Diversity Scale by Supervised Learning for Privacy Preserved and Informative Data Publishing

MD. Riyazuddin[1] Dr. V.V.S.S.S. Balaram[2]

[1]Muffakham Jah College of Engineering & Technology
Banjara Hills, Hyderabad, Telangana State, India
[1]riyaz.mdr1@gmail.com
[2]Sreenidhi Institute of Science & Technology
Ghatkeser, Hyderabad, Telangana State, India

Abstract

The data digitization is the backbone of the distributed computer and communication systems of the current era, which enabled the phenomenal opportunities and benefits in almost all contexts of the new age requirements of human life. However, the scope of vulnerabilities is also critically significant to breach the digital data that evinces the data leakage, privacy fissure and other unethical and unauthorized usage of the data available in the digital platform. In regard to defend the data breaching in the data publishing process, many contemporary models endeavoured toimprove the anonymization and diversification of the sensitive elements of the data used in the publishing process.Nevertheless, these methods could not be generous towards voluminous data. Contemporarily, some of the scalable methods are recognized for preserving privacy data published in literature. Moreover, most of them were on the basis of 1-diversity & k-anonymity. Nonetheless, these methods are vulnerable to data breaches or publishes data that mostly uninformative, which lightens the research scope to derive privacy-preserving methods for data publishing that prevents the vulnerabilities caused for data breaches or results uninformative data to publish. Hence,the contribution of this manuscript intended to define a novel supervised learning-based privacy-preserving technique for secure yet informative data publishing. The proposed method that titled as “Diversity Scale by Supervised Learning (DSSL) for Privacy Preserved and Informative Data Publishing” considers the values projected for the sensitive data elements as label, and verifies the diversity of the other elements as n-gram quasi identifiers of the corresponding data, which are causing the expose of the values (considered as labels) of sensitive elements. Experimental study scaled the significance of the proposal by comparing with the contemporary diversity based techniques.

Keywords: Netflix, PPDP (privacy-preserving data publishing) & PPDM (privacy-preserving data mining),

1 Introduction

Recently, there is an effort in assuring the confidentiality of data & the integrity of its outsourced database. Various proposals of research recommended data encryption before transferring to cloud [1], [2]. Whereas encryption could offer confidentiality of data, it might be less productive in inference deterring attacks. Therefore, it demands novel privacy-enhancing schemes, which can provide confidentiality of data & evade inference attacks simultaneously because of the accumulative query response.

The procedure of anonymizing person-related information to protect the privacy of the individual while handling effective data utility level for data mining is called PPDP (privacy-preserving data publishing). Distinct PPDP privacy methods offer different privacy protection types [3]. The work [4] presents that differential privacy could be dynamic privacy method, which does not predict knowledge of adversary's background. The private differential technique assures that any output probability is equally same from total approximate equivalent input data-sets & hence assures that overall outcomes were insensitive towards any data of an individual.

In this manuscript, the architecture proposed is query processing based on cloud, which preserves requests of query & data confidentiality simultaneously, when different privacy is assured on results of a query for protecting against internal attacks. Now, let us deliberate the following instance. The data BC of the population is a non-beneficial organization responsible for handling and storing patient-related health information received from various hospitals, organizations of health, & agencies of government in British Colombia Province, Canada. The explicit identifiers are utilized by pop-data for data integration, and later hide the data integrated by isolating explicit identifiers from remaining contents of data. The Miners of data who were interested in data querying primarily sign an agreement that is non-recognized for preventing them to release research data, which could be utilized for re-identifying individuals. Here, when the request of data access is received by pop-data, then data miner is authenticated initially, examines that she was researching on an approved project, and later query is executed on re-identified data, and finally the outcome is forwarded back to data Miner. The identical organization could be recognized in other countries.

Data privacy is a significant concern in this instance. Even though the information is re-identified, the Miners of data still execute attribute or record linkage attacks &individual’s re-identification as exhibited in AOL [5]& Netflix [6]. On another dimension, for lessening pop-data workload, the services of cloud could be utilized for managing, answering, & storing queries on data integrated. Nevertheless, this exhibits other 2 concerns. Among them, one is the confidentiality of data, where patient outsourced particular data need to be stored for preventing cloud from queries answering from unauthorized data miners. It also protects in averse to possible multi-tenancy issues because of services sharing, physical architecture & resources among manifold tenants that are independent on the cloud [7]. The other important concern is the confidentiality of query, where cloud needs to be capable of executing the requests of a query

from data miners who are authorized without the capability of knowing what attributes & values of attributes are given for every query.

2 Related work

The work [8] proposed or projected a decomposition method by utilizing distortion of data through random addition of noise. It deliberates sensitive attributes to be an initial sensitive attribute. Here, this contribution projects novel privacy attack & it projects two solutions such as diversity& decomposition method. This method is not appropriate for rigorous releases. Moreover, it offers numerous directions on manifold numerical & categorical sensitive-attributes.

