A Classification of non-Cryptographic Anonymization Techniques ensuring Privacy in Big Data

Zakariae El Ouazzani[1] and Hanan El Bakkali[1]

[1]Rabat-IT Center, ENSIAS, Mohammed V University, Rabat, Morocco

***Abstract*:** Recently, Big Data processing becomes crucial to most enterprise and government applications due to the fast growth of the collected data. However, this data often includes private personal information that arise new security and privacy concerns. Moreover, it is widely agreed that the sheer scale of big data makes many privacy preserving techniques unavailing. Therefore, in order to ensure privacy in big data, anonymization is suggested as one of the most efficient approaches. In this paper, we will provide a new detailed classification of the most used non-cryptographic anonymization techniques related to big data including generalization and randomization approaches. Besides, the paper evaluates the presented techniques through integrity, confidentiality and credibility criteria. In addition, three relevant anonymization techniques including k-anonymity, l-diversity and t-closeness are tested on an extract of a huge real data set.

***Keywords*:** Big Data, Privacy, Anonymization, K-anonymity, L-diversity, T-closeness

1. Introduction

Due to the immense data growth, the concept of big data has definitely gathered momentum in recent years. Big data refers to the explosive quantity of data generated in today's society. Compared to conventional databases, big data is generally described by a 3Vs model containing Volume, Variety and Velocity [1], [2]. Later, two other features in terms of Value and Veracity were added to the previous ones [3]. Further, big data is extended up to 9V's including Visualization, Viscosity, Virality and Validity [4].

A big amount of the collected data is private such as medical records, financial credentials, Web usage, emails, photos, videos, etc. Unfortunately, at every step of managing such data, there is a possibility that private information may be disclosed due to the extraordinary scale of data. Therefore, the challenge of ensuring privacy is considered as one of the major hurdles in big data applications. At this level, it is generally believed that most of conventional data privacy techniques don't support the sheer scale of big data. Thus, many recent techniques have been proposed in the literature to preserve privacy.

The aim of this paper is to focus on the best way to sanitize data by using anonymization techniques which strike the right balance between preserving data utility and ensuring privacy in big data. Our scientific contribution consists of providing a detailed classification of recent anonymization techniques supporting large scale of data. In the literature, there exist two general approaches which are cryptographic and non-cryptographic ones. However, in this paper, the main focus is to propose a classification dealing with non-cryptographic techniques. The proposed classification is divided into two main categories, and each category contains various anonymization techniques dealing with a specific type of data. The whole mentioned techniques in the proposed classification were discussed, evaluated and compared according to three main criteria: confidentiality, integrity and credibility. In addition, we presented several cases where some anonymization techniques could be applied according to the type of managed data, the degradation of the system's speed, etc; besides, the most appropriate anonymization technique is proposed to each treated case. We focus more on three relevant anonymization techniques belonging to generalization approach which are k-anonymity, l-diversity and t-closeness. In fact, the combination of these techniques ensures privacy, preserves utility and treats both Quasi-Identifier and sensitive attributes. These techniques were tested on an extract of a real huge data set.

The reminder of the paper is organized as follows: in section 2, the main characteristics of big data and its privacy concerns are provided; in section 3, the difference between pseudonymization and anonymization is explained; in section 4, a classification of non-cryptographic anonymization techniques into two main different categories is presented; in sections 5 and 6, different techniques of generalization and randomization approaches are listed respectively. In section 7, some solutions and recommendations are suggested. Finally, a conclusion and future research directions are given.

2. Big data privacy

Big data makes reference to data sets whose size exceeds the capacity of traditional relational databases in order to handle and treat the data within a reasonable response time.

2.1. Big data background

In the literature, there are different types of data such as structured, semi-structured and unstructured data [2], [4], [5]. Structured data is a type where values are classified into records; each record is identified with a unique value and has an identical number of attributes [6]. For instance, structured data could be stored in relational databases, enterprise data warehouses and NoSQL databases [7]. However, semi-structured data is considered as a form of structured data which is not conforming to the formal structure of data models associated with relational databases [8]. Semi-structured data could be used to query more data formats including data types like Extensible Markup Language (XML), Comma-Separated Values (CSV) data, flat files, etc [7], [9]. Whereas, unstructured data is a complex type of big data existing in each organization's data set [6]. About 80% of the total data in the world is unstructured and, usually composed of audio data, images, videos or data from social media [4]. The unstructured data is very difficult to represent

since it cannot be easily stored into a tabular form, thus new mechanisms were introduced such as non-relational database (NOSQL) and enterprise data warehouses [7].

Big data is pertinent to the entire society; the industry uses big data technologies for effective business operations. The government is also concerned in using big data in order to enhance services to citizens. Healthcare is another important field to which big data can offer new opportunities. Big data in healthcare has become an emerging and outstanding field of research; in addition, many other fields of science and engineering are actually facing the growth of the data volume generated by various sources [5].

Big data is measured by the following characteristics which are summarized as "9Vs" [4].

- **Volume:** refers to the enormous quantity of data created by organizations or individuals in a unit of time, either second, minute, hour, or day. It is almost difficult for data providers to manage the whole data they actively or passively furnish to others. As a result, the large volume of data increases the risk of information leakage which may violate individuals' privacy [3].
- **Variety:** illustrates the huge diversity of data formats and sources. The data format covers structured, semi-structured, and unstructured data; for instance, voice and video chat, images, text messages and data gathered from social media discussions. Therefore, not only the huge data infrastructure management up to petabyte level needs to be secured, but also the data management methods dealing with the source of data [4], [10].
- **Velocity:** indicates the speed of generating new data, the continuity and the elevated frequencies of data. This characteristic makes security and privacy issues harder. Fast data growth demands non-relational databases, thus security and privacy must be considered when developing distributed programming frameworks [3], [10].
- **Value:** points to the output gained from big data. Individuals, organizations and companies would get advantage from big data predictive analysis by determining the value inside the data. Generally, the value is obtained by examining certain patterns from the user's list of activities in the data set, analyzing their preferences, behavior or feelings. Therefore, the tradeoff between privacy and utility must be taken into consideration [3], [4].
- **Veracity:** shows the trustworthiness, abnormality, noise, bias, applicability and other properties of the data. In other words, veracity means that the quality and efficiency of the data could not be at the highest level. Besides, veracity is considered as one of the challenging issues in big data analysis when checking the integrity of the mined data, and also when verifying the credibility of the published data [3], [5].
- **Visualization:** allows seeing several dimensions of big data. This characteristic is so important since big data is usually presented as a black box. Despite the fact that new technologies have allowed analysts to get a deeper vision to the vast amount of data, unprecedented visualization is still needed to better understand the data node [4].
- **Viscosity:** represents the resistance when processing certain data sets and also the complexity of its processing. Due to the diversity of data, the complexity of processing every set changes when handling big data. Thus, a fast processing of complex cases is needed to eliminate the resistances related to data sets [4].
- **Virality:** indicates the velocity of data propagation in the network. Knowing the virality content in advance would be helpful in various applications and also for viral markets. In addition, it can be useful for organizations in order to improve the network performances [4].
- **Validity:** refers to both the coherence and power of data. Even though the analysis and processing of the data are fast, accurate decisions could be made [4].

