ZIBELINE INTERNATIONAL™ PUBLISHING
ISSN: 2616-5961 (Online)
CODEN: IMCSBZ

Information Management and Computer Science (IMCS)

DOI: http://doi.org/10.26480/imcs.01.2023.17.21

RESEARCH ARTICLE

STUDY OF BIG DATA BASED PROBLEMS FOR DATA ANONYMIZATION

Monika Singh

Faculty of information technology, Gopal Narayan Singh University
****Corresponding Author Email: singhmoni@gmail.com***

ARTICLE DETAILS

Article History:

Received 15 December 2022
Revised 18 January 2023
Accepted 21 February 2023
Available online 23 February 2023

ABSTRACT

In order to defeat typical attacks like the similarity attack, the probabilistic inference attack, and others that are possible with anonymized data, we have studied different techniques. To anonymize the data set and disseminate the anonymized data set on a distributed environment without endangering data privacy, a privacy-preserving distributed framework is suggested in most of the techniques. It is possible to achieve a better balance between privacy and data utility, and the data utility is demonstrated in terms of conventional measures. The privacy-preserved data set is also subjected to the application of several classifiers in order to measure the utility of the data When sharing and processing data in a distributed setting or with the Internet of Things, data privacy is a crucial requirement. High communication and computational costs are involved in collaborative privacy-preserving data mining based on secured multiparty computation. Data protection against identity revelation is achieved by the use of data anonymization, a promising technology in the field of privacy-preserving data mining. Anonymization faces significant difficulties, including information loss and frequent attacks that may be made on the anonymized data. Utilizing data mining techniques, data anonymization has recently demonstrated a considerable increase in data value. Still, the methods now in use are ineffective for dealing with attacks. Therefore, a clustering-based anonymization approach that is resistant to similarity attacks and attacks based on inference is suggested in this study. On the Hadoop Distributed File System, the anonymized data is dispersed. The technique creates a better balance between utility and privacy.

KEYWORDS

Communication, framework, Anonymization, crucial requirement

1. INTRODUCTION

Today, there is a natural distribution of data over many geographically dispersed sites. The knowledge derived from the comprehensive collection of data available at these sites is of interest to researchers. Think about a sudden epidemic disease that is sweeping through society, and the researchers are trying to find its causes, signs, and cures. Data mining on patient information available from hospitals across the country or the world would be more effective than data mining on information from a single hospital. Data and other things are shared through a network of objects in the growing field known as the Internet of Things (IoT). IoT has applications in many different industries, including inventory management, traffic control, and patient monitoring. The user's identity and sensitive information pertaining to their interaction and mobility should be protected in all of these applications. As a result, IoT data privacy for individuals is ensured. Moreover, every hospital keeps a copy of each patient's medical record. In accordance with the privacy law, information should be kept private (HIPAA, 1999). This necessitates an efficient mechanism to protect people's privacy while sharing data in distributed environments, the Internet of Things etc., (Yuksel et al., 2017; Sicari et al., 2015; Yao et al., 2015; Henze et al., 2016). To publish their microdata or share their knowledge with third parties, data owners need privacy-preserving data mining or privacy-preserving data publishing techniques (Fung et al., 2010; Fahad et al., 2014; Aggarwal and Yu, 2008).

Some of the key methods used in the field of privacy-preserving data mining or data publishing include data anonymization, data randomization, and cryptography (Sweeney, 1997; Samarati and Sweeney, 1998; Samarati, 2001; Sweeney, 2002; Sweeney, Chen et al., 2007; Chen and Liu, 2009; Chen and Liu, 2011; Islam and Brankovic, 2011; Pinkas, 2002; Liu et al., 2015). K-anonymization is the process of making records anonymous so that k different people can no longer be distinguished from one another. This mechanism guards against identity theft or linkage attacks on the record. The possibility that an attacker will reveal a person's sensitive attribute value in conjunction with the Quasi Identifier (QID) attributes' known values is known as a linking attack or identity disclosure. Anonymization is accomplished by either suppressing or generalising the QID attributes (Sweeney, 2002). By substituting more generalised values for the specific QID values, generalised equivalence classes are created. By substituting a "*" for the original value, suppression completely eliminates the QID values. Generalization is an NP-hard problem, and information is severely lost when it is suppressed. The method of choice to protect health care data from identity disclosure is known as k- anonymization (Reddy and Aggarwal, 2015).

