Journal of Big Data

SURVEY PAPER

Open Access

Privacy preservation techniques in big data analytics: a survey

P. Ram Mohan Rao[1,4*], S. Murali Krishna[2] and A. P. Siva Kumar[3]

*Correspondence:
rammohan04@gmail.com
[1] Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, India
Full list of author information is available at the end of the article

Abstract

Incredible amounts of data is being generated by various organizations like hospitals, banks, e-commerce, retail and supply chain, etc. by virtue of digital technology. Not only humans but machines also contribute to data in the form of closed circuit television streaming, web site logs, etc. Tons of data is generated every minute by social media and smart phones. The voluminous data generated from the various sources can be processed and analyzed to support decision making. However data analytics is prone to privacy violations. One of the applications of data analytics is recommendation systems which is widely used by ecommerce sites like Amazon, Flip kart for suggesting products to customers based on their buying habits leading to inference attacks. Although data analytics is useful in decision making, it will lead to serious privacy concerns. Hence privacy preserving data analytics became very important. This paper examines various privacy threats, privacy preservation techniques and models with their limitations, also proposes a data lake based modernistic privacy preservation technique to handle privacy preservation in unstructured data.

Keywords: Data, Data analytics, Privacy threats, Privacy preservation

Introduction

There is an exponential growth in volume and variety of data as due to diverse applications of computers in all domain areas. The growth has been achieved due to affordable availability of computer technology, storage, and network connectivity. The large scale data, which also include person specific private and sensitive data like gender, zip code, disease, caste, shopping cart, religion etc. is being stored in public domain. The data holder can release this data to a third party data analyst to gain deeper insights and identify hidden patterns which are useful in making important decisions that may help in improving businesses, provide value added services to customers [1], prediction, forecasting and recommendation [2]. One of the prominent applications of data analytics is recommendation systems which is widely used by ecommerce sites like Amazon, Flip kart for suggesting products to customers based on their buying habits. Face book does suggest friends, places to visit and even movie recommendation based on our interest. However releasing user activity data may lead inference attacks like identifying gender based on user activity [3]. We have studied a number of privacy preserving techniques which are being employed to protect against privacy threats. Each of these techniques has their own merits and demerits. This paper explores the merits and demerits of each

of these techniques and also describes the research challenges in the area of privacy preservation. Always there exists a trade off between data utility and privacy. This paper also proposes a data lake based modernistic privacy preservation technique to handle privacy preservation in unstructured data with maximum data utility.

Privacy threats in data analytics

Privacy is the ability of an individual to determine what data can be shared, and employ access control. If the data is in public domain then it is a threat to individual privacy as the data is held by data holder. Data holder can be social networking application, websites, mobile apps, ecommerce site, banks, hospitals etc. It is the responsibility of the data holder to ensure privacy of the users data. Apart from the data held in public domain, knowing or unknowingly users themself contribute to data leakage. For example most of the mobile apps, seek access to our contacts, files, camera etc. and without reading the privacy statement we agree for all terms and conditions, there by contributing to data leakage.

Hence there is a need to educate the smart phone users regarding privacy and privacy threats. Some of the key privacy threats include (1) Surveillance; (2) Disclosure; (3) Discrimination; (4) Personal embracement and abuse.

Surveillance

Many organizations including retail, e-commerce, etc. study their customers buying habits and try to come up with various offers and value added services [4]. Based on the opinion data and sentiment analysis, social media sites does provide recommendations of the new friends, places to visit, people to follow etc. This is possible only when they continuously monitor their customer's transactions. This is a serious privacy threat as no individual accepts surveillance.

Disclosure

Consider a hospital holding patient's data which include (Zip, gender, age, disease) [5–7]. The data holder has released data to a third party for analysis by anonymizing sensitive person specific data so that the person cannot be identified. The third party data analyst can map this information with the freely available external data sources like census data and can identify person suffering with some disorder. This is how private information of a person can be disclosed which is considered to be a serious privacy breach.

Discrimination

Discrimination is the bias or inequality which can happen when some private information of a person is disclosed. For instance, statistical analysis of electoral results proved that people of one community were completely against the party, which formed the government. Now the government can neglect that community or can have bias over them.

