A Dynamic Approach to Health Data Anonymization by Separatrices

Kristtopher K. Coelho*, Maurício M. Okuyama*, Michele Nogueira†,
Alex B. Vieira‡, Edelberto F. Silva‡ and José Augusto Miranda Nacif*
*Federal University of Viçosa UFV – Florestal, Brazil
Email: {kristtopher.coelho, mauricio.okuyama, jnacif}@ufv.br
†Federal University of Juiz de Fora UFJF – Juiz de Fora, Brazil
Email: alex.borges@ufjf.edu.br and edelberto@ice.ufjf.br
‡Federal University of Minas Gerais UFMG – Belo Horizonte, Brazil
Email: michele@dcc.ufmg.br

2024 IEEE Symposium on Computers and Communications (ISCC) | 979-8-3503-5423-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCC61673.2024.10733590

***Abstract*—Technological advances enable the integration of Internet of Things (IoT) devices to perform continuous and proactive patient monitoring. These devices collect a large volume of sensitive data that requires privacy. Anonymization provides privacy by removing or modifying information that identifies an individual. However, traditional anonymization techniques, such as k-anonymity, depend on a fixed and pre-defined k value, susceptible to attribute disclosure attacks. This article presents Dynamic Anonymization by Separatrices (DAS), an approach for defining the ideal value k and for dynamic grouping of data to be anonymized using separatrices measurements. Results show that the proposed approach efficiently mitigates attribute disclosure attacks.**

***Index Terms*—Data Anonymization, Separatrices, Privacy, IoHT**

I. Introduction

There has been massive growth in adopting Healthcare Information Technologies (HIT) in recent years. Technological advances and the miniaturization of devices (wearable and implantable) of the Internet of Things (IoT), when applied to health (Internet of Health Things – IoHT), have provided a significant increase in the generation of health data [1]. Healthcare big data analysis allows us to improve the accuracy of diagnoses and clinical decision-making, in addition to promoting remote monitoring of adverse events, reducing costs, and optimizing disease treatment [2]–[4]. However, these analyses take place on sensitive data that requires privacy by law while being shared between institutions.

Traditional privacy methods, such as encryption, authentication, and biometric schemes, allow the sharing of health information and data, providing high security and privacy [5]. However, such methods are computationally expensive to integrate all IoHT devices. In contrast, anonymization techniques transform dataset records into generic and indistinguishable data, using simple and computationally efficient operations [6]. The approaches based on k-anonymity stand out, as they always guarantee that at least k individuals will have the same combination of quasi-identifying attributes in the dataset. Complementary approaches such as l-diversity and t-proximity bring greater diversity to groups of similar data, preventing identity disclosure. However, they are still susceptible to attribute disclosure attacks. Furthermore, the literature has yet to converge to define an ideal value of k, which makes its application to private health data unfeasible [7].

In the literature, clustering-based approaches to solving the leakage and exposure of private health data information through anonymization techniques stand out. Among them, the proposal θ–Sensitive [8] adds noise to anonymized data tuples to increase the variety between equivalence classes. [9] present work in which quasi-identifier groups undergo generalization and suppression. Adaptive k-anonymity [10] adopts high computational complexity clustering methods (k-member, C-means, One-Pass k -mean–OKA and, Efficient Systematic Clustering–ESC) to anonymize sensitive data. In [6], the authors propose anonymizing numeric attributes using a fixed-range approach, replacing the original healthcare data values with a calculated equivalent value. The latter requires the value of k to be known in advance, which reduces the efficiency of privacy in the case of mistaken choices.

This article presents Dynamic Anonymization by Separatrices (DAS), a new approach to data anonymization based on groups ordered into parts defined by k-percentile. It can modify information identifying a person (numeric quasi-identifying attributes) to ensure privacy preservation and greater fidelity to raw data. The dynamic approach defines the ideal value for the number of groups (k) using the Elbow method. Then, the groups are delimited by their respective percentiles. All values from a group belonging to a k-percentile are generalized, replacing them with the mean values from that range. In summary, the DAS approach has two main contributions. First, it defines the ideal value for the number of groups (k). Second, it anonymizes numerical attributes with low computational cost and groupings defined by separatrices. The dynamic definition of the ideal value of k and the application of anonymization by separatrices in each type of numerical attribute provide more significant diversity/heterogeneity between the attribute

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil), by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/Brazil), by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant #2018/23098-0.

equivalence classes.

