A Density-Driven Anonymization Framework for Privacy Preservation and Data Utility Optimization

Ayush Sharma
Department of CSE
IIIT Vadodara, Gujarat, India
ayushsharma@iiitv.ac.in

Naveen Kumar
Department of CSE
IIIT Vadodara, Gujarat, India
naveen_kumar@iiitv.ac.in

Mandadi Sriya Reddy
Department of CSE
IIIT Vadodara, Gujarat, India
202463006@iiitv.ac.in

***Abstract*—Maintaining individual privacy and data utility is a primary challenge in contemporary data analytics. Classic anonymization models like k-anonymity and l-diversity ensure strong privacy levels but usually involve high information loss, impairing their analytical power. This paper presents a Density-based Optimal Partitioning (DOP) framework that strikes a better balance between privacy and utility. The new approach unifies a privacy-sensitive, density-based clustering stage with an Normalized Certainty Penalty-optimized recursive partitioning stage to facilitate adaptive anonymization consistent with data's native structure. DOP is more task-agnostic and does not need predefined target attributes or cluster numbers compared to target-specific approaches, making it more flexible for handling different datasets. Experimental tests on several real-world datasets prove that DOP always has lower Normalized Certainty Penalty (NCP) values than the classical Mondrian algorithm. The findings attest that DOP is capable of efficiently reducing information loss while ensuring strong privacy protection, and thus it is a scalable and general-purpose solution for privacy-preserving data publishing and analytics.**

***Index Terms*—Data Privacy, Anonymization, k-Anonymity, l-Diversity, Information Loss, Density-Based Clustering**

I. Introduction

Privacy is essentially the right of a person to manage access to their own personal data, name, and decisions, such as what information are disclosed, to whom, and under what circumstances. With the era of big data where tremendous amounts of personal information are constantly generated and transmitted, issues on data protection and privacy have become more important. Data privacy is concerned with safeguarding personal or sensitive data from unauthorized use, exploitation, or disclosure by means of mechanisms like anonymization, encryption, and fine-grained access control.

Examples of such practical use include hospital information systems limiting access to patient data by authorized healthcare providers and e-commerce websites blocking unauthorized dissemination or sale of users' information. With the growth of digital technology at an exponential pace, a larger percentage of such data now includes Personally Identifiable Information (PII) or other sensitive data. Thus, the size, sensitivity, and sharing of such data have in themselves increased privacy threats quite substantially. Achieving an optimal balance between data privacy protection and data utility has thus become a key research problem in current data management and analytics.

Underlying every dataset are four fundamental types of attributes: identifiers, quasi-identifiers (QIs), sensitive attributes, and non-sensitive attributes. Identifiers (e.g., name, student ID, mobile number, or email address) explicitly disclose an individual's identity. QIs (e.g., age, gender, address, ZIP code, or date of birth) might not identify anyone by themselves but can facilitate re-identification when integrated with auxiliary data. Sensitive attributes (e.g., medical condition, income, religion, biometric data, or criminal record) bear personal information that requires robust protection against disclosure, whereas non-sensitive attributes will usually include generic or public data.

A data set that includes one or more sensitive attributes is called a private data set, and its protection without loss of analytical usefulness is one of the main challenges. While encryption is helpful in achieving confidentiality in full, it makes data useless for analytis. Therefore, anonymization and delinking mechanisms are widely followed to alter quasi-identifiers and sensitive attributes so that re-identification is avoided at the cost of usability in data. Strong privacy models like k-anonymity [1], l-diversity [2], and t-closeness [3] are commonly used to make this trade-off, each providing different balances between data utility and privacy strength.

State-of-the-art works have introduced target-aware anonymization models with decision tree–based partitioning for balancing privacy and analytic performance [4]. The methods often suffer from target dependency, where performance deteriorates when analytical goals or dataset properties shift. In an effort to mitigate these constraints, this research presents a generic anonymization framework intended to provide strong privacy protection while sustaining high utility of the data for a wide range of analytical purposes and varied datasets. The proposed solution is meant to provide a scalable, versatile, and application-independent privacy-preserving mechanism for large-scale data domains.

