RESEARCH ARTICLE

Verification of K-Anonymity Model using Mapreduce for Big Data Privacy-Preserving in D2D Communication

Shelly Bhardwaj[1]*, Abhishek Kumar Mishra [2], and Rahul Kumar Mishra [3]

[1]Research Scholar, Department of Computer Science and Engineering, IFTM University, Moradabad, Uttar Pradesh, India.

[2]Associate Professor, Department of Computer Science and Engineering, IFTM University Moradabad, Uttar Pradesh, India

[3]Professor (Director), School of Computer Science and Applications, IFTM University Moradabad, Uttar Pradesh, India

Received: 26 Feb 2023 Revised: 05 Apr 2023 Accepted: 08 May 2023

***Address for Correspondence**
Shelly Bhardwaj
Research Scholar,
Department of Computer Science and Engineering,
IFTM University, Moradabad,
Uttar Pradesh, India.
E. Mail: shellybhardwaj29@gmail.com

ABSTRACT

Due to the extraordinary advantages of quick transmission as well as the reaction on data delivery, including its variety of applications, device-to-device (D2D) communication has attracted a lot of research interest as a distinctive and exciting innovation for 5G communications networks throughout the world. Big data analytics offers new advantages but also poses a significant barrier to D2D communications since the data frequently contains sensitive personal or corporate information that is vulnerable to leaking. Today's climate has almost made privacy protection a must for D2D services, yet technology advancements have left a substantial research gap. By implementing the (k)-anonymity model using Map Reduce, the authors developed a novel method for ensuring privacy in huge data in D2D communication. In this approach, enormous D2D datasets are handled via the (k)-anonymity paradigm, and various massive datasets are categorised and grouped using Map Reduce. The findings of this experiment and analysis demonstrate that this recommended prototype is more durable and workable than earlier approaches in terms of huge data secrecy, lowest data or information losses, and less computation time.

Keywords: Big Data, Data Security, D2D Communication, K- anonymity, Privacy Preserving.

INTRODUCTION

Modern communication devices and cutting-edge technology are drastically altering people's lives all around the world. These modern technology, such as cell phones, computers, and tablets, are assisting people in a variety of ways, including texting and calling, exchanging image files, and many other things. With the aid of the internet, individuals share this kind of massive amounts of data on a daily basis [1], [2]. A potential approach for 5G cellular networks is device-to-device (D2D) communications. D2D communications have been shown to improve network performance in terms of spectrum efficiency, power dissipation, cellular coverage, and communication capacity and latency. The amount of data and traffic generated by mobile networks has increased significantly in recent years due to the improvement in the quality and availability of multimedia services. To transfer interesting files locally, users choose to use wireless short-range D2D connection.

According to recent studies based on their social and mobile behaviours, users prefer to communicate content offline via D2D communication. On the other hand, earlier studies on the subject have relied on small-scale data analysis and the development of algorithms for particular user groups. Given the rapid growth of mobile users and devices, D2D technology should be able to adapt to the transmission of enormous amounts of data to a large number of users. This paper introduces a (k)-anonymous D2D big data privacy-preserving architecture based on Map Reduce to provide speedy sharing, high accuracy on deliveries, efficient and intelligent distribution, and right content promotion to a large number of users. Big data technology presents new opportunities for sharing D2D communication capabilities, but it also presents challenges for the conventional data analysis of mobile user groups. The dimensionality, heterogeneity, and complexity of the data worsen the security and privacy problems associated with D2D communication. Big D2D data frequently includes a user's or an event's private information. Private user information may be leaked as a result of the mining, analysis, and processing of D2D big data. There are a large number of sensing nodes in D2D communication systems that continuously convey a large amount of data about citizens, organisations, and national infrastructures, much of which contains sensitive information. If they are not sufficiently protected throughout data mining, analysis, and processing, these sensitive data may be disclosed.

The next phase of cellular technology, which will include wireless telecommunication infrastructure, will have cutting-edge innovations like D2D telecommunications. Datasets used in D2D telephony must be protected in order to head off damaging attacks. Meanwhile, encrypted D2D transmissions between portable phones continue to be difficult. In this paper, researchers describe a technique for controlling D2D data accessibility through the use of an attribute-rooted encryption strategy to ensure the privacy of large datasets during the communication process. A General-Trust grade provided by central networking, anLocal-Trust degree determined by a device, or perhaps both in real-time, might be used to accomplish this new concept.

