This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3424399

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

On the Complexity of Optimal k-Anonymity: A New Proof based on Graph Coloring

YAVUZ CANBAY

Sutcu Imam University, Department of Computer Engineering, Kahramanmaras, Turkey

Corresponding author: Yavuz CANBAY (e-mail: yavuzcanbay@ksu.edu.tr).

ABSTRACT Privacy is a complex balancing problem between risks and the utility of data. K-anonymity, a fundamental model for preserving privacy, guarantees that an item cannot be differentiated from at least k-1 other items. Due to the k-anonymity is a hard problem, which means obtaining an optimal solution within a reasonable time is not possible, researchers endeavor to create near-optimal solutions. There are some researches in literature demonstrate the NP-Hardness of achieving k-anonymity. The problems of k-dimensional perfect matching, edge partition into triangles, minimum vertex covering, and maximum k-dimensional matching with k-occurrences are some examples of NP-Complete problems commonly used for reduction to prove the NP-Hardness of k-anonymity. This study presents a significant contribution by providing a new proof for the NP-Hardness of k-anonymity. The proof is achieved using a reduction from the graph coloring problem, which is being provided for the first time. The proof enabled us to enhance both the alphabet size and the number of suppressed cells.

INDEX TERMS graph coloring, k-anonymity, np-hardness, proof

I. INTRODUCTION

The digitalization of society enables the handling of an increasing amount of data pertaining to the real world, machines, individuals, and so on. Currently, numerous institutions or parties, referred as data curators, gather and retain data from various individuals and entities such as clients, patients, users, firms, and institutions. The primary objectives of these acts include fulfilling their goals, enhancing services such as customer modeling, identifying behavioral patterns, diagnosing diseases, formulating plans, establishing regulations, and constructing decision-making procedures, etc. In some cases, it is necessary to publish or share the data in order to maximize the benefits. Through this approach, one can achieve outcomes that have a direct and positive impact on respondents at all levels, ranging from the individual to the country as a whole [1-3]. On the other hand, one of the most crucial concerns with data publishing is privacy.

The concept of privacy was initially presented by Warren and Brandeis in 1890 [4] and defined as the "right to be let alone". Today, the protection of this right is ensured by legal measures, as it is considered both essential and hot topic. The preservation of privacy of individuals, who may be encounter with numerous cyber-attacks, can be defined as the right to maintain one's individuality in both physical and digital domains, wherein individuals establish their personal boundaries. The limits exhibit variability throughout different cultures, countries, religions, and even among individuals [5]. In addition to the definitions of personal privacy, it may be helpful to consider certain definitions of data privacy in order to better grasp the concept of privacy in the context of data.

In the literature, some works provide definitions for data privacy, such as the concept of "informational self-determination" [6] and “the appropriate use of responders’ information and the ability to decide what information of a responder goes where” [7]. Impressive definitions can be provided such as, the ability of the data owner to selective control the borders of data sharing, including with whom, for what purpose, and to what extent. Another meaning can be given as "the right to be data". In the light of the growing accumulation of personalized data, protecting data privacy has emerged as a critical necessity and an essential prerequisite for conducting privacy-preserving data analysis [8, 9]. Any attempt to direct publishing of raw data may violate the privacy of responders. Hence, it is vital to employ strategies that eradicate breaches of privacy.

Anonymization and cryptography are the primary methods used for protecting the privacy of data [5]. Cryptography employs cryptographic keys to encrypt data, whereas anonymization masks the identity of respondents. Due to the lack of usefulness of encrypted data, encryption is not a recommended method for publishing of data. Anonymization is the method of grouping data records in such a way that each member of a group cannot be differentiated from others based on certain features. This strategy is often favored for maintaining data privacy while yet ensuring the data with utility [10].

Several privacy-preserving approaches have been established to resist privacy-focused disclosure attacks, such as record linkage attack, attribute linkage attack and probability attack. k-anonymity, l-diversity and t-closeness are well-known and important models for protecting privacy. k-anonymity is a method that addresses the issue of record linkage attack by guaranteeing that a record or tuple is can not be discriminated from at least k-1 other records [11]. l-diversity ensures the presence of diverse sensitive information within equivalence classes, effectively mitigating record linkage attack and attribute linkage attack [12]. t-closeness is a method that addresses the issues of attribute linkage attack and probability attack. It ensures a balance between the distribution of sensitive data both inside each equivalence class and over the entire table [13]. A comprehensive list and their explanations can be find in [14]. While previous studies related to this topic acknowledge that each record is associated with a distinct individual, subsequent researches have begun to acknowledge the possibility that one human may have many records [15-17].

While k-anonymity offers a solution to privacy concerns, the complexity of k-anonymity is an additional matter that needs to be handled. Previous researches have shown that brute force methods for achieving k-anonymity exhibit an exponential relationship between input size and the number of possible solutions. Hence, the literature emphasizes that achieving k-anonymity is a computationally challenging task classified as NP-Hard, necessitating the use of near-optimal methods [18, 19].