Personal embracement and abuse

Whenever some private information of a person is disclosed, it can even lead to personal embracement or abuse. For example, a person was privately undergoing medication for some specific problem and was buying some medicines on a regular basis from a

medical shop. As part of their regular business model, the medical shop may send some reminder and offers related to these medicines over phone. If any family member has noticed this, it will lead to personal embracement and even abuse [8].

Data analytics activity will affect data Privacy. Many countries are enforcing Privacy preservation laws. Lack of awareness is also a major reason for privacy attacks. For example many smart phones users are not aware of the information that is stolen from their phones by many apps. Previous research shows only 17% of smart phone users are aware of privacy threats [9].

Privacy preservation methods

Many Privacy preserving techniques were developed, but most of them are based on anonymization of data. The list of privacy preservation techniques is given below.

- K anonymity
- L diversity
- T closeness
- Randomization
- Data distribution
- Cryptographic techniques
- Multidimensional Sensitivity Based Anonymization (MDSBA).

K anonymity [10]

Anonymization is the process of modifying data before it is given for data analytics [11], so that de identification is not possible and will lead to K indistinguishable records if an attempt is made to de identify by mapping the anonymized data with external data sources. K anonymity is prone to two attacks namely homogeneity attack and back ground knowledge attack. Some of the algorithms applied include, Incognito [12], Mondrian [13] to ensure Anonymization. K anonymity is applied on the patient data shown in Table 1. The table shows data before anonymization.

K anonymity algorithm is applied with k value as 3 to ensure 3 indistinguishable records when an attempt is made to identify a particular person's data. K anonymity is applied on the two attributes viz. Zip and age shown in Table 1. The result of applying anonymization on Zip and age attributes is shown in Table 2.

Table 1 Patient data, before anonymization

Sno	Zip	Age	Disease
1	57677	29	Cardiac problem
2	57602	22	Cardiac problem
3	57678	27	Cardiac problem
4	57905	43	Skin allergy
5	57909	52	Cardiac problem
6	57906	47	Cancer
7	57605	30	Cardiac problem
8	57673	36	Cancer
9	57607	32	Cancer

Table 2 After applying anonymization on Zip and age

Sno	Zip	Age	Disease
1	576**	2*	Cardiac problem
2	576**	2*	Cardiac problem
3	576**	2*	Cardiac problem
4	5790*	>40	Skin allergy
5	5790*	>40	Cardiac problem
6	5790*	>40	Cancer
7	576**	3*	Cardiac problem
8	576**	3*	Cancer
9	576**	3*	Cancer

Table 3 L diversity privacy preservation technique

Sno	Zip	Age	Salary	Disease
1	576**	2*	5k	Cardiac problem
2	576**	2*	6k	Cardiac problem
3	576**	2*	7k	Cardiac problem
4	5790*	>40	20k	Skin allergy
5	5790*	>40	22k	Cardiac problem
6	5790*	>40	24k	Cancer

The above technique has used generalization [14] to achieve Anonymization. Suppose if we know that John is 27 year old and lives in 57677 zip codes then we can conclude John to have Cardiac problem even after anonymization as shown in Table 2. This is called Homogeneity attack. For example if John is 36 year old and it is known that John does not have cancer, then definitely John must have Cardiac problem. This is called as background knowledge attack. Achieving K anonymity [15, 16] can be done either by using generalization or suppression. K anonymity can optimized if the minimal generalization can be done without huge data loss [17]. Identity disclosure is the major privacy threat which cannot be guaranteed by K anonymity [18]. Personalized privacy is the most important aspect of individual privacy [19].

L diversity

To address homogeneity attack, another technique called L diversity has been proposed. As per L diversity there must be L well represented values for the sensitive attribute (disease) in each equivalence class.

Implementing L diversity is not possible every time because of the variety of data. L diversity is also prone to skewness attack. When overall distribution of data is skewed into few equivalence classes attribute disclosure cannot be ensured. For example if the entire records are distributed into only three equivalence classes then semantic closeness of these values may lead to attribute disclosure. Also L diversity may lead to similarity attack. From Table 3 it can be noticed that if we know that John is 27 year old and lives in 57677 zip, then definitely John is under low income group because salaries of all

three persons in 576** zip is low compare to others in the table. This is called as similarity attack.

