Enhancing Privacy Preservation: Multi-Attribute Protection with P-Sensitive K-Anonymity

Twinkle Patel
Department of Computer Engineering, SAL College of Engineering, Ahmedabad, Gujarat, India

Dr Kiran Amin
Computer Engineering, Ganpat University, Mehsana, Gujarat, India

ABSTRACT

In recent years, the proliferation of extensive personal data has sparked concerns over privacy infringement and data misuse. This data encompasses various facets of individuals' lives, including shopping patterns, criminal records, medical histories, and credit profiles. While the exchange and analysis of such data offer substantial benefits for businesses and governments, privacy apprehensions can hinder data sharing.

To address these concerns, privacy-preserving data publishing techniques have emerged. Our approach focuses on p-sensitive k-anonymity, a method that extends traditional k-anonymity to consider multiple sensitive attributes simultaneously. By anonymizing data in this manner, individuals' identities are protected, mitigating the risk of re-identification while still enabling meaningful analysis. Our proposed approach aims to strike a balance between data utility and privacy protection, facilitating informed decision-making without compromising individual privacy rights.

General Terms

1st General Term, 2nd General Term

Keywords

Privacy preservation, K Anonymity, p-sensitive, P+ Sensitive

1. INTRODUCTION

In today's world, where digital technologies permeate nearly every aspect of our lives, an enormous amount of personal data is being collected, processed, and shared. From tracking shopping preferences and online behavior to storing medical records and financial histories, data has become the cornerstone of decision-making for businesses, governments, and institutions. This data-driven landscape holds great promise, offering insights that can lead to better-targeted services, improved public policies, and enhanced operational efficiencies [6].

However, the conveniences and advantages afforded by this data revolution are accompanied by a parallel concern—privacy. With the accumulation of vast volumes of personal data, the potential for misuse or unauthorized sharing of sensitive information has raised apprehensions among individuals [1]. Worries about identity theft, surveillance, and unauthorized access to personal details have given rise to the pressing need to strike a balance between data utilization and the protection of personal privacy.

In response to these challenges, researchers and professionals have been dedicated to developing techniques that facilitate responsible data sharing while safeguarding individual privacy [4]. The quest for methods that uphold the utility of data for analysis while maintaining the confidentiality of personal details has led to the exploration of privacy-preserving data publishing techniques. These techniques aim to obfuscate or modify data in ways that prevent the identification of individuals while still enabling meaningful analysis.

At the forefront of these efforts is the concept of ”P-Sensitive K-Anonymity.” This method integrates the principles of ”k-anonymity” and ”sensitivity” to provide a robust framework for privacy preservation [12]. While k-anonymity involves grouping data in such a way that each group includes at least k similar records, sensitivity adds an additional layer by considering the impact of revealing certain attributes, often referred to as sensitive attributes. What makes P-Sensitive K-Anonymity especially intriguing is its potential applicability to scenarios involving multiple sensitive attributes, an aspect that distinguishes it from other privacy-preserving techniques [12].

The primary focus of this research paper is to delve into the intricacies of P-Sensitive K-Anonymity, particularly when applied to situations where individuals have more than one sensitive attribute. In today's data-driven landscape, the demand for such methods has never been more pronounced. As the volume and diversity of data continue to grow, so too does the imperative to extract valuable insights while respecting the privacy rights of individuals.

Throughout the subsequent sections, this paper will take you on a journey through the basics of P-Sensitive K-Anonymity, offering clear examples to aid comprehension. It will delve into the mathematical model that forms the foundation of this technique and provide insights into its real-world application. Moreover, practical experiments will be conducted, using real datasets, to gauge the effectiveness of P-Sensitive K-Anonymity in preserving privacy while maintaining the utility of data.

By examining the nuances of P-Sensitive K-Anonymity and its relevance in an era marked by data-driven decision-making, this research paper contributes to the ongoing discourse on how to strike a harmonious balance between data utilization and individual privacy concerns.

1.1 Motivation

In an era defined by the rapid proliferation of digital technologies and the pervasive nature of data collection, privacy has emerged as a paramount concern for individuals, institutions, and societies at large. The vast reservoirs of personal data being amassed and analyzed for various purposes have underscored the critical need to strike a delicate balance between reaping the benefits of data-driven insights and upholding the sanctity of individual privacy. This delicate equilibrium between utility and confidentiality is central to the motivation behind this research.