This work proofs the NP-Hardness of k-anonymity by demonstrating a reduction from the graph coloring problem. The paper is organized as follows. Section 2 provides some briefs about previous proofs. In Section 3, a novel proof utilizing graph coloring was introduced to establish the NP-Hardness of k-anonymity. The conclusion and discussions were provided in Section 4.

II. PREVIOUS PROOFS ON THE HARDNESS OF K-ANONYMITY

Several studies in the literature specifically address the NP-Hardness of k-anonymity.

Meyerson and Williams [18] employed a reduction from k-dimensional perfect matching problem to examine the complexity of k-anonymity. They stated that if there is no limitation on the size of the alphabet, achieving k-anonymity becomes computationally difficult and falls under the category of NP-Hard problem for $k \geq 3$. Additionally, they mentioned that the maximum number of suppressed cells is also $n(m-1)$.

However, Aggarwal et al. [19, 20] decreased the size of the alphabet to 2, while the number of suppressed cells were remained as $n(m-1)\lceil \log_2(n(m-1)/3) \rceil$ (in [19], the number of suppressed cells is determined as $9m$ where m indicates the number of triangles. Nevertheless, in order to do a comparison between the number of suppressed cells in all proofs, we generalized and assumed $9m$ as $n(m-1)$, where n represents the number of rows and m indicates the number of columns for $n(m-1)$. Therefore we accepted that $9mt$ equals to $n(m-1)\lceil \log_2(n(m-1)/3) \rceil$). They employed a reduction technique to transform the edge partition problem into triangles problem.

In a similar manner, Sun et al. [21] reported an alphabet size of 2, while the number of repressed cells was determined as nm (in [21], the number of suppressed cell is presented as $48m$. However, in order to do a comparision between the number of suppressed cells in all proofs, we generalized and assumed $48m$ as nm, where n indicates the number of rows and m indicates the number of columns. Consequently, we accepted that $48m$ equals to nm). In their proof, they utilized edge partitioning into 4-cliques and claimed that (p,α)-sensitive k-anonymity is NP-Hard.

In addition, Bonizzoni et al. [22], specifically examined two restricted instances of the k-anonymity problem. The researchers demonstrated that achieving 3-anonymity is APX-Hard under the constraint of a binary alphabet. Furthermore, they established that achieving 4-anonymity remains APX-Hard even when the number of rows has a length of 8. The problem of minimum vertex cover on a cubic graph was also employed.

In a different study, Blocki and Williams [23] presented proof that utilizes a reduction from the problem of maximum 3-dimensional matching with 3 occurrences. The number of attributes in each record was limited to 27, and it was demonstrated that with this constraint, achieving 3-anonymity is a computationally hard problem known as Max SNP-Hard.

Further, the proof provided by Scott et al. [24] indicates that the process of anonymizing k-attribute is also NP-Hard, for $k \geq 2$. They employed a reduction from c-Hitting Set problem.

Finally, LeFevre et al. [25] employed a reduction from partition to demonstrate the NP-Hardness of optimal k-anonymous multidimensional partitioning. They used discernibility metric to approximate optimal solution.

Based on the aforementioned proofs, we conducted a thorough review of two studies considering the hardness of k-anonymity. The reviews are provided below.

A. PROOF USING A REDUCTION FROM K-DIMENSIONAL PERFECT MATCHING

In [18], the authors employed a reduction from k-dimensional perfect matching to prove that k-anonymity is NP-Hard for $k \geq 3$. To understand the proof better, some notations used in the related paper were given and described in Table 1.

suppressor function: let t be a map from V to $(\sum \cup \{*\})^m$. For all $v \in V$ and $j = 1,...,m$; if $t(v)[j] \in \{v[j],*\}$ then t is a suppressor function.

k-anonymizer: let t be a suppressor function on $V = \{v_1,...,v_n\} \subseteq \sum^m$. $t(V)$ is k-anonymous if for all $v_i \in V$, there exists $k-1$ indices $i_1, i_2, ..., i_{k-1} \in \{1,...,n\}$

ensuring that $t(v_{i_1}) = t(v_{i_2}) = \ldots = t(v_{i_{k-1}}) = t(v_i)$. Therefore, t is called as a k-anonymizer on V.

TABLE I
A DESCRIPTIVE REPRESENTATION OF THE NOTATIONS USED IN [18]

Notation	Description
m	dimension
$\sum$	alphabet of attribute values
V	table, $V \subseteq \sum^m$
t	suppressor function
v	record in $V = \{v_1, \ldots, v_n\}$
v'	suppressed data record in $V' = \{v'_1, \ldots, v'_n\}$
V'	anonymized table
i	row indices $\{i_1, \ldots, i_n\}$
j	column indices $\{j_1, \ldots, j_m\}$
n	number of records
k	minimum cardinality in an equivalence class
$*$	symbol used in suppression
l	maximum number of suppressed vector coordinate in V'
H	hypergraph
U	set of vertices
E	set of edges
S	subset of edge set
u	each vertex in $U = \{u_1, \ldots, u_n\}$
e	each edges in $E = \{e_1, \ldots, e_n\}$
$j(\mathrm{i})$	indices of unique hyper edge containing u_i

k-anonymity as a decision problem: for a given table V and an integer number $l \in N$, is there a suppressor t which makes V k-anonymous and suppresses maximum l coordinates?

