INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Novel algorithm for efficient privacy preservation in data analytics

P Ram Mohan Rao[1]*, S Murali Krishna[2], A P Siva Kumar[3]

1 Department of Computer Science and Engineering, JNTUA, Anantapuramu, Andhra Pradesh, India
2 Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Tirupathi, Andhra Pradesh, India
3 Department of Computer Science and Engineering, JNTUA, Anantapuramu, Andhra Pradesh, India

OPEN ACCESS

Received: 29.09.2020
Accepted: 12.01.2021
Published: 19.02.2021

Citation: Ram Mohan Rao P, Murali Krishna S, Siva Kumar AP (2021) Novel algorithm for efficient privacy preservation in data analytics. Indian Journal of Science and Technology 14(6): 519-526. https://doi.org/10.17485/IJST/v14i6.1773

* **Corresponding author**.

rammohan04@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To address the modern privacy threats in data analytics by designing an efficient privacy preserving data analytics technique. **Methods**: The method applied is a non anonymized method that uses the concepts of synthesizing quasi identifiers and application of differential privacy. The proposed method was applied to three data sets viz. Adult data set, Statlog data set and Indian Liver Patient data set. All the data sets are freely available in the UCI repository. **Findings**: The study presents "Synthesize Quasi Identifiers and apply Differential Privacy" (SQIDP) which is proved to be a more efficient and scalable algorithm. Compared to anonymity based algorithms SQIDP is not prone to similarity attacks, background knowledge attacks, attribute disclosure, and inference attacks. Anonymization, cryptographic, SWARM, and randomization methods will reduce data utility whereas SQIDP offers 100% data utility. Hence it is more efficient than other techniques. SQIDP was applied on three different data sets with 270, 583, and 48842 records but the execution time of the algorithm remained the same for all three data sets. SQIDP is proved to be a better privacy preservation technique with 100% data utility because it is not anonymized that abides by the recommendation in many privacy legislations like GDPR (General Data Protection Regulation) of the European Union and PDP (Personal Data Protection bill) of India.

Keywords: Data privacy; privacy regulations; privacy preservation; synthetic data; differential

1 Introduction

The majority of the privacy preservation methods developed in the past were based on anonymization techniques which will reduce data utility [(1)]. Cryptographic techniques were also proved to be inadequate especially when the data is voluminous. Swarm based anonymization techniques for privacy preservation is a recent development in the field of study but privacy legislations recommend non-anonymization based solutions for ensuring data utility [(2)]. Application of K-anonymity together with perturbation techniques is also studied but suffers from the data utility problem [(3)]. Data mining

techniques are also used for privacy preservation to overcome similarity and inference attacks with an improved trade-off between data utility and data anonymization[4]. However, to achieve maximum data utility, a non anonymized solution is preferred. In this paper, we examined the key aspects of privacy legislations, modern privacy threats and proposed a privacy preservation algorithm called SQIDP to offer privacy preservation in data analytics. The key features of the algorithm and the main contributions of the study are listed below.

1. A non-anonymization-based solution to privacy preservation problem as recommended in GDPR.
2. Identity disclosure and attribute disclosure is not possible, because the quasi identifiers are synthesized and cannot be mapped with external data sources.
3. Sensitive data is tokenized before analytics, differential privacy is applied on synthetic data which will prevent background knowledge and homogeneity attack.
4. Strong and coherent privacy protection is guaranteed because the original data set is not involved in data analytics and instead a synthetic data set is used which is statistically similar to the original data set such that the analytical results of synthetic data can be related or mapped to the original data set.

2 Modern privacy threats and concerns

The nature of privacy threats has changed due to the emergence of applications like recommendation systems, e-commerce, etc. Conventional data analysis included a statistical analysis of data especially using aggregate queries where data was analyzed as a whole[5,6]. Applications like recommendation systems will analyze personal data like buying habits, social media posts and try to predict suitable recommendations that are possible only through constant surveillance. Recommendation systems may lead to the disclosure of sensitive data leading to personal embarrassment and inference attacks. Another important source of privacy breach is the usage of smartphones.[7] Most of the smartphone apps demand permissions to access network, location, contacts, and storage which can be shared by the app developer with third parties and adversaries causing serious privacy breaches and threats to sensitive data. The percentage of users aware of privacy threats of using smartphone apps is very less[8]. The nature of privacy threats is changing every day and some of the modern privacy threats include.