LITERATURE REVIEW

In [3], U. N. Kar *et al.* conducted a review on D2D communication within a mobile cellular networking environment. D2D telecommunication, which offers much-reduced delay for consumer interaction, is anticipated to serve a large part in emerging cellular infrastructures. Either licensed or unlicensed bandwidth could be used by this novel concept. It represents a fresh improvement on the established wireless telecommunication concept. Nevertheless, despite its advantages, there are several economic but also technological problems that must be overcome before it is integrated into the wireless environment. The basic traits of D2D telecommunication are covered throughout this article, along with its use contexts, framework, and technological aspects, including current investigation fields. The usage of D2D within cellular networking has indeed been examined across several researches. For instance, to lower the same expenses of ground controlling headquarters resulting from the requirement of organizing cars throughout vast numbers as coding peddlers, the researchers of [4] suggested a machine-learning-rooted coding dissemination system. This method chooses cars with a greater covering ratio as well as dependability as coding distributors. Through suggesting a process orchestration as well as datasets aggregation architecture which may provide solutions for organizing the datasets including merging data packages. Liu *et al.* minimized the operation reply latency as well as duplication of datasets in [5]. In [6], B. Yang *et al.* discussed an automated repairing scheme for D2D telecommunication routing in real-time. For improving the secrecy along with the communication quality effective maintenance of the cellular network is very essential. Owing to increased client activity, communication datasets easily cross the buffer's edge, decreasing the amount of protected dataset information. The issue of inadequate telecommunication secrecy is caused by the fact that previous restoration techniques primarily focus on the features of covering dataset information, neglecting the influence of networking infrastructure data transfer latency as well as packet failure throughout estimation. However, this approach is inefficient in the case of larger datasets translation on the limited bandwidth channels owing to the constant increment in the number of consumers all around the world. Many adaptations to the k-anonymity model have been developed since Sweeney *et al.* [7] first presented it to solve its flaws. For instance, [8] suggests t-closeness and [8] suggests l-diversity, both of which defend against inference-based attacks. None of the models that have been suggested take into account combining different anonymization processes.

Zhaohao *et al.* [9] explained about Privacy and security in the big data age have drawn significant attention in academia and industry. This article examines privacy and security in the big data paradigm by proposing a model for privacy and security in the big data age and classification of big data-driven privacy and security. It extends the big data body of knowledge, highlights important research topics, and identifies critical gaps through statistical analysis of big data and its impacts on privacy and security based on literature data published from 1916 to 2016. It also presents state-of-the-art privacy and security based on the analysis of SCOPUS data from 2012 to 2016. The result shows that privacy and security face new challenges and require new policies, technologies, and tools for protecting privacy in the big data paradigm. The proposed approach might facilitate the research and development of privacy and security, and big data-driven privacy and security in terms of technology, governance, and policy development. In [10], Jordi Domingo-Ferrer *et al.* explained the challenges raised by big data in privacy-preserving data management. First, we examine the conflicts raised by big data concerning preexisting concepts of private data management, such as consent, purpose limitation, transparency, and individual rights of access, rectification, and erasure. Anonymization appears as the best tool to mitigate such conflicts, and it is best implemented by adhering to a privacy model with precise privacy guarantees. For this reason, we evaluate how well the two main privacy models used in anonymization (k-anonymity and ε-differential privacy) meet the requirements of big data, namely composability, low computational cost, and linkability. Shuai Li et al.[11], discussed about the Internet of things (IoT) has become a significant part of our daily life. Composed of millions of intelligent devices, IoT can interconnect people with the physical world. With the development of IoT technology, the amount of data generated by sensors or devices is increasing dramatically. IoT-based big data has become a very active research area. One of the key issues in IoT-based big data is ensuring the utility of data while preserving privacy. In this paper, we deal with the protection of big data privacy in the data storage phase and propose a searchable encryption scheme satisfying personalized privacy needs. Our proposed scheme works for all file types including text, audio, image, video, etc., and meets different privacy needs of different individuals at the expense of high storage costs. We also show that our proposed scheme satisfies index In distinguish ability and trapdoor In distinguish ability.