It is claimed that if the alphabet size is unlimited, optimal k-anonymity is a hard problem for $k \geq 3$.

Theorem: k-anonymity is NP-Hard for $k \geq 3$ even $|\sum| \geq |V|$

A reduction from k-dimensional perfect matching: let $H = (U, E)$ be a k-hypergraph, n and m be the number of vertices and the number of edges, respectively. In this case, in n/k hyperedges, is there a subset $S \in E$, such that each vertex of U is covered by one hyperedge of S?

Assume H is a k-dimensional simple hypergraph, $U = \{u_1, \ldots, u_n\}$ and $E = \{e_1, \ldots, e_m\}$ are vertices and edges of H, respectively and finally $\sum = \{0, 1, \ldots, n\}$ is the alphabet. Table V can be constructed as below and for each u_i; m dimensional vector $v_i \in \sum^m$ can be defined as;

$$v_i[j] := \begin{cases} 0 & \text{if } u_i \in e_j, \\ 1 & \text{otherwise.} \end{cases}$$

Suppose V includes the series of v_i such as $V := \{v_1, \ldots, v_n\}$. Assume that t suppresses minimum number of vector coordinates and ensures k-anonymity. It was claimed that the number of coordinates suppressed by t is maximum $n(m-1)$ if there exists a k-dimensional perfect matching in H.

This claim is proved for $k = 3$. Firstly, it was accepted that there exists a perfect 3-dimensional matching M in H. For $i = 1, \ldots, n$, let $j(i)$ be the indices of $e_{j(i)}$ which is the unique hyperedge in M containing vertex u_i.

Suppressor t is defined as follows;

$$t(v_i)[j'] := \begin{cases} 0 & \text{if } j' = j(i), \\ * & \text{otherwise.} \end{cases}$$

Because of u_i is on the hyperedge $e_{j(i)}$, the following states occur by definition, $v_i[j(i)] = 0$ and all other coordinates are $*$. Therefore, t is a suppressor on V.

Consider any three nodes $u_i, u_{i'}, u_{i''}$ on $e_{j(i)}$ and each node has identical anonymized vectors such as $t(v_i) = t(v_{i'}) = t(v_{i''})$. Therefore, $t(V)$ has three identical vectors, which shows that $t(V)$ is 3-anonymous.

Since every $t(v) \in t(V)$ has at most one non-$*$ coordinate, the value of the solution is $n(m-1)$. Therefore, the optimum solution of 3-anonymity includes at most $n(m-1)$ number of $*$'s in these vectors.

B. PROOF VIA EDGE PARTITION INTO TRIANGLES

In [19, 20], it was shown that k-anonymity is NP-Hard for $k \geq 3$ edge partition into triangles reduction. To understand the proof in detail, some notations used in the paper were presented in Table 2.

TABLE II
A DESCRIPTIVE REPRESENTATION OF THE NOTATIONS USED IN [19, 20]

Notation	Description
m	number of disjoint triangles
$\sum$	alphabet of attribute values
T	preliminary table for vertex-edge relationship
T'	replication of T for 4-stars
$r_{a,b}$	the row corresponding to edge (a, b)
a, b, c	vertices of triangles
d	central vertices
$*$	symbol used in suppression
t	number of blocks
n	number of columns
G	complete graph
V	set of vertexes
E	set of edges
v	common vertex
e	edges in $E = \{e_1, \ldots, e_n\}$
$conf_i$	configuration $i = \{0, 1\}$
i	indices of blocks
k	minimum number of cardinality in an equivalence class

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3424399

Theorem: k-anonymity is NP-Hard even for $\sum = \{0,1,2\}$.

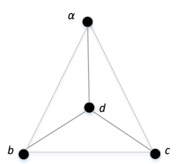
Edge partition into triangles reduction: for a given graph $G=(V,E)$ with $|E|=3m$ for any integer m, can the edges of G be partitioned into m triangle whose edges are disjoint?

The proof starts with the reduction of edge partition into triangles and 4-stars before explaining the reduction of edge partition into triangles.

The proof has two phases. In the first phase, it will be showed that for a graph $G=(V,E)$ with $|E|=3m$, if and only if G can be partitioned into triangle and 4-stars, the optimal 3-anonymity solution for T is $9m$. In the second phase, again it will be showed that if and only if G can be partitioned into m disjoint-triangles, the 3-anonymity solution is maximum $9m\lceil \log_2(3\mathrm{m}) \rceil$ for table T'.

