Received: 19 October 2019 | Revised: 27 July 2020 | Accepted: 28 July 2020

DOI: 10.1002/widm.1387

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

WILEY

Privacy preserving big data analytics: A critical analysis of state-of-the-art

M. Ileas Pramanik[1] | **Raymond Y. K. Lau**[2] | **Md Sakir Hossain**[3] | **Md Mizanur Rahoman**[1] | **Sumon Kumar Debnath**[4] | **Md Golam Rashed**[5] | **Md Zasim Uddin**[1]

[1] Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh

[2] Department of Information Systems, City University, Hong Kong, China

[3] Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh

[4] Department of Electrical and Electronics Engineering, Begum Rokeya University, Rangpur, Bangladesh

[5] Department Information and Communication Engineering, Rajshahi University, Rajshahi, Bangladesh

Correspondence

M. Ileas Pramanik, Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh.
Email: mpramanik2-c@my.cityu.edu.hk

Funding information

BRUR Research Project, Grant/Award Number: 2019-20/40; the CityU Shenzhen Research institute; the NSFC Basic Research Program, Grant/Award Number: 71671155; the Research Grants Council of the Hong Kong Special Administrative Region, Grant/Award Number: CityU 11525716

Abstract

In the era of "big data," a huge number of people, devices, and sensors are connected via digital networks and the cross-plays among these entities generate enormous valuable data that facilitate organizations to innovate and grow. However, the data deluge also raises serious privacy concerns which may cause a regulatory backlash and hinder further organizational innovation. To address the challenge of information privacy, researchers have explored privacy-preserving methodologies in the past two decades. However, a thorough study of privacy preserving big data analytics is missing in existing literature. The main contributions of this article include a systematic evaluation of various privacy preservation approaches and a critical analysis of the state-of-the-art privacy preserving big data analytics methodologies. More specifically, we propose a four-dimensional framework for analyzing and designing the next generation of privacy preserving big data analytics approaches. Besides, we contribute to pinpoint the potential opportunities and challenges of applying privacy preserving big data analytics to business settings. We provide five recommendations of effectively applying privacy-preserving big data analytics to businesses. To the best of our knowledge, this is the first systematic study about state-of-the-art in privacy-preserving big data analytics. The managerial implication of our study is that organizations can apply the results of our critical analysis to strengthen their strategic deployment of big data analytics in business settings, and hence to better leverage big data for sustainable organizational innovation and growth.

This article is categorized under:

Commercial, Legal, and Ethical Issues > Security and Privacy
Fundamental Concepts of Data and Knowledge > Big Data Mining
Fundamental Concepts of Data and Knowledge > Data Concepts

KEYWORDS

big data, business analytics, information privacy, privacy preservation

1 | INTRODUCTION

Privacy-preserving big data analytics has received much attention from academic and industrial communities. The problem of privacy preservation in data mining has become increasingly prevalent because of the wide spread storage of electronic documents on the Internet, the explosive growth of digitized signals generated from sensor networks, the rapid generation of transactional data via online-to-offline commerce, and the proliferation of user-contributed dialogs in the social web. Given that the huge volume of digitized data (e.g., user online profiles, check-in data, Global Positioning System signals of location-based e-commerce, etc.) often capture users' personal information, the risk of breaching user privacy while applying business analytics to big data has become an increasingly serious concern (Li, 2014; Torra & Navarro-Arribas, 2014). Thus, various techniques, such as randomization, k-anonymity, and distributed privacy preserving, have been proposed to ensure appropriate data protection while applying business analytics to digitized data. In the past few years, researchers have responded to the big data challenge by developing some new models and algorithms for privacy-preserving big data analytics. However, a systematic and thorough study about the existing privacy-preserving approaches in big data analytics is missing in the literature. Sound principles and guidance are lacking for empirical researchers and industrial practitioners who want to apply privacy-preserving big data analytics to business settings. In this article, we perform a systematic and critical review on privacy-preserving methodologies in general and privacy-preserving big data analytics in particular to fill the current research gap.

This article makes important contributions to existing theories and practice of privacy-preserving data analytics in fourfold. First, we conduct a critical analysis toward different paradigms of privacy-preserving data mining techniques in general and a systematic study about state-of-the-art privacy-preserving methodologies in big data analytics in particular. Second, we propose a novel four-dimensional (4D) framework for the systematic evaluation and effective design of the next generation of privacy-preserving methodologies in big data analytics. Third, we identify the potential opportunities and challenges of applying privacy-preserving big data analytics to business settings. Finally, we provide five recommendations to guide empirical researchers and industrial practitioners who are involved in the governance of new initiatives related to privacy-preserving big data analytics. To the best of our knowledge, this systematic study is the first to discuss about state-of-the-art methodologies in privacy-preserving big data analytics. The managerial implication of this study is that organizations can apply the results of the critical analysis to strengthen their strategic deployment of big data analytics in business settings and to achieve sustainable organizational innovation and growth.

The article is organized as follows. Section 2 provides a thorough literature review on privacy-preserving data mining methods in general and privacy-preserving big data analytics in particular. Section 3 performs a critical analysis of the pros and cons of state-of-the-art privacy-preserving approaches in big data analytics. Section 4 proposes a novel 4D framework for evaluating and designing the next generation of privacy-preserving methodologies in big data analytics. Section 5 explores the opportunities and challenges of applying privacy-preserving big data analytics to business settings and provides guidance to empirical researchers and industrial practitioners who are involved in privacy-preserving projects in big data analytics. Section 6 provides the conclusions and summarizes the directions for future research work.

2 | LITERATURE REVIEW

2.1 | Privacy preservation data mining methods

A US-based research project (2014) no. 708 demonstrated that firms are investing millions of dollars for their privacy unit to assure the protection of business data. A survey was conducted by TRUSTe among 200 privacy professionals in various companies with more than 1,000 employees. Results showed that 30% of organizations allocated more than $1 million budget in privacy concern in 2014, and 71% of respondents stated that data privacy management is "very important" or "important" for their company. Privacy issues are highly concerned in all aspects of information utilization, such as data gathering and preserving and data analysis and release (Duan & Canny, 2014). For preserving privacy, some data mining privacy methods are deployed in big data area. Most methods for privacy computation use some form of transformations of the data to achieve privacy preservation. In this section, we describe different strategies for preserving data privacy.

2.1.1 | Cryptography

The cryptography-based technique usually guarantees very high-level data privacy because it provides strong and powerful primitives. These primitives can be applied to the design of information security systems, which are rigorously proven as a strong privacy method either unconditionally or under some rational assumptions (Duan & Canny, 2014). Modern cryptography provides other important security tools, such as secure multiparty computation (MPC) and zero-knowledge proof, which are considered as privacy proof mechanisms (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006). In an MPC, suppose some participants p_1, p_2, …, p_m, where each of them has private data d_1, d_2, …, d_m. Participants want to compute the value of a public function f on m variables at the point (d_1, d_2, …, d_m). An MPC protocol is secure if any participant cannot gain information more from the description of the public function and the result of the global calculation that anyone can gain from own entry under particular conditions depending on the model used (Lindell, 2005). Here, the privacy of clients' inputs and the correctness of the outputs are guaranteed although some players are corrupted by the same adversary.

2.1.2 | Statistical privacy

Privacy study recommends that query responses need to be perturbed by random noise with sufficient variance to maintain privacy. Different statistical learning algorithms, such as support vector machine (SVM), decision tree learning, singular value decomposition (SVD), and k-means, are proposed for statistical data privacy preservation. These algorithms and all algorithms in the statistical query model are based on expectation–maximization (EM; Machanavajjhala et al., 2006).

2.1.3 | Existing privacy methods

Various types of privacy-preserving methods, such as randomization, anonymization (k-anonymity, l-diversity, t-closeness), partition-based privacy, and differential privacy methods, are commonly used to solve the problem of de-identification. All these methods have different vulnerabilities, and researchers are continuing their research for updating them to adopt contemporary data. The computational details of different privacy preservation techniques are described as follows:

Randomization method

For the randomization method, noisy signals are introduced to the plain data to hide the actual values of the individual records. The added noise in the data is significantly large to protect the individual value of the records. However, the aggregate behavior of the data distribution can be reconstructed by subtracting the noise from the data. The reconstructed distribution is often sufficient for various data mining tasks. The addition of A and B creates a new distribution C. Given that the distribution of B is publicly known, we can estimate the distribution obtained by subtracting B from C.

Thus, we have $C = A + B$; then, $A = C - B$.

In two studies, a pair of closely related iterative methods were discussed to approximate the corresponding probability distributions (Agrawal & Aggarwal, 2001; Agrawal & Srikant, 2000). The first study used the Bayes rule for distribution approximation (Agrawal & Srikant, 2000), whereas the second study used the EM method for distribution approximation (Agrawal & Aggarwal, 2001). Such iterative methods typically have a higher accuracy than the sequential solution of first approximating C and then subtracting B from it. The above-mentioned technique is an additive strategy for randomization. Another strategy called multiplicative strategy is used for randomization, where the records with random vectors can be multiplied to provide the final representation of the data. In the randomization approach, all records are not preserved in equal rank of privacy. The outlier records are more susceptible to adversarial attacks compared with records in denser regions in the data (Aggarwal, 2007). To prevent this condition, one may need to be unnecessarily aggressive in adding noise to all the records in the data that will lead to information loss. This loss usually reduces the utility of the data for mining purposes. Table 1 shows the comparison between different studies on randomization.