A two-step clustering-based k-anonymization algorithm suitable for both categorical and numerical data was proposed (Han et al., 2014). This technique combines generalisation and microaggregation, two popular techniques for anonymization. Fast data-oriented microaggregation algorithm for anonymization was proposed by Mor-tazavi and Jalili (Mortazavi, 2014). In order to minimise information loss and to satisfy the anonymization parameter k, partitions are created. Gkoulala-Divanis and others demonstrated that anonymization is the best known method to prevent identity disclosure (Gkoulalas-Divanis et al., 2014). The algorithm was proposed to preserve privacy and utility. Wimmera and others suggested using k-anonymization for multiagent privacy-preserving medical data decision making (Wimmera et al., 2016). An adequate level of accuracy is achieved in the anonymization of the data combined from

Quick Response Code	Access this article online	
	Website: www.theimcs.org	**DOI:** 10.26480/imcs.01.2023.17.21

Cite the Article: Monika Singh (2023). Study of Big Data Based Problems for Data Anonymization. *Information Management and Computer Science, 6(1): 17-21.*

several sources. The homogeneity attack, similarity attack, background knowledge attack, and probabilistic inference attack are some of the frequent attacks that can be made against k-anonymization.

The l-diversity principle was put forth in order to include l sensitive values that are well represented in each equivalence class group (Machanavajjhala et al., 2007). To get around the limitations of the l-diversity principle, a group researchers proposed the t-closeness principle. When some sensitive values are more common than others, entropy diversity is impossible to achieve and the distinct l-diversity is insufficient to defend against a probabilistic inference attack (Li et al., 2007). According to the t-closeness principle, each equivalence class should have a sensitive value distribution that is as similar as possible to the distribution of sensitive values in the original data set. But figuring out t-close equivalence classes is practically impossible. A stochastic t-closeness principle combining k-anonymity and -differential privacy was put forth by Domingo-Ferrer and Soria-Comas (Domingo-Ferrer and Soria-Comas, 2015). The authors circumvent the deterministic t-closeness principle's drawbacks. To achieve k-anonymous t-close equivalence classes, three different microaggregation algorithms were put forth (Soria-comas et al., 2015). The clustering algorithm that was used to achieve t-close equivalence among those classes first displays the least amount of information loss. The algorithms only work with numerical data.

The conventional approaches are vulnerable to frequent attacks and fall short of a better privacy-utility trade-off. The accuracy of the knowledge extracted from these data would be im- proved by the use of data mining techniques to achieve anonymization (Ghinita et al., 2011; Kisilevich et al., 2011). With the help of the greedy clustering algorithm, group research proposed an MS (k, 2) privacy model for Adverse Drug Reaction data (Wen-Yang et al., 2016). The algorithm creates anonymized data that satisfies parameter k for anonymization while maintaining the usefulness of the data. Similarity attack is a very popular serious data anonymization issue. It is uncommon to use the available k-anonymization and l-diversity methods to protect data privacy in a distributed setting. Data anonymization would be a comparably better method to achieve privacy in a fully distributed setting (Chen and Liu, 2009). Hence in this paper we intend to employ k-anonymization to preserve the data privacy in a distributed environment. We pro- posed a clustering algorithm to achieve k-anonymization and l- diversity resilient to similarity attack and probabilistic inference attack. Later the privacy-preserved anonymized data sets are dis- tributed on Hadoop (Hadoop, 2012; Polato et al., 2014).

2. Additional Work

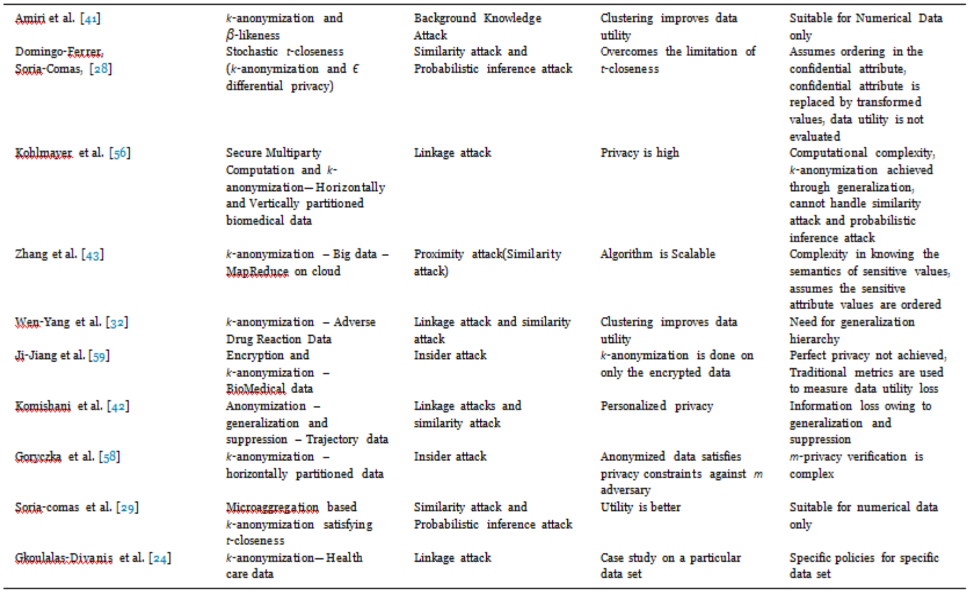
The methods created to defend against similarity attacks and other types of attacks that are possible with anonymised equivalence classes are discussed in this section. The benefits and drawbacks of employing secured multiparty computation with privacy-preserving data mining methods on dispersed data are examined. We have also researched various recent hybrid and anonymization-based methods for protecting data privacy. addressing typical assault techniques. In order to enable data publishing that protects user privacy, (, k) anonymity achieves the k-anonymity property and the - deassociation property. The user defined threshold must be less than or equal to the relative frequency of the most frequent sensitive value in each equivalence class. The authors demonstrate that it is similarly NP-hard to achieve (k, k) anonymization (Wong et al., 2006). For data publishing, a group researchers proposed (p,) and (p,) sensitive k-anonymity (Traian et al., 2007). To produce p different sensitive values in each equivalence class with a total weight of at least, the (p,) sensitive property is proposed. Even this attribute might not be enough to safeguard data from similarity attacks. In order to generate equivalence classes with different sensitive values from each category, [37] first divides the sensitive attribute values into four categories, ranging from Top Secret to Non Secret (Sun et al., 2008).