T closeness

Another improvement to L diversity is T closeness measure where an equivalence class is considered to have 'T closeness' if the distance between the distributions of sensitive attribute in the class is no more than a threshold and all equivalence classes have T closeness [20]. T closeness can be calculated on every attribute with respect to sensitive attribute.

From Table 4 it can be observed that if we know John is 27 year old, still it will be difficult to estimate whether John has Cardiac problem or not and he is under low income group or not. T closeness may ensure attribute disclosure but implementing T closeness may not give proper distribution of data every time.

Randomization technique

Randomization is the process of adding noise to the data which is generally done by probability distribution [21]. Randomization is applied in surveys, sentiment analysis etc. Randomization does not need knowledge of other records in the data. It can be applied during data collection and pre processing time. There is no anonymization overhead in randomization. However, applying randomization on large datasets is not possible because of time complexity and data utility which has been proved in our experiment described below.

We have loaded 10k records from an employee database into Hadoop Distributed File System and processed them by executing a Map Reduce Job. We have experimented to classify the employees based on their salary and age groups. In order apply randomization we added noise in the form of 5k records which are randomly added to make a database of 15k records and following observations were made after running Map Reduce job.

- More number of Mappers and Reducers were used as data volume increased.
- Results before and after randomization were significantly different.
- Some of the records which are outliers remain unaffected with randomization and are vulnerable to adversary attack.
- Privacy preservation at the cost of data utility is not appreciated and hence randomization may not be suitable for privacy preservation especially attribute disclosure.

Table 4 T closeness privacy preservation technique

Sno	Zip	Age	Salary	Disease
1	576**	2*	5k	Cardiac problem
2	576**	2*	16k	Cancer
3	576**	2*	9k	Skin allergy
4	5790*	>40	20k	Skin allergy
5	5790*	>40	42k	Cardiac problem
6	5790*	>40	8k	Flu

Data distribution technique

In this technique, the data is distributed across many sites. Distribution of the data can be done in two ways:

i. Horizontal distribution of data
ii. Vertical distribution of data

Horizontal distribution When data is distributed across many sites with same attributes then the distribution is said to be horizontal distribution which is described in Fig. 1.

Horizontal distribution of data can be applied only when some aggregate functions or operations are to be applied on the data without actually sharing the data. For example, if a retail store wants to analyse their sales across various branches, they may employ some analytics which does computations on aggregate data. However, as part of data analysis the data holder may need to share the data with third party analyst which may lead to privacy breach. Classification and Clustering algorithms can be applied on distributed data but it does not ensure privacy. If the data is distributed across different sites which belong to different organizations, then results of aggregate functions may help one party in detecting the data held with other parties. In such situations we expect all participating sites to be honest with each other [21].

Vertical distribution of data When Person specific information is distributed across different sites under custodian of different organizations, then the distribution is called vertical distribution as shown in Fig. 2. For example, in crime investigations, the police officials would like to know details of a particular criminal which include health, profession, financial, personal etc. All this information may not be available at one site. Such a distribution is called vertical distribution where each site holds few set of attributes of a person. When some analytics has to be done data has to be pooled in from all these sites and there is a vulnerability of privacy breach.

In order to perform data analytics on vertically distributed data, where the attributes are distributed across different sites under custodian of different parties, it is highly