TABLE 1 Comparison between different randomization-based techniques proposed by different studies

Study	Main feature	New software framework	Scalability	Privacy breach/ effectiveness analysis	Privacy model	Data distortion	Time efficiency	Considerable data type	Precise scheme	Comparison with existing approach	Experiment on real datasets	Security operation /applied methods	Example application
Agrawal and Aggarwal (2001)	Design and quantification of privacy preserving data mining algorithms	No	N/A	No	Perturbation approach	Moderate	N/A	Arbitrary data	Privacy metrics	Yes	No	Uniform and Gaussian distribution	No
Aggarwal (2007)	Comprehensive analysis of the randomization method in the presence of public information	Yes	High	Yes	Randomization with perturbation distribution	High but not suppress	High	Arbitrary but uniformly distributed data	Randomization approach	Yes	Yes	Uniform and Gaussian distribution	No
Huang, Du, and Chen (2005)	How much correlations among the attributes can help disclose private information	No	High	Yes	Randomization by using correlated random noise	High	High	Categorical data	UDR,PCA-DR,BE-DR, and SF	Yes	Yes	Data reconstruction method	No
Kargupta, Datta, Wang, and Sivakumar (2003)	Develop a spectral filter for extracting the hidden data from the perturbed data	No	N/A	Yes	Perturbation approach	N/A	High	Discrete and continuous data	Spectral filtering	Yes	Yes	Random matrix theory	No
Evfimievski (2002)	Randomized private information in terms of monetary value.	No	Moderate level	Yes	Randomization model	Moderate	High	Numerical and categorical data	Randomization, sampling and swapping	No	No	Additive randomization	No
Liu, Kargupta, and Ryan (2005)	Use of random projection matrices as a tool for privacy preserving data mining	Yes	High	Yes	Perturbed by inner matrix product model	High	High	Boolean and discrete data	Value distortion approach	Yes	Yes	Multiplicative perturbation based model	Yes
Kim and Winkler (2003)	Prove multiplying noise better protect the confidentiality	No	Moderate	No	Masking scheme	Low	N/A	Continuous data	Logarithmic transformation, and noise multiplication	Yes	Yes	Multivariate normal distribution	No
Oliveira and Zaiane (2004)	Design a privacy preserving method named “rotation based transformation(RBT) method”	Yes	Low	No	Perturbation approach	High	Low	Discrete and real data set	Isometric transformation	No	Yes	Normalization, distortion and suppression.	No
Chen and Liu (2005)	Proposed a rotation-based perturbation technique	Yes (random iterative algorithm)	High	Yes	Multidimensional perturbation approach	High	High	Arbitrary datasets	Unified privacy model	Yes	Yes	Applied multi-column privacy metrics	No
Dalenius and Reiss (1982)	Propose a new technique for control the data disclosure.	Yes (mathematical framework)	Low	No	Swapping through statistical approach	Moderate	Low	Categorical data	T-order statistics	No	No	Swapping approach	No

2.1.4 | Vulnerabilities on randomization

In study (Huang et al., 2005; Kargupta et al., 2003), the spectral filtering or principal component analysis (PCA)-based techniques were proposed to reconstruct the distribution of the dataset. The broad idea in those techniques, such as PCA (Huang et al., 2005), is that the correlation structure in the original data can be estimated fairly and accurately (in big data) with the addition of noise. In these techniques, the noise removal results are extremely close to the original data (Huang et al., 2005; Kargupta et al., 2003). Thus, additive perturbation is vulnerable because those techniques can significantly reduce the privacy of perturbation. Adversarial attacks are significant attacks with the use of public information, where value reconstruction and subject identification are required. Consider a record $A = \{a_1,a_2,\ldots\ldots.a_n\}$ that is perturbed to $C = \{c_1,c_2,\ldots\ldots.c_n\}$. We can use a maximum likelihood fit of the potential perturbation of C to a public record by using the known value of distribution of the perturbations. Different perturbing distributions can have significant role on privacy-preserving data mining (Gambs, Kégl, & Aïmeur, 2007). The use of uniform perturbations is experimentally shown to be more effective in low-dimensional data, whereas time Gaussian perturbations are more effective for high-dimensional data. The work in study (Gambs et al., 2007) characterized the amount of perturbation required for a particular dimensionality with each type of perturbing distribution. The two types of perturbations tend to be less effective with increasing data dimensionality. Most perturbation mechanisms attain one of two goals, privacy and data utility, at the expense of the other. In study (Chamikara et al., 2019), a privacy-preserving algorithm for big data was proposed to bring a tradeoff between them. It uses Laplacian noise and Chebyshev interpolation to propose a Secure and Efficient data perturbation Algorithm utilizing local differential privacy.

2.1.5 | Applications of randomization

The randomization method has been extended to various data mining problems. Methods for privacy-preserving data mining over different classifiers have been proposed in a recent study (Zhu & Liu, 2004). A number of other techniques, such as association rules, have been proposed in other papers (Evfimievski, 2002). The randomization technique needs to be modified slightly by considering the quantification. In study (Aggarwal et al., 2010), the randomization approach was extended to other applications, such as online analytical processing and SVD-based collaborative filtering in a recent study (Polat & Du, 2005). A number of methods have been proposed for privacy-preserving classification with randomization. In study (Aggarwal et al., 2010), a method was discussed for decision tree classification with the use of the aggregate distributions reconstructed from the randomized distribution. They used an additive perturbation technique, where a random perturbation is added to the original value of the row, and the perturbation is picked from Gaussian probability distribution function. They showed that building accurate decision tree classification models is possible on the perturbed data and that aggregate queries on multiple columns can be used for privacy-preserving construction of decision trees. Various collaborative filtering methods have been proposed using randomized perturbed techniques in two different studies (Polat & Du, 2003; Polat & Du, 2005). The collaborative filtering problem is used in business intelligence systems. In the collaborative filtering problem, we aim to predict the ratings of products for a particular user through user ratings with similar profiles. Such ratings are useful for making recommendations that the user may like. Another method using randomized perturbation techniques for collaborative filtering based on SVD was proposed in a study (Polat & Du, 2005). Multiplicative perturbations can be used for distributed privacy-preserving data mining (Liu et al., 2005). A number of techniques for multiplicative perturbation, such as rotation, projection, and geometric perturbations, were found in study (Kim & Winkler, 2003; Liu et al., 2005). Data randomization indicates noise addition or noise multiplication with data, and another process called data swapping is used for data perturbation. In data swapping, the values across different data records are swapped to perform privacy preservation (Fienberg & McIntyre, 2004). Data swapping does not follow the general rules of randomization because perturbation of one record is dependent from other records. In data swapping technique, certain types of aggregate computation can be exactly performed without compromising privacy.

Group-based anonymization methods

Many privacy transformations are to construct groups among anonymous records that are transformed in a group-specific manner. A number of techniques for group anonymization, such as k-anonymity, l-diversity, and t-closeness methods, have been proposed in different studies. Table 2 shows the comparison between different studies on group-

TABLE 2 Comparison between different group-based anonymization techniques proposed by different studies

Study	Main feature	New software framework	Scalability	Privacy breach/ effectiveness analysis	Privacy model	Data distortion	Time efficiency	Considerable data type	Precise scheme	Comparison with existing approach	Experiment on real datasets	Security operation/ applied methods	Example application
Machanavajjhala et al. (2006)	Shown the oretical and experimental proof that a k-anonymity allows strong attacks due to lack of diversity in data	Yes (l-diversity)	High	Yes	Entropy l-diversity and recursive l-diversity	Moderate (generalization)	High	Arbitrary data	Generalization with maintaining diversity	Yes	Yes	Bayes optimal privacy	No
Li et al. (2007)	Proposed t-closeness for removing the limitation of k-anonymity and l-diversity.	No	High	Yes	t-closeness with generalization approach	Moderate (generalization)	High	Categorical and numerical data	Anonymization approach with threshold t	Yes	Yes	EMD with generalization and subset property	No
Gedik and Liu (2004)	Proposed a customizable k-anonymity model for providing Location privacy	Yes (ClicqueCloaking algorithm)	High	Yes	Customizable k-anonymity	High	High	Arbitrary data	Encryption and decryption approach	Yes	Yes	Applied spatial clocking and temporal clocking	No
Jiang and Clifton (2005)	Proposed a protocol for generating k-anonymous data	Yes (DPP$_2$GA)	High	No	k-anonymous on vertically partitioned data	High	High	Arbitrary (vertically partitioned) data	Anonymization through generalization approach	Yes	No	Applied distributed privacy preserving two party generic Anonymizer	No
Xiao and Tao (2007)	Proposed a method for privacy preserving re-publishing of a fully dynamic dataset	Yes (CG algorithm)	High	Yes	m-invariance generalization approach	Moderate (generalization)	High	Arbitrary data	Local and global recoding generalization scheme	Yes	Yes	Applied counterfeited generalization algorithm for computing publishable relations	No
Bayardo and Agrawal (2005)	Proposed an algorithm on anonymization	Yes	High	Yes	Optimal k-anonymity	High (suppression and generalization)	Low	Arbitrary data	Anonymization approach	Yes	Yes (census dataset)	Applied cost metrics (discernibility and classification metric)	No
Zhang, Liu, et al. (2013a)	Multidimensional anonymization	Yes (MRMONDRAIN DRIVER)	High	Yes	k-anonymity	Low	Low	Continuous and nominal data	Multidimensional scheme	Yes (good in scalability, cost effectiveness but not in time efficiency)	Yes	Generalization	No

TABLE 2 (Continued)

Study	Main feature	New software framework	Scalability	Privacy breach/ effectiveness analysis	Privacy model	Data distortion	Time efficiency	Considerable data type	Precise scheme	Comparison with existing approach	Experiment on real datasets	Security operation/ applied methods	Example application
Wong, Li, Fu, and Wang (2006)	Propose the (α,k) anonymity model to protect both identifications and relationships To sensitive information in data.	Yes (global and local re-coding algorithm)	High	No	k-anonymity with α-dissociation criteria	Low	High	Arbitrary data	k-anonymization approach	Yes	Yes	Optimal global re-coding method and efficient local-encoding based method	No
LeFevre, DeWitt, and Ramakrishnan (2006)	Introduce a multidimensional recoding model for k-anonymity	Yes (Greedy algorithm)	High	No	Top down greedy approximation model	High	High (more time efficient than singles dimension)	Categorical and numerical data	Multidimensional anonymization approach	Yes	Yes	General-purpose metrics for contracting generalization hierarchy	No
Rebollo-Monedero, Forne, and Domingo-Ferrer (2009)	Propose an idea for producing privacy like t-closeness	No	Moderate	No	Post randomization method	High	Low	Arbitrary data	Randomization approach	Yes	No	Information theoretic concept	No

TABLE 3 Comparison between k-anonymity, l-diversity, and t-closeness methods

Methods / serial	k-anonymity method	l-diversity method	t-closeness method
1.	Sensitive data disclosure	Semantic similarity of sensitive data	Distance measures
2.	Vulnerability under homogeneous attack	Vulnerability under skewness attack	Vulnerability under attribute linkage attack
3.	Strong against record linkage only	Strong against record linkage and attribute linkage	Strong against attribute linkage and probabilistic attack

based anonymization techniques. Before describing these methods, we will explain five key terms that are related to these models in the privacy field.