The motivation to explore privacy-preserving techniques, particularly P-Sensitive K-Anonymity for scenarios involving multiple sensitive attributes, is grounded in the urgent necessity to address the challenges posed by the contemporary data landscape. Traditional privacy preservation methods, while effective to some extent, often fall short when confronted with diverse datasets that encompass multiple facets of an individual's personal attributes. The growing awareness of the interplay between various attributes and the potential risks posed by sensitive data disclosure has driven the need for more sophisticated approaches. By delving into the realm of P-Sensitive K-Anonymity with a focus on multiple sensitive attributes, this research endeavors to contribute to the development of comprehensive privacy-preserving solutions that can cater to the complexity of modern datasets.

1.2 Contribution

This research paper makes several significant contributions to the field of privacy preservation and data anonymization:

Novel Application of P-Sensitive K-Anonymity: While P-Sensitive K-Anonymity is a known approach in privacy preservation, this paper extends its application to scenarios involving multiple sensitive attributes. This extension is crucial, as real-world datasets often contain a myriad of attributes that can collectively lead to the identification of individuals. By exploring the effectiveness of P-Sensitive K-Anonymity in such scenarios, this research augments the applicability of the technique to diverse and complex datasets.

—Real-World Experimentation and Analysis: The research paper offers practical experimentation using real-world datasets to evaluate the efficiency and efficacy of P-Sensitive K-Anonymity with multiple sensitive attributes. Through these experiments, the paper provides valuable insights into the strengths and limitations of the approach in preserving privacy while maintaining data utility. This empirical evaluation contributes to the understanding of the practical implications of the technique and informs its potential adoption in various domains.

—Contribution to Privacy-Preserving Discourse: By delving into the nuances of P-Sensitive K-Anonymity and its relevance in contemporary data contexts, this research paper contributes to the broader discourse on privacy preservation and data anonymization. The exploration of multiple sensitive attributes aligns with the evolving needs of industries, governments, and institutions grappling with complex datasets. As such, the findings of this research paper provide a stepping stone for informed discussions and informed decision-making regarding privacy-preserving strategies.

In conclusion, this research paper seeks to make a significant contribution to the ongoing dialogue on data privacy and utility. By extending the application of P-Sensitive K-Anonymity to the realm of multiple sensitive attributes and conducting practical experiments, this research aims to equip stakeholders with insights and tools to navigate the intricate landscape of data-driven decision-making while safeguarding the privacy of individuals.

The paper is structured as follows: Section 2 delves into related work within the privacy preservation field. In Section 3, the paper introduces the utilized Preliminaries. Detail about the proposed approach is outlined in Section 4, while Section 5 covers result analysis. Finally, the paper concludes in Section 6.

2. RELATED WORK

This section presents the different related work done in the filed of privacy preservation. An investigation into the disclosure of sensitive information when there's prior knowledge is presented in [8]. The analysis assumes bounds on the attacker's background knowledge in terms of basic units. Basic implications are chosen as these units. Although calculating the probability of disclosure from a set of basic implications is complex, the paper outlines an effective approach to determine the worst-case scenario considering all possible sets of implications. It also demonstrates how to identify a secure grouping resilient to various implications [8]. The approach's outcomes align with the l-diversity concept but guard against a wider range of background knowledge. It's highlighted that the method might be conservative against attackers with extensive background knowledge expressed through many basic implications. To enhance efficiency, future research could explore enriching the language of basic units. Further directions involve extending the model for probabilistic background knowledge, studying cost-based disclosure, and adapting the findings to other anonymization techniques like data swapping and anonymized summaries [8].

Unique approach to enhancing the privacy of deep learning model publication through three novel contributions is presented in [14]. Firstly, as training neural networks involves numerous iterations, study utilize CDP (Concentrated Differential Privacy) to precisely estimate the privacy loss, thereby achieving accurate privacy accounting. Secondly, study address two distinct data batching techniques and propose privacy accounting methods for each, allowing precise estimation of privacy loss. Lastly, study implement dynamic privacy budget allocation techniques to enhance model accuracy, setting it apart from conventional uniform budget allocation strategies. Experiments on diverse datasets underscore the effectiveness of dynamic privacy budget allocation in improving model accuracy [14].