In [12], A. Ozhelvaci *et al.* discussed another article on handover secrecy as well as D2D telecommunication within the 5G (5th Generation) HetNets. Technical specifications for these currently being developed coming-generation cellular communication technologies are set by the 3GPP (Third-Generations and Partnership-Project (3GPP). This is indeed known as 5G cordless cellular networking, which has emerged as the model for bringing not just answers to the growing need for vast amounts of dataset transmission but also enormously linked objects, for example, the IoT (Internet-of-things) and many other additional activities. Additionally, 5G is anticipated to provide the quickest, best dependable networking connection to accommodate vast dataset traffic as well as terminals that are heavily linked with minimal delay as well as excellent capacity. However, this approach contains numerous limitations namely the more computational complexity. In [13], X. Chen *et al.* discussed the investigation growth scheme on the big datasets secrecy technique. A big dataset has a tremendous impact on folks' life as a fresh but dynamic area of financial growth, a creative accelerator of societal growth, as well as a smart instrument for defining country competence. Increased adoption of big dataset uses is, nevertheless, being hampered more and more by big data protection due to increased societal knowledge of the worth of dataset and the rapid growth of big dataset platforms. A single big data privacy paradigm has not yet been developed, but as big dataset technologies as well as architecture continue to advance, academics continue to have divergent views on that fundamental concept as well as essential elements of big dataset security.

In [14], Yi Liu *et al.* explained the rapid development of 5G networks, big data, and IoT, data in many environments is often continuously and dynamically generated with high growth rates, just like a stream. Thus, we call it a big data stream, which plays an increasingly important role in all walks of life. However, how to verify its authenticity becomes a challenge when this big data stream is in an untrusted environment such as a cloud platform, for it faces the problems just like delay-sensitive, unpredictable data size and privacy leaks caused by third-party audits. To solve these problems, we propose a new authenticate data structure named privacy-preserving adaptive trapdoor hash authentication tree (P-ATHAT) by introducing trapdoor hash and BLS signature to the Merkle hash tree. The P-ATHAT scheme realizes real-time verification of the data stream and can dynamically expand its structure as the data stream arrives. These characteristics not only shorten the authentication path but also solve the single point failure problem of the conventional authentication trees and enhance the robustness of the scheme. Moreover, we construct a homomorphic verification scheme above the tree structure to solve the privacy leakage problem in the third-party audit. Finally, security analysis and detailed experimental evaluation are performed on the proposed scheme, both results demonstrate that it is desirable for big data stream authentication and privacy-preserving in practical application. The authors of [15] make a suggestion on how to combine differential privacy and k-anonymity in a single data release. The authors suggest selecting a sample at random from the initial dataset and processing this sample to achieve k-anonymity before publication in order to add a stochastic component to k-anonymity (a deterministic system). Differential privacy can also be achieved thanks to the uncertainty introduced by randomly choosing users to be included in the disclosed (yet anonymized) dataset.If the information for each person contained in the release cannot be discerned from at least k-1 other individuals whose information appears in the release, the release is said to have the k-anonymity [16, 17] property. A database, in the context of k-anonymization problems, is a table with n rows and m columns, where each row of the table represents a record related to a specific person from a population and the columns of the table serve as the data.

A novel set of k- anonymity rules with roots in clustering for secrecy maintenance was the subject of research by S. Ni et al.[18] K anonymity is a practical idea that may be used in a variety of ways to protect privacy while disclosing information. They favour local generalisation because it results in less data loss. Nevertheless, these techniques struggle to work well when dealing with enormous amounts of data since they take a lot of time. To address these issues, the research offers a different swarm K-anonymity technique that is likewise multiplexing optimised. Our results demonstrate that such methodology performs best in terms of data loss as well as speed when compared to conventional strategies and Incognito methods .Another adaptive k-Anonymity strategy was investigated by K. Arava *et al.* [19] for the confidentiality of cloud datasets. Cloud services require a high level of information security for information exchange. Some consumers find it challenging to use contemporary technological platforms for encrypted communication in business healthcare applications. Although k-anonymity and "data analysis employing data mining technologies" were important, extensive research has been done on encrypted datasets of sensitive data. The term "k-anonymity" in the context of vulnerable data protection refers to delicate private information that has been released and could, therefore, be linked to the identification of the k1 and other users.To achieve k-anonymity, clustering techniques are used. Yet, it is challenging to identify the right seeding numbers for capturing relevant data that can be anonymous at the same time to reduce data loss. The research employs the flexible k-anonymity strategy and adheres to a thorough methodology for seeding the selection of how to group all the data. With the goal of reducing data loss as well as runtime, early work is compared and assessed.