Identifiers: These attributes include information that uniquely and directly identify individuals, such as identifier name, driver license, and social security number.

Quasi-identifiers: These attributes denote a set of information, such as gender, age, date of birth, and zip code that are not unique. However, quasi-identifiers can be combined with other external data to identify an individual.

Sensitive attributes: These attributes include private and personal information, such as sickness and salary.

Insensitive attributes: These attributes refer to general and innocuous information.

The k-anonymity model was introduced by Sweeney in 2002 and was developed because of the possibility of indirect identification of records from public databases. A table is called k-anonymous if every record in the table is similar to at least k − 1 other records with respect to every set of quasi-identifiers. Here, we reduce the granularity of data representation with the use of different techniques, such as generalization and suppression. In the anonymous method, the values in the quasi-identifiers are generalized by reducing their accuracy and representing them as hierarchies. The k-anonymity method successfully blocks the record linkage attack (Verykios & Christen, 2013). However, k-anonymity incompletely protects the users' privacy when homogeneity of the sensitive values is found within a group. For protecting the homogeneity attack in k-anonymity, the l-diversity model was proposed by Ashwin Machanavijjhala in 2007 (Machanavajjhala et al., 2006). The l-diversity method maintains every quasi-identifier group to contain at least l "well-represented" sensitive values. The simple understanding of "well-represented" is to ensure that at least l distinct values are found for the sensitive attributes in each quasi-identifier group. The concept of intragroup diversity of sensitive values is promoted within the anonymization scheme (Machanavajjhala et al., 2006). The l-diversity only guarantees the diversity of sensitive values but a problem is found that different values may belong to the same category. Therefore, l-diversity does not prevent attribute linkage attacks when the overall distribution of a sensitive attribute is skewed. The limitations of l-diversity are resolved by another method named t-closeness, which was first introduced in a study in 2007 (Li, Li, Venkatasubramanian, 2007). In the t-closeness method, the distribution of sensitive attributes in any equivalence class is close to the distribution of the attributes in the entire table. By definition, the distance between the two distributions is smaller than a threshold t. The t-closeness uses the earth mover's distance (EMD) function to measure the closeness between distributions of sensitive values and requires the closeness to be within t. A simple comparison between each method of group-based anonymization is given in Tables 3 and 4.

Table 5 shows the hierarchies generalized for zip code from Table 6. At least k records that share the same combination of generalized quasi-identifiers are found. If we apply k-anonymity method with $k = 5$ to Table 6 on quasi-identifiers (zip code, gender, and age), the result is similar to Table 7, which contains three equivalence classes.

The l-diversity ensures each equivalence class has diversity (at least l) of different sensitive attributes, such as the first and second equivalence classes in Table 7. A dataset satisfies l-diversity if each group of records shares a combination of key attributes, with at least l "well-represented" values for each confidential attribute. The term "well-represented" can be defined in several means, such as distinct l-diversity, entropy l-diversity, and recursive (c, l)-diversity (Machanavajjhala et al., 2006). Therefore, a table is said to have t-closeness when its equivalence classes have t-closeness.

2.1.6 | Shortcoming of group-based anonymization

Although k-anonymity method successfully blocks the linkage attack, it fails to protect the users' privacy because of sensitive attribute disclosure. As shown in Table 7, the equivalence class contains the quasi-identifier combination of

TABLE 4 Medical record (raw) from a hospital

No	Name	SSN	Zip	Sex	Age	Disease
1	Aly	354	45112	F	29	Heart Di
2	Toma	123	44216	F	66	Lung Ca
3	Dev	234	45111	M	44	Diabetes
4	Jack	345	44125	M	70	Lung Ca
5	Mimi	456	45203	F	52	Hepatitis
6	Goni	567	45112	M	23	Diabetes
7	Chu	678	45115	F	28	Kidney
8	Jose	890	44203	M	69	Lung Ca
9	Jery	987	45125	F	55	Diabetes
10	Don	876	44211	M	63	Lung Ca
11	Jhon	765	45121	M	22	Hepatitis
12	Wei	654	45112	M	60	Diabetes
13	Rony	543	45112	F	47	Hepatitis
14	Mary	321	45125	F	26	Lung Ca
15	Moni	440	44112	M	80	Lung Ca

TABLE 5 Anonymization level

Zip code				
Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
45112	4511*	451**	45***	*
44216	4421*	442**	44***	*
45111	4511*	451**	45***	*
44125	4412*	441**	44***	*
45203	4520*	452**	45***	*
45112	4511*	451**	45***	*
45115	4511*	451**	45***	*
44203	4420*	442**	44***	*
45125	4512*	451**	45***	*
44211	4421*	442**	44***	*
45121	4512*	451**	45***	*
45112	4511*	451**	45***	*
45112	4511*	451**	45***	*
45125	4512*	451**	45***	*

44***, *, >60, and the sensitive attribute (diagnosis) has the same value with lung cancer. If an adversary knows anybody who has the zip code 44***, either male or female, and is older than 60, he can easily learn that the individual has lung cancer. This condition is called the homogeneity attack.

The l-diversity method remains vulnerable to the similarity attack. A similarity attack mainly matches the similar values among sensitive attributes. In the l diversity method, sensitive attributes are only diverse but they may be under the same categories, such as the same categories of sickness have different types of diseases. Thus, similarity attack is a potential problem for the l-diversity method. The t-closeness method is introduced to address the protection of similarity attack.

However, some criticisms can be made to t-closeness: (Li et al., 2007)

- Although several methods are used to check t-closeness (using several distances between distributions), no computational procedure is available to enforce this given property.

No	Zip	Sex	Age	Disease
1	45112	F	29	Heart Di
2	44216	F	66	Lung Ca
3	45111	M	44	Diabetes
4	44125	M	70	Lung Ca
5	45203	F	52	Hepatitis
6	45112	M	23	Diabetes
7	45115	F	28	Kidney
8	44203	M	69	Lung Ca
9	45125	F	55	Diabetes
10	44211	M	63	Lung Ca
11	45121	M	22	Hepatitis
12	45112	M	60	Diabetes
13	45112	F	47	Hepatitis
14	45125	F	26	Lung Ca
15	44112	M	80	Lung Ca

TABLE 6 Record without identification data

No	Zip	Sex	Age	Disease
1	45***	*	20–30	Heart Di
2	45***	*	20–30	Lung Ca
3	45***	*	20–30	Diabetes
4	45***	*	20–30	Kidney
5	45***	*	20–30	Hepatitis
6	45***	*	31–60	Diabetes
7	45***	*	31–60	Diabetes
8	45***	*	31–60	Diabetes
9	45***	*	31–60	Hepatitis
10	45***	*	31–60	Hepatitis
11	44***	*	>60	Lung Ca
12	44***	*	>60	Lung Ca
13	44***	*	>60	Lung Ca
14	44***	*	>60	Lung Ca
15	44***	*	>60	Lung Ca

TABLE 7 k-anonymization table ($k = 5$)

- The t-closeness will greatly damage the usefulness of information. In study (Li et al., 2007), they acknowledged that t-closeness limits the utility of released information. However, enforcing t-closeness destroys the correlations between key attributes and confidential attributes, and previous limitations are slightly considered. On the basis of the definition of t-closeness, the values of a confidential attribute have the same distribution for any combination of values of sensitive attributes. In the t-closeness method, a system can tradeoff between utility and privacy by tuning threshold t.

2.1.7 | Application of group-based Anonymization

Group-based anonymization methods are widely used to solve different security issues of data mining in large-scale environments. A number of potential anonymization methods have been proposed in different articles (Gedik & Liu, 2004; Jiang & Clifton, 2005; Xu & Yung, 2004), and their performance is above satisfactory level. Other techniques, such as anatomy and m-invariance, have been proposed in other papers (Xiao & Tao, 2007). Considering the tradeoff between information and privacy losses, group-based anonymization techniques need to be reformed and extended to updated versions. In study (LeFevre, DeWitt, & Ramakrishnan, 2005), an incognito method was proposed to compute a k-minimal generalization with the use of bottom-up aggregation along the domain generalization hierarchies. Two methods, namely top-down specialization and bottom-up generalization, using group-based anonymization were proposed by Fung (2005), Fung, Wang, and Yu (2005)).

A technique was proposed in study (Winkler, 2002), where a simulated annealing algorithm is used to generate k-anonymous representation of the data. In study (Iyengar, 2002), the authors proposed another privacy technique using genetic algorithm to construct k-anonymization. The group-based anonymization approach has been extended to other techniques, such as wavelet-based and fast Fourier transformation-based techniques (Liu, Wang, & Zhang, 2008; Xu & Lai, 2007). A sparsified SVD method was introduced in study (Xu, Zhang, Han, & Wang, 2006) for data distortion. They constructed few matrices to measure the difference between the distorted dataset and the original dataset and the degree of privacy protection. They showed the experimental results using synthetic and real-world datasets, fairly presenting that the sparsified SVD method works well in preserving privacy and maintaining the utility of the datasets.

Partition-based privacy preservation

Partitioning is a popular and favorable application in distributed database management systems. Such partitioning may be horizontal, where records are distributed across multiple entities, or vertical, where attributes are distributed across multiple entities. Partitioning improves the performance for sites that have regular transactions involving certain views of data while maintaining the availability and privacy of data. The partitioning technique uses various sets of instructions for sharing their entire datasets because the overall effect is to preserve privacy for each individual.

Table 8 demonstrates the horizontal partitioning of data, where two banks, HSBC and Hang Seng, collect personal information of their clients.

Table 9 shows the vertical partitioning of data, where it presents two censuses of different organizations that collect different information for the same populations. Different attributes, such as SSN, name, gender, age, occupation, income, and religion, are stored in different databases. Table 10 demonstrates the arbitrary partitioning of data.

Table 11 summarizes the relevant references specifying privacy-preserving data mining problems in horizontally partitioned, vertically partitioned, and arbitrary partitioned methods.

2.1.8 | Shortcomings of partition-based techniques

Partition-based techniques depend on the level of trust between two participants because the protocol may be affected by various adversaries, such as semihonest and malicious adversaries.

