ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun1862

A Survey: Privacy Preservation in Data Publishing

Hanumanthappa S[1]; Prashantha S J[2]; Vishwanath B R[3]; Sathisha M S[4]

[1]Professor, [2]Professor, [3]Associate Professor, [4]Professor
[1]Dept. of ISE, Kalpataru Institute of Technology, Tiptur, Karnataka, India
[2]Dept. of AI & DS, Rajeev Institute of Technology, Hassan, Karnataka, India
[3]Dept. of E & CE, Rajeev Institute of Technology, Hassan, Karnataka, India
[4]Dept. of AI & ML, Navkis College of Engineering, Hassan, Karnataka, India

Publication Date: 2025/07/07

Abstract: Many organizations like small and medium business (SMB), the datasets are being actively collected and stored by businesses. The majority of them have acknowledged the potential significance of this data as a source of information for corporate decision-making. Privacy preservation in data publishing is required to protecting sensitive information. There are several ways that the personal data might be utilized improperly. This study presents a brief overview of a number of strategies, including generalization and bucketization, both of which have been developed for privacy preservation in micro data publishing. Recent research has demonstrated that generalizing to high-dimensional data will result in significant information loss, and bucketization doesn't prevent membership disclosure, so it can't be applied to data where there isn't a distinct distinction between sensitive and quasi-identifying attributes. The generalization and bucketization approaches for anonymization are designed to protect your privacy when creating micro data. These methods can be applied to privacy preservation in data publishing. Also, we look at a game theory model and compare the RSA and ECC algorithm.

Keywords: *Privacy Preservation, Data Publishing, Data Mining, ECC Algorithm.*

How to Cite: Hanumanthappa S; Prashantha S J; Vishwanath B R; Sathisha M S (2025) A Survey: Privacy Preservation in Data Publishing. *International Journal of Innovative Science and Research Technology*, 10(6), 2669-2673. https://doi.org/10.38124/ijisrt/25jun1862

I. INTRODUCTION

The field of data mining is multidisciplinary which includes wide areas such as Information retrieval, High performance computing and data visualization, knowledge-based system, machine learning and database technology. In order to keep information in the sector, data mining has been used, because knowledge is widely available and because there is always a demand, information are transformed into valuable data and information, information acquired could be utilized for different applications going from market investigation, misrepresentation location and client maintenance. Since there is so much information, The industry's dedication to privacy has been seen as being threatened by data mining. Preserving individual information is required for owners of data to protect his privacy and plays a significant part in data publishing [2].

Now days, due to increased possibility of storing personal data of users and giving more importance to the privacy preservation in data mining [1]. Sharing information that contains sensitive personal information violates an individual's privacy and creates serious problems for the privacy of an individual's sensitive information. Privacy preservation is a method of publishing data while individual privacy can be preserved [3].

Today, Data publishing can be used for making strong consumer relationship with many organizations like marketing, retail and financial etc. Data privacy, also known as information privacy, is a branch of security that focuses on personal data. It regulates the gathering, sharing, and use of sensitive data, including financial and intellectual property data, and ensures that it is handled properly. Data privacy concerns might result from information from a variety of sources, such medical records. Financial institutions and transactions, Criminal justice investigations and processes, biological characteristics, such as genetic information, regional data, and place of residence Geo location and location-based services, invasion of privacy Using permanent cookies and academic research, determine user preferences or web browsing habits [6].

The acquired data and its transactions are someplace documented in this information era. The privacy of the individual is protected by a variety of data security-enhanced

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun1862

approaches that have been created for the purposes of data collection and data mining. When data has been collected, there is desire for it to be shared and published among multiple parties. The data would be gathered from various sources and couldn't be directly shared. In order to mine the data and transfer it to the receiver, the owner may gather data and analyze it using different anonymization techniques. To protect privacy and confidentiality, the obtained data would be prepared in advance of publication [2].

Micro data is included in data publication to protect privacy. We assume that micro data is kept in a table with one identity for each entry (row). Each entry has the following form: D (Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes).

Identifier (I): Some attributes in the published table clear that uniquely identifies an individual. For Examples name and social security number.

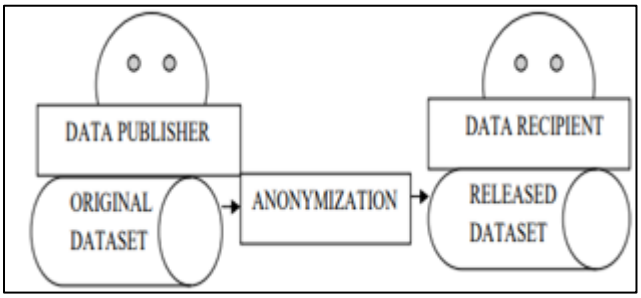

Fig 1: Simple Privacy Model

Sensitive Attribute (SA): attributes contain some sensitive information about a person, such as their salary, disability status and disease. These attributes unknown to adversary.

Non-Sensitive Attributes (NSA): Attributes which do not fall into the I, QA, SA.

In most publications, it is believed that the four groups of attributes are not connected. The bulk of works make the assumption that each table row has a separate owner.

In the cryptographic field, privacy has traditionally been studied. The designed interest of things resulted from recent research on data publishing. In Section: II. We will describe numerous methods for privacy preservation in data publishing and provide an overview of them. Section: III. The Elliptic Curve Cryptography algorithm is described. Section: IV. Concludes this paper.

II. RELATED WORK

A. Generalization

One common anonymized method is generalization, which substitutes less precise but semantically equivalent values for quasi-identifiers. To minimize the level of representational granularity, the values in this case are generalized to a range. The group's QI values would all be generalized to the group's full extent in QID space[6]. The data analyst must make the uniform distribution assumption while doing data analysis on the generalized table since no alternative distribution assumption can be justified. Because each attribute is generalized separately, correlations between various attributes suffer, which significantly reduces the data utility of the generalized data. be supported by evidence. As each feature is generalized independently, relationships between various attributes are lost, which dramatically diminishes the data value of the generalized data [7]. The data analyst must make the assumption that any potential combination of attribute values is equally feasible in order to examine attribute co-relation. Since no alternative distributional premise can be supported. When each feature is generalized independently, correlations across several characteristics suffer, which dramatically diminishes the data value of the generalized data. be supported by evidence. As each feature is generalized independently, relationships between various attributes are lost, which dramatically diminishes the data value of the generalized data. The data analyst must make the assumption that any potential combination of attribute values is equally feasible in order to examine attribute co-relation. Records belonging to the same bucket should be close to one another for generalization to work well and prevent information loss.

➢ *Limitations:*

- Due to dimensionality's curse, it fails with high-dimensional data [8].
- Because of the assumption of uniform distribution, too much information is lost.

B. Bucketization

Bucketization is to divide the tuples in 'T' into buckets, then isolate the non-sensitive ones by randomly permuting the values of the sensitive attribute in each bucket. After that buckets containing permuted sensitive values make up the sanitized data. And using bucketization we are constructing published data from original table. Yet, all of our conclusions also hold for full-domain generalization. We now formalize the definition of our bucketization concept. A collection of buckets each containing permuted SA values make up the anonymized data. Bucketization, high-dimensional data has been anonymized in particular through bucketization. Nonetheless, their approach clearly distinguishes between quasi-identifiers and sensitive attributes.

➢ *Limitations:*

- It does not prevent disclosure of membership. As bucketization can determine if a person has a record of I disseminated information or not and distributes QI values according to their specific structures.
- The attribute co-relation between QIs and SAs is broken by bucketization, which separates SA attributes from QI attributes.
- The separation of QIs and SAs is necessary for bucketization. Nevertheless, many data sets do not make it apparent which qualities are QIs and which are SAs.

C. Slicing

Slicing is a brand-new method that divides data both horizontally and vertically. Generalization does not preserve data with the same utility. Compared to bucketization, it

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun1862

preserves more attribute relationships with the SAs. It can also handle data with high dimensions and data without a clear separation between sensitive attributes and quasi-identifiers.

We think about protecting against attribute and membership disclosure during the slicing process. Due to the fact that each tuple is contained within a bucket, it is unclear which identity disclosure policy should be applied to the sliced data. The associations between the various columns are hidden within a bucket. Since attribute disclosure follows identity disclosure.

Table 1: Original data.

Age.	Sex.	Zip code.	Disease.
23	M	45706	gastric
23	F	45706	pneumonia
34	M	45705	gastric
50	M	45705	dyspepsia
52	F	45601	flu
55	F	45601	flu
55	F	45602	pneumonia
60	M	45603	pneumonia

Table 2: Sliced Data.

(Age, Sex).	(Zip code, Disease)
(23, M) (23, F) (34, M) (50, M)	(45706, gastric) (45706, pneumonia) (45705, gastric) (45705, dyspepsia)
(52, F) (55, F) (55, F) (60, M)	(45601, flu) (45601, flu) (45602, pneumonia) (45603, pneumonia)

➢ *Limitations:*

- We take into account exactly one column of slicing for each attribute; The concept of overlapping slicing, which repeats an attribute in more than one column, is an expansion. As a result, more attribute correlations are produced. Therefore, privacy implications must be thoroughly investigated and comprehended.
- Since random grouping is ineffective, we intend to develop tuple grouping algorithms that are more efficient. As a result, we investigate membership disclosure security in depth.
- By separating the relationship of uncorrelated characteristics by splitting attributes into columns, we secure privacy while keeping the value of the data by maintaining the association between highly correlated attributes.
- This may lose data utility because we generate associations between bucket column values at random.

D. K- Anonymity

The accuracy of the published data is guaranteed by K-Anonymity. The K-anonymity proposal focuses specifically on two methods: suppression and generalization [9]. Data owners usually remove or encrypt explicit identifiers like names and social security numbers before sharing micro data to safeguard respondents' identities. However, de-identifying data do not guarantee anonymity. K-anonymity, which has recently been proposed as One of the new ideas in micro data protection is a property that encapsulates the defense of a micro data table against the potential re-identification of the responders to whom the data pertain. The first approach to privacy preservation in data publishing was changing the input before it was mined. The development of a second methodology for data mining that protects privacy involved the use of cryptographic methods.

➢ *Limitations:*

- It doesn't mask if a person is stored in the database.
- It displays an individual's sensitive attributes
- It is incapable of defending against threats based on prior information.
- Privacy can be violated just by knowing about the k-anonymization algorithm.
- It cannot be used on highly dimensional data without losing all of its usefulness.
- Additional approaches are needed if the dataset is anonymized and released more than once.

E. Game Theory

Game Theory gives a conventional way to deal with model circumstances where a gathering of records (agents) needs to pick ideal activities thinking about the shared impacts of different specialist's choices. Players, actions, payoffs, and information are a game's essential components [10]. Throughout the game, there are actions that players can perform at predetermined times. Players receive rewards for the actions they take. The outcome for each player is determined by their own actions as well as those of the other players. Modeling information makes use of the notion of an information set, which symbolizes a player's comprehension of the values of numerous game variables. The outcome of the game is a collection of items selected from the values of the actions, payoffs, and other factors. If a player acts in a way that maximizes his payoff, he or she is considered rational. A player's strategy is a guide that instructs him on given his information set, what to do at each moment of the game. A strategy profile is an ordered set of tactics with one strategy for each gamer. A strategy profile that includes the best strategy for each player in the game is called equilibrium. Nash equilibrium serves as the game's primary guiding principle for harmony. If no other players depart from a strategy profile and no player has an incentive to do so, it is a Nash equilibrium.

Game theory has been used effectively in a variety of disciplines, like computer science, political science, and economics, among others. Data publishing-related privacy concerns have been addressed using this game theory strategy.

The research paradigm of the game theory approach is as follows:

- Define the game's components: specifically, players, actions, and payoffs.
- Identify the type of game: Information that is either complete or not at all.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun1862

- Resolve the game to identify equilibriums.
- Identify the equilibriums to explore some relevant applications.

While real-world issues can be extremely complex, the preceding paradigm appears to be clear and straightforward. Data publishing involves the user roles of data supplier and data collector. A data collector may need to negotiate if they want to get information from data suppliers who place a high value on their private information. In information distributing, information client or digger, who needs to purchase an informational collection from the information gatherer, makes a value deal to the gatherer at start of the game. In the event that the data collector accepts the proposal, he will reward the publisher of the data in exchange for their private information. Before selling the data to the user, the data collector uses anonymization techniques to the data in order to protect the data providers' privacy to some level. The data miner asks for a degree of privacy protection since he knows the data will be anonymized and wants to balance the amount and quality of the data. Also, the data collector informs data sources of the extent of privacy protection.

A data provider bases his decision on the degree of protection and incentives offered by the data collector. The degree of privacy protection affects each player's actions and rewards in the game of data collecting. By resolving the subgame perfect Nash equilibrium of the proposed game, a consensus on the degree of privacy protection is possible. This technique might be useful for applications that require aggregate queries. They show that resolving game equilibriums can provide stable combinations of revelation level, data retention period, data item price, and data provider incentives.

➢ *Limitations:*

- Demonstrates no personal privacy protection. i.e., privacy has different repercussions for different people. For instance, some people consider salary information to be private, while others do not. Some people are more concerned with privacy than others. When anonymizing the data, the "Personality" of privacy has to be considered as a result.
- Inability of the adversary to use background knowledge, which allows the adversary to use a variety of types of knowledge to extract the target's information from published data. Published data can be useful if the data collector lacks a clear understanding of the adversary's capabilities, such as the knowledge the adversary can learn from other sources and how that knowledge can be applied to make inferences about the target's data. The adversary will almost probably de-anonymize the material. So, the information gatherer has to conduct a comprehensive examination of the adversary's background information and develop legal models to formalize the assaults in order to build a successful protection model for thwarting various assaults.

F. Elliptic Curve Cryptography

An effective method for cryptography is ECC, an alternative to RSA. The math of elliptic bends is used to establish security between key matching for public key encryption. Due to its ability to retain security and decreased key size, ECC has quietly gained favour recently. In contrast, RSA performs a similar function using prime numbers rather than elliptic curves. This tendency is expected to continue as the demand for secure devices develops due to the rising size of keys and the associated strain on constrained mobile resources. Understanding the context of elliptic curve cryptography is crucial for this reason. In contrast to RSA, ECC uses algebra to arrange elliptic curves across finite fields in order to approach public key cryptography schemes. ECC generates keys that are consequently harder to decrypt theoretically. As a result, ECC is thought of as the public key cryptography's next generation and is considered to be more secure than RSA.

ECC may be used to maintain high levels of performance and security. An elliptic curve, for the purposes of the present ECC, is a plane curve over a finite field made up of points that meet the following equation: In this cryptography example, any point on the elliptic curve can be mirrored along the x-axis and the outcome will still be the same: $y2 = x3 + axe + b$. Any non-vertical line will encounter the curve in no more than three places.

The size and security yield of the RSA and ECC encryption keys are very different. The table 3 lists the key sizes needed to retain the same degree of security. Or to put it another way, a 7680-bit RSA key is as secure as a 384-bit elliptic curve cryptography key.

Table 3. Key Length Comparison of RSA and ECC

Security Level(bits)*	**RSA key size(bits)***	**ECC key size(bits)***
80	1024	160-223
112	2048	224-255
128	3072	256-283
192	7680	384-511
256	15360	512-571

III. CONCLUSION

In this work, we presented an overview of many methods for privacy preservation in data publishing; also, various anonymization methods, like generalization and bucketization Slicing and game theory methods to distributed privacy-preserving data publishing are also discussed. The proposed algorithm Elliptic curve cryptography is more efficient than any other encryption algorithm and ECC guarantees quicker encryption and decryption, as well as effectiveness even on tiny devices and providing more security.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun1862

REFERENCES

[1]. CH. Srikanth Reddy, MD. John Saida" A Novel Approach to Privacy Preserving Data Publishing Using Slicing Technique" International Journal of Research Studies in Science, Engineering and Technology Volume 1, Issue 8, November 2014, PP 55-63.

[2]. S. Gokila, Dr. P. Venkateswari "A SURVEY ON PRIVACY PRESERVING DATA PUBLISHING", International Journal on Cybernetics & Informatics (IJCI) Vol. 3, No. 1, February 2014.

[3]. V. Arul, C. Vairavel, M. Prakash and N.V. Kousik "Privacy Preservation Of Micro Data Publishing Using Fragmentation " Ictact Journal On Soft Computing, April 2019, Volume: 09, Issue: 03. Doi: 10.21917/Ijsc.2019.0271

[4]. Shah, Hitarth, Kakkad, Vishruti, Patel, Reema, Doshi, Nishant" A survey on game theoretic approaches for privacy preservation in data mining and network security". 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019), Volume 155, page 686-691, https://doi.org/10.1016/j.procs.2019.08.098

[5]. Qinghai Liu, Hong Shen, Ying Peng Sang "A Privacy-preserving Data Publishing Method for Multiple Numerical Sensitive Attributes via Clustering and Multi-Sensitive Bucketization" 2014 Sixth International Symposium on Parallel Architectures, Algorithms and Programming2168-3034/14© 2014 IEEE.

[6]. Amar Paul Singh, Ms. Dhanshri Parihar" A Review of Privacy Preserving Data Publishing Technique" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-2, Issue-6).

[7]. Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[8]. G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[9]. Benjamin C M Fung, Ke Wang, Ada wai-cheefu and Philip S Yu, "Privacy preservation in data publishing concepts and techniques", Data mining and knowledge discovery series (2010).

[10]. E. Rasmusen, "Games and Information: An introduction to game theory", vol.2, Cambridge, MA, USA: Blackwell,1994

[11]. https://avinetworks.com/glossary/elliptic-curve-cryptography.

[12]. Dindayal Mahto, Dilip Kumar Yadav "RSA and ECC: A Comparative Analysis" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 19 (2017) pp. 9053-9061.

[13]. Xiaokui Xiao, Yufei Tao "Personalized Privacy Preservation" SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA.

[14]. Jianfeng Xia, Min Yu1, Ying Yang, Hao Jin "Personalized Privacy-Preserving with high performance: anonymity" 2018 IEEE Symposium on Computers and Communications (ISCC) 978-1-5386-6950-1/18.

[15]. Jasmina N Vanasiwala, Nirali R Nanavati" Privacy preserving data publishing of multiple sensitive attributes by using various anonymization techniques "Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020) ISBN:978-1-7281-4889-2.

[16]. Michael Wooldridge, "AI and Game Theory "University of Oxford: 1541-1672/12 © 2012 IEEE.

[17]. E. Poovammal, Dr. M. Ponnavaikko "Dr. M. Ponnavaikko" 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.

[18]. www.wikipedia.org

[19]. Ge, YF., Wang, H., Cao, J. *et al.* Privacy-preserving data publishing: an information-driven distributed genetic algorithm. *World Wide Web* **27**, 1 (2024). https://doi.org/10.1007/s11280-024-01241-y

[20]. Nidhi Desai, Manik Lal Das, Payal Chaudhari, Naveen Kumar, "Background knowledge attacks in privacy-preserving data publishing models, Computers & Security" Volume 122,2022,102874, ISSN 0167-4048,https://doi.org/10.1016/j.cose.2022.102874.