

SLICING: A NEW APPROACH TO PRIVACY PRESERVING DATA PUBLISHING

¹Chinthagunta.Anitha, ²Mr.B.Balaji

¹(Student, Dept. of Master of Computer Applications, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.

Email: chinthaguntaanitha@gmail.com)

²(Asst.Prof, Dept. of Computer Science & Engineering, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.

Email: bbalaji0407@gmail.com)

Abstract:

In the last decade, anonymization techniques such as generalisation, bucketization and slicing have appeared in privacy-preserving data publishing. Protecting identity privacy means using generalisation, but this often lowers the usefulness of the data, particularly with datasets containing many features. Although useful for privacy, bucketization can still lead to membership disclosure and becomes less reliable when it's hard to separate quasi-identifying from sensitive attributes. For dealing with these restrictions, a new method referred to as slicing is now introduced. Both horizontal and vertical partitioning of data save more useful information than if the data is generalised and also keep information about membership private. Besides, slicing is equipped to handle lots of data features and matches ℓ -diversity standards for confidentiality. A good algorithm for doing sliced data lines up with privacy rules. We found that slice units work better than generalised units at keeping the importance of data when dealing with sensitive attributes. In addition, the use of slicing ensures that no connexion can reveal membership. This work adds the t-closeness slicing (TCS) method which helps secure transactional data, paying special attention to issues of membership privacy, identity privacy and attribute privacy. On large datasets, TCS efficiently works using the time complexity $O(n \log n)$. The results from three different transactional datasets demonstrate that TCS improves privacy and gives better value from data than other approaches.

Keywords — *Privacy-preserving data publishing; transactional data; slicing technique; t-closeness slicing (TCS);*

1. Introduction

1.1 Overview

Because of more concerns about data leaks, it is now particularly important to ensure the private publishing of microdata. Traditional methods for hiding data such as generalisation and bucketization as part of k-anonymity and ℓ -diversity, do not protect or provide useful results for high-dimensional data knowledge.

By generalising, birthdate, zip code and sex are changed to avoid anyone being identified. As a result, both information and the connexion between attributes are usually reduced. Bucketization stops all QIs together, yet separates SAs so that we can learn if a person belongs to the data.

To handle these problems, we propose slicing which cuts up data tuples and attributes. It works better at preserving activity patterns, stronger

repeatability and safer privacy than regularisation or sketching.

1.2 Motivation for Slicing

Because high-dimensional data leads to the curse of dimensionality, generalisation requires all attributes to have the same value and ignores connexions between them. Bucketization works well in some cases but remains unable to stop membership disclosure when other people's knowledge is public. Three of the QIs are sufficient to distinguish 85% or more of Americans, according to research.

By using a slicing approach, these matters are solved directly.

- Putting groups of highly connected attributes together in one column (so the key connexions are upheld).
- Sorting data within groups to keep people from being matched with particular records
- Using ℓ -diverse slicing so there is strong confidence no sensitive value will be revealed

1.3 Key Contributions

- Slicing was introduced as a privacy-preserving tool that suits dataset with many dimensions and is complex.
- We compared data over ℓ -diverse slices to hide important attributes.
- Created a good slicing algorithm that helps to keep the results useful while protecting privacy.
- Revealed that slicing a table can help by masking how information relates to other records with a range of matching combinations.

- In practise, clustering showed up better than generalisation and bucketization and even surpassed the original data in some cases because it helped reduce overfitting.

2. Literature Review

With big data, lots of data on transactions is created today from activities such as webpage visits, online shopping, reservations, posting reviews of products and receiving medical treatments. McKinsey Global Institute points out (Woetzel et al., 2014) that sharing such data leads to valuable results such as rapid development of new drugs and enhanced service to consumers based on analysing patterns.

There is a significant privacy problem associated with sharing data. Due to the anonymous logs released by AOL in 2006, we found information about user No. 4417749 (Barbaro & Zeller, 2006). A similar weakness was found in Netflix's use of anonymous ratings: it inadvertently helped to reveal who users were (Narayanan & Shmatikov, 2008).

According to Loukides et al. (2010), almost every patient in a medical dataset could be distinguished by diagnosis codes, even when the group contained more than a million people. As a result, publishing transactional information must prioritise privacy, as any leaks could risk individuals and expose the owners of the data.