Currently, privacy preservation in the realm of text-based deep learning is still in its early stages. Ttrike a balance between the accessibility and security of deep learning is given in [13]. Initially, this study delve into the privacy threats within deep learning, categorizing various attack methods based on sample status. Subsequently, this study introduce a blend of techniques such as k-anonymity, homomorphic encryption, differential privacy, and adversarial learning in the context of deep learning to ensure privacy protection. In contrast to conventional privacy methods, this approach leverages deep learning's capabilities to integrate heterogeneous data from multiple sources, introducing more precise noise or vague attribute distinctions for privacy enhancement. Although text-specific privacy experiments remain limited, the methods outlined serve as foundational principles for advancing privacy pro-

tection in text-related information within deep learning. Lastly, the study outline current challenges and suggest directions for further exploration in this evolving field.
An enhanced privacy models, namely enhanced identity-reserved l-diversity and enhanced identity-reserved (α,β) is given in [11]. Where, anonymity, to address the limitations of existing identity-reserved (k, l)-anonymity and identity-reserved (α,β) and anonymity in preventing attribute disclosure. The proposed general anonymization algorithm, DAnonyIR, incorporates clustering and tailored decision functions to mitigate information loss. Comparative experiments against the GeneIR method reveal that our enhanced models offer more robust privacy protection, resulting in decreased information loss and relative error ratios in query responses.
The Internet of Things (IoT) has greatly impacted the digitization of Electronic Health Records (EHR), collecting patient data that is subsequently vulnerable to privacy breaches. To address this, privacy protection methods have been explored, including p+-sensitive k-anonymity and balanced p+-sensitive k-anonymity [5]. However, these approaches exhibit certain vulnerabilities, leading to the identification of new attacks: sensitive variance and categorical similarity. In response, a novel privacy model, the θ sensitive k-anonymity, is proposed to counter these attacks by creating more diverse k-anonymous groups. Formal analysis and experimentation validate the effectiveness of the proposed model, showcasing its superiority in achieving privacy security compared to existing methods (14.64%).

3. PRELIMINARIES

In the landscape of privacy-preserving data publishing, several fundamental concepts serve as the bedrock for safeguarding individual privacy while allowing for meaningful data analysis. These concepts provide a framework that balances the utility of data with the protection of personal information. In this section, study introduce some of these key concepts: K-anonymity, L-Diversity, T-Closeness, P-Sensitive K-Anonymity, and P+-Sensitive K-Anonymity.

3.1 K-Anonymity

K-anonymity is a cornerstone privacy-preserving technique that focuses on preventing the identification of individuals in a dataset. The concept is rooted in grouping records together such that each group contains at least k similar records [7]. By doing so, the identity of any specific individual within the group remains hidden, thereby providing a level of anonymity. K-anonymity achieves this through data transformation techniques like generalization and suppression. For instance, sensitive attributes can be generalized or suppressed to protect individual information while still maintaining useful data patterns. Table 1 and 2 presents the Inpatient Microdata and Anonymous Inpatient Microdata, respectively.

3.2 L-Diversity

L-Diversity builds upon the principles of K-anonymity by addressing a limitation associated with it—namely, the potential for attribute disclosure within the anonymized groups [2]. L-Diversity ensures that each group not only contains k similar records but also includes at least l distinct sensitive attribute values. This diversity of sensitive information fortifies privacy by mitigating the risk of attribute disclosure and preventing adversaries from inferring specific attributes within a group. Table 3 presents the diverse inpatient microdata.

Table 1. : Inpatient Microdata [2]

	Non-Sensitive	Sensitive		
	Zipcode	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	Ameliean	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Table 2. : Anonymous Inpatient Microdata [2]

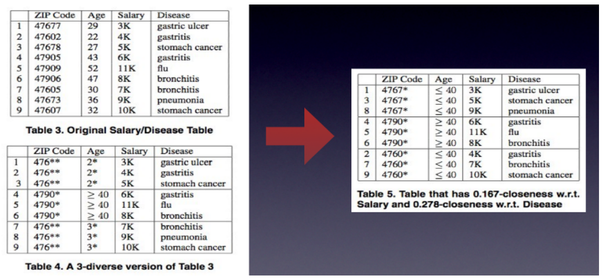
	Non-Sensitive	Sensitive		
	7ip Code	Age	Nationality	Condition
1	130* *	< 30	*	Heart Disease
2	130* *	<30	*	Heart Disease
3	130* *	< 30	*	Viral Infection
4	130* *	< 30	*	Viral Infection
5	1485 *	≥40	*	Cancer
6	1485 *	≥40	*	Heart Disease
7	1485 *	≥40	*	Viral Infection
8	1485 *	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Table 3. : Diverse Inpatient Microdata [2]

2*	Non- Sensitive	Sensitive		
	ZipCode	Age	Nationality	Condition
1	1305 *	≤ 40	*	Heart Disease
4	1305 *	≤ 40	*	Viral Infection
9	1305 *	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485 *	>40	*	Cancer
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Viral Infection
8	1485*	>40	*	Viral Infection
2	1306 *	≤40	*	Heart Disease
3	1306*	≤40	*	Viral Infection
11	1306*	≤40	*	Cancer
12	1306*	≤40	*	Cancer

3.3 T-Closeness

T-Closeness introduces yet another layer of privacy enhancement [10]. It focuses on reducing the distance between the distribution of sensitive attributes within a group and the overall distribution of the attribute in the entire dataset. By maintaining a certain threshold of closeness, T-Closeness guarantees that the sensitive attribute's distribution in a group closely resembles its distribution in the complete dataset. This prevents an adversary from discerning an individual's sensitive attributes based on group membership [10].