In spite of the various opportunities provided by big data, theoretically, the old 3Vs including Volume, Variety and Velocity are of particular importance because they constitute the basis of big data [4]. Besides, the growth of data has increased privacy concerns. For instance, Facebook preserves all the information concerning the social relationships and personal life and Google can get information about shopping choices and browsing habits.

2.2. Privacy concern

Privacy is frequently confused with security. Security deals with the CIA triad composed by Confidentiality, Integrity and Availability, whereas Privacy is ensured when it is possible to hide the real identity of the person [2]. In other words, privacy aims to preserve the data from divulgation. From a legal point of view, several laws for protecting personal information were proposed, for example, in Australia (Privacy Act, 1988), the European Union (1995) and Canada (PIPEDA, 2000). In this context, the United States established separate laws and operates according to the principle of the portability and accountability for Health Information Act (HIPAA, 1996), protection of children's data (COPPA, 1998) and finance (Gramm-Leach-Bliley, 1999). Otherwise, the privacy principles deal with the fundamental rules on how organizations should treat personal information, for example, the Fair Information Practices (FTC, 2000) and OECD Privacy Principles (2010) [10]. A latest, law was proposed in European Union (EU), called General Data Protection Regulation (GDPR) which became efficient in May 2018. This law concerns all the organizations of the EU, the European Economic Area (EEA) as well as organizations belonging to other countries processing European citizen's data [1], [11].

In the case of big data, and in order to take advantage of the benefits of the services offered by web services and information platforms, individuals themselves furnish Personally Identifiable Information (PII). These personal information such as name, social security IDs and credit card numbers may be used alone or with other information to single out a person. The privacy of individuals may be compromised and no longer under control when information is made disposable [10]. Therefore, it is mandatory to ensure privacy by making sure that all the attempts to identify an individual will fail [12].

The main purpose of this paper is presenting, discussing and evaluating the majority of non-cryptographic anonymization techniques ensuring privacy in big data.

3. Pseudonymization Vs Anonymization

Both pseudonymization and anonymization approaches protect the identity of a person. Although anonymization is

the most used approach, there are some cases where pseudonymization is the most suitable since it keeps certain information unencrypted which may cause some problems to the person's life if they are hidden.

3.1. Pseudonymization

Pseudonymization is a method for ensuring privacy even if it doesn't represent a fully anonymization and it can be considered as a data minimization measure [1]. Pseudonymization replaces an identifier with a randomly generated identifier called pseudonym and it generates several identification keys in order to create a connection between distinct information related to individuals [13]. In the following, the most used pseudonymization techniques as mentioned in [14].

- **Encryption with secret key:** Since the data set still contains personal data, the owner of the key can trivially identify each data subject throughout decrypting the same data set. However, Public key encryption is not able to preserve privacy of sensitive data belonging to individuals [11].
- **Hash function:** It's a one-way function that returns an output data with fixed size from an unfixed input size. However, this function has the possibility to replace the range of known input values in order to deduce the right value for a special record. Hash functions are generally designed to be relatively fast in the calculation processes, although they fail against brute force attacks.
- **Keyed-hash function with stored key:** It is a special hash function using a secret key as an additional data entry. A data controller can apply the function on the attribute by employing this secret key. However, when an adversary applies the function without knowing the key, then this technique becomes much harder and impractical since the number of possibilities to be analyzed is large.
- **Deterministic encryption:** It may be assimilated to a technique that chooses a pseudonym random number for every attribute in the data set. This pseudonym is then removed from the matching table in order to reduce the risk of linkability. This technique will be computationally difficult for an adversary to decrypt the function, because he or she has to test every possible key whenever the tested one is not correct.
- **Tokenization:** It is specially employed in financial industry to substitute the numbers of an ID card by other values which reduce the utility for an adversary. This technique is based on applying unidirectional encryption mechanisms throughout an indexed function of random produced numbers which are mathematically not determined from the source data.

Generally, pseudonymization does not have a negative effect on the data mining process [1]. However, the reversibility of pseudonymized data could be very significant. For example, in the context of clinical drug trials, it is important that patients' pseudonymized trial data could be reversed if necessary in order to inform the patients about a medically undesirable event. That is why; fully anonymized data in this context might be dangerous and irresponsible. Therefore, in most of the time, a fully anonymization is needed since the objective of this approach is privacy protection. In the following, we will present the anonymization process and will mention the required tasks to get an anonymizer system.

3.2. Anonymization

In order to protect sensitivity or confidentiality of shared data, data sanitization is often made before the distribution and analysis processes. And, when the intention of sanitization is privacy, then it is often called data anonymization, or de-identification. Anonymization is considered as a utility-based approach ensuring privacy in big data. It maintains the identity of records existing in the published data set protected against identity disclosure attacks by applying some anonymization techniques such as generalization, suppression, etc [15]. In fact, the anonymization techniques goal is providing a balance between preserving data utility and ensuring privacy [16]. There are many open source tools made for anonymization such as ARX (Powerful Data Anonymization), the cornell anonymization and hadoop anonymization toolkit [17].

Anonymization is a process used to prevent a person's identity from being connected with other information. Depersonalization, masking or even obfuscation are other forms of anonymization. Furthermore, the anonymization process is considered effective if it is impossible to deduce original data from the anonymized data set by using a mathematical process [18].

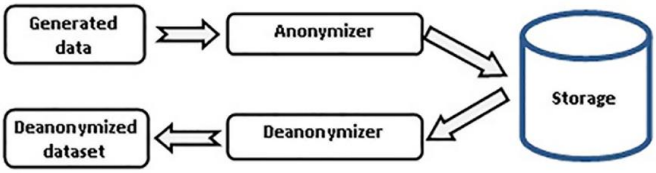

Figure 1. General anonymization architecture

While safeguarding privacy, the anonymization process is expected to enable big data tools to analyze the data in meaningful ways. So, as shown in Figure 1, once the data is anonymized, it can be safely moved to Hadoop File System (HDFS) based storage for example, where it would be disposable for analysts to examine the output. The architecture gives also the possibility to identify and reconstruct the original data set in order to control the unaccustomed behaves of some persons [17]. According to Križan et al. in [18], an anonymizer system is expected to obey a number of requirements, mainly security and speed.

- **Security** should be very strong to make a backup of the original information, fairly difficult or even impossible. Security is not strong enough even if the used keys are adequately protected, thus, anonymization techniques should be involved. Otherwise, the possibility to apply the masking only on authenticated users would improve the security of the system.
- **Speed** is measured by the number of hidden records in a unit of time. The speed of the used technique gives the ability to determine if the system is able to work online or offline only. Thus, any process should be done within a reasonable time in all cases.