At least p separate categories of sensitive attribute values with a total weight of at least are proposed for the (p,) sensitive property. Sun et al(L, .'s) diversity privacy models were built on the classification of sensitive attribute values into various levels of confidentiality and the l-diversity. Utilizing functional (, l) diversity improve l-diversity. Until the (, l) diversity is satisfied, the sensitive values and QID characteristics are both generalized (Sun et al., 2011; Tian and Zhang, 2011). All of the user-defined parameters listed in the aforementioned methods are utilised to determine the proper thresholds for the size of the equivalence class and the number of distinct values of the sensitive attribute value. It is impossible to defeat a similarity attack using the current methodologies when the input data set only comprises a small number of possible values for the sensitive property or if some sensitive values occur more frequently than others.

Anonymization based on generalisation with sampling was suggested as a way to hide the adversary's trust in a person's sensitive information (Sattar et al., 2013). Amiri et alk-anonymous .'s -likeness model was put forth to safeguard data from both identity and attribute exposure (Amiri et al., 2016). In order to establish a k-anonymous likeness privacy model and counter a background knowledge attack, the authors have devised two clustering methods. A tailored privacy model based on trajectory data anonymization was suggested (Komishani et al., 2016). While the trajectory data is suppressed, the sensitive attribute is generalized. The technique defeats linking and similarity attacks. In order to protect data privacy in the cloud using anonymization, a group researchers presented a two-phase clustering approach (Zhang et al., 2015). Sensitive attribute values with different semantic diversities are thought to generate equivalence classes.

The privacy of the data in a distributed system was the subject of research at the same time. A cryptographic method called secured multiparty computing is used to protect data privacy against malicious and semi-decent adversaries (Lindell and Pinkas, 2009). In this method, participants only securely transmit the data necessary for a specific distributed computation, such as the frequency of occurrence of an attribute value or statistics for a specific attribute. Many data mining algorithms, including those for horizontally and vertically partitioned databases [45,46], have been improved so that they can be used with distributed databases (Kantarcioglu and Clifton, 2004; Xiao et al., 2006). In a completely distributed environment, suggested secure protocols to carry out frequent itemset mining, Naive Bayes Classification, and ID3 classification (Yang et al., 2005). In the aforementioned class of algorithms, nodes in a distributed environment exchange all of the data. Additionally, the data transferred is restricted to a single data mining task.

Table 1: Comparison of current data mining methods that protect privacy



Amiri et al. [41]	k-anonymization and β-likeness	Background Knowledge Attack	Clustering improves data utility	Suitable for Numerical Data only
Domingo-Ferrer, Soria-Comas, [28]	Stochastic t-closeness (k-anonymization and ϵ differential privacy)	Similarity attack and Probabilistic inference attack	Overcomes the limitation of t-closeness	Assumes ordering in the confidential attribute, confidential attribute is replaced by transformed values, data utility is not evaluated
Kohlmayer et al. [56]	Secure Multiparty Computation and k-anonymization— Horizontally and Vertically partitioned biomedical data	Linkage attack	Privacy is high	Computational complexity, k-anonymization achieved through generalization, cannot handle similarity attack and probabilistic inference attack
Zhang et al. [43]	k-anonymization – Big data – MapReduce on cloud	Proximity attack(Similarity attack)	Algorithm is Scalable	Complexity in knowing the semantics of sensitive values, assumes the sensitive attribute values are ordered
Wen-Yang et al. [32]	k-anonymization – Adverse Drug Reaction Data	Linkage attack and similarity attack	Clustering improves data utility	Need for generalization hierarchy
Ji-Jiang et al. [59]	Encryption and k-anonymization – BioMedical data	Insider attack	k-anonymization is done on only the encrypted data	Perfect privacy not achieved, Traditional metrics are used to measure data utility loss
Komishani et al. [42]	Anonymization – generalization and suppression – Trajectory data	Linkage attacks and similarity attack	Personalized privacy	Information loss owing to generalization and suppression
Goryczka et al. [58]	k-anonymization – horizontally partitioned data	Insider attack	Anonymized data satisfies privacy constraints against m adversary	m-privacy verification is complex
Soria-comas et al. [29]	Microaggregation based k-anonymization satisfying t-closeness	Similarity attack and Probabilistic inference attack	Utility is better	Suitable for numerical data only
Gkoulalas-Divanis et al. [24]	k-anonymization— Health care data	Linkage attack	Case study on a particular data set	Specific policies for specific data set

