THINK INDIA JOURNAL
ISSN: 0971-1260
Vol-22-Issue-16-August-2019

A Novel Approach To Privacy Preserving Using Anonymization & Differential Privacy

Vinay Bhardwaj

vinay.23825@lpu.co.in

Abstract- Globally, a huge amount of data is collected and published on daily basis which include a lot of sensitive information about the users. Thus, the privacy of data is a biggest concern and it is important to maintain the secrecy of such information. The data collected from centralized servers is processed under various data mining techniques for getting useful information, analysis and decision making. The trust of the users isentirely dependent on maintaining the privacy which may be violated while performing mining techniques on such a big data. Thus it is mandatory to develop an effective technique for data mining which preserve the data as well as information. It is further required that the data quality remains as per standard for further classification and analysis purposes. This paper is an attempt to study various privacy issues with a precise and systematic review of existing techniques of privacy preservation (Anonymization and Differential Privacy) and to put forward a concept which results into nominal information loss and better utility of data.

Keywords: Big data, Privacy, Anonymization, Differential Privacy.

1 Introduction

In the era of technology, there is an overflow of data. Government agencies, Corporate industries, Social networking sites are the major sources of data generation. The large proportion of data is contributed via online as well offline by Social media (i.e., Facebook, LinkedIn, and Twitter), telecommunication networks. With availability of huge amount of information, individuals hev entered are into an era of Big data. *Big data* is basically a problem statement. The conventional statistical method is not capable of manage such a big quantum of information. Thus, big data term came into existence to refer all the data generated around the world. This data could be structured, unstructured or semi-structured. Information technology helps in creating digital data and exchanging it through internet. For instance, hundreds of Petabytes (PB) of information was processed by Google, 10PB per month of log data is created by Facebook, tens of Terabytes (TB) information is generated of online trading by Taobao, a subsidiary of Alibaba. Such a huge data create the issues of privacy for the people. When such a large collection of data is published and used for mining purposes, this often leads to disclosing the sensitive information which a person doesn't want to disclose. For example, an individual can be identified from published microdata by simply joining the quasi-identifiers (QID) attributes with an external data source. The most reliableand appropriate source of data is Social networking sites. For instance, every minute on an average a video of 72 hours is being uploaded on Youtube. Therefore, the major concern is to maintain the privacy while collecting and compiling the data. To avert the disclosure of all information i.e. personal and sensitive are collected from large microdata being published is called as *Big data privacy*. Various privacy preserving techniques of data mining are being used to preserve the critical information while extracting the useful information for some other purposes in a manner that the real data remain safe with the cloud providers. The various methods are being used such as Cryptography, Perturbation, Anonymization, Swapping, Condensation, Differential Privacy and so on. This paper focuses on Anonymization and Differential Privacy because the proposed concept is based on these two techniques.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

2 Privacy and Issues of Privacy

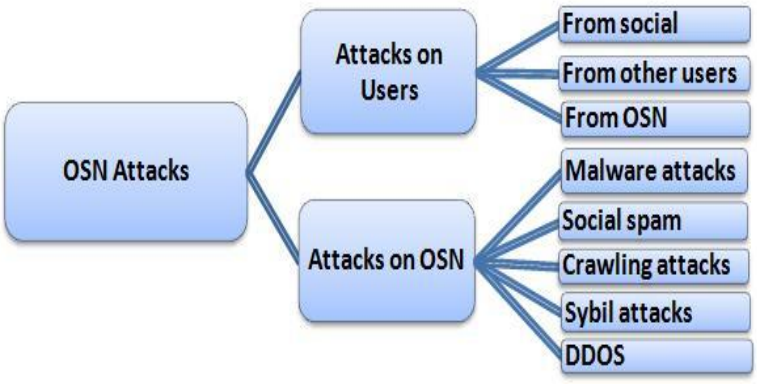
Privacy is the main security concern and an objective in today's age of networks because privacy include both the protection of personal information as well as the published social data. Privacy and security of information are the current issues. It is a big challenge for big data as far as the privacy of individual data as well as organizational data is concerned. For example, social data contain information like age, sex, gender and also other quasi-identifiers including individuals relationships which when linked to some other external data sources like voter card or census data or medical records may reveal a person's identity and also his sensitive information which he/she doesn't want to disclose like health problem, salary, location and so on. A personal profile of users on Social network is significant for third parties like advertisers for better segmentation, targeting and positioning of product.