The work [9] proposed a Decomposition+ method for overcoming the confines of partitioning method. It produces dataset with divergent diversity-1 over manifold sensitive-attributes. Here, this method is appropriate for rigorous release. However, it is not appropriate for datasets of high-dimension.

The work [10] proposed novel k-anonymity method on manifold sensitive attributes. Here, this method segregates sensitive attributes into extreme & degraded sensitive attributes. Moreover, tuples were arranged as per extreme sensitive attributes. Further, the association is interrupted amid sensitive attributes for overcoming attack. Besides, it utilizes entropy information to detect equivalence class diversity. This method can lessen the ratio of suppression; however, it could not be appropriate for huge data-sets.

The work [11] proposed a method on manifold sensitive attributes on the basis of preserving privacy data mining by utilizing anonymity. Here, it utilizes a method on non-homogenous anonymization based on clustering for manifold sensitive attributes. Moreover, this method deliberates sensitive attributes sensitiveness. Further, it could not be applied on practical datasets & dataset that is published might be cracked through knowledge background attack.

The work [12] proposed a method SLOMS. It could be preserving privacy publishing data model for manifold microdata of sensitive attributes. This model anonymizes several amounts of attributes in the dataset. Here, the sensitive & quasi attributes are generalized. This model segments manifold sensitive attributes vertically into various tables & tuples are packetized by utilizing 1-diversity.

The work [13] projected MNSACM model by utilizing multi-sensitive packetization & clustering to anonymize datasets. Here, real-time data-sets comprise both categorical & numerical sensitive-attributes; however, it is confined to only numerical-sensitive-attributes and could not be applied over real-time datasets. Further, it offers comparative analysis for 3 divergent packetization methods such as MSDCF, MNSACM & MSB. This method could not be appropriate for enhanced data-sets.

The work [14] projected a model on the anatomization by slicing: a novel privacy preservation method for manifold sensitive attributes. Here, the anatomization method segments sensitive attributes into quasi & sensitive attributes into quasi-table. The MFA model is utilized to bucketize tuples. Here, this contribution utilizes 1-diversity & k-anonymity principles. Published data is protected from the linkage attribute attack, identity attack & membership attack; however, the algorithm of slicing need to be implemented on ST & QIT separately.

The work [15] projected an effective method to publish microdata for manifold sensitive-attributes. Here, it utilizes angelization method for anonymization of manifold sensitive-attributes. Moreover, published privacy data is preserved by evading background threat attack knowledge & membership-attack. Here, it is confined to case, where there could be one record associated with an individual presence in specified dataset.

The work [16] proposed KC-slice method. It could be rigorous in preserving privacy published data scheme for manifold sensitive attributes by integrating LKC privacy method & slicing features [17]. This method implements the scheme on only 1 unique value amid various sensitive attribute values. The KC-slice implements similar threshold quantity for entire sensitive attributes regardless of sensitiveness because of which it might result to a superior ratio of suppression.

The anonymization of data initiated with k-anonymity, proposed in [18], [19]. Consequently, the contribution is extended for diverse data types like graph data [20], [21], value data set [22], streaming-data [23], [24] & relational-data [25], [26], [27], [28]. Numerous anonymization methods like t-closeness [29], (c,t)-isolation [30], [31], differential privacy [32], [33], [34], δ-presence [35], l-diversity [36], [37], & manifold-association k-anonymity [38], [35], [27] is proposed by divergent researchers. The anonymization methods could be classified as PPDP (privacy preserving data publishing) & PPDM (privacy preserving data mining) [39], [40]. Further, PPDP methods publish entire dataset in the anonymized way, while PPDM publishes anonymized queried data. The l-diversity, k-anonymity & t-closeness were some of well-recognized PPDP methods. Differential privacy, (c,t)-isolation & δ-presence are some of well-recognized PPDM schemes. The work [41] identifies k-anonymity as an optimal candidate for addressing entire big-data 3Vs by comparing with contemporary PPDP method with entire big data 3Vs.

The work [42], [43], [44] presents that, review associated with PPDP scalability methods are confined. Nevertheless, we found some of PPBDP method that is scalable & research with a distributed framework of programming. The work [45] proposed top-down 1-phase specialization (TPTDS), which could be an extension of privacy-preserving method. The work [46] proposed TDS (top-down-specialization) method by utilizing the architecture of MapReduce. Moreover, it could be perceived that, for minimum k values, the TDS method faces a maximum time of running, while BUG (bottom-up-generalization) method projected by [47] provides a maximum time of running for maximum k-values.

Hence, a hybrid model for anonymization of scalable sub-tree is projected, where anonymization model is chosen on the basis of k-values [48]. The prominent confine of these methods could be data-distribution, where huge data crowd impact is utilized & as an alternative, it becomes distributed anonymization of data. The work [49] projected multi-dimensional anonymization method. The work [26] proposed an algorithm for the framework of MapReduce. They proposed 2 versions of MapReduce on the basis of anonymization (MRA). In the primary version, the specified data could be deliberated as unique equivalence class & partition later performed in regard to tuples with every attribute till novel generated class fulfil the condition of k-anonymity.