3. Hybrid Methods

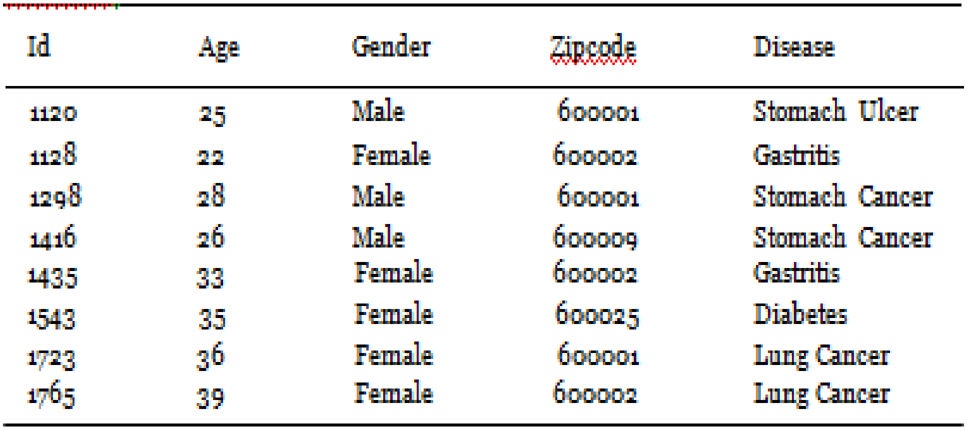
Recently, researchers have focused on hybrid approaches that combine anonymization with other privacy-enhancing methods. An adaptable method to anonymize distributed data was put forth (Kohlmayer et al., 2014). In order to protect the privacy of the data, a combination of secure multiparty computing and anonymization is used. To prevent linking attacks, the final integrated data is anonymized after each participating party sends the encrypted data to the next party. Permutation-based data anonymization was proposed by Domingo-Ferrer and Muralidhar taking into account the data subject, intruder, and transparency of anonymization (Domingo-Ferrer and Muralidhar, 2016). It takes into account the variety of the data that belong to the same equivalence class and can therefore withstand a similarity attack But at the moment, it only applies to numerical data.

For collaborative data publishing proposed an m-privacy model based on anonymization (Goryczka et al., 2014). The anonymized data are protected by the m-privacy model from m adversaries who are aware of both their data and the anonymized data. To maintain data privacy in a cloud environment, proposed a hybrid approach combining encryption and anonymization (Ji-Jiang et al., 2015). Explicit identifiers, quasi-identifier qualities, and medical information make up the three sections of the data. While quasi-IDs and the explicit identifiers are both encrypted, the medical data is available in plain text. However, identity and attribute disclosure will result if the decrypted identifiers or anonymized quasi-identifiers are connected to medical data. To distort the data, proposed a reversible data transform algorithm (Chen-Yi, 2016). Watermarking is then used to determine whether the distorted data has been tampered with.

3.1 Contrasting Modern Methods

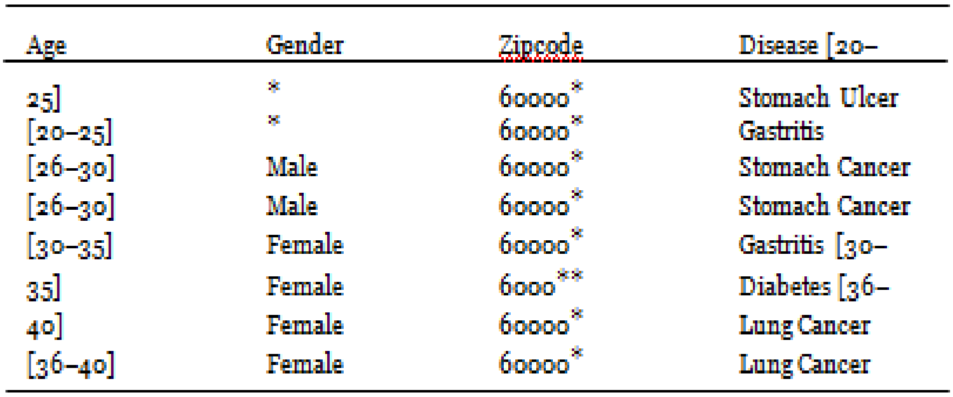
The comparison of some of the existing methods in this field is shown in Table 1. The present methods for protecting data privacy based on k-anonymization, which defeats similarity attacks and probabilistic inference attacks, either apply to numerical attributes or presume that the values of sensitive attribute values are sorted in an inherent way. There might not be any implied ordering between the attribute values in some applications. There are also several hybrid strategies that combine safe multiparty computing and k-anonymization; however these techniques are either computationally challenging or can only withstand linkage attacks on the anonymized data. Additionally, these techniques mainly focus on the secure integration of data from many sources and are unable to shield data against similarity attacks and attacks based on probabilistic inference. In our earlier work, we suggested a privacy strategy to create anonymised equivalence classes with the widest range of sensitive attribute values conceivable. This would defend against similarity attacks on the anonymised data. By creating a uniform distribution of sensitive attribute values in both the original data and the anonymized groups, we hope to safeguard the anonymized data from attacks using probabilistic inference in this study.