TABLE 8 Horizontally partitioned data

Sl. no	Client ID	Gender	Age	Occupation	Account type
HSBC Bank (100 entities with 6 attributes)					
1	A-1034	Male	30	Doctor	Saving
2	A-1190	Female	46	Teacher	Current
3	Y-1230	Female	29	Labor	Current
....	....				
100	Z-9560	Male	65	Student	Saving
Hang Seng Bank (150 entities with 6 attributes)					
1	19560	Female	30	Officer	Current
2	21450	Male	27	Engineer	Saving
3	45569	Male	37	Merchantman	Saving
4	57683	Female	55	Driver	Current
	....			....	
150	98970	Male	50	Teacher	Current

TABLE 9 Vertically partitioned data

Grand dataset							
Sl. no	**SSN**	**Disease**	**Gender**	**Age**	**Occupation**	**Income**	**Religious**
1	123–564	Cancer	Male	34	Farmer	US$520	Buddhist
2	234–678	Diabetes	Male	45	Engineer	US$1670	Christian
3	123–432	Fever	Female	39	Doctor	US$2000	Hindu
4	234–678	Diarrhea	Female	67	Driver	US$670	Christian
5	124–876	Cancer	Male	74	Student	US$1200	Buddhist
....	....						
19999	999–989	Back pain	Male	25	Teacher	US$3500	Muslim

Census by Health Ministry (1,999 entities with 4 attributes)				Census by Finance Ministry (1,999 entities, 3 attributes)		
SSN	**Disease**	**Gender**	**Income**	**SSN**	**Occupation**	**Income**
123–564	Cancer	Male	US$520	123–564	Farmer	US$520
234–678	Diabetes	Male	US$1670	234–678	Engineer	US$1670
123–432	Fever	Female	US$2000	123–432	Doctor	US$2000
234–678	Diarrhea	Female	US$670	234–678	Driver	US$670
124–876	Cancer	Male	US$1200	124–876	Student	US$1200
....				....		
999–989	Back pain	Male	US$3500	999–989	Teacher	US$3500

TABLE 10 Arbitrary partitioned data (two matrix)

Grand dataset							
Sl. no	**SSN**	**Disease**	**Gender**	**Age**	**Occupation**	**Income**	**Religious**
1	🟨 123–564	🟨 Cancer	🟨 Male	34	Farmer	US$520	Buddhist
2	🟨 234–678	🟨 Diabetes	🟨 Male	45	Engineer	US$1670	Christian
3	🟨 123–432	🟨 Fever	🟨 Female	39	Doctor	US$2000	Hindu
4	🟨 234–678	🟨 Diarrhea	🟨 Female	67	Driver	US$670	Christian
5	124–876	Cancer	Male	74	Student	US$1200	Buddhist
....	....			🟩	🟩	🟩	🟩
19999	999–989	Back pain	Male	🟩 25	🟩 Teacher	🟩 US$3500	🟩 Muslim

Note: 🟨Matrix with 4 entities, 3 attributes; 🟩Matrix with 99 entities, 4 attributes.

2.1.9 | Semihonest adversaries

Two-party computations are assumed to be semihonest, that is, the two parties faithfully follow their specified protocols. However, we assume that they record intermediate messages in an attempt to infer as much information about the other party's data as possible. In many situations, this condition may be considered a realistic model of adversarial behavior.

2.1.10 | Malicious adversaries

Two-party settings may deviate from the protocol and may send sophisticated inputs to one another because they learn in an unfair manner about each other, making them to be under malicious adversaries.

TABLE 11 Comparison between privacy preservation methods on partitioned data proposed by different authors

Study	Main feature	Communication cost	New software framework	Scalability	Privacy breach/ effectiveness analysis	Privacy model	Data distortion	Time efficiency	Considerable data type	Precise scheme	Comparison with existing approach	Experiment on real datasets	Security operation/ applied methods	Example application
Polat (2005)	Proposed a scheme to provide top-N recommendations without compromising privacy	O(m)	Yes (private similarity computation protocol)	High	No	PPTN on HPD*	N/A	Low	Numeric datasets	Horizontal partitioned scheme	Yes	Yes	Collaborative filtering and binary rating method	No
Dwork and Nissim (2004)	Proposed some strategies for providing privacy on partitioned data	N/A	Yes	High	Yes	Cryptographic and perturbation approach	High (adding noise)	High	Arbitrary data	Vertical partition scheme	No	No	Semantic security	No
Inan et al. (2007)	Proposed a method for privacy preserving clustering on partitioned data	$O(p*n^{2})$	Yes (clustering algorithm)	High	Yes	Distributed clustering approach	Low	Low	Numeric, alphanumeric and categorical data	Horizontal partition Scheme	Yes	Yes	Secure multi-party computation of a dissimilarity matrix	No
Jagannathan and Wright (2005)	Proposed a protocol for privacy preserving k-means clustering on partitioned data	O(nlck)	Es (K-means clustering algorithm)	N/A	No	Cryptographic and K-means clustering approach	Low	High	Numeric data	Arbitrary (horizontal and vertical) Partition scheme	Yes	No	Homomorphic (semantically secure) encryption	No
Kantarcioglu and Clifton (2004)	Addressed secure mining of association rules over partitioned data	$O(t^{3*\ \vert CG}{}_{(k)\vert}*N^{2})$ and $O(\vert CG_{(k)}\vert*m*t^{3})$	Yes (privacy in semi honest model)	High	No	Association rule and cryptographic approach	Low	High	Arbitrary datasets	Horizontal partition scheme	Yes	Yes	Fast distributed mining of association rules (FDM)	No
Kantarcıoglu, Vaidya, and Clifton (2003)	Presented Naive Bayes classifier for partitioned data	Number of bits O(log(p). (t + k)) Bit transfer O(log(p))	Yes (protocol for numeric and nominal attributes)	Moderate	No	Cryptographic approach	High	High	Numeric and nominal datasets	Horizontal partition scheme	Yes	No	Secure summation and logarithm model	No
Lakshmi and Rani (2012)	Proposed a method to compute association rules without disclosing entity values	Round 4 Bitwisecost 2*N*MaxValSz O(N)	Yes (association rule mining algorithm)	High	No	Association rule and distributed approach	N/A	Low	Boolean datasets	Vertically partitioned scheme	Yes	No	Scalar product method with Boolean association rule	No
Yakut and Polat (2010)	Proposed a method for privacy preserving through SVD based CF on partitioned data	N/A	Yes (privacy preserving protocol-1 & II)	High	No	Homomorphic cryptosystem	N/A	Low	Arbitrary dataset	Integrating horizontal and vertical partition scheme	Yes	Yes	Collaborative filtering methods with model based scheme	No
Vaidya and Clifton (2003)	Proposed a method for privacy preserving in distributed and partitioned data	roundO(r + k) and bitcostO(rnk)	Yes (K-means clustering)	High	No	Homomorphic cryptosystem	Low	High	Arbitrary dataset	Vertical partition scheme.	Yes	No	Secure multi party computation	No
Vaidya and Clifton (2005)	Proposed a method for building a decision tree on partitioned data.	$O(nk^{2}*bitsize)$	Yes (decision tree construction)	N/A	No	Cryptographic and secure multiparty computation approach	Low	High	Arbitrary dataset	Vertical partition scheme	Yes	No	ID3 algorithm on vertically partitioned data	No
Yu, Jiang, and Vaidya (2006)	Proposed a solution on PP-SVM nonlinear classification on partitioned data	Linear to the parties	Yes (SVM nonlinear classification algorithm)	High	No	Cryptographic approach	Low	High	Binary dataset	Horizontal partition scheme	Yes	Yes (SPECT datasets)	SVM in remote method invocation	No

Differential privacy methods

Dwork proposed an insightful privacy notion: the risk to the owners' record privacy should not substantially increase because of participating in the statistical database. Hence, participations in the social network does not pose a threat to their privacy if the data are released using a differential privacy mechanism because the statistics will not look extremely different in the absence of any participation. Dwork proposed a privacy model called ε-differential privacy to ensure that the inclusion or exclusion of a single database record does not potentially affect the outcome of any analysis (Dwork, 2011).

ε -differential privacy can be defined as follows:

A randomized function F satisfies ε-differential privacy if all datasets D_1 and D_2 differ at most one element and any subset S of possible outcomes is in range (F),

$$P(F(D_1)\epsilon S) \leq \exp(\varepsilon) \times P(F(D_2)\epsilon S).$$

Here, a user profile is an element, and $D_1 \subseteq V$ and $D_2 \subseteq V$.

2.2 | Privacy-preserving methods for big data

In data mining, the data are typically large, possibly distributed and cannot be fitted into the primary memory. Therefore, common functions rely on parallel and aggregate data from different sources and use parallel computing programming to conduct the mining process (Zhao et al., 2020). However, the data have unexpectedly larger volume in data mining than the capacity that a single personal computer can handle. Thus, big data mining framework relies on clustering with high-performance computing platform. For the case of big data, data mining task is deployed by running some parallel programming, dimensionality reduction, and distributed tools, such as MapReduce. Recently, researchers give significant attention to handle the volume and velocity of big data by upgrading the scope of MapReduce and extending the real-time nature of high-volume data processing units (Pramanik et al., 2017). For obtaining the statistics to solve or optimize model parameters, the training data are scanned using existing data mining algorithms. However, big data mining needs intensive computing to frequently access large-scale data (Han et al., 2020; Wu et al., 2013). For controlling the emerging nature of large-scale data, general purpose parallel processing algorithms are extended to large number of machine learning algorithms based on the simple MapReduce programming model on multicore processors.

Information sharing and provision in big data mining present new challenges for adopting new tools and rules, protecting privacy and confidentiality in sensitive data, and new opportunities by providing different useful datasets (Sarathy & Muralidhar, 2006). Privacy-preserving data mining has been studied for a long period, but the existing literature in big data mining privacy have significant research gaps. This problem has attracted much attention from the academic and industrial research communities, with different dimensional proposals.