After iteration, data-file could be upgraded according to the novel class of equivalence, and it is specified as an input for further iteration. Global-file requirements are to share amid entire nodes for upgrading the information of equivalence class after the iteration is deliberated to be a significant confine of the projected method. Here, global-file becomes more and more after every iteration. The 2nd version of anonymization based on MapReduce is projected for overcoming the 1st version confine. In spite of producing a unique global file for entire nodes, files chunks were distributed & generated among entire nodes. In a step of mapping, every node appends individual id of file for every file to identify. Therefore, in further iterations, every node requires to be accessed only files that have been upgraded by its respective reducer.

A load of global file maintenance could be eradicated in this approach. Nevertheless, manifold iteration & management of file were 2 important confines of this method. As there is an increment in the number of iterations, then system performance decreases & file-management by utilizing MapReduce becomes an intricate task. The work [50], [51] projected novel method called SKA (Scalable k-anonymization) by utilizing MapReduce. The term SKA could be reverse method entirely for performing an algorithm called Mondrain [26]. In MRA [49] & Mondrain methods, the input dataset could be deliberated as a single class of equivalence & segmented into sub-classes till the condition of k-anonymity is true, while in an instance of SKA method, the data set could be arranged & segmented into small possible classes by identical values within it. Moreover, it is combiner iteratively until it satisfies the condition of k-anonymity. The enhanced version of SKA has explored in the following sections. Further, we found that SKA could be extended towards 1-diversity. Hence, we proposed an enhanced, scalable 1-diversity method to be an extension of enhanced SKA version.

The contemporary contributions “unfiltered data anonymization” [52], and “information gain based anonymization with slicing Model” [53] found to be other competent methods to preserve privacy in data publishing. These two methods have portrayed the improvisation of the diversity and anonymization techniques in respective order. The first one is improvised l-diversity technique that aimed to achieve privacy preserving of the dynamically publishing data, with frequent changes. The other one is improvised anonymization technique that aimed to achieve preserving privacy with minimal distortion of the publishing data. However, these contemporary methods haven’t defined their ability to handle the crux of high dimensionality

appeared in the values assigned to corresponding attributes of the publishing data.The contribution of this manuscript aimed to deal the crux of dimensionality in values of the sensitive parameters to preserve privacy in publishing data.

3 Methods and Materials

This section exhibits the description of the method proposed for privacy preserved and informative data publishing, which includes methods used and materials required to perform the privacy-preserving of the subjects represented by the data to be published.

3.1 Model Description

The objective of the model proposed is to identify the quasi-identifiers sensitive to exhibit the subject of the data to be published. The data format that competent to apply the proposed model is a set of records, such that each record represents the diversified elements, which appears in one or more records of the corresponding data. The initial phase of the proposal lists out the sensitive values appearing in one or more records as subject-labels from a given dataset of multiple records.

Further, performs a novel search toidentify dynamic n-gram quasi-identifiers and the diversity of these quasi n-gram identifiers and the subject labels exhibited by the corresponding quasi identifier. After that, estimates the sensitivity of the quasi n-gram identifiers and corresponding subject labels by using the diversity observed for both quasi-identifiers and subject labels.

3.2 The Data

Let the notation DC denotes the data corpus having a set of records of the count $|DC|$. Let the notation sL denotes the set that consists of subject-labels, which often provided along with the data corpus. If the set sL is empty, then the proposed model considers each entry that exists in one or more records of the given data corpus as subject-labels, which performs as follows

$\forall_{i=1}^{|DC|}\{(sL = sL \cup r_i)\exists r_i \in DC\}$ // iterates on each record r_i of the data corpus and moves the entries of the record r_i, which are not existing in the subject-label list sL

3.3 Subject Diversity Weights

This factor denotes the sensitivity of the subject-label, which is the empirical probability of the corresponding subject-label

$\forall_{i=1}^{|sL|}\{s_i \exists s_i \in sL\}$ // iterates on each subject-label $\{s_i \exists s_i \in sL\}$ listed in the set sL

$s_{dw} = \left(\sum_{j=1}^{|DC|}\{1 \exists s_i \in r_i \wedge r_j \in DC\}\right) * (|DC|)^{-1}$ // finding the empirical probability of each subject label towards the given data corpus DC

End

3.4 Quasi Identifiers and corresponding diversity weights

//Initially, the quasi n-gram identifiers and corresponding diversity weights for n=1 have to find as follows

$\forall_{i=1}^{|DC|}\{r_i \exists r_i \in DC\}$ // iterates on each record r_i of the data corpus and moves the entries of the record r_i, which are not existing in n-gram list nGl

$r_i \setminus (nGl \cap r_i)$ // discard the entries of the record r_i, which exist in both record r_i and list nGl

$\forall_{j=1}^{|r_i|}\{e_j \exists e_j \in r_i\}$ Begin // for each leftover element of the record r_i