1. Digital Profiling
2. Social media privacy and cyberstalking
3. Image analytics and privacy hazards

2.1 Digital profiling

Digital Profiling is the automated processing of person-specific data to evaluate certain attributes relating to a person, particularly to analyze and predict an individual's economic situation, buying habits, health, preferences, interests, behaviour, etc. Digital Profiling also influences group privacy wherein an individual may be a member of one or more groups[9]. Digital Profiling is widely used in direct digital marketing businesses. Profiling without the consent of the individual is a privacy breach. Google has recently announced to end support for third-party cookies in its Chrome browser which will make it very difficult for digital marketing companies to build a user profile.[10]. Article 22 of GDPR facilitates the right to the individual that, no automated data processing including profiling is allowed without consent from the user.

2.2 Social media privacy and cyber stalking

Social media platforms are highly vulnerable to stalking attacks. One of the common stalking techniques involves an online mob of anonymous self-organized groups to target individuals causing defamation, threats of violence, and technology-based attacks. Social media are used to build trust between the perpetrator and the victim. When the victim transmits confidential data including pictures and videos, the perpetrator abuses them for blackmail purposes[11]. Social media firms are also responsible to identify the user with malicious intentions.

2.3 Image data and privacy hazards

Image data analytics is widely used in health care, social media, and e-commerce applications. In social media applications like Facebook and Instagram, users upload a lot of images every day. An image is worth more than a thousand words and hence it may reveal the emotional state of a person[12]. Some of the key privacy hazards in image data analytics include

1. Attempt to analyse the emotional state of people and exploit them. Facebook and Whatsapp status updates can be studied using machine learning models and sentiment analysis can help analyse the social and emotional wellbeing of a person and in turn, exploit them.
2. Disclosure of secret medication being taken by a person by virtue of promotional offers on medicine.
3. Another important privacy concern is identity theft because copies of permanent account number (PAN) cards, passports and driving licenses are kept in digital form and shared. Insurance and banking firms and third parties will extract a lot of sensitive data which is a serious privacy hazard [13,14].
4. Medical imaging deals with a visual representation of the internal structure of organs and tissues. Medical imaging may lead to leakage of personal and sensitive medical data of a person. [15]

3 Methods

3.1 SQIDP algorithm (Synthesize Quasi Identifier and apply Differential Privacy)

Data Privacy has gained paramount importance in recent times and it is evident from the privacy legislation passed in more than 100 countries. Firms dealing with data sensitive applications need to abide by the privacy legislation of respective regions. In the recent past, a lot of promising work has been done in privacy preserving data analytics. Swarm based algorithms were also applied to the data sets alongside perturbation techniques. Swarm based algorithm developed for privacy preservation uses k-anonymity as the building block. Even though swarm algorithms are promising, they suffer from the traditional flaws of anonymization [2]. Researchers have tried to apply a map reduce framework to process data sets using a perturbation mechanism along with probabilistic anonymity [3]. However, the application of anonymization is not recommended in privacy legislation since it reduces the data utility and hence non anonymized solution has to be designed. SQIDP is a non anonymized technique where the original data set D is transformed into D' without any anonymization and the new data set D' is used for analytics. Step by Step procedure to generate D' from D is described in sections 3.2, 3.3and 3.4.

3.2 Synthetic data in data analytics

Synthetic Data is one of the data sanitization methods where original data is replaced with synthetic data ensuring privacy preserving data analytics. Data can be fully synthetic or partial and various types of synthetic data generation methods were studied and compared in the previous literature [16]. Synthetic data can be a generative and additive approach for generating a near replica of original data but care must be taken to ensure the reliability of the synthetic data and also data utility must not be reduced. Quasi Identifiers are the attributes that can be linked with external data sources that may reveal sensitive data and therefore they are synthesized to prevent linkage attacks. The statistical similarity between quasi identifiers and synthetic quasi identifiers is ensured by generating synthetic data with close statistical properties of original quasi identifiers.