The key contributions of this work are as follows:

- A target-independent anonymization model that best combines the privacy and data utility through the joint use of k-anonymity and l-diversity models.
- An in-depth comparison of data utility and information loss across various datasets with different k-anonymity and l-diversity parameters. The outcomes are compared and examined with the baseline k–l anonymization strat-

979-8-3315-7896-1/26/$31.00 ©2026 IEEE

egy [5] to show the performance gain of the proposed framework.

In what follows, Section II summarizes the related work. Section III discusses the problem statement and the goals of the proposed framework. Section IV gives a detailed overview of the proposed framework. Section V outlines the experimental evaluation and results. Section VI concludes this work.

II. Related Works

The k-anonymity framework proposed by Sweeney [1] is a cornerstone achievement in data anonymization. This guarantees every record in a dataset is unidentifiable from a minimum of $k-1$ records through their quasi-identifiers (QIs) to prevent individual re-identification [6]. This ensures identity disclosure protection and has been the building block for many privacy-preserving schemes. However, k-anonymity suffers from a critical limitation: if considering a k-anonymous group where all records share the same sensitive attribute value, such as a particular medical condition, attribute disclosure can still occur.

To overcome this limitation, l-diversity was proposed by Machanavajjhala et al. [2], introducing the requirement that each k-anonymous group must contain at least l distinct sensitive values. This improvement minimizes the likelihood of homogeneity attacks by maintaining diversity of sensitive attributes. Despite the enhancement of k-anonymity through l-diversity, both are still prone to attacks in skewed or imbalanced distributions of sensitive attributes. As a result, the demand for more utility-aware and adaptive privacy models has continued. This current research is specifically interested in these two seminal anonymization models—k-anonymity and l-diversity as a foundation for the development of an enhanced generalized framework.

A number of early anonymization algorithms implemented these models using generalization and suppression methods. The Datafly algorithm [7] was the first to take this approach, using full-domain generalization to substitute precise attribute values with coarse categories (e.g., reducing all 5-digit ZIP codes to 3-digit prefixes). Although successful at providing privacy assurances, this method tended to incur too much information loss due to uniform abstraction of entire attributes regardless of data distribution.

To overcome these inefficiencies, partitioning-based algorithms were developed. Among them, Mondrian [5] was a prominent top-down, multidimensional partitioning algorithm. It recursively partitions the dataset to create localized equivalence classes for context-sensitive generalization that preserves much higher data utility than full-domain generalization. Expanding on this method, the k–l Mondrian variant [5], [8] extended the model to simultaneously meet both k-anonymity and l-diversity. Each of these equivalence classes includes a minimum of k records and l sensitive values, thus balancing identity and attribute disclosure threats. Owing to their efficiency and adaptability, the Mondrian and k–l Mondrian algorithms have been universally acclaimed as benchmark methods in privacy-preservation data publishing and have been taken up as baseline models in the current work.

Later, Target-Aware Data Anonymization (TADA) [4] proposed a target-based partitioning approach that adapts anonymization to a particular predictive task. While this approach boosts utility for a particular target efficiently, it is still target-dependent, and re-anonymization is needed when the analytical task changes—therefore, not particularly useful in general cases. Likewise, utility-oriented anonymization by local recoding [9] optimizes data for specific query loads but varies pervasively across diverse analytical settings. Techniques such as sparse high-dimensional anonymization [10] try to maintain utility in high-dimensional data but still come with significant information loss, particularly in datasets that have irregular or skewed distributions since they rely on uniform partitioning schemes.

Unlike these heuristic-based or task-oriented techniques, this work introduces an overall-purpose anonymization model based on natural clustering that understands data's inherent structure. Unlike other models, the introduced technique doesn't depend on pre-specified targets or cluster numbers, supporting flexible anonymization with improved information preservation and utility and robust privacy assurance on a wide range of data and analytics tasks.

III. Problem Statement and Objectives

The central problem tackled in this research is that of finding privacy preservation versus data utility balance in anonymization. While k-anonymity and l-diversity models guarantee effective prevention of identity and attribute disclosure, they tend to result in significant information loss and, therefore, lower the usefulness of the dataset for general analytical or machine learning purposes. Therefore, the underlying issue addressed in this research is: *How do we provide k-anonymity and l-diversity for a dataset with minimal information loss to maintain maximum data utility for diverse analysis purposes?*

The goals of this work are as follows:

- To maximize data usefulness throughout k-anonymity and l-diversity model applications to ensure efficient privacy protection with minimal analytical performance degradation.
- To create a general-purpose anonymization framework that is standalone of particular target attributes or analytical tasks to enhance flexibility and usability in diverse data-driven contexts.