$e_j^{dw} = \left(\sum_{k=1}^{|DC|}\{1 \exists e_j \in r_k \wedge r_k \in DC\}\right) * (|DC|)^{-1}$ // discovering the diversity weight e_j^{dw} of element $\{e_j \exists e_j \in r_i\}$

End

$nGl \leftarrow \{r_i \exists |r_i| > 0\}$ // add all the entries if any of the record r_i

End

//Further discovers the n-gram quasi identifiers of dynamic size and their respective diversity weights

Do

$tnGl = null$ // an empty set

$\forall_{i=1}^{|nGl|}\{ng_i \exists ng_i \in nGl\}$ // iterates on each quasi n-gram identifier $\{ng_i \exists ng_i \in nGl\}$ of the list nGl

$\forall_{j=1}^{|nGl|}\{ng_j \exists ng_j \in nGl \wedge i \neq j\}$ // iterates on each quasi n-gram identifier $\{ng_j \exists ng_j \in nGl \wedge j \neq i\}$ of the list nGl, which is not equal to the quasi n-gram identifier ng_i,

$ng = ng_i \cup ng_j$ // finding new n-gram quasi identifier

$if\ (ng \notin nGl \wedge ng \notin tnGl)$ Begin

$ng_{dw} = \left(\sum_{k=1}^{|DC|}\{1 \exists ng \subseteq r_k \wedge r_k \in DC\}\right) * (|DC|)^{-1}$ // discovering the diversity weight ng_{dw} of the quasi n-gram identifier ng

$if\ (ng_w > 0)\ tnGl \leftarrow ng$

End

End

End

$if\ ((|tnGl| > 0) \wedge ((tnGl \cap nGl) \neq tnGl))$ Begin

$\forall_{i=1}^{|tnGl|}\{ng_i \exists ng_i \in tnGl\}$ Begin // iterates on each quasi n-gram identifier $\{ng_i \exists ng_i \in tnGl\}$ of the list $tnGl$

$ng_i^{dw} = \left(\sum_{j=1}^{|DC|}\{1 \exists ng_i \subseteq r_j \wedge r_j \in DC\}\right) * (|DC|)^{-1}$ // discovering the diversity weight ng_i^{dw} of quasi n-gram identifier $\{e_j \exists e_j \in r_i\}$

$nGl \leftarrow (tnGl \cup nGl)$ // adding new n-gram quasi-identifiers have added to the list nGl

End

End

$while((|tnGl| > 0) \wedge ((tnGl \cap nGl) \neq tnGl))$ // while new n-gram quasi-identifiers have generated

3.5 Privacy weights

The term privacy weight has coined to define the correlation between the n-gram quasi-identifiers and the subject-labels. The privacy weight of the n-gram quasi identifiers nGl listed and the subject-labels listed sL have to estimate as follows.

Initially finds the quasi identifier level diversity of subject-labels as follows

$\forall_{i=1}^{|sL|}\{s_i \exists s_i \in sL\}$ // iterates on each subject-label $\{s_i \exists s_i \in sL\}$ listed in the set sL

$s_i^{qdw} = \left(\sum_{j=1}^{|nGl|}\{1 \exists s_i \subseteq ng_j \wedge ng_j \in nGl\}\right) * (|nGl|)^{-1}$ // discovering the quasi identifier level diversity weight of each subject-label

Further discovers, privacy weights of the quasi-identifiers as follows

$\forall_{j=1}^{|nGl|}\{ng_j \exists ng_j \in nGl\}$ Begin // for each n-gram quasi identifier

$ng_j^{pw} = 1$

$\forall_{i=1}^{|sL|}\{s_i \exists s_i \in sL\}$ Begin // for each subject label

$if\ (s_i \in ng_j)\ ng_j^{pw} = ng_j^{pw} * s_i^{qdw}$

End

$ng_j^{pw} = 1 - ng_j^{pw}$ // diversity has normalized to its maximum weight, which is in between 0 and 1.

End

Further discovers, privacy weights of the subject-labels as follows

$\forall_{i=1}^{|sL|}\{s_i \exists s_i \in sL\}$ Begin // for each subject label

$s_i^{pw} = 1$

$\forall_{j=1}^{|nGl|}\{ng_j \exists ng_j \in nGl\}$ Begin // for each n-gram quasi identifier

$if\,(s_i \in ng_j)\ s_i^{pw} = s_i^{pw} * ng_j^{dw}$

End

$s_i^{pw} = 1 - s_i^{pw}$ // privacy weight s_i^{pw} of the subject s_i has normalized to its maximum privacy weight, which is in between 0 and 1.

End

Further, the n-gram quasi-identifiers and subject-labels having privacy weight less than the given privacy threshold shall conclude as the most vulnerable to a privacy breach.

