International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.253-262
http://dx.doi.org/10.14257/ijdta.2015.8.6.23

# Bigdata Anonymization Using One Dimensional and Multidimensional Map Reduce Framework on Cloud

Shalin Eliabeth S and Sarju S

*Department of Computer Science and Engineering*
*SJCET Palai, Kerala, India.*
*shalinelizabeth9@gmail.com, sarju.s@sjcetpalai.ac.in*

***Abstract***

*Data privacy preservation is one of the most disturbed issues on the current industry. Data privacy issues need to be addressed urgently before data sets are shared on cloud. Data anonymization refers to as hiding complex data for owners of data records. In this paper investigate the problem of big data anonymization for privacy preservation from the perspectives of scalability and time factor etc. At present, the scale of data in many cloud applications increases tremendously in accordance with the big data trend. Here propose a scalable Two Phase Top-Down Specialization (TPTDS) approach to anonymize large-scale data sets using the MapReduce framework on cloud. For the data anonymization-45,222 records of adults information with 15 attribute values was taken as the input big data. With the help of multidimensional anonymization on map reducing framework, here implemented the proposed Two-Phase Top-Down Specialization anonymization algorithm on hadoop will increases the efficiency of the big data processing system. In both phases of the approach, deliberately design multidientional MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way. Data sets are generalized in a top-down manner and the better result was shown in multidmientional MapReduce framework by compairing the onedimentional MapReduce framework anonymization job. The anonymization was performed with specialization operation on the taxonomy tree. The experiment demonstrates that the solutions can significantly improve the scalability and efficiency of big data privacy preservation compared to existing approaches. This work has great applications to both public and private sectors that share information to the society.*

***Keywords****: Data Anonymization, Big Data, Cloud Computing, MapReduce Privacy Preservation, Top Down Specialization*

## 1. Introduction

This paper is part of this project which sprecifically works on preserving the privacy of increasing personal information. The disclosure of personal information without using proper means of hiding informations could lead to misuse/abuse of personal information and be used by third parties.To avoid the disclosure of personal informations unique personal identifiers like personal numbers, social security number or any other unique numbers can easily be deleted from datasets before releasing them publicly. Personal data can be protected by using cryptographic algorithms to hide them from adversaries. Researchers and analysts require data which is consistent and coherent; encrypting these data with cryptographic algorithms will not give them the data with their completeness/truthfulness. This data need to be properly studied on how much information could be discovered by linking this data with other publicly available informations. Though there are different kinds of anonymization methods, their central

ISSN: 2005-4270 IJDTA

objective is to protect the De identifyability of individuals from datasets, and keep the data utility for further studies.

Big data is a large amount of information refers to data collection in applications have been growing tremendously and complicatedly so that traditional data processing tools are incapable of handling the data processing pipeline including collection, storage, processing, mining, sharing, *etc.*, within a tolerable elapsed time. Nowadays, many companies and organizations have been collecting huge amount of data containing various personal information via their products or services such as social network websites, online healthcare services and location-based services. There are three research challenges of privacy -preserving big data publishing in cloud computing, from perspectives of scalability, monetary cost and compatibility.

At present, parallel and distributed paradigms like MapReduce are widely adopted in big data processing applications. Accordingly, an increasing number of data mining or analytical tools and platforms are built on top of MapReduce, e.g., scalable machine learning library Apache Mahout. However, none of traditional anonymisation has been built on such a paradigm, while the published or shared data are usually consumed by big platforms or tools mentioned above. As a result, traditional anonymisation approaches lack the compatibility to be integrated with the state-of-the-art big data mining or analytical tools and platforms seamlessly. These organizations use personal informations to process the number of customers using the smart home application. This thesis paper specifically studies on how to anonymize the data that is securely collected by conducting a survey. A number of different privacy preserving algorithms have been invented to solve the issues of linking quasi-identifiers that uniquely identifies customers. But the first issues with such kind of customer vs. service provider issues are to secure the network between the two parties, storing and processing them. The previous way of solving linking problems does not comply with the real time requirement of smart home applications.

## 2. Related Work

In Bigdata applications, the privacy preservation for data analysis, share and processing is a challenging research issue due to increasingly larger volumes of data sets, thereby requiring intensive investigation. A wide verity of privacy models and anonymization approaches have been put forth to preserve the privacy sensitive informational data sets. Data privacy is one of the most concerned issues because processing large-scale privacy-sensitive data sets for big data applications.