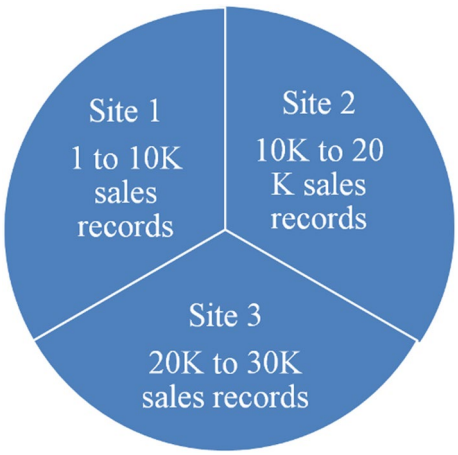

Fig. 1 Distribution of sales data across different sites

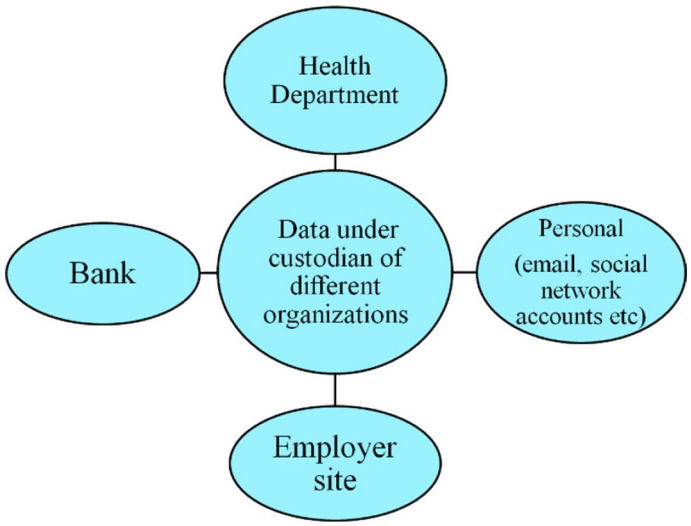

Fig. 2 Vertical distribution of person specific data

difficult to ensure privacy if the datasets are shared. For example, as part of a police investigation, the investigating officer wants to access some information about the accused from his employer, health department, bank to gain more insights about the character of the person. In this process some of the personal and sensitive information of the accused may be disclosed to investigating officer leading to personal embarrassment or abuse. Anonymization cannot be applied when entire records are not needed for analytics. Distribution of data will not ensure privacy preservation but it closely overlaps with cryptographic techniques.

Cryptographic techniques

The data holder may encrypt the data before releasing the same for analytics. But encrypting large scale data using conventional encryption techniques is highly difficult and must be applied only during data collection time. Differential privacy techniques have already been applied where some aggregate computations on the data are done without actually sharing the inputs. For example, if x and y are two data items then a function F(x, y) will be computed to gain some aggregate information from both x and y without actually sharing x and y. This can be applied on when x and y are held with different parties as in the case of vertical distribution. However, if the data is at single location under the custodian of a single organization, then differential privacy cannot be employed. Another similar technique called secure multiparty computation has been used but proved to be inadequate in privacy preservation. Data utility will be less if encryption is applied during data analytics. Thus encryption is not only difficult to implement but it reduces the data utility [22].

Multidimensional Sensitivity Based Anonymization (MDSBA)

Bottom up Generalization [23] and Top down Generalization [24] are the conventional methods of Anonymization which were applied on well represented structured data records. However, applying the same on large scale data sets is very difficult leading to

issues of scalability and information loss. Multidimensional Sensitivity Based Anonymization is a improved version of Anonymization proved to be more effective than conventional Anonymization techniques.

Multidimensional Sensitivity Based Anonymization is an improved Anonymization [25] technique such that it can be applied on large data sets with reduced loss of information and predefined quasi identifiers. As part of this technique Apache MAP REDUCE [26] framework has been used to handle large data sets. In conventional Hadoop Distributed Files System, the data will be divided into blocks of either 64 MB or 128 MB each and distributed across different nodes without considering the data inside the blocks. As part of Multidimensional Sensitivity Based Anonymization [27] technique the data is split into different bags based on the probability distribution of the quasi identifiers by making use of filters in Apache Pig scripting language.

Multidimensional Sensitivity Based Anonymization makes use of bottom up generalization but on a set of attributes with certain class values where class represents a sensitive attributes. Data distribution was made effectively when compared to conventional method of blocks. Data Anonymization was done using four quasi identifiers using Apache Pig.

Since the data is vertically partitioned into different groups, it can protect from background knowledge attack if the bag contains only few attributes. This method also makes it difficult to map the data with external sources to disclose any person specific information.

In this method, the implementation was done using Apache Pig. Apache Pig is a scripting language, hence development effort is less. However, code efficiency of Apache Pig is relatively less when compared to Map Reduce job because ultimately every Apache Pig script has to be converted into a Map Reduce job. Multidimensional Sensitivity Based Anonymization [28] is more appropriate for large scale data but only when the data is at rest. Multidimensional Sensitivity Based Anonymization cannot be applied for streaming data.

