs
3: **end for**
4: **for** each $PM_i$ **do**
5: Find the position $p_k$ with highest value in each row in the matrix. where $1 \le k \le j$
6: **if** $P_i$ location is already taken **then**
7: go for next highest location
8: **end if**
9: **if** Two or more location with same highest value **then**
10: Choose the left hand side value
11: **end if**
12: Rearrange the values of QIDs based on $p_k$
13: **end for**

| | | | | |
|---|---|---|---|---|
| 0.8 | 0.2 | 0.3 | 0.8 | 0.7 |
| 0.7 | 0.9 | 0.4 | 0.6 | 0.5 |
| 0.6 | 0.5 | 0.9 | 0.8 | 0.2 |
| 0.8 | 0.7 | 0.6 | 0.6 | 0.5 |
| 0.3 | 0.9 | 0.4 | 0.6 | 0.5 |

Figure. 4. An Example for Probability Matrix

## IV. RESULTS AND DISCUSSION

As the Figure. 5. implies, the execution time is reduced as the number of partitions increases. There is not a steep decrease, but somehow it is possible to say that the time elapsed when using no partition is higher compared to time consumed when without using partitions.

Figure. 6. shows the relation between anonymity level and execution time. Figure. 7. implies, the TPTDS which is the generalization module takes longer time compared to the other two modules. TPTDS is a tedious process while the execution time of suppression and randomization are negligible.

One of the main concern while proposing the system was the reduction in tuple loss. Compared to all other methods existing, it is possible to reduce the TLoss to almost 0. This is depicted in Figure. 8. This approach is chosen with a fact that outliers may provide relevant information.

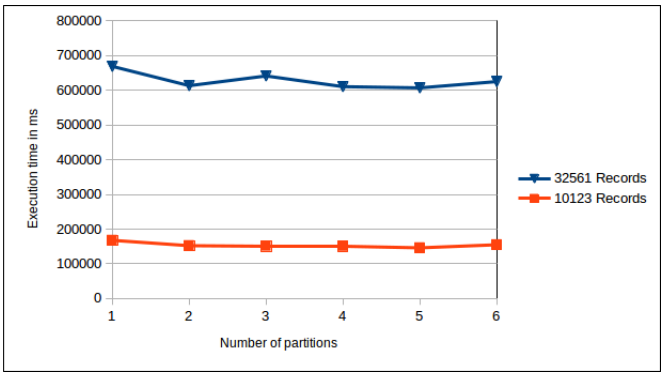


Figure. 5. Partitions VS Execution Time

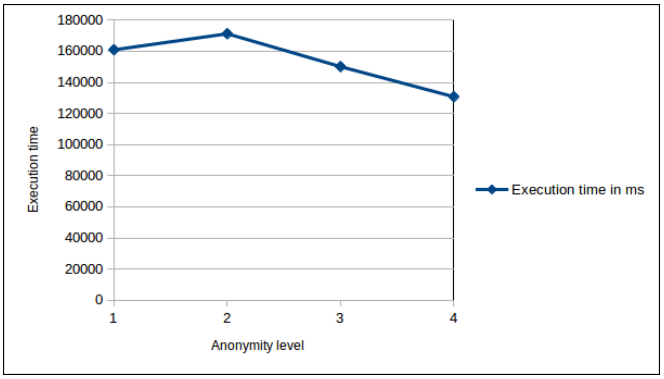


Figure. 6. Anonymity Level VS Execution Time

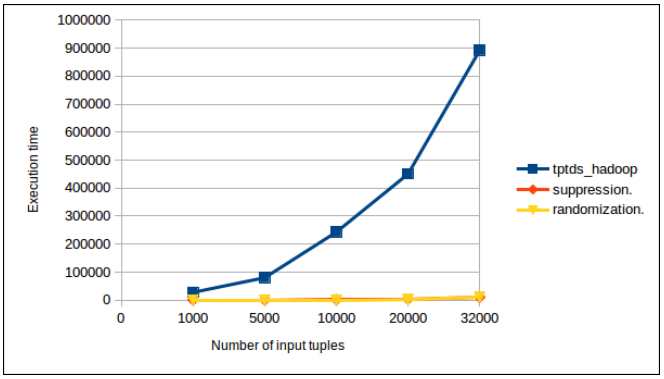


Figure. 7. Number of Input Tuples VS Execution Time for Different Modules

 