4 Experimental results

The experimental study has carried out on a mechine having the following specifications. The operating system used is windows 10, which installed on Intel processor of 1.6GHz with a RAM of 4.0 GB and storage of hard disk to be 1TB. The projected or proposed DSSL model and contemporary models hasprogramed by utilizing java. Here, from UCI (University of California Irwin) adult dataset the repository of machine learning has utilized as input [54].

We assess our method known as DSSL that intended to preserve the privacy of rigorously published datasets, which determined by comparing with the contemporary models ENC (unfiltered data –anonymization) [52] & IGASM (information gain based anonymization with slicing Model) [53] associated towards data through measuring NCP (normalized certainty penalty) & IL (information loss) for the transaction & relational attributes. Also, we performed a simulation assessment regarding the impact of 2 datasets on computational-cost for 3 models. Moreover, to evaluate the tradeoff of data information & disclosure risk in anonymization of the transaction of data,

Further, we compared the proposed method DSSL with contemporary methods IGASM [53] & ENC [52] on a benchmark dataset. Notice that IGASM could not normalize the attributes of a transaction as high as ENC that implements more normalization towards the attributes of the transaction to preserve more relational attributes information. Here, in either instance, the objective of AMT-dataset anonymizing is to enable a significant analysis of transaction & relational attributes that defeated.

Comparatively, the impact of on IL of a graph (see Figure 1) on the basis of DSSL could be insignificant,, and it is detected to outperform others consistently on entire. This could be mainly due to IGASM method [52], [53] forces huge amount of items for generalizing together when the DSSL based on graph modifies count ofsensitive subjects graph for preserving the relational rules among items. In the mean-time, the associations among distinct attributes types were preserved by a count of sensitive subjects in architecture. Nevertheless, our method is constant and dataset that means our method is more appropriate for spare transactional-data.

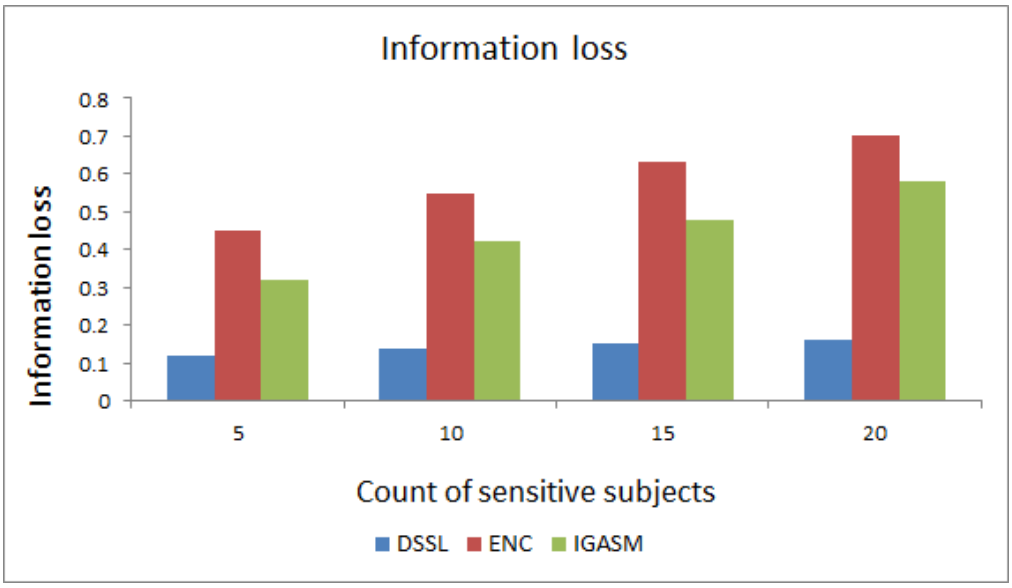

Figure 1: Information loss for the DSSL, ENC & IGASM methods

In Figure 1, the graph is drawn between information loss and count of sensitive subjects over the proposed method DSSL and contemporary methods ENC & IGASM. Here, the proposed method is compared with contemporary methods. From the statistics as depicted in the above figure exhibits that, the information loss for the proposed DSSL model is less when compared with contemporary ENC & IGASM models.