IV. Proposed DOP Framework

Current anonymization models like k-anonymity and l-diversity well protect individual privacy but frequently accomplish this at the cost of data utility. The newer approaches like Target-Aware Data Anonymization (TADA) and utility-based local recoding improve analytical performance by adapting anonymization to a particular target or query. But their reliance on pre-defined attributes or analytical goals confines them to apply and adapt in different datasets.

To overcome these constraints, this research introduces a Density-based Optimal Partitioning (DOP) framework, a general-purpose anonymization solution that balances privacy protection with maximum analytical useability. The framework exploits natural data clustering to discover the inherent structure of datasets and thus facilitate adaptive generalization without reference to fixed target attributes or pre-specified cluster numbers. Such design guarantees compliance with privacy requirements like k-anonymity and l-diversity while minimizing distortion caused by anonymization.

The overall design of the proposed framework includes two main parts: anonymization and grouping (clustering) of data. Its modular nature facilitates integration with heterogeneous datasets, analytical frameworks, as well as privacy models. In the current implementation, k–l-based clustering and k–l-based anonymization are used, as shown in Fig. 1.

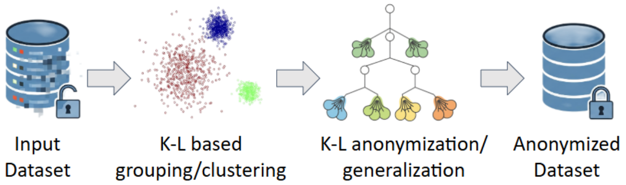

Fig. 1: System architecture of the proposed DOP framework.

To be adaptable, the underlying partitioning algorithm should adjust to the natural, generally unseen, distribution of the data instead of inflexible assumptions. This encouraged the choice of a density-based clustering method over conventional methods like k-means [11], that enforce spherical cluster shapes and are not efficient with arbitrarily shaped data structures. The DOP framework works by a two-phase mechanism. Phase 1 sees the framework detect natural, dense structures in the data to construct initial partitions. Phase 2 sees it perform optimized partitioning to meet privacy requirements while preserving information loss to a minimum. This method circumvents the inflexibility of traditional top-down anonymization algorithms by initially modifying to the natural data distribution prior to imposing privacy assurances.

A. K-L-based Clustering

For accommodating flexible partitioning, the design uses natural clustering algorithms with no a-priori number of clusters. These algorithms discover the inherent data structure and are better able to deal with outliers compared to conventional clustering algorithms. k-means [11], hierarchical clustering [12], and DBSCAN [13] are candidate algorithms. k-means is inefficient for dealing with sparse or irregular data because of its requirement of spherical cluster shapes, whereas hierarchical clustering is more flexible but computational expensive on large datasets.

As opposed to DBSCAN, which is more suited to scalability and capable of identifying clusters of any shapes and effectively locating outliers, it is especially well-suited to anonymization, where outliers typically disproportionately affect information flows. The framework thus uses DBSCAN for its merits in finding clusters naturally without knowing their numbers in advance and in separating outliers as “noise” in conformance with the goal of having general, data-driven anonymization.

Density-Based Macro-Partitioning and Noise Handling: The initial phase utilizes an adapted DBSCAN algorithm to perform density-aware, privacy-sensitive clustering. The process is made QI-aware by giving each attribute A_i a weight w_i equal to its cardinality.

$$w_i = \frac{|A_i|}{\sum_{j=1}^{|\mathrm{QI}|} |A_j|} \tag{1}$$

The distance $d(x, y)$ between two records x and y is then computed using a weighted Euclidean distance.

$$d(x, y) = \sqrt{\sum_{i=1}^{|\mathrm{QI}|} w_i \cdot (x_i - y_i)^2} \tag{2}$$

To process datasets containing both numerical and categorical quasi-identifiers (QIs), categorical attributes are preprocessed by converting them to numerical values via ordinal encoding. This conversion maps every category to a distinct integer (e.g., assigning integers to “job” values like “admin” or “technician” levels), allowing for the application of Euclidean distance in the case of clustering. The weights are put in place to prevent attributes with greater cardinality (e.g., “job” with numerous categories) from overwhelming the distance computation disproportionately. Yet, this process relies on the assumption that numerical differences between the encoded categorical values bear meaning, which might not always yield semantic relations between categories.