Analysis

Various privacy preservation techniques have been studied with respect to features including, type of data, data utility, attribute preservation and complexity. The comparison of various privacy preservation techniques is shown in Table 5.

Table 5 Comparison of privacy preservation techniques

Features	Privacy preservation techniques				
	Anonymization techniques	Cryptographic techniques	Data distribution	Randomization	MDSBA
Suitability for unstructured data	No	No	No	No	Yes
Attribute preservation	No	No	No	Yes	Yes
Damage to data utility	No	No	Yes	No	Yes
Very complex to apply	No	Yes	Yes	Yes	Yes
Accuracy of results of data analytics	No	Yes	No	No	No

Results and discussions

As part of systematic literature review, it has been observed that all existing mechanisms of privacy preservation are with respect to structured data. More than 80% of data being generated today is unstructured [29]. As such, there is a need to address following challenges.

i. Develop concrete solution to protect privacy in both structured and unstructured data.
ii. Scalable and robust techniques to be developed to handle large scale heterogeneous data sets.
iii. Data should be allowed to stay in its native form without need for transformation and data analytics can be carried out while ensuring privacy preservation.
iv. New techniques apart from Anonymization must be developed to ensure protection against key privacy threats which include identity disclosure, discrimination, surveillance etc.
v. Maximizing data utility while ensuring data privacy.

Conclusion

No concrete solution for unstructured data has been developed yet. Conventional data mining algorithms can be applied for classification and clustering problems but cannot be used in privacy preservation especially when dealing with person specific information. Machine learning and soft computing techniques can be used to develop new and more appropriate solution to privacy problems which include identity disclosure that can lead to personal embarrassment and abuse.

There is a strong need for law enforcement by governments of all countries to ensure individual privacy. European Union [30] is making an attempt to enforce privacy preservation law. Apart from technological solutions, there is a strong need to create awareness among the people regarding privacy hazards to safeguard themselves form privacy breaches. One of the serious privacy threats is smart phone. Lot of personal information in the form of contacts, messages, chats and files are being accessed by many apps running in our smart phone without our knowledge. Most of the time people do not even read the privacy statement before installing any app. Hence there is a strong need to educate people on the various vulnerabilities which can contribute to leakage of private information.

We propose a novel privacy preservation model based on Data Lake concept to hold variety of data from diverse sources. Data lake is a repository to hold data from diverse sources in their raw format [31, 32]. Data ingestion from variety of sources can be done using Apache Flume and an intelligent algorithm based on machine learning can be applied to identify sensitive attributes dynamically [33, 34]. The algorithm will be trained with existing data sets with known sensitive attributes and rigorous training of the model will help in predicting the sensitive attributes in a given data set [35]. Accuracy of the model can be improved by adding more layers of training leading to deep learning techniques [36]. Advanced computing techniques like Apache Spark can be used in implementing privacy preserving algorithms which is a distributed