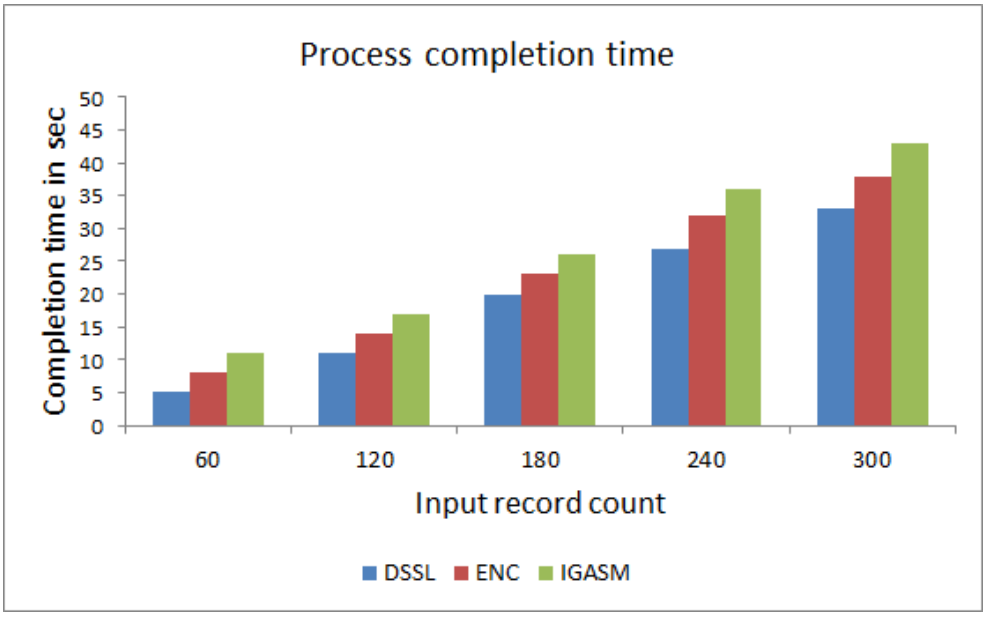

Figure 2: Process completion time for the DSSL, ENC & IGASM methods

In Figure 2, the graph is drawn between completion time in seconds and input record count over the proposed method DSSL and contemporary methods ENC & IGASM. Here, the proposed method is compared to contemporary methods. From the statistics as depicted in the above figure

exhibits that, the completion time for the proposed DSSL model is less when compared with contemporary ENC & IGASM models.

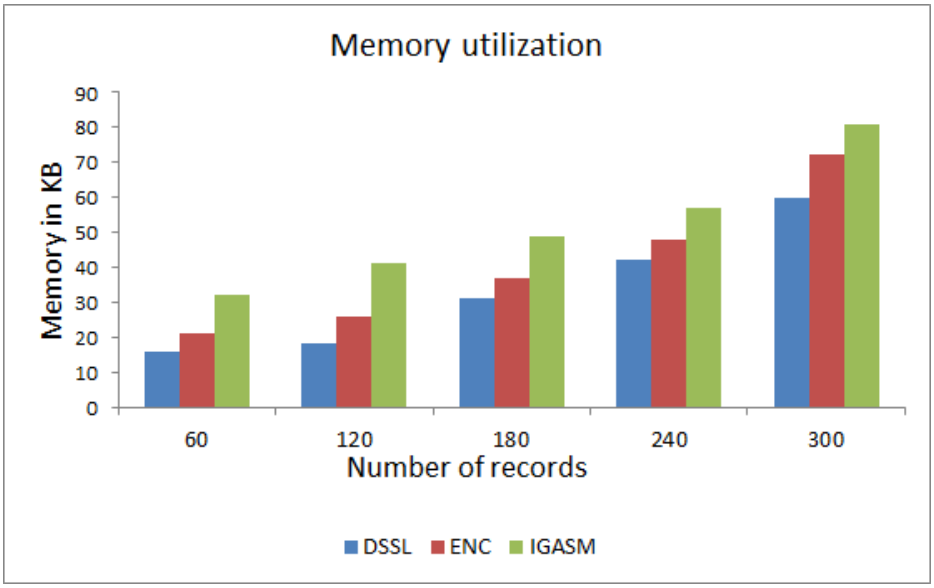

Figure 3: Memory utilization for the DSSL, ENC & IGASM methods

In Figure 3, the graph is drawn between memory utilization and number of records over the proposed method DSSL and contemporary methods ENC & IGASM. Here, the proposed method is compared with contemporary methods. From the statistics as depicted in above figure exhibits that, the memory utilization in KB for the proposed DSSL model is less when compared with contemporary ENC & IGASM models.

By integrating the above simulation outcomes, the anonymized data quality exhibits that our method is always more compared to other methods by utilizing entire metrics. Our method is devised for anonymizing for preventing a multi-objective attack. Here, special focus is towards information enhancement while providing robust assurance for the manifold sensitive attributes. Our method not only includes divergent kinds of privacy pre-requisites yet also combines divergent information pre-requisites. We finalize that, our multifold method based on a graph to anonymize AMT datasets is protective & resourceful practically in the applications.

5 Conclusion

The contribution of this manuscript has aimed to defend the data breach in the publishing process. Adopting sensitive subjects to anonymize is the significance of the proposed model. The proposal is able to protect the data from breaches in data, without having inputs about quasi-identifiers and sensitive subjects. In contrast to the proposal, thecontemporary models require information about a sensitive subject and quasi-identifiers. The significant performance advantage of the proposed model DSSL has noticed by comparing the information loss noticed from the other contemporary models at diversified sensitive subjects count. The other performance metrics related to memory usage and process completion time also verified in an

experimental study, which signifying the proposed model DSSL over the contemporary models. The usage of evolutionary techniques to identify dynamic n-gram quasi-identifiers and sensitive subjects would be the significant scope for future research contributions.