As part of our research, we employed a novel algorithm called SQIDP in which quasi identifiers (QI) are synthesized, sensitive attribute(s) are tokenized, and finally, differentially privacy is applied to generate a new dataset from the original data set.

Algorithm: SQIDP
1. Start
2. Given a Dataset D with attributes D{a1,a2,a3...an}
3. Choose Quasi identifiers (QI) example **QID{a3,a4,a5}**
4. for each QI
5. do
6. Synthesize each QI using **rnorm** function to generate
7. rnorm (column size, desired mean, desired standard dev.) to create synthetic data for QI.
Example. SQID {a3', a4', a5'}.
8. end for
9. Tokenize the sensitive attribute (SA).Example SA {a6}. In tokenization each discrete value of the attribute is replaced with a token.
10. Merge non quasi identifiers of D, SQID and tokenized SA to generate new data set D'.
11. End

3.3 Application of SQIDP algorithm:

The algorithm was initially applied on the adult dataset, downloaded from the University of California, Irvine (UCI) machine learning repository [17]. The Quasi Identifiers are fnlwgt, age, capital-gain, and capital-loss attributes. The data set size was 32561 records. Initially, we found the mean and standard deviation of fnlwgt attribute as 189778.4 and 105550 respectively. "rnorm" is a function in R used to generate multivariate random values that are normally distributed. Using rnorm we have generated synthetic data for fnlwgt attribute which is referred to as fnlwgt_syn. The mean and standard deviation of fnlwgt_syn is 189777 and 105551 respectively which is very close enough to the mean and standard deviation of the original attribute. The fnlwgt_syn is shown in Figure 1.

Pseudo code: To generate synthetic data for Quasi Identifiers (R language)

Data set used: http://archive.ics.uci.edu/ml/datasets/Adult

```
adult<-read.csv("adult.csv", header = TRUE)
adult_synthetic<-adult
print(mean(adult$fnlwgt))
print(sd(adult$fnlwgt))
fnlwgt_syn <- rnorm(32561,189778.9,10555.5)
fnlwgt_syn
print(mean(adult$capitalgain))
print(sd(adult$capitalgain))
capitalgain_syn <- rnorm(32561,1078,7385.8)
```

Fig 1. Pseudo code for synthetic data generation

Figure 1 analysis: Synthetic data is generated for all the quasi identifiers. The above figure depicts synthetic code generation using the rnorm function which simulates random variates having special normal distribution by considering mean and standard deviation of the attribute values. "rnorm" function accepts three parameters viz. number of values, mean and standard deviation. Figure 1 shows how synthetic data for the attributes fnlwgt and capital gain are generated. Similarly, the remaining quasi identifiers are also synthesized.

The attribute marital-status is the sensitive attribute (SA) which is tokenized. Marital-status enumerates {Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed}. The SA attribute is tokenized using a numerical vector. The new data set D' is created by combining the non quasi identifiers of D, SQID, and tokenized sensitive attribute (SA). D' is released instead of D for data analytics. D' contains synthetic data that has a very close resemblance with original data but it is not the original data. The mathematical transformations done on the quasi identifiers (QI) will ensure that the analytical results of D' can be applied on D without releasing the original data set. The comparison of synthesized attributes in D' with original values in D is shown in Figure 2.