	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

Fig. 1: T-Closeness

3.4 P-Sensitive K-Anonymity

P-Sensitive K-Anonymity advances privacy-preserving data publishing by combining k-anonymity and the diversity of sensitive attributes [2]. In this approach, the modified microdata table T' satisfies both k-anonymity and a requirement that within each quasi-identifier group, the number of distinct sensitive attribute categories is equal to or greater than p. By enforcing both anonymity and diversity, P-Sensitive K-Anonymity offers a comprehensive mechanism for safeguarding individual privacy while enabling meaningful data analysis. This technique is particularly useful for datasets where specific sensitive attributes could lead to re-identification, ensuring a balanced approach to data utility and confidentiality [2]. Effective parameter selection and attribute consideration are essential for its successful implementation. Table 4 and 5 presents raw microdata and 2 sensitive 4 anonoymous data, respectively.

Table 4. : Raw microdata [2]

ID	AGE	Zip Code	Country	Disease
1	27	14248	USA	HIV
2	28	14207	Canada	HIV
3	26	14306	USA	Cancer
4	25	14249	Canada	Cancer
5	41	13053	China	Phthisis
6	48	13074	Japan	Hepatitis
7	45	14064	India	Obesity
8	42	14062	India	Asthma
9	33	14248	USA	Flu
10	37	14204	Canada	Flu
11	36	14005	Canada	Flu
12	35	14248	USA	Indigestion

Table 5. : 2 sensitive 4 anonoymous [2]

a				
ID	**AGE**	**Zip Code**	**Country**	**Disease**
1	<30	142**	America	HIV
2	<30	142**	America	HIV
3	<30	142**	America	Cancer
4	<30	142**	America	Cancer
5	>40	130**	Asia	Phthisis
6	>40	130**	Asia	Hepatitis
7	>40	130**	Asia	Obesity
8	>40	130**	Asia	Asthma
9	3*	142**	America	Flu
10	3*	142**	America	Flu
11	3*	142**	America	Flu
12	3*	142**	America	Indigestion

3.5 P+ Sensitive K-Anonymity

P+ Sensitive K-Anonymity enhances privacy-preserving data publishing by extending the foundation of k-anonymity and incorporating the protection of sensitive attributes [2]. In this context, the modified microdata table T' adheres to k-anonymity and introduces an additional criterion: within every quasi-identifier group, the number of distinct categories for each sensitive attribute is equal to or exceeds p. By enforcing this condition, P+ Sensitive K-Anonymity ensures not only anonymity through grouping but also the robust protection of multiple sensitive attributes. This approach is particularly valuable in scenarios where the disclosure of combined sensitive attributes could lead to unintended re-identification. By striking a balance between data utility and comprehensive privacy protection, P+ Sensitive K-Anonymity contributes to the evolving landscape of secure data analysis [2]. Effective parameter choice and thoughtful consideration of attributes remain vital to its successful application. Table 6 and 7 presents the categories of diseases and 2+ Sensitive 4-anonymous data, respectively.

Table 6. : Categories of diseases [2]

Categories of Diseases		
Category ID	**Sensitive Values**	**Sensitivity**
1	HIV, Cancer	Top Secret
2	Phthisis, Hepatitis	Secret
3	Obesity, Asthma	Less Secret
4	Flu, Indigestion	Non Sectet

Table 7. : 2^+ Sensitive 4-anonymous [2]

ID	AGE	Zip Code	Country	Disease	Category
1	<40	142**	America	Cancer	1
2	<40	142**	America	Cancer	1
3	<40	142**	America	HIV	1
4	<40	142**	America	Flu	4
5	>40	130**	Asia	Phthisis	2
6	>40	130**	Asia	Hepatitis	2
7	>40	130**	Asia	Obesity	3
8	>40	130**	Asia	Asthma	3
9	<40	14***	America	HIV	1
10	<40	14***	America	Flu	4
11	<40	14***	America	Indigestion	4
12	<40	14***	America	Indigestion	4

4. PROPOSED APPROACH

The primary focus of this research is to extend the application of P-Sensitive K-Anonymity to scenarios involving multiple sensitive attributes. This approach aims to enhance the privacy preservation techniques while accommodating the intricate nature of modern datasets that encompass diverse attributes. The proposed approach can be broken down into several key steps, each contributing to the overall objective of preserving privacy without sacrificing data utility.