The main objective of big data privacy is to assure privacy and achieve systematic protection of personal data and radical personal control over how they are collected and used. However, existing privacy-preserving data mining methods are unsuitable to cope with big data because of operational complexity and efficiency issues. Table 12 shows the comparison and summary of recent studies on big data privacy. Big data privacy issues are expanded by the velocity, volume, value, veracity, and variety of big data, such as large-scale datasets and diversity of data sources and formats. IS researchers have ignored the core technological issues in big data. After reviewing the literature of big data privacy, we have identified two major challenges, namely, data privacy and management and integrated security features, in organization security. The first challenge is to assure secure computation with scalable and various data storage and transaction, and the second challenge is to setup real-time security monitoring system with end point input veracity. The two major challenges can be solved through two types of approaches, including (a) developing special privacy models by enhancing the dimensions of privacy-preserving algorithm for controlling large-scale datasets and (b) enhancing the communication protocol security for secure data sharing mainly developed on hardware and software infrastructures.

A neural network is an excellent less-computationally sophisticated alternative to many systems that experience high computational overhead. The development in neural network research enables us to use more than one hidden layer learning by a network with more than one hidden layer is called deep learning, which can process a bulk amount of data and is suitable for big data (Hao et al., 2019). Recently, deep learning has been extensively applied in all areas of

TABLE 12 Comparison between different studies on privacy-preserving methods in big data context

Study	Used privacy method	New software framework	Measure running time and communication cost	Comparison with existing approach	Measure scalability	Experiments on real datasets	Key contributions
Hao, Li, Xu, Liu, and Yang (2019)	Federated deep learning, stochastic, differential privacy, additively homomorphic encryption	Yes	Yes	No	Yes	Yes	Present an efficient and secure gradients aggregation scheme in federated deep learning with lightweight ho-momorphic encryption.
Duan and Canny (2014)	Statistical learning methods (SVD, machine learning, ID3, PCA, K-means)	Yes (peers for privacy-P4P)	Yes	Yes	Yes	Yes	Show SVD can be performed efficiently with privacy in peers for privacy.
LeFevre, DeWitt, and Ramakrishnan (2008)	Anonymization method including SVM	Yes (workload aware anonymization based privacy)	Yes	Yes	Yes	Yes	Present algorithms for incorporating workloads into the anonymization process.
Gilburd et al. (2004)	Distributed data mining method	Yes (K-privacy in trusted third party)	No	No	Yes	No	Present new privacy model through generalizing trusted third party (TTP).
Magkos, Maragoudakis, Chrissikopoulos, and Gritzalis (2009)	Homomorphic encryption with horizontal partition	Yes (privacy preserving random forest algorithm)	No	Yes	Yes	No	Present different security requirements for large-scale privacy preserving data mining in the client to server setting.
Narayanan and Shmatikov (2019)	Anonymization, perturbation methods.	Yes (de-anonymization algorithm scoreboard-RH)	No	No	Yes	Yes	Presents the fundamental limits of privacy in public micro-data

(Continues)

TABLE 12 (Continued)

Study	Used privacy method	New software framework	Measure running time and communication cost	Comparison with existing approach	Measure scalability	Experiments on real datasets	Key contributions
Zhang, Liu, et al. (2013b)	Anonymization method	Yes (hybrid approach for subtree anonymization)	Yes	Yes	Yes	No	Propose a combine approach of top-down specialization (TDS) and bottom-up generalization (BUG) for efficient sub-tree anonymization over big data.
Zhang, Yang, et al. (2013)	Anonymization method	Yes (MRMONDRIAN)	Yes	Yes	Yes	Yes	Presents a scalable multidimensional anonymization approach for big data privacy preservation.
Huang & Du (2014)	Image encryption method	Yes	Yes	Yes	Yes	Yes	Achieve image data privacy via hybrid cloud.
Lu, Zhu, Liu, Liu, and Shao (2014)	Privacy-preserving aggregation, operations over encrypted data, and de-identification techniques	Yes (propose a protocol)	Yes	Yes	No	No	Introduced a protocol in response to the efficiency and privacy requirements of data mining in the big data era.

science and engineering for efficient handling of big data and substantial computational overhead. Deep learning is also used in privacy preservation of big data (Hao et al., 2019).

To address big data privacy, researchers have explored different privacy methodologies in their recent studies. In the big data context, existing approaches experience severe scalability or IT cost issues. MapReduce, a parallel and distributed large volume data processing paradigm, has attracted much attention from the research community and has been widely used for big data application. Multidimensional anonymization methods can preserve data privacy in big data context by using MapReduce (Zhang, Yang, et al., 2013). A MapReduce-based subtree anonymization approach is applied in another study (Zhang, Liu, et al., 2013b), where MapReduce is leveraged to achieve high scalability and cost-effectiveness. In study (Huang & Du, 2014), hybrid cloud is used to achieve big data privacy and significantly accelerates the substitution and diffusion processes. Simultaneously, a number of statistical deanonymization approaches are proposed to demonstrate the basic limits of privacy in anonymized public microdata, but most of them only focus on the volume of big data (Narayanan & Shmatikov, 2019). A large number of statistical learning algorithms, including SVD, PCA, k-means, ID3, and machine learning algorithms, are widely used to design privacy methods in large-scale datasets (Duan & Canny, 2014; LeFevre et al., 2008).

Academic and industrial researchers are exploiting new privacy challenges of big data, especially devoting attention toward efficient and privacy-preserving computing in big data era. Although many security and privacy techniques have been developed in big data context, they are insufficient for the newly emerging big data scenarios because its new dimensions are included, making it extremely challenging (Lu et al., 2014). For designing an effective and efficient privacy method, researchers should simultaneously consider data privacy and data utility. Most privacy-preserving methods use some form of transformation on the data to perform privacy preservation. Such methods typically diminish the granularity of representation for achieving the desired privacy level. However, this reduction leads to information loss that also causes the loss of effectiveness of data management or mining algorithm (Chen & Zhang, 2014). This condition is a natural tradeoff between data utility and data privacy. Therefore, proper trade-off between information loss and privacy level is required for developing an effective privacy model depending on the intended application.

3 | CRITIQUES OF COMMON PRIVACY-PRESERVING METHODS IN BIG DATA ANALYTICS

In this section, we examine some existing privacy-preserving techniques for big data and explain the problems that may arise by applying these techniques to big data. The common privacy-preserving techniques that we will evaluate include de-identification, encryption, perturbation, aggregation, and suppression.

3.1 | De-identification for privacy preservation

De-identification is a usual technique used to prevent a person's identity from being connected with other information. The data should be first sanitized (to remove noise and inconsistent data, to handle the missing data fields, etc.) and then integrated (to combine data from multiple sources) before publishing to provide individual privacy. Masking or deleting identifiers (IDs), such as name, passport ID, and SSN, and generalizing quasi-identifiers, such as age and date of birth, are the common strategies used for de-identifying datasets. De-identification may include preserving identifying information that can be relinked by a trusted party in certain situations. Compared with other privacy techniques, including encryption (Jagannathan & Wright, 2005; Magkos et al., 2009), synthetic and swapping (Dalenius & Reiss, 1982; Fienberg & McIntyre, 2004), aggregation, and suppression (Bayardo & Agrawal, 2005), de-identification (Lu et al., 2014; Wang, Li, Guo, Zhang, & Cui, 2019) can make more effective and flexible data analytics and mining. For preserving privacy in big data, some traditional de-identification techniques, such as k-anonymity (Machanavajjhala et al., 2006), l-diversity (Machanavajjhala et al., 2006), and t-closeness (Li et al., 2007), have been deployed after enhancing their functionalities (Wang et al., 2019; Wong et al., 2006). In the big data era, an attacker can collect a number of information of an entity (individual) because it has many attributes, thereby usually providing rich background knowledge and increasing the risk of re-identification. Therefore, de-identifications are insufficient to preserve big data privacy due to different limitations. De-identifying the genetic information using these techniques is difficult (McGuire & Gibbs, 2006; Yang et al., 2018).

3.2 | Privacy preservation by aggregation

Data aggregation in big data finds relevant search query data and individual data findings in a summarized format that is meaningful and useful for the end user or application. In privacy context aggregation, which is built on some homomorphic encryptions (Magkos et al., 2009), data aggregation is a popular technique for collecting data from a data provider. Data aggregation reduces disclosure risks by turning atypical records (Zhu et al., 2009), which are most at risk into typical records, and usually works on big data or data marts that do not deliver much information value as a whole. Aggregation can protect privacy in the phase of data collection and preservation but does not have a major role in the phase of data publishing (Xiao & Tao, 2007). Two major problems are found in privacy-preserving aggregation. The first problem is confusing inference, which represents that aggregation is generic but does not apply at individual level. The second problem is stubbornness, which indicates that aggregation is purpose-specific, inflexible for general applications. For smooth extraction of knowledge, such biased technique is unsuitable for big data scenarios.

3.3 | Privacy preservation by suppression

Suppression is the process of replacing some values with a special value and sometimes is the process of deleting cell values or entire tuples (Oliveira & Zaiane, 2004). In agencies, suppression operations are used to delete sensitive values from the transformed dataset. Typical suppression scheme includes cell, record, and value suppressions. The use of suppression in big data can severely degrade the quality of data and can alter the overall statistics, rendering the data practically useless (Aggarwal, 2007). The main problem of cell suppression is the distortion of information in the table by intentionally selecting cells to suppress. Thus, data providers can obtain deceptive and biased inference on the basis of the cell values that are reported (Zhu et al., 2009).

3.4 | Privacy preservation by perturbation

Perturbation is the process of replacing the original data values with some synthetic data values, which is the replacement of original data values at high risk of disclosure with values simulated from probability distributions (Polat & Du, 2005). Two data perturbation approaches are used in large-scale datasets. The first approach is known as the probability distribution approach (Chen & Liu, 2005), where it takes the data and replaces it from the same distribution sample or from the distribution itself. The second approach is called the value distortion approach (Liu et al., 2005), where it perturbs data by multiplicative or additive noise or other randomized processes (Kargupta et al., 2003). In case of multiplicative or additive noise, the degree of confidentiality protection depends on the nature of noise distribution where great variance provides high protection. The perturbation family includes swapping values between records in the dataset (Agrawal & Aggarwal, 2001; Fienberg & McIntyre, 2004). Some limitations are found in perturbation methods when they are applied in big data environment. Velocity is an important challenge for providing real-time response in large-scale datasets, and perturbation over large-scale datasets is usually complex and time consuming. Hence, some perturbation methods cannot provide quick response to users or customers. A misclassification problem is found when perturbation is used for clustering big datasets (Oliveira & Zaiane, 2004).