### 2.1. Data Anonymization Using One-Dimensional Mapreduce Framework

In this paper [1] the proposed data anonymization algorithm was implemented using one dimensional mapreduce framework. If the dataset is so high, then the proposed anonymization algorithm does not work with the one dimensional mapreduce operation. A large number of cloud services require users to share private data like electronic health records for data analysis or mining, bringing privacy concerns. In this paper, propose a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using a single MapReduce framework on cloud. In both phases of our approach, a single MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way.

### 2.2. Datafly Algorithm for the Data Anonymization

The Datafly algorithm [4] is one of the first practical applications of K-anonymity. The algorithm approaches the K-anonymity by using full-domain generalization until every combination of the quasi-identifier is at least K times. The Datafly algorithm performs its generalization and suppression steps to make data ready for release. The generalizing is a

recursive task until the required level of k or less tuples have distinct sequences in frequency. Step 3 is to suppress those tuples who have a frequency of less than k. The final step is to construct the table based on the number of occurences. The domain generalization hierarchy is done on the attribute of the birth date.

### 2.3. Mondrian Algorithms for the Data Anonymization

Mondrian [2] is a multidimensional k-anonymity algorithm that anonymizes data through recursively partitioning using the quasi-identifier attribute dimensions with a median-partition methodology. This model is very fast, scalable and it produces better results than most other recoding models [3]. The Mondrian algorithm uses strict partitioning and relaxed partitioning methods instead of domain generalization or value generalization methods. If we have d quasi-identifiers, then we need a the dimension representation of Mondrian. There are two types of partitioning algorithms in Mondrian, the relaxed and the strict partitioning.

## 3. Methodology

Two-phase Top-Down Specialization (TPTDS)
There are 3 components present in the TPTDS approach, *i.e*.
1) Data partition,
2) Anonymization level merging
3) Data specialization

### 3.1. Sketch of Two-Phase Top-Down Specialization

The TPTDS method to conduct the computation which are required in TDS in a highly scalable and efficient way. Generally Map Reduce on the cloud has two levels of parallelization IE, job level and task level [4]. for example, the Amazon Elastic Map-Reduce service [5]. Task level parallelization is refers to that multiple mapper/reducer tasks in a Map-Reduce job is executed concurrently over data splits [6]. To obtain finally consistent anonymous data sets, the second phase is important to integrate the intermediate results [7] and further anonymize entire data sets. The subroutine is a Map Reduce edition of centralized TDS (MRTDS) [7] which concretely conducts the computation is essential in TPTDS.

### 3.2 Data Partition

In the Data is partitioned [8], Data cut into number of pieces required that the distribution of data records in DI is similar to D. A data record here can be treated as a point in a m-dimensional space, where m is the number of attributes. Random sampling technique is adapted to partition. The number of Reducers should be equal to ‘p’, so that each Reducer handles one value of the brand, exactly producing p resultant files. Each file contains a random sample of D.

### 3.3 Anonymization Level Merging

All middle anonymization levels merge into one in the second phase. The merging of anonymization levels [9] is completed by merging cuts. For the case of multiple anonymization levels [10], can merge them in the same way by iteratively fashion.

### 3.4 Data Specialization

An original data set D is concretely specialized for anonymization [11] in a one-iteration in Map Reduce job. When I obtain the merged intermediate anonymization level AL*,

run MRTDS Driver (D, k, AL*) on the entire data set D, and get the final anonymization level AL*. Then Reduce function simply

The IGPL Update job dominates the efficiency and the scalability of MRTDS [13], when it is executed iteratively as given in this method Thus, Hadoop variations [14] like Hadoop and Twist have been proposed recently to support efficient iterative Map Reduce computation.

Input: Data set *D*, anonymity parameters *k, k'* and the number of partitions *p*.

Output: Anonymous data set *D**.