2.1 Privacy Threats:

A retail transactional dataset is used to explain both common privacy problems and their solutions. The processing of algorithms that guard privacy is made faster by converting records to a binary format. A row is the details for a purchase and a column lists all the items included. During any transaction, the cell's value is 1 only if the item appears and 0 if it is not present.

Releasing this kind of transaction data, while anonymizing named individuals, means there's a

risk of members being exposed, identities being disclosed or attributes being revealed.

Let’s quickly look at what these three threats are about:

- Attackers are able to see if a person’s information is part of the data in the dataset.
- Reidentification: A person is recognisable from the information in the data.
- Attribute Disclosure: An intruder gets access to some of an individual’s personal sensitive data, although the record cannot be identified.

Name	Apple (I1)	Egg (I2)	Milk (I3)	Sugar (I4)	Towel (I5)	Soap (I6)	Pregnancy Test Kit
Alice	1	1	1	0	0	0	0
Bob	1	0	0	0	1	1	0
Cindy	1	0	0	0	1	0	1
Elena	0	1	1	1	0	0	0
Frank	0	0	0	1	1	1	0
Gilbert	1	1	1	0	1	1	1
Helen	1	1	1	1	0	0	1
Ian	0	0	0	0	1	1	1
Jimmy	0	1	1	1	0	0	1

Fig1: Transactional-Dataset

Those risks are all linked together. When identity is disclosed, both the individual’s inclusion in the dataset and the values of their sensitive information become public. But when membership disclosure happens, the attacker doesn’t yet know who the victim is. Things can get sensitive if multiple individuals with the same attributes purchase an item—the adversary will learn something, even if they don’t identify who they are.

Threats	Definitions	Examples
Membership Disclosure	An adversary infers that a data subject’s record is included in a published dataset.	An adversary infers that Cindy’s transaction is included in the dataset; he may or may not know the exact record.
Identity Disclosure	An adversary infers that a data subject is linked to a specific record in the published dataset with a high probability.	An adversary knows with high confidence that the third record of the dataset is Cindy’s transaction.
Attribute Disclosure	New information about a data subject is revealed in the published dataset.	An adversary infers that Cindy purchased a Pregnancy Test Kit.

Fig2: Privacy Threats- commonType

3. Project Description

3.1 PROBLEM DEFINITION

The UL (Utility-Privacy Layered) method is presented here to help guard data privacy without affecting its usefulness when it is published. Ensuring no one can link data from different public datasets is the challenge the initiative hopes to solve.

The main aim is to get to a required level of privacy and also to make sure data is still good for mining and analysis. The proposed UL method is organised into four main steps, all of which help balance privacy and utility. The different stages are explained in the next set of subsections.

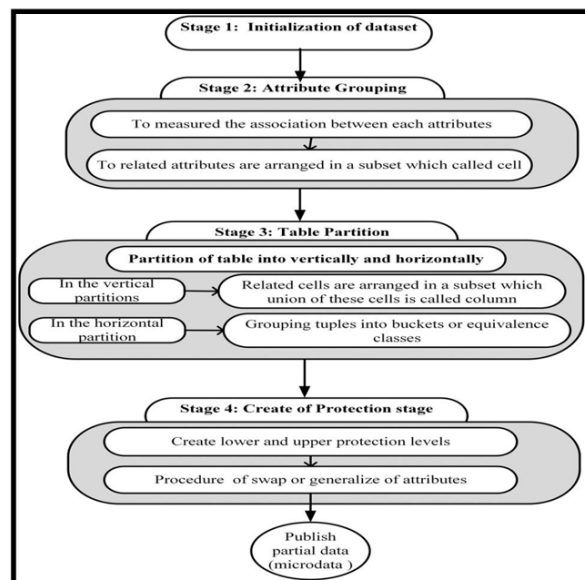


Fig3: Block-Diagram

3.2 Objective Of The Project (Slicing Method):

The main aim of this project is to put slicing into action, since it is a current approach to publish data that gives good results for both privacy and usefulness. While generalisation and bucketization may be less accurate with the data, slicing both keeps data accurate and protects information by separating the data into different subsets using privacy-related attributes.