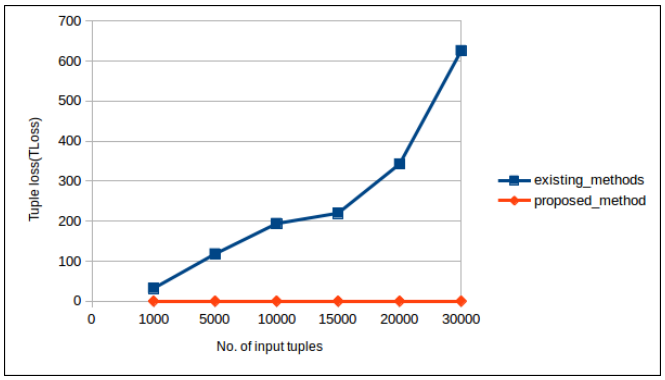


Figure. 8. Number of Input Tuples VS TLoss

## V. CONCLUSION AND FUTURE SCOPE

Anonymization is a hot research topic nowadays. There are different anonymization algorithms. They all have a common drawback which is the background knowledge attack and ignorance of outliers. As we are not able to predict the level of background knowledge an attacker is having about an individual, we need to compromise slightly with the information loss. In that way, if we could generalize the sensitive attribute also such that it can reduce the background knowledge attack.

The outliers are treated as information banks, such that they are to be preserved for publishing instead of just ignoring them. The proposed method make use of TPTDS approach along with randomization in such a way that the outliers are handled efficiently and also the work uses Hadoop MapReduce framework as an execution model thereby reducing the execution time of TPTDS which is the most tedious process among other modules.

## REFERENCES

[1] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, *"Information security in big data: Privacy and data mining*", pp. **1149-1176**, **2014**.

[2] B.C.M. Fung, K.Wang, R. Chen and P. S. Yu, "*Privacy preserving data publishing: A survey of recent developments*", ACM Computing surveys, pp. **1-53**, **2010**.

[3] Xuyun Zhang, Laurence T.Yang, Chang Liu, and Jinjun Chen, "*A scalable top down specialization approach for data anonymization using Map Reduce on cloud",* IEEE Transactions on parallel and distributed systems, pp. **363-373**, **2014**.

[4] Benny Pinkas ,"*Cryptographic techniques for privacy preserving data",* SIGKDD Explorations, pp. **12-19**.

[5] L. Sweeney,"*K-anonymity, A model for protecting privacy*",International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, pp. **557-570**, **2012**.

[6] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, *"l-Diversity: Privacy beyond k- anonymity"*, **2005**.

[7] Ahmed Ali Mubark, Hatem Abdulkader, *"Semantic anonymization in publishing categorical sensitive attributes*", IEEE International. Conference, pp. **89-95**, **2016**.

[8] B.C.M. Fung, K. Wang, and P. S. Yu, "*Anonymizing classification data for privacy preservation",* IEEE Transactions Knowledge and Data Engineering, pp. **711-725**, **2007**.

[9] Savita Lohiya, LataRagha, "*Privacy preserving in data mining using hybrid approach",* IEEE International Conference on Computational Intelligence and Communication Networks, pp. **743-746**, **2012**.

## AUTHOR PROFILE

Ms. Athiramol. S pursued Bachelor of Technology in Computer Science and Engineering from CUSAT university in 2015 and Master of Technology from Kerala Technological university in 2017.She is currently working as Assisatnt professsor in Department of Computer Science since 2018. She has published papers in the area of big data security in reputed international journals and IEEE conference in 2017. Her main research work focuses on Big data security and Data mining.

 