1. Partition *D* into *Di,* $1 \leq i \leq p$.
2. Execute MRTDS *(Di, k', AL )* ->*ALi*, $1 \leq i \leq p$ in parallel as multiple MapReduce jobs.
3. Merge all intermediate Anonymization levels into one, merge $(AL_1, AL_2, \ldots AL_p)$ -> $AL^I$.
4. Execute MRTDS $(D, k, AL^I)$ -> *AL** to achieve k-anonymity.
5. Specialize *D* according to *AL**, Output *D**

[1] The value of AL varies in Driver according to the output of the IGPL Initialization or IGPL Update jobs.

Hadoop [15] used as an open-source implementation of MapReduce. The distributed cache mechanism is use to pass the content of AL to each Mapper or Reducer node as shown in Figure 1. MD5[17] (Message Digest Algorithm) is employed to compress the records transmitted for anonymity.

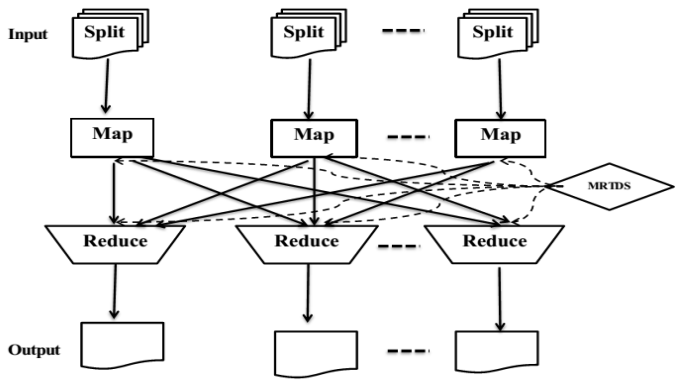



**Figure 1. The Working of MRTDS Anonymization Algorithm on One Dimensional MapReducing Framework**

### 3.2. Mapreduce with Multidimensional Anonymization

Here propose a scalable MapReduce based approach for multidimensional anonymisation over big data sets [18] can see in Figure 2. In this paper, propose a scalable MapReduce based approach for multidimensional anonymisation over big data sets [19].My basic and intuitive idea to partition a large data set recursively into smaller data partitions using MapReduce until all partitions can fit in the memory of a computation

node. Multidimensional anonimyzation with three mapreduce framework was proposed with seed initialization and seed updation algorithm [18]. But in this paper proposes themultidimensional anonimyzation with three mapreduce framework on the MRTDS anonimyzation algorithm is shown in Figure 2.