Key Objectives:

- Data usefulness is kept better when applying this method than with generalisation methods.
 - Do not reveal people's names or identities in the data.
 - Stop Hacking Attacks: Secure personal data without affecting the use of other information.
 - Further improve k-Anonymity by providing greater protection against attacks on privacy.
 - Help researchers do their work together securely and protect confidential information.
- Retains more real information from the data than is shown by generalisation.
 - Enables users to analyse data from many different data slices.
 - Lawmakers want to improve how members are required to share their positions.
 - Relies on l-diversity to obscure any connexions between a person's sensitive attributes and their identity.
 - It's an excellent method for working with data that has a lot of features.
 - Can process complex datasets smoother through two methods of partitioning.

Wide Applicability:

- Helps keep users' attributes and personal identifying information confidential.
- Design for sharing data at any scale, controlled by privacy.
- Dynamic Data Publishing is supported by the API.
- It keeps pace with new data sets and still protects privacy year after year.
- Much Better Than Generalisation & Bucketization.
- Gives you a better, more flexible and more secure choice.

3.3 Scope Of The Project:

By using this technique, data privacy is enhanced, more than with conventional approaches, for high-dimensional datasets. It opens the door to protection and better usage through both organising data across and through various layers.

Scope Highlights:

- Traditional tanks help to better preserve and conserve utility materials.

4. Computational Environment

4.1 Hardware Configuration

Component	Specification
Processor	Pentium – III
Speed	1.1 GHz
RAM	256 MB (Minimum)
Hard Disk	20 GB

TABLE1:HARDWARE-CONFIG

4.2 Software Configuration

Component	Specification
Operating System	Windows 95 / 98 / 2000 / XP
Application Server	Tomcat 5.0 / 6.X
Front End	HTML, Java, JSP, AJAX
Scripts	JavaScript
Server-side Script	Java Server Pages (JSP)
Database	MySQL

TABLE2:SOFTWARE-CONFIG

5. Feasibility Study

5.1 Feasibility Overview

To find out if the planned upgrades are feasible and worthwhile, a feasibility study forms an important step in this initial analysis. It checks for safety, efficiency, affordability and effectiveness so the solution makes sense.

Fact is, any solution is possible with as many resources and time as we like, though that's not the world we inhabit. That's why this study stays down to earth.

Key Areas:

- Technical Feasibility
- Operational Feasibility
- Economic Feasibility

5.2 Technical Feasibility

You need to confirm that the solution can be done using present tools and infrastructure.

Questions addressed:

- ❖ Can you get the technology you require and will it work correctly?
- ❖ Do the existing systems have enough power to support the added changes and new information?
- ❖ Should you expect strong performance when groups access your network simultaneously?
- ❖ The system can it be made bigger or improved subsequently?
- ❖ Are your records secure, free of errors and easy for you to access?

The system covers all technical requirements and can be put into action using present infrastructure.

5.3 Operational Feasibility

To cheque operational feasibility, we need to know if the system will be usable in real situations and if users will value it.

Key Considerations:

- Do stakeholders and users agree to use your team's insights?
- Could it deliver value once introduced into the organisation?
- Could factors exist to lower the system's impact?

6. System Analysis

6.1 Existing System:

While existing methods like k-means cheque for centroids, we can't do that in our situation since attributes are represented by their own points. Moreover, k-medoids is able to handle outliers and produces independent cluster assignments, regardless of the data point order. At present, existing methods such as column generalisation and Mondrian fail to meet all the needs of both data evaluation and privacy. Such approaches find it difficult to deal with valuable or large amounts of data.

Disadvantages:

- With existing algorithms, it's only possible to give privacy to a single column.
- The ways we analyse data right now are not effective when dealing with anonymized or sliced data.
- Protecting privacy, through methods like differential privacy, is not very useful in practise.
- Having to use too many calculations and information that doesn't really help.

6.2 Proposed System:

We demonstrate a new approach to removing personal details from data using "slicing," which ensures that data remains more useful than if it is generalised. Membership disclosure protection is improved by using slicing and its effectiveness applies to data of any dimension. It meets the ℓ -diversity criteria which keeps attribute information confidential.