Edge partition into triangles and 4-stars: for a given graph $G=(V,E)$ with $|E|=3m$ and $|V|=n$, a table T with $3m$ row and n attribute is created. The row $r_{a,b}$ corresponding to edge (a,b) has 1 in positions corresponding to attributes a and b, and 0 otherwise. Assume that graph G can be partitioned into m disjoint-triangles and 4-stars. Let a,b and c be the vertices of triangle. In the rows $r_{a,b}$, $r_{b,c}$ and $r_{a,c}$, by suppressing the positions of a,b and c, three identical rows, each containing 3 *s and 0, are obtained. Now consider a 4-star has the vertices a,b,c,d and edges (a,d), (b,d), (c,d). In the rows $r_{a,\mathrm{d}}$, $r_{b,\mathrm{d}}$, $r_{\mathrm{c,d}}$, by suppressing corresponding positions of a,b and c, three identical rows each containing 3 *s, with a single 1 and 0 anywhere else. Hence, for every triangle and 4-stars, 3 identical generalized records are obtained by suppressing 9 cells. In conclusion, table T is 3-anonym with cost $9m$. Figure 1 shows an illustration about this phase.

Edge partition into triangles: Assume $t=1+\lceil \log_2(3\mathrm{m}) \rceil$. Let T' is a table whose each row has t blocks and n columns. For an arbitrary order of edges in E, the rank of an edge $e=(a,b)$ can be expressed in a binary form such as $b_1,\ldots,\mathrm{b}_t$. In tuples corresponding to edge e, each block is 0 except a and b. Any block can be in two configurations based on the values of a and b. $conf_0$: 1 in position a and 2 in position b, $conf_1$: 2 in position a and 1 in position b. i^{th} block in corresponding row of e, has $conf_{b_i}$. For example, the edges in Figure 1 is ranked from 1 to 6 and T' is presented in Figure 2. It can be understand that if and only if E is partitioned into m disjoint-triangles, optimal 3-anonymity cost is at most $9mt$ for T'.



	a	b	c	d
(a,b)	1	1	0	0
(a,c)	1	0	1	0
(b,c)	0	1	1	0
(a,d)	1	0	0	1
(b,d)	0	1	0	1
(c,d)	0	0	1	1

T

	a	b	c	d
(a,b)	*	*	*	0
(a,c)	*	*	*	0
(b,c)	*	*	*	0
(a,d)	*	*	*	1
(b,d)	*	*	*	1
(c,d)	*	*	*	1

T_A

FIGURE 1. Anonymizing of triangle and 4-stars obtained from graph

		a	b	c	d	a	b	c	d	a	b	c	d
001	(a,b)	1	2	0	0	1	2	0	0	2	1	0	0
010	(a,c)	1	0	2	0	2	0	1	0	1	0	2	0
011	(b,c)	0	1	2	0	0	2	1	0	0	2	1	0
100	(a,d)	2	0	0	1	1	0	0	2	1	0	0	2
101	(b,d)	0	2	0	1	0	1	0	2	0	2	0	1
110	(c,d)	0	0	2	1	0	0	2	1	0	0	1	2

		a	b	c	d	a	b	c	d	a	b	c	d
001	(a,b)	*	*	*	0	*	*	*	0	*	*	*	0
010	(a,c)	*	*	*	0	*	*	*	0	*	*	*	0
011	(b,c)	*	*	*	0	*	*	*	0	*	*	*	0
100	(a,d)	*	*	*	1	*	*	*	*	*	*	*	*
101	(b,d)	*	*	*	1	*	*	*	*	*	*	*	*
110	(c,d)	*	*	*	1	*	*	*	*	*	*	*	*

FIGURE 2. Anonymized table obtained from graph in Figure 1.

III. A NEW PROOF BASED ON GRAPH COLORING

A new proof based on graph coloring was presented in this section. In the literature, there are some studies using different types of NP-Complete problems for reductions. We preferred to use graph 3-coloring problem for the reduction in our work. Our aim is to examine whether graph coloring problem can be used for a reduction to prove the NP-Hardness of k-anonymity and whether it enables us to improve both alphabet size and number of suppressed cells. In addition, graph coloring based representation also presents simplicity for a better understanding.

Garey and Johnson [26, 27] proved that graph 3-colorability with no vertex degree exceeding 4 is NP-Complete. In our study, we borrowed and adopted this idea and then investigated the availability of graph 3-coloring problem to prove the NP-Hardness of k-anonymity

A. REVISIT OF THE PROOF OF GAREY AND JOHNSON

Garey and Johnson [26, 27] proposed to restrict the maximum vertex degree of graph will be colored. If the maximum vertex degree is restricted with a small enough degree, many graph coloring problem is solved in polynomial time. Hence, finding the most powerful constraint of vertex degree that will keep the problem in NP-Complete is very important. Table 3 presents the complexity classes of problems that a sub problem can be belong to any of them with a degree constraint.

TABLE III
CLASSIFICATION OF SUBPROBLEMS BASED ON COMPLEXITY CLASSES THAT ANY D VERTEX DEGREE LIMITED GRAPH CAN BE BELONG

Problem	P ($D \leq$)	NP-Complete ($D \geq$)
Exact Cover	2	3
Hamilton Cycle	2	3
Graph 3-colorability	3	4
Feedback vertex set	2	3

In this context, the maximum vertex degree in graph 3-colorability problem is presented as 4. This implies that, subproblem is still in NP-Complete class even the constraint of vertex degree of graph 3-colorability problem is determined as 4. Hence, in the proof, maximum vertex degree is limited with 4.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3424399

In order to prove the results of degree limited NP-Completeness, vertex substitute approach is used. Vertex substitution is defined as substituting a vertex with a subgraph that meets some certain criteria.