METHODOLOGY

Design

The term "5G" refers to the fifth generation of cellular networking, which mainly utilises device-to-device (D2D) communication. Once the communication channel has been established, datasets can be transferred directly without the use of intermediary devices. This can improve spectrum use, reduce the demand for data on the telecommunication platform's network architecture, and significantly enhance internet bandwidth. D2D communication was developed with the purpose of broadcasting important or interesting info to numerous other phone devices as well as requesting nearby peers for desired material. Throughout this procedure, a sizable amount of varied data is gathered. The possibilities, including the research value of looking through and using such huge, intricate databases, are vast. Threats to privacy are a major factor in D2D communication. The several stages of the life-cycle of huge datasets are depicted in Fig. 1, including dataset generation, storage, and processing. The framework for massive datasets and a threat prototype for D2D (Device to Device) communication for privacy preservation are shown in Fig. 2.

Instrument

The verification and secrecy assessment of the suggested scheme is done by utilizing Hadoop, which is indeed a Map Reduce-rooted software paradigm. The entire testing procedure of the proposed scheme is done by utilizing the described system structure: AMD Ryzen 7, 64-bit operating system, SSD (Solid State Drive) 512 GB and integrated with the 16 GB RAM. Map Reduce is indeed a pragmatic and faster software paradigm as well as a programming prototypical utilized for processing gigantic datasets. This Map Reduce package functions within two diverse stages such as Map as well as Reduce. The overall mapping jobs deal with the breaking as well as mapping of entire datasets while reducing the jobs shuffle as well as minimization of the entire data for effective segregation of a large amount of the datasets.

Data Collection

In this part, the researchers provided motivational examples of how to apply the proposed strategy. The major goal of the author is to increase secrecy in massive datasets D2D communication contexts, particularly since handling enormous amounts of data becomes more complicated and challenging in the present era. Below is a description of the whole implementation as well as the verification scenario. Table 1 illustrates the sources of the datasets as well as the outcomes of the mapper. As a result, a viable privacy-preserving architecture for D2D multimedia applications with massive data is necessary. Table 2 illustrates the grouping of the outcome data of Table 1 as likeness class.

Data Analysis

Equivalence class (EC) and information are divided into a number of dataset block segments by Map Reduce, which iterates the subsequent steps concurrently. When q increases, the projected sample size decreases until all data is attributed to the same EC. It's important to keep in mind that each EC can only have a fixed number of q parameters throughout this scenario. These ECs are included in master files, along with newly produced ECs to which all Map Reduce tasks are allocated. Every EC in the approach is included in this global document, and the drivers add any newly created EC to the end iterations. Each cycle's mapper, combiners, and reducer each serve a different purpose.

RESULT AND DISCUSSION

The experiment was carried out using Hadoop, a software framework that implements Map Reduce, and the characteristics of the data transmission of the algorithms studied were examined by looking at the validity of data. This experiment uses two datasets, Poker Hand Datasets. The poker hand dataset consisted of 11 numeric attributes, with the odd and even QI having ranges of 1-4 and 1-13, respectively, and the SA being variable classes. The dataset was divided into tiny blocks using the pre-processed Map Reduce. Synthetic Datasets Using these two pieces of data, a synthetic dataset [10] was created. One had 10 million data records totalling 1.4 GB in size, whereas the other had

10 million records totalling 14 GB. A collection of data had 10 clusters with a random mean and bias, 15 dimensions, and 15 clusters. Each dimension in the mapper was scaled to 5 M100 M, and datasets of 10 M and 100 M were split into 70 copies and 300 fragments, respectively. The cut was made using the midpoint value and the longest side of the bounding rectangle, which were both set to q 2.