In the following, we will present the proposed classification of non-cryptographic anonymization techniques.

4. Classification of non-cryptographic anonymization techniques

Before presenting the proposed classification, it is necessary to give an insight on the classifications made in some works dealing with anonymization as an approach to ensure privacy. For Li and Zhang in [19], the anonymization could

be achieved through cryptographic or non-cryptographic techniques. According to Križan et al. in [18] and Working party in [14], anonymization techniques are divided into two categories; the first one is called randomization while the second is called generalization. Otherwise, encryption, randomization, bucketization and k-anonymity constitute the classification adopted by Patil and Ingale in [20]. Besides, authors in [2] assume that the classification could be realized by treating cryptographic techniques, k-anonymity, l-diversity, t-closeness and Differential Privacy techniques. Venifa Mini and Angel Viji in [11] propose an architecture that ensures privacy against potential security breaches in the cloud through the anonymization of encrypted data. Furthermore, the classification of anonymization techniques made by Wang and Li in [21] includes generalization, suppression, bucketization and perturbation. In this paper, we will divide the anonymization techniques into two main classes, generalization and randomization approaches as shown in Figure 2.

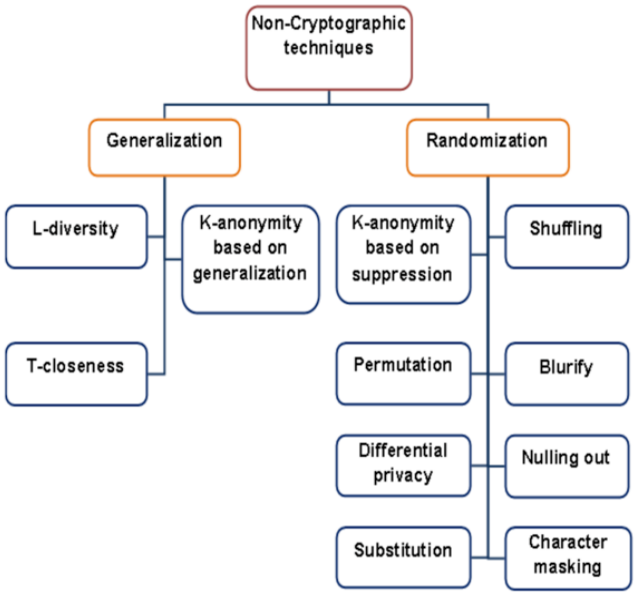

Figure 2. Diagram of the proposed classification of non-cryptographic anonymization techniques

Figure 2 represents a diagram of non-cryptographic anonymization techniques helping to ensure privacy in big data. This paper will give much information concerning all the techniques mentioned in the diagram. Table 1 presents non-cryptographic techniques already mentioned in Figure 2 to evaluate the anonymized published output through integrity, confidentiality and credibility criteria.

As shown in the Table 1, the integrity is ensured when a copy of the original data set is saved before the anonymization process or when the whole original data still exist in the published anonymized data set. According to the same table, l-diversity, t-closeness, permutation, differential privacy and shuffling are the only techniques where the integrity criterion is intact. In fact, l-diversity and t-closeness techniques leave the data as close as possible to its original form. Moreover, the distribution of values using permutation technique remains unmodified. In addition, a copy of the original data set will be maintained when using differential privacy; besides, when employing shuffling technique, the data still exists in the anonymized data set. Regarding the confidentiality criterion, it is ensured when the published anonymized data set doesn't contain real information that could lead to identify a specific person. In fact, the confidentiality is guaranteed when using k-anonymity and l-diversity only if the thresholds k and l are high enough respectively.

Table 1. The evaluation of non-cryptographic anonymization techniques

Approaches	Techniques	Integrity	Confidentiality	Credibility
Generalization	K-anonymity based on generalization [22], [23], [24], [25]	No	No	Yes
	L-diversity [2], [14], [23], [25]	Yes	No	Yes
	T-closeness [2], [14], [26]	Yes	Yes	Yes
Randomization	K-anonymity based on suppression [20], [24], [27]	No	No	Yes
	Permutation [14], [28]	Yes	Yes	No
	Differential Privacy [29], [14] [30], [31], [32]	Yes	Yes	Yes
	Substitution [18], [33], [34], [35]	No	Yes	No
	Shuffling [18], [33], [35]	Yes	No	No
	Blurify [18], [33], [35]	No	Yes	Yes
	Nulling out [18], [33], [35]	No	Yes	No
	Character masking [18], [33]	No	Yes	No

However, this criterion could not be maintained when using shuffling technique if the used algorithm in the anonymization process isn't appropriate. Concerning the credibility criterion, a published anonymized data is credible when it truly represents what it is supposed to represent. Actually, the credibility is not ensured when using permutation, substitution, shuffling, Nulling out and character masking techniques. In permutation technique, the values in the data set are substituted by other ones. In substitution, the data is replaced by falsified hidden information. Otherwise, in shuffling technique, the values are randomly rearranged inside one data set column. Besides, a column of the data set is removed through a substitution using NULL values when using Nulling out technique and the values in Character masking technique are substituted by a special constant character. In the following two sections, we will discuss every technique mentioned in Table 1. In addition, k-anonymity, l-diversity and t-closeness techniques will be tested through an extract of a real huge data set.

5. Generalization Techniques

The generalization is the first family of non-cryptographic anonymization techniques and it consists of making the attributes of data subjects more widespread by changing their scale or order of size [12].The generalization could help in preventing singling out, but it is not helpful in all cases. In fact, the generalization needs specific and sophisticated quantitative techniques for preventing linkability and inference attacks [14]. Besides, Generalization is considered as the most common approach to anonymize a data set in order to ensure privacy in big data [36]. Moreover, values in generalization algorithms are replaced with more general ones based on Value Graph Hierarchy, either on Taxonomy tree as mentioned in [22] and [36].

In this section, we present the generalization techniques which are k-anonymity based on generalization, l-diversity and t-closeness.

5.1. K-anonymity based on Generalization

Anonymization changes the format of the original data in order to protect personal or private information. In a broad sense; there exist two main attribute types including Quasi-Identifier (QI) and sensitive attributes. QI attributes may lead to identify an individual in a data set, but only when they are linked with other attributes in external data sets.