Theorem: Graph 3-colorability with no vertex degree exceeding 4 is an NP-Complete problem.

Proof: Assume $G=(V,E)$ is an arbitrary graph of a general problem, $G'=(V',E')$ is a restricted instance of G that no vertex of G' have the degree exceeding 4. If and only if G is 3-colorable, G' is also 3-colorable.

In this vertex substitution approach, a graph with eight vertex is considered. In Figure 1.a, graph H_3 has three outlets with labels 1, 2 and 3. x is the number of outlets, for $x \geq 4$, H_x which is a vertex substitution with x number of outlets is formed with adjoining a copy of H_3 to substitution H_{x-1}. In H_5, which is used to prove NP-Completeness of graph 3-colorability with degree restriction, is presented in Figure 1.b. In these graphs, all outlets have the same color.

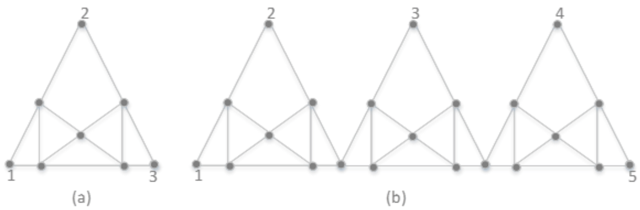


FIGURE 3. Graph H_3 (a) and vertex substitution H_5 (b)

As indicated in Figure 3, for $x \geq 3$, the following situations appear;

1. H_x has $7(x-2)+1$ number of outlets, including x number of labelled outlets.
2. H_x has no vertex whose degree exceed 4.
3. The degree of each outlets of H_x is 2.
4. H_x is 3-colorable, but not 2-colorable. In every different 3-coloring way, outlets of H_x have the same color.

Assume an arbitrary graph G, composing from s number of vertices $v_1,v_2,...,v_s$ and containing vertexes with degree exceeding 4, composed from graph array as shown below;

$$G=G_0,G_1,...,G_s=G'$$

Each G_l, $1 \leq l \leq s$, is constructed from G_{l-1}. Let d be the degree of v_l in G_{l-1} and $\{v_1,v_l\},\{v_2,v_l\},...,\{v_3,v_l\}$ be the edges containing v_l. To form G_l, v_l is deleted from G_{l-1} and is replaced with a copy of H_d. Each edge $\{v_m,v_l\}$ is replaced with an edge joining outlet m and v_m. In the new construction, for $0 \leq x \leq s$, the number of vertex of G_x exceeding 4 is $s-x$, if and only if graph G is 3-colorable, then G_x is 3-colorable. Therefore, $G'=G_s$ is obtained.

The overall approach is illustrated in Figure 4. In graph G, if there exist any vertex whose degree exist 4, by replacing those vertices with vertices in G', G becomes 3-colorable. Because of the graph, which belongs to a general problem and was showed in Figure 4.a, has vertexes with degree exceeding 4, it is replaced with the substitution in Figure 4.b. Thus, graph G becomes 3-colorable. As a result, graph 3-colorability problem with 4- degree restriction is NP-Complete problem.

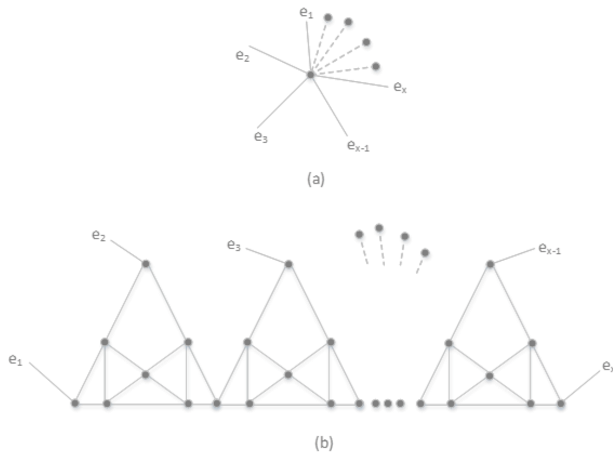


FIGURE 4. An example of vertex with degree exceeding 4 (a) and structure of vertex substitution (b)

B. THE PROPOSED PROOF BASED ON GRAPH COLORING

The NP-Hardness of k-anonymity for $k \geq 3$ was proved using a reduction from degree-limited graph 3-coloring. Some notations used in the proof was presented in Table 4.