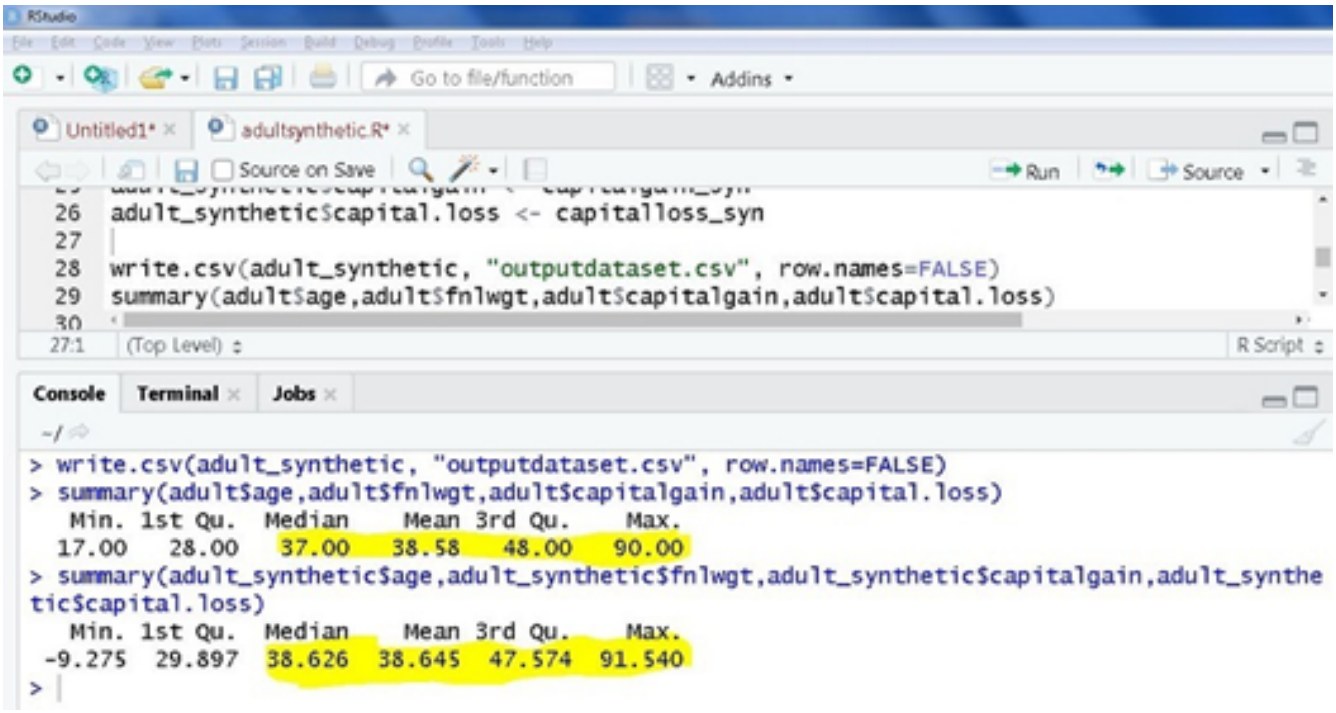

Fig 2. Summary of D and D' representing statistical properties of quasi identifiers and synthetic quasi identifiers

Figure 2 analyses: The mean, median values of the datasets show that the statistical properties of both D and D' are almost the same which are highlighted in Figure 2. Linkage attacks were possible in most anonymized methods where the quasi identifiers can be mapped with external data sets. In SQIDP algorithm D' contains synthesized values of quasi identifiers that cannot be mapped to external data sets and linkage attacks can be prevented. Since D' and D have close statistical properties like mean and standard deviation, the results of data analytics made on D' can be applied back on to original data set D. The same process was repeated to create two more partially synthetic datasets viz. Indian Liver patients data and Statlog heart data set which is described in Table 1.

Table 1. Datasets description

S.no	Data set name	No. of attributes	No. of records
1	Adult Data set	14	48842
2	Statlog Data set	13	270
3	Indian Liver Patient records	11	583

All the datasets are available for free at UCI ML repository https://archive.ics.uci.edu/ml/datasets/.

3.4 Application of differential privacy

In Section 3 we have demonstrated the generation of partially synthetic data with strategic changes made to quasi identifiers. The dataset thus generated (D') can be released for analytics and the results can be applied back to the original data set. However, to make the dataset more robust to privacy attacks, an additional differential privacy algorithm is employed on D'. Laplace mechanism of differential privacy is applied on D' to generate a differentially private data set which makes it very difficult to predict whether an individual record was present in the data set or not. Package diffpriv[18] contains an implementation of different mechanisms of differential privacy. Laplace mechanism is one of the differential privacy mechanisms where Laplace noise is added to the dataset using Laplace distribution which is the probability density function.