This step separates the dataset into dense clusters and outliers (noise points), as shown in Fig. 2. Dense clusters with less than k records are attached to their closest valid neighbor to satisfy privacy requirements. Noise points are handled independently—if the noise group is k-anonymous and l-diverse, it is kept as a final partition; otherwise, their QIs are masked (replaced with “*”) to avoid compromising overall anonymization quality.

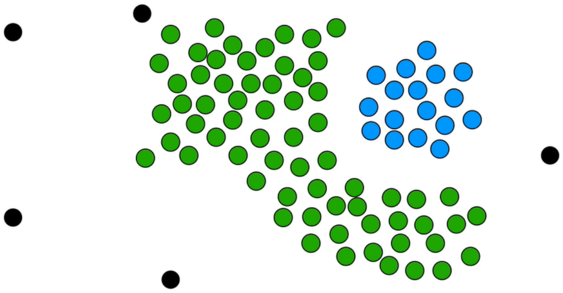

Fig. 2: An example of DBSCAN’s output. It picks out dense, arbitrarily shaped clusters (blue and green) and separates them from outlier points (black), which are regarded as noise.

B. K-L-based Anonymization

During the second phase, all dense clusters from Phase 1 are anonymized using Modified (k,l) Mondrian anonymization to impose the necessary privacy constraints in an information-loss-minimizing way. In contrast to the basic Mondrian algorithm, which is based on a naive heuristic, the approach in this paper performs an exhaustive search of the best split. To facilitate this, a Local Information Loss (LIL) measure is proposed as a cost function to estimate the generalization needed within partition P. For numerical quasi-identifiers, LIL is defined as:

$$LIL(P) = \sum_{j \in \text{Numerical QIs}} \frac{\max(P_j) - \min(P_j)}{\text{Global Range}_j} \tag{3}$$

Here, max(P_j) and min(P_j) represent the range of attribute j within the partition, while Global Range$_j$ denotes its range in the entire dataset.

For categorical quasi-identifiers, splits are ranked according to their encoded values, which come from ordinal encoding in preprocessing. Each unique value is taken as a candidate for split points, partitioning the partition into two subsets, with both being compliant with k-anonymity and l-diversity constraints. In contrast with numerical quasi-identifiers, no direct LIL cost is calculated for splits among categorical quasi-identifiers; however, the overall LIL cost for a split solely relies on the numerical quasi-identifiers in the derived partitions. This ensures that splits on categorical attributes are privacy-preserving and possible, with their information loss quantified after anonymization during Normalized Certainty Penalty (NCP) computation, as explained in Section V.

In recursive partitioning, all possible splittings of a partition P into subpartitions P_1 and P_2 are computed by the total cost $LIL(P_1) + LIL(P_2)$. The algorithm takes the split of minimum total cost, and thus the current decision is one that minimizes loss of information. It repeats this process until no split can be made without a violation of k or l constraints.

The final, correct partitions are formed by the resulting leaf nodes. A minimal generalization is then used to generate the anonymized dataset, for example, substituting numerical values with their respective [min, max] intervals. This two-stage methodology allows the DOP framework to obtain effective privacy protection while maintaining optimal analytical usefulness in diversified data settings.

V. Experimental Evaluation

In order to evaluate the utility of the suggested DOP framework, a set of experiments were conducted on various benchmark datasets to compare its performance with the basic Mondrian algorithm. The assessment was meant to capture the privacy-data utility trade-off for different k and l values based on the NCP metric [14].

A. Datasets Used

The evaluation employed following three publicly available privacy-related datasets from the UCI Machine Learning Repository [15]:

- **Bank Marketing**:
  - Quasi-identifiers: age, job, marital, education, balance
  - Sensitive attribute: y (binary: Subscribed/Not Subscribed)
  - Numerical QIs: age, balance
  - Categorical QIs: job, marital, education
- **Heart Disease**:
  - Quasi-identifiers: cp, trestbps, chol
  - Sensitive attribute: target (multi-class: 0–4 levels)
  - Numerical QIs: trestbps, chol
  - Categorical QIs: cp
- **Student Performance**:
  - Quasi-identifiers: age, Medu, Fedu, traveltime, studytime
  - Sensitive attribute: G3 (numeric: final grade, 0–20)
  - Numerical QIs: age, Medu, Fedu, traveltime, studytime
  - Categorical QIs: none

A short summary of the three datasets is given in Table I.