TABLE IV
A DESCRIPTIVE REPRESENTATION OF NOTATIONS USED IN OUR PROOF

Notation	Definition
m	dimension
$\sum$	alphabet of attribute values
R	table (adjacency matrix), $R \subseteq \sum^m$
t	suppressor function
r	record in $R=\{r_1,...,r_n\}$
r'	suppressed data record in $R'=\{r'_1,...,r'_n\}$
R'	anonymized version of R
i	row indices $\{i_1,...,i_n\}$
j	column indices $\{j_1,...,j_m\}$
n	number of records
k	minimum cardinality in an equivalence class
$*$	symbol used in suppression
l	maximum number of suppressed vector coordinates in R'
G	graph
V	set of vertices
E	set of edges
C	set of colors

v_i	vertices in $V=\{v_1,...,v_n\}$
$e_{i,j}$	edges in $E=\{e_{1,2},...,e_{q,z}\}$
$M_{x,y,z}$	candidate matrix of corresponding coordinates of x,y,z
x,y,z	some indices
P	color sharing table
P'	anonymized version of P
c_i	colors in $C=\{c_1,...,c_r\}$

Definition: Let t be a map from R to $(\sum\cup\{*\})^m$. For all $r\in R$ and $j=1,...,m$; if $t(r)[j]\in\{r[j],*\}$ then t is a suppressor function.

Each vector $r\in R$ has corresponding $t(r)=r'$ in anonymized table $R'\subseteq(\sum\cup\{*\})^m$. In addition, the coordinates of r' are the same as with the coordinates of r.

In order to work on vector sets in R, t can be extended as follows. $t(R)$ is accepted as a multiple set when one or more vectors in r is mapped to the same suppressed vectors. For instance, $r_1\neq r_2\neq r_3\in R$ but when t is applied, $t(r_1)=t(r_2)=t(r_3)$ is obtained.

Definition: Let t be a suppressor on $R=\{r_1,...,r_n\}\subseteq\sum^m$. $t(R)$ is k-anonymous if for all $r_i\in R$, there exist $k-1$ number of indices $i_1,i_2,...,i_{k-1}\in\{1,...,n\}$ providing $t(r_{i_1})=t(r_{i_2})=...=t(r_{i_{k-1}})=t(r_i)$. Therefore, t can be called as a k-anonymizer on R.

k-anonymity: for a given table R and an integer $l\in N$, is there a suppressor t which makes table R k-anonymous and suppresses maximum l number of coordinates?

Theorem: k-anonymity is NP-Hard for $k\geq 3$ even $\sum=\{0,1\}$.

A reduction from degree-limited graph 3-coloring: given a graph $G=(V,E)$, is G 3-colorable such that for every edge $e_{i,j}\in E$, the color c_i of v_i is different from the color c_j of v_j and there exist no vertex degree exceeding 4.

We investigated this situation for $|V|=6w$, for any integer w. Assume G is a simple graph, $V=\{v_1,...,v_n\}$ and $E=\{e_{1,2},...,e_{q,z}\}$ are vertices and edges of G, respectively and $\sum=\{0,1\}$ is the alphabet. A table R can be created as follows. For every $e_{i,j}$, m-dimensional $r_i\in\sum^m$ vector can be defined as;

$$r_i[j]:=\begin{cases}1 & \text{if } e_{i,j}\in E,\\ 0 & \text{otherwise.}\end{cases}$$

Set $R:=\{r_1,...,r_n\}$. The relations between each vertices can be obtained through R. A color sharing table P can be created as follows. Firstly, one-complement of R is taken and then principal diagonal elements are assigned to zeros. Hereby, P facilitates us to determine which vertices cannot share the same color with v_i. P can be obtained as follows;

$$P=diag(\overline{R})_{zeros}$$

Assume that t suppresses minimum number of vector coordinates and provides k-anonymity. In our proof, it is claimed that if and only if G is 3-colorable, the number of suppressed vector coordinates by t is at most $n(m-3)$. This claim was proved for $k=3$. Firstly it is assumed that G is 3-colorable and has $6w$ number of vertices for any integer w, and V can be partitioned into some disjoint groups each containing triple vertices with colors c_1,c_2,c_3.

For $0\leq x,y,z\leq n$, if any x,y,z coordinates of any 3 vertices are zeros, then these parts are left as they are, otherwise they are replaced with *s. In other words, if $M_{x,y,z}$ is a non-overlapping zeros matrix it saves the original form, but if it is not then it is suppressed with *s. In this case, candidate matrix $M_{x,y,z}$ and suppressor t on $M_{x,y,z}$ can be defined as;

$$M_{x,y,z}=P_{x,y,z}\begin{bmatrix}r_{x,x} & r_{x,y} & r_{x,z}\\ r_{y,x} & r_{y,y} & r_{y,z}\\ r_{z,x} & r_{z,y} & r_{z,z}\end{bmatrix}$$

$$t(M_{x,y,z}):=\begin{cases}0 & \text{if } M_{x,y,z}=0,\\ * & \text{otherwise.}\end{cases}$$

Because of each v_i meets the condition of 3-colorability, the following states occur by definition, each candidate matrix $M_{x,y,z}=0$ and all other coordinates is *s. Hence, t is a suppressor on P.