The anonymization performed by DAS provides more realistic, helpful data, and prevents the data from being disclosed. Therefore, the evaluation focuses on the protection offered by the DAS approach against the disclosure of attributes, considering two distinct databases, one composed exclusively of medical data. Furthermore, the information loss of the techniques is evaluated using the Normalized Certainty Penalty (NCP) metric. The performance is compared with other proposals based on k-anonymity and fixed-interval anonymization [6]. Furthermore, the computational complexity analysis of the DAS approach is presented. The results indicate that the DAS approach is more efficient against attribute disclosure attacks than relevant proposals in the literature. The source code of the DAS algorithm is available at GitHub [1].

This paper is organized as follows: Section II presents related work. In Section III, the DAS approach is detailed. Section IV describes the evaluation methodology, including the description and discussion of the results. Finally, Section V concludes the article.

II. Related Work

The widespread development and deployment of IoT devices in healthcare threatens the protection of sensitive data. Several studies address the leakage of sensitive data by applying low computational cost techniques. For example, to ensure privacy, Ouazzani and Bakkali [11] proposed an algorithm for k-anonymity without prior knowledge of the maximum k value. The algorithm iteratively groups identical combinations concerning the chosen quasi-identifying attributes until all individuals in the same group have data with the same characteristics. Despite focusing on large volumes of data, the proposal is evaluated for a fictitious quasi-identifier attribute test table containing nine rows with three quasi-identifiers. This resulted in $k = 3$. Although this work does not delve deeper into the discussion, the authors follow the literature, inferring that a low k value implies less privacy.

The Adaptive k-anonymity (AKA) [10] method promises application for cloud healthcare services. As the cloud service is charged according to use, redoing the anonymization to obtain optimal *clusters* and minimal loss of information is costly. AKA adopts established clustering methods (*k-member*, *C-means*, *One-Pass k-mean* – OKA and *Efficient Systematic Clustering*–ESC) to anonymize sensitive data. However, these groups have high computational costs. Furthermore, finding a good initial value of k is challenging by choosing randomly. Therefore, the authors rely on *Enhanced Clustering Method* – KOC to define the k value in k-anonymity.

The proposal θ–Sensitive [8] mitigates the sensitive variance and categorical similarity attacks. The proposed approach obtains the value of θ by multiplying the variance (σ^2) of a different equivalence class by an observed value (μ). Furthermore, if the desired privacy is not achieved, it adds small amounts of noise tuple(s) to increase the variability in an equivalence class. This solution prevents the risk of attribute disclosure, increasing the privacy of the dataset considered. However, parameters, such as μ and the amount of noise, significantly interfere with the levels of privacy and usefulness of the data.

Onesimu et al. [6] present a privacy-preserving data publishing scheme focusing on numeric and categorical attributes. For numeric attributes, they adopt the fixed range approach, in which the original values of the health data are replaced with a calculated equivalent value. Categorical attributes are protected by l-diverse slicing of the data, horizontally and vertically, generalizing them to prevent privacy leaks. This approach requires that the value of k be provided in advance, which can directly interfere with privacy efficiency. Additionally, slicing increases information loss.

Another anonymization method, based on clustering, groups similar quasi-identifiers together to protect the privacy of data collected by wearable devices [9]. The authors consolidate quasi-identifiers into the same groups through generalization and suppression, ensuring all records in the same equivalent set are similar. However, recent privacy models based on k-anonymity, including MDAV [12] and Mondrian [13], have been found to share fundamental vulnerabilities to attribute disclosure [7]. MDAV [14] is a heuristic multivariate microaggregation method for n-dimensional spaces ($>$1). Mondrian [15] involves the recursive division from bottom to top (*bottom-up*) of the database in two until each *cluster* has between k and $2k - 1$ records, also with a user-defined value of k.

Therefore, to advance state of the art for healthcare data anonymization, particularly for devices with limited computational power, this article introduces a novel approach called Dynamic Anonymization by Separatrices (DAS). DAS leverages the Elbow [16] method to identify the optimal cluster size k, thereby minimizing the risk of attacks. As a result, each cluster contains anonymous quasi-identifiers corresponding to the respective k-percentile. The proposal underwent rigorous assessment for attribute disclosure risk [7] and information loss [17], [18] to validate its robustness and efficiency in safeguarding data privacy.