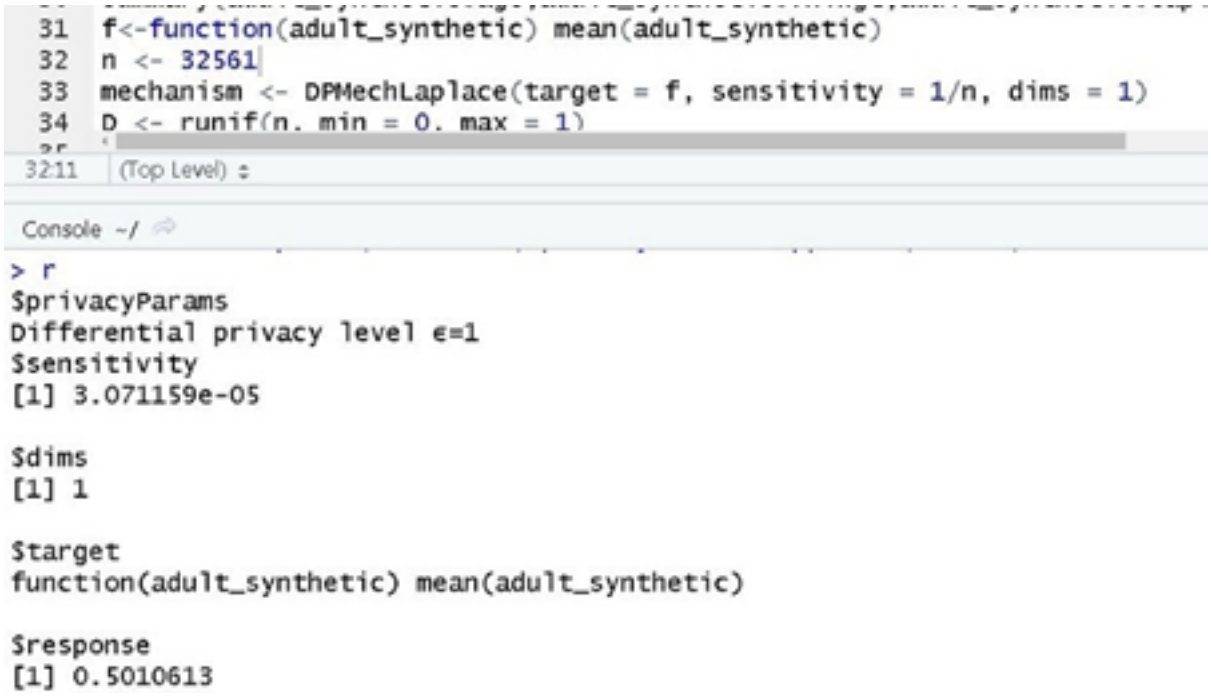

Fig 3. After application of Laplace Mechanism of Differential Privacy on D'

Figure 3 analysis: After synthesizing quasi identifiers and tokenizing the sensitive attribute, the data set adult_synthetic is passed to a function that applies the Laplace mechanism of differential privacy on the data set D'. The idea is to add enough noise to hide the contribution of any individual irrespective of the dataset. It is difficult to predict whether a single person is in the dataset or not if the dataset is differentially private.

4 Results

In SQIDP, the quasi identifiers were replaced with synthetic data generated using random variates having specified normal distribution. The mean and standard deviation of the synthetic data will be very close to the mean and standard deviation of the original quasi identifiers. This will ensure the results of data analytics on synthetic data can be mapped to original data sets.

Advantages of SQIDP:

1. The execution time of SQIDP was same on all three data sets with different sizes and hence it is scalable.
2. SQIDP addressed the background knowledge attack and homogeneity attack because the quasi identifiers are synthesized and cannot be mapped with any external data sources.
3. SQIDP is a non anonymized method and offers 100% data utility.

5 Discussions

SQIDP is an innovative method of privacy preservation where quasi identifiers are synthesized to ensure no scope of linkage attacks which was a common problem in previous privacy preservation techniques. SQIDP is found to be more efficient than existing techniques and a detailed comparison is given in Table 2 and performance metrics of SQIDP are listed in Table 3. Figure 4 illustrates the comparative graphic analysis of three datasets along with execution time of SQIDP which is found to be uniform across all datasets.