4.1 Dataset Review

This study study enhance analysis by merging the 'Adult' dataset with the 'Italia' dataset, incorporating an extra sensitive attribute labeled 'disease' into the 'Adult' data. The integration of these datasets is executed using a randomized algorithm, ensuring a controlled amalgamation. This process is guided by the frequency of each individual disease ('Diseasei') within the 'Italia' dataset, denoted as F_i. The probability of a specific disease within the amalgamated dataset is determined by the ratio of its frequency to the summation of all frequencies across the diseases, encompassing a range from 1 to N. This meticulous procedure ensures a balanced and representative inclusion of the 'disease' attribute, enriching our dataset with valuable insights into health-related aspects. The probability of this disease in combined dataset will be given as per equation 1.

$$P_i = \frac{F_i}{\sum_{i=1}^{N} F_i} \quad \{1, 2, \ldots, N\} \tag{1}$$

Comprising a total of 32,561 tuples, this dataset serves as a representative sample for our privacy-preserving techniques [9]. Our focus centers on a carefully selected set of quasi-identifiers, namely Zipcode, Sex, and Race. These attributes collectively constitute our nominated set, serving as the basis for applying privacy-preserving methodologies.

P(S) Analysis:
This study delve deeper into the dataset's attributes through the lens of P(S), where the focus extends beyond individual attributes to encompass various combinations:

—(Zip, Sex, Race): This triple combination encapsulates a holistic view of an individual's location, gender, and race, offering multifaceted insights into their profile.

—(Zip, Sex): The pairing of Zipcode and Sex provides insights into geographical and gender-related aspects, unveiling trends in different regions.

—(Zip, Race): The interplay between Zipcode and Race reveals the distribution of ethnic backgrounds across geographical regions.

—(Sex, Race): This combination unveils patterns related to gender and race, contributing to a comprehensive understanding of social dynamics.

—(Zip): The Zipcode itself serves as a significant quasi-identifier, shedding light on geographical distributions.

—(Sex): Isolating Sex as a quasi-identifier allows us to explore gender-related patterns.

—(Race): Lastly, the standalone Race attribute aids in examining ethnic demographics independently.

The summary for the dataset statistics is presented in Table 8.

Table 8. : Dataset Statistics

Adult- Number of Tuples	
Element	**Number of Tuples**
Sex	2
Race	5
Sex, Race	10
ZIP	21648
Zip, Sex	22019
ZIP. Race	21942
ZIP. Sex. Race	22188

4.2 Step 1: Data Preprocessing

The initial step involves preprocessing the raw data to prepare it for the privacy-preserving procedure. This might include removing any identifying information or personally identifiable attributes, leaving only the quasi-identifiers (QIs) and the sensitive attributes intact. It's essential to strike a balance between retaining the data's analytical value and eliminating potential sources of data leakage.

4.3 Step 2: Selecting Quasi-Identifiers and Sensitive Attributes

In this step, a careful analysis of the dataset is conducted to select the appropriate quasi-identifiers (QIs) and sensitive attributes [9]. The choice of QIs is pivotal, as they play a central role in forming the equivalence classes used in the anonymization process [9]. Additionally, the sensitive attributes that warrant protection must be identified, as their disclosure can lead to re-identification of individuals. The step by step procedure is explained here [9].

—Step 1: Nominate a comprehensive set of person-dependent attributes sourced from various data owners.

—Step 2: Calculate the probability P(S) of the nominated attribute set.

—Step 3: Create a table for each element/elements in P(S), containing their respective distinct values.

—Step 4: Identify the element in P(S) with the highest number of tuples, which will constitute the set of quasi-identifiers (QI) attributes. In case of multiple such elements, choose the one with the least number of attributes.

4.4 Step 3: Applying P-Sensitive K-Anonymity

The core of the proposed approach lies in applying the P-Sensitive K-Anonymity technique. By integrating k-anonymity and sensitivity considerations, this approach ensures that individuals in the dataset cannot be singled out based on their quasi-identifiers and sensitive attributes. In the context of multiple sensitive attributes, the approach requires that each quasi-identifier group exhibits a sufficient level of diversity in the categories of the sensitive attributes. This diversity contributes to a higher degree of anonymity and prevents potential adversaries from inferring individual information.