III. Dynamic Anonymization by Separatrices

The DAS approach is based on Separatrices construction. Separatrices are values that occupy certain places in a frequency distribution [19]. We can classify them according to the number of equal parts the data is partitioned into. This way, we have, for example, quartiles (4 parts), deciles (10 parts), and percentiles (100 parts) [19]. In this sense, separatrices allow dividing a distribution into n equal parts.

The DAS approach deals with the problem of choosing the value of k since it is chosen dynamically. Testing multiple values of k is an expensive solution, and choosing an arbitrary value of k poses a significant risk regarding attribute disclosure. Examples in the literature indicate a value of $k = 3$. However, some authors [7], [20] suggest the value of $k = 5$. However, the higher the k, the lower the risk, a natural

[1]https://github.com/mauriciokuyama/das

consequence of the dimensionality in the grouping. Therefore, combined with dynamic choice, the DAS approach contributes to ensuring efficiency in terms of data privacy, increasing the value associated with k. In addition to the efficiency inherent to privacy, the DAS approach has an anonymous data grouping construction with low computational cost, which allows it to operate on hardware with limited resources.

In the context of data privacy, attributes of a dataset are categorized into identifiers, quasi-identifiers, and sensitive attributes. *identifiers* are attributes that uniquely identify an individual, such as name and ID. Therefore, to protect patients' privacy, these attributes must be removed from the dataset [6]. *quasi-identifiers (QIs)* consist of attributes that, on their own, do not identify an individual but can reveal their identity when combined [1]. Examples of quasi-identifiers include characteristics such as gender, age, and zip code. Lastly, *sensitive attributes* refer to information that becomes sensitive when linked to an individual [1]. Medication, clinical conditions, and physiological signals are examples of sensitive attributes in the context of healthcare applications.

The DAS method expects QI attributes to be provided in a potentially heterogeneous tabular data structure with labeled axes (rows and columns) (indices and numbers). In addition to categorization in terms of privacy, attributes can also be classified as numerical or categorical. *Numerical attributes* are represented by continuous values, while *categorical attributes*, can have only a finite number of values [21]. In medical data, for example, we find categorical values such as nationality, gender, and education and numerical values such as age, height, and weight [21].

The method applies the Elbow method that is traditionally used to find the ideal number of *clusters* in clustering techniques, and this can be extended to anonymization. Therefore, the Elbow method is responsible for dynamically defining the ideal value of k, which represents the respective k-percentile according to the distribution of attributes. The Elbow [16] method is a graphical method used to determine a suitable value for the number of *clusters* k so that the addition of another *cluster* does not significantly reduce the cost function to be minimized. The idea is to start with $k = 2$ and increase it at each step, computing the cost for each value. Eventually, the cost decreases drastically and then stabilizes to higher values [22]. This point represents the optimal value of k, which can be automatically detected using algorithms such as *Kneedle* [23]. Although there are other approaches in the literature to determine the ideal number of *clusters*, the Elbow method stands out for its shorter execution time, compared to other methods in the literature (*Gap Statistic*, *Silhouette Coefficient* and *Canopy*) [24]. The optimal value of k defines the separatrices measures, which occupy certain places in a frequency distribution, thus delimiting the k groups of attributes with characteristics to be generalized. Figure 1 illustrates the distribution in quartiles.

The Algorithm 1 describes the instructions to perform dynamic anonymization of numeric attributes using separatrices. The value of k, which will be used as a threshold to define the

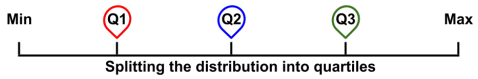

Fig. 1. Representation of the division of data into quartiles

percentiles, is defined individually for each QI. Then, the algorithm processes each sorted QI. To obtain each separator value, the q^{th} percentile of the data along the specified axis (QI) is calculated. The DAS approach uses the "closest_observation" parameter, which estimates the closest observation to an ideal value for the number of percentiles [25].

The algorithm processes each quasi-identifier individually. In line 2, the Elbow method defines the ideal k (k-percentile) value for QI. In Line 3, the algorithm orders the QI in ascending order. In Line 6, the value of the q^{th} percentile is calculated for all records belonging to QI. Next, in Line 7, the highest index corresponds to the respective q^{th} percentile, thus delimiting the range of samples that will be anonymized by generalization. Line 8 calculates the average value of the samples belonging to the respective q^{th} percentile. In Line 9, all samples belonging to the range will be replaced by the average value calculated in the previous step. In Line 12, the values dynamically anonymized by grouping based on the q^{th} percentile are assigned to the $QIs\_anonymous$ table. This flow repeats until all QIs are processed. Finally, the algorithm returns the anonymized table $QIs\_anonymous$, in line 14.