Table 2. The original test table

Id	Gender	Age	Zip Code	Disease	Treatment	Date of diagnosis
1	M	52	67025	Pulmonary emphysema	Home oxygen therapy	08/02/2003
2	F	37	75983	Asthma	Inhaled steroid therapy	13/08/2006
3	M	43	67012	Pulmonary emphysema	Smoking cessation therapy	24/02/2003
4	F	40	69300	Chronic obstructive bronchitis	Chronic obstructive pulmonary disease clinical management plan	06/09/2005
5	F	39	75918	Non-small cell carcinoma of lung TNM stage 4	Chronic pain control management	04/09/2013
6	M	70	57011	Primary small cell malignant neoplasm of lung TNM stage 4	Cancer education	27/09/2016
7	M	67	67069	Alzheimer	Demential management	30/08/2012
8	M	66	57200	Concussion injury of brain	Recommend ation to rest	14/10/2008
9	F	15	75900	Concussion injury of brain	Head injury rehabilitatio n	14/10/2008
10	M	68	57140	Stroke	Stress management	17/02/2016
11	F	65	69470	Non-small cell carcinoma of lung TNM stage 4	Terminal care	31/12/2009
12	F	46	69200	Chronic obstructive bronchitis	Smoking cessation therapy	20/10/2014

Whereas, sensitive attributes include confidential information belonging to a specific individual, these attributes need more protection compared to QI ones [23], [37]. The k-anonymity is achieved when all the records belonging to a set of QI attributes cannot be distinguished from at least k−1 other records in the data set [23], [25]. Moreover, every record in a k-anonymized data set has a maximum probability 1/k of being identified [25]. In addition, the confidentiality of the published data is better ensured when the value of the threshold k is high enough [17], [24].

The k-anonymity based on generalization protocol works as follows; in the first step, it separates QI attributes from sensitive ones. After that, it makes sure that QI attributes are generalized according to the threshold k. Later, it verifies if the generalized QI are indistinguishable from at least k-1 other records, then, it inserts them into the anonymized resulted table, otherwise the procedure is repeated [24].

Table 2 represents the original test table including the QI and sensitive attributes. This test table is an extract from a huge real data set called “Careplans” with respect to “Disease”, “Treatment” and “Date of diagnosis” attributes [38]. However, “Gender”, “Age”, and “Zip code” attributes were chosen randomly in order to show the importance of using k-anonymity and t-closeness techniques.

Table 3 shows the result of applying k-anonymity based on generalization through an algorithm called “A new technique ensuring privacy in big data: k-anonymity without prior value of the threshold k”, proposed in [24]. The algorithm generalizes the attribute “Age” before applying the principle of k-anonymity with respect to both “Gender” and “Age” attributes.

Table 3. K-anonymity based on generalization with k=3

Id	Gender	Age	Zip Code	Bucket
1	M	[41,70]	67025	1
3	M	[41,70]	67012	1
7	M	[41,70]	67069	1
2	F	[11,40]	75983	2
5	F	[11,40]	75918	2
9	F	[11,40]	75900	2
4	F	[41,70]	69300	3
11	F	[41,70]	69470	3
12	F	[41,70]	69200	3
6	M	[61,70]	57011	4
8	M	[61,70]	57200	4
10	M	[61,70]	57140	4

The main idea of k-anonymity based on generalization technique is ensuring that there are identical values within each bucket when making a horizontal partitioning. By applying k-anonymity principle, an adversary is not able to detect the real values corresponding to a certain individual. As shown in Table 3, a new column representing the resulting buckets is added. In this case, there are at least 3 records within each bucket with respect to “Gender” and “Age” attributes. For instance, the combination {M, [41-70]} is repeated 3 times in bucket 1. Besides, the combination {F, [11-40]} is repeated 3 times in bucket 3. It is clearly shown that there are at least 3 identical values with respect to “Gender” and “Age” attributes within each bucket of Table 3. Therefore, this table is considered as a 3-anonymity table. Finally, the “Id” column must be removed from the published anonymized data set in order to hide the real order of tuples.

In practice, optimal k-anonymity is a Non-deterministic Polynomial-time (NP) hard problem, thus, different approaches come to address the k-anonymity limitation like l-diversity and t-closeness models [18]. Next, the second non-cryptographic anonymization technique called l-diversity will be presented.

5.2. L-diversity

The model of l-diversity is introduced to address the shortcomings of k-anonymity. L-diversity is a form of group-based anonymization and it aims to ensure privacy by partitioning the data sets into several buckets. Thus, the huge scale of big data is minimized in terms of representation [23]. This technique ensures that each sensitive attribute has at least l different values within each bucket [12], [15]. L-diversity technique is achieved when it is able to resist against background knowledge attack [25], [39]; besides, l-diversity can ensure that sensitive attributes would have actually the same frequency [40]. In addition, it is impossible to implement the inference attacks against an 'l-diverse' data set with certitude of 100% [14].

In the literature, there exist three models of l-diversity which are Distinct, Entropy and Recursive models [24], [26]. However, the distinct l-diversity technique is the most used where each bucket in the data set contains only distinct values.

Table 4. 3-diversity applied on sensitive attributes

Id	Age	Disease	Treatment	Date of diagnosis	Bucket
1	[36-55]	Pulmonary emphysema	Home oxygen therapy	08/02/2003	1
2	[36-55]	Asthma	Inhaled steroid therapy	13/08/2006	1
4	[36-55]	Chronic obstructive bronchitis	Chronic obstructive pulmonary disease clinical management plan	06/09/2005	1
3	[36-65]	Pulmonary emphysema	Smoking cessation therapy	24/02/2003	2
5	[36-65]	Non-small cell carcinoma of lung TNM stage 4	Chronic pain control management	04/09/2013	2
6	[36-65]	Primary small cell malignant neoplasm of lung TNM stage 4	Cancer education	27/09/2016	2
7	[66-75]	Alzheimer	Demential management	30/08/2012	3
8	[66-75]	Concussion injury of brain	Recommendation to rest	14/10/2008	3
10	[66-75]	Stroke	Stress management	17/02/2016	3
9	[15-65]	Concussion injury of brain	Head injury rehabilitation	14/10/2008	4
11	[15-65]	Non-small cell carcinoma of lung TNM stage 4	Terminal care	31/12/2009	4
12	[15-65]	Chronic obstructive bronchitis	Smoking cessation therapy	20/10/2014	4

We will illustrate this technique through an example with l=3 as shown in Table 4 by doing the same analysis as shown in the case of k-anonymity based on generalization. However, l-diversity principle ensures that each bucket will include distinct values with respect to the chosen sensitive attributes. The technique is achieved by making sure of having at least 3 distinct values within each bucket with respect to the combination of attributes "Disease", "Treatment" and "Date of diagnosis".

In Table 4, the l-diversity technique applies a horizontal partitioning on the original test Table 2 with respect to "Disease", "Treatment" and "Date of diagnosis" sensitive attributes. For instance, all the 4 buckets include at least 3 distinct values, thus, Table 4 would be considered as a 3-diverse table. The published anonymized data set will not include the "Id" column to not give the adversary the possibility of retrieving the original data set. Although, the l-diversity technique gives good results in terms of data anonymization, it cannot resist against skewness and similarity attacks [2].

The following t-closeness technique comes to address the l-diversity technique limitation.

5.3. T-closeness

T-closeness is a refinement of l-diversity and aims to create equivalent classes, also called buckets, which look like the initial distribution of attributes in the original table. This technique is efficient when it is necessary to remain the data as close as possible to their original form [14].