Table 2: Microdata



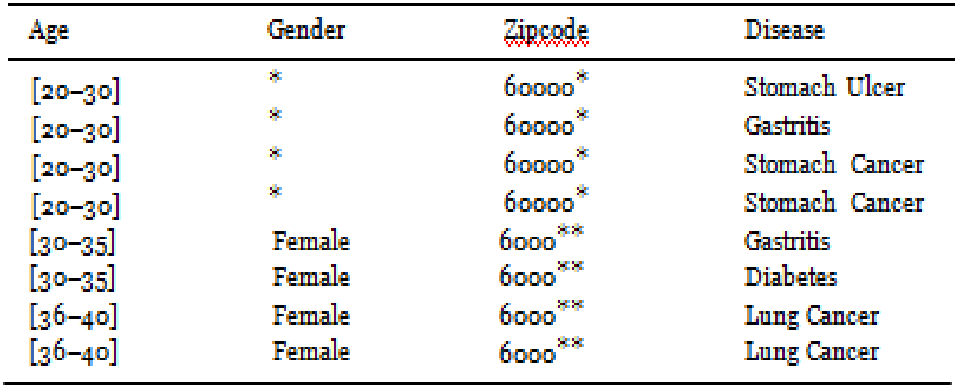
Id	Age	Gender	Zipcode	Disease
1120	25	Male	600001	Stomach Ulcer
1128	22	Female	600002	Gastritis
1298	28	Male	600001	Stomach Cancer
1416	26	Male	600009	Stomach Cancer
1435	33	Female	600002	Gastritis
1543	35	Female	600025	Diabetes
1723	36	Female	600001	Lung Cancer
1765	39	Female	600002	Lung Cancer

Table 3: Anonymous group



Age	Gender	Zipcode	Disease [20–
25]	*	60000*	Stomach Ulcer
[20–25]	*	60000*	Gastritis
[26–30]	Male	60000*	Stomach Cancer
[26–30]	Male	60000*	Stomach Cancer
[30–35]	Female	60000*	Gastritis [30–
35]	Female	6000**	Diabetes [36–
40]	Female	60000*	Lung Cancer
[36–40]	Female	60000*	Lung Cancer

Table 4: Anonymous 3-diverse groups



Age	Gender	Zipcode	Disease
[20–30]	*	60000*	Stomach Ulcer
[20–30]	*	60000*	Gastritis
[20–30]	*	60000*	Stomach Cancer
[20–30]	*	60000*	Stomach Cancer
[30–35]	Female	6000**	Gastritis
[30–35]	Female	6000**	Diabetes
[36–40]	Female	6000**	Lung Cancer
[36–40]	Female	6000**	Lung Cancer

4. Preliminaries

Both public and private attributes can be found in the microdata that data owners have made available. Public qualities are those that are widely recognised and accessible to everyone. The term "QID attributes" refers to characteristics that a hacker uses to reveal a person's identity. The values of sensitive attributes (SA) should remain secret while the data is published. Age, Gender, and Zipcode are QID properties shown in Table 2, while Id is a public attribute and Disease is a sensitive attribute. Using his or her understanding of the QID attribute value of the target, the attacker attempts to reveal the sensitive attribute value of the target. Consider a burglar named A who seeks treatment at a specific hospital because he is acquainted with a 35-year-old woman who lives in his neighbourhood. If he has access to the microdata that was made available by the hospital, he may be able to determine that she has diabetes. Identity disclosure or linkage attacks are terms used to describe this kind of assault.