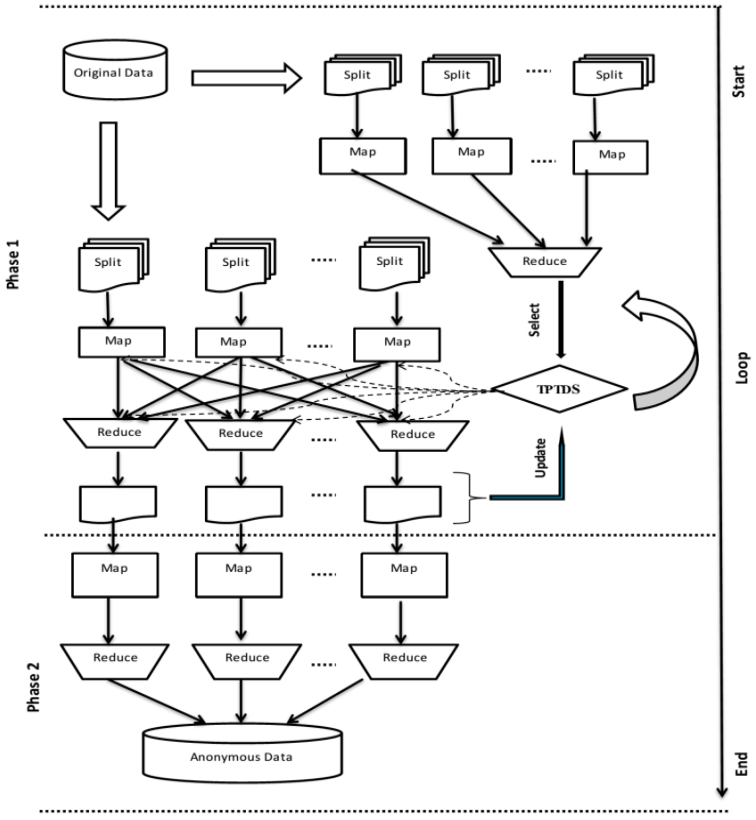


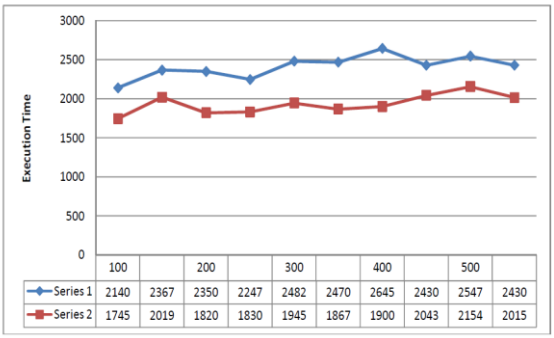

**Figure 2. The Working of MRTDS Anonymization Algorithm on Multidimensional MapReducing Framework**

## 4. Results and Evaluations

In this section evaluates both the anonymization and information loss models. I have used adult dataset to show how we can achieve the required perturbed data and quantify the information loss. Unlike the existing papers of K-anonymity. Here have concluded that TPTDS multidimensional anonymization method results are better in both utility and information loss. Thus, all results presented in this thesis paper are a result of the TPTDS multidimensional algorithm by comparing the one dimensional TPTDS anonymization. By comparing the base paper the proposed TPTDS anonymization algorithm was implemented in multidimensional mapreduce framework. But in case of base paper the TPTDS algorithm was implemented in one dimensional map reduce framework. The values are based on,

- Size of the dataset in MB
- Number of data partitioning
- Execution time
- Anonymity parameter k value
- IGPL values

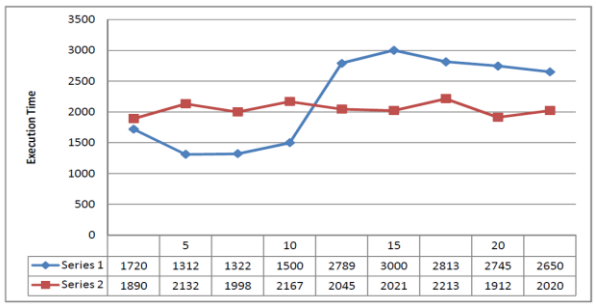
These parameters are plotted in X and Y axis and compared the values in the base paper, By comparing the values in the graph, obtained the performance evaluation in one and multidimensional map reduce framework for the TPTDS anonymization.




Graph 1

X-Axis: Size of the dataset in MB and the Y-Axis: Execution time in seconds

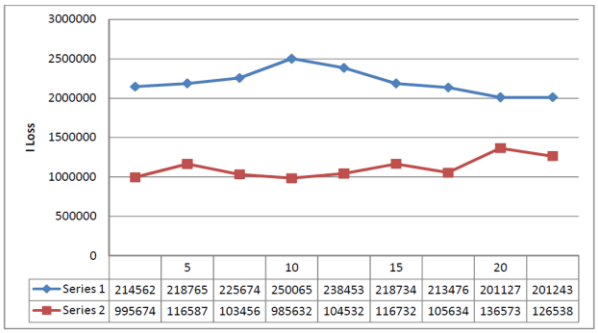
The topmost line series 1 shown in the graph indicates the TPTDS algorithm execution with one dimensional map reduce framework (In base paper). Bottom line series 2 indicates the TPTDS algorithm execution with multi dimensional map reduce framework (In proposed work). By comparing the two lines series in the 1st graph, the performance is better in multidimensional anonymization in series 2 by decreasing the execution time.




Graph 2

X-Axis: Execution time in seconds and the Y-Axis: Number of data partitions.

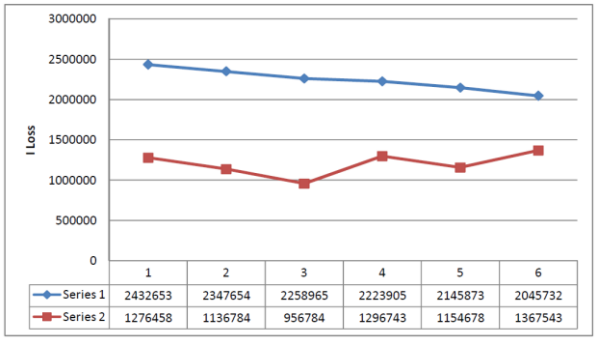
The line series 1 on the graph 2 indicates the TPTDS algorithm execution with one dimensional map reduce framework (In base paper). The line series 2 indicates the TPTDS algorithm execution with multidimensional map reduce framework (In proposed work). By comparing the two lines series in the 2nd graph, the performance is better in multidimensional anonymization in the line series 2 by decreasing the execution time.




Graph 3

X-Axis: Number of data partitions and the Y-Axis: I Loss values obtained in TPTDS algorithm.

The series 1 line in graph 3 indicates the TPTDS algorithm execution with one dimensional map reduce framework (In proposed work). The series 2 line indicates the TPTDS algorithm execution with multidimensional map reduce framework (In base paper). By comparing the two lines in the 3rd graph, the performance is better in multidimensional anonymization in the series 2 line by decreasing the I Loss.




Graph 4

X-Axis: K-Anonymity parameter and the Y-Axis: I Loss in TPTDS algorithm.

The line series 1 indicates the TPTDS algorithm execution with one dimensional map reduce framework (In base paper). The line series 2 indicates the TPTDS algorithm execution with multidimensional map reduce framework (In proposed work). By comparing the two lines in the 4th graph, the performance is better in multidimensional anonymization shown in the series 2 line by decreasing the I Loss. If the IGPL value increases then ILoss will also increases then degrades the performance on data anonymization.

Advantage on the proposed Multidimentional anonimyzation framework

1. Multidimensional scheme that recursively partitions the domain space to improve flexibility with regard to single-dimensional anonymisation.
2. Improve the scalability and efficiency by indexing anonymous large data records by comparing the traditional methods.
3. It improves in time efficiency and which are cost effective in execution.
4. A method to reduce the required memory footprint.
5. Accurate in scheduling than the traditional MapReduce framework.

Disadvantages

1. This approach is only effective in systems with high throughput, low-latency.
2. Data analysis, share and mining is a challenging research issue due to increasingly larger volumes of data sets
3. High data splitting cause transmission overhead
4. User constraints such as deadlines are important requirements which are not considered.

## 5. Conclusion

In order to overcomethe existing problems, it is essential that a solution is devised. For existing single dimensional anonymization technique, the proposed multidimensional anonymization technique shows the better result. For providing security on the large dataset, here proposes a Two Phase Top-Down Specialization anonymization approach using multidimensional mapreduce framework on hadoop. K-anonymization works by making each tuple in a data set identical to at least k-1 other tuples. The results proved

that as the privacy level of users is low, the anonymization results in more information loss than those with high privacy level. The performance of the system is also evaluated based on the execution time. In the TPTDS anonymity technique is capable of anonymizing data to a reasonable level of privacy while still retaining the data utility in which if the dataset is so high. The experimental result is that the better result shows in anonymization with multidimensional mapreduce framework by comparing the one dimensional mapreduce framework. The intermediate results are merged and further anonymized to produce consistent k- anonymous data sets in the second phase. It have creatively applied MapReduce on hadoop to data anonymization and deliberately designed a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way.

## 6. Future Work

There are several conditions that could be studied in the future. Risk evaluation could give an insight into how much information adversaries could dig out. In future it is expected to design the TPTDS anonymization by the bottom up generalization scheme on taxonomy tree for the anonymization. As future work a combination of top-down and bottom up approach generalization is contributed for data anonymization in which data Generalization hierarchy is utilized for anonymization. To handle iterative MapReduce jobs, in future, can use Twister- a distributed in-memory MapReduce runtime optimized for iterative MapReduce computations. Many Cloud computing vendors like Cloudera now a days are offering HaaS(Hadoop as a Service). This can ease the conFigureuration labor and can provide on-click elasticity of computational resources too.