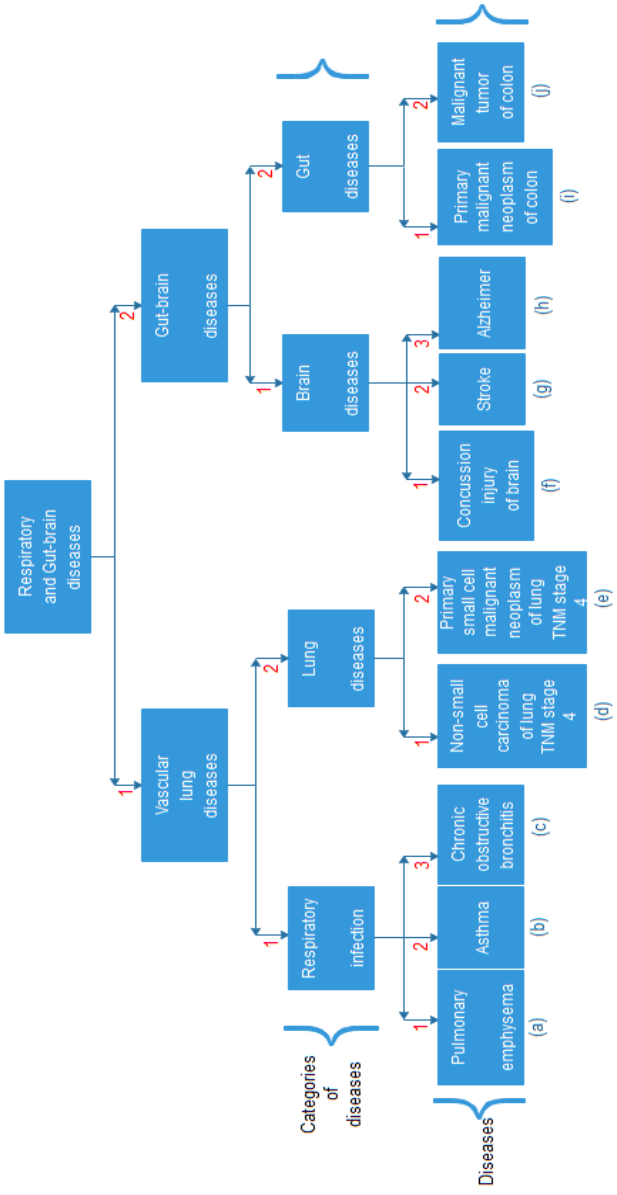

Figure 3. The hierarchy of diseases

When the distance between the distribution of a sensitive attribute in the equivalence class and the distribution of the attribute in the whole table is less than a threshold t, then the equivalence class is called "have t-closeness" [2], [26].

As said before, l-diversity technique is not able to resist against several attacks. Therefore, the most critical one is called similarity attack. Despite the fact that the sensitive values are distinct within each bucket after applying l-diversity technique, the semantic significance of these distinct values is similar and thus, the information can be disclosed [2]. Figure 3 shows the hierarchy of diseases related to health sector including categorical values. This hierarchy is adopted in order to show that l-diversity technique cannot resist against similarity attack and therefore; thus, t-closeness technique is needed to address l-diversity limitation.

Table 5 shows the result after applying t-closeness technique on Table 4 corresponding to l-diversity technique.

Table 5. T-closeness technique applied on Table 4

Id	Age	Disease	Treatment	Date of diagnosis	Bucket
8	[66-75]	Concussion injury of brain	Recommendation to rest	14/10/2008	1
2	[36-55]	Asthma	Inhaled steroid therapy	13/08/2006	1
4	[36-55]	Chronic obstructive bronchitis	Chronic obstructive pulmonary disease clinical management plan	06/09/2005	1
3	[36-65]	Pulmonary emphysema	Smoking cessation therapy	24/02/2003	2
5	[36-65]	Non-small cell carcinoma of lung TNM stage 4	Chronic pain control management	04/09/2013	2
6	[36-65]	Primary small cell malignant neoplasm of lung TNM stage 4	Cancer education	27/09/2016	2
7	[66-75]	Alzheimer	Demential management	30/08/2012	3
1	[36-55]	Pulmonary emphysema	Home oxygen therapy	08/02/2003	3
10	[66-75]	Stroke	Stress management	17/02/2016	3
9	[15-65]	Concussion injury of brain	Head injury rehabilitation	14/10/2008	4
11	[15-65]	Non-small cell carcinoma of lung TNM stage 4	Terminal care	31/12/2009	4
12	[15-65]	Chronic obstructive bronchitis	Smoking cessation therapy	20/10/2014	4

Although the buckets in Table 4 contain distinct values, they may correspond to a specific category. Thus, an adversary would easily know the category of diseases belonging to a specific bucket. For instance and based on the hierarchy in Figure 3, bucket 3 of Table 4 includes "Alzheimer", "Concussion injury of brain" and "Stroke" which all of them correspond to "Brain diseases" category.

Therefore, if an adversary knows that a 68 years old person exists in Table 4, he/she would easily deduce that this person belongs to bucket 3 and consequently, this person certainly suffers from a brain disease. Thus, t-closeness comes to resist against similarity attack. Based on the algorithm proposed in [28], buckets 2 and 4 of Table 4 will remain intact since they include values belonging to different categories. However, buckets 1 and 3 of Table 4 will be changed to address the problem of similarity attack since they are corresponding to "Respiratory infection" and "Brain diseases" categories respectively. Besides, a permutation process is used to diversify the values within buckets 1 and 3 of Table 4 semantically. As a result, all the buckets belonging to Table 5 correspond to more than one category and the similarity attack is no more a threat. In the published anonymized data set, the "Id" attribute will be removed in order to hide the real order of tuples.

6. Randomization Techniques

Randomization is the second family of non-cryptographic anonymization techniques. It alters the veracity of the data in order to remove the strong relationship between the data and the individual. If the adversary has enough confusion concerning the data, then, he can no longer identify an individual [14]. The randomization has the advantage of providing protection against inference attacks. Randomization techniques can be applied when collecting the data and also during the data pre-processing steps [41]. However, it will not reduce the singularity of records itself since each record will always be derived from a single data subject. The randomization approach could be combined with the generalization approach in order to produce stronger privacy guarantees [14].

In the next section, we will list some techniques related to the randomization approach.

6.1. K-anonymity based on Suppression

The main idea of suppression based algorithm is hiding the values of some attributes by using an asterisk "*" while the concept of k-anonymity is ensured with respect to the chosen attributes [24], [27].