Table 3 displays the 2-anonymous table created using k-anonymization from the provided microdata. Four classes of equivalence are obtained via generalisation and suppression, each having the same values for the QID characteristics. The values that are suppressed during the anonymization process are denoted by an asterisk (*). Each two-person group defeats linking attack. Now, a 35-year-old female can only be assumed to be diabetic with a 50% likelihood by the same intruder who knows her. She also has an equal likelihood of developing gastritis. Although a connecting attack cannot succeed against Table 3, there is still a potential for the other attacks described below. Homogeneity Attack: The k-anonymization method would be useless if every record in an anonymized group had the same values for the sensitive attribute. Homogeneity attack is the name given to this type of assault.

Homogeneity attack is caused by the second and fourth equivalence classes in Table 3 having the same values for their sensitive property. The diversity of The homogeneity attack is largely defeated by principle. Table 4 displays four groups of three different equivalence classes, each with three different sensitivity values (l = 3). Attack based on similarity: Although the sensitive values in an anonymised group aren't identical, they could be similar to one another. Despite the fact that the values are l-diverse, this will result in the revealing of sensitive attributes. For the sensitive values of both records, the first equivalence class in Table 3 has values that are comparable. A stranger would quickly surmise that the subject is experiencing stomach-related issues. Similar to Table 3, Table 4 also has three diverse sensitive values in the first equivalence class that are all equivalent to one another. Attack using Background Knowledge: An attacker can reduce the number of potential sensitive values in an equivalence class group by using background knowledge about the sensitive attribute or the general features of the material. Background knowledge assault is the term used to describe this type of attack. There is a probability for a probabilistic inference attack when one sensitive attribute value in an equivalence class occurs more frequently than the others. This aids the intrusive party in learning which sickness is more prevalent within a specific set of QID values.

5. Conclusion

By using different data anonymization techniques on, the data owners can share the data in a secure manner. All the sensitive data are corrupted due linking assault, homogeneity attack, similarity attack, and probabilistic inference attack. The Hadoop environment distributes the set of data with privacy protected. T. By generating the ideal number of clusters that satisfy the criteria, a better trade-off between privacy and utility can be made.

References

Aggarwal, C.C., Yu, P.S., 2008. Privacy-Preserving Data Mining: Models and Algorithms, Springer Publication, Berlin. http://dx.doi.org/10.1007/978-0- 387-70992-5.

Amiri, F., Yazdani, N., Shakery, A., Chinaei, A.H., 2016. Hierarchical anonymization algorithms against background knowledge attack in data releasing. Knowl. Based Syst., 101, Pp. 71–89.

Chen, K., Liu, L., 2009. Privacy-preserving multiparty collaborative mining with geometric data perturbation. IEEE Trans. Parallel Distrib. Syst., 20, Pp. 1764–1776. http://dx.doi.org/10.1109/TPDS.2009.26.

Chen, K., Liu, L., 2011. Geometric data perturbation for privacy preserving out- sourced data mining. Springer-Knowl. Inf. Syst., 29, Pp. 657–695. http://dx.doi.org/10.1007/s10115-010-0362-4.

Chen, K., Sun, G., Liu, L., 2007. Towards attack-resilient geometric data perturbation. in: Proceedings of the Seventh SIAM International Conference on Data Mining, Pp. 78–89. http://dx.doi.org/10.1137/1.9781611972771.8.

Chen-Yi, L., 2016. A reversible data transform algorithm using integer transform for privacy-preserving data mining. J. Syst. Softw., 117, Pp. 104–112.

Domingo-Ferrer, J., Muralidhar, K., 2016. New directions in anonymization: Permu- tation paradigm, verifiability by subjects and intruders, transparency to users. Inform. Sci., 337–338, Pp. 11–24.

Domingo-Ferrer, J., Soria-Comas, J., 2015. From t-closeness to differential privacy and vice versa in data anonymization. Knowl.-Based Syst., 74, Pp. 151–158. http://dx.doi.org/10.1016/j.knosys.2014.11.011.

Fahad, A., Tari, Z., Almalawi, A., Goscinski, A., Khalil, I., Mahmood, S., 2014. PPFSCADA: Privacy preservation framework for SCADA data publishing. Future Gener. Comput. Syst., 37, Pp. 496–511.

Fang, W., Yang, B., 2008. Privacy preserving decision tree learning over ver- tically partitioned data, in: Proceedings of the International Conference on Computer Science and Software Engineering, Pp. 1049–1052. http://dx.doi.org/10.1007/11535706_11.

Fung, B.C.M., Wang, K., Chen, R., Yu, S., 2010. Privacy preserving data publishing: A survey on recent developments. ACM Comput. Surv., 42, Pp. 1–53. http://dx.doi.org/10.1145/1749603.1749605.

Ghinita, G., Kalnis, P., Tao, Y., 2011. Anonymous publication of sensitive transactional data, IEEE Trans. Knowl. Data Eng., 23, Pp. 161–174. http://dx.doi.org/10.1109/TKDE.2010.101.

Gkoulalas-Divanis, A., Loukides, G., Sun, J., 2014. Toward smarter healthcare: Anonymizing medical data to support research studies. IBM J. Res. Dev., 58 (1).

Goryczka, S., Xiong, L., Fung, B.C.M., 2014. M-privacy for collaborative data publishing. IEEE Trans. Knowl. Data Eng., 26 (10).

Han, J., Yu, J., Mo, Y., Lu, J., Liu, H., 2014. MAGE: A semantics retaining K-anonymization method for mixed data. Knowl.-Based Syst., 55, Pp. 75–86. http://dx.doi.org/ 10.1016/j.knosys.2013.10.009.

Henze, M., Hermerschmidt, L., Kerpen, D., Häußling, R., Rumpe, B., Wehrle, K., 2016. A comprehensive approach to privacy in the cloud-based Internet of things. Future Gener. Comput. Syst., 56, Pp. 701–718.

HIPAA. 1999. Health Insurance Portability and Accountability Act of 1999. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule (accessed 20.06.15).

Islam, M.Z., Brankovic, L., Privacy preserving data mining: A noise addition framework using a novel clustering technique. Knowl. -Based Syst., 24, Pp. 1214–1223.

Ji-Jiang, Y., Jian-Qiang, L., Niu, Y., 2015. A hybrid solution for privacy preserving medical data sharing in the cloud environment. Future Gener. Comput. Syst., 43–44, Pp. 74–86.

Kantarcioglu, M., Clifton, C., 2004. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. Knowl. Data Eng., 16, Pp. 1022–1037. http://dx.doi.org/10.1109/TKDE.2004.45.

Kantarcioglu, M., Vaidya, J., 2003. Privacy preserving Naive Bayes classifier for horizontally partitioned data. in: IEEE Workshop on Privacy Preserving Data Mining, Pp. 3–9.

Kisilevich, S., Rokach, L., Elovici, Y., Shapira, B., 2011. Efficient multidimensional suppression for K-anonymity. IEEE Trans. Knowl. Data Eng., 22, Pp. 334–347. http://dx.doi.org/10.1109/TKDE.2009.91.

Kohlmayer, F., Prasser, F., Eckert, C., Kuhn, K.A., 2014. A flexible approach to distributed data anonymization. J. Biomed. Inform., 50, Pp. 62–76.

Komishani, E.G., Abadi, M., Deldar, F., 2016. PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression, Knowl. Based Syst., 94, Pp. 43–59.

Li, N., Li, T., Venkatasubramanian, S., 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. in: IEEE International Conference on Data Engineering, Pp. 106–115. http://dx.doi.org/10.1109/ICDE.2007.367856.

Lindell, Y., Pinkas, B., 2009. Secure multiparty computation for privacy-preserving data mining. J. Priv. Confidentiality, 1, Pp. 59–98.

Liu, H., Huang, X., Liu, J.K., 2015. Secure sharing of personal health records in cloud computing: Ciphertext-policy attribute-based signcryption. Future Gener. Comput. Syst., 52, Pp. 67–76.

Machanavajjhala, A., Gehrke, J., Kifer, D., Venkatasubramaniam, M., 2007. l-diversity: Privacy beyond k-anonymity, ACM Trans. Knowl. Discov. Data 1, Pp. 1–47. http://dx.doi.org/10.1145/1217299.1217302.

Mortazavi, R., Jalili, S., 2014. Fast data-oriented micro aggregation algorithm for large numerical datasets, Knowl. Based Syst., 67, Pp. 195–205. http://dx.doi.org/ 10.1016/j.knosys.2014.05.011.

Pinkas, B., 2002. Cryptographic techniques for privacy-preserving data mining. ACMSIGKDD Explor., Newsl., 4, Pp. 12–19. http://dx.doi.org/10.1145/772862.

Polato, I., Re, R., Glodman, A., Kon, F., 2014. A comprehensive view of hadoop research– a systematic literature review, J. Netw. Comput. Appl., 46, Pp. 1–25. http://dx.doi.org/10.1016/j.jnca.2014.07.022.

Reddy, C.K., Aggarwal, C.C., 2015. Healthcare Data Analytics, in: Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC.

Samarati, P., 2001. Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng., 13, Pp. 1010–1027. http://dx.doi.org/10.1109/69. 971193.

Samarati, P., Sweeney, L., 1998. Protecting privacy when disclosing information: k- anonymity and its enforcement through generalization and suppression, in: Proceedings of IEEE Symposium on Research in Security and Privacy, Pp. 188–206.

Sattar, A.H.M.S., Li, J., Ding, X., Liu, J., Vincent, M., 2013. A general framework for privacy preserving data publishing, Knowl. Based Syst., 54, Pp. 276–287. http://dx.doi.org/10.1016/j.knosys.2013.09.022.

Sicari, S., Rizzardi, A., Grieco, L.A., Coen-Porisini, A., 2015. Security, privacy & trust in the Internet of things: The road ahead. Comput. Netw., 76, Pp. 146–164.

Soria-comas, J., Domingo-Ferrer, J., Sanchez, D., Martinez, S., 2015. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. IEEE Trans. Knowl. Data Eng., 27 (11), Pp. 3098–3110.

Sun, X., Li, M., Wang, H., 2011. A family of enhanced (L, α) diversity models for privacy preserving data publishing. Elsevier J. Future Gener. Comput. Syst., 27, Pp. 348–356. http://dx.doi.org/10.1016/j.future.2010.07.007.

Sun, X., Wang, H., Li, J., Truta, T.M., 2008. Enhanced P-sensitive K-Anonymity models for privacy preserving data publishing. Trans. Data Priv., 1, Pp. 53–66.

Sweeney, L., 1997. Datafly: A system for providing anonymity in medical data, in: Proceeding of Eleventh International Conference on Database Security, Pp. 356–381.

Sweeney, L., 1997. Guaranteeing anonymity when sharing medical data, the Datafly system. in: Proceedings of AMIA Annual Fall Symposium American Medical Informatics Association, Pp. 51–55.

Sweeney, L., 2002. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl -Based Syst. 10, Pp. 571–588. http://dx.doi.org/10.1142/S021848850200165X.

Sweeney, L., 2002. K-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst., 10, Pp. 557–570.

Tian, H., Zhang, W., 2011. Extending l-diversity to generalize sensitive data. Elsevier J. Data Knowl. Eng., 70, Pp. 101–126. http://dx.doi.org/10.1016/j.datak.2010. 09.001.

Traian, T.M., Campan, A., Meyer, P., 2007. Generating microdata with P-sensitive k-anonymity property, in: Secure Data Management. in: Lecture Notes in Computer Science, 4721, Pp. 124–141.

Vaidya, J., 2004. Privacy Preserving Data Mining over Vertically Partitioned Data (Dissertation), Purdue University.

Vaidya, J., Clifton, C., 2002. Privacy preserving association rule mining in vertically partitioned data, in: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pp. 639–644. http://dx.doi.org/10.1145/775047.775142.

Vaidya, J., Clifton, C., 2003. Privacy preserving K-means clustering over vertically partitioned data, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pp. 206–215. http://dx.doi.org/10.1145/956750.956776.

Vaidya, J., Clifton, C., 2009. Privacy-preserving Kth element score over verti- cally partitioned data. IEEE Trans. Knowl. Data Eng., 21, Pp. 253–258. http://dx.doi.org/10.1109/TKDE.2008.167.

Vaidya, J., Kantarcioglu, M., Clifton, C., 2008. Privacy-preserving Naive Bayes classifier. VLDB J., 17, Pp. 879–898. http://dx.doi.org/10.1007/s00778-006-0041-y.

Wen-Yang, L., Duen-Chuan, Y., Jie-Teng, W., 2016. Privacy preserving data anonymiza- tion of spontaneous ADE reporting system dataset, BMC Med. Inf. Decis. Mak.

White, T., 2012. Hadoop: The Definitive Guide, O'Reilly Media Publishers.

Wimmera, H., Yoon, V.Y., Sugumaran, V., 2016. A multi-agent system to support evidence-based medicine and clinical decision making via data sharing and data privacy. Decis. Support Syst., 88, Pp. 51–66.

Wong, R.C., Li, J., Fuand, A.W., Wang, K., 2006. (α, k) Anonymity: An enhanced k- anonymity model for privacy-preserving data publishing, in: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pp. 733–744. http://dx.doi.org/10.1145/1150402. 1150499.

Xiao, M.J., Han, K., Huang, L.S., Li, J.Y., 2006. Privacy preserving C4.5 algorithm over horizontally partitioned data, in: Proceedings of the Fifth Interna- tional Conference on Grid and Cooperative Computing, Pp. 78–85. http://dx.doi.org/10.1109/GCC.2006.73.

Yang, Z., Zhong, S., Wright, R.N., 2005. Privacy-preserving classification of customer data without loss of accuracy. in: Proceedings of the Fifth SIAM International Conference on Data Mining, Pp. 92–102.

Yao, X., Chen, Z., Tian, Y., 2015. A lightweight attribute-based encryption scheme for Internet of things. Future Gener. Comput. Syst., 49, Pp. 104–112.

Yu, H., Vaidya, J., Jiang, X., 2006. Privacy-preserving SVM classification on vertically partitioned data, in: Proceedings of the 10th Pacific-Asia Conferences on Advances in Knowledge Discovery and Data Mining, Pp. 647–656. http://dx.doi.org/10.1007/11731139_74.

Yuksel, B., Kupcu, A., Ozkasap, O., 2017. Research issues for privacy and security of electronic health services. Future Gener. Comput. Syst., 68, Pp. 1–13.

Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., Chen, J., 2015. Proximity-aware local recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. IEEE Trans. Comput., 64, Pp. 8.