CONCLUSION

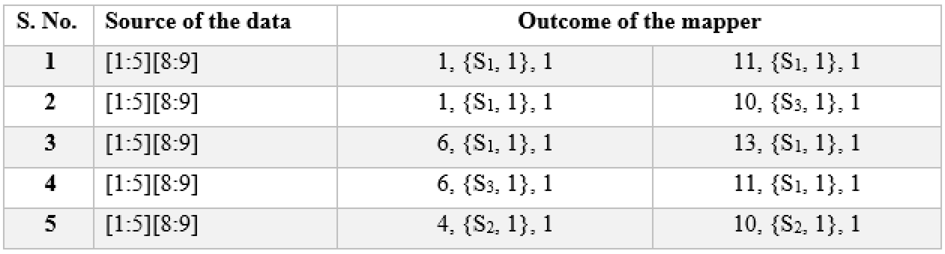
For categorising and organising the huge data datasets, the authors uses distributed Map Reduce, which increased computer efficiency and cut down on calculation times. We used the (k)-anonymity prototype for the secrecy framework to defend ourselves against various adversaries around the world. Results of the recommended paradigm show that this framework is more effective in terms of D2D communication secrecy for large volume datasets. The study and experimentation results show that this proposed prototype is more practical and robust than the existing techniques in terms of huge data secrecy, lowest data or information losses, and fastest computing times.

REFERENCES

1. L. A. Tawalbeh and G. Saldamli, “Reconsidering big data security and privacy in cloud and mobile cloud systems,” J. King Saud Univ. - Comput. Inf. Sci., 2021, doi: 10.1016/j.jksuci.2019.05.007.
2. S. Venkatraman and R. Venkatraman, “Big data security challenges and strategies,” AIMS Mathematics. 2019. doi: 10.3934/math.2019.3.860.
3. U. N. Kar and D. K. Sanyal, “An overview of device-to-device communication in cellular networks,” *ICT Express*. 2018. doi: 10.1016/j.icte.2017.08.002.
4. M. G. Sarwar Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, and F. Hussain, “Machine Learning at the Network Edge: A Survey,” *ACM Computing Surveys*. 2022. doi: 10.1145/3469029.
5. G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, “Toward an Intelligent Edge: Wireless Communication Meets Machine Learning,” *IEEE Commun. Mag.*, 2020, doi: 10.1109/MCOM.001.1900103.
6. B. Yang and K. Jiang, “Automatic Repair Method for D2D Communication Routing Buffer Overflow Vulnerability in Cellular Network,” *Sci. Program.*, 2021, doi: 10.1155/2021/3963574.
7. L. Sweeney, “k-Anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002
8. N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy beyond k-anonymity and `-diversity,” in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007, pp. 106–115.
9. Z. Sun, K. D. Strang, and F. Pambel, “Privacy and security in the big data paradigm,” *J. Comput. Inf. Syst.*, 2020, doi: 10.1080/08874417.2017.1418631.
10. J. Soria-Comas and J. Domingo-Ferrer, “Big Data Privacy: Challenges to Privacy Principles and Models,” *Data Sci. Eng.*, 2016, doi: 10.1007/s41019-015-0001-x.
11. S. Li, M. Li, H. Xu, and X. Zhou, “Searchable encryption scheme for personalized privacy in IoT-based big data,” *Sensors (Switzerland)*, 2019, doi: 10.3390/s19051059.
12. A. Ozhelvaci and M. Ma, “ Security for Handover and D2D Communication in 5G HetNets ,” in *Wiley 5G Ref*, 2020. doi: 10.1002/9781119471509.w5gref262.
13. X. Chen, Y. Gao, H. Tang, and X. Du, “Research progress on big data security technology,” *Scientia Sinica Informationis*. 2020. doi: 10.1360/N112019-00077.
14. Y. Sun, Q. Liu, X. Chen, and X. Du, “An Adaptive Authenticated Data Structure with Privacy-Preserving for Big Data Stream in Cloud,” *IEEE Trans. Inf. Forensics Secur.*, 2020, doi: 10.1109/TIFS.2020.2986879.
15. N. Li, W. Qardaji, and D. Su, “On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy,” in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. ACM, 2012, pp. 32–33.
16. Li N, *et al.* t-Closeness: privacy beyond k-anonymity and L-diversity. In: Data engineering (ICDE) IEEE 23rd international conference; 2007.