The k-anonymity based on suppression protocol is described as follows; first, it separates QI attributes from sensitive ones; after that, it substitutes some QI attributes by the special character "*". Later, it checks if the suppressed QI attributes are equal to the non-suppressed ones. In the end, they are inserted in the table, otherwise the procedure is repeated. One advantage of using the k-anonymity based on suppression is the impact of substitution of the actual value with "*" which makes unauthorized users confused. However, it becomes impossible to make a backup of the original data set [20]. Besides, it causes a huge amount of information loss and therefore, the data utility is not preserved [42]. Table 6 illustrates an example of applying k-anonymity based on suppression with k=3 on all the QI attributes already existing in Table 2. The k-anonymity based on suppression technique is similar to k-anonymity based on generalization. The difference is that the first one modifies the values within a column by substituting for example, some

digits with asterisk in "Zip code" attribute as shown in Table 6.

Table 6. K-anonymity based on suppression with k=3

Id	Gender	Age	Zip Code	Bucket
1	M	52	67***	1
3	M	43	67***	1
7	M	67	67***	1
2	F	37	75***	2
5	F	39	75***	2
9	F	15	75***	2
4	F	40	69***	3
11	F	65	69***	3
12	F	46	69***	3
6	M	62	57***	4
8	M	66	57***	4
10	M	68	57***	4

And the second one focuses on dispersing the range of values by making them more general. However, the principle of k-anonymity is the same as done before when treating the generalization case as mentioned in Table 3. Table 6 represents the result after applying k-anonymity based on suppression where all the buckets include similar values with respect to "Gender" and "Zip code" attributes. In the published data set, the "Id" column is deleted to avoid retrieving the real order of tuples.

6.2. Permutation

Permutation can be considered as a particular way of adding noise to data. However, the generation of an exact amount of noise could be a challenging task. Also, changing the values of attributes may slightly not provide enough privacy. Alternatively, permutation technique modifies only the values in the data set by substituting a record with another one. Permutation technique could be applied during the anonymization process between minimum values corresponding to each consecutive buckets existing in the data set [28]. Such permutation between records will ensure that the range and the distribution of values will remain unchanged, but the link between values and individuals will not. If two or more attributes have a logical or statistical relationship and are swapped independently, then, such correlation will be destroyed. Therefore, it might be important to permute a set of linked attributes while breaking the logical correlation; otherwise, an attacker may identify the permuted attributes and reverse the permutation [14]. In addition, it is essential to isolate sensitive attributes from the original data set and then, apply the permutation process on the corresponding sensitive values in order to take benefit from protecting personal data and to prevent the adversary from retrieving valuable information.

Permutation itself is not able to ensure privacy in big data. However, it must be combined with other generalized anonymization techniques such as t-closeness [14], [28].

6.3. Differential Privacy

Differential privacy is an anonymization technique which is very suitable for big data as it does not allow the degradation of system's speed. In addition, it is very hard for an attacker to deduce the presence or absence of an individual. Furthermore, when two different data sets produce almost the same output, then the opponent is unable to determine the real targeted data set [29], [30]. When the amount of data becomes important, the differential privacy becomes less efficient since the original data could be estimated from the perturbed data [30]. Figure 4 illustrates the required process in order to achieve differential privacy.

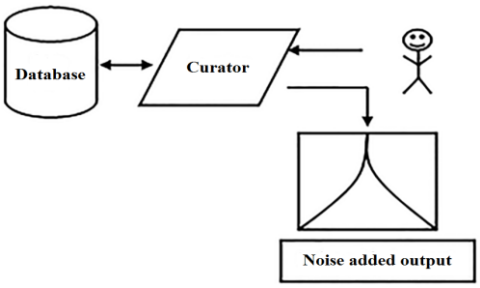

Figure 4. Achieving differential privacy [30]

As shown in Figure 4, differential privacy is achieved by making a curator between the database and the user/analyst. Once the user or analyst makes a request, it is received by the curator that accesses the impact of privacy through calculating the sensitivity of information, after that, the curator sends the request to the database and waits to receive the clean response [30].

One of the strengths of using differential privacy is highlighted when the data sets are delivered to authorized third parties in order to reply to a particular request instead of releasing a single data set. Therefore, Differential Privacy technique ensures privacy by adding noise to the output of a given function, and consequently an adversary cannot deduce if a specific record is involved in the data set [31]. However, Differential Privacy technique has some weaknesses, for instance, when making numerous requests; an attacker could be able to identify a particular individual through two or more answers [14]. Besides, the technique is not efficient for privacy preserving when processing a data set including highly correlated attributes [32]. The ability to generate the right quantity of noise to be added to the output is considered as a challenging issue when using Differential Privacy technique [14].

6.4. Substitution

Substitution is an anonymization technique which consists of substituting the values into a data set in a random way or even through a list of data similar to the original data set values [34], [35]. The substituted values can be selected either from a given pseudonymization list containing falsified values [18]. Substitution is highly adequate when the anonymization intends to preserve the appearance and the feel of current data [35]. However, preparing a considerable amount of substitutable information to be accessible for every substitution is a challenging task. For instance, to sanitize names, a fairly extensive random list of names must be prepared; and to sanitize the phone numbers, a huge list of fake phone numbers is needed; nevertheless, the capacity to produce an invalid data is very difficult [33].

The integrity in substitution technique seems to be preserved; but the original table remains inaccessible. However, an efficient substitution requires a list whose size is equal to or larger than the size of data requiring substitution. So, if the data set contains a huge amount of data without having enough substitutable data, then, the substitution technique will not be the best technique for anonymization.

6.5. Shuffling or Data Swapping

Shuffling or data swapping technique is similar to substitution technique but the anonymized data is derived from the column itself. It randomly rearranges values inside one data set's column while maintaining the order in the other columns [18]. Shuffling technique is useful when it is essential to keep the aggregated values in their original form. Moreover, it could process columns with a single constraint [35]. The data migrates between lines until there is no possible correlation in the data available in the data set. However, there is a risk when using the shuffling technique since the source data still exists. So, an adversary with some significant information can deduce the original data. Another problem is the selection of the algorithm used to shuffle the data; at that time, the data may be simply unshuffled if the adversary could deduce the shuffling algorithm. For instance, if the shuffling algorithm works by swapping the data existing in a column between every two lines, then, the interested party would not make a big effort to get a backup of the original data set. It is true that shuffling is quick; however, a high attention should be paid when using a modern and advanced algorithm to randomize the lines in the data set [33]. In fact, it is more secure to apply shuffling on huge data set because tracing the original values is harder [18]. Although shuffling technique preserves the data integrity, it may be insufficient especially when the amount of records in the data set is tiny.

6.6. Blurify

Blurify technique gives the opportunity to dissemble the data in a reasonable way. It involves modifying each value in a column by a particular variance which represents a random percentage of the original value [18]. Blurify technique considerably changes the data in order to make it untraceable by any adversary. For instance, a salary details column could have a random variance of ±10%. Certain values might be higher; some of them lower, however all the values would not be too far away from their original range [33]. Blurify technique is also called "The Number and Variance" technique in some literature researches; besides, it is generally useful on numeric or date data [35]. For example, financial data like salaries are increased or decreased randomly for a particular variance percentage [18], and birth dates data could be used through an arbitrary range of ±120 days. Actually, this range conceals the personally identifiable information, whereas the distribution is still preserved [33].