Consider graph G given in Figure 5. Assume three vertices v_1,v_2,v_3 of graph G is colored with three different colors c_1,c_2,c_3. If P is anonymized, we have the same anonymized vectors $t(r_1)=t(r_2)=t(r_3)$. Therefore, $t(P)$ contains three identical vectors with respect to v_1,v_2,v_3, and this situation shows that $t(P)$ is 3-anonymous. $t(P)$ includes 9 non-*s in each triple anonymized vectors. Therefore, an optimal 3-anonymous solution has at most $n(m-3)$ number of *s.

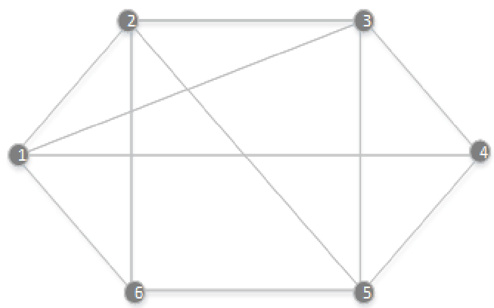


FIGURE 5. **A sample graph** G

With regard to Figure 5, Table R can be created as follows;

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3424399

TABLE V
TABLE R (OR ADJACENCY MATRIX)

	1	2	3	4	5	6
1	0	1	1	1	0	1
2	1	0	1	0	1	1
3	1	1	0	1	1	0
4	1	0	1	0	1	0
5	0	1	1	1	0	1
6	1	1	0	0	1	1

If we take one complement of R and then change the values of principal diagonal elements with zeros, we obtain color sharing table P as presented in Table 6.

Table 6 guides us to obtain the following statements. In Figure 5, v_3, v_4 and v_5 cannot share the same color and each of these vertices is colored with one out of three different colors. Similarly, v_1, v_2 and v_6 have the same condition, and also there may be many other different selections. Table 6 shows that G can be divided into two groups and each group includes exactly three elements with three different colors. For this example, group one contains 1, 2 and 6 while group two includes 3, 4 and 5. A possible 3-coloring of G was presented in Figure 6.

TABLE VI
COLOR SHARING TABLE P

	1	2	3	4	5	6
1	0	0	0	0	1	0
2	0	0	0	1	0	0
3	0	0	0	0	0	1
4	0	1	0	0	0	1
5	1	0	0	0	0	0
6	0	0	1	1	0	0

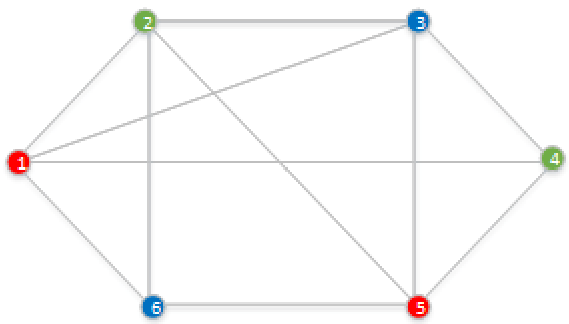


FIGURE 6. A possible 3-coloring of G

Table 7 indicates the anonymized version of P. Therefore, as it was stated in the proof, maximum number of $*$s is $n(m-3)$.

TABLE VII
ANONYM TABLE P'

	1	2	3	4	5	6
1	0	0	*	*	*	0
2	0	0	*	*	*	0
3	*	*	0	0	0	*
4	*	*	0	0	0	*
5	*	*	0	0	0	*
6	0	0	*	*	*	0

Within the context of existing assumptions, the number of suppressed cells of 3-anonymization of G is at most $n(m-3)$. Since the previous proofs present $n(m-1)$, $n(m-1)\lceil \log_2(n(m-1)/3) \rceil$ and nm number of suppressed cells, respectively, our proof reduces it to $n(m-3)$. Hence, the average information losses for each result is obtained as follows;

$$\frac{n(m-3)}{nm} < \frac{n(m-1)}{nm} < \frac{n(m-1)\lceil \log_2(n(m-1)/3) \rceil}{nm} < \frac{nm}{nm}$$

It can be clearly seen that our proof presents a lower average information loss and alphabet size in comparison with other previous proofs and this result may be a good reason for employing graph coloring approach for reductions.

In Table 8, we listed a number of studies available in the literature for the hardness of k-anonymity. We tabularized these studies based on some criteria such as reduction methods, alphabet sizes and average information loss. The results show that our proof provides acceptable outcomes and better results.

TABLE VIII
COMPARISON TABLE FOR THE PROPOSED PROOF AND THE PREVIOUS STUDIES (N/A* = A CERTAIN VALUE IS NOT DEFINED)

Paper	Reduction Method	Alphabet Size	Average Information Loss
Meyerson and Williams [18]	k-dimensional perfect matching	n	$\frac{n(m-1)}{nm}$
Aggarwal et al. [19, 20]	edge partition into triangles	3	$\frac{n(m-1)\lceil \log_2(n(m-1)/3) \rceil}{nm}$
Bonizzoni et al. [22]	minimum vertex cover on cubic graph	2	N/A*
Scott et al. [24]	maximum 3-	N/A*	N/A*

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3424399

	dimensional matching with 3 occurrences		
Sun et al. [21]	edge partition into 4-cliques	2	$\frac{nm}{nm}$
Chen et al. [28]	vertex cover	N/A*	N/A*
Our proof	graph 3-coloring	2	$\frac{n(m-3)}{nm}$

IV. CONCLUSION AND DISCUSSION

As introduced earlier, this paper focuses on the computational complexity of k-anonymity and introduces a new approximation approach. Since k-anonymity is an NP-Hard problem and optimal solutions cannot be achieved in a reasonable time, near-optimal solutions are always required. To prove the NP-Hardness of k-anonymity, especially graph problems are employed for reduction frequently.