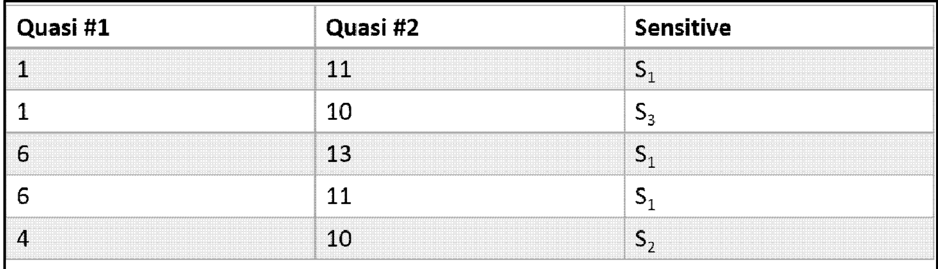
17. Ton A, Saravanan M. Ericsson research. [Online]. http://www.ericsson.com/_research-blog/data-knowledge/big-data-privacy preservation/2015.
18. S. Ni, M. Xie, and Q. Qian, "Clustering based k-anonymity algorithm for privacy preservation," *Int. J. Netw. Secur.*, 2017, doi: 10.6633/IJNS.201711.19(6).23.
19. K. Arava and S. Lingamgunta, "Adaptive k-Anonymity Approach for Privacy Preserving in Cloud," *Arab. J. Sci. Eng.*, 2020, doi: 10.1007/s13369-019-03999-0.
20. S. Khan, K. Iqbal, S. Faizullah, M. Fahad, J. Ali, and W. Ahmed, "Clustering based privacy preserving of big data using fuzzification and anonymization operation," *Int. J. Adv. Comput. Sci. Appl.*, 2019, doi: 10.14569/ijacsa.2019.0101239.

Table 1 illustrates the sources of the datasets as well as the outcomes of the mapper.



S. No.	Source of the data	Outcome of the mapper	
1	[1:5][8:9]	1, {S_1, 1}, 1	11, {S_1, 1}, 1
2	[1:5][8:9]	1, {S_1, 1}, 1	10, {S_3, 1}, 1
3	[1:5][8:9]	6, {S_1, 1}, 1	13, {S_1, 1}, 1
4	[1:5][8:9]	6, {S_3, 1}, 1	11, {S_1, 1}, 1
5	[1:5][8:9]	4, {S_2, 1}, 1	10, {S_2, 1}, 1

Table 2 Illustrates the grouping of the outcome data of Table 1 as likeness class.



Quasi #1	Quasi #2	Sensitive
1	11	S_1
1	10	S_3
6	13	S_1
6	11	S_1
4	10	S_2

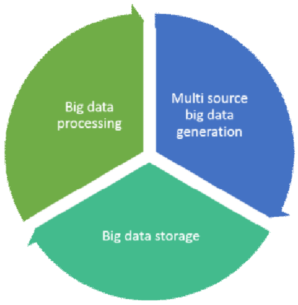


Fig. 1Illustrates the diverse phases of the life-cycle of the big datasets, for instance, generation of the datasets, storage as well as processing.

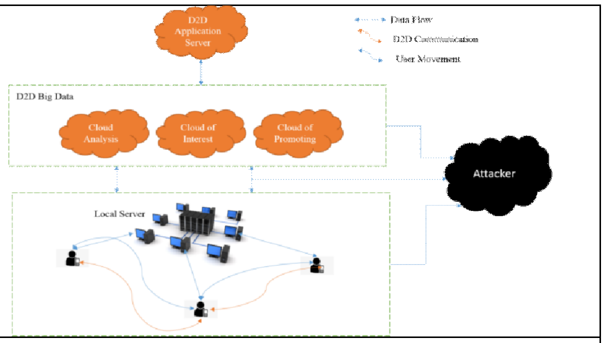


Fig. 2 Illustrates the big datasets framework as well as threat prototypical for the D2D (Device to Device) communication for privacy-preserving.

