A New Anonymization Model for Privacy Preserving Data Publishing: CANON

Yavuz Canbay, Seref Sagiroglu, Yilmaz Vural

***Abstract*—Data privacy is a challenging trade-off problem between privacy preserving and data utility. Anonymization is a fundamental approach for privacy preserving and also a hard trade-off problem. It enables to hide the identities of data subjects or record owners and requires to be developed near-optimal solutions. In this paper, a new multidimensional anonymization model (CANON) that employs vantage-point tree (VP-tree) and multidimensional generalization for greedy partitioning and anonymization, respectively, is proposed and introduced successfully for the first time. The main concept of CANON is inspired from Mondrian, which is an anonymization model for privacy preserving data publishing. Experimental results have shown that CANON takes data distribution into consideration and creates equivalence classes including closer data points than Mondrian. As a result, CANON provides better data utility than Mondrian in terms of GCP metric and it is a promising anonymization model for future works.**

***Index Terms*—Data privacy, anonymization, data publishing, CANON.**

I. Introduction

With the development in technology, the amount of data is increasing day by day. Internet of things, smart grid, wearable devices, mobile applications, social media, smart cities, e-commerce, health applications, smartphones etc. enable to collect more data than ever.

Today, privacy is a hot topic especially in the digital world. Any violation on sensitive data causes harm on the reputation of individual and they also may lead to discrimination. Hence, protecting privacy of individual in real and digital world is important and requires more effort [1].

Data holders or curators publish data publicly or with a limited set of researchers [2–4]. However, if the data contains sensitive information about individuals (e.g. genomic information [5]), privacy concern becomes one of the major issues to be addressed [6]. "Informational self-determination" [7] and "the appropriate use of responders' information and the ability to decide what information of a responder goes where" [8] are some of the definitions for data privacy in the literature. Recently, due to the increase in the collection of person-specific information, data privacy has become a major need and a requirement for data publishing or data mining [9, 10].

Anonymization is a utility-based privacy preserving approach that hides the identity of data subject and, in meantime, provides data utility [11]. In the literature, there exist some anonymization models which enable data curators to publish sensitive data while preserving data privacy. k-anonymity, l-diversity and t-closeness are the most known and frequently used privacy preserving models in data publishing [12]. These models are explained briefly as below.

- k-anonymity ensures that a record in any equivalence class is similar to at least $k-1$ other records within the same equivalence class and it provides a solution for record linkage attack [13].
- l-diversity guarantees the diversity of sensitive data in each equivalence class and proposes a solution for attribute linkage attack [14].
- t-closeness provides a balance for the distribution of sensitive data between an equivalence class and entire table, and also presents a solution for skewness attack [15].

The reference [16] provides a comprehensive list of attacks and preserving models for privacy. In addition to those, differential privacy is introduced by Dwork et al. [17] as a recent solution. Differential privacy guarantees the results of any analysis will be almost the same if an individual participates the dataset or not, and it presents a solution to background attacks. While k-anonymity, l-diversity and t-closeness are directly proposed for privacy preserving data publishing, differential privacy is mainly used to perturb the results of statistical queries.

k-anonymity, l-diversity and t-closeness are mostly used in privacy preserving data publishing. Since we think that CANON is a base model for future works, we preferred to apply k-anonymity in this paper. It is well-known that, k-anonymity has an exponential relation between the input size and solution space. This situation proves that k-anonymity is an NP-Hard problem and near-optimal solutions are always required as stated in [18–25].

In the literature, there exist many algorithms or models achieving k-anonymity which are compared and presented in Table I. The evaluations of these algorithms or models are summarized and criticized as given below;

- Optimal algorithms require exponential search spaces. Hence, these given algorithms do not provide acceptable solutions in a reasonable time if the size of records increases.

Yavuz CANBAY (corresponding author) is with the Department of Computer Engineering, Faculty of Engineering and Architecture, Sutcu Imam University, Kahramanmaras, Turkey e-mail: yavuzcanbay@ksu.edu.tr

Seref SAGIROGLU is with the Department of Computer Engineering, Faculty of Engineering, Gazi University, Ankara, Turkey e-mail: ss@gazi.edu.tr

Yilmaz VURAL is with the Department of Computer Science, College of Engineering, University of California, Santa Barbara, USA, e-mail: yilmazvural@ucsb.edu

Manuscript received Jan 23, 2022; accepted July 17, 2022.
DOI: 10.17694/bajece.1061910

TABLE I: Comparing models/algorithms achieved k-anonymity

Model/Algorithm	Optimality	Dimension	Direction	Partitioning Strategy	Splitting Value Det.
MinGen [26]	Optimal	Single	Bottom-up	Hierarchical	N/A
Incognito [27]	Optimal	Single	Bottom-up	Hierarchical	N/A
Flash [28]	Optimal	Single	Bottom-up	Hierarchical	N/A
Datafly [29]	Near-Optimal	Single	Bottom-up	Hierarchical	N/A
BUG [30]	Near-Optimal	Single	Bottom-up	Hierarchical	N/A
TDS [31]	Near-Optimal	Single	Top-Down	Hierarchical	N/A
Mondrian [32]	Near-Optimal	Multiple	Top-Down	Splitting	Frequency based
CANON (this study)	Near-Optimal	Multiple	Top-Down	Splitting	Distance based

- Near-optimal algorithms provide acceptable solutions in a reasonable time. However, working on single dimension causes more information loss compared to multiple dimensions.
- Hierarchical partitioning requires hierarchy trees. These trees are being constructed by researchers according to their needs under some constraints. Hence these trees may cause undesired information loss because of having a fixed structure. Note that these trees do not present a utility-aware partitioning, they just define the ranges of any related values.
- Among these algorithms, Mondrian stays one step ahead because of supporting multiple dimension. However, it splits data space by employing a frequency based approach which does not consider distribution of data and hence absorbs potential data utility.
- Finally, the presented model (CANON) provides a distance based approach for data space partitioning and increasing data utility by considering data distribution.

CANON is a distribution-aware anonymization model which splits data space by considering the distribution of data points. It creates equivalence classes by grouping closer data points than Mondrian. Hence, CANON provides more data utility compared to Mondrian.