Algorithm 1: Dynamic Anonymization by Separatrices for numeric attributes

```
   Input: QIs
   Output: QIs_anonymous
   begin
1      for QI ∈ QIs do
2          k = Elbow (Unique(QI));
3          QI = Sort(QI);
4          id_start = 0;
5          for i ← 1 to k + 1 do
6              Separatrices = Percentile(QI, ((100 / k) * i),
                   method="closest_observation");
7              id_end = Max(in(QI == Separatrices));
8              QI_average = Average(QI, id_start, id_end);
9              QI = Replace_average(QI, QI_average,
               id_start, id_end);
10             id_start = id_end + 1;
11         end
12         Insert(QIs_anonymous, QI)
13     end
14     return QIs_anonymous
   end
```

Table I illustrates how the raw data is represented, which feeds the processing by the DAS approach. Table II presents the data anonymized by the proposed approach. In this example (for $k = 3$), the QIs are anonymized in the order they

appear (age, height in centimeters, and weight in kilograms). From the indices (ID) 1 and 4 of Table II, it is evident that the algorithm does not satisfy k-anonymity for the value $k = 3$, defined by the Elbow method. Specifically, the anonymized table does not satisfy the condition of that for each record in the table, there must be at least $k - 1$ other records with the same QI values. By not requiring the k-anonymity condition, DAS generates more significant heterogeneity between data in respective percentile groupings, resulting in better data utility.

TABLE I
RAW DATA

ID	Age	Height (cm)	Weight (Kg)
0	21	160	50.55
1	24	154	60.60
2	25	158	48.80
3	30	170	76.80
4	34	169	54.70
5	33	176	67.90
6	38	183	79.00
7	41	190	80.60
8	39	180	83.10

TABLE II
ANONYMIZED DATA

ID	Age	Height (cm)	Weight (Kg)
2	23	157	51.35
0	23	157	51.35
4	32	171	51.35
1	23	157	68.43
5	32	171	68.43
3	32	171	68.43
6	39	184	80.90
7	39	184	80.90
8	39	184	80.90

IV. METHODOLOGY

This section describes the methodology and datasets used to evaluate the DAS approach. Furthermore, it publishes and discusses the results accomplished.

A. Database

Adult [26] is a database built to predict whether an individual's salary exceeds $50 thousand per year, based on sense data, including quasi-identifiers such as age, sex, education, occupation, and race, in addition to the sensitive attribute salary. It is used for evaluation in most proposals for k-anonymity [6]–[8], [27] methods, including in IoHT [10] application scenario. The Adult database is made up of 48,842 records with 14 attributes, 7 of which are categorical and 7 of which are numeric. As DAS is intended to anonymize numerical data, only these were considered in this article. The data was preprocessed to remove entries with missing attributes. Thus, the final file contains 30,162 records.

The evaluation of the proposed technique using real health data was also considered, thus proving its usefulness. Therefore, besides the Adult database, the wearable-exercise-frailty – WEF [28] health database is also used in this work. The WEF database contains real health data, including quasi-identifying attributes such as age, sex, height, and weight, and sensitive attributes such as electrocardiogram (ECG) and tri-axial acceleration (ACC). The database consists of 80 records and 45 attributes, these being numeric and categorical. Only the available numerical attributes, age, height, and weight, were considered for this database.

B. Metrics and Experiments

The Dynamic Separatrices Anonymization proposal was implemented in Python, with support from the sci-kit-learn, yellow brick, numpy, pandas, and matplotlib libraries. The experiments were conducted using a machine with 20 GB of DDR4 3200MHz RAM and an AMD Ryzen 5 5500u processor. The ideal value of k was identified through the average of 10 interactions of the Elbow method for each QI attribute. Then, the performance of the DAS approach was evaluated by comparing attribute disclosure and information loss risk metrics described below.