4.5 Step 4: Privacy-Utility Trade-off Analysis

After applying the P-Sensitive K-Anonymity approach, a crucial analysis of the privacy-utility trade-off ensues. This involves assessing how well the approach preserves privacy by minimizing the risk of individual identification while simultaneously evaluating the impact on data utility. Striking the right balance between privacy and utility is essential to ensure that the data remains useful for meaningful analysis without compromising the confidentiality of the individuals being studied.

4.6 Step 5: Experimental Evaluation

To gauge the effectiveness of the proposed approach, experiments are conducted using real-world datasets. These experiments involve measuring the degree of privacy achieved, as well as assessing the impact on data utility. Metrics such as anonymity level, information loss, and the effectiveness of the approach in masking sensitive attributes are quantitatively analyzed. The experimental results provide insights into the real-world applicability of the proposed approach and its potential benefits across different domains.
By following these steps, the proposed approach seeks to provide a comprehensive solution for privacy preservation in the face of multiple sensitive attributes. It leverages the strengths of P-Sensitive K-Anonymity and extends its application to complex datasets, contributing to the ongoing efforts to reconcile the demands of data analysis with the paramount need to safeguard individual privacy.

4.7 Mathematical Model

Privacy Parameter:
Let k be the desired level of k-anonymity and p be the desired level of p-sensitivity.
Range Calculation:
For each feature dimension j $(1 \leq j \leq m)$, calculate the range $R(j)$ as Equation 2.

$$R(j) = \max(D(j)) - \min(D(j)), \tag{2}$$

where $D(j)$ represents the values of feature j in the dataset D.
Recursive Partitioning:
Define a recursive partitioning process to split the dataset D based on the feature dimensions. At each recursive step, choose a feature dimension j $(1 \leq j \leq m)$ to split. Select a splitting point $s(j)$ within the range $R(j)$ for feature j. Split the dataset D into two partitions, D_{left} and D_{right}, based on the selected feature and splitting point:

$$D_{left} = \{d \in D \mid d(j) \leq s(j)\}, \tag{3}$$

$$D_{right} = \{d \in D \mid d(j) > s(j)\}, \tag{4}$$

where $d(j)$ represents the value of feature j for instance d.
Stopping Criteria:
Repeat the recursive partitioning process until one of the stopping criteria is met:

—Each resulting partition satisfies the desired k-anonymity and p-sensitivity level.
—The maximum partition size reaches a predefined threshold.
—No further split is possible (e.g., the range for all features is zero).

Steps to implement k-anonymity, p-sensitive k-anonymity, and p+-sensitive k-anonymity for a single sensitive attribute in algorithm manner is presented below.
[h!] [1] Load the necessary packages. Read the dataset. the data by removing the key attribute. Identify categorical attributes among all attributes. Implement a function that computes spans for all columns in a partition of the dataset. For numerical columns, calculate the difference between max and min; for categorical columns, count unique values. Create a split function that takes a partition, a column, and a median value as input. The function divides the partition into two sub-partitions based on whether values are below or above the median. Implement the partitioning algorithm, incorporating a P-sensitive k-anonymous criterion. Ensure that each partition satisfies both k-anonymity and p-sensitivity requirements.
[h!] [1] Load the required packages. Read the dataset. Preprocess the data, removing the key attribute. Identify categorical attributes from the dataset. Create a function to calculate the spans for each column in a partition. For numerical columns, compute the difference between the maximum and minimum values; for categorical columns, count the distinct values. Implement a splitting function that takes a partition, a column, and a median value as inputs. The function partitions the data based on values below and above the median. Develop the partitioning algorithm using a p+-sensitive k-anonymous criterion. Ensure that partitions meet the requirements of both k-anonymity and p+-sensitivity.

5. RESULT ANALYSIS

Experiments were conducted on a system equipped with an Intel Core i5 2.39 GHz processor and 4 GB of RAM, operating on the Windows 10 platform. The algorithm implementation was carried out using Python 3.7.
The dataset employed for these experiments was the Adults database, available for public use through the UC Irvine Machine Learning Repository at[1].
In this study, the quasi-identifiers encompassed age, zip code, and sex, while income and disease were regarded as sensitive attributes.