Advantages:

- When data is sliced, anonymization is more effective than generalisation.

- It hides each user's sensitivity and offers privacy thanks to ℓ -diversity.
- Data is both sliced and protected by an efficient algorithm.
- From our experiments, we see that slicing does better than generalisation and bucketization, mainly on sensitive data classification tasks.

7. System Architecture

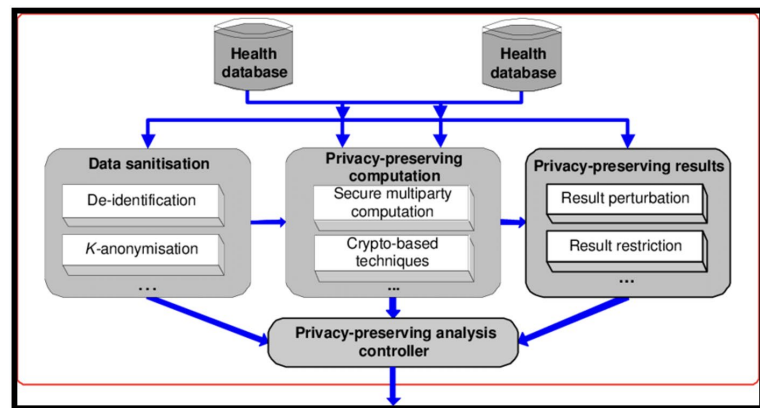


Fig4: System-Architecture

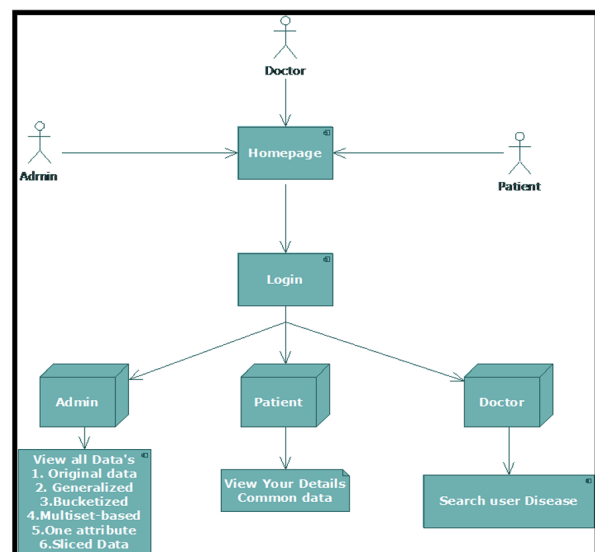


Fig5: DataFlow-Diagram

8. Conclusion

In short, these results present several potential directions for additional research. Another direction is to introduce overlapping columns, where one attribute is repeated in different columns and more attribute connexions are then found. Even though greater data utility can be gained, it brings new privacy problems that need to be looked at closely. In addition, we need to explore how much privacy the approach is able to give up in exchange for better utility.

We need to investigate how to stop the revelation of member information in future research. Our results suggest that randomly grouping tuples may not be very efficient, so new algorithms are necessary. It is also noteworthy that slicing has potential when analysing high-dimensional information, as this area has also recently attracted interest in the study of anonymizing transaction databases.

Despite the many options for keeping data anonymous, using that information is still a difficult task. The experiments we performed showed that random relationships between column values can reduce the usefulness of data, so future efforts should concentrate on using anonymized data with data mining tasks.

Even so, while k-anonymity lowers the possible revealing of one's identity, it does not handle the risk of revealing personal data points as thoroughly. Even though 'l-diversity tries to solve this, our demonstrations point out its drawbacks by attacking the model. We suggest a new way for privacy and utility to be balanced, called closeness which includes two versions: t-closeness and a more flexible edition. Even though the model helps preserve privacy, more can be done to improve its effectiveness using generalisation of any sensitive information.

Acknowledgement

I am thankful to the Management of Amrita Sai Institute of Science and Technology for giving me an opportunity to work with his project.

I would like to thank **Dr. M. Sasidhar**, Principal, Amrita Sai institute of science and technology, for his constant encouragement and support during the progress of this work.

I am deeply grateful to **Dr. P. Chiranjeevi**, Professor and Head of the Department, for his valuable guidance and consistent support during the course of the project.

A special note of thanks to my internal guide, **Mr.B.Balaji**, for his exceptional guidance, constant motivation, and continuous encouragement, which played a crucial role in the successful completion of this project.

CHINTHAGUNTA.ANITHA

References

[1] Acoria, "Rock web server and load balancer," [Online]. Available: <http://www.acoria.com>. [Accessed: May 22, 2025].

[2] Amazon Web Services (AWS), "Amazon Web Services," [Online]. Available: <http://aws.amazon.com>. [Accessed: May 22, 2025].

[3] M. Cardellini, M. Colasanti, and P. S. Yu, "Dynamic load balancing on web-server systems," IEEE Internet Computing, vol. 3, no. 3, pp. 28-39, 1999.

[4] S. Cherkasova, "FLEX: Load balancing and management strategy for scalable web hosting service," IEEE Symposium on Computers and Communications, pp. 8, 2000.

[5] F5 Networks, "F5 Networks," [Online]. Available: <http://www.f5.com>. [Accessed: May 22, 2025].

- [6] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext transfer protocol – HTTP/1.1," IETF RFC 2616, 1999.
- [7] Google Inc., "Google App Engine," [Online]. Available: <http://code.google.com/appengine/>. [Accessed: May 22, 2025].
- [8] HaProxy, "HaProxy load balancer," [Online]. Available: <http://haproxy.1wt.eu/>. [Accessed: May 22, 2025].
- [9] G. Hunt, E. Nahum, and J. Tracey, "Enabling content-based load distribution for scalable services," Tech. Rep., 1997.
- [10] E. Katz, M. Butler, and R. McGrath, "A scalable HTTP server: The NCSA prototype," in Proc. First International Conference on the World Wide Web, Apr. 1994.
- [11] G. Ghinita, P. Kalnins, and Y. Tao, "Anonymous publication of sensitive transactional data," IEEE Trans. Knowledge Data Eng., vol. 23, no. 2, pp. 161-174, Feb. 2011.
- [12] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in Proc. ICDE, pp. 205-216, 2005.
- [13] G. Ghanta, Y. Tao, and P. Kalnis, "On the anonymization of sparse high-dimensional data," in Proc. ICDE, pp. 715-724, 2008.
- [14] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in Proc. ICDE, 2009.
- [15] L. Kaufman and P. Rousseau, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
- [16] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in Proc. SIGMOD, pp. 217-228, 2006.
- [17] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate query answering on anonymized tables," in Proc. ICDE, pp. 116-125, 2007.
- [18] P. Samarati, "Protecting respondent's privacy in microdata release," IEEE Trans. Knowledge Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov. 2001.
- [19] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 429-440, 2009.
- [20] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in Proc. ICDE, p. 25, 2006.
- [21] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization," in Proc. KDD, pp. 277-286, 2006.
- [22] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in Proc. SIGMOD, pp. 473-486, 2008.
- [23] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ -diversity," in Proc. ICDE, pp. 106-115, 2007.
- [24] T. Li and N. Li, "Injector: Mining background knowledge for data anonymization," in Proc. ICDE, pp. 446-455, 2008.
- [25] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in Proc. KDD, pp. 517-526, 2009.
- [26] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, " ℓ -diversity: Privacy beyond k-anonymity," in Proc. ICDE, p. 24, 2006.
- [27] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vil Huber, "Privacy: Theory meets practice on the map," in Proc. ICDE, pp. 277-286, 2008.
- [28] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in Proc. ICDE, pp. 126-135, 2007.
- [29] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proc. S&P, 2008.
- [30] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in Proc. SIGMOD, pp. 665-676, 2007.
- [31] V. Rastogi, D. Suciu, and S. Hong, "The boundary between privacy and utility in data publishing," in Proc. VLDB, pp. 531-542, 2007.
- [32] P. Samarati, "Protecting respondent's privacy in microdata release," TKDE, vol. 13, no. 6, pp. 1010-1027, 2001.
- [33] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," Int. J. Uncertain. Fuzz., vol. 10, no. 6, pp. 571-588, 2002.