6.7. Nulling out

Nulling out is an anonymization technique that deletes the sensitive data contained in the data set by removing the whole corresponding column and replacing it with NULL values [35]. Nulling out technique cannot usually be employed on non nullable columns of the data set [18]. In general, the test teams require a non nullable data for their processing. Although, this technique is simple, it is not much desirable and it may not be appropriate if an assessment has to be conducted on the data [33], [35]. For instance, it would be impossible to query accounts of customers if vital information like names and other customer details are null values. Nulling out could also be called as truncating data technique, and it is helpful in some situations where the data is not very important [33].

6.8. Character masking

Character masking technique is similar to Nulling out technique; it changes the initial value with a special constant character [18]; and it changes certain fields by using a mask character. This technique strongly hides the contents of the data while maintaining the same format and reports [33]. For example, a credit card number could be viewed such as: 4346 6454 0020 5379 and after applying the masking, the information would appear like: 4346 XXXX XXXX 5379.

The Character masking technique efficiently eliminates a great part of the sensitive content of the record while conserving the appearance and feel. Thus, much care has to be provided to make sure that a sufficient amount of data is masked in order to insure privacy. An operation of masking like: XXXX XXXXXXXX 5379 would remove much information about the credit card number; thus, character masking technique would be a strong and rapid technique when dealing with a data in a particular and unchanging format [33]. If many appropriate cases should be treated, then character masking could be slow, very difficult to manage and could possibly leave some data without being masked.

7. Solutions and Recommendations

In order to ensure privacy in big data, many anonymization techniques were presented in this paper including generalization and randomization ones. However, choosing an anonymization technique instead of another one is a challenging issue. In the following, numerous cases are cited:

- If the used data set contains only QI attributes; then, the most adequate technique is k-anonymity based on generalization or suppression or even both of them. Besides the more the threshold k of k-anonymity is high the more the technique is powerful.
- If the data set includes only sensitive attributes; then, l-diversity technique is suggested among other anonymization techniques. However, since l-diversity technique cannot resist against similarity attack, t-closeness technique must be involved in the anonymization process.
- When the handled data is qualitative; the use of generalization techniques is advisable since they do not remove the data from the original data set.
- Since the encryption increases the size of data, the system's speed is degraded; thus, the employment of non-cryptographic techniques is appropriate when the speed of anonymization is a priority for the user.
- When the manipulated data is quantitative; it would be recommended to use a randomization technique since this type of data involves removing or aggregating variables.
- Anonymization techniques such as Nulling out and Character masking may be favored when the data set contains a trivial or powerless content.
- Differential privacy would be the most suggested technique when the data set contains secret information which needs to be disrupted by adding a particular noise to the data.
- When it is necessary to save a copy of the original data set; then the most suitable techniques to use are t-closeness, permutation, differential privacy and shuffling.

- If the data set includes both QI and sensitive attributes; then, the combination of k-anonymity, l-diversity and t-closeness techniques would be useful and will make a balance between privacy and data utility.

8. Conclusions

In this paper, we made a classification of different anonymization techniques ensuring privacy in big data. All the techniques belong to non-cryptographic category which is divided into two main approaches including generalization and randomization. The generalization is considered as the most appropriate approach to ensure privacy in big data. Three main anonymization techniques belonging to this approach (k-anonymity, l-diversity and t-closeness) were discussed and evaluated since they treat both Quasi-Identifier and sensitive attributes. Besides, the combination of these three techniques makes a balance between ensuring privacy and preserving data utility. According to the proposed classification, it seems clear that generalization approach is preferred when dealing with qualitative data, while randomization is generally a better choice when processing quantitative one. As a future work, we plan to elaborate a hybrid anonymization technique, which consist of applying l-diversity technique on highly correlated attributes only because it will preserve the data utility gained from the strong relationship between attributes. Besides, l-diversity technique will ensure data privacy. In addition to l-diversity, the utilization of t-closeness principle will be interesting since it can resist against similarity attack especially when treating categorical sensitive attributes.

References

[1] N. Gruschka, V. Mavroeidis, K. Vishi, M. Jensen, "Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR," IEEE Proceeding of the International Conference on Big Data, USA, pp. 5027-5033, 2018.

[2] S. Sangeetha, G. S. Sadasivam, "Privacy of Big Data: A Review," Springer International Publishing A. Dehghantanha, K. K. R. Choo (eds.), Book: Handbook of Big Data and IoT Security, India, pp. 5-23, 2019.

[3] T. L. Nguyen, "A Framework for Five Big V's of Big Data and Organizational Culture in Firms," IEEE Proceeding of the International Conference on Big Data, USA, pp. 5411-5413, 2018.

[4] K. Khurshid, A. A. Khan, H. Siddiqi, I. Rashid, "Big Data-9Vs, Challenges and Solutions," Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan, Vol. 23, No. 3, pp. 28-34, 2018.

[5] F. Soleimani-Roozbahani, A. Rajabzadeh Ghatari, R. Radfar, "Knowledge discovery from a more than a decade studies on healthcare Big Data systems: a scientometrics study," Springer Journal of big data, Vol. 6, No. 8, 2019.

[6] K. Mungai, A. Bayat, "The Impact of Big Data on the South African Banking Industry," Proceeding of the 15th International Conference on Intellectual Capital, Knowledge Management and Organizational Learning, South Africa, pp. 225-236, 2018.

[7] A. Oussous, F. Benjelloun, A. AitLahcen, S. Belfkih, "Big Data technologies: A survey" Journal of King Saud University–Computer and Information Sciences, Vol. 30, pp. 431-448, 2018.

[8] Y. Gahi, M. Guennoun, T. Mouftah, "Big Data Analytics: Security and Privacy Challenges," IEEE Proceeding of the Symposium on Computers and Communication (ISCC), Italy, pp. 952-957, 2016.

[9] Q. Qi, F. Tao, "Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison," IEEE Access, Vol. 6, pp. 3585-3593, 2018.

[10] T. Basso, R. Matsunaga, R. Moraes, N. Antunes, "Challenges on Anonymity, Privacy and Big Data," IEEE Proceeding of the Seventh Latin-American Symposium on Dependable Computing (LADC), Colombia, pp.164-171, 2016.

[11] G. Venifa Mini, K. S. Angel Viji, "A Comprehensive Cloud Security Model with Enhanced Key Management, Access Control and Data Anonymization Features," International Journal of Communication Networks and Information Security (IJCNIS), Vol. 9, No. 2, pp. 263-273, August 2017.

[12] K. Abouelmehdi, A. Beni-Hessane, H. Khaloufi, "Big healthcare data: preserving security and privacy," Springer Journal of big data, Vol. 5, No. 1, 2018.