5.1 Performance Metric

Generalized Information Loss Metric:
This metric quantifies the penalty incurred when a specific attribute is generalized by calculating the fraction of domain values that undergo generalization. Let L_i and U_i represent the lower and upper bounds of attribute i. An entry in the attribute i is generalized to an interval $[L_{ij}, U_{ij}]$, defined by the endpoints L_{ij} and U_{ij} [3].

$$GenILoss(T*) = \frac{1}{|T| \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i} \tag{5}$$

Here, T stands for the original table, n signifies the number of attributes, and $|T|$ represents the number of records.
Discernibility Metric (DM Score):
The DM score evaluates how distinguishable a record is from others by assigning a penalty to each record, equivalent to the size of the equivalence class (EQ) to which it belongs. If a record is suppressed, its penalty is equal to the size of the input table. The overall DM score for a k-anonymized table T^* is determined by Equation 6 [3].

$$DM(T*) = \sum_{\forall EQs.t.|EQ| \geq k} |EQ|^2 + \sum_{\forall EQs.t.|EQ| < k} |T| \cdot |EQ| \tag{6}$$

Where T is the original table, $|T|$ is the number of records, and $|EQ|$ represents the size of the equivalence classes created after anonymization.
Average Equivalence Class Size Metric: This metric assesses how well the creation of equivalence classes (EQs) approximates the ideal scenario, where each record is generalized within an EQ of k records. The goal is to minimize the penalty: an ideal anonymization would result in a value of 1, indicating that the EQs' size

[1] https://archive.ics.uci.edu/ml/datasets

matches the specified k value. The overall CAV G score for an anonymized table T^* is computed as Equation 7 [3].

$$C_{AVG}(T*) = \frac{|T|}{|EQs| \cdot k} \tag{7}$$

Where T represents the original table, $|T|$ is the number of records, $|EQs|$ stands for the total number of equivalence classes created, and k denotes the privacy requirement.

In the comprehensive analysis of our proposed approach, study examined three vital metrics that shed light on its efficacy. The Generalized Information Loss, measured at 0.169521, quantifies the extent to which the anonymization process led to the loss of original data attributes. A lower value signifies successful preservation of information. The Discernibility Metric (DM Score) yielded a value of 202547, indicating the level of distinction maintained between sensitive attributes in the anonymized dataset. A higher DM Score demonstrates the effectiveness of the approach in obfuscating individual characteristics. Moreover, the Average Equivalence Class Size, averaging at 1.8346, gauges the grouping of similar records during anonymization. This metric reflects the technique's ability to strike a balance between record grouping efficiency and privacy preservation. Together, these metrics affirm the promising performance of our approach in achieving a harmonious trade-off between data privacy and utility.

The experimental results for proposed approach including generalized information loss, discernibility score (DM Score), and average equivalence class size is presented in Figures 2, 3, and 4, respectively.

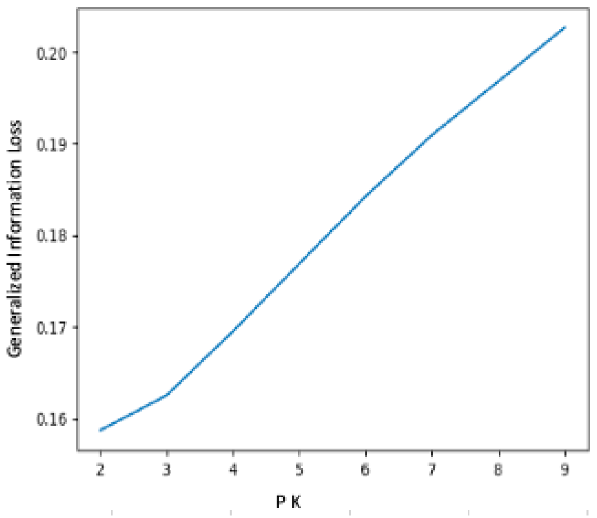


Fig. 2: Generalized Information Loss

6. CONCLUSION

The proliferation of electronic data held by corporations has prompted data publishing to be perceived as a privacy risk. This heightened awareness has resulted from growing apprehensions about the safeguarding of data privacy.

The study explored diverse approaches to anonymity. This study's focus lies on achieving p-sensitive k-anonymity and p+-sensitive k-anonymity in datasets containing multiple sensitive attributes. Furthermore, study demonstrate the trade-off between information loss and privacy enhancement.

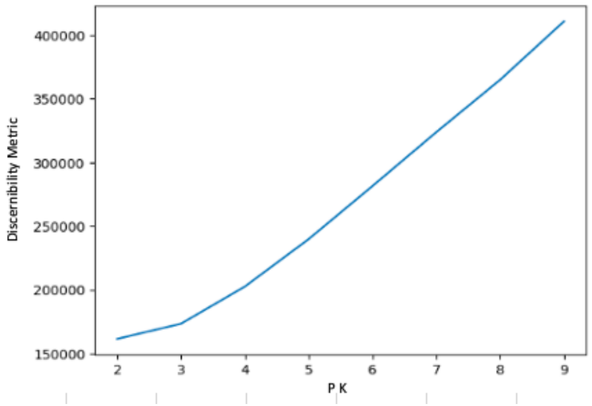


Fig. 3: Discernibility Score (DM Score)