The various privacy issues regarding the privacy concern are discussed below:

Privacy issue in healthcare data: Analytics and researchersare having immediate access to medical records which is being used by the doctors for diagnostic purpose. Electronic Health Record (EHR) helped in digitalization of healthcare system. Moreover,EHR program promotes hospitals to develop a precise and complete EHR. However, EHR having personal information of patients is a privacy concern. Pathology report analysis of patients is containing sensitive attributes which are difficult to be identified.

Privacy issue in mobile data: Mobile phones have accelerated the pace at whichpeople access, acquire and generate the data. Through the details provided by the service provider, one can easily identify a person who is using the data. Also, the data contain more private information about individuals such as habit, location and health condition. This data can be analyzed and collected by the data managers in order to share with researchers, governments and/or companies. This privacy issue badly calls for its preservation before it is shared widely.

Privacy issue in online social network (OSN) data: The exponential growth of social media, which is one of the techniques to share a lot of information (e.g. Twitter, Facebook, LinkedIn) cause serious effects for traditional data analysis algorithms and techniques. It is because of the computational complexity for larger datasets [4]. Public network information contain private and sensitive data regarding people and their relationship. Adversaries can exploit various kinds of techniques (e.g. quasi-identifiers) to analyse the data of social networks and hence can compromise the privacy of a user. Online social network (OSN) privacy and security problems can be divided into two different types [6]. First is the attack on users and second is the attack on OSN itself. Attacks on individuals are those attacks which are generally unapproachable and are targeted on a small-scale population of irregular or target individual. These can be attacks from social applications, inference attacks, de-identification attacks, attacks from OSN etc. Atacks on the OSN are aimed on the service purveyoritself, by intimidatingits central business. Attacks on OSN can be crawling attacks, malware attacks, Sybil attacks, DDOS (Distributed denial of service attacks) and so on.



ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Figure 1: OSN attacks categories

Privacy issues in Web data: Privacy concern in web data is biggest concern. Intel intends to create its internal websitedynamic based on web usage data of all the users of the website. In these settings, the looks of the website have been decided according to the pattern of search by the users and the link searched by the most of the users should be shown as suggestion on the first page as it will reduce the time of search as well as enhance the efficiency. From the browsing history, the Ip address can be found and their activities on the internet could be traced. Such a procedure are a breach to the privacy. Models containing symmetric key encryption can be used to anonymize the sensitive data identified based on predefined tags like IP address and users ID from semi-structured web usage data [5].

3 Anonymization

Anonymization is also known as de-identification. It is a type of information sanitization whose goal is privacy preservation. In other words, it is the process that either encrypt or remove personal information from datasets. For example, patient's personal information (name, age, sex) can be removed from the recipient. Generalization and perturbation are two popular approaches of anonymization. In 2000, it was found that 87% of the US population was uniquely identified by their five-digit zip code, gender, and date of birth [8]. In 2006 [11], when simple demographics in US population were revisited by the researchers, it was found that the count was 63% in uniquely identifying the information. We call this attack *Record linkage* where given quasi-identifier information; we can point to one or very few records in the dataset. Like relational data privacy preservation methods for social networks, several privacy preservation anonymization methods have been developed. *K-anonymity* can be used by bulding each independent homogeneous from at least K-1 other ones with similar structure. There must always be at least K records for each equivalence group present in the dataset (where equivalence group is equal to all records with the same combination of quasi-identifiers). *K-anonymity* is a property which is possessed by certain anonymized data. If we are given the data having a specific set of fields, the data become practically useful without disclosing the identity of the individuals who are the subjects of the data. *K-anonymity* is generally classified into two categories [7] [9], as follows-

- *Cluster or generalization-based approach:* This focus on identical nodes and edges are clustered and generalized to besimilar, with each group containing at the minimum of k vertices. Or in other words, vertices and edges are grouped into cluster and then anonymizing a subgraph within a super-vertex.
- *Graph improvement approach*: The topological layout of the graph is changes to make every node in frame to be similar by adding upor deleting edges from vertices in a graph.