[13] L. El Haourani, A. A. Elkalam, A. A. Ouahman, "Knowledge Based Access Control a Model for Security and Privacy in the Big Data," ACM Proceeding of the 3rd International Conference on Smart City Applications (SCA), Morocco, pp. 1-8, 2018.

[14] Working party, "The working party on the protection of individuals with regard to the processing of personal data (0829/14/EN WP216)," Brussels, Belgium: Data Protection Working Party. http://statewatch.org/news/2014/apr/eu-art-29-dp-wp-216.pdf, 2014.

[15] Y. Canbay, Y. Vural, S. Sagiroglu, "Privacy Preserving Big Data Publishing," IEEE Proceeding of The International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, Turkey, pp. 24-29, 2018.

[16] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, Z. Huo, "Privacy Preservation in Big Data from the Communication Perspective-A Survey," IEEE Communications Surveys & Tutorials, Vol. 21, No. 1, 2019.

[17] J. Sedayao, R. Bhardwaj, N. Gorade, "Making Big Data, Privacy, and Anonymization work together in the Enterprise: Experiences and Issues," IEEE Proceeding of the International Congress on Big Data (BigData Congress), Alaska, USA, pp. 601-607, 2014.

[18] T. Križan, M. Brakus, D. Vukelić, "In-Situ Anonymization of Big Data," IEEE Proceeding of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, pp. 292-298, 2015.

[19] X. Li, Z. Zhang, "Exploit The Scale Of Big Data For Data Privacy: An Efficient Scheme Based on Distance-Preserving Artificial Noise and Secret Matrix Transform," IEEE Proceeding of the China Summit & International Conference on Signal and Information Processing, Xi'an, China, pp. 500-504, 2014.

[20] M. Patil, S. Ingale, "Privacy Control Methods for Anonymous & Confidential Database Using Advance Encryption Standard," International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 2, No. 8, pp. 224-229, 2013.

[21] Li-e. Wang, X. Li, "A Hybrid Optimization Approach for Anonymizing Transactional Data," ACM Proceeding of the International Workshops and Symposiums on Algorithms and Architectures for Parallel Processing (ICA3PP), Zhangjiajie, China, pp. 120-132, 2015.

[22] S. Kavitha, S. Yamini, P. Raja Vadhana, "An Evaluation on Big Data Generalization Using k-Anonymity Algorithm on Cloud," IEEE Proceeding of the 9th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, pp. 1-5, 2015.

[23] P. Jain, M. Gyanchandani, N. Khare, "Big data privacy: a technological perspective and review," Springer Journal of Big Data, Vol. 3, No. 25, 2016.

[24] Z. El Ouazzani, H. El Bakkali, "A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k," Elsevier Proceeding of The First International

Conference on Intelligent Computing in Data Sciences, Morocco, pp. 52-59, 2018.

[25] U. P. Rao, B. B. Mehta, N. Kumar, "Scalable l-Diversity: An Extension to Scalable k-Anonymity for Privacy Preserving Big Data Publishing," IGI Global International Journal of Information Technology and Web Engineering (IJITWE), Vol. 14, No. 2, pp. 27-40, 2019.

[26] A. Rahmani, A. Amine, R. M. Hamou, "Combination of Access Control and De-Identification for Privacy Preserving in Big Data," IGI Global International Journal of Information Security and Privacy, Vol. 10, No. 1, 2016.

[27] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, S. Guo, "Protection of Big Data Privacy. Special section on theoretical foundations for big data applications: Challenges and opportunities," Vol. 4, pp. 1821-1834, 2016.

[28] Z. El Ouazzani, H. El Bakkali, "Proximity Test for Sensitive Categorical Attributes in Big Data," IEEE Proceeding of The 4th International Conference on Cloud Computing Technologies and Applications (Cloud'tech), Belgium, 2018.

[29] O. Hasan, B. Habegger, L. Brunie, N. Bennani, E. Damiani, "A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case," IEEE Proceeding of the International Congress on Big Data, CA, USA, pp. 25-30, 2013.

[30] K. M. PdShrivastva, M.A. Rizvi, S. Singh, "Big Data Privacy Based On Differential Privacy a Hope for Big Data," IEEE Proceeding of the 6th International Conference on Computational Intelligence and Communication Networks (CICN), Bhopal, India, pp. 776-781, 2014.

[31] L. Cui, Y. Qu, S. Yu, L. Gao, G. Xie, "A Trust-Grained Personalized Privacy-Preserving Scheme for Big Social Data," IEEE Proceeding of the International Conference on Communications (ICC), USA, pp. 1-6, 2018.

[32] M. Du, K. Wang, Z. Xia, Y. Zhang, "Differential Privacy Preserving of Training Model in Wireless Big Data with Edge Computing," IEEE Proceeding of The Transactions on Big Data, 2018.

[33] Data Masking, A Net 2000 Ltd. White Paper, "Data Masking: What You Need to Know, What You Really Need To Know Before You Begin", 2016.

[34] S. Arfaoui, A. Belmekki, A. Mezrioui, "Privacy Enhancement of Telecom Processes Interacting with Charging Data Records," Springer Proceeding of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR), pp. 268-277, 2018.

[35] K. Sharmila, A. C. S. Borgia, V.S. Sreeja, "A comprehensive Study of Data Masking Techniques on cloud," International Journal of Pure and Applied Mathematics, Vol. 119, No. 15, pp. 3719-3728, 2018.

[36] A. Raj, R. G. L. D'Souza, "Big Data Anonymization in Cloud using k-Anonymity Algorithm using Map Reduce Framework," International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), Vol. 5, No. 1, pp. 50-56, 2019.

[37] A. Anjum, N. Ahmad, S.U.R. Malik, S. Zubair, B. Shahzad, "An efficient approach for publishing microdata for multiple sensitive attributes," Springer The Journal of Supercomputing, Vol. 10, 2018.

[38] Careplans Real Big data set, Retrieved from https://storage.googleapis.com/synthea-public/synthea_sample_data_csv_sep2019.zip, (2019).

[39] K. Sharma, A. Jayashankar, K. SharmilaBanu, B.K. Tripathy, "Data Anonymization through Slicing Based on Graph-Based Vertical Partitioning," Springer Proceeding of the 3rd International Conference on Advanced Computing, Networking and Informatics (ICACNI), India, Vol. 44 of the series Smart Innovation, Systems and Technologies, pp. 569-576, 2015.

[40] K. Dhivakar, S. Mohana, "A Survey on Privacy Preservation Recent Approaches and Techniques," International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 2, No. 11, pp. 6559-6566, 2014.

[41] P. R. M. Rao, S. M. Krishna, A. P. S. Kumar, "Privacy preservation techniques in big data analytics: a survey," Springer Journal of Big Data, Vol. 5, No. 33, 2018.

[42] W. Zheng, Z. Wang, T. Lv, Y. Ma, C. Jia, "K-Anonymity Algorithm Based on Improved Clustering," Proceeding of the International Conference on Algorithms and Architectures for Parallel Processing, China, pp. 462-476, 2018.