TABLE I: Summary of Dataset Attributes

Datasets	Total	Numerical	Categorical	Sensitive
Datasets	Attributes	QI	QI	Attributes
Bank Marketing	17	2	3	1
Heart Disease	14	2	1	1
Student Performance	33	5	0	1

B. Setup and Metrics Used

All the algorithms were coded in Python 3.8 with the Pandas and Scikit-learn libraries. Performance was measured against the NCP, which measures the amount of information loss during anonymization. For a dataset T' anonymized from T, NCP is computed as:

$$\text{NCP}(T') = \frac{1}{N \cdot |\text{QI}|} \sum_{i=1}^{N} \sum_{j=1}^{|\text{QI}|} \frac{\text{span}(T'_{ij})}{\text{span}(T_j)} \tag{4}$$

For numerical quasi-identifiers, span(T'_{ij}) is the normalized range of the generalized interval (e.g., $[\min, \max]$) against the global range of attribute j. For categorical quasi-identifiers, the span is defined as:

$$\text{span}(T'_{ij}) = \frac{|\text{cat}(T'_{ij})| - 1}{|\text{cat}(T_j)| - 1} \tag{5}$$

where $|\text{cat}(T'_{ij})|$ is the size of distinct categories in the generalized set for record i and quasi-identifier j, and $|\text{cat}(T_j)|$ is the size of the set of possible categories for quasi-identifier j. The lower NCP values, the less information distortion and, therefore, the higher the data utility.

C. Results

The suggested DOP algorithm was rigorously compared with the Standard Mondrian anonymization technique under the same (k, l) settings on a series of benchmark data sets. The comparative results, presented in Table II, uniformly show that DOP generates substantially lower NCP scores in all test scenarios (illustrated in Figures 3-5). This decrease in NCP captures DOP's capability to preserve greater data utility while guaranteeing the desired degrees of k-anonymity and l-diversity.

As an example, in the Bank Marketing dataset, DOP had an NCP of 0.2790 at (k=5, l=2) against Mondrian's 0.3959, representing a remarkable 29% enhancement in preserving utility. Similar performance improvements were noted in the Heart Disease and Student Performance datasets, where DOP had consistently lower NCP scores on rising k values. The improvements that were noticed were stronger for mixed or continuous attribute datasets, which is indicative that DOP is optimally suited for varied data types and distributions.

TABLE II: NCP Score Comparison Across Datasets

Datasets	**Methods**	(k, l)			
		(1,1)	***(2,2)***	***(5,2)***	***(10,2)***
Bank Marketing	Mondrian	0.2054	0.4572	0.3959	0.4477
	Proposed	**0.0000**	**0.3659**	**0.2790**	**0.3744**
Heart Disease	Mondrian	0.0308	0.2488	0.3250	0.4193
	Proposed	**0.0000**	**0.1683**	**0.2718**	**0.3652**
Student Performance	Mondrian	0.2115	0.2348	0.2834	0.3496
	Proposed	**0.0000**	**0.0416**	**0.1226**	**0.2075**

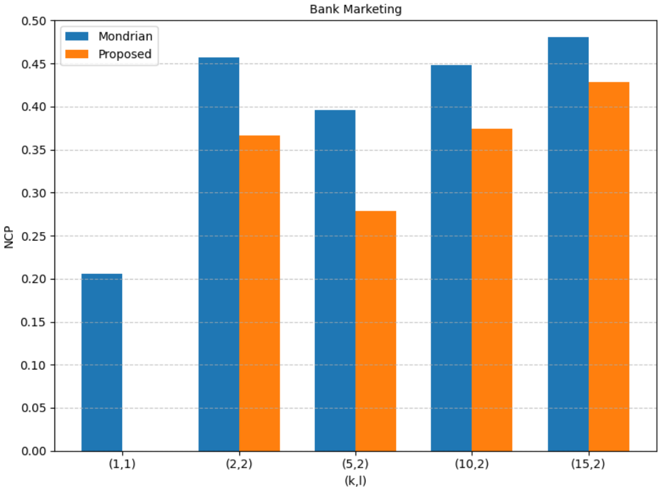

Fig. 3: NCP Score vs. (k, l) on the Bank Marketing Dataset.