References

[1] Ge T, Zdonik S. Answering aggregation queries in a secure system model. InProceedings of the 33rd international conference on Very large data bases 2007 Sep 23 (pp. 519-530).

[2] Popa RA, Redfield CM, Zeldovich N, Balakrishnan H. CryptDB: protecting confidentiality with encrypted query processing. InProceedings of the Twenty-Third ACM Symposium on Operating Systems Principles 2011 Oct 23 (pp. 85-100).

[3] Fung BC, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (Csur). 2010 Jun 23;42(4):1-53.

[4] DWORK C. Differential Privacy. Lecture notes in computer science. 2006.

[5] Barbaro M, Zeller T, Hansell S. A face is exposed for AOL searcher no. 4417749. New York Times. 2006 Aug 9;9(2008):8.

[6] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In2008 IEEE Symposium on Security and Privacy (sp 2008) 2008 May 18 (pp. 111-125). IEEE.

[7] Dillon T, Wu C, Chang E. Cloud computing: issues and challenges. In2010 24th IEEE international conference on advanced information networking and applications 2010 Apr 20 (pp. 27-33). Ieee.

[8] Ye Y, Liu Y, Wang C, Lv D, Feng J. Decomposition: privacy preservation for multiple sensitive attributes. InInternational Conference on Database Systems for Advanced Applications 2009 Apr 20 (pp. 486-490). Springer, Berlin, Heidelberg.

[9] Das D, Bhattacharyya DK. Decomposition+: improving ℓ-diversity for multiple sensitive attributes. InInternational Conference on Computer Science and Information Technology 2012 Jan 2 (pp. 403-412). Springer, Berlin, Heidelberg.

[10] Liu F, Jia Y, Han W. A new k-anonymity algorithm towards multiple sensitive attributes. In2012 IEEE 12th International Conference on Computer and Information Technology 2012 Oct 27 (pp. 768-772). IEEE.

[11] Usha P, Shriram R, Sathishkumar S. Multiple sensitive attributes based privacy preserving data mining using k-anonymity. Int. J. Sci. Eng. Res. 2014;5(4).

[12] Han J, Luo F, Lu J, Peng H. SLOMS: A Privacy Preserving Data Publishing Method for Multiple Sensitive Attributes Microdata. JSW. 2013 Dec 1;8(12):3096-104.

[13] Liu Q, Shen H, Sang Y. Privacy-preserving data publishing for multiple numerical sensitive attributes. Tsinghua Science and Technology. 2015 Jun 19;20(3):246-54.

[14] Susan VS, Christopher T. Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes. SpringerPlus. 2016 Dec 1;5(1):964.

[15] Anjum A, Ahmad N, Malik SU, Zubair S, Shahzad B. An efficient approach for publishing microdata for multiple sensitive attributes. The Journal of Supercomputing. 2018 Oct 1;74(10):5127-55.

[16] Onashoga SA, Bamiro BA, Akinwale AT, Oguntuase JA. KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes. Information Security Journal: A Global Perspective. 2017 May 4;26(3):121-35.
[17] Mohammed N, Fung BC, Hung PC, Lee CK. Anonymizing healthcare data: a case study on the blood transfusion service. InProceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining 2009 Jun 28 (pp. 1285-1294).
[18] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. InPODS 1998 Jun 1 (Vol. 98, No. 10.1145, pp. 275487-275508).
[19] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.
[20] Liu K, Terzi E. Towards identity anonymization on graphs. InProceedings of the 2008 ACM SIGMOD international conference on Management of data 2008 Jun 9 (pp. 93-106).
[21] Hay M, Miklau G, Jensen D, Towsley D, Weis P. Resisting structural re-identification in anonymized social networks. Proceedings of the VLDB Endowment. 2008 Aug 1;1(1):102-14.
[22] Xue M, Karras P, Raïssi C, Vaidya J, Tan KL. Anonymizing set-valued data by nonreciprocal recoding. InProceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012 Aug 12 (pp. 1050-1058).
[23] Zakerzadeh H, Osborn SL. Delay-sensitive approaches for anonymizing numerical streaming data. International journal of information security. 2013 Oct 1;12(5):423-37.
[24] Zhou B, Han Y, Pei J, Jiang B, Tao Y, Jia Y. Continuous privacy preserving publishing of data streams. InProceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology 2009 Mar 24 (pp. 648-659).
[25] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. InProceedings of the 2005 ACM SIGMOD international conference on Management of data 2005 Jun 14 (pp. 49-60).
[26] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. In22nd International conference on data engineering (ICDE'06) 2006 Apr 3 (pp. 25-25). IEEE.
[27] Nergiz ME, Clifton C, Nergiz AE. Multirelational k-anonymity. IEEE Transactions on Knowledge and Data Engineering. 2008 Oct 17;21(8):1104-17.
[28] Wong WK, Mamoulis N, Cheung DW. Non-homogeneous generalization in privacy preserving data publishing. InProceedings of the 2010 ACM SIGMOD International Conference on Management of data 2010 Jun 6 (pp. 747-758).
[29] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In2007 IEEE 23rd International Conference on Data Engineering 2007 Apr 15 (pp. 106-115). IEEE.
[30] Chawla S, Dwork C, McSherry F, Smith A, Wee H. Toward privacy in public databases. Lecture Notes in Computer Science. 2005;3378:363-85.