3.5 | Privacy preservation by encryption

In encryption, a message is encoded in such a way that only authorized parties can read it. Encryption does not prevent access but denies to give the plain message. Two types of encryption, namely, data (image/text) transit (Huang & Du, 2014) and data storage, are used in applications (Yang, Zheng, Guo, Liu, & Chang, 2019). In big data analytics, operations over encrypted data are usually complex and complicated (Lu et al., 2014). High volume is a major challenge in big data, and operation on high-volume encrypted data is time consuming and a potential barrier of real-time response in big data analytics.

4 | FOUR-DIMENSIONAL FRAMEWORK FOR ANALYZING PRIVACY-PRESERVING BIG DATA ANALYTICS

The systematic assessment of the strengths and weaknesses of state-of-the-art privacy-preserving big data analytics methods prompts for a suitable evaluation framework and appropriate assessment criteria (benchmarks). Different methods may perform better than other methods based on specific criteria, such as data utility, robustness, complexity, and efficiency. A primary list of evaluation dimensions to be used for assessing privacy-preserving methods in big data analytics in the settings of business applications is summarized as follows.

- *Data utility*: it refers to the extent that the original information contents can be preserved when a privacy-preserving method is applied to a business dataset.
- *Robustness*: it refers to the reliability and accuracy of preserving users' privacy from a business dataset; an effective privacy-preserving method should be strong against different types of attacks and linkages.
- *Complexity*: it refers to the computational complexity of a privacy-preserving method; a polynomial complexity is preferred, and the communication costs incurred during the exchange of information between a number of collaborating computing units should be minimized.
- *Efficiency*: it refers to the computational efficiency of executing a privacy-preserving method. In the context of big business data, efficiency is an important dimension that determines the possible practical application of a specific privacy-preserving method.

We illustrate each of the proposed dimensions in the following paragraphs. We systematically classify the existing methods based on the proposed analytical framework, as shown in Table 13. According to our analysis, none of the existing methods perform well along all the four dimensions.

In Table 13, we evaluate 48 papers on data privacy using the 4D evaluation framework, where we classify all papers into eight classes. Although none of the studies performs better along all dimensions, different studies cover one or more dimensions successfully. In case of data utility, approximately 10% of studies presents high information loss, whereas other studies ensure high-level data utility. Only one-fourth of the studies are strongly robust because most privacy methods are better against a specific attack or privacy gap. A research work (Dwork, 2011) provides different and more robust privacy guarantees than others. The computation complexity levels of studies with robust privacy guarantees are comparatively high. Two-third of studies have high computational complexity similar to the present study. In big data context, computational efficiency is an important dimension of the data privacy model, and velocity is an important challenge of big data that researchers should consider in designing a privacy-preserving method. Table 13 shows that two-thirds of privacy methods are computationally efficient. Although no privacy methods can be used in big data context, all privacy methods in the last row of Table 13 can be deployed for privacy-preserving big data analytics after reducing the computational complexity. Most of the privacy methods are inadequate and cannot provide the desired privacy level in big data analytics.

4.1 | Data utility

In a study in KDD (2008), Brickell and Shmatikov proposed an evaluation methodology by comparing privacy gain with the resulting utility gain and concluded that "even modest privacy gains require approximately complete destruction of the data mining utility." Privacy and utility have inverse relation between them. Therefore, we can achieve 100% privacy by not releasing any data, but this solution has no utility. In data utility, evaluation parameters should be the aggregate of information that is lost after the application of privacy-preserving method. Privacy is an individual concept, and utility is an aggregate concept. A privacy method is good when it is applied to a dataset that is safe to be published only when the privacy for each individual is protected. Utility gain adds up when multiple pieces of knowledge are learned from the published dataset. Information loss in different techniques are dissimilar, and data utility for different techniques are diverse. Deviation from the prior perception of any data that might be false or correct leads to privacy loss, whereas only correct information contributes to utility. Although developing an ideal privacy-preserving method that will provide zero privacy and zero utility losses is practically impossible, the maximum utility of the data should be maintained without compromising the underlying privacy constraints. For assessing any privacy method, only the quantification of privacy is insufficient without quantifying the utility of the data created by the privacy method.

TABLE 13 Four-dimensional evaluation framework for privacy-preserving data analytics

Dimensions studies	Data utility	Robustness	Complexity	Efficiency
Liu et al. (2008)	Low	Weak	Low	Low
Bayardo and Agrawal (2005); Yakut and Polat (2010)	High	Strong	High	Low
Zhang, Liu, et al. (2013b); Zhang, Yang, et al. (2013)	Low	Weak	High	Low
Inan et al. (2007); Rebollo-Monedero et al. (2009)	High	Weak	Low	Low
Duan and Canny (2014); Huang and Du (2014); Lakshmi and Rani (2012); Oliveira and Zaiane (2004); Polat (2005); Zhang, Liu, et al. (2013a)	High	Weak	High	Low
Dwork and Nissim (2004); Fung et al. (2005); Jagannathan and Wright (2005); Jiang and Clifton (2005); Kantarcıoglu et al. (2003); Lu et al. (2014); Polat (2003); Wang et al. (2019); Wong et al. (2006); Xu et al. (2006); Xu and Lai (2007)	High	Weak	High	High
Aggarwal (2007); Aggarwal et al. (2010); Chen and Liu (2005); Evfimievski (2002); Fienberg and McIntyre (2004); Huang et al. (2005); LeFevre et al. (2008); Liu et al. (2005); Vaidya and Clifton (2003); Yu et al. (2006)	High	Weak	Low	High
Agrawal and Srikant (2000); Dwork (2011); Kantarcioglu and Clifton (2004); Kargupta et al. (2003); LeFevre et al. (2005); LeFevre et al. (2006); Li and Li (2007); Machanavajjhala et al. (2006); Magkos et al. (2009); Narayanan and Shmatikov (2019); Xiao and Tao (2007)	High	Strong	High	High

4.2 | Efficiency

Some new challenges, including how to handle large-scale datasets efficiently, how to quickly process streaming and near-real-time data, and how to analyze structured and unstructured data to maximize the value of big data accurately, are found in privacy-preserving big data analytics. The efficiency of any privacy-preserving method is the processing frequency to large volume datasets. With the tremendous increase in global data, computational efficiency is an important challenge for traditional privacy-preserving methods. In big data mining, real datasets often contain millions of high-dimensional tuples, and highly scalable algorithms are needed to handle them. The efficiency of a privacy method is measured in terms of three main factors, namely, CPU time, space requirements (related to memory usage and the required storage capacity), and communication requirements. Therefore, we should ensure high efficiency with proper utility of data in evaluating any privacy method.

4.3 | Robustness

The robustness and sustainability of a privacy method are defined by the degree of protection against different types of attacks. The robustness of any proposed method in practice depends on different reasons. In the distributed privacy-preserving method, the robustness of the protocol depends on the level of trust among participants because this type of protocol may be subjected to various types of adversarial attacks. In study (Chen & Liu, 2005), the robustness of privacy-preserving classification method exhibits a better balance between privacy and information losses without performance penalty. In the group-based anonymization method, the robustness of anonymity characterizes the stability of the overall quality of anonymity against dynamic setting (e.g., a strong attacker or different probability distributions). In case of big data analytics, a robust system continues to improve services, better functions with increasing data dimensionality, with smooth degradation, and performs well in undesired events, such as hardware failures and software bugs.

4.4 | Complexity

Complexity is an important dimension used to evaluate any privacy-preserving method. An optimal algorithm has time and space complexity in $O(n^2)$, where n is the number of data records. For the evaluation of computational cost based on the time requirements of a privacy method, an algorithm having a $O(n^2)$ polynomial complexity is more efficient than the one with $O(e^n)$ exponential complexity (Verykios et al., 2004). The high complexity of any privacy method is the reason of service quality degradation. Complexity is the ability of a privacy-preserving methods to execute with good performance in terms of all the resources implied by the method. Thus, any privacy-preserving algorithm has to be examined to reduce its computational complexity before running any algorithm. The communication cost should be considered when exchanging information between the number of collaborating sites. In big business data, a large volume of data is often distributed and exchanged between different computing units. Accordingly, the communication cost of distributed privacy methods should be kept to a minimum.

5 | BUSINESS CHALLENGES AND OPPORTUNITIES OF PRIVACY-PRESERVING BIG DATA ANALYTICS

In big data-related applications, such as business, health, and administration, privacy and data security issues are significant for several reasons (Md et al., 2020). Implementing the protection approaches to unprecedentedly large volume of datasets is mostly complicated. Secondly, a huge workload is assigned to individual and organizational levels for maintaining privacy and security (Chen & Zhang, 2014). However, we chose three business domains, namely, health business, retail business, and manufacturing, for exploring how big data can create value and its potential size. The three business domains represent one-third of the global gross domestic product (GDP) in 2010 (Manyika, 2011). The healthcare centers or businesses are some collection of segments within the economic system that provides logistic supports and services to treat patients with curative, preventive, rehabilitative, and palliative care. The healthcare sector earned 7% GDP of the global economy in 2010 (Healthcare's Digital Future, 2014). At the same trend in 2012, the Organization for Economic Co-operation and Development countries spent 9.3% of GDP on average in health, which was 9.2% in 2011 (OECD Health Statistics, 2020). In the healthcare sector, big data can improve the quality of care and treatment by providing global patient and treatment database. Another business domain is retail business, which is mainly the sale of products and services from individuals or businesses to end users. A large number of categories of retail business products, such as food and beverage, clothing and accessories, sporting goods, toys, books, movies and musical instruments, electronic equipment, small mechanical products, furniture and home furnishings, and miscellaneous, are found in the market. In study (Manyika, 2011), the McKinsey Global Institute reported that the retail sector contributed 6% of the total GDP of the global economy in 2010. Big data is a useful domain to interact between retailers and consumers because consumers can use it to search and compare the prices and quality of products and retailers can measure the market demand, analyze the consumer perceptions, and control the supply chain. In the manufacturing business domain, the merchandises for use or sale are produced by integrating row materials, labor, manufacturing tools, and a sort of processes and theories. The manufacturing sector, which contributed 18% of the total GDP in 2010, is another important driver of GDP in the global economy and is the main factor of many developed economies. In the digital and online business era, with the requirements of high-quality production, manufacturers are encouraged to utilize the advantages of big data for driving efficiency across the extended enterprise and designing high-quality product (Digital Insights, 2020). Manufacturing industries are using large pools of data to infer actionable business intelligence in real time. The interactive use of big data is an extremely effective forward-looking initiative because manufacturers can improve their demand forecasting and supply planning. At present, travel and transportation is a significant business area where big data analytics can help to achieve three crucial business objectives: exploit the availability of assets and infrastructures; enhance the services to provide considerable revenues and capacities; and improve customer experience (Explore IBM Software and Solutions, n.d.). The telecommunication business provides 2% GDP in the world economy. Therefore, big data ensures to promote growth and increase efficiency and profitability across the entire telecom area. Using big data, the telecommunication sector can receive different opportunities, such as optimizing routing and quality of service by network analysis in real time, analyzing call data records in real time for fraud detection, allowing call center representatives to flexibly and profitably modify network and channel distribution strategies, modifying marketing campaigns to individual customers using location-based and social networking technologies, and using insights into customer behavior and usage to develop new products and services.