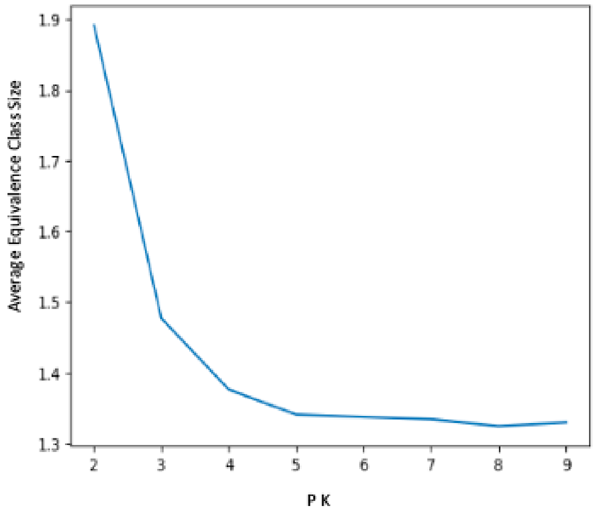


Fig. 4: Average Equivalence Class Size

Looking ahead, data privacy offers promising future research directions. These include advancing differential privacy techniques, exploring machine learning-driven anonymization, preserving contextual privacy, dynamic anonymization strategies, privacy-preserving deep learning, enhancing usability and user education, developing quantifiable privacy metrics, addressing real-world deployment challenges, fostering interdisciplinary collaboration, and investigating adversarial attacks and defenses. Each avenue holds the potential to shape more robust and comprehensive strategies for safeguarding data privacy in an evolving digital landscape.

7. REFERENCES

[1] Mansoor Ali, Faisal Naeem, Muhammad Tariq, and Georges Kaddoum. Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. *IEEE journal of biomedical and health informatics*, 27(2):778–789, 2022.

[2] Adeel Anjum, Kim-Kwang Raymond Choo, Abid Khan, Asma Haroon, Sangeen Khan, Samee U Khan, Naveed Ahmad, Basit Raza, et al. An efficient privacy mechanism for electronic health records. *computers & security*, 72:196–211, 2018.

[3] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, Liam Murphy, et al. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Priv.*, 7(3):337–370, 2014.

[4] Mahawaga Arachchige Pathum Chamikara, Peter Bertok, Dongxi Liu, Seyit Camtepe, and Ibrahim Khalil. Efficient privacy preservation of big data for accurate data mining. *Information Sciences*, 527:420–443, 2020.

[5] Razaullah Khan, Xiaofeng Tao, Adeel Anjum, Tehsin Kanwal, Saif Ur Rehman Malik, Abid Khan, Waheed Ur Rehman, and Carsten Maple. θ-sensitive k-anonymity: An anonymization model for iot based electronic health records. *Electronics*, 9(5):716, 2020.

[6] Jun Liu, Yuan Tian, Yu Zhou, Yang Xiao, and Nirwan Ansari. Privacy preserving distributed data mining based on secure multi-party computation. *Computer Communications*, 153:208–216, 2020.

[7] Waranya Mahanan, W Art Chaovalitwongse, and Juggapong Natwichai. Data privacy preservation algorithm with k-anonymity. *World Wide Web*, 24:1551–1561, 2021.

[8] David J Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke, and Joseph Y Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 126–135. IEEE, 2006.

[9] Amani Mahagoub Omer and Mohd Murtadha Bin Mohamad. Simple and effective method for selecting quasi-identifier. *Journal of Theoretical and Applied Information Technology*, 89(2):512, 2016.

[10] Mohammad Hosein Panahi Rizi and Seyed Amin Hosseini Seno. A systematic review of technologies and solutions to improve security and privacy protection of citizens in the smart city. *Internet of Things*, 20:100584, 2022.

[11] Jinyan Wang, Kai Du, Xudong Luo, and Xianxian Li. Two privacy-preserving approaches for data publishing with identity reservation. *Knowledge and Information Systems*, 60:1039–1080, 2019.

[12] Nan Wang, Haina Song, Tao Luo, Jinkao Sun, and Jianfeng Li. Enhanced p-sensitive k-anonymity models for achieving better privacy. In *2020 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 148–153. IEEE, 2020.

[13] Siran Yin, Leiming Yan, Yuanmin Shi, Yaoyang Hou, and Yunhong Zhang. A survey on recent advances in privacy preserving deep learning. *Journal of Information Hiding and Privacy Protection*, 2(4):175, 2020.

[14] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 332–349. IEEE, 2019.