[31] Chawla S, Dwork C, McSherry F, Talwar K. On privacy-preserving histograms. arXiv preprint arXiv:1207.1371. 2012 Jul 4.
[32] Cynthia D. Differential privacy. Automata, languages and programming. 2006 Jul 9:1-2.
[33] Dwork C. Ask a better question, get a better answer a new approach to private data analysis. InInternational Conference on Database Theory 2007 Jan 10 (pp. 18-27). Springer, Berlin, Heidelberg.
[34] Dwork C. Differential privacy: A survey of results. InInternational conference on theory and applications of models of computation 2008 Apr 25 (pp. 1-19). Springer, Berlin, Heidelberg.
[35] Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. InProceedings of the 2007 ACM SIGMOD international conference on Management of data 2007 Jun 11 (pp. 665-676).
[36] Machanavajjhala A, Gehrke J, Kiefer D, Venkitasubramanian M. *l*-Diversity: Privacy beyond *k*-anonymity. InProc. IEEE Int. Conf. Data Eng.(ICDE) 2006 (p. 24).
[37] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007 Mar 1;1(1):3-es.
[38] Nergiz ME, Clifton C, Nergiz AE. MultiRelational k-Anonymity. In2007 IEEE 23rd International Conference on Data Engineering 2007 Apr 15 (pp. 1417-1421). IEEE.
[39] Cormode G, Srivastava D. Anonymized data: generation, models, usage. InProceedings of the 2009 ACM SIGMOD International Conference on Management of data 2009 Jun 29 (pp. 1015-1018).
[40] Clifton C, Tassa T. On syntactic anonymity and differential privacy. In2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW) 2013 Apr 8 (pp. 88-93). IEEE.
[41] Mehta BB, Rao UP, Kumar N, Gadekula SK. Towards privacy preserving big data analytics. InProc. 2016 Sixth Int. Conf. Advanced Computing and Communication Technologies, Ser. ACCT-2016, Rohtak, India: Research Publishing 2016 Sep (pp. 28-35).
[42] Mehta B, Rao UP, Gupta R, Conti M. Towards privacy preserving unstructured big data publishing. Journal of Intelligent & Fuzzy Systems. 2019 Jan 1;36(4):3471-82.
[43] Sangeetha S, Sadasivam GS. Privacy of big data: a review. InHandbook of Big Data and IoT Security 2019 (pp. 5-23). Springer, Cham.
[44] Canbay Y, Vural Y, Sagiroglu S. Privacy preserving big data publishing. In2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT) 2018 Dec 3 (pp. 24-29). IEEE.
[45] Zhang X, Yang LT, Liu C, Chen J. A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. IEEE Transactions on Parallel and Distributed Systems. 2013 Feb 25;25(2):363-73.

[46] Fung BC, Wang K, Philip SY. Anonymizing classification data for privacy preservation. IEEE transactions on knowledge and data engineering. 2007 Mar 26;19(5):711-25.

[47] Wang K, Yu PS, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection. InFourth IEEE International Conference on Data Mining (ICDM'04) 2004 Nov 1 (pp. 249-256). IEEE.

[48] Zhang X, Liu C, Nepal S, Yang C, Dou W, Chen J. A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. Journal of Computer and System Sciences. 2014 Aug 1;80(5):1008-20.

[49] Zakerzadeh H, Aggarwal CC, Barker K. Privacy-preserving big data publishing. InProceedings of the 27th international conference on scientific and statistical database management 2015 Jun 29 (pp. 1-11).

[50] Mehta BB, Rao UP. Privacy preserving big data publishing: a scalable k-anonymization approach using MapReduce. Iet Software. 2017 Jul 31;11(5):271-6.

[51] Mehta BB, Rao UP. Toward Scalable Anonymization for Privacy-Preserving Big Data Publishing. InRecent Findings in Intelligent Computing Techniques 2018 (pp. 297-304). Springer, Singapore.

[52] Temuujin O, Ahn J, Im DH. Efficient L-Diversity Algorithm for Preserving Privacy of Dynamically Published Datasets. IEEE Access. 2019 Aug 19;7:122878-88.

[53] Gachanga E, Kimwele M, Nderu L. Feature Based Data Anonymization with Slicing Method for Data Publishing. InProceedings of the 2019 11th International Conference on Machine Learning and Computing 2019 Feb 22 (pp. 274-279).

[54] Newman DJ. UCI repository of machine learning databases, University of California, Irvine. http://www. ics. uci. edu/mlearn/MLRepository. html. 1998.