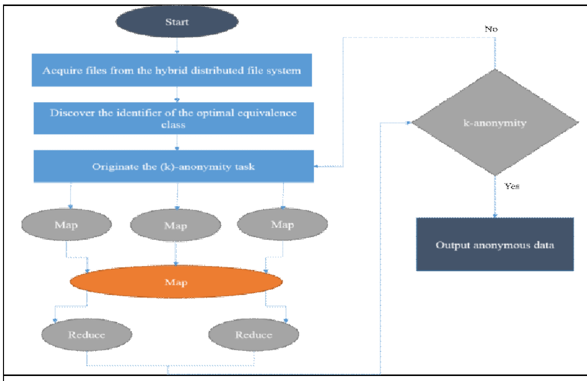
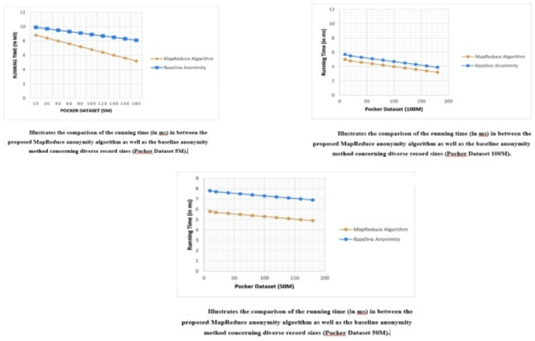

Start
Acquire files from the hybrid distributed file system
Discover the identifier of the optimal equivalence class
Originate the (k)-anonymity task
No
k-anonymity
Yes
Output anonymous data
Map
Map
Map
Map
Reduce
Reduce

Fig. 3. Illustrates the suggested (k)-anonymity set of the rules-based framework using MapReduce.

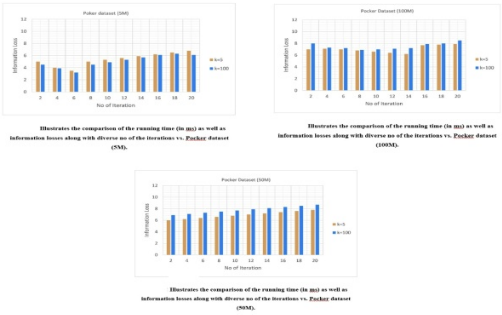


Illustrates the comparison of the running time (in ms) in between the proposed MapReduce anonymity algorithm as well as the baseline anonymity method concerning diverse record sizes (Pocker Dataset 5M).

Illustrates the comparison of the running time (in ms) in between the proposed MapReduce anonymity algorithm as well as the baseline anonymity method concerning diverse record sizes (Pocker Dataset 100M).

Illustrates the comparison of the running time (in ms) in between the proposed MapReduce anonymity algorithm as well as the baseline anonymity method concerning diverse record sizes (Pocker Dataset 50M).

Fig. 4. **Illustrates the comparison of the running time (in ms) in between the proposed MapReduce anonymity algorithm as well as the baseline anonymity method concerning diverse record sizes (Pocker Dataset 5M, 50M and 100M)**

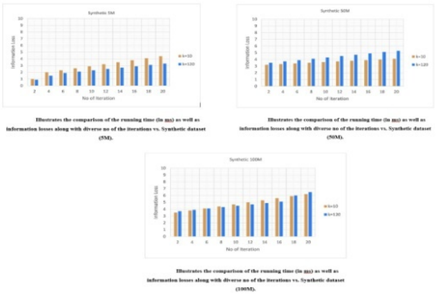


Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Pocker dataset (5M).

Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Pocker dataset (100M).

Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Pocker dataset (50M).

Fig. 5 Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Pocker dataset (5M, 50M and 100M).



Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Synthetic dataset (5M).

Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Synthetic dataset (50M).

Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Synthetic dataset (100M).

Fig. 6 Illustrates the comparison of the running time (in ms) as well as information losses along with diverse no of the iterations vs. Synthetic dataset (5M, 50M and 100M).