Attribute disclosure occurs when an adversary attempts to acquire more knowledge about an individual (e.g., diagnosis). When the adversary can identify an individual's QI records and correlate them with some prior knowledge, linking can occur and consequently identity disclosure and exposure of sensitive attributes [6]. One metric to assess the risk of this type of attack is distance-based record linkage [29]. This work uses the implementation and description presented in [30]. The algorithm attacks to calculate the number of records that can be disclosed. For each record $r1$ in the anonymized database $DB1$, the distance of $r1$ is calculated concerning each record $r2$ in the original database $DB2$ (the Euclidean distance function was used in this work). Then, the record $r'(r1)$ most similar (close) to $r1$ is selected. If the selected record $r'(r1)$ matches the anonymized record $r1$, then the two records have been linked correctly. Thus, the metric reflects the number of correct links obtained by the algorithm in relation to the total number of records. The correspondence between the records, known as the databases $DB1$ and $DB2$, was generated in this work.

The information loss of the techniques is evaluated using the *Normalized Certainty Penalty* [17], [18] metric. Let min_i and max_i be the minimum and maximum value of a numeric attribute i, respectively. A record j belongs to an equivalence class with maximum value max_{ij} and minimum value min_{ij}, for attribute i. The NCP of an anonymized table T^* is defined as:

$$NCP(T^*) = \frac{1}{|T| \times n} \times \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{max_{ij} - min_{ij}}{max_i - min_i} \quad (1)$$

where $|T|$ is the number of records and n the number of attributes. The metric is based on the concept that values representing a larger range are less accurate than values representing smaller ranges [17]. The NCP value varies between 1 and 0, where 0 represents no information loss (original data), and 1 represents full information loss. Therefore, values close to 0 are desirable [17].

C. Results and Discussions

The literature recommends $k = 3$ or $k = 5$. However, this estimate depends on the characteristics of the data to be anonymized, which directly implies the level of privacy protection and the usefulness of the data for the WEF dataset when applying the dynamic definition for the value of k, based on the Elbow method, the ideal value of $k = 6$ for weight and height is found, as identified by the intersection of the line dashed with the blue line in Figure 2. The figure also displays the amount of *cluster* model tuning time for each k as a dashed green line. In general, the larger value of k is demonstrably safer. However, inferring an arbitrarily high value affects the compromise between anonymization and the usefulness of the data since total generalization does not represent the characteristics of all individuals. Table III displays information about disclosing records for the values of k suggested in the literature and the value obtained dynamically. For the WEF database, the Separatrices-based dynamic anonymization approach achieved similar performance to the fixed interval with $k = 6$, eliminating the user's responsibility for choice and the respective cost. Regarding the loss of information from the techniques, Table IV shows information that supports the application of DAS even on smaller datasets, with a loss rate of just 0.101 considering dynamic values of k.

The Adult dataset is considerably larger and complies well with evaluations of data anonymization techniques. For this database, the ideal value assigned by the Elbow method considering the hours_per_week attribute is $k = 9$, as illustrated by Figure 3. It is essential to highlight that the larger the database, the greater the possibility of disclosing attributes [7]. In Table V, it is possible to observe the discovery of attributes for MDAV and Mondrian for $k = 3$ and $k = 5$, which is significantly reduced when applying the value of $k = 9$. Furthermore, DAS performed better than other methods, with $k = 9$. Regarding the information loss from the techniques, Table VI illustrates the efficient performance of the DAS anonymization method under a vast dataset, with a loss rate of just 0.033 considering dynamic values of k.

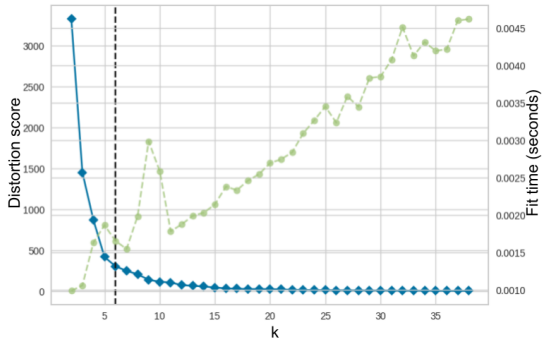

Fig. 2. Definition of the ideal value of k using Elbow for the WEF database considering the QI weight

The computational performance of the DAS is intrinsically linked to the Elbow method's computational complexity, the algorithm's most expensive stage. However, Elbow stands out positively concerning the other methods (*Gap Statistic*, *Silhouette Coefficient* and *Canopy*). DAS uses the implementation