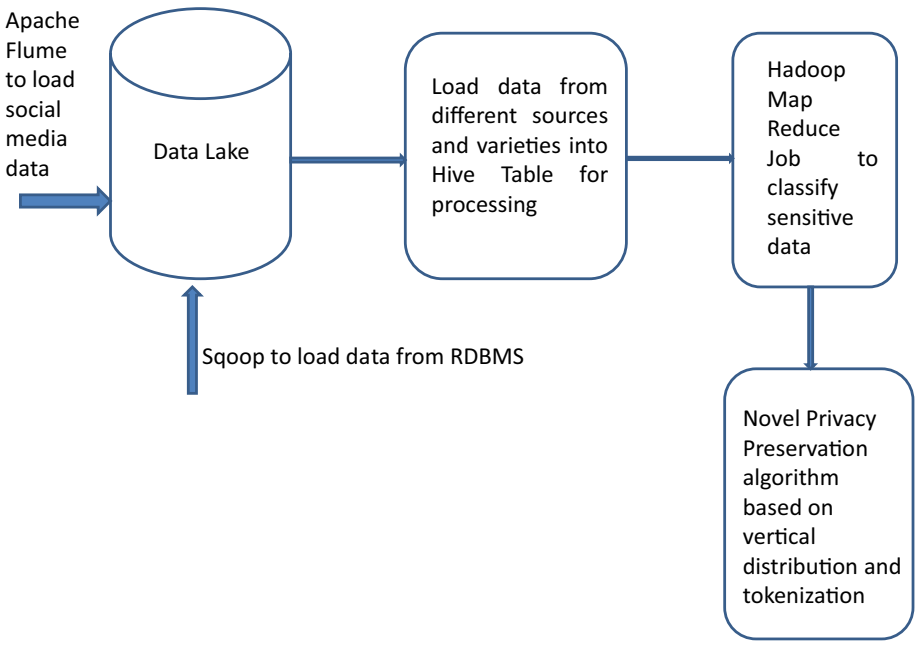

Fig. 3 A Novel privacy preservation model based on vertical distribution and tokenization

massive parallel computing with in memory processing to ensure very fast processing [37]. The proposed model is shown in Fig. 3.

Data analytics is done on the data collected from various sources. If an ecommerce site would like to perform data analytics, they need transactional data, website logs and customers opinion through social media pages. A Data lake is used to collect data from different sources. Apache Flume is used to ingest data from social media sites, website logs into Hadoop Distributed File System(HDFS). Using SQOOP relational data can be loaded into HDFS.

In Data lake the data can remain in its native form which is either structured or unstructured. When data has to be processed, it can be transformed into HIVE tables. A Hadoop map reduce job using machine learning can be executed on the data to classify the sensitive attributes [38]. The data can be vertically distributed to separate the sensitive attributes from rest of the data and apply tokenization to map the vertically distributed data. The data without any sensitive attributes can be published for data analytics.

Abbreviations

CCTV: closed circuit television; MDSBA: Multidimensional Sensitivity Based Anonymization.

Authors' contributions

PRMR: as part of Ph.D. work I have done my literature survey and submitted my work in the form of a paper. SMK: supported me in compiling the paper. APSK: suggested necessary amendments and helped in revising the paper. All authors read and approved the final manuscript.

Author details

[1] Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, India. [2] Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Tirupati, Andhra Pradesh, India. [3] Department of Computer Science and Engineering, JNTU Anantapur, Anantapuramu, Andhra Pradesh, India. [4] JNTU Anantapur, Anantapur, Andhra Pradesh, India.

Acknowledgements

I would like to thank my guides, for supporting my work and for suggesting necessary corrections.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
If any one is interested in our work, we are ready to provide more details of the map reduce job which we have executed and the data processing techniques applied. However the data is used in our work, is freely available in many repositories.

Funding
No Funding.

Publisher's Note

Received: 21 March 2018 Accepted: 4 September 2018
Published online: 22 September 2018