However, various challenges are found in big data. Obtaining access to big data and its advanced analytics is crucial for handling and finding the right outcome, incentivizing the right behavior, and driving efficiencies (Demirkan & Delen, 2013). Reluctance to provide personal health information can impede the success of web-based healthcare services. At the same time, patients are mostly concerned about their individual privacy because it is extremely sensitive to them (Bansal & Gefen, 2010). Similarly, the retail industry needs to deal with a number of barriers where privacy challenge is a vital issue when it starts to gain the potential value from the use of big data. Therefore, the security and privacy of the involved consumers are the significant challenging issues. Another major problem of big data is to update IT infrastructure systems because many IT systems are settled several years ago. Present datasets have different characteristics and they cannot be readily integrated, accessed, and analyzed by present systems.

To obtain sustainable benefit from the manufacturing business, a company needs to have different types of challenges, including technological, managerial, ethical, and marketing challenges. To achieve business value from large-scale datasets, manufacturing companies need to considerably invest in IT sectors, such as updating the product lifecycle management platform, which is linked to various technical process, and making organizational changes, such as increasing the investment in research and development. The profit of the manufacturing sector depends on the rate of technology adoption and the ability to use the technology for remaining competitive and adding value (Virginia Beach Economic Development, n.d.). During the big data application, consumers and manufacturers are challenged by privacy issues. Breaching the security of consumers' data and manufacturers' plan and policy can have severe negative business consequences.

However, many types of businesses exist in the world, and every business somehow benefits from big data. Although big data provides organizations with new opportunities, they make some challenges that need special attention. Covering all the challenges and opportunities of privacy-preserving big data analytics in the business area is difficult due to the space limitation. To preserve the privacy of big business data effectively, we provide the following recommendations for businesses:

1. *Recommendation for policymakers in an organization*: The big data ecosystem requires the development and implementation of comprehensive, adaptable policy, and technology frameworks for building and maintaining consumer trust. Organizations should include mechanisms to hold entities by collecting and analyzing data accountable for complying with the rules and best practices. Organizations should provide supports for adopting the privacy-enhancing technical architectures/models to collect and share data in a secure manner.
2. *Recommendation for organizational practice on data privacy*: Organizations that participate in the big data ecosystem should appoint a team that is responsible for three key areas: (a) Analyzing the Consumer Privacy Bill of Rights and its underlying principles; (b) Extracting new ideas to incorporate and reflect these principles; (c) Providing a concrete report and some recommendations that are forwarded to policy makers for final action.
3. *Recommendation for consumer empowerments in the organization*: Organizations should develop the settlement of consumer empowerments in the three following important points. (a) Individual control, where consumers are allowed to exercise control over what an individual data organization collects from consumers and how they use it. (b) Transparency, where consumers are allowed to assess privacy and security practices of their organization. (c) Accuracy control, where consumers have right to correct personal data for avoiding the risk of adverse consequences for inaccurate data.
4. *Recommendation to recruit skilled professionals and run training programs*: For managing big data security, organizations should focus on recruiting, training, and hiring data privacy analysts, data security scientists, and data privacy architects, who can develop the applications for data exploration and data analytics in a secure manner. Cross trainings in different data disciplines, such as data warehousing, data integration, data quality, content management, and database administration, are required for most business intelligence/data warehousing professionals.

5. *Recommendation to develop and apply harmonic collaboration*: Organizations should develop collaborative strategies because big data has diversity and diverse technology teams. Big data should be managed with broad access and leveraged by multiple business units and stakeholders. For securely managing every operation, considerable harmonic collaborations are required among all business units. New collaborations are required to control big data privacy and to adopt for governing the use of personal data in the business process. A strong legislative big data privacy framework must be introduced to provide consumer privacy satisfaction and achieve competitive market conditions.

6 | CONCLUSIONS

With the explosive growth of all types of digitized information, the existing privacy preservation methods fail to scale up with big business data. Privacy-preserving methods in big data analytics are still in the infant stage, and these methods may not ensure user privacy because of operational and efficiency issues. In this article, we examine privacy-preserving big data analytics through a critical review of the different paradigms of privacy-preserving methodologies. The main contribution of this article is a thorough and systematic study of the state-of-the-art privacy-preserving methods, where their pros and cons are better understood by academic and industrial communities. Thus, a better roadmap for the design of the next generation privacy-preserving big data analytics can be developed by academic researchers and industrial practitioners. Specifically, we identify the main weaknesses of existing privacy-preserving approaches in big data analytics and propose a 4D framework for a systematic evaluation and a better design of the next generation privacy-preserving methods in big data analytics. To the best of our knowledge, this article is the first to report a systematic analysis of state-of-the-art privacy-preserving methods in big data analytics. This study conducts the first thorough study about the potential challenges and opportunities of applying privacy-preserving big data analytics to business settings. Accordingly, we provide five recommendations to organizations where they can better leverage these opportunities to achieve sustainable competitive advantages and continuous growth. These recommendations can benefit empirical researchers and industrial practitioners who may engage in big data analytics projects. In particular, these recommendations provide clear guidance to business managers who are involved in the governance of new initiatives related to privacy-preserving big data analytics. Future work should conduct an interdisciplinary study about privacy-preserving big data analytics where the inherent challenges can be better understood from different perspectives (e.g., legislative and legal aspects).

ACKNOWLEDGMENT

Our research work was partly supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Projects: CityU 11525716), the NSFC Basic Research Program (Project No. 71671155), BRUR Research Project 2019-20/40, and the CityU Shenzhen Research institute.

CONFLICT OF INTEREST

The authors have declared no conflict of interest for this article.

AUTHOR CONTRIBUTIONS

Ileas Pramanik: Conceptualization; formal analysis; visualization; writing-original draft; writing-review and editing. **Y.K. Lau Raymond:** Supervision; writing-review and editing. **Md Hossain:** Conceptualization; investigation; resources; supervision; writing-review and editing. **Md Rahoman:** Formal analysis; writing-original draft; writing-review and editing. **Sumon Debnath:** Conceptualization; formal analysis; writing-review and editing. **Md Rashed:** Conceptualization; formal analysis; writing-review and editing. **Md Uddin:** Conceptualization; funding acquisition; project administration.

RELATED WIREs ARTICLES

Verykios VS, Christen P. Privacy-preserving record linkage. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2013 Sep;3(5):321-32.

Torra V, Navarro-Arribas G. Data privacy. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014 Jul;4(4):269-80.

Further Reading

Cavoukian, A., & Jonas, J. (2012). *Privacy by design in the age of big data*. Canada: Information and Privacy Commissioner of Ontario.

Hansen, M., Schwartz, A., & Cooper, A. (2008). Privacy and identity management. *IEEE Security & Privacy*, *6*, 38–45.

Sawires, A., Tatemura, J., Po, O., Agrawal, D., & Candan, K. S.. 2005. *Incremental maintenance of path-expression views*. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data; Baltimore Maryland, pp. 443–454.

REFERENCES

Aggarwal, C. C.. 2007. *On randomization, public information and the curse of dimensionality*. IEEE 23rd International Conference on Data Engineering; Istanbul, Turkey, pp. 136–145.

Aggarwal, G., Panigrahy, R., Feder, T., Thomas, D., Kenthapadi, K., Khuller, S., & Zhu, A. (2010). Achieving anonymity via clustering. *ACM Transactions on Algorithms (TALG)*, *6*, 49.

Agrawal, D., & Aggarwal, C. C.. 2001. *On the design and quantification of privacy preserving data mining algorithms*. Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; pp. 247–255.

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ACM SIGMOD Record*, *29*, 439–450.

Bansal, G., & Gefen, D. (2010). The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems*, *49*, 138–150.

Bayardo, R. J., & Agrawal, R.. 2005. *Data privacy through optimal k-anonymization*. 21st IEEE International Conference on Data Engineering; Tokoyo, Japan, pp. 217–228.

Chamikara, M. A. P., Bertók, P., Liu, D., Camtepe, S., & Khalil, I. (2019). An efficient and scalable privacy preserving algorithm for big data and data streams. *Computers & Security*, *87*, 101570.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*, 314–347.

Chen, K., & Liu, L. 2005. *Privacy preserving data classification with rotation perturbation*. 5th IEEE International Conference on Data Mining. Houston, TX, p. 4.

Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, *6*, 73–85.

Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, *55*, 412–421.

Digital Insights. 2020. Available from http://www.mckinsey.com/insights/business_technology/

Duan, Y., & Canny, J. (2014). Practical distributed privacy-preserving data analysis at large scale. In *Large-scale data analytics* (pp. 219–252). New York, NY: Springer.

Dwork, C. (2011). Differential privacy. In *Encyclopedia of cryptography and security* (pp. 338–340). New York Dordrecht Heidelberg London: Springer.

Dwork, C., & Nissim, K.. 2004. *Privacy-preserving datamining on vertically partitioned databases*. Annual International Cryptology Conference; Santa Barbara, CA, pp. 528–544.

Evfimievski, A. (2002). Randomization in privacy preserving data mining. *ACM Sigkdd Explorations Newsletter*, *4*, 43–48.

Explore IBM Software and Solutions. n.d.. Available from http://www-01.ibm.com/software/data/bigdata/industry-travel.html

Fienberg, S. E., & McIntyre, J. 2004. Data swapping: Variations on a theme by dalenius and reiss. *International Workshop on Privacy in Statistical Databases*. Barcelona, Spain, pp. 14–29.