TABLE III
NUMBER OF ANONYMIZED RECORDS CORRECTLY LINKED TO THE 80 RECORDS IN THE ORIGINAL WEF DATABASE

WEF	Fixed Interval	MDAV	Mondrian	DAS
$k = 3$	17	21	13	-
$k = 5$	38	15	13	-
$k = 6$	48	12	8	-
$age : k = 4$ $height : k = 6$ $weight : k = 6$	-	-	-	51

TABLE IV
LOSS OF ORIGINAL DATA INFORMATION FOR THE WEF DATABASE

WEF	Fixed Interval	MDAV	Mondrian	DAS
$k = 3$	0,184	0,094	0,160	-
$k = 5$	0,107	0,158	0,160	-
$k = 6$	0,086	0,185	0,240	
$age : k = 4$ $height : k = 6$ $weight : k = 6$	-	-	-	0,101

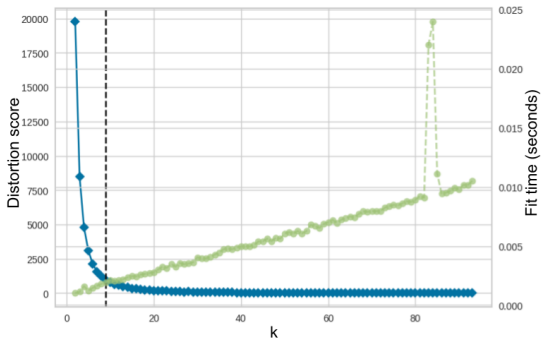

Fig. 3. Definition of the ideal value of k using Elbow for the Adult database considering QI hours per week

TABLE V
NUMBER OF ANONYMIZED RECORDS CORRECTLY LINKED TO THE 30,162 RECORDS IN THE ORIGINAL ADULT DATABASE

Adult	Fixed interval	MDAV	Mondrian	DAS
$k = 3$	26	3367	2200	-
$k = 5$	85	2219	1186	-
$k = 9$	447	1471	683	-
$age : k = 8$ $education : k = 5$ $hours\_per\_week : k = 9$	-	-	-	111

TABLE VI
LOSS OF ORIGINAL DATA INFORMATION FOR THE ADULT DATABASE

Adult	Fixed Interval	MDAV	Mondrian	DAS
$k = 3$	0,318	0,008	0,021	-
$k = 5$	0,195	0,016	0,037	-
$k = 9$	0,092	0,028	0,056	-
$age : k = 8$ $education : k = 5$ $hours\_per\_week : k = 9$	-	-	-	0,033

of the Elbow method *KElbowVisualizer* from the yellowbrick library, which applies the Kneedle algorithm to find the ideal value of k, with a known complexity of $O(n^2)$ depending on the amount of records [23]. As attributes have one-dimensional characteristics, the input for the Elbow method is limited to unique attribute values, significantly reducing the number of inputs and processing time.

The performance of the DAS approach in the data anonymization stage is obtained through low computational complexity by combining the Kneedle algorithm's cost with the sorting function's cost, which has an upper limit $NlogN$, multiplied by the number of attributes. On the other hand, solutions based on k-anonymity belong to the NP-hard [6] class of problems. A determining factor in achieving excellent performance regarding the attribute disclosure attack is that the dynamic anonymization proposal based on separatrices generates groupings or unbalanced equivalence classes. In this way, it is possible to prevent attribute disclosure attacks effectively. In addition to the security provided, data anonymization allows it to be publicly disclosed to supply the scientific community with data equivalent to raw data. Consequently, it makes it possible to extract precise statistics and process them equivalent to those obtained with raw data, which is ideal when considering an IoHT scenario.

V. Conclusion

Sharing data is essential to advance research in various areas, especially healthcare. However, concerns about data exposure and privacy make sharing between institutions difficult. Therefore, in recent years, several techniques have been proposed to guarantee the privacy of sensitive data and mitigate problems of public exposure. However, the trade-off between security and data destruction is tenuous. In this sense, the approach proposed in this article, dynamic anonymization using separatrices, guarantees the privacy of numerical attributes, keeping them as close to raw characteristics as possible. Such commitment to data privacy is achieved through dynamic identification of the best clustering configuration. Furthermore, the DAS is applied individually to each attribute, generating greater security and usefulness/fidelity to the data due to the heterogeneity between the groupings. In future work, we aim to extend DAS to cover categorical attributes, proposing a dynamic and efficient anonymization approach.

