EDAMS: Efficient Data Anonymization Model Selector for Privacy-Preserving Data Publishing

Tehreem Qamar
Department of Computer Science and Software Engineering
Jinnah University for Women
Karachi, Pakistan
tehreem.qamar@juw.edu.pk

Narmeen Zakaria Bawany
Department of Computer Science and Software Engineering
Jinnah University for Women
Karachi, Pakistan
nsb@juw.edu.pk

Najeed Ahmed Khan
Department of Computer Science & Information Technology
NED University of Engineering & Technology
Karachi, Pakistan
najeed@neduet.edu.pk

***Abstract*—The evolution of internet to the Internet of Things (IoT) gives an exponential rise to the data collection process. This drastic increase in the collection of person's private information represents a serious threat to his/her privacy. Privacy Preserving Data Publishing (PPDP) is an area that provides a way of sharing data in their anonymized version, i.e. keeping the identity of a person undisclosed. Various anonymization models are available in the area of PPDP that guard privacy against numerous attacks. However, selecting the optimum model which balances utility and privacy is a challenging process. This study proposes an Efficient Data Anonymization Model Selector (EDAMS) for PPDP which generates an optimized anonymized dataset in terms of privacy and utility. EDAMS inputs the dataset with required parameters and produces its anonymized version by incorporating PPDP techniques while balancing utility and privacy. EDAMS is currently incorporating three PPDP techniques, namely k-anonymity, l-diversity, and t-closeness. It is tested against different variations of three datasets. The results are validated by testing each variation explicitly with the stated techniques. The results show the effectiveness of EDAMS by selecting the optimum model with minimal effort.**

Keywords-data anonymization; privacy-preserving data publishing; k-anonymity; l-diversity; t-closeness

I. INTRODUCTION

The advent of IoT, high processing speed hardware, and cloud storage with high bandwidth communication produces vast amounts of data which would be unthinkable a couple of decades ago. Due to these advancements, around 2.5 quintillion bytes of data are created each day [1]. Such huge production of information not only advances users' quality of life, but also enhances various vital administrations. The data collection process is not governed by a single entity [2]. The applications used in order to perform daily routine activities efficiently are constantly saving, collecting, and tracking user data. Moreover, companies are encouraged to release their micro-data in order to facilitate data analysis that eventually supports providing new business opportunities [3, 4]. However, the release of micro-data results in tracking the public and private lives of concerned individuals, thus putting their privacy at risk [3, 5, 6]. A typical data collecting and publishing scenario is depicted in Figure 1. In the data collection phase, data holders gather data from individuals, i.e. record owners (e.g. Ahmed, Haris, Laraib, Sana). In the publishing phase the data are provided to data recipients who can be data miners or other third parties that can make use of that data for their own purposes.

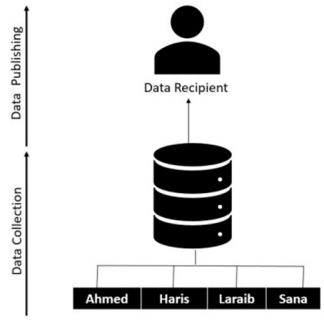

Fig. 1. Data collection and data publishing

The published records may contain sensitive information [7-11]. To secure data owners' privacy and to avoid data exploitation, eradicating identifiable attributes like name, address, telephone number, and social security numbers is a common practice prior to data release. However, this simplistic technique is not sufficient to guarantee the protection of record owners. Data publishing in a way that they contain no sensitive information and the privacy of record owners remains intact is termed as PPDP [7]. Typically, PPDP deals with publishing of data in an anonymized way, i.e. the data contain sensitive information but that information cannot be linked with its owner, while being still useful for the interested parties. Various methods have been proposed [12-15] for transforming data into their anonymized version. These methods differ in their capabilities of preventing linking of data owners which can eventually harm their privacy. There is no standard method for selecting a particular anonymization technique. Technique selection is highly dependent on the type of dataset and its sensitive attributes. The publisher has to anonymize data by using multiple techniques in order to select the most suitable. This is not only expensive in terms of time and resources but also requires sufficient knowledge in order to choose the appropriate method to convert the actual data into their

Corresponding author: Tehreem Qamar

anonymized version. Selection of an inappropriate method may cause data loss therefore it is necessary to select a method which could provide results at the optimum level of its utility with the least possible loss of data. Keeping in view the aforementioned problems, this study aims to propose a model that can identify the most suitable technique for anonymizing a certain dataset with minimum information loss. The main contributions of the current study are:

- The development of a model that helps the data holder who has no particular knowledge of data anonymization techniques to release data anonymously.
- The selection of the most appropriate method according to the nature of the respective dataset.
- The generation of an anonymized dataset with least information loss and maximized utility.

II. LITERATURE REVIEW

Various real world attacks indicate the significance of preserving individual privacy when distributing personal information. Many times data released by companies for research purposes ended up with hurting individual privacy. The re-identification of individuals happens when they get linked with some other available external information is termed as linking attack [12]. Some reported incidents regarding released data that got linked with external information are summarized in Table I.

TABLE I. UNCIDENTS REPORTED AGAINST LINKING ATTACKS

Privacy breach	Dataset used	Results
[16]	Health dataset from Washington State	43% identification by linking the dataset with newspaper stories containing the word "hospitalized".
[17]	Prescription data of South Korean residents	100% individuals in the dataset were re-identified. Data were encrypted prior release.
[12]	Medical records of state employees of Massachusetts	Governor of Massachusetts was identified when the dataset was linked with the publicly available voter enrollment list.
[18]	Three month credit card records	90% identification by analyzing buying patterns
[9]	AOL dataset	One of the users was identified and interviewed by New York Times within three days of data release
[19]	Netflix dataset	99% of records were identified with 8 movie ratings

Authors in [16] collected a health dataset from Washington State, which did not contain patient names. However, 43% of the individuals were successfully identified by linking the dataset with the newspaper stories containing the word "hospitalized". Authors in [17] conducted experiments on the encrypted prescription data of 23,163 South Korean Resident Registration Numbers (RRNs). They claimed that they were able to re-identify 100% of the data and concluded that encrypted data are also vulnerable. Author in [12] described the re-identification of the dataset released by Group Insurance Commission (GIC) that included medical records of the state employees of Massachusetts and was intended to facilitate medical research. The dataset contained demographic data, for example, birth date, gender, and zip code. It was explained how easily William Weld (the then governor of Massachusetts) was identified by linking the Massachusetts voter enrollment list with the information given by GIC. Authors in [18] studied a credit card report of 3 months consisting of 1.1 million individuals and uniquely identified 90% of them via analyzing only four spatiotemporal points. They reported that the buying patterns with a use of a credit card make an individual's privacy vulnerable. A similar incident has been reported in 2006 when AOL released 20 million search queries of its users and within three days of its release one of its users was identified and interviewed by New York Times [9]. A few months later, Netflix also faced re-identification of its users in the dataset it released for the development of an accurate movie recommendation algorithm. The data were attacked by authors in [19], and they showed that external information can be linked to identify or to link the data with the respective individual.

PPDP is a way of releasing anonymized data while preserving individual privacy [6]. In PPDP, the data are generally represented as a Table of Explicit Identifiers, Quasi Identifiers, Sensitive Attributes, and Non-Sensitive Attributes, where Explicit Identifiers is a set of attributes that explicitly identifies the individual, and Quasi Identifiers are those that could potentially identify the individual. Sensitive person-specific information such as salary, real time location and disability status are considered as Sensitive Attributes while the term Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories. Numerous techniques and models have been proposed in PPDP for producing anonymized data such as k-anonymity [12], l-diversity [13], and t-closeness [20], which have become the foundation of many other models [22-26] and are therefore used in EDAMS.

III. THE EDAMS MODEL

A. Preliminaries

Let T be an original data table of the following form:

T = {DI1, DI2,……, DIj, QI1, QI2,……, QIk, SA1, SA2,… SAn}

where, DIs are Direct Identifiers, the attributes which should be removed prior data publishing, QIs are Quasi Identifiers, the non-sensitive attributes which when linked with external data can reveal the identity of a record owner, and SAs are Sensitive Attributes, the private information related to a record owner.