A deviation from the overall trend was detected in the Bank Marketing dataset at (k=2, l=2), where both algorithms produced larger NCP values than other settings. This divergence is due to the binary sensitive attribute of the dataset, which limits potential l-diverse groupings and causes rash generalization in partitions with smaller sizes. Nevertheless, as k grows larger, the DOP framework competently balances by creating more balanced partitions that naturally contain both sensitive attribute values, thereby minimizing NCP.

The findings verify that the DOP framework delivers a stronger and more flexible anonymization technique, which always outperforms the Standard Mondrian approach in preserving data utility, flexibility, and privacy.

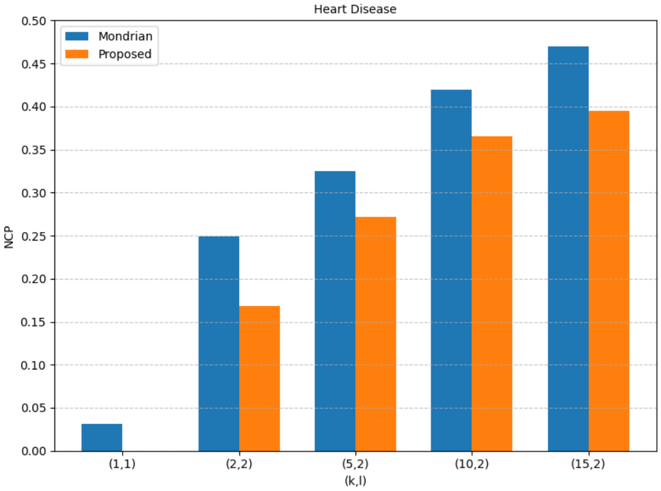

Fig. 4: NCP Score vs. (k, l) on the Heart Disease Dataset.

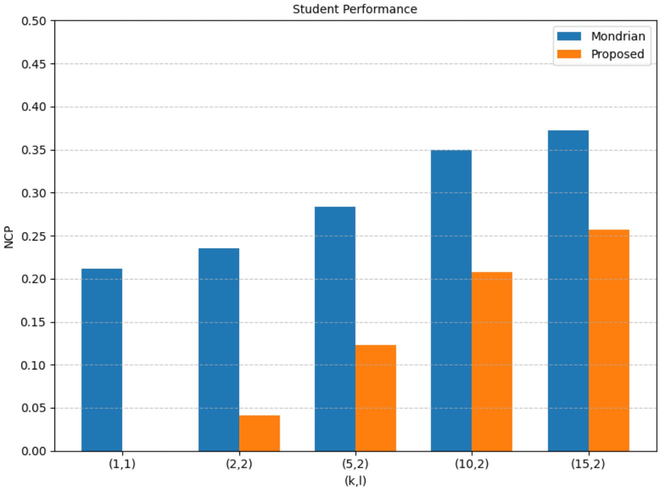

Fig. 5: NCP Score vs. (k, l) on the Student Performance Dataset.

D. Discussion

The DOP framework improves anonymization performance through a two-stage design that overcomes the limitations of heuristic-based methods. First, cardinality-weighted DBSCAN clustering adaptively partitions the dataset using density-sensitive macro-partitioning, leveraging intrinsic data distributions to form utility-efficient and privacy-compliant clusters with arbitrary shapes. This avoids fixed axis-aligned splits, such as those by Mondrian, and excessive generalization by isolating outliers, thus preserving statistical relationships and minimizing information loss. Second, Optimal Mondrian achieves recursive partitioning by exhaustive search and selects the splits that minimize cumulative Normalized Certainty Penalty defined by the Local Information Loss metric. This provides highly homogeneous partitions with reduced distortion of the quasi-identifiers, in contrast to heuristic approaches that pick the largest span attribute, thus giving superior utility consistently across various experiments.

Experimental observations bring out a number of key observations on how anonymization mechanisms behave with different characteristics in the dataset. Binary sensitive attributes in datasets like Bank Marketing show anomalies in NCP evolution due to restricted diversity space that leads to high NCP at low k values. In contrast, datasets with sensitive multi-valued attributes, like Heart Disease and Student Performance, demonstrate predictable monotonic increases in NCP as privacy constraints are tightened, thus manifesting the intrinsic privacy–utility trade-off of such models.