Table 2. Comparison of SQIDP with other privacy preservation techniques

Techniques Features	K anonymity	Cryptographic techniques	Randomization	SWARM Based techniques	Multi Dimensional Sensitivity Based Anonymization (MDSBA)	SQIDP
Attribute disclosure and linkage attacks	Vulnerable	Not Vulnerable	Vulnerable	Not Vulnerable	Vulnerable	Not Vulnerable
Background knowledge attack	Vulnerable	Not Vulnerable	Vulnerable	Not Vulnerable	Vulnerable	Not Vulnerable because of synthesis of Quasi Identifiers and tokenization of sensitive attribute
100% Data Utility	No	No	No	No	No	Yes
Scalability	No	No	No	No	Yes	Yes

Table 3. Performance Metricsof SQIDP

S.no.	Performance Metric	Description
1	Data Utility	SQIDP offers 100% data utility because it is non anonymized.
2	Robust	SQIDP is robust because it is not vulnerable to Linkage attacks - because of differential privacy Background knowledge attack - because of synthetic quasi identifiers Attribute and Identity disclosure - because of tokenization of sensitive attribute.
3	Compliance	SQIDP complies to privacy regulations and does not use anonymization as recommended in GDPR.
4	Execution time	All the privacy preserving techniques including SQIDP will have O(n) time complexity. However, SQIDP can be executed in a distributed computing platform to gain better performance.
5	Accuracy	Anonymization leads to data loss and in turn affects the results of the analytics. SQIDP is a non anonymized and hence offers accurate analytics.

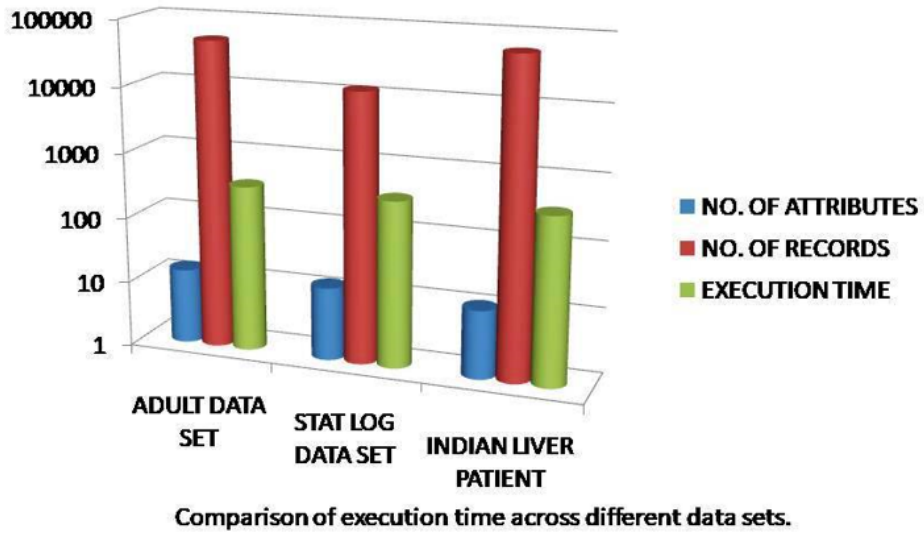

Fig 4. SQIDP execution time on three different datasets.

Even though Differential privacy is employed in US Census 2020 [19], Google Chrome and iOS 11, etc., differential privacy alone cannot guarantee privacy preservation because differential privacy has its limitations. Differential privacy will fail in a few aggregate queries performed on the data. For example, if we want to find the average salary of all woman employees of an

organization and if there is an employee with a very high salary whose presence or absence in the data makes a significant change in the average. In such cases, a huge amount of noise has to be added to ensure privacy but excess noise may affect the data utility. Another important observation is differential privacy cannot handle background knowledge attacks. If the adversary is aware of certain information about a person, then his presence or absence in the data set does not mean anything. SQIDP is a more efficient privacy preservation technique when compared to conventional anonymization techniques, randomization techniques, and cryptographic techniques. Anonymization techniques will reduce data utility whereas SQIDP does not suffer from data utility and it is also in line with GDPR. SQIDP is more efficient when compared to cryptographic techniques because the application of cryptographic techniques adds processing overhead and also declines data utility. Differential privacy has its own limitations and Differential privacy alone cannot address the privacy threats involved in data analytics. SQIDP has proved to be efficient than Differential privacy alone because synthesizing quasi identifiers and tokenizing sensitive attribute will prevent background knowledge attacks. Irrespective of any number of queries, there is no chance of any privacy breach which was noticed in the application of Differential privacy. Hence SQIDP is an efficient mechanism of privacy preservation in data analytics and a useful contribution to the field of privacy preserving data analytics.