To the best of our knowledge, this article proved the NP-Hardness of k-anonymity using a reduction from degree-limited graph 3-coloring for the first time. We also improved both alphabet size and average information loss in comparison with some previous proofs which were listed in Table 8. The results showed that reductions utilize graph coloring presents better results than others. However, in the future, other NP-Complete problems can be examined in terms of whether they present better results.

REFERENCES

[1] A. T. Hasan and Q. Jiang, "A general framework for privacy preserving sequential data publishing," in *Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on*, 2017: IEEE, pp. 519-524.

[2] M. M. Almasi, T. R. Siddiqui, N. Mohammed, and H. Hemmati, "The Risk-Utility Tradeoff for Data Privacy Models," presented at the International Conference on New Technologies, Mobility and Security, Larnaca, Cyprus, 2016.

[3] X. Chen and V. Huang, "Privacy Preserving Data Publishing for Recommender System," presented at the IEEE Annual Computer Software and Applications Conference Workshops, Izmir, Turkey, 2012.

[4] S. D. Warren and L. D. Brandeis, "The Right to Privacy," *Harvard Law Review*, pp. 193-220, 1890.

[5] W. Fang, X. Z. Wen, Y. Zheng, and M. Zhou, "A Survey of Big Data Security and Privacy Preserving," *IETE Technical Review*, vol. 34, no. 5, pp. 544-560, 2017.

[6] M. Chibba and A. Cavoukian, "Privacy, Consumer Trust and Big Data: Privacy by Design and the 3 C's," presented at the ITU Kaleidoscope: Trust in the Information Society, Barcelona, Spain, 2015.

[7] P. Jain, M. Gyanchandani, and N. Khare, "Big Data Privacy: A Technological Perspective and Review," *Journal of Big Data*, vol. 3, no. 1, p. 25, 2016.

[8] J. J. V. Nayahi and 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.

[9] Q. Tang, Y. Wu, S. Liao, and X. Wang, "Utility-based k-Anonymization," presented at the International Conference on Networked Computing and Advanced Information Management, Seoul, South Korea, 2010.

[10] B. C. Fung, K. Wang, A. W. Fu, and S. Y. Philip, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. USA: CRC Press, 2010.

[11] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557-570, 2002.

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

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," presented at the IEEE International Conference on Data Engineering, Istanbul, Turkey, 2007.

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

[15] Q. Gong, J. Luo, M. Yang, W. Ni, and X.-B. Li, "Anonymizing 1: M Microdata with High Utility," *Knowledge-based Systems*, vol. 115, pp. 15-26, 2017.

[16] Y. Tao, Y. Tong, S. Tan, S. Tang, and D. Yang, "Protecting the Publishing Identity in Multiple Tuples," in *Annual Conference on Data and Applications Security and Privacy*, London, UK, 2008: Springer, pp. 205-218.

[17] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos, "Anonymizing Data with Relational and Transaction Attributes," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Prague, Czech Republic, 2013: Springer, pp. 353-369.

[18] A. Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," presented at the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Paris, France, 2004.

[19] G. Aggarwal *et al.*, "Approximation Algorithms for k-Anonymity," *Journal of Privacy Technology*, pp. 1-18, 2005. Paper No. 20051120001.

[20] G. Aggarwal *et al.*, "Anonymizing Tables," in *International Conference on Database Theory*, Edinburgh, UK, 2005: Springer, pp. 246-258.

[21] X. Sun, L. Sun, and H. Wang, "Extended k-Anonymity Models Against Sensitive Attribute Disclosure," *Computer Communications*, vol. 34, no. 4, pp. 526-535, 2011.

[22] P. Bonizzoni, G. Della Vedova, and R. Dondi, "Anonymizing Binary and Small Tables is Hard to Approximate," *Journal of Combinatorial Optimization*, vol. 22, no. 1, pp. 97-119, 2011.

[23] J. Blocki and R. Williams, "Resolving the Complexity of Some Data Privacy Problems," in *International Colloquium on Automata, Languages, and Programming*, Bordeaux, France, 2010: Springer, pp. 393-404.

[24] A. Scott, V. Srinivasan, and U. Stege, "k-Attribute-Anonymity is Hard Even for k=2," *Information Processing Letters*, vol. 115, no. 2, pp. 368-370, 2015.

[25] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," presented at the International Conference on Data Engineering, Atlanta, USA, 2006.

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, information security and artificial intelligence are his main research interests. He is the director of Data Vision Lab (DEVLAB).