Another factor influencing performance is quasi-identifier composition. Datasets with largely numeric QIs-for instance, Student Performance-obtain smoother generalizations with less NCP, while those containing categorical QIs-for example, Bank Marketing-lose more information through hierarchy-based generalization. Furthermore, those datasets where the number of QIs is small when compared to the overall number of attributes see higher utility maintained as the non-QI attributes are not affected by anonymization. These results indicate that the selection and weight assignment of QIs depend on datasets and should be carefully performed in order to optimally trade off between privacy and analysis utility.

Although DOP has apparent strengths in utility preservation, the computation scalability issue originating from an exhaustive search in Optimal Mondrian, makes the algorithm less applicable to high-dimensional data. However, this limitation is mitigated by parallel processing, GPU acceleration, or approximation-based heuristics preserving near-optimal performance at much lower cost. Similarly, the DBSCAN clustering step, is highly sensitive to parameter choice—most notably the neighborhood radius (`eps`).

Finally, the current implementation maximizes general information loss which is achieved through NCP-without taking into consideration the actual task-aware utility. Generalizing the LIL objective function to a multi-objective optimization problem that incorporates domain-conscious metrics, such as Information Gain or classification accuracy, could facilitate task-sensitive anonymization. This would enable DOP to dynamically balance privacy and performance based on the analytical context for which the data are intended, thus finding wider use in both privacy-preserving machine learning and domain-specific data publishing. The DOP framework suggested in this proposal shows that the integration of density-aware clustering with cost-driven optimal partitioning results in significant utility preservation gains while providing good privacy guarantees.

VI. Conclusions and Future Work

The research introduced the Density-based Optimal Partitioning (DOP) framework as a new anonymization technique that efficiently addresses the disadvantages of conventional partitioning-based methods. By categorically combining a density-based privacy-conscious clustering step with an NCP-optimized recursive partitioning step, the resulting framework possesses superior privacy–utility balance. Experimental studies on various real datasets confirm that DOP tends to incur lower information loss than the Standard Mondrian across multiple real datasets and consistently produces more useful data, thus validating its power to retain statistical integrity while maximizing the data utility for various analytical purposes.

Parallel or distributed configurations of the Optimal Mondrian component would significantly enhance scalability for high-dimensional datasets. The addition of robust privacy models like t-closeness or differential privacy would increase resistance to more advanced inference attacks. In addition, applying DOP to multi-modal and temporal datasets like those produced by IoT systems, social networks, or financial streams would expand its usage. Lastly, incorporating task-agnostic utility measures into the NCP-guided optimization procedure may facilitate tuneable anonymization approaches that optimally balance overall data utility with the downstream machine learning task performance demands.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 557–570, Oct. 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, pp. 3–es, Mar. 2007.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. 23rd Int. Conf. Data Eng. (ICDE)*, pp. 106–115, 2007.

[4] S. Barezzani, S. D. C. di Vimercati, S. Foresti, V. Ghirimoldi, and P. Samarati, "TADA: Target-Aware Data Anonymization," *IEEE Trans. Privacy*, 2025.

[5] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, pp. 25–36, Apr. 2006.

[6] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, pp. 1010–1027, Nov. 2001.

[7] L. Sweeney, "Datafly: A system for providing anonymity in medical data," in *Database Security, XI: Status and Prospects* (T. F. Lin and S. Jajodia, eds.), pp. 1–19, London, UK: Chapman & Hall, 1998.

[8] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proc. 21st Int. Conf. Data Eng. (ICDE'05)*, pp. 205–216, 2005.

[9] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, pp. 785–790, Aug. 2006.

[10] G. Ghinita, Y. Tao, and P. Kalnis, "On the anonymization of sparse high-dimensional data," in *Proc. 24th Int. Conf. Data Eng. (ICDE)*, pp. 715–724, Apr. 2008.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, vol. 1, pp. 281–297, 1967.

[12] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining (KDD-96)*, pp. 226–231, Aug. 1996.

[14] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, pp. 279–288, July 2002.

[15] D. Dua and C. Graff, "Uci machine learning repository." http://archive.ics.uci.edu/ml, 2017.