B. Methodology

The proposed data anonymization model initially makes use of k-anonymity, l-diversity and t-closeness as privacy models, and generalization and suppression as PPDP operations. The utility that guarantees the optimum information loss is Information Loss (ILoss) metric [27], which measures the loss of information by calculating the uncertainty that occurred in generalizing a value which relies upon how many other values cannot be distinguished from it. The overall anonymization process is depicted in Figure 2.

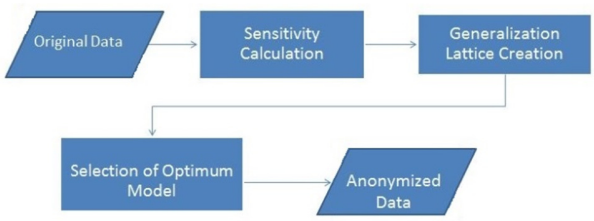

Fig. 2. EDAMS's data anonymization process

The process comprises of 5 steps. In the first step, the original data are taken as input that clearly marks the DIs, QIs and SAs. After realizing the attribute's nature, the sensitivity of the overall dataset is calculated. As the sensitivity is computed, the generalization hierarchy of the QIs is generated. And, on the basis of sensitivity of the dataset, the optimum privacy model is selected for its anonymized version. The sensitivity of the dataset is calculated by:

$$\text{Sensitivity percentage} = \frac{\text{Number of SA}}{\text{Number of QI}} \text{ x } 100 \quad (1)$$

If the sensitivity is 0 that means no sensitive attribute is present in the dataset and k-anonymity privacy model will be used. Applying k-anonymity requires the value of k to be used optimally because it is responsible for the utility ratio of the dataset [28]. EDAMS makes use of two PPDP operations, i.e. generalization and suppression. The generalization lattice is created for each QI. DI and the attributes that cannot be generalized will get suppressed in the resulting anonymized table. When the above two steps are completed then the optimum model is chosen on the basis of sensitivity. The information loss is calculated via ILoss metric [7] and the data holder will get the anonymized version of the data with least cost. Figure 3 depicts the applied algorithm.

Algorithm: Anonymizing Data

Input: Dataset T, and the parameters (DI, QI and SA)
Output: Anonymized Dataset T*

1. Calculate sensitivity percentage of T
2. Determine generalization lattice for each QI
3. Let p = optimized model on the basis of sensitivity percentage
4. anonymize(T,p) and calculate ILoss
5. Return T* with ILoss

Fig. 3. Data anonymization algorithm for EDAMS

IV. EXPERIMENTS

EDAMS is developed using Java that run on a 2.4GHz Intel Core i5 Processor with 6GB RAM. Three datasets with their customized versions were examined for the assessment of the model, namely UCI Adult dataset [29], Employee's Salary dataset [30], and Crime Incident dataset [31] along with their different variations. Each dataset has been evaluated twice. Firstly with EDAMS and secondly with each method separately applied to it in order to get the optimal result in a process termed as Hit and Trial. Its results are shown in Table II.

A. Case 1: Adult Dataset

The dataset in [29] contains 30,162 records. It consists of 9 attributes in total: sex, age, race, marital-status, education, native-country, work class, occupation, and salary-class. Three variants of this dataset were considered of having no DI. The first variation took all attributes as QIs. The second variation considered occupation as an SA and the rest of them as QIs, and the third variation included six of them as QIs and two of them, i.e. marital-status and occupation, as sensitive.

1) Selection via EDAMS

When dataset is taken as input to EDAMS, its sensitivity is calculated, i.e. the ratio of SA over Quasi Attributes. Considering the first variation, when there is no SA the sensitivity between SA and QI becomes 0% which means although the dataset has no direct sensitive information, it can serve as a tool for linking attacks. In this case, EDAMS suggested k-anonymity for the respective dataset with maximum information loss of 60%. Table II represents the chosen models with maximum information loss when the same procedure was applied to all of its variants.

TABLE II. RESULT OF ADULT DATASET VIA EDAMS

DI	QI	SA	Privacy model	Sensitivity ratio	Max. information loss
0	9	0	k-anonymity	0%	60.37%
0	8	1	l-diversity	11%	62.50%
0	7	2	l-diversity	22%	57.14%

2) Selection via Hit and Trial Model

For the verification of the results obtained by EDAMS each variant of the dataset is tested with each privacy model in order to find the best suitable model for the respective dataset. The threshold values are selected from the lowest possible values to the values where change in threshold values does not affect the result. Considering the same variant, the data holder has to try each and every possible combination of different methods which demands a substantial amount of time. Table III depicts the results obtained from different combination of methods employed through hit and try model.

TABLE III. HIT AND TRIAL MODEL RESULTS

Privacy model	Threshold value	Max. information loss
k-anonymity	3	17.08%
	5	57.31%
	7	57.31%
l-diversity	3	54%
	5	61%
	7	61%
t-closeness	0.02	100%
	0.2	91%
	0.8	54%
k-anonymity and l-diversity	k=3, l=2	54%
	k=5, l=3	61%
	k=5, l=7	61%
k-anonymity with t-closeness	k=3, t=0.002	100%
	k=5, t-0.2	91%
	k=5, t=0.8	61%
k-anonymity with l-diversity and t-closeness	k=3, l=2, t=0.002	100%
	k=5, l=3, t=0.2	91%
	k=7, l=5 t=0.9	61%

It can be seen that minimum information loss occurs when k-anonymity model is applied. But the identification of this least information loss method became possible after trying each model and their combinations with different threshold values. However, the same model is recommended by the EDAMS without requiring any extra effort.

B. Case 2: Employee's Salary Dataset

This dataset [30] contains 1,999 records and comprises on five attributes (name, gender, telephone number, zip code, salary). Two variations were created, in which two attributes (name and telephone number) were considered as DIs. The first variant considers the rest of the three attributes as QIs while the second variation considers salary as SA and rest of the two, i.e. gender and zip code as QI.

1) Selection via EDAMS

The process of selection of privacy model through EDAMS will remain the same for every dataset. Considering its second variation, there are two DIs and one SA. The DIs were removed out rightly from its anonymized version while l-diversity was selected as the privacy model. Table IV shows its results.

TABLE IV. SELECTION THROUGH EDAMS

DI	QI	SA	Privacy model	Sensitivity ratio	Max. information loss
2	3	0	k-anonymity	0%	8.12%
2	2	1	l-diversity	20%	0.00%

2) Selection via Hit and Trial Model

Analyzing the same dataset yields the results shown in Table V. The two models are providing the same results, however one of them has already been suggested by EDAMS (l-diversity).

TABLE V. HIT AND TRIAL MODEL RESULTS

Privacy model	Threshold value	Max. information loss
k-anonymity	3	8.11705%
	5	8.11705%
	7	8.11705%
l-diversity	3	0.00007%
	5	0.00007%
	7	0.00007%
t-closeness	0.02	100%
	0.2	100%
	0.8	0.0023%
k-anonymity and l-diversity	k=3, l=2	0.00007%
	k=5, l=3	0.00007%
	k=5, l=7	0.00007%
k-anonymity with t-closeness	k=3, t=0.002	100%
	k=5, t-0.2	100%
	k=5, t=0.8	0.0023%
k-anonymity with l-diversity and t-closeness	k=3, l=2, t=0.002	100%
	k=5, l=3, t=0.2	100%
	k=7, l=5 t=0.02	100%

C. Case 3: Crime Incident Dataset

This dataset [31] contains a total of eight attributes, namely last name, first name, block, gender, race, date of birth, case number, and crime_code with 1,058 records. Four versions of this dataset were formed. Last name, first name, and date of birth served as DIs in the first two versions and the remaining five attributes were taken as QIs in the first variant while crime_code was taken as sensitive attribute in the second version and the rest as QIs. The third and fourth variant took only the first name and date of birth as DIs and the rest of the structure remained the same.

1) Selection via EDAMS

Analyzing its second variant there is one SA along with four QI and three DI. The sensitivity is calculated to 13% and l-diversity is suggested by EDAMS. Table VII summarizes the results of all variants.

TABLE VI. SELECTION MODEL IN CRIME DATASET FROM EDAMS

DI	QI	SA	Privacy model	Sensitivity ratio	Max. information loss
3	5	0	k-anonymity	0%	60.54%
3	4	1	l-diversity	13%	55.68%
2	6	0	k-anonymity	0%	15.26%
2	5	1	l-diversity	13%	64.54%

2) Selection via Hit and Trial Model

Table VII shows the results in finding the appropriate method for the second variant of this dataset. It is evident from this example that EDAMS chose the most appropriate model required for the respective dataset.

TABLE VII. SELECTION FROM HIT AND TRIAL MODEL

Privacy model	Threshold value	Max. information loss
l-diversity	3	55%
	5	75%
	7	75%
t-closeness	0.002	100%
	0.2	75%
	0.9	50%
k-anonymity and l-diversity	k=3, l=2	100%
	k=5, l=3	100%
	k=5, l=7	100%
k-anonymity and t-closeness	k=3, t=0.002	100%
	k=5, t=0.2	100%
	k=5, t=0.8	100%
k-anonymity with l-diversity and t-closeness	k=3, l=2, t=0.002	100%
	k=5, l=3, t=0.2	100%
	k=7, l=5, t=0.9	100%

V. DISCUSSION

The cost of producing anonymized data via hit and trial model is high as there is no standard method for anonymizing data. Data holder has to keep checking different models over different thresholds to achieve data anonymity with greater utility. Moreover, absence of knowledge regarding privacy models makes it more difficult for the data holder to modify the data into their unidentified version. However, EDAMS is capable of selecting the appropriate model for the respective dataset by applying some initial effort thereby minimizing the overall cost with good efficiency.

VI. CONCLUSION AND FUTURE WORK

Available vast data can provide immense benefits when analyzed carefully. Many companies are sharing their data for research or other purposes. However, the data are becoming highly personalized as everything becomes automated, thus the companies need to make the necessary arrangements to protect their clients' privacy. PPDP is a promising approach that can be used to publish data while preserving individual privacy to a great extent. Many techniques are available in this domain for the generation of anonymized data but choosing one is a challenging decision. This study presented the data anonymization model selection tool EDAMS that is capable of generating anonymized data with minimal effort. EDAMS requires the dataset and the nature of attributes to proceed with the selection of the optimal method among k-anonymity, l-diversity and t-closeness. The results were validated by applying the techniques separately one by one on the same datasets and the conclusion was that EDAMS efficiently selects the most appropriate method.

PPDP is still in its development stage as the researchers are coming up with more efficient algorithms. EDAMS is currently providing limited anonymization algorithm selection, however it has the capability to work as a classifier when trained rigorously. As a result, it will be capable of anonymizing any type of data by selecting the most efficient algorithm. EDAMS is dealing with linking attacks using generalization and suppression as PPDP techniques and k-anonymity, l-diversity, and t-closeness as anonymization algorithms. However in the future it is planned to accommodate more anonymization techniques to protect individual privacy against probabilistic attacks.

REFERENCES

[1] B. Marr, "How much data do we create every day? The mind-blowing stats everyone should read", available at: www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read

[2] R. Madge, "Five loopholes in the GDPR", available at: medium.com/mydata/five-loopholes-in-the-gdpr-367443c4248b

[3] J. Li, Y. Tao, X. Xiao, Preservation of proximity privacy in publishing numerical sensitive data, Chinese University of Hong Kong, 2008

[4] L. Gomes, "Data analysis is creating new business opportunities", available at: www.technologyreview.com/s/423897/data-analysis-is-creating-new-business-opportunities

[5] J. Liu, "Privacy preserving data publishing: Current status and new directions", Information Technology Journal, Vol. 11, No. 1, pp. 1–8, 2012

[6] S. Chawla, C. Dwork, F. Mcsherry, A. Smith, H. Wee, "Toward privacy in public databases", available at: www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tcc05-cdmsw.pdf, 1948

[7] B. C. M. Fung, K. Wang, R. Chen, P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys, Vol. 42, No. 4, Article ID 14, 2010

[8] A. Anjum, N. Ahmad, S. U. R. Malik, S. Zubair, B. Shahzad, "An efficient approach for publishing microdata for multiple sensitive attributes", The Journal of Supercomputing, Vol. 74, pp. 5127–5155, 2018

[9] M, Barbaro, T. Zeller Jr., "A face is exposed for AOL searcher no. 4417749", available at: www.nytimes.com/2006/08/09/technology/09aol.html

[10] D. Vatsalan, P. Christen, C. M. O 'Keefe, V. S. Verykios, "An evaluation framework for privacy-preserving record linkage", Journal of Privacy and Confidentiality, Vol. 6, No. 1, pp. 35-75, 2014

[11] K. E. Emam, S. Rodgers, B. Malin, "Anonymising and sharing individual patient data", BMJ, Vol. 350, Article ID h1139, 2015

[12] L. Sweeny, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 5, pp. 557–570, 2002

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, pp. 1-12, 2007

[14] N. Li, T. Li, S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, CERIAS Tech Report 2007-78, 2007

[15] J. Liu, K. Wang, "On optimal anonymization for l+-diversity", IEEE 26th International Conference on Data Engineering, Long Beach, USA, March 1-6, 2010

[16] L. Sweeney, Matching known patients to health records in Washington state data, Harvard University, 2013

[17] L. Sweeney, J. S. Yoo, "De-anonymizing South Korean resident registration numbers shared in prescription data", Technology Science, Article ID 2015092901, 2015

[18] Y. A. D. Montjoye, L. Radaelli, V. K. Singh, A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata", Science, Vol. 347, No. 6221, pp. 536–539, 2015

[19] A. Narayanan, V. Shmatikov, "Robust de-anonymization of large sparse datasets", IEEE Symposium on Security and Privacy, Oakland, USA, May 18-22, 2008

[20] N. Li, T. Li, S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 7, pp. 943–956, 2010

[21] P. Samarati, L. Sweeney, "Generalizing data to provide anonymity when disclosing information", 17th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, Washington, USA, June, 1998

[22] Z. E. Ouazzani, H. E. Bakkali, "A new technique ensuring privacy in big data: Variable t-closeness for sensitive numerical attributes", 3rd International Conference of Cloud Computing Technologies and Applications, Rabat, Morocco, October 24-26, 2017

[23] S. S. Devi, R. Indhumathi, "A study on privacy-preserving approaches in online social network for data publishing", in: Data Management, Analytics and Innovation, pp. 99–115, Springer, 2011

[24] H. Zhu, H. B. Liang, L. Zhao, D. Y. Peng, L. Xiong, "τ-Safe (l,k)-diversity privacy model for sequential publication with high utility", IEEE Access, Vol. 7, pp. 687–701, 2018

[25] A. S. M. T. Hasan, Q. Jiang, "A general framework for privacy preserving sequential data publishing", 31st International Conference on Advanced Information Networking and Applications Workshop, Taipei, Taiwan, March 27-29, 2017

[26] S. Hamid, N. Z. Bawany, S. Khan, "AcSIS: Authentication system based on image splicing", Engineering, Technology & Applied Science Research, Vol. 9, No. 5, pp. 4808-4812, 2019

[27] M. O. A. Dwairi, A. Y. Hendi, Z. A. AlQadi, "An efficient and highly secure technique to encrypt and decrypt color images", Engineering, Technology & Applied Science Research, Vol. 9, No. 3, pp. 4165-4168, 2019

[28] A. H. A. Omari, "Lightweight dynamic crypto algorithm for next internet generation", Engineering, Technology & Applied Science Research, Vol. 9, No. 3, pp. 4203-4208, 2019

[29] UCI, Adult data set, available at: archive.ics.uci.edu/ml/datasets/adult

[30] Employee Salary dataset, available at: www.kaggle.com/varungitboi/employee-salary-dataset

[31] Open Data Philly, Crime incidents, available at: www.opendataphilly.org/dataset/crime-incidents