References

1. Ducange Pietro, Pecori Riccardo, Mezzina Paolo. A glimpse on big data analytics in the framework of marketing strategies. Soft Comput. 2018;22(1):325–42.
2. Chauhan Arun, Kummamuru Krishna, Toshniwal Durga. Prediction of places of visit using tweets. Knowl Inf Syst. 2017;50(1):145–66.
3. Yang D, Bingqing Q, Cudre-Mauroux P. Privacy-preserving social media data publishing for personalized ranking-based recommendation. IEEE Trans Knowl Data Eng. 2018. ISSN (Print):1041-4347, ISSN (Electronic):1558-2191.
4. Liu Y et al. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans Ind Inf. 2018.
5. Duncan GT et al. Disclosure limitation methods and information loss for tabular data. In: Confidentiality, disclosure and data access: theory and practical applications for statistical agencies. 2001. p. 135–166.
6. Duncan GT, Diane L. Disclosure-limited data dissemination. J Am Stat Assoc. 1986;81(393):10–8.
7. Lambert Diane. Measures of disclosure risk and harm. J Off Stat. 1993;9(2):313.
8. Spiller K, et al. Data privacy: users' thoughts on quantified self personal data. Self-Tracking. Cham: Palgrave Macmillan; 2018. p. 111–24.
9. Hettig M, Kiss E, Kassel J-F, Weber S, Harbach M. Visualizing risk by example: demonstrating threats arising from android apps. In: Smith M, editor. Symposium on usable privacy and security (SOUPS), Newcastle, UK, July 24–26, 2013.
10. Bayardo RJ, Agrawal A. Data privacy through optimal k-anonymization. In: Proceedings 21st international conference on data engineering, 2005 (ICDE 2005). Piscataway: IEEE; 2005.
11. Iyengar S. Transforming data to satisfy privacy constraints. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2002.
12. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. New York: ACM; 2005.
13. LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd international conference (ICDE'06) on data engineering, 2006. New York: ACM; 2006.
14. Samarati, Pierangela, and Latanya Sweeney. In: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.
15. Sweeney Latanya. Achieving k-anonymity privacy protection using generalization and suppression. In J Uncertain Fuzziness Knowl Based Syst. 2002;10(05):571–88.
16. Sweeney Latanya. k-Anonymity: a model for protecting privacy. Int J Uncertain, Fuzziness Knowl Based Syst. 2002;10(05):557–70.
17. Williams R. On the complexity of optimal k-anonymity. In: Proc. 23rd ACM SIGMOD-SIGACT-SIGART symp. principles of database systems (PODS). New York: ACM; 2004.
18. Machanavajjhala A et al. L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd international conference on data engineering (ICDE'06), 2006. Piscataway: IEEE; 2006.
19. Xiao X, Yufei T. Personalized privacy preservation. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data. New York: ACM; 2006.
20. Rubner Y, Tomasi T, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vision. 2000;40(2):99–121.
21. Aggarwal CC, Philip SY. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining. Springer: US; 2008. p. 11–52.
22. Jiang R, Lu R, Choo KK. Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. Future Gen Comput Syst. 2018;78:392–401.
23. Wang K, Yu PS, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection. In: Fourth IEEE international conference on data mining, 2004 (ICDM'04). Piscataway: IEEE; 2004.
24. Fung BCM, Wang K, Yu PS. Top-down specialization for information and privacy preservation. In: Proceedings 21st international conference on data engineering, 2005 (ICDE 2005). Piscataway: IEEE; 2005.
25. Zhang X et al. A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. In: Third international conference on cloud and green computing (CGC), 2013. Piscataway: IEEE; 2013.

26. Zhang X, et al. A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. IEEE Trans Parallel Distrib Syst. 2014;25(2):363–73.
27. Al-Zobbi M, Shahrestani S, Ruan C. Improving MapReduce privacy by implementing multi-dimensional sensitivity-based anonymization. J Big Data. 2017;4(1):45.
28. Al-Zobbi M, Shahrestani S, Ruan C. Implementing a framework for big data anonymity and analytics access control. In: Trustcom/BigDataSE/ICESS, 2017 IEEE. Piscataway: IEEE; 2017.
29. Schneider C. IBM Blogs; 2016. https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.
30. TCS. Emphasizing the need for government regulations on data privacy; 2016. https://www.tcs.com/content/dam/tcs/pdf/technologies/Cyber-Security/Abstract/Strengthening-Privacy-Protection-with-the-European-General-Data-Protection-Regulation.pdf.
31. He X, et al. Qoe-driven big data architecture for smart city. IEEE Commun Mag. 2018;56(2):88–93.
32. Ramakrishnan R et al. Azure data lake store: a hyperscale distributed file service for big data analytics. In: Proceedings of the 2017 ACM international conference on management of data. New York: ACM; 2017.
33. Beheshti A et al. Coredb: a data lake service. In: Proceedings of the 2017 ACM on conference on information and knowledge management. New York: ACM; 2017.
34. Shang T et al. A DP Canopy K-means algorithm for privacy preservation of Hadoop platform. In: International symposium on cyberspace safety and security. Cham: Springer; 2017.
35. Jia Q et al. Preserving model privacy for machine learning in distributed systems. IEEE Trans Parallel Distrib Syst. 2018;29(8).
36. Psychoula I et al. A deep learning approach for privacy preservation in assisted living. arXiv preprint arXiv:1802.09359. 2018.
37. Guller M. Big data analytics with spark: a practitioner's guide to using spark for large scale data analysis. New York: Apress; 2015.
38. Fung BCM, Wang K, Philip SY. Anonymzing classification data for privacy preservation. IEEE Trans Knowl Data Eng. 2007;19(5):711–25.