## References

[1] X. Zhang, L. T. Yang, C. Liu and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization using MapReduce on Cloud", IEEE Transactions on Parallel and Distributed Systems (TPDS), ISSN: 1045-9219. (A*, IF: 1.796), vol. 25, no. 2, **(2014)**, pp. 263-373.
[2] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, "A PrivacyLeakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud", IEEE Trans. Parallel and Distributed Systems, to be published, **(2012)**.
[3] X. Zhang, C. Liu, S. Nepal, W. Dou and J. Chen, "Privacy-preserving Layer over MapReduce on Cloud and Green Computing (CGC 2012), Xiangtan, China, November, **(2012)**, pp. 304-310
[4] P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers", Proc. 23rdAnn. IFIP WG 11.3 Working Conf.Data and Applications Security XXIII (DBSec '09), **(2009)**, pp. 191-207.
[5] Liu H. and Orban D., "Cloud MapReduce: a MapReduce implementation on top of a cloud operating system", In: IEEE/ACM international symposium on cluster, cloud and grid computing, **(2011)**, pp 464–474.
[6] Candan K. S., Kim J. W., Nagarkar P., Nagendra M. and Yu R., "RanKloud: scalable multimedia data processing in server clusters", IEEE MultiMed, vol. 18, no. 1, **(2010)**, pp. 64–77.
[7] Dean J. and Ghemawat D. S. "MapReduce: simplified data processing on large clusters", Commun ACM, vol. 51, **(2008)**, pp. 107–113.
[8] B. C. M. Fung, K. Wang and P. S. Yu, "Anonymizing Classification Data for Privacy Preservation", IEEE Trans.Knowl..Data Eng., vol. 19, no. 5, **(2007)**, pp. 711-725.
[9] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi and A. W. Fu., "Utility based anonymization using local recoding", in ACM SIGKDD, **(2006)**.
[10] W. Jiang and C. Clifton, "A Secure Distributed Framework forAchieving k-Anonymity," VLDB J., vol. 15, no. 4, **(2006)**, pp. 316-333.
[11] Amazon Web Services, "Amazon Elastic Mapreduce,"http://aws.amazon.com/elasticmapreduce/, accessed on: January 05, **(2013)**.
[12] I. Roy, S. T. V. Setty, A. Kilzer, V. Shmatikov and E. Witchel, "Airavat: Security and Privacy for Mapreduce", Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI'10), **(2010)**, pp. 297-312.
[13] Brodsky A., Farkas C. and Jajodia S., "Secure databases: Constraints, inference channels, and monitoring disclosures", IEEE Transactions on Knowledge and Data Engineering, vol. 12, **(2000)**, pp. 900–919.

[14] N. Cao, C. Wang, M. Li, K. Ren and W. Lou, "PrivacyPreservingMulti-Keyword Ranked Search over Encrypted Cloud Data", Proc. IEEE INFOCOM, **(2011)**, pp. 829-837.
[15] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia, "A View of Cloud Computing", Comm. ACM, vol. 53, no. 4, **(2010)**, pp. 50-58.
[16] P. Mohan, A. Thakurta, E. Shi, D. Song and D. Culler, "Gupt: Privacy Preserving Data Analysis Made Easy", Proc. ACMSIGMOD Int'l Conf. Management of Data (SIGMOD '12), **(2012)**, pp. 349-360.
[17] L. H. Ying and W. G. Tzeng, "A Secure Erasure Code-BasedCloud Storage System with Secure Data Forwarding", IEEE Trans. and Distributed Systems, vol. 23, no. 6, **(2012)**, pp. 995-1003.
[18] X. Zhang and W. Dou, "Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud", IEEE transactions on computers, tc-2013-12-0869.
[19] "UCI Machine Learning Repository", ftp://ftp.ics.uci.edu/pub/machine-learnng-databases/.

## Authors

**Shalin Elizabeth S**, she has completed B.Tech from NSS College of Engineering Palakkad, and currently doing M.Tech in St Josephs College Of Engineering and technology,Palai. Her area of intrest includes bigdata analytics,cloud computing,biometrics. She published several papers in reputed national and international journals including IEEE in the area cloud computing, bigdata analytics and biometrics.

**Sarju S**, he is currently working as Assistant Professor in Computer Science and Engineering at St Josephs College Of Engineering and Technology,Palai. He has done B.Tech in CSE from Younus College of Engineering and Technology,Kollam and completed ME from KCG college of Technology(National Institute of Technology and Science)Chennai. Mr.Sarju's research interests include data mining, big data analytics and web security. He has many research publications in reputed national and international journals in the area data mining and bigdata analytics and websecurity. He is the Faculty Adviser in IEEE Computer Society and Life time Member in Indian Society for Technical Education (ISTE).