References

[1] I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A review of anonymization for healthcare data," *Big data*, 2022.

[2] L. M. Fernandes, M. O'Connor, and V. Weaver, "Big data, bigger outcomes," *Journal of AHIMA*, vol. 83, no. 10, pp. 38–43, 2012.

[3] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: preserving security and privacy," *Journal of big data*, vol. 5, no. 1, pp. 1–18, 2018.

[4] K. Batko and A. Ślęzak, "The use of big data analytics in healthcare," *Journal of big Data*, vol. 9, no. 1, p. 3, 2022.

[5] K. K. Coelho, E. T. Tristão, M. Nogueira, A. B. Vieira, and J. A. Nacif, "Multimodal biometric authentication method by federated learning," *Biomedical Signal Processing and Control*, vol. 85, p. 105022, 2023.

[6] J. A. Onesimu, J. Karthikeyan, J. Eunice, M. Pomplun, and H. Dang, "Privacy preserving attribute-focused anonymization scheme for healthcare data publishing," *IEEE Access*, vol. 10, pp. 86 979–86 997, 2022.

[7] V. Torra and G. Navarro-Arribas, "Attribute disclosure risk for k-anonymity: the case of numerical data," *International Journal of Information Security*, vol. 22, no. 6, pp. 2015–2024, 2023.

[8] R. Khan, X. Tao, A. Anjum, T. Kanwal, S. U. R. Malik, A. Khan, W. U. Rehman, and C. Maple, "θ-sensitive k-anonymity: An anonymization model for iot based electronic health records," *Electronics*, vol. 9, no. 5, p. 716, 2020.

[9] F. Liu and T. Li, "A clustering k-anonymity privacy-preserving method for wearable iot devices," *Security and Communication Networks*, vol. 2018, pp. 1–8, 2018.

[10] K. Arava and S. Lingamgunta, "Adaptive k-anonymity approach for privacy preserving in cloud," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 2425–2432, 2020.

[11] Z. El Ouazzani and H. El Bakkali, "A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k," *Procedia Computer Science*, vol. 127, pp. 52–59, 2018, proceedings Of The First International Conference On Intelligent Computing In Data Sciences, ICDS2017.

[12] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Transactions on Knowledge and data Engineering*, vol. 14, no. 1, pp. 189–201, 2002.

[13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005, pp. 49–60.

[14] M. Templ, "Statistical disclosure control for microdata using the r-package sdcmicro," *Transactions on Data Privacy*, vol. 1, no. 2, pp. 67–85, 2008.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Multidimensional k-anonymity," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[16] P. Bholowalia and A. Kumar, "Ebk-means: A clustering technique based on elbow method and k-means in wsn," *International Journal of Computer Applications*, vol. 105, no. 9, 2014.

[17] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy *et al.*, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners." *Trans. Data Priv.*, vol. 7, no. 3, pp. 337–370, 2014.

[18] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *Proceedings of the 33rd international conference on Very large data bases*, 2007, pp. 758–769.

[19] S. Correa, "Probabilidade e estatística," 2003.

[20] N. Victor and D. Lopez, "Privacy preserving sensitive data publishing using (k, n, m) anonymity approach," *Journal of communications software and systems*, vol. 16, no. 1, pp. 46–56, 2020.

[21] D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Information Sciences*, vol. 571, pp. 418–442, 2021.

[22] T. M. Kodinariya, P. R. Makwana *et al.*, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[23] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a" kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st international conference on distributed computing systems workshops*. IEEE, 2011, pp. 166–171.

[24] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

[25] N. Developers, "numpy.percentile," Fev 2024. [Online]. Available: https://numpy.org/doc/stable/reference/generated/numpy.percentile.html

[26] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[27] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[28] D. Sokas, M. Butkuvienė, E. Tamulevičiūtė-Prascienė, A. Beigienė, R. Kubilius, A. Petrėnas, and B. Paliakaitė, "Wearable-based signals during physical exercises from patients with frailty after open-heart surgery," *PhysioNet*, 2022.

[29] P. Christen, T. Ranbaduge, and R. Schnell, "Linking sensitive data," *Methods and techniques for practical privacy-preserving information sharing. Cham: Springer*, 2020.

[30] L. Jiang and V. Torra, "Data protection and multi-database data-driven models," *Future Internet*, vol. 15, no. 3, 2023.