3.1 Generalization & Suppression

Generalization means put back the values that shows in the database in the subsets of the values, so that each entry $R_i(j)$, $1<=i<=n,1<=j<=r$, where element of A_j, is changed by a subset of A_j that incorporate that element. Or in simple words, individual values of attributes are replaced by with a broader category. *Suppression* refers to the removal of data from the table so that they are not released. Suppression is applied at row level and a row can be suppressed only in its entirety. It happens many a time when we publish a data which is not properly anonymized. For example, a famous data breach in Massachusetts [10], i.e. a dataset containing health information like patient's date of birth, sex, zip code and health problem he is facing was analysed. The dataset didn't contain any other information like names, social security numbers or any other identifying information. Still, an adversary or a researcher was able to cross-reference this dataset with the voter list and identified the Governor William Weld of Massachusetts from this dataset. This kind of problem is illustrated

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

in Figure 2 showing a table of released medical data discovering by repressing names and Social Security Numbers (SSNs) so that the identities of persons to whom the data belongs are not disclosed. As illustrated in Figure 2, ZIP, DOB, and sex are linked to their voter list and reveal the name, address, and city. Also, race and marital status also linked to other public available population registers. In the Figure 2,of medical data where female, who born on 9/15/60 and resident in the 02141 area. This identified Sue J. Carlson living in Boston has a problem of shortness of breath.

Medical Data Released as Anonymous

SSN	Name	Race	DOB	Sex	ZIP	Marital status	Problem
		Asian	09/30/65	Female	02140	Divorced	Obesity
		Asian	09/15/65	Female	02140	Divorced	Hypertension
		Asian	04/18/65	Male	02140	Married	Chest pain
		Black	03/15/64	Male	02139	Married	Hypertension
		Black	03/18/64	Male	02139	Married	Shortness of breath
		Black	09/13/65	Female	02140	Female	Shortness of breath
		White	05/14/60	Male	02139	Married	Chest pain
		White	05/08/60	Male	02139	Married	Obesity
		White	09/15/60	Female	02141	Female	Shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....
Sue J. Carlson	1459 Main St.	Boston	02141	09/15/60	Female	Democrat	.....
.....	.....	.....	.....	.....	.....	.....	.....

Figure 2: An Example of anonymous data by linking to external information

We have given an example of a dataset in Table 1, to anonymize a data and then see the utility of anonymized data. The table contains a list of people, their sex, age, location, zip code, profession and the party they vote for. From the table, we recognise that k=1, i.e. one can easily identify that a particular person will vote for a particular party by cross-referencing it with the voter list. Now we will anonymize it and get k=2 or k=3. A person living in Baba Bakala and Ajnala can be generalized to Amritsar, zip code isgeneralised by using asterisk symbol and the age attribute can also be generalised to ranges instead of writing the exact values. We generalise the profession of hardware engineer and software engineer to just an engineer. Likewise, we generalised the professions plumber and carpenter to the craftsperson, ranged their age and set their sex to 'any' as they didn't seem to be much similar. There was a person with profession botanist whose record doesn't fit with anyone so its record is suppressed. This generalization can be understood from the hierarchy diagram in Figure 3 and the table 2 showing generalised and suppressed data.

Table 1: Dataset to be anonymized

Age	Sex	Profession	Location	ZIP	Party Affiliation
25	Male	Teacher	Ajnala	143102	C
52	Female	Software Engineer	Ludhiana	141001	A
55	Male	Teacher	Baba Bakala	143202	A
34	Female	Hardware Engineer	Ludhiana	141001	A
62	Female	Carpenter	Gurdaspur	143521	C
43	Male	Plumber	Gurdaspur	143521	A
23	Female	Botanist	Samrala	141114	B

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Table 2: Generalised and suppressed version of table 1 with k=3 anonymity

Age	Sex	Profession	Location	ZIP	Party affiliation
20-29	Male	Teacher	Amritsar	14****	C
45-58	Female	Engineer	Ludhiana	14****	A
48-57	Male	Teacher	Amritsar	14****	A
30-38	Female	Engineer	Ludhiana	14****	A
60-68	Any	Craftsman	Gurdaspur	14****	C
39-46	Any	Craftsman	Gurdaspur	14****	A
...	...	...	...	...	...

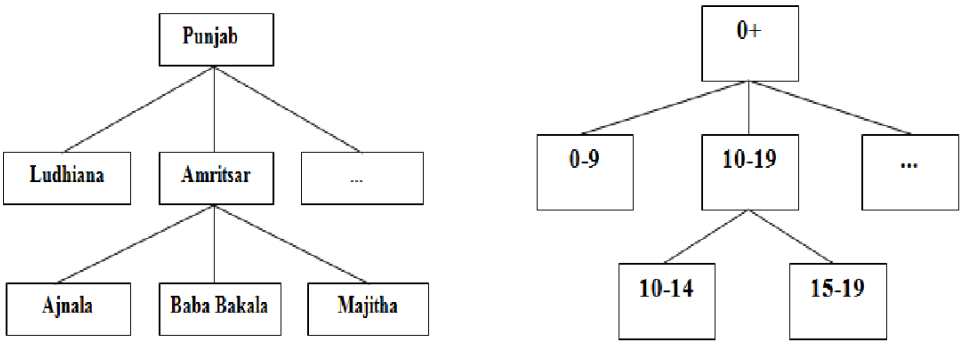


Figure 3: Hierarchy for generalisation

Table 2 showed k=3 anonymity. We have a problem with it. We still have information that people having profession engineer will vote for A. This attack is called an *attribute linkage*. Target is vulnerable because some sensitive values dominate target's equivalence group. So a solution to this is to use *l-diversity*.

l-diversity [12] says that there must be at least l distinct values for each sensitive attribute and an equivalence group(includes k-anonymity with k>=1). Table 2 has l=1 anonymity. It is then anonymized so that l=2. So we will generalize further as shown in Table 3.

Table 3: Anonymized data with l=2 success!

Age	Sex	Profession	Location	ZIP	Party Affiliation
20-29	Male	Teacher	Amritsar	14****	C
45-58	Any	Any	Punjab	14****	A
48-57	Male	Teacher	Amritsar	14****	A
30-38	Any	Any	Punjab	14****	A
60-68	Any	Any	Punjab	14****	C
39-46	Any	Any	Punjab	14****	A
...	...	...	...	...	...

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

We lost many information contents as shown in Table 3 but still, our data is not yet safe. We achieved l=2 success! but we have an information that 73% of people will still vote for A. We still have a dominating for the equivalence group. So an alternative to this is ***t-closeness*** [13] which says that distribution of a sensitive attribute in any equivalence group must be close to this attribute'sdistribution in the whole dataset. As a result, adversary can't figure out anything from the dataset because the distribution of equivalence group can be quite similar to the overall distribution.

4 Differential Privacy

When dealing with statistical databases, the privacy risk should not substantially increase. An alternative approach to anonymization is Differential Privacy, going away from quasi- identifiers and equivalence groups.Valuable information like individual's identifying information can be gained by Differential Privacy and hence can be protected. It assures that an individual cannot be recognized whenever this individual is deleted from the dataset or is in the dataset and so the results will not change much. Differential Privacy is an efficient way to achieve maximum accuracy from analysis of statistical databases along with minimizing the instances of record disclosures. Each record of a dataset is independent of the rest. When Differential Privacy is applied to two different datasets, the probability of the output will be nearly same.

Definition 1: A private function F givesε-differential privacy if for database D1 and database D2, differing on at most by one element, andall S ⊆ Range (F),

$$\Pr [F (D1) \in S] \leq \Pr [F (D2) \in S] \times e^{\varepsilon}$$

Where S is a database containing all outcomes such that S ⊆ Range (F) and parameter ε allows us to control the level of privacy.

Lower the value of ε, stronger will be the privacy. This is because of the limitations provided by the influence of a record on the outcome of a calculation. The values of ε are considered to be smaller than 1 [15], e.g. 0.01 or 0.1 (for smaller values we have $e^{\varepsilon} \approx 1+ \varepsilon$).

Definition 2: Laplace noise mechanism (lap (σ)), in which magnitude is associated to the variation that the elimination of asingle participant that cause on the output. The Laplace distribution (let μ=0) wherescale b is the distribution of probