# A Knn Based Multiple Forms of Attack Prevention Algorithm for Non-Numerical Big Data in Medical Domain

**Mahwish Abid[1], Muhammad Sheraz Arshad Malik[2], Muhammad Usman[3]**
**Muhammad Mashhood ul Hasan[4], Zainab Khalid[5]**

Department of Computer Science, Riphah International University Faisalabad Campus, Pakistan[1,3,4,5]
Department of Information Technology, Government College University Faisalabad, Pakistan[2]

**Abstract**
The data which is in the large quantity and complicated to a certain level that the traditional data processing tools are unsuccessful to be applied on them is known as big data. Big data offers assistance in many fields such as IT, healthcare, customer care, e-commerce and many more. But it provide major benefits in the field of healthcare. But with the advancement of technology and internet, big healthcare data privacy has become a major concern these days. Intruder can perform numerous attacks or get some sensitive information about the patient which can be misused or mishandled. Various techniques and methodologies including anonymization has been proposed for big healthcare data security. But still data suffers from various attacks. While on the other hand such techniques only works for numerical data. Therefore, to handle the background knowledge and homogeneity attack an algorithm is proposed to handle the non-numerical data in this research. Results were calculate with the help of a tool where success rates are found and data re-identification rates are seen.

***Key words***
*K-anonymization, background knowledge attack, homogeneity attack, generalization, microaggregation*

## 1. Introduction

'Big data' term is used to represent those data sets which are very large and so complicated to be handled by the softwares which are used to process the data [1]. As the data sets are growing very quickly, the reason is that the data is collected from various different and inexpensive resources like mobile devices, different networks, softwares, cameras and much more [2]. Big data is mostly categorized with the help of 3Vs which are the large volume of the data, extensive variety of the data types and velocity at which the data is being processed. But with the passage of time various other Vs are also included in the details of big data [3].

During the current times, big data and analytics are used efficiently in various disciplines for example in IT, healthcare, services for the customer care, online banking, management of risks, the astronomy and much more [4]. During the past decades, the data which is stored in health care databases have been increased a lot which makes that data a considerably big healthcare data. This large amount of the data makes sure to support an extensive collection of medical and healthcare facilities which includes clinically decision support system, health management with the help of sensors, detection of diseases and so on[5]. But by using the big data, the major problem is to protect privacy. The large quantity of the data can result in breaching the privacy of various clients [6]. According to a report, till 2017 there are 2,181 attacks are reported on healthcare data [7]. Therefore, security of big data in the field of healthcare is a very important issue during the modern times. For this purpose only removing the attributes, which are also known as identifiers, is not enough specially when quasi identifiers are also existing in the data [8].

Attributes/ identifiers are divided into four different type of categories:

- Attributes that are individually identified they are known as **explicit identifiers**. Names, addresses, social security numbers etc. are included in them [9].
- Attributes which can give information about an individual when they are joint with other attributes are known as **quasi identifiers**. Zip-codes, date of births, and gender are included in them [10].
- And the attributes are known as **sensitive attributes** if they are carrying the sensitive information at individual level[10]. They include disease, salary etc. while publishing the data, sensitive information about an individual must need to be protected [9].
- **Non-sensitive identifiers** are general attributes that do not belong to any of the groups.

So there is a need to protect these attributes so that sensitive data cannot be unveiled or miscalculated. Furthermore, there are two forms of disclosure. Identity disclosure and the Attribute disclosure. If any individual could be recognized from the released data then this one is

Manuscript received December 5, 2018
Manuscript revised December 20, 2018

known as **Identity disclosure**. The phenomenon in which the one's essential information is inferred from data which is published on some particular platform is known as **Attribute disclosure** [11].

Similarly, privacy problems can also arise when transferring the data to the third party which can lead to various privacy attacks. Generally privacy attacks can be classified into following categories:

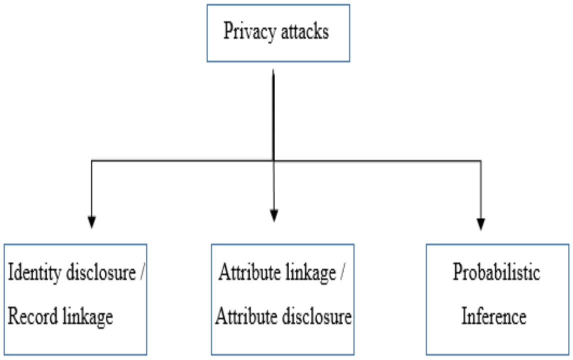

Privacy attacks
Identity disclosure / Record linkage
Attribute linkage / Attribute disclosure
Probabilistic Inference

Fig 1 classification of privacy attacks

**Record linkage** occurs when record owner is identified by linking the entities with background knowledge. It yields in identifying an individual's data or detecting an owner of that particular record. **Attribute linkage** is identifying the attributes about an owner of the record by combining released records with background knowledge. It yields in deduction of sensitive information. **Probabilistic inference** is about changing the beliefs of attacker about sensitive attributes of record owner, before and after accessing the data[12].

Hence, to propose the promising methodologies to preserve the privacy of healthcare data has gained a lot of attention during current times[9]. A number of techniques have been offered in the past which aim at generation of an anonymized form of original data so that there must not be any disclosure risk when it is published[6]. Specially, k-anonymization helps in making the records re-identification impracticable by hiding all the subjects within a group of k- subjects[13].

Anonymization is a technique which is used for preserving the privacy of an individual by making the removal of the attributes present in data which is in the publishing process, but to maintain the originality of the information is also the key part of the method [8]. As k-anonymity implementation is easy and chances to data re-identification is lesser when value of the k is high. But it is not successful for the prevention of background knowledge attack and the homogeneity attack when data is in non-numerical form and undergo from the attribute linkage and the record linkage[14].

Two clustering algorithms have [13]been established for creating the k-anonymity β- likeness model for data privacy and to prevent the background knowledge attack for numerical data[9]. A personalized model for data privacy which is based upon the anonymizing technique for data related with flight paths is proposed. For this purpose, the sensitive data attributes are being generalized while suppression is performed on the trajectory data. With the help of this technique linking attack and similarity attack can be handled and minimized to some extent[15].

As k-anonymization provide protection against identity disclosure but it does not offer adequate protection against attribute disclosure and two major attack are being identified homogeneity attack and background knowledge attack [16]. Further, three different microaggregation based algorithms were also proposed for achieving the k-anonymity t-close based equivalence classes. Clustering based algorithms were also used for this purpose which results in the less loss of information. But they works only for numerical data. So, still k-anonymization suffers from attacks for example homogeneity attack, similarity attack, background knowledge attack in non-numerical form of the data[17].

In this research some of the basic concepts and benefits of big data are being mentioned, also various different attacks are also explained and also there results are being described. This research also contributes to know about the achievements and limitations of k-anonymization. And to handle the problem and algorithm is proposed and results are being calculated with the help of a tool.

## 2. Literature Review

With the advancement of the interne and technologies used to process the data have enhanced the speed of collecting and distributing the data [8]. So the healthcare records are saved in the electronic form in the information systems. Furthermore, these systems also provide help to share the data with others to make the improvements in efficiency and quality of healthcare [18]. But these systems containing the healthcare data also have data which could be privacy sensitive because of having the personal information and medical information of an individual, for example Electronic Health Records (EHR).So the unwanted accessibility to the private data can be resulted into the usage pf the data for other purposes rather than healthcare essential processes[19].

Healthcare data are in large volume and heterogeneous and the main reason is that the data is being collected from various different internal and the external sources which are available at different localities in various legacy and applications. Moreover, data can be in different forms for example, files, .csv, table in relational databases, ASCII/texts and so on. Some of the common examples of the sources which could be internal or external are mentioned below:

**External source:** They could be web or social media data which is the data from various sites like Facebook, Twitter, LinkedIn, blogs and much more. They also could be machine to machine data which contains readings from remotic sensors, meters and the other vital sign devices.

**Internal sources:** they could be biometric data which includes finger prints, genetics, handwritings, scan of retinas, X-rays, medical images, blood pressure and many more. They also includes Human- generated data which is the medical data being collected from various electronically medical records, notes taken by the physicians, interviews done with the patients and so on. They could be unstructured or semi-structured or could be both [4].

As the health records are being digitized and interconnected, there are the various significant advantages have been accomplished today. Some of the major benefits include the quality management, reducing the load of the work, to save the consulting time, detection of various diseases in their earlier times and efficiently reducing the cost, detection of the healthcare records and much more. Healthcare information also provide help to the patients to make the correct decisions on the right time. More over analysis on health records can also be applied for identification of the individuals to provide the proactive care or to make changes in their living style to lower the risk of health degradation. Which results in the improvement of the health but also reduced the cost of care[20].

Patient records which are collected from different resources also assist in research and development for making the improvement in the research quality. In actual R&D can also propose the new algorithms to the detection of the new diseases [21]. Moreover, healthcare providers can also propose the new techniques for takin the care of the patients through which excessive hospitalization can be reduced. After analyzation of the patterns of the diseases, tracking the occurrences and the transmitions make sure the improvement in health observation of public and the speed responses[22].

Health care strategies are also capable of predicting the viral diseases in their earlier times before they are spread all over. As it could not be possible by making the analyzation of the patients' social logs which suffers from the disease by living in a specific geo location[23]. Misusing the medical information of a person to get the wrongful health benefits or the funds can also be identified by analyzing the healthcare records. For this purpose various predictive models can use used [24].With all these benefits, still the healthcare data are often of a sensitive nature, which means the data to be protected appropriately. Many attack patterns can be applied to published healthcare data and can reveal either the identity of an individual or a sensitive attribute of a person by combining different data sets.

**Preliminaries:** Some of main attacks on data privacy are following:

**Linking attack:** It takes place when published data is combined with background knowledge by an attacker for identification of record owner which yields in record linkage or identity disclosure.

**Collusion attack:** Any attacker even without having the background knowledge can perform collusion attack just by combining different k-anonymous versions of identical data for identification of record owner. This attack yields in identity disclosure or record linkage [25].

**Neighborhood Attack:** When any attacker who is having information about the neighborhood of an individual and relationships amongst neighbors, makes the usage of that information for identifying that particular individual from any social site [26].

**Skewness attack:** When the sensitive attributes from equivalence classes of an anonymized table are relating with each other, target could be predicted by any attacker with assurance. Thus the attack is persuaded with the help of skewed sensitive attributes is known as skewness attack. Which yields in attribute linkage or attribute disclosure[16].

**Proximity attack:** When some of sensitive attributes that lies in a range are recurrently available in data, then attacker can predict the sensitive attributes with assurance. This attack yields in attribute linkage or attribute disclosure [27].

**Freeform attack:** It takes place when any attacker makes the usage of any attribute of record owner of data and guesses a sensitive value confidently. The values with are used to make the combinations are not essentially quasi-identifiers. This attack yields in attribute disclosure or attribute linkage [28].

**Minimality attack:** This attack takes palace whenever any attacker makes the usage of the information about algorithm of anonymization to make the anonymity reverse and tries to get the original data from anonymized values [29].

**Homogeneity attack/ Consistency attack:** It occurs over an anonymized table which is not having diverse values for its sensitive attributes and an attacker infers a pattern from anonymous table even without combing with an external table. This attack yields in attribute linkage or attribute disclosure[30].

**Background knowledge attack:** Itself, it does not disclose the privacy but when opponent associates some of the attributes with other information for getting more detailed implication on someone's sensitive information then background knowledge attack takes place. It yields in attribute linkage or attribute disclosure [30].

As when the data is not being misused, people do not find that their privacy has breached. The problem is that when the information is published it could be difficult to avoid its misuse. A number of different techniques provide the solution to this privacy issue to some extent [9].Furthermore, according to the privacy laws, hospital records regarding the patient's health should must be confidential [31]. For this purpose, there must be some efficient mechanisms for maintaining the privacy when data is being shared in a distributed environment [32].

Various data privacy preservation techniques have been introduced to anonymize the data. For protecting the privacy leakage due to publishing the data two different models are available [12]. Syntactic based model depicted by k-anonymization and semantic based proposed by differential privacy models [33]. To publish the data, k-anonymization was the first privacy model to be introduced. For improving the limitations of k-anonymization, various privacy models has introduced like l-diversity, t-closeness [12].

K-anonymous β-likeness has been introduced for protecting data from identity disclosure and attribute disclosure. Two clustering based algorithms were designed for creating k-anonymous β-likeness model to handle the background knowledge attack [8]. A personalized model for data privacy which is based upon the anonymizing technique for data related with flight paths is proposed. For this purpose, the sensitive data attributes are being generalized while suppression is performed on the trajectory data. With the help of this technique linking attack and similarity attack can be handled and minimized to some extent [15].

A clustering based algorithm having two phases for preserving the privacy of data on cloud by making usage of the anonymization is introduced. To create the equivalence classes of the diverse sensitive values, semantic diversity is being considered [34]. A flexible methodology for anonymizing the distributed data is proposed which is a combined technique for preserving the privacy of data by making use of the secure multiparty communication and anonymization which helps in protecting the data from linking attack [35].

Another, anonymizing method based on permutation has been introduced which takes data subject, intruder and anonymizing transparency under consideration. It uses the data diversity and handles the similarity attack. But it only works for numerical form of the data [36]. Another model which is m-privacy model based upon anonymization is proposed which is used to publish the collaborative data. It provide the protection to anonymized data from m opponents which are having information about original and the anonymized data [36].

Further, a hybrid technique has been proposed which combines the encryption of the data and anonymizing technique for preserving the privacy of the data in a cloud environment. But it does not provide a complete data privacy and can lead towards the identity disclosure and the attribute disclosure [37]. A k-anonymization based clustering method consisting of two steps is being introduced which is appropriate for both numerical and categorical data. It associates the two techniques of anonymization which are microaggregation and generalization [38]. A fast data oriented microaggregation algorithm is formed for anonymization. It is proposed to minimize the loss of information and to satisfy the anonymizing parameter k [39].

Another multi-agent medical data decision making system to preserve the privacy is offered by using the k-anonymity. The data integrated from various different resources are normalized to an acceptable level of accuracy. But still k-anonymity does not cover some attacks, for example, homogeneity attack, background knowledge attack, similarity attack, probability inference attack [17].

## 3. Methodology

The proposed methodology provides the new anonymization technique to handle the background knowledge attack and homogeneity attack in non-numerical data.

Take the table T = {e1, e2, ..., en}
For each entity ei there is a corresponding individual vi, known as entity respondent.
Every ei of T holds d QI attributes (which are denoted by A1,A2, ...,Ad)
Sensitive attribute => S
D[Ai], $1 \leq i \leq d$ denotes the attribute domain of Ai
D[S] = {s1, s2, .., sm} denotes the attribute domain of S.
Suppose that ei[Aj ] indicates the Aj value from ei present in T and ei[QI] indicates QI values.

**Anonymization algorithm**

At first, all the explicit identifiers should must be removed. All the identifiers must have some attributes. Next non-numeric data is labelized. For this purpose, numerical labelization is used for non-numeric letters. Next the anonymization hierarchies are applied and entities are generalized.

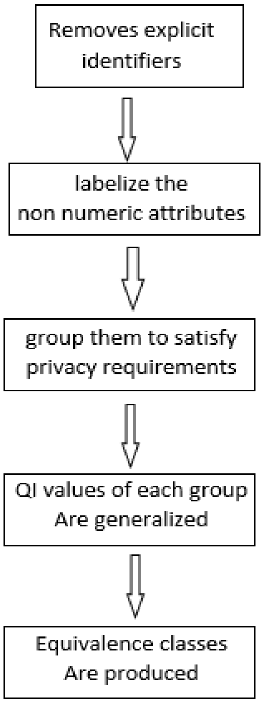

Removes explicit identifiers
labelize the non numeric attributes
group them to satisfy privacy requirements
QI values of each group Are generalized
Equivalence classes Are produced

Fig. 2 Anonymization algorithm

### Clustering based generalization algorithm

In this phase, equivalence classes are created through generalization to perform the anonymization and to control the background knowledge attack.

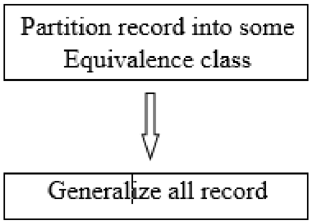


Fig. 3 Generalization

### The Hierarchical Anonymization Algorithm

```
Input:
T = {e1, e2, ..., en}: original microdata
Output: set of equivalence classes
    1. generate clusters set
    2. N = {N_1, N_2, ...,N_n}
    3. For cluster Ni such that |Ni| ≥ k do
    4. (Clustering phase) partition records within Ni
       into equivalence classes.
    5. end for
```

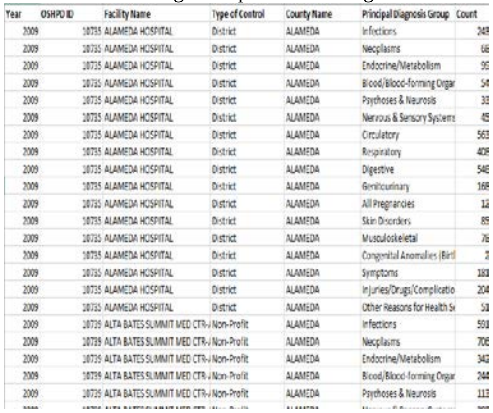
### K-anonymization Algorithm

```
Input:
    • Ni = {e_1, e_2, ..., e_n|}: set of all entities.
    • k: k-anonymity level
Output: set of equivalence classes E = {E_1, E_2, ...,E_j}
    1. while N' ≠ {}
            a. e_a=average entity of N' according to the
               QI values.
            b. C_k=GenerateCluster;
            c. clear = clear u C_k
            d. N'=N' \C_k
    2. end while
    3. if clear ≠ {}
            a. generate the equivalence classes using
               entities in clear
    4. end if
    5. return E
    6. end
    7. function GenerateCluster
apply k-anonymity level
apply microaggregation hierarchy level
```

## 4. Results & Discussion

For calculating the results ARX tool is used. ARX- is an open source tool which can be used to anonymize the sensitive data of any individual. It can be used in various different aspects and it can handle large datasets easily and different anonymizing approaches can be used easily by using a GUI interface which is very user friendly. Further data is collected from the internet. This dataset is having the distribution of inpatients discharge by a group named principal diagnosis for each hospital of California. This dataset contains almost 50000 entries from 2009 to recent available year.

Table 1: Original inpatient discharge data



| Year | OSHPD ID | Facility Name | Type of Control | County Name | Principal Diagnosis Group | Count |
|---|---|---|---|---|---|---|
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Infections | 249 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Neoplasms | 66 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Endocrine/Metabolism | 95 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Blood/Blood-forming Orgar | 54 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Psychoses & Neurosis | 33 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Nervous & Sensory Systems | 45 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Circulatory | 563 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Respiratory | 409 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Digestive | 548 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Genitourinary | 169 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | All Pregnancies | 12 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Skin Disorders | 85 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Musculoskeletal | 76 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Congenital Anomalies (Birt | 2 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Symptoms | 183 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Injuries/Drugs/Complicatio | 204 |
| 2009 | 10735 | ALAMEDA HOSPITAL | District | ALAMEDA | Other Reasons for Health Se | 51 |
| 2009 | 10739 | ALTA BATES SUMMIT MED CTR-A | Non-Profit | ALAMEDA | Infections | 591 |
| 2009 | 10739 | ALTA BATES SUMMIT MED CTR-A | Non-Profit | ALAMEDA | Neoplasms | 706 |
| 2009 | 10739 | ALTA BATES SUMMIT MED CTR-A | Non-Profit | ALAMEDA | Endocrine/Metabolism | 342 |
| 2009 | 10739 | ALTA BATES SUMMIT MED CTR-A | Non-Profit | ALAMEDA | Blood/Blood-forming Orgar | 244 |
| 2009 | 10739 | ALTA BATES SUMMIT MED CTR-A | Non-Profit | ALAMEDA | Psychoses & Neurosis | 113 |

Below the reidentification risks of attributes after applying the generalization is depicted. Of the left hand side reidentification risks of original data is shown and on the right hand side reidentification risks of the data after performing the generalization method is shown. As it is visible that records before generalization was at the highest risk and after applying the generalization highest risk record has been decreased at the rate 0.061%.

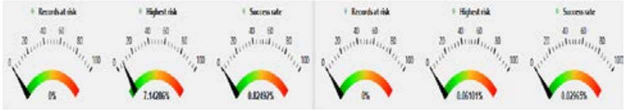


Fig. 4 Reidentification risks after generalization

Here the success rate of the generalized data is shown and from the figure it can be seen that data success rate of data re-identification has been decreased to 0.029%.

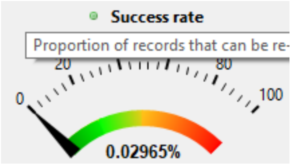


Fig. 5 Success rate of OD after Gen.

Here the data re-identification risks after clustering and microaggregation are shown. On the left hand side re-identification risks of the data is shown before applying clustering and microaggregation. On the right hand side re-identification risks of data after applying clustering and microaggregation are shown. After applying clustering and microaggregation risks has been reduced to 0.01%.

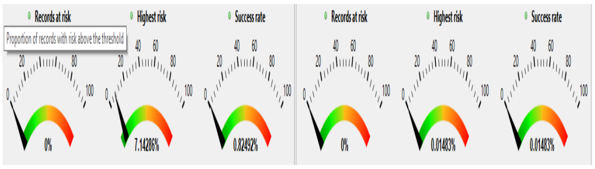


Fig. 6 Re-iden. Risks after CLu. & Mic.

Here the success rate of re-identification of attributes is shown before applying clustering and microaggregation which is high.

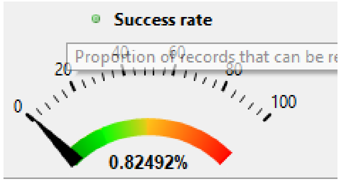


Fig. 7 Success rate of OD before CLu. & Mic.

Here the success rate of re-identification of attributes is shown after applying clustering and microaggregation which has been reduced to 0.014%.

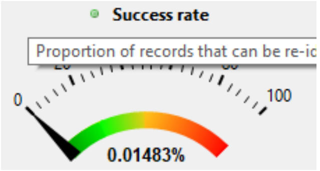


Fig. 8 Success rate of OD after CLu. & Mic.

## 5. Conclusion and Future work

Big healthcare data privacy is the major concern during the modern era. The data must have some security mechanism while published on the internet or shared with third party. So, the intruder can perform attacks on the data to get sensitive information which results in the privacy breaching of the patients. Therefore, sensitive information of the data needs to be protected from a number of attacks like linking attack, similarity attack, homogeneity attack, background knowledge attack. To protect the data which is in the non-numeric form k-anonymity is being used to handle the background knowledge attack and homogeneity attack. An algorithm is proposed using generalization and microaggregation. Results are being calculated with the help of ARX tool where the success rate of data reidentification is calculated. With the help of this proposed method data can be secured when it is in the non-numeric form.

As this algorithm only works for non-numerical data therefore, for future, it can be considered to enhance this algorithm for other forms of the medical data. Moreover, a simulation can be proposed for this algorithm to get better results.

## References

[1] "Big Data," Wikipedia, the free encyclopedia, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Big_data. [Accessed: 02-Dec-2018].

[2] J. Hellerstein, "Parallel Programming in the Age of Big Data," Gigaom, 2008. [Online]. Available: https://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/. [Accessed: 10-Dec-2018].

[3] M. Rouse, "big data," TechTarget. [Online]. Available: https://searchdatamanagement.techtarget.com/definition/big-data. [Accessed: 10-Dec-2018].

[4] B. K. Sarkar, "Big data for secure healthcare system: a conceptual design," Complex Intell. Syst., vol. 3, no. 2, pp. 133–151, 2017.

[5] L. Fernandes, "Big data, bigger outcomes," J. AHIMA, no. November, pp. 37–43, 2014.

[6] J. A. P, "Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data," vol. 10, no. 2, pp. 247–253, 2017.

[7] “Healthcare Data Breach Statistics,” The HIPAA Journal. [Online]. Available: https://www.hipaajournal.com/healthcare-data-breach-statistics/. [Accessed: 19-Sep-2018].

[8] F. Amiri, N. Yazdani, A. Shakery, and A. H. Chinaei, “Hierarchical anonymization algorithms against background knowledge attack in data releasing,” Knowledge-Based Syst., vol. 101, pp. 71–89, 2016.

[9] I. K. Gayki and A. S. Kapse, “Privacy Preservation of Published Data Using Anonymization Technique,” vol. 5, no. 2, pp. 2355–2357, 2014.

[10] X. Jin, M. Zhang, N. Zhang, and G. Das, “Versatile publishing for privacy preservation,” Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’10, p. 353, 2010.

[11] T. Karle and D. Vora, “Privacy preservation in big data using anonymization techniques,” 2017 Int. Conf. Data Manag. Anal. Innov., pp. 340–343, 2017.

[12] BENJAMIN C. M. FUNG, K. WANG, R. CHEN, and P. S. Y. University, “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” ACM Comput. Surv., vol. 42, no. 4, pp. 1–53, 2010.

[13] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, “T-closeness through microaggregation: Strict privacy with enhanced utility preservation,” 2016 IEEE 32nd Int. Conf. Data Eng. ICDE 2016, pp. 1464–1465, 2016.

[14] L. SWEENEY, “k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,” Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 05, pp. 557–570, 2002.

[15] E. Ghasemi Komishani, M. Abadi, and F. Deldar, “PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression,” Knowledge-Based Syst., vol. 94, pp. 43–59, 2016.

[16] N. Li, T. Li, and S. Venkatasubramania, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” IEEE 23rd Int. Conf., no. 3, pp. 106–115, 2007.

[17] H. Wimmer, V. Y. Yoon, and V. Sugumaran, “A multi-agent system to support evidence based medicine and clinical decision making via data sharing and data privacy,” Decis. Support Syst., vol. 88, pp. 51–66, 2016.

[18] G. Perera, A. Holbrook, L. Thabane, G. Foster, and D. J. Willison, “Views on health information sharing and privacy from primary care practices using electronic medical records,” Int. J. Med. Inform., vol. 80, no. 2, pp. 94–101, 2011.

[19] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, “Big data security and privacy in healthcare: A Review,” Procedia Comput. Sci., vol. 113, pp. 73–80, 2017.

[20] A. Gartenberg, “IBM Predictive Analytics to Detect Patients at Risk for Heart Failure,” Adam Gartenberg’s Blog Business Analytics and Optimization, IBM and Social Marketing, 2014.

[21] K. R. Ghani, K. Zheng, J. T. Wei, and C. P. Friedman, “Harnessing big data for health care and research: Are urologists ready?,” Eur. Urol., vol. 66, no. 6, pp. 975–977, 2014.

[22] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The parable of google flu: Traps in big data analysis,” Science (80-. )., vol. 343, no. 6176, pp. 1203–1205, 2014.

[23] Y. Ren, R. Pazzi, and a Boukerche, “Monitoring patients via a secure and mobile healthcare system,” Wirel. Commun. IEEE, vol. 17, no. February, pp. 59–65, 2010.

[24] V. Reddy, K. M. Sc, M. Biswas, P. Krishnan, and K. M. Sc, “Healthcare Fraud Management using Big Data Analytics,” vol. 2011, no. Hhs, 2012.

[25] J. Pei, Y. Tao, J. Li, and X. Xiao, “Privacy preserving publishing on multiple quasi-identifiers,” Proc. - Int. Conf. Data Eng., pp. 1132–1135, 2009.

[26] B. Zhou and J. Pei, “Preserving privacy in social networks against neighborhood attacks,” IEEE 24th Int’l Conf. Data Eng. (ICDE ’08), vol. 00, no. d, pp. 506–515, 2008.

[27] Y. Li, X. He, W. Wang, H. Chen, and Z. Wang, “Preservation of proximity privacy in publishing numerical sensitive data,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7808 LNCS, pp. 563–570, 2013.

[28] K. Wang, Y. Xu, A. W. C. Fu, and R. C. W. Wong, “FF-anonymity: When quasi-identifiers are missing,” Proc. - Int. Conf. Data Eng., pp. 1136–1139, 2009.

[29] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, “Minimality attack in privacy preserving data publishing,” Proc. 33rd Int. Conf. Very Large Data Bases, pp. 543–554, 2007.

[30] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “ℓ-Diversity: Privacy beyond k-anonymity,” Proc. - Int. Conf. Data Eng., vol. 2006, p. 24, 2006.

[31] “The HIPAA Privacy Rule,” U.S. Department of Health & Human Services. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. [Accessed: 26-Nov-2018].

[32] K. Alptekin, “Research Issues for Privacy and Security of Electronic Health Services,” Futur. Gener. Comput. Syst., 2016.

[33] P. Sui and X. Li, “A privacy-preserving approach for multimodal transaction data integrated analysis,” Neurocomputing, vol. 253, pp. 56–64, 2017.

[34] X. Zhang et al., “Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud,” IEEE Trans. Comput., vol. 64, no. 8, pp. 2293–2307, 2015.

[35] F. Kohlmayer, F. Prasser, C. Eckert, and K. A. Kuhn, “A flexible approach to distributed data anonymization,” J. Biomed. Inform., vol. 50, pp. 62–76, 2014.

[36] J. J. Vedha and V. Kavitha, “Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop,” Futur. Gener. Comput. Syst., 2016.

[37] J. J. Yang, J. Q. Li, and Y. Niu, “A hybrid solution for privacy preserving medical data sharing in the cloud environment,” Futur. Gener. Comput. Syst., vol. 43–44, pp. 74–86, 2015.

[38] J. Han, J. Yu, Y. Mo, J. Lu, and H. Liu, “MAGE: A semantics retaining K-anonymization method for mixed data,” Knowledge-Based Syst., vol. 55, pp. 75–86, 2014.

[39] R. Mortazavi and S. Jalili, “Fast data-oriented microaggregation algorithm for large numerical datasets,” Knowledge-Based Syst., vol. 67, no. September, pp. 195–205, 2014.