Fung, B. C., Wang, K., & Yu, P. S.. 2005. *Top-down specialization for information and privacy preservation*. 21st International Conference on Data Engineering. Tokoyo, Japan, pp. 205–216.

Gambs, S., Kégl, B., & Aïmeur, E. (2007). Privacy-preserving boosting. *Data Mining and Knowledge Discovery*, *14*, 131–170.

Gedik, B., & Liu, L. (2004). *A customizable k-anonymity model for protecting location privacy*. Atlanta, Georgia: Georgia Institute of Technology.

Gilburd, B., Schuster, A., & Wolff, R.. 2004. *K-TTP: A new privacy model for large-scale distributed environments*. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Seattle, Washington, pp. 563–568.

Han, Z., Wu, J., Huang, C., Huang, Q., & Zhao, M. (2020). A review on sentiment discovery and analysis of educational big-data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*, 1–22.

Hao, M., Li, H., Xu, G., Liu, S., & Yang, H. 2019. *Towards efficient and privacy-preserving federated deep learning*. ICC 2019–2019 IEEE International Conference on Communications; Shanghai, China, pp. 1–6.

Healthcare's Digital Future. 2014. Available from https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/healthcares-digital-future

Huang, X., & Du, X.. 2014. *Achieving big data privacy via hybrid cloud*. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); Toronto, Ontario, Canada, pp. 512–517.

Huang, Z., Du, W., & Chen, B.. 2005. *Deriving private information from randomized data*. Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore Maryland, pp. 37–48.

Inan, A., Kaya, S. V., Saygın, Y., Savaş, E., Hintoğlu, A. A., & Levi, A. (2007). Privacy preserving clustering on horizontally partitioned data. *Data & Knowledge Engineering*, *63*, 646–666.

Iyengar, V. S.. 2002. *Transforming data to satisfy privacy constraints*. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Edmonton Alberta Canada, pp. 279–288.

Jagannathan, G., & Wright, R. N.. 2005. *Privacy-preserving distributed k-means clustering over arbitrarily partitioned data*. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining; Chicago Illinois, pp. 593–599.

Jiang, W., & Clifton, C. 2005. *Privacy-preserving distributed k-anonymity*. IFIP Annual Conference on Data and Applications Security and Privacy. Storrs, CT, pp. 166–177.

Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge & Data Engineering*, *16*(9), 1026–1037.

Kantarcıoglu, M., Vaidya, J., & Clifton, C. 2003. *Privacy preserving naive bayes classifier for horizontally partitioned data*. IEEE ICDM Workshop on Privacy Preserving Data Mining; Melbourne, Florida, pp. 3–9.

Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2003). On the privacy preserving properties of random data perturbation techniques. *ICDM*, *3*, 99–106.

Kim, J., & Winkler, W. (2003). Multiplicative noise for masking continuous data. *Statistics*, *1*, 1–18.

Lakshmi, N. M., & Rani, K. S. (2012). Privacy preserving association rule mining in vertically partitioned databases. *International Journal of Computer Applications*, *39*, 29–35.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R.. 2005. *Incognito: Efficient full-domain k-anonymity*. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore Maryland, pp. 49–60.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R.. 2006. *Mondrian multidimensional k-anonymity*. Proceedings of IEEE International Conference on Data Engineering; Atlanta, GA, pp. 25–25.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2008). Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems (TODS)*, *33*, 17.

Li, N., Li, T., & Venkatasubramanian, S.. 2007. *t-closeness: Privacy beyond k-anonymity and l-diversity*. 2007 IEEE 23rd International Conference on Data Engineering; Istanbul, Turkey, pp. 106–115.

Li, Y. (2014). The impact of disposition to privacy, website reputation and website familiarity on information privacy concerns. *Decision Support Systems*, *57*, 343–354.

Lindell, Y. (2005). Secure multiparty computation for privacy preserving data mining. In *Encyclopedia of data warehousing and mining* (pp. 1005–1009). England: IGI Global.

Liu, K., Kargupta, H., & Ryan, J. (2005). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, *18*, 92–106.

Liu, L., Wang, J., & Zhang, J.. 2008. *Wavelet-based data perturbation for simultaneous privacy-preserving and statistics-preserving*. 2008 IEEE International Conference on Data Mining Workshops; Pisa, Italy, pp. 27–35.

Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, *28*, 46–50.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M.. 2006. *l-diversity: Privacy beyond k-anonymity*. 22nd International Conference on Data Engineering (ICDE'06). Atlanta, GA, pp. 24–24.

Magkos, E., Maragoudakis, M., Chrissikopoulos, V., & Gritzalis, S. (2009). Accurate and large-scale privacy-preserving data mining using the election paradigm. *Data & Knowledge Engineering*, *68*, 1224–1236.

Manyika, J. Big data: The next frontier for innovation, competition, and productivity. 2011. Available from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

McGuire, A. L., & Gibbs, R. A. (2006). No longer de-identified. *Science*, *312*, 370–371.

Md, I. P., Lau, R. Y., Md, A. K. A., Md, S. H., Md, K. H., & Karmaker, B. K. (2020). Healthcare informatics and analytics in big data. *Expert Systems with Applications*, *152*, 113388.

Narayanan, A., & Shmatikov, V. Robust de-anonymization of large sparse datasets: a decade later, 2019.

OECD Health Statistics 2020. Available from http://www.oecd.org/els/health-systems/health-data.htm

Oliveira, S., & Zaiane, O. Data perturbation by rotation for privacy-preserving clustering, 2004.

Polat, H., & Du, W.. 2003. *Privacy-preserving collaborative filtering using randomized perturbation techniques*. 3rd IEEE International Conference on Data Mining; Melbourne, pp. 625–628.

Polat, H., & Du, W.. 2005. *Privacy-preserving top-n recommendation on horizontally partitioned data*. The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). Compiegne, France, pp. 725–731.

Pramanik, M. I., Lau, R. Y., Yue, W. T., Ye, Y., & Li, C. (2017). Big data analytics for security and criminal investigations. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*, e1208.

Rebollo-Monedero, D., Forne, J., & Domingo-Ferrer, J. (2009). From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1623–1636.

Sarathy, R., & Muralidhar, K. (2006). Secure and useful data sharing. *Decision Support Systems*, *42*, 204–220.

Torra, V., & Navarro-Arribas, G. (2014). Data privacy. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*, 269–280.

Vaidya, J., & Clifton, C.. 2003. *Privacy-preserving k-means clustering over vertically partitioned data*. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Washington, D.C., pp. 206–215.

Vaidya, J., & Clifton, C. 2005. *Privacy-preserving decision trees over vertically partitioned data*. IFIP Annual Conference on Data and Applications Security and Privacy; Storrs, CT, pp. 139–152.

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, *33*, 50–57.

Verykios, V. S., & Christen, P. (2013). Privacy-preserving record linkage. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*, 321–332.

Virginia Beach Economic Development. n.d.. Available from www.yesvirginiabeach.com/business-sectors/pages/advanced-manufacturing.aspx

Wang, J., Li, H., Guo, F., Zhang, W., & Cui, Y.. 2019. D2D big data privacy-preserving framework based on (a, k)-anonymity model. Mathematical Problems in Engineering.

Winkler, W., 2002. Using simulated annealing for k-anonymity. Research Report 2002-07, US Census Bureau Statistical Research Division.

Wong, R. C. W., Li, J., Fu, A. W. C., & Wang, K.. 2006. (*α, k*)*-anonymity: an enhanced k-anonymity model for privacy preserving data publishing*. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Philadelphia PA, pp. 754–759.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*, 97–107.

Xiao, X., & Tao, Y., 2007. *M-invariance: Towards privacy preserving re-publication of dynamic datasets*. Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing China, pp. 689–700.

Xu, S., & Lai, S.. 2007. *Fast Fourier transform based data perturbation method for privacy protection*. 2007 IEEE Intelligence and Security Informatics; New Brunswick, New Jersey, pp. 221–224.

Xu, S., & Yung, M.. 2004. *K-anonymous secret handshakes with reusable credentials*. Proceedings of the 11th ACM Conference on Computer and Communications Security; Washington DC, pp. 158–167.

Xu, S., Zhang, J., Han, D., & Wang, J. (2006). Singular value decomposition based data distortion strategy for privacy protection. *Knowledge and Information Systems*, *10*, 383–397.

Yakut, I., & Polat, H. (2010). Privacy-preserving SVD-based collaborative filtering on partitioned data. *International Journal of Information Technology & Decision Making*, *9*, 473–502.

Yang, Y., Zheng, X., Guo, W., Liu, X., & Chang, V. (2018). Privacy-preserving fusion of IoT and big data for e-health. *Future Generation Computer Systems*, *86*, 1437–1455.

Yang, Y., Zheng, X., Guo, W., Liu, X., & Chang, V. (2019). Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. *Information Sciences*, *479*, 567–592.

Yu, H., Jiang, X., & Vaidya, J.. 2006. *Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data*. Proceedings of the 2006 ACM Symposium on Applied Computing; pp. 603–610.

Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., & Chen, J. (2013a). SaC-FRAPP: A scalable and cost-effective framework for privacy preservation over big data on cloud. *Concurrency and Computation: Practice and Experience*, *25*, 2561–2576.

Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., & Chen, J.. 2013b. *Combining top-down and bottom-up: Scalable sub-tree anonymization over big data using MapReduce on cloud*. 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications; Melbourne, Australia, pp. 501–508.

Zhang, X., Yang, C., Nepal, S., Liu, C., Dou, W., & Chen, J.. 2013. *A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud*. 2013 International Conference on Cloud and Green Computing; Karlsruhe, Germany, pp. 105–112.

Zhao, Y., Tarus, S. K., Yang, L. T., Sun, J., Ge, Y., & Wang, J. (2020). Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives. *Information Sciences*, *515*, 132–155.

Zhu, D., Li, X. B., & Wu, S. (2009). Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. *Decision Support Systems*, *48*, 133–140.

Zhu, Y., & Liu, L.. 2004. *Optimal randomization for privacy preserving data mining*. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Seattle, Washington, pp. 761–766.

How to cite this article: Pramanik MI, Lau RYK, Hossain MS, et al. Privacy preserving big data analytics: A critical analysis of state-of-the-art. *WIREs Data Mining Knowl Discov*. 2021;11:e1387. https://doi.org/10.1002/widm.1387