6 Conclusions

The SQIDP algorithm can be applied only to text data by ensuring privacy preservation and protection from background knowledge attack and linkage attacks. SQIDP is a useful contribution to the field of privacy preserving data analytics that ensures data utility along with privacy preserving data analytics. However, SQIDP is limited to text data and cannot be applied to image or video data. Extensive usage of social media has led to the creation of a huge amount of image and video data that are prone to various cyber security vulnerabilities and have enough research scope.

References

1) Madan S. A Literature Analysis on Privacy Preservation Techniques. In: Advances in Computing and Intelligent Systems. Springer. ;p. 2020–223. Available from: https://doi.org/10.1007/978-981-15-0222-4_19.
2) Madan S, Goswami P. A privacy preservation model for big data in map-reduced framework based on k-anonymisation and swarm-based algorithms. *International Journal of Intelligent Engineering Informatics*. 2020;8:38–53. Available from: https://doi.org/10.1504/IJIEI.2020.105433.
3) Eyupoglu C, Aydin M, Zaim A, Sertbas A. An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques. *Entropy*. 2018;20(5). Available from: https://dx.doi.org/10.3390/e20050373.
4) Nayahi JJV, Kavitha V. Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop. *Future Generation Computer Systems*. 2017;74:393–408. Available from: https://dx.doi.org/10.1016/j.future.2016.10.022.
5) Rao PRM, Krishna SM, Kumar APS. Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data*. 2018;5(1). Available from: https://dx.doi.org/10.1186/s40537-018-0141-8.
6) Tang E. A quantum-inspired classical algorithm for recommendation systems. In: and others, editor. Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing. 2019. Available from: https://doi.org/10.1145/3313276.3316310.
7) Zhou Y, et al. User attitudes and behaviors toward personalized control of privacy settings on smartphones. *Concurrency and Computation: Practice and Experience*. 2019;31(22). Available from: https://doi.org/10.1002/cpe.4884.
8) Barth S, et al. Putting the privacy paradox to the test: Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources. *Telematics and informatics*. 2019;41:55–69. Available from: https://doi.org/10.1016/j.tele.2019.03.003.
9) Mavriki P, Karyda M. Automated data-driven profiling: threats for group privacy. *Information & Computer Security*. 2019;28(2):183–197. Available from: https://dx.doi.org/10.1108/ics-04-2019-0048.
10) Google chrome to end support for third party cookies within two years. . Available from: https://www.cnbc.com/2020/01/14/google-chrome-to-end-support-for-third-party-cookies-within-two-years.html.
11) March E, Litten V, Sullivan DH, Ward L. Somebody that I (used to) know: Gender and dimensions of dark personality traits as predictors of intimate partner cyberstalking. *Personality and Individual Differences*. 2020;163. Available from: https://dx.doi.org/10.1016/j.paid.2020.110084.
12) Chen L, Gong T, Kosinski M, Stillwell D, Davidson RL. Building a profile of subjective well-being for social media users. *PLOS ONE*. 2017;12(11). Available from: https://dx.doi.org/10.1371/journal.pone.0187278.
13) Yang J, Wu J, Wang X. Convolutional neural network based on differential privacy in exponential attenuation mode for image classification. *IET Image Processing*. 2020. Available from: https://dx.doi.org/10.1049/iet-ipr.2020.0078.
14) Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circulation: Cardiovascular Quality and Outcomes*. 2019;12(7). Available from: https://dx.doi.org/10.1161/circoutcomes.118.005122.
15) Wang P, Chen T, Wang Z. Research on privacy preserving data mining. *Journal of Information Hiding and Privacy Protection*. 2019;1(2):61–68. Available from: https://doi.org/10.32604/jihpp.2019.05943.
16) Walters A. . Available from: https://patents.google.com/patent/US20200012902A1/en.
17) Dua D, Graff C. 2019. Available from: http://archive.ics.uci.edu/ml.
18) Rubinstein B, Ip F, Alda. . Available from: https://cran.r-project.org/web/packages/diffpriv/vignettes/diffpriv.pdf.
19) Abowd JM. The US census bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.