This paper was organized as follows. In Section II, a literature review was presented. In Section III, Mondrian anonymization model was briefed. Some preliminary information about this study was provided in Section IV. Section V introduced VP-tree based greedy partitioning algorithm. The proposed model was presented and introduced in Section VI. Experimental results conducted in this paper was provided in Section VII. Finally, the conclusion was given in Section VIII.

II. Literature Review

This section briefs some current studies about k-anonymity and Mondrian.

Although k-anonymity was proposed in 2002 by Sweeney [13], it is still employed in many current studies. Kacha et al. [33] proposed a metaheuristic method by using black hole algorithm to provide k-anonymity. They employed adult data set and NCP metric in the experiments and presented a comparison for the results. Finally, it was seen that their model provided more utility than the others. Bhati et al. [34] focused on the anonymization of transport user data. A k-anonymity based anonymization model considering both numerical and categorical data was proposed. In addition, they introduced a new normalization technique and some utility metrics. A real-world dataset and information distortion metric were utilized in the experiments and then successful results were obtained. Mahanan et al. [35] presented a new k-anonymity algorithm based on a heuristic approach. Their algorithm enhanced the performance of existing optimization algorithms by providing a heuristic search for generalization lattice. Uber, Jester and T-drive datasets were used in the experiments and they provided a comparison for the results of the proposed algorithm and other existing algorithms. Finally, they reported that their algorithm provided an efficient anonymization. Adrew and Karthikeyan [36] combined k-anonymity and laplace differential privacy for big data anonymization. They firstly generalized tabular data and then applied laplace noise to provide differential privacy. Experiments are performed on Adult dataset and then NCP metric was used to measure data utility. They reported that the proposed model achieves better utility than other existing models.

Mondrian, which was introduced in 2006, is still being used by some current studies. For example, Wang et al. [37] enhanced Mondrian by applying Self-Organizing Map, Andrew et al. [38], Nezarat et al. [39] and Ashkouti [40] applied Mondrian for privacy preserving big data publishing. In addition, Tang et al. [41] proposed an extended version of Mondrian. They optimized the lowest value of partitions and provided equivalence classes with lower sizes than classical Mondrian. In the experiments, CM, DM and NCP metrics were employed to measure data utilities and their model presented more utility than classical Mondrian. Liu et al. [42] optimized multidimensional k-anonymity model by enhancing Mondrian model. They focused on attribute weighting and then provided a new algorithm. Census dataset was employed and NCP metric was preferred to measure data utility. In the experiments, the proposed model provided better results than classical Mondrian. Gong et al. [43] considered incomplete data anonymization and proposed two algorithms to achieve high data utility on incomplete data. They used NCP metric to measure data utility and Adult dataset. Finally, it was observed that the proposed algorithm gave better utility than classical Mondrian. Nergiz et al. [44] proposed a multidimensional hybrid k-anonymization algorithm based on Mondrian. They aimed to decrease the negative effect of generalization process on anonymization and introduced a hybrid approach using both generalization and data relocation. Census dataset was employed in the experiments and they obtained successful results. LeFevre et al. [45] proposed three different models based on

Mondrian. In their models, they focused on enhancing classical Mondrian with entropy, least square deviance and imprecision. They employed syntactic, Census and Adult datasets in the experiments and obtained better results than the other models in the literature.

This section indicates that k-anonymity is one of the anonymization models used by current studies and Mondrian is still being used to achieve k-anonymity. Since anonymization is one of the hot topics today, there is always a need for novel models and algorithms. This paper proposes a novel anonymization model, called CANON, which is an alternative to Mondrian. We hope that CANON will be a base model for further studies.

III. Mondrian Anonymization Model

Mondrian [32] is a frequently used near-optimal anonymization model in the literature. It splits data space by using two different partitioning strategies; strict and relaxed. While strict partitioning creates non-overlapping regions, relaxed partitioning creates potentially overlapping regions. Both strategies accept that the lower bound is k. In addition, $2d(k-1)+t$ and $2k-1$ are accepted as the upper bounds of strict and relaxed partitioning, respectively [32].

Mondrian is a multidimensional anonymization approach that exploits multidimensional generalization and k-Dimensional tree (KD-tree) space partitioning approach. Multidimensional generalization provides higher data utility than other anonymization operators and KD-tree provides an acceptable time complexity of $O(nlogn)$. However, KD-tree has some general disadvantages or weaknesses [46–49] which can be listed as below;

1) frequency based splitting value determination,
2) non-flexible space partitioning,
3) inefficiency in high dimensional data and,
4) producing a high unbalanced tree due to skewed data.

If these weaknesses are evaluated in the context of anonymization, it can be clearly seen that only (1) directly affects data utility. Because, frequency based splitting value determination does not consider data distribution and this causes a decrease in data utility. Therefore, if this weakness can be eliminated, a more powerful anonymization model may be obtained. Hence, CANON focuses to eliminate this weakness and provides more data utility than Mondrian.

IV. Preliminary

This section briefs the hardness of k-anonymity, gives some background information about KD-tree and VP-tree, and finally introduces some definitions used in this paper.

A. On the hardness of k-anonymity

In order to reveal the hardness of anonymization problem, we reviewed some papers and presented a summary of these works in this section. In the literature, there exist a number of papers focusing on the NP-Hardness of k-anonymity. Meyerson and Williams [18] investigated the computational complexity of k-anonymity by using a reduction from k-dimensional perfect matching problem. They indicated that if there is no restriction on the alphabet size, k-anonymity is NP-Hard for $k \geq 3$ and the maximum number of suppressed cells is $n(m-1)$, where n and m are the number of vertexes and edges, respectively. However, Aggarwal et al. [19, 20] employed a reduction from edge partition into triangles and reduced the alphabet size as 2, but the number of suppressed cells remained as $9m$, where m represents number of triangles. Similarly, Sun et al. [21] employed edge partition into 4-clique and represented an alphabet with the size of 2, but the number of suppressed cells was obtained as $48m$, where m is some integer. On the other hand, the proof presented by Scott et al. [24] used a reduction from c-hitting set problem and indicated that anonymization of k-attribute is also NP-Hard even for $k \geq 2$. Finally, LeFevre et al. [32] used another reduction to prove the NP-Hardness of optimal k-anonymous multidimensional partitioning. They utilized discernibility metric to approximate the optimal solution.

From the literature review, it can be clearly understood that anonymization is a hard problem and near-optimal solutions are always required.

B. KD-tree and VP-tree

Mondrian, which is a frequently preferred near-optimal anonymization model, employs KD-tree for data space partitioning. KD-tree [50] spits data points based on their projections into some lower dimensional spaces. A node in KD-tree contains four information as presented below;

- Dimension: is the axis of dataset will be divided,
- Splitting value: median value of dataset on dimension,
- Left-hand side: the left subtree including the data points which are smaller than or equal to the splitting value,
- Right-hand side: the right subtree including the data points that are greater than the splitting value.

The main steps of a KD-tree construction are shown as below [50];

1) choose a dimension,
2) calculate frequencies of a dataset on the dimension,
3) find median of these frequencies,
4) accept this median as a splitting value,
5) partition data space into two subspaces either horizontally or vertically according to the splitting value,
6) go to step 1 until no data point is left.

A VP-tree [51] is a metric tree and a balanced binary tree that recursively divides the space into two partitions based on a median of the distances between a vantage-point and the others. A hypersphere is employed for splitting data into n-dimensional metric space. In this tree structure, a node contains four information as;

- Vantage-point: a point which is selected from dataset,
- Radius: a distance defining the range of vantage-point,
- Left-hand side: the left subtree including the data points which are smaller than or equal to the radius of a vantage-point and,
- Right-hand side: the right subtree including the data points that are greater than the radius of a vantage-point.

The main steps of a VP-tree construction are presented as below [46], [51–55];

1) choose a vantage-point,
2) calculate the distances between the vantage-point and the others,
3) find the median of these distances,
4) accept the median as a splitting value,
5) according to the splitting value, partition data space into two subspaces,
6) go to step 1 until no data point is left.

The illustrations of KD-tree and VP-tree decompositions on a sample dataset are presented in Figure 1. A sample dataset, a KD-tree decomposition and a VP-tree decomposition are shown in Figure 1a, Figure 1b and Figure 1c, respectively. In KD-tree, the frequencies of data on any dimension are calculated and then the median of these frequencies is determined to split dataset. Therefore, a splitting operation which is regardless of data distribution is performed and this situation leads a decrease in potential data utility. But, VP-tree calculates the distances between a vantage-point and the other data points then partitions data space according to the median of these distances. Hence, a distribution-aware splitting is performed to dataset and better data utility is obtained.

In the literature, there are some works comparing KD-tree and VP-tree [46, 49, 56–58]. Based on the results of these works, it can be seen that VP-tree data structure is more successful than KD-tree. Hence, we proposed a new anonymization model that employs VP-tree for space partitioning and also adopts some basic functions of Mondrian. The proposed model is inspired and borrowed some functionalities from Mondrian, but has some advantages compared to Mondrian.

C. Some definitions for the proposed model

In this section, we borrowed some definitions from [32], redefined them in order to provide the theoretical background of the proposed model. These definitions are given below.

A Multidimensional Cut; is a vantage-point centered cutting that produces two disjoint multisets of points, for a multiset of points P.

Allowable Multidimensional Cut; is the state of being multidimensional partibility. In a d-dimensional space, a cut for the region R_i with radius r_i is allowable if and only if $Count(R_i.P_i > r_i) \geq k$ and $Count(R_i.P_i < r_i) \geq k$.

Non-allowable Multidimensional Cut; in a d-dimensional space, a cut for the region R_i with radius r_i is non-allowable if $Count(R_i.P_i > r_i) < k$ or $Count(R_i.P_i > r_i) < k$.

Multidimensional Partitioning; means cutting the space into multidimensional sub-regions $R_i, ..., R_n$ that cover all attribute domains.

Minimal Multidimensional Partitioning; let R_i be the i^{th} region created by a multidimensional partitioning and contains multiset P_i of points. If $|R_i.P_i| \geq k$ then this partitioning is minimal and there exist no allowable multidimensional cut for R_i.

In the proposed model, we accepted that the upper bound of minimal multidimensional partitioning is $2k-1$ and the lower bound as k since our model inspires from relaxed partitioning strategy of Mondrian [32].

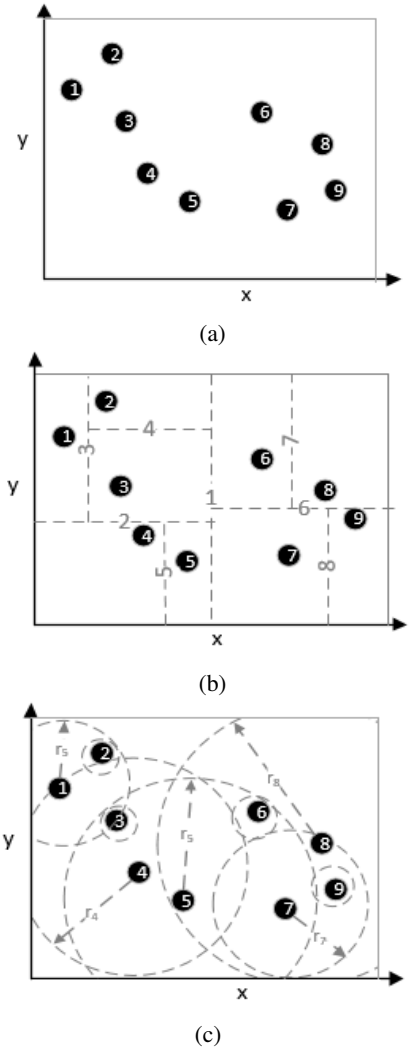


Fig. 1: a) A spatial representation of sample dataset, b) A possible KD-tree decomposition, c) A possible VP-tree decomposition

V. VP-Tree Based Greedy Partitioning Algorithm

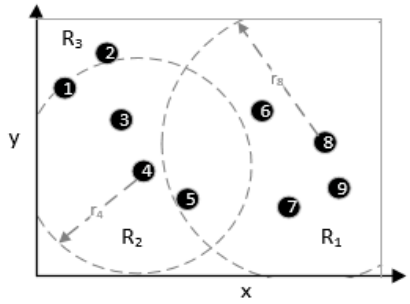
In order to partition data space with a proximity based approach, we employed VP-tree in our anonymization model. VP-tree works as the following manner. Firstly, it takes a set of entire records of partition S. If there exists an allowable cut for S, a random element p (vantage-point) is then selected from S. For each $s \in S$, distance $d(p, s)$ is calculated and the median of these distances is then assigned to μ. After that, if an element is smaller than or equal to μ, it is included into the left partition, otherwise the right partition. This iteration continues until no point is left. After the execution of the algorithm, a set of multidimensional regions that the size of each is between k and $2k-1$ are obtained. The time complexity of the partitioning algorithm is $O(nlogn)$.

Consider any multiset of points as illustrated in Figure 1a. They are labeled with some numbers and have some coordinates such as (x_i, y_i) which can be generalized for higher dimensions. If KD-tree and VP-tree partitioning are applied on these points, two potential partitions can be obtained as

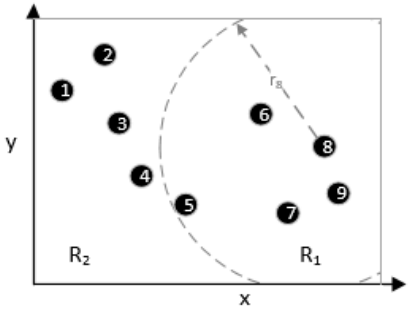
presented in Figure 1b and Figure 1c, respectively. In Figure 1b, random axises are selected and possible partitionings are presented, and in Figure 1c, vantage-points are selected randomly and then possible partitionings are represented roughly.

As illustrated in Figure 1, constructing a VP-tree on a sample dataset can be achieved successfully. However, in order to adopt this construction to our model, we had to set two important rules, non-allowable cutting and allowable cutting, which are detailed in Section IV-C. Because we have to verify that the points are available for a potential partitioning.

In Figure 2a, a non-allowable cut is indicated. The data points with the labels of 8 and 4 are selected for vantage-points for the related partitions, respectively. After the partitioning process, although r_8 partitioning meets the requirements of allowable multidimensional cutting, r_4 partitioning fails since it does not satisfies $Count(P.R_i \geq r_i) \geq k$, for $k \geq 3$. In contrast to a non-allowable cutting, an allowable partition may be available if the predefined conditions are met. Figure 2b indicates an example of allowable cutting. In this example, in order to provide an allowable cut, the vantage-point with the number of 8 is selected for illustration.



(a)



(b)

Fig. 2: a) A sample representation of a non-allowable cut, b) A sample representation of allowable cut, for $k \geq 3$.

VI. The Proposed Anonymization Model: CANON

CANON, is a multidimensional anonymization model applying k-anonymity, consist of two phases which are VP-tree phase and Generalization phase. In VP-tree phase, the data space is partitioned into two subsets recursively and then data partitions including minimum k and maximum $2k-1$ elements are obtained. In the generalization phase, each partition is generalized. The algorithm of CANON is shown in Figure 3.

Input: dataset $X = \{x_1, x_2, \dots, x_N\}$; $x_N \in R^D$, k parameter of k-anonymity
Initialize: $data = X$
1) Apply VP-tree over *data*
Iterate:
- Choose vp; $vp = \{x \in X,\ x = choose\_vp(data)\}$
- Calculate distance; $dist = \{\{d_1, d_2, \dots, d_N\} = distance(vp, data)\}$
- Find the median; $\mu = \{m \in dist, m = med(dist)\}$
- Split data; $lhs = \{l \in data,\ l_{dist} \leq \mu\}$
  $rhs = \{l \in data,\ l_{dist} > \mu\}$
- Repeat for; $data = lhs$
  $data = rhs$, until the stopping criteria is met

Return:
- Return partitions $\{P_1, P_2, \dots, P_T\} \in X$

2) Generalize each P_i in $\{P_1, P_2, \dots, P_T\} \in X$
Iterate:
- Generalize; $P_i^G = gen(P_i)$
- Repeat for each P_i

Return:
- Return anonymized partitions $\{P_1^G, P_2^G, \dots, P_T^G,\} \in X^G$

Output: anonymized dataset X^G

Fig. 3: The algorithm of CANON

A general view for working of CANON is presented in Figure 4, which is created based on the assumptions listed below;

- 3-anonymization is applied,
- square boxes are employed to represent data points,
- each color shows the proximity of data points, in other words, same colored points (boxes) are closer to each other than the others,
- each group including a number of same colored points (boxes) represents equivalence classes,
- transformation from one color to another equals generalization (for example in Figure 4, blue and orange colors in Partition1 transform to green).

The proposed model illustrated in Figure 4 works as follow;

1) choose a vantage-point,
2) calculate the distances between vantage-point and the other points,
3) find the median of distances,
4) accept the median as the splitting value,
5) split data space into two subspaces (LHS and RHS) with regard to the median value,
6) check if there exist allowable cutting,
7) if yes, go to step 1,
8) if no, obtain partitions that satisfies lower and upper bounds,
9) generalize all partitions and obtain equivalence classes,
10) collect and combine partitions, and then obtain anonymized dataset.

VII. Experimental Results

In order to achieve the experiments, a number of steps have been followed. These steps are given below.

A. Determination of datasets

In the experiments, we employed Adult and Diabetes datasets. A brief information about these datasets is given below.

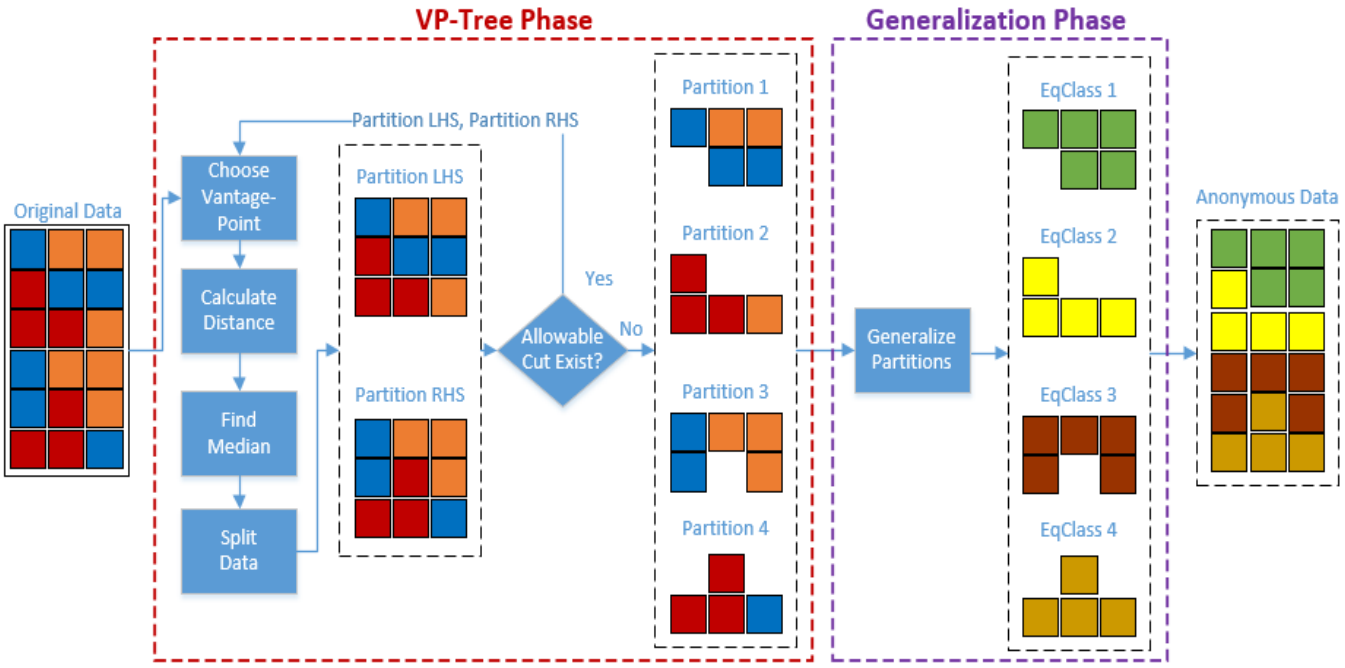


Fig. 4: The block diagram of the proposed model CANON

Adult dataset [59] is frequently used in anonymization studies and includes total 48,842 records. 18,680 records are incomplete and by removing them 30,162 records are obtained and employed in the experimental studies. Since we considered only numerical attributes through this study, we determined the quasi-identifiers as age, $final\_weight$, $capital\_loss$, $capital\_gain$ and $hours\_per\_week$.

Diabetes dataset [60] was employed tı verify the proposed model. It contains 101,766 records and only numerical attributes are considered. $mean\_age$, $num\_medications$, $number\_outpatient$, $num\_lab\_procedures$, $num\_procedures$, $number\_emergency$, $number\_impatient$ and $number\_diagnosis$ were employed as quasi-identifiers.

B. Giving data utility metrics

In this study, we used Discernibility Metric (DM) [61], Average Equivalence Class Size (AECS) [32] and Generalized Certainty Penalty (GCP) [62] metrics to evaluate data utilities produced by the models. In addition, we evaluated the number of equivalence classes (ES) created by the models.

Consider that D is a dataset, EQ is a set of equivalence classes, qid is a quasi-identifier, k is an anonymity level, E is any equivalence class, NCP is normalized certainty penalty metric, n is the number of records in D and d is the dimension of D. DM, AECS and GCP metrics are calculated according to Eq(1), Eq(2) and Eq(3), respectively.

$$DM(D) = \sum_{qid_i} |D[(qid)_i]|^2 \tag{1}$$

$$AECS(D) = \frac{|D|}{(|EQ| * k)} \tag{2}$$

$$GCP(D) = \frac{\sum_{E \in D} |D| * NCP(E))}{(d * n)} \tag{3}$$

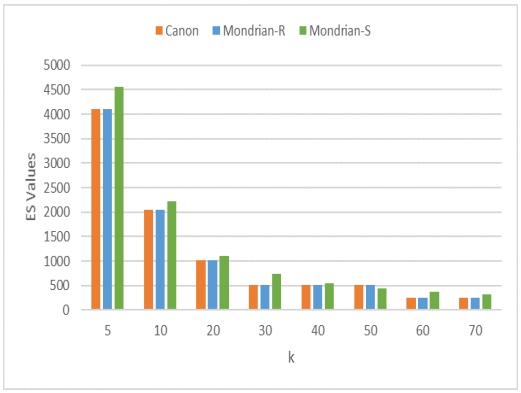
DM, AECS and GCP metrics are used to measure the size of equivalence classes, the size of average equivalence classes and the perimeter of equivalence classes, respectively. It should be emphasized that DM, AECS and GCP metrics are inversely proportional to data utility. Hence lower values of these metrics represent higher data utility and vice versa. In addition, ES values indicate the number of equivalence classes created by algorithm.

Note that CANON almost inspires from relaxed partitioning strategy of Mondrian. Hence, comparing the results of CANON and Mondrian with Relaxed partitioning (Mondrian-R) mainly reveals the success of this paper. But, it is worth to compare CANON to Mondrian with Strict partitioning (Mondrian-S).

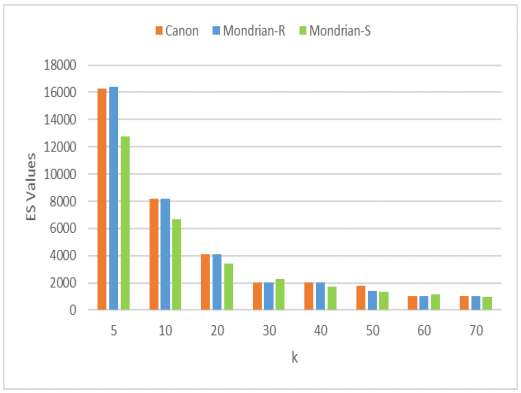
C. Conducting experiments

Two experiments were performed in this section. Experiment 1 and Experiment 2 are conducted to test and verify CANON.

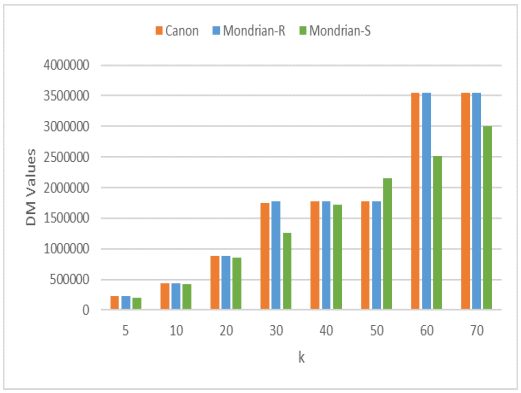
Experiment 1: Adult dataset is employed to test CANON, Mondrian-R and Mondrian-S. As presented on the left of Figure 5, DM, AECS and ES values for Mondrian-R and CANON are almost the same. The main reason is that both CANON and Mondrian-R have the same lower and upper bounds as mentioned earlier. Since DM, AECS and ES measure the size of equivalence classes, the size of average equivalence classes and the number of equivalence classes, these metrics should give almost the same results for both models. The main contribution of this paper is that CANON presents a proximity-aware anonymization and also considers data distribution with



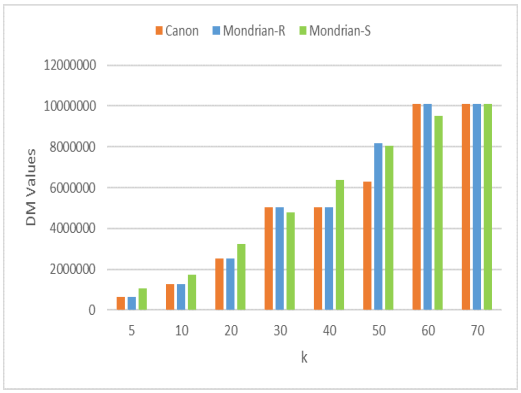
(a) ES Values for Adult Dataset



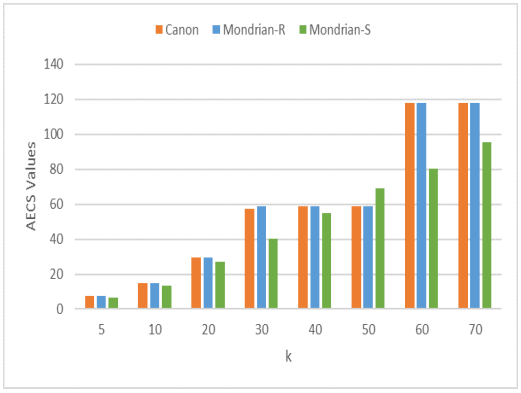
(b) ES Values for Diabetes Dataset



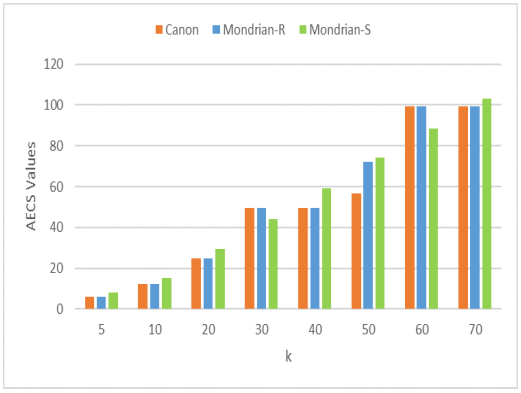
(c) DM Values for Adult Dataset



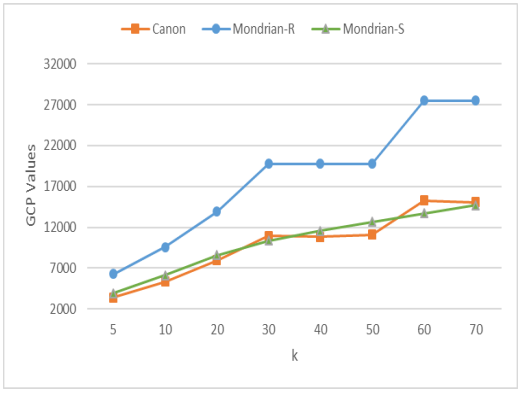
(d) DM Values for Diabetes Dataset



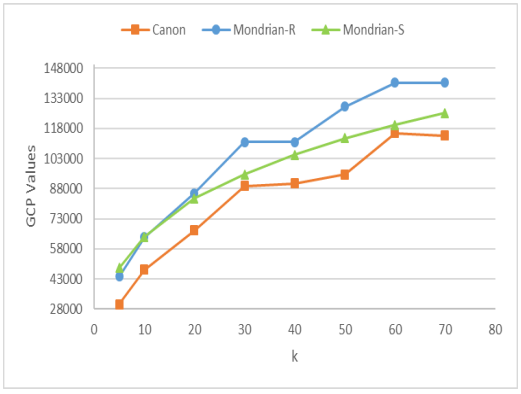
(e) AECS Values for Adult Dataset



(f) AECS Values for Diabetes Dataset



(g) GCP Values for Adult Dataset



(h) GCP Values for Diabetes Dataset

Fig. 5: Experimental Results

a distance based manner. Therefore, CANON proposes higher data utility than Mondrian-R in terms of GCP metric. However, Mondrian-S provides generally higher ES and lower DM and AECS values than other two models. If a comparison is performed for GCP metric, it can be seen that CANON also presents higher data utility than Mondrian-S, for some k values.

Experiment 2: Diabetes dataset was employed to verify the performance of the proposed model. In the right of Figure 5, the results have shown that CANON and Mondrian-R present almost the same ES, DM and AECS values, but Mondrian-S provides different values than these two models. Again, GCP metric is the key factor to evaluate the performance of CANON. The GCP results of CANON show lower values than other two models. Hence, CANON provides better data utility than Mondrian-R and Mondrian-S for different k values.

Lower GCP values represent that closer data points are located in each equivalence classes. Therefore, CANON creates equivalence classes with closer data points compared to Mondrian-R and Mondrian-S.

VIII. Conclusion

This paper successfully introduces a new multidimensional anonymization model called CANON. The proposed model uses VP-tree and multidimensional generalization for greedy partitioning and anonymization, respectively.

CANON is a VP-tree based anonymization model by employing a distance-based partitioning and distribution-aware splitting. As can be seen from the results of two datasets that the proposed model provides higher data utility than Mondrian for both strategies in terms of GCP metric. CANON takes data distribution into consideration and creates equivalence classes including closer data points compared to Mondrian, and provides better data utility.

For Adult dataset; CANON presents 45.47% and 13.55% higher data utility than Mondrian for both strategies in terms of GCP metric. For Diabetes dataset; CANON presents 31.01% and 37.04% higher data utility than Mondrian for both strategies in terms of GCP metric.

The proposed model introduced in this article is promising for future works, it can be accepted as a base model for anonymization and further studies may extend the proposed model to obtain higher utilities.

Acknowledgment

Code availability: https://github.com/ycanbay/canon
Authors' contributions: Yavuz Canbay: Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review and editing. Seref Sagiroglu: Supervision, Methodology, Writing - original draft, Writing - review and editing. Yilmaz Vural: Software, Validation, Writing - original draft, Writing - review and editing.

References

[1] W. Fang, X. Wen, Y. Zheng, M. Zhou, "A Survey of Big Data Security and Privacy Preserving", *IETE Technical Review*, vol. 34, pp. 544-560, 2017.

[2] A. Hasan, Q. Jiang, "A General Framework for Privacy Preserving Sequential Data Publishing", International Conference On Advanced Information Networking And Applications Workshops, pp. 519-524, Taipei, Taiwan, 2017.

[3] M. Almasi, T. Siddiqui, N. Mohammed, H. Hemmati, "The Risk-Utility Tradeoff for Data Privacy Models", International Conference On New Technologies, Mobility And Security, pp. 1-5, Larnaca, Cyprus, 2016.

[4] X. Chen, V. Huang, "Privacy Preserving Data Publishing for Recommender System", IEEE Annual Computer Software And Applications Conference Workshops, pp. 128-133, Izmir, Turkey, 2012.

[5] M. Akgun, "An Active Genomic Data Recovery Attack", *Balkan Journal of Electrical and Computer Engineering*, pp. 417-423, 2019.

[6] R. Wang, Y. Zhu, C. Chang, Q. Peng, Q, "Privacy-preserving High-dimensional Data Publishing for Classification", *Computers And Security*,vol 93. 2020.

[7] M. Chibba, A.Cavoukian, "Privacy, Consumer Trust and Big Data: Privacy by Design and the 3 C's", *ITU Kaleidoscope: Trust In The Information Society*, pp. 1-5, 2015.

[8] P. Jain, M. Gyanchandani, N. Khare, "Big Data Privacy: A Technological Perspective and Review", *Journal Of Big Data*, vol. 3, no. 25, 2016.

[9] J. Nayahi, V. Kavitha, "Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop". *Future Generation Computer Systems*, vol. 74, pp. 393-408, 2017.

[10] Q. Tang, Y. Wu, S. Liao, X. Wang, "Utility-based k-Anonymization", International Conference On Networked Computing And Advanced Information Management, pp. 318-323, Seoul, 2010.

[11] B. Fung, K. Wang, A. Fu, S. Philip, "Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques", CRC Press, 2010.

[12] B. Fung, K. Wang, R. Chen, P. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments", *Computing Surveys*, vol. 42, no. 14, 2010.

[13] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*, vol. 10, pp. 557-570, 2002.

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity", International Conference on Data Engineering, Atlanta, USA, 2006.

[15] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", IEEE International Conference On Data Engineering, pp. 106-115, Istanbul, Turkey, 2007.

[16] S. Abdelhameed, S. Moussa, M. Khalifa, "Privacy-Preserving Tabular Data Publishing: A Comprehensive Evaluation From Web to Cloud", *Computers And Security*, vol. 72, pp. 74-95, 2017.

[17] C. Dwork, "Differential Privacy", International Colloquium On Automata, Languages And Programming, pp. 1-12, Venice, Italy, 2006.

[18] A. Meyerson, R. Williams, "On the Complexity of Optimal k-Anonymity", ACM SIGMOD-SIGACT-SIGART Symposium On Principles Of Database Systems, pp. 223-228, Paris, France, 2004.

[19] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D., Thomas, A. Zhu, "Approximation Algorithms for k-Anonymity", *Journal Of Privacy Technology*, pp. 1-18, 2005.

[20] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D., Thomas, A. Zhu, "Anonymizing Tables", International Conference On Database Theory, pp. 246-258, Edinburgh, UK, 2005.

[21] X. Sun, L. Sun, H. Wang, "Extended k-Anonymity Models Against Sensitive Attribute Disclosure", *Computer Communications*, vol. 34, pp. 526-535. 2011.

[22] P. Bonizzoni, G. Della Vedova, R. Dondi, "Anonymizing Binary and Small Tables is Hard to Approximate", *Journal Of Combinatorial Optimization*, vol. 22, pp. 97-119, 2011.

[23] J. Blocki, R. Williams, "Resolving the Complexity of Some Data Privacy Problems", International Colloquium On Automata, Languages, And Programming, pp. 393-404, Bordeaux, France, 2010.

[24] A. Scott, V. Srinivasan, U. Stege, "k-Attribute-Anonymity is Hard Even for k=2", *Information Processing Letters*, vol. 115, pp. 368-370, 2015.

[25] R. Chen, B. Fung, N. Mohammed, B. Desai, K. Wang, "Privacy-preserving trajectory data publishing by local suppression", *Information Sciences*, vol. 231, pp. 83-97, 2013.

[26] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*, vol. 10, pp. 571-588, 2002.

[27] K. LeFevre, D. DeWitt, R. Ramakrishnan, "Incognito: Efficient Full-domain k-Anonymity", ACM SIGMOD International Conference On Management Of Data, pp. 49-60, Baltimore, Maryland, 2005.

[28] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, K. Kuhn, "Flash: Efficient, Stable and Optimal k-Anonymity", International Conference On Privacy, Security, Risk And Trust And International Confernece On Social Computing, pp. 708-717, Amsterdam, Netherlands, 2012.

[29] L. Sweeney, "Datafly: A System for Providing Anonymity in Medical Data", *Database Security XI*, pp. 356-381, 1998.

[30] K. Wang, P. Yu, S. Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection", IEEE International Conference On Data Mining, pp. 249-256, Brighton, UK, 2004.

[31] B. Fung, K. Wang, P. Yu, "Top-Down Specialization for Information and Privacy Preservation", International Conference On Data Engineering, pp. 205-216, Tokyo, Japan, 2005.

[32] K. LeFevre, D. DeWitt, R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity", International Conference On Data Engineering, pp. 25-25, Atlanta, USA, 2006.

[33] L. Kacha, A. Zitouni, M. Djoudi, "KAB: A new k-anonymity approach based on black hole algorithm", *Journal Of King Saud University-Computer And Information Sciences*, in press, 2021.

[34] B. Bhati, J. Ivanchev, I. Bojic, A. Datta, D. Eckhoff, "Utility-Driven k-Anonymization of Public Transport User Data", *IEEE Access*, vol. 9, pp. 23608-23623, 2021.

[35] W. Mahanan, W. Chaovalitwongse, J. Natwichai, "Data privacy preservation algorithm with k-anonymity", *World Wide Web*, vol. 24, pp. 1551-1561, 2021.

[36] J. Andrew, J. Karthikeyan, "Privacy-preserving big data publication:(K, L) anonymity", Intelligence In Big Data Technologies—Beyond The Hype, pp. 77-88, 2021.

[37] P. Wang, P. Huang, Y. Tsai, R. Tso, "An Enhanced Mondrian Anonymization Model based on Self-Organizing Map", Asia Joint Conference On Information Security, pp. 97-100, Taipei, Taiwan, 2020.

[38] J. Andrew, J. Karthikeyan, J. Jebastin, "Privacy preserving big data publication on cloud using mondrian anonymization techniques and deep neural networks", International Conference On Advanced Computing And Communication Systems, pp. 722-727, Coimbatore, India, 2019.

[39] A. Nezarat, K. Yavari, "A distributed method based on mondrian algorithm for big data anonymization", International Congress On High-Performance Computing And Big Data Analysis, pp. 84-97, Tehran, Iran, 2019.

[40] F. Ashkouti, A. Sheikhahmadi, "DI-Mondrian: Distributed improved Mondrian for satisfaction of the L-diversity privacy model using Apache Spark", *Information Sciences*, vol. 546, pp. 1-24, 2021.

[41] Q. Tang, Y. Wu, X. Wang, "New Algorithm with Lower Upper Size Bound for k-Anonymity", International Conference On Communication Systems, Networks And Applications, pp. 421-424, Hong Kong, China, 2010.

[42] K. Liu, C. Kuo, W. Liao, P. Wang, "Optimized Data de-Identification Using Multidimensional k-Anonymity", International Conference On Trust, Security And Privacy In Computing And Communication, International Conference On Big Data Science And Engineering, pp. 1610-1614, New York, USA, 2018.

[43] Q. Gong, M. Yang, Z. Chen, J. Luo, "Utility Enhanced Anonymization for Incomplete Microdata", nternational Conference On Computer Supported Cooperative Work In Design, pp. 74-79, Nanchang, China, 2016.

[44] M. Nergiz, M. Gök, "Hybrid k-Anonymity", *Computers And Security*, vol. 44, pp. 51-63, 2014.

[45] K. LeFevre, D. DeWitt, R. Ramakrishnan, "Workload-aware Anonymization", ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, pp. 277-286, 2006.

[46] N. Kumar, L. Zhang, S. Nayar, "What is a Good Nearest Neighbors Algorithm for Finding Similar Patches in Images", Springer, 2008.

[47] M. Dolatshah, A. Hadian, B. Minaei-Bidgoli, "Ball*-Tree: Efficient Spatial Indexing for Constrained Nearest-Neighbor Search in Metric Spaces", Preprint ArXiv:1511.00628, 2015.

[48] I. Witten, E. Frank, M. Hall, C. Pal, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 2016.
[49] J. Wang, N. Wang, Y. Jia, J. Li, G. Zeng, H. Zha, X. Hua, "Trinary-Projection Trees for Approximate Nearest Neighbor Search", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 36, pp. 388-403, 2014.
[50] J. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching", *Communications Of The ACM*, vol. 18, pp. 509-517, 1975.
[51] P. Yianilos, "Data Structures and Algorithms for Nearest Neighbor Search In General Metric Spaces", Symposium On Discrete Algorithms, vol. 93, pp. 311-321, Texas, USA, 1993.
[52] A. Fu, P. Chan, Y. Cheung, Y. Moon, "Dynamic Vp-Tree Indexing for N-Nearest Neighbor Search Given Pair-Wise Distances", *International Journal On Very Large Data Bases*, vol. 9, pp. 154-173, 2000.
[53] T. Bozkaya, M. Ozsoyoglu, "Distance-Based Indexing for High-Dimensional Metric Spaces", ACM SIGMOD International Conference On Management Of Data, pp. 357-368, 1997.
[54] T. DeFreitas, H. Saddiki, P. Flaherty, "GEMINI: A Computationally-Efficient Search Engine for Large Gene Expression Datasets", *BMC Bioinformatics*, vol. 17, pp. 102, 2016.
[55] C. Böhm, S. Berchtold, D. Keim, "Searching in High-Dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases", *ACM Computing Surveys*, vol. 33, pp. 322-373 2001.
[56] A. Kuznetsov, E. Myasnikov, "Copy-Move Detection Algorithm Efficiency Increase Using Binary Space Partitioning Trees", CEUR Workshop Proceedings, pp. 373-378, 2016.
[57] F. Zhang, P. Di, H. Zhou, X. Liao, J. Xue, "RegTT: Accelerating Tree Traversals on GPUs by Exploiting Regularities", International Conference On Parallel Processing, pp. 562-571, Philadelphia, USA, 2016.
[58] T. Kristensen, C. Pedersen, "Data Structures for Accelerating Tanimoto Queries on Real Valued Vectors", International Workshop On Algorithms In Bioinformatics, pp. 28-39, Liverpool, UK, 2010.
[59] D. Dheeru, E. Taniskidou, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml
[60] B. Strack, J. DeShazo, C. Gennings, J. Olmo, S. Ventura, K. Cios, J. Clore, "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records", BioMed Research International, 2014.
[61] A. Skowron, C. Rauszer, "The Discernibility Matrices and Functions in Information Systems", Intelligent Decision Support, pp. 331-362, 1992.
[62] G. Ghinita, P. Karras, P. Kalnis, N. Mamoulis, "Fast Data Anonymization with Low Information Loss", International Conference On Very Large Databases, pp. 758-769, Vienna, Austria, 2007.

Yavuz CANBAY received his Ph.D. degree from Department of Computer Engineering, Gazi University in 2019. Currently, he is an assistant professor in Kahramanmaras Sutcu Imam University. Data privacy, big data analytics and information security are his main research interests.

Seref Sagiroglu is a professor in the Department of Computer Engineering at Gazi University. He is the editor of International Journal of Information Security Science. His research interests include machine learning, intelligent system identification, recognition, modeling and control, artificial intelligence, software engineering, information and computer security, biometry, malware detection, big data analytics, cyber security.

Yilmaz Vural is a researcher in the Department of Computer Science at University of California, Santa Barbara. He received his Ph.D. Degree in Computer Engineering from Hacettepe University. His research interests include data privacy, information and computer security.