# Differential Privacy Preserving of Training Model In Wireless Big Data with Edge Computing

Karthika.B[1], Dr.Subha.R[2], Venkatesh.V[3]
[1,2,3] *Department of CSE, Sri Eshwar College of Engineering, Coimbatore, India*

***Abstract*- In this project implement a machine learning strategy for smart edges using differential privacy. In existing system focus attention on privacy protection in training datasets in wireless big data scenario. Moreover, to guarantee privacy protection by adding Laplace mechanisms, and design two different algorithms Output Perturbation (OPP) and Objective Perturbation (OJP), which satisfy differential privacy. In addition, consider the privacy preserving issues presented in the existing literatures for differential privacy in the correlated datasets, and further provided differential privacy preserving methods for correlated datasets, guaranteeing privacy by theoretical deduction. This approach converts the original sample data sets into a group of Non-Sensitive data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an accurate analysis can be built directly from those unreal data sets. This novel approach can be applied directly to the data storage as soon as the first sample is collected. The Relevant Columns Values Swapping approach is compatible with other privacy preserving approaches, such as without cryptography, for extra protection.**

**Index Terms- OJP, OPP, Privacy Preserving, Geometric Perturbation model, Row swapping**

## I. INTRODUCTION

The use of Big Data is becoming a crucial way for leading companies to outperform their peers. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value. Indeed, we found early examples of such use of data in every sector we examined. In healthcare, data pioneers are analyzing the health outcomes of pharmaceuticals when they were widely prescribed, and discovering benefits and risks that were not evident during necessarily more limited clinical trials. Other early adopters of Big Data are using data from sensors embedded in products from children's toys to industrial goods to determine how these products are actually used in the real world. Such knowledge then informs the creation of new service offerings and the design of future products.

Big Data will help to create new growth opportunities and entirely new categories of companies, such as those that aggregate and analyse industry data. Many of these will be companies that sit in the middle of large information flows where data about products and services, buyers and suppliers, consumer preferences and intent can be captured and analyzed.

In addition to the sheer scale of Big Data, the real-time and high-frequency nature of the data are also important. For example, 'nowcasting,' the ability to estimate metrics such as consumer confidence, immediately, something which previously could only be done retrospectively, is becoming more extensively used, adding considerable power to prediction. Similarly, the high frequency of data allows users to test theories in near real-time and to a level never before possible.

- Big Data is Timely – 60% of each workday, knowledge workers spend attempting to find and manage data.
- Big Data is Accessible – Half of senior executives report that accessing the right data is difficult.
- Big Data is Holistic – Information is currently kept in silos within the organization. Marketing data, for example, might be found in web analytics, mobile analytics, social analytics, CRMs, A/B Testing tools, email marketing systems, and more… each with focus on its silo.
- Big Data is Trustworthy – 29% of companies measure the monetary cost of poor data quality. Things as simple as monitoring multiple systems for customer contact information updates can save millions of dollars.

- Big Data is Relevant – 43% of companies are dissatisfied with their tools ability to filter out irrelevant data. Something as simple as filtering customers from your web analytics can provide a ton of insight into your acquisition efforts.
- Big Data is Secure – The average data security breach costs $214 per customer. The secure infrastructures being built by big data hosting and technology partners can save the average company 1.6% of annual revenues.
- Big Data is Authoritive – 80% of organizations struggle with multiple versions of the truth depending on the source of their data. By combining multiple, vetted sources, more companies can produce highly accurate intelligence sources.
- Big Data is Actionable – Outdated or bad data results in 46% of companies making bad decisions that can cost billions.

A large body of research has been devoted to the protection of sensitive information when samples are given to third parties for processing or computing. It is in the curiosity of research to disseminate samples to a large audience of researchers, without making strong assumptions concern their trustworthiness.

Even if information collectors ensure that information are released only to third parties with non-malicious intent (or if a privacy preserving approach can be applied before the data are released), there is always a chance that the information collectors may inadvertently disclose samples to malicious parties or that the samples are actively stolen from the collectors.

Samples may be leaked or stolen anytime during the storing process or while residing in storage. This paper focuses on anticipating such attacks from third parties for the whole lifetime of the samples.

Contemporary research in privacy preserving data mining mainly falls into one of two categories: perturbation and randomization-based approaches, and secure multiparty computation (SMC)-based approaches. SMC approaches use cryptographic tools for collaborative data mining computation by multiple parties. Samples are distributed among completely different parties and they take part in the information computation and communication process. SMC analysis focuses on protocol development for protecting privacy among the involved parties or computation efficiency; however, centralized processing of samples and storage privacy is out of the scope of SMC.

The amount of data being collected and stored every day by private and public sectors increased dramatically**.** Almost all industries, organizations and hospitals are maintaining personal information about individuals for decision making or pattern recognition. Security risk is very high while sharing this personal sensitive information among different data collectors.

Therefore, privacy-preserving processes have been developed to sanitize private information from the samples while keeping their utility. For that safe and secure distributed computation new privacy preserving data mining algorithm has been developed. The main goal of these algorithms is to prevent that sensible information from hackers, during knowledge extraction from huge amount of data.

This project introduces a privacy preserving approach that can be applied to decision tree learning, without concomitant loss of accuracy. This approach converts the original sample data sets into a group of unrealized data sets, from which the original data samples cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an efficient and accurate decision tree can be built directly from those unreal data sets.

This novel Relevant Columns Values Swapping approach can be applied directly to the data storage as soon as the first sample is collected. The approach is adaptable with other privacy preserving approaches such as cryptography for extra protection

- To make the machine learning process Ranking Swapping Model
- To generate the Relevant Columns Values Swapping by the different dataset.
- Geometric perturbation mechanism is used for group key authentication
- Usage of data set gives, efficient and real time results here
- Multiparty privacy-preserving collaborative mining.
- Privacy preserving mining service to multiple data providers.
- Converts the original samples into some unrealized data

## II.RELATED WORKS

Linghe Kong and Daqiang Zhang [1] describe a new-generation industries heavily rely on big data to improve their efficiency. Such big data are commonly collected by smart nodes and transmitted to the cloud via wireless. Due to the limited size of smart node, the shortage of energy is always a critical issue, and the wireless data transmission is extremely a big power consumer. Aiming to reduce the energy consumption in wireless, this article introduces a potential breach from data redundancy. If redundant data are no longer collected, a large amount of wireless transmissions can be cancelled and their energy saved. Motivated by this breach, this article proposes a compressive-sensing-based collection framework to minimize the amount of collection while guaranteeing data quality. This framework is verified by experiments and extensive real- trace-driven simulations

Sihai Zhang and Dandan Yin et al [2] describe the important aspects in this topic, including data set information, data analysis techniques, and two case studies. We categorize the data set in the telecommunication networks into two types, user-oriented and network-oriented, and discuss the potential application. Then, several important data analysis techniques are summarized and reviewed, from temporal and spatial analysis to data mining and statistical test. Finally, we present two case studies, using Erlang measurement and call detail record, respectively, to understand the base station behavior. Interestingly, the night burst phenomenon of college students is revealed by comparing the base stations location and real-world map, and we conclude that it is not proper to model the voice call arrivals as Poisson process

Matt Fredrikson and Somesh Jha et al [3] describe a machine-learning (ML) algorithms are increasingly utilized in privacy-sensitive applications such as predicting lifestyle choices, making medical diagnoses, and facial recognition. In a model inversion attack, recently introduced in a case study of linear classifiers in personalized medicine by Fredrikson, adversarial access to an ML model is abused to learn sensitive genomic information about individuals. Whether model inversion attacks apply to settings outside theirs, however, is unknown. Author develops a new class of model inversion attack that exploits confidence values revealed along with predictions. Our new attacks are applicable in a variety of settings, and we explore two in depth: decision trees for lifestyle surveys as used on machine-learning-as-a-service systems and neural networks for facial recognition. In both cases confidence values are revealed to those with the ability to make prediction queries to models. We experimentally show attacks that are able to estimate whether a respondent in a lifestyle survey admitted to cheating on their significant other and, in the other context, show how to recover recognizable images of people's faces given only their name and access to the ML model. We also initiate experimental exploration of natural countermeasures, investigating a privacy-aware decision tree training algorithm that is a simple variant of CART learning, as well as revealing only rounded confidence values. The lesson that emerges is that one can avoid these kinds of MI attacks with negligible degradation to utility.

Yi Wang And Qixin Chen [4] describe a competitive retail market, large volumes of smart meter data provide opportunities for load serving entities to enhance their knowledge of customers' electricity consumption behaviors via load profiling. Instead of focusing on the shape of the load curves, this paper proposes a novel approach for clustering of electricity consumption behavior dynamics, where "dynamics" refer to transitions and relations between consumption behaviors, or rather consumption levels, in adjacent periods. First, for each individual customer, symbolic aggregate approximation is performed to reduce the scale of the data set, and time-based Markov model is applied to model the dynamic of electricity consumption, transforming the large data set of load curves to several state transition matrixes. Second, a clustering technique by fast search and find of density peaks (CFSFDP) is primarily carried out to obtain the typical dynamics of consumption behavior, with the difference between any two consumption patterns measured by the Kullback–Liebler distance, and to classify the customers into several clusters. To tackle the challenges of big data, the CFSFDP technique is integrated into a divide-and- conquers approach toward big data applications. A numerical case verifies the effectiveness of the proposed models and approaches

Quan Geng and Pramod Viswanath [5] study the (nearly) optimal mechanisms in (E, δ) -differential privacy for integer-valued query functions and vector-valued (histogram-like) query functions under a utility- maximization/cost-minimization framework. Within the classes of mechanisms oblivious of the database and the queries beyond the global sensitivity, we characterize the tradeoff between and δ in utility and privacy analysis for histogram-like query functions, and show that the (□, δ) -differential privacy is a framework not much more general than the ( , 0 ) -differential privacy and ( 0 ,δ) -differential privacy in the context of 1 and 2 cost functions, i.e., minimum expected noise magnitude and noise power. In the same context of 1 and 2 cost functions, we show the near-optimality of uniform noise mechanism and discrete Laplacian mechanism in the high privacy regime (as (□, δ) → ( 0 , 0 ) ). We conclude that in (□, δ) -differential privacy, the optimal noise magnitude and the noise power are □( min (( 1 /□), ( 1 /δ))) and ( min (( 1 /□ 2 ), ( 1 /δ 2 ))) , respectively, in the high privacy regime.

Engin Zeydan And Mehdi Bennis [6] describe order to cope with the relentless data tsunami in 5 G wireless networks, current approaches such as acquiring new spectrum, deploying more base stations (BSs) and increasing nodes in mobile packet core networks are becoming ineffective in terms of scalability, cost and flexibility. In this regard, context-aware 5 G networks with edge/cloud computing and exploitation of big data analytics can yield significant gains to mobile operators.

In this article, proactive content caching in 5 G wireless Dong Liu and Binqiang Chen describes Caching and wireless edge promising way of boosting spectral efficiency and reducing energy consumption of wireless systems. These improvements are rooted in the fact that popular contents are reused, asynchronously, by many users. In this article author first introduce methods to predict the popularity distributions and user preferences, and the impact of erroneous information. Design aspects to discuss the two aspects of caching systems, namely content placement and delivery. An expound the key differences between wired and wireless caching, and outline the differences in the system arising from where the caching takes place, e.g., at base stations, or on the wireless devices themselves. Special attention is paid to the essential limitations in wireless caching, and possible tradeoffs between spectral efficiency, energy efficiency and cache size.

Y. Hong and j. Vaidya [8] describe a severe privacy leakage in the AOL search log incident has attracted considerable worldwide attention. However, all the web users' daily search intents and behavior are collected in such data, which can be invaluable for researchers, data analysts and law enforcement personnel to con duct social behavior study, criminal investigation and epidemics detection . Thus, an important and challenging research problem is how to sanitize search logs with strong privacy guarantee and sufficiently retained utility. Existing approaches in search log sanitization are capable of only protecting the privacy under a rigorous standard or maintaining good output utility. To the best of our knowledge, there is little work that has perfectly resolved such tradeoff in the context of search logs, meeting a high standard of both requirements.

## III. METHODOLOGY

The perturbation and randomization-based approach that protects centralized sample data sets utilized for decision tree data mining. Privacy preservation is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or theft. In contrast to other sanitization methods, the approach does not affect the accuracy of data mining results. The decision tree can be built directly from the sanitized data sets, such that the originals do not need to be reconstructed. Moreover, this approach can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected. But the following assumptions are made in the existing system: first, as is the norm in data collection processes, a sufficiently large number of sample data sets have been collected to achieve significant data mining results covering the whole research target.

Second, the number of data sets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for decision making. Fourth, all data

collected are discretized; continuous values can be represented via ranged-value attributes for decision tree data mining

The existing ID3 (Iterative Dichotomiser 3) decision tree learning algorithm which covers the discrete-valued attributes are implemented also in the proposed system. To preserve privacy when datasets are given to multiple parties, the proposed system finds the solution for the key problem of applying geometric data perturbation in multiparty collaborative mining which securely unify multiple geometric perturbations that are preferred by different parties, respectively. In the proposed system, the privacy is preserved even if the data is spread across multi parties. Suppose a bank issues the account holders' some of the attributes to more than one insurance agency. Then from the attributes of the table along with the records given to one insurance agency, other agency could not guess or identify the facts regarding the account holders. Likewise, if two agencies give their data set (retrieved from the bank) to other parties, they must not identify the facts by combing both data sets

On the privacy issues of the training datasets for smart edges, and proposed a machine learning strategy using differential privacy in wireless big data scenario. One rank swapping is applied to test data set and two different algorithms are designed, i.e., Output Perturbation (OPP) and Objective Perturbation (OJP) to satisfy differential privacy. The following modules are present in the project.

- Rank swapping
- Output perturbation algorithm
- Objective perturbation algorithm
- Identification of non-sensitive columns for perturbation
- Relevant columns value swapping and perturbation

RANK SWAPPING

In this module, from the dataset, all the records are taken. Then all the columns are sorted such as second column values are sorted, then third column values and so on up to last column. Then each pair of rows is taken first. Then from second column to last column, all the column values are interchanged. If the row count is odd, then previous row of last row and last row values are interchanged. This perturbation is made to alter the values with no loss in data, i.e., no modification in data values.

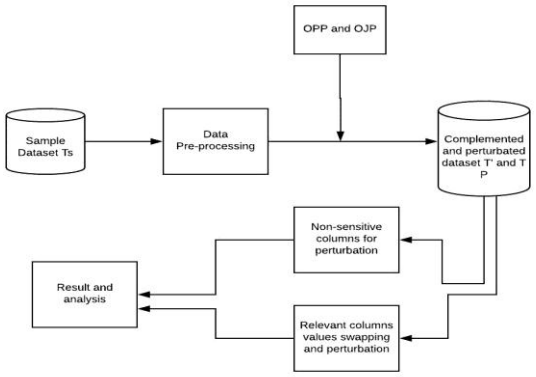
OUTPUT PERTURBATION ALGORITHM

The datasets are taken such that D = {(x1,y1),(x2,y2),...,(xn,yn)} and D0 = {(x1,y1),(x2,y2),...,(x0n,y0 n)}, and suppose that the outputs of the algorithm are W(D) and W(D0), respectively. For the objective function W(D) = argminK(u,D), the output of the algorithm is W(D) + q, where q is a random Laplace noise.

In order to measure the quality of the prediction function u : X → Y in the training datasets, the empirical risk minimization method is employed to minimize the K(u,D) with a nonnegative loss function s : Y ×Y → R.

$K(u,D) = 1/n \sum_{i=1}^{n} s(u(x_i),y_i) + \lambda Z(u)$

where Z(u) represents the smoothness of a function, u is a linear prediction function, and λ is an adjustable parameter. Here equilidiean distance is found out such that distance from origin (0,0) to (x1,y1), (x2,y2) up to (xn,yn) and summed out to prepare K(u,D). Like that five values are prepared and minimum value index is taken. That laplacian noise is used to perturbate the dataset.




OBJECTIVE PERTURBATION ALGORITHM

Like the previous module, the datasets are taken such that D = {(x1,y1),(x2,y2),...,(xn,yn)} and D0 = {(x1,y1),(x2,y2),...,(x0n,y0 n)}, and suppose that the outputs of the algorithm are W(D) and W(D0), respectively. For the objective function W(D) = argminK(u,D). Here the laplacian noise is added to K(u,D)s and the minimum value laplacian noise is selected. That perturbated records are given as new dataset.

IDENTIFICATION OF NON-SENSITIVE COLUMNS FOR PERTURBATION

In this module, the column with more similar values is identified as well as with very less similar values is also identified. Those column values are not perturbated and given with original values in the perturbated (other columns perturbated) dataset. This is carried out to eliminate the burden of perturbating less important columns.

RELEVANT COLUMNS VALUE SWAPPING AND PERTURBATION

In this module, the columns with more similar values (but not exact) are identified. Those column values are swapped in the same row. This is carried out to non-loss perturbation of the dataset. Present the Relevant Columns Values Swapping approach which helps to multiparty privacy-preserving collaborative mining.

- From the two given data sets, the original facts cannot be guessed.
- Privacy is preserved even if the data is spread across multi parties.
- Consider multiple service providers collaboratively providing the privacy preserving mining service to multiple data providers

## IV. CONCLUSION

Privacy preservation via data set complementation fails if all training data sets are leaked because the data set reconstruction rank swapping algorithm is generic. This project covers the application of this new privacy preserving approach with the OPP and OJP algorithm and discrete-valued attributes only. The norm in data collection processes, a sufficiently large number of sample data sets have been collected to achieve significant data mining results covering the whole research target. Second, the number of data sets leaked to potential attackers constitutes a small portion of the entire sample database.

- This project covers the applications of this new privacy preserving approach with the OPP and OJP algorithm
- In addition Rank Swapping mechanism is used so that the data is secured even distributed to more than one parties.
- Suitable for multiparty Swapping data distribution

## V. FUTURE ENHANCEMENTS

The project provides a best assistance in converting data sets into unrealized data sets and decision tree learning. The application become useful if the below enhancements are made in future.

If the application is designed as web site, it can be access from anywhere. In addition, different records can be converted into different formats of unrealized data sets and given to different parties. The application is developed such that above said enhancements can be integrated with current modules. The further research is required to overcome this limitation. As it is very straight forward to apply a cryptographic privacy preserving approach, such as the (anti)monotone frame-work, along with data set complementation, this direction for future research could correct the above limitation.

## REFERENCES

[1] X. Ding, Y. Tian, and Y. Yu, “A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations,” IEEE Transactions on Industrial Informatics , vol. 12, no. 3, pp. 1232-1242, 2016.

[2] L. Kong, D. Zhang, and Z. He, “Embracing big data with compressive sensing: a green approach in industrial wireless networks,” IEEE Communications Magazine, vol. 54, no. 10, pp. 53-59, 2016.

[3] F. Xu, Y. Lin, and J. Huang, “Big data driven mobile traffic understanding and forecasting: a time series approach,” IEEE Transactions on Services Computing , vol. 9, no. 5, pp. 796-805, 2016.

[4] S. H. Zhang, D. D. Yin, and Y. Q. Zhang, “Computing on base station behavior using erlang measurement and call detail record,” IEEE Transactions on Emerging Topics in Computing , vol. 3, no. 3, pp. 444-453, 2015.

[5] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit onfidence information and basic countermeasures,” In Proceedings of the 22nd ACM SIGSAC

Conference on Computer and Communications Security , 2015, pp. 1322-1333.

[6] O. Denas and J. Taylor, "Deep modeling of gene expression regulation in an erythropoiesis model," In Representation Learning , ICML Workshop, 2013.

[7] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art, challenges, and implementation in next generation mobile networks(vEPC)," IEEE Network , vol. 28, no. 6, pp. 18-26, 2014.

[8] Y. Wang, Q. Chen, and C. Kang,"Clustering of electricity consump- tion behavior dynamics toward big data applications," IEEE Transactions on Smart Grid , vol. 7, no. 5, pp. 2437-2447, 2016.

[9] Q. Liu, C. C. Tan, J. Wu, and G. Wang, "Towards differential query services in cost-efficient clouds," IEEE Transactions on Parallel and Distributed Systems , vol. 25, no. 6, pp. 1648-1658, 2014.

[10] Y. Hong, J. Vaidya, H. Lu, P. Karras, and S. Goel, "Collaborative search log sanitization: Toward differential privacy and boosted utility," IEEE Transactions on Dependable and Secure Computing , vol. 12, no. 5, pp. 504-518, 2015.

[11] T. Zhang and Q. Zhu, "Dynamic differential privacy for ADMM-based distributed classification learning,"IEEE Transactions on Information Forensics and Security , vol. 12, no. 1, pp. 172-187, 2017.

[12] Q. Geng and P. Viswanath, "Optimal noise adding mechanisms for approximate differential privacy," IEEE Transactions on Information Theory, vol. 62, no. 2, pp. 952-969, 2016.

[13] E. Zeydan et al., "OBig data caching for networking: Moving from cloud to edge," IEEE Communication Magazine , vol. 54, no. 9, pp. 36-42, 2016.

[14] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," IEEE Communication Magazine, vol. 54, no. 9, pp. 22-28, 2016.

[15] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the accuracy of differentially private histograms through consistency," in Proc. Of VLDB Endowment , 2010, pp. 1021-1032