Journal of Computer Science

Review

A Survey of Data Anonymization Techniques for Privacy-Preserving Mining in Bigdata

Helen Wilfred Raj and Santhi Balachandran

School of Computing, SASTRA Deemed University, Thanjavur, Tamilnadu, India

Article history
Received: 15-07-2019
Revised: 08-10-2019
Accepted: 12-12-2019

Corresponding Author:
Santhi Balachandran
School of Computing,
SASTRA Deemed University,
Thanjavur, Tamilnadu, India
Email: santhi@cse.sastra.edu

Abstract: Bigdata era is seeing the data burst occurring in a multitude of angles that are better expressed in terms of the 4Vs (Volume, Velocity, Velocity, Veracity). While trying to infer information from data, care should be exercised as not to reveal the identity of the data owner, which breaches the privacy rights. Leakage of information can happen right from the data collection point, at the data storage area, followed by the distribution of data to data users/miners and finally with published results. A cross-matching of all these points with the 4Vs (growing still) of big data, puts a huge challenge on how to extract the maximum possible information, without compromising on the privacy of the data owner. Anonymization of the original data should be done at one or more of the above-mentioned stages before the data are given for the mining process. This work makes a survey of the various anonymization techniques followed to transform the data in such a way that the privacy of the data owner is not compromised. Also, the sample data drawn should resemble and represent the original dataset in the maximum possible number of dimensions. The results of the various methodologies have been analyzed and the observations have been presented.

Keywords: Privacy-Preserving, Anonymization, Perturbation, Generalization, Dimensionality Reduction

Introduction

With the huge volume of data being generated by a variety of sources like ubiquitous hand-held devices, social networking sites, communication networks and the like, it has already been established that 90% of the present data have been generated in the past 2 years (stated by IBM). Retrieving information from such a voluminous data needs to concentrate on how to collect, store, organize, classify, categorize, identify and pull out only the relevant data, so that the appropriate information is properly identified. This will help a lot in faster processing and producing more accurate and fine tuned results moving close to our objectives.

In order to achieve this, the data has to be collected, stored, cleaned and then processed. Collecting all the data first and then to process, does not make good sense, due to the very nature of the speed at which it is generated, which entails that the data collection will never end and hence one has to be satisfied with a subset of the dataset that will well represent the entire dataset in all its characteristics (Velocity nature). Providing a centralized storage for the data again poses a problem (owing to its Volume nature). Trying to arrange all the data with a common approach is a big issue due to the variety of the data (Variety nature). Producing the results with as much accuracy as possible is a tedious task, due to the fast evolving nature of the domain and new techniques are evolving daily (Veracity nature).

Phases in Data Manipulation

Ensuring the privacy of the data at all the three phases namely - data collection, data storage and data processing has to be ensured. The process of transforming the data can be done at several places starting from the data gathering till the stage of information collection. Data transformation is achieved by anonymizing the data by means of generalization, specialization, suppression, perturbation and similar other techniques. Data mining models like k-anonymity, l-diversity, t-closeness etc., are applied on the data set that complements the above mentioned techniques.

While attempting to protect the privacy, the very aim of information disclosure also have to be borne in mind that the mining techniques applied on the transformed dataset does disclose a reasonable amount of information as well. A strict implementation of privacy preserving strategies will rank high in terms of privacy protection.

But at the same time, if the amount of information gained after the application of data mining algorithms is very less, then it is not worth the effort. On the reverse, with the greed to gain the maximum information possible, if the data protection efforts are relaxed, then it will become easy for an intruder to infer the private information pertaining to the data owners. The survey made here presents some of the works going on in this field to protect the privacy of the data owner. Several works done on the above perspective are discussed, with the restriction of the domain on which they are designed to work. An analysis of the results is presented at the end with some suggestions as how they can be enhanced further.

Data Sets and Sources

This survey on the impact of data mining techniques on the privacy of data, involves both synthetic and real datasets. Datasets containing numerical, categorical, binary attributes and of high dimensional nature have been involved in the study. The statistics of the datasets used in this analysis have been tabulated in Table 1.

Privacy Preserving Mining Approaches

FRamework for Accuracy in Privacy-Preserving (FRAPP), is a framework designed for randomized perturbation based privacy preserving mining (Agrawal and Haritsa, 2005). The perturbation parameters themselves are randomized here. This increases the degree of anonymity achieved to a greater extent, but at the same time reduces the possibility of reconstructing the model. Focus is on categorical attributes where the domain values are limited. It has been proved that not only the choice of perturbation matrix, but the dataset size also has a considerable impact on the accuracy achieved through the mining model.

The framework follows a data model and a perturbation model. For data model, consider a database D containing N records. Each record has M categorical attributes. Each attribute has a domain of values and all such domain values of all the categorical attributes are represented together for each record. For example, consider that D has the set of categorical attributes as given in Table 1.

Mapping is done across the various values of each attribute and an index is assigned with each such unique combination of values, as shown in Table 2.

The distribution of values is not private but that does not pose any threat to the information in the records to be disclosed to other data users.

The perturbation model chooses a randomization operator and perturbs the indices by applying the operator over the data model values. Proof of the privacy guarantees according to the perturbation level has been provided by Agrawal and Haritsa (2005).

The approach tends to preserve the probability of the private information of a data owner before and after the perturbation. There is no undue difference between the prior and post information of perturbation. The perturbation matrix is randomized, which further increases the privacy of the data owner. FRAPP framework finds further utility in association rule mining in exploring interesting associations among the database attributes.

A Tree Based Approach

A tree based approach was developed which focuses on data owned by an individual organization and hence does not work for distributed databases (Li and Sarkar, 2006). It is meant for numeric data alone. A sample of the dataset is shown in Table 3. The kd-tree structure used in this approach makes use of a recursively partitioning approach. At every point of partition, the attribute with the maximum variance is selected. Partition the dataset based on the median or mid-range of the values. After every partition, the records within the new subsets move closer to each other. The process stops when the subset size drops below a user defined minimum threshold. The records of the resulting dataset are perturbed by replacing the confidential with their average. For multiple attributes, the values corresponding to each attribute is averaged out, which is used to replace the original values.

The approach uses an approximation about the conditional expectation of the confidential attributes in the subset of the dataset. Using the conditional expectation ensures the preservation of the relationships between the confidential attributes and its counterpart.

The data thus generated by building the perturbation trees possess the following properties-their mean always equals that of the initial dataset and its variance will be lesser than that of the original dataset. By this approach, all confidential attributes and a subset of the non-confidential attributes can be perturbed. The limitation arises while trying to balance between the weightage given for confidential and non-confidential attributes and the corresponding amount of risk involved in disclosing and the loss of information.

Utility Based Anonymization

Another approach that decides the level of anonymization based on the utility of the data, which works both for numerical and categorical data was developed by (Xu *et al*., 2006). Generalization is the technique adopted here to anonymize the data making use of the range of values, but using median and mean also gives the same results. The utility based metric used here considers both-amount of information lost and the importance of the attribute. Numeric attributes are associated with weighted certainty penalty and categorical attributes with normalized certainty penalty, which is

arrived at based on the hierarchy of values identified among the categorical attributes. The anonymization process follows two greedy methods-bottom-up search and top down mode. The bottom-up mode is a local recoding process, which does only searching and no splitting is done. Top down approach follows binary partitioning.

In the dataset shown in Table 4, tuples a, b and c correspond to that of a quasi-identifier A and tuples d, e and f to that of B. By calculating discernability metric and normalized average a and b are generalized into ([10,20], [60,65]) group, c and d are generalized into ([20,20], [45,50]) group and e and f are generalized into ([10,10], [50,55]) group. The uncertainty level of all the tuples are summed up which gives an idea about the amount of information loss. Obviously, lesser the uncertainty level, the lesser will be the information loss also. Normalized certainty penalty is calculated for categorical attributes to estimate the amount of information loss. It was proved that as finding the optimal solution for utility based anonymization is a NP-hard problem (Xu *et al*., 2006).

The quality of anonymization achieved by both these methods comes with a high cost in computation time. The top-down method is much faster than the bottom-up mode. This is due to a heuristic followed in top down method to make the split. The pair of tuples with the highest certainty penalty is chosen and hence this heuristic gives good approximation of the maximum value. More often, in anonymization, computational intensity always takes the priority, moving the quality to back seat.

Anonymizing Classification Data

An approach for privacy preservation using classification technique was coined by (Benjamin *et al*., 2007). Raw data generally contains a lot of noise information and unused redundant structures. Such structures have been made use of here to hide the data effectively without having to compromise on the degree of classification achieved. A Top-Down Approach (TDR) is followed here, starting from the top of the masked table and drill down refining the values, further. The method can be applied for both categorical (both with and without taxonomy) and continuous attributes. Instead of concentrating on a single quasi-identifier with all attributes, multiple quasi identifiers are used that has the obvious advantage of avoiding unnecessary distortion to data. For the top-down refinement of the table, generalization, suppression and discretization are applied over the table. Also, the process can be stopped at any intermediate level of the tree, once it is felt that the desired level of anonymization has been achieved. The refinement process results in a rise in the amount of information gained and a proportionate loss in anonymity also.

The work has been proven to be efficient, as it applies the two types of works in all iterations. The first type of work accesses the records by making a best split of the records. Making a split requires that the records be in sorted order. Due to the top-down refinement process, since the records are already sorted, the desired record is accessed in a single scan. The second type of work computes the score of the record which is done without accessing the data records. This is made possible because the count statistics are already maintained while the records were scanned in the first stage.

Table 5 illustrates how to create the compressed table. It emphasizes the significance of data generalization by way of compressing the data. Table 6 and the tree shown in Fig. 1 explain the various means of compressing the data by combining the attributes. Anonymization can be achieved via generalization and generating the taxonomy tree over the generalized table.

A sample taxonomy tree based on qualification attribute is shown below in Fig. 1. Here, all at school level can be divided into two groups-Junior School and Senior School. Both UG and PG may be combined together as college level. A further level generalization up the tree can be done as School and University category.

Table 1: Data set statistics

Reference	Source	Dimensions	Data Points
Aggarwal *et al*. (2013)	Synthetic datasets	100	10,000
	Arrythmia data set(UCI Machine learning repository)	279	50,000
Benjamin *et al*. (2007)	Adult data set from UCI repository	six continuous attributes, eight categorical attributes	30,000+records-training set; 15,000+records-testing
Benjamin *et al*. (2007)	CRX data set	six continuous attributes, nine categorical attributes and a binary class attribute	465 and 188 records for the presplit training and testing, respectively
Fouad *et al*. (2014)	Item description table of Wal-Mart database	30 attributes	400,000 records
Xu *et al*. (2006)	Adults census data set from the UC Irvine machine learning Repository	15 attributes	30, 162 tuples
Wong *et al*. (2010)	CENSUS dataset	15 attributes	32561individuals

Table 2: Attribute distribution for Data model

Attribute	Domain values
Category	Under 18, 18 - 40, Above 40
Experience	Novice, skilled, highly skilled
Education	Primary, Secondary, Graduate, Post graduate

Table 3: Attribute indexing done

	Attribute values		Index
Under 18	Novice	Primary	1
Under 18	Skilled	Secondary	5
Above18	Skilled	secondary	9

Table 4: Example dataset

S No	Age	Experience	Actual income	Perturbed income
1	28	5	46	43
2	37	10	54	57
3	41	14	57	53
4	49	18	62	66
5	30	8	50	54

Table 5: Sample dataset

A	(20,65)
B	(10,60)
C	(20,45)
D	(50,20)
E	(55,10)
F	(50,10)

Table 6: Compressed table

Qualification	Gender	Hours worked	Category	No. of recs
Std X	F	43	0Y4N	4
Std XII	M	45	1Y4N	5
Std XI	M	56	3Y2N	5
U G	M	39	5Y3N	8
U G	F	44	4Y2N	6
U G	F	51	2Y5N	7
P G	F	35	1Y0N	1
P G	M	47	0Y5N	5
PhD	M	60	1Y1N	2
PhD	F	55	0Y2N	2
Total			17Y28N	45

Regarding memory requirement, the approach assumes that the table is compressed and hence it will fit in main memory. If it doesn't fit, then the leaves are kept in memory and other partitions moved to disk. The size of the leaf is decided in such a manner that it will not lead to fragmentation and hence it enables a partition to be fetched in a single access. TDR is much more efficient in handling the continuous attributes because it requires only less number of record splitting. But bottom up approach needs several merge operations. Also, in TDR process, the records that cannot be further refined can be dropped. But bottom up process has to necessarily carry such records till the end.

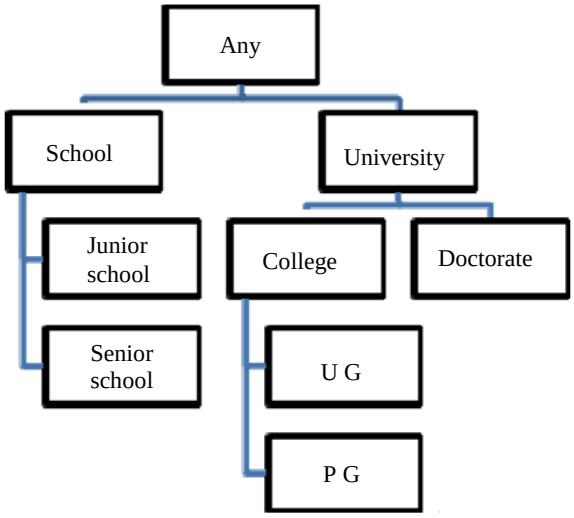

Fig. 1: Taxonomy tree for the compressed table

Original table		Anonymized table	
Tuple ID	QID	Tuple ID	QID
t_1	1	t_1'	1-5
t_2	2	t_2'	2-3
t_3	3	t_3'	2-3
t_4	5	t_4'	1-5
t_5	4	t_5'	3-4

Fig. 2: Original and anonymized tables

Non-Homogeneous Generalization

A technique that generalizes partitions in a non-homogenous manner and still ensures that they satisfy k-anonymity was proposed by (Wong *et al*., 2010). This can further be enhanced by combining with any existing partitioning based approach to improve its usage further.

The table in Fig. 2 illustrates the non-homogenous generalization being applied to the data tuples. However, the process does not satisfy 2-anonymity. The tuple t_5 still remains being capable of getting identified. A generalization is applied in the process of randomization in order to achieve the desired k-anonymity.

A problem associated with the non-homogenous generalization is that, if the algorithm adopted is deterministic, then it tends to produce the same set of values for the data tuples in every run. Hence, if any adversary knows the deterministic algorithm, he can apply that on the data set and with repeated runs, it is possible to get the quasi identifier attribute values and hence the anonymity gets violated. In order to address this problem, a randomization is applied along with the

generalization. The generalization procedure is accomplished in two steps-QID is generated with a generalized approach and the generation of random assignment. Homogenous generalizations are not enough, as the results produced by them are not of good quality.

The framework does the partitioning of tuples, generalizes the partition with the QID of each tuple and with random assignment the tuples are named with generalized QIDs. Choosing a good strategy for partitioning plays a major role in non-homogenous generalization.

Dimension Based Techniques

Dimension of data plays a crucial role in deciding the scalability, efficiency and the ability to sustain the privacy level of the data mining algorithms. An increase in dimension will have an immediate downward effect on the above said factors.

Multidimensional Suppression

kACTUS algorithm anonymizes the dataset and ensures the privacy without affecting the mining results (Kisilevich *et al*., 2010). K-anonymity is used as the standard to test, how far the algorithm adheres to the privacy constraint. The algorithm follows a wrap-around formula wherein, the classification is done by a decision tree inducer. The obvious advantage lies in adopting the existing, standard and proven practice, which protects the accuracy of the results without much of deviation.

It produces an anonymized data set as output which is given the data miners for information retrieval. kACTUS applies multidimensional suppression to produce the anonymized output. The k-anonymity algorithm transforms the original dataset into a k-anonymity compliant transformed dataset. The algorithm do not necessitate the generation of domain generalization taxonomy for categorical attributes which needs some amount of initial knowledge about the domain. For the sake of ensuring the k-anonymity, kACTUS assumes a univariate classification tree and all the internal nodes refer to quasi-identifier attributes.

Adding more attributes to the internal nodes can still increase the accuracy of classification. But at same time, it can induce over anonymity as well. In order to tackle this, generalization may be applied in place of suppression. Selecting the data tuples is randomized here. Instead, a heuristic approach may be tried for where it gives an opportunity to choose the domain of tuples. The heuristics applied may again be changed according to the current subset of dataset at hand.

High-Dimensional Randomization

Applying randomization at the time of data collection does neither give a measure of the level of anonymization applied, nor does it allow having control over it (Aggarwal, 2013). The log likelihood fit is used to estimate the probability that a public database record corresponds to a particular perturbed data. So, larger perturbations favor the log likelihood and thus increase the chance of similar record being identified in the same dataset, which ultimately enhances the privacy of records. Regarding the distribution of data, both Gaussian and uniform distribution both have their effect on the level of randomization achieved. With huge data sets, even after perturbing the data, the randomization method will be able to reconstruct the original distribution.

The results shown by the implementation of the above algorithm shows that the revised classification tree obtained after the successive iterations makes the result set move towards the desired level of anonymity. Comparisons of the result in terms of accuracy have been made with suppression based anonymization and anonymization with multidimensional dataset.

The Role of Dimensionality

An increase in the dimensionality of the dataset immensely affects the effectiveness of any privacy preserving attempt. It has been proved that the level of randomization expected reduces with a rise in dimensionality for a given level of perturbation. Taking proportionality into consideration, the perturbing distribution should be in linear relationship with the increase in dimensionality. Clusters help in increasing the level of randomization. Data sets with oscillating density distribution tend to have a lower level of worst-case randomization than the average randomization level. Presence of outliers too, plays a similar role with respect to the randomization level. Even with the well established perturbation functions, a rise in dimensionality affects the randomization levels.

A Hybrid Scalable Approach

A highly scalable approach, where the anonymization of sub tree is done by choosing either Top-Down Specialization (TDS) or Bottom Up Generalization (BUG), was framed by (Zhang *et al*., 2014). The approach is hybrid in that it uses a combination of generalization and specialization, but one of the two chosen at a time. Decision of which technique to be chosen is taken based on the nature of the data set then. A threshold is derived from the characteristics of the current subset of data in hand and the amount of workload. TDS or BUG is chosen based on the tradeoff between information gain and privacy loss. Quasi identifier attributes are chosen in the process, whose sensitivity lie in between that of confidential and non confidential attributes.

A large variance in the distribution of the QI attributes implies the odd distribution of the records among the values domain. The coefficient of variation is calculated for the attribute set and the adjustment is done in the k-anonymity. But, the problem arises when multiple iterations are carried out for generalizing, which increases the amount of computation to be done. While the adjustment of k-anonymity is done to modify the work load balancing point, its reverse effect on privacy gain and information loss was not probed into.

Privacy-Preserving Computing for Big Data

Several privacy preserving techniques play a major role in protecting the sensitive data from being disclosed to the data miners (Lu *et al*., 2014). Some of them are privacy-preserving aggregation, de-identification techniques and applying the operations on encrypted data. Privacy-preserving aggregation plays a major role during the collection and storage of big data. But it is inflexible, because data collected under one purpose may not be useful for another. Operating over encrypted data though complex, but they are safe. Applying the same on big data becomes a tedious process owing to its complexity and the volume of data to be handled.

De-identification first generalizes the data and then applies suppression techniques over it, before it is released for the mining activity. But the anonymity of the data even after applying this two-step process was not up to the expected level. When applied to big data, with the bulk of data available, it becomes easier to re-identify the data. Privacy Preserving Cosine Similarity (PCSC) protocol has been developed which combines lightweight multiparty random masking and polynomial aggregation techniques. PCSC does not consume much time since it does not need exponential operations.

Differential Privacy Preserving Algorithm

Differential privacy emphasizes that a change in the input data should not make any significant difference in the distribution of the outcome (Fouad *et al*., 2014). The data generalization model proposed here shows that the optimization of the objective can be done in polynomial time, by loading the objective with the constraint on utility. The approximation algorithm proposed here produces a data transformation with optimal constant guarantees. Also, a modified version of ARUBA algorithm produces a data transformation algorithm of polynomial time.

Geng *et al*. (2015) proposed an alternative noise distribution that can replace Laplacian noise in each instance in the literature and for the same privacy level add lesser amount of noise and for the same level of differential privacy, the performance in each instance improves. In the work, it is shown that staircase mechanisms are extremal points of the (convex) space of differentially private mechanisms and optimality of a large class of utility maximization problems is achieved by one of these staircase mechanisms.

Kairouz *et al*. (2016) introduced a combinatorial family of extremal privatization mechanisms, called staircase mechanisms and showed that it contains the optimal privatization mechanisms for a broad class of information theoretic utilities such as mutual information and f-divergences. They introduced binary and randomized response mechanisms, privatization mechanism and staircase mechanism.

Comparison of the Techniques

A comparative study of the above analyzed techniques has been tabulated below in Fig. 3 as quick reference. With respect to the memory needs, all the techniques show a positive result as not to load the memory much in most of the cases. Classification techniques make use of the decision tree method. Usage of attributes is a major factor that always has a major impact on the quality of results. Quasi-identifier attributes always tend to be more sensitive and critical, in that, it becomes easier to extrapolate this information on public data to easily infer the original information about the data owner. Being applied for big data scenario, scalability issue has to be analyzed thoroughly in order for the results to be capable of being applied to any data. The study also has concentrated on the computational intensive part of the procedure that which needs to be carefully designed and that which also has scope for improvement as well. Finally, the scope of extending the work further also has been highlighted at the end.

Analysis and Observation

The multi-distortion method was proposed which tries to make up for the lack of data where the data collection possibilities are limited with the help of distorted data (Agrawal and Haritsa, 2005). By collecting distorted versions of data to compensate for lack of data, no compromise is done on the part of security. But, its effect on the information gain should be compared with that achieved without any distortion of data. With regard to the time taken to complete the mining process, the algorithm performs the same on both the original and perturbed databases as when compared to the standard Apriori algorithm. Adding some noise to the leaf level data can compensate for the negative variance (Li and Sarkar, 2006). But, when the degree of dissimilarity between the data points is high, then the dataset has to be divided into chunks using the perturbation trees and then the noise addition should be tried on.

Type of Anonymization Applied	Memory requirement	Strategy followed	Type of attribute	Scalability	Computational intensity	Scope for extension
Generalization and Specialization	Good scaling with MapReduce	Automated choosing of BUG or TDS	Quasi-identifier attributes	Tree construction becomes tedious	BUG - more intensive	Impact of l-diversity can be studied be studied
Suppression	Based on levels in the induction tree	Decision tree based k-anonymity	Quasi identifiers used	Scales well for large datasets		Multivariates can be considered
Perturbation	Negligible overhead	Perturbation matrix constructed	Categorical converted to continuous valued	Scales well with large datasets	No additional burden	Mining the distorted DB directly
Randomization	Increases with raising dimension	Uniform and Gaussian distribution	All types of attributes	Outliers affect randomization at worst case	Affected by dimensionality factor	Using sample datasets for better results

Fig. 3: Comparison table showing the results of the analysis

The distribution of dataset plays a significant role in determining the quality of anonymization achieved through the approach developed by Xu *et al*. (2006). With respect to performance, both the top-down and bottom-up approaches give the similar results, without much of deviation among themselves. Similarly, both the methods consume more time as when compared to the MultiDim method. The non-homogenous generalization, (Wong *et al*., 2010), concludes that the method can be further extended to l-diversity as well. While it is extended to l-diversity, if the dataset is randomized, a thorough analysis should be done on whether the process will lead to over anonymity, thus reducing the information gain to the minimum. With k-anonymity, the results are acceptable.

kACTUS algorithm was tested on the standard dataset, for which a generalized taxonomy is already available (Kisilevich *et al*., 2010). But if suppression has to be applied for anonymization, then the existing taxonomy cannot be applied. While the approach scales well with generalization to achieve the desired k-anonymity, the same should be proved if suppression is followed. The analytical results produced by (Aggarwal, 2013) give a wide opening for the algorithmic implementation of the randomization method and sets a platform to compare the theoretical and the experimental results. The dimensionality factor obviously will have an impact on the level of privacy achieved, the relationship being inversely proportional. The outliers will have a significant role to play in case of randomization, which should be proved by implementing the analytical approach suggested here.

Conclusion

The study has taken into consideration the various techniques used for anonymization of data for data privacy. The comparative table clearly highlights the scope available for the researchers to carry on the work further.

With the voluminous data to be manipulated to explore the information inside it, the appropriate data mining techniques need to be applied. Privacy of the data owner has to be preserved during all the phases by applying the various anonymization techniques as applicable, which depends on a lot of factors. The level of anonymization also has to be analyzed such that the anonymized data is still capable of giving the maximum information possible and at the same time protecting the privacy of the data owner.

Acknowledgement

The authors would like to record their gratitude for SASTRA University for providing the Data Science lab to conduct the experiments.

Author's Contributions

Both the authors have equally contributed in the review and analysis of the research articles.

Ethics

This article is an original and contains unpublished materials. All authors have read and approved this manuscript and no ethical issues involved.

References

Aggarwal, C.C., 2013. On the analytical properties of high-dimensional randomization. IEEE Tran. Knowl. Data Eng., 25: 1628-1642. DOI: 10.1109/TKDE.2012.98

Agrawal, S. and J.R. Haritsa, 2005. A framework for high-accuracy privacy-preserving mining. Proceedings of the 21st International Conference on Data Engineering, (CDE' 05), Apr. 5-8, IEEE Xplore Press, Tokoyo, Japan, pp: 193-204. DOI: 10.1109/ICDE.2005.8

Benjamin, C.M.F., K. Wang and P.S. Yu, 2007. Anonymizing classification data for privacy preservation. IEEE Tran. Knowl. Data Eng., 19: 711-725. DOI: 10.1109/TKDE.2007.1015

Fouad, M.R., K. Elbassioni and E. Bertino, 2014. A supermodularity-based differential privacy preserving algorithm for data anonymization. IEEE Tran. Knowl. Data Eng., 26: 1591-1601. DOI: 10.1109/TKDE.2013.107

Geng, Q., P. Kairouz, S. Oh and P. Viswanath, 2015. The staircase mechanism in differential privacy. IEEE J. Selected Topics Signal Process., 9: 1176-1184. DOI: 10.1109/JSTSP.2015.2425831

Kairouz, P., S. Oh and P. Viswanath, 2016. Extremal mechanisms for local differential privacy. J. Mach. Learn. Res., 17: 492-542.

Kisilevich, S., L. Rokach, Y. Elovici and B. Shapira, 2010. Efficient multidimensional suppression for k-anonymity. IEEE Tran. Knowl. Data Eng., 22: 334-347. DOI: 10.1109/TKDE.2009.91

Li, X.B. and S. Sarkar, 2006. A tree-based data perturbation approach for privacy-preserving data mining. IEEE Tran. Knowl. Data Eng., 18: 1278-1283. DOI: 10.1109/TKDE.2006.136

Lu, R., H. Zhu, X. Liu, J.K. Liu and J. Shao, 2014. Computing in big data era. IEEE Netw., 28: 46-50. DOI: 10.1109/MNET.2014.6863131

Wong, W.K., N. Mamoulis and D.W.L. Cheung, 2010. Non-homogeneous generalization in privacy preserving data publishing. Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 06-10, ACM, New York, USA, pp: 747-758. DOI: 10.1145/1807167.1807248

Xu, J., J. Pei, X. Wang, B. Shi and A.W.C. Fu, 2006. Utility-based anonymization for privacy preservation with less information loss. ACM SIGKDD Explorat. Newsletter, 8: 21-3. DOI: 10.1145/1233321.1233324

Zhang, X., L.T. Yang, C. Liu and J. Chen, 2014. A scalable two-phase topdown specialization approach for data anonymization using mapreduce on cloud. IEEE Tran. Parallel Distributed Syst., 25: 363-373. DOI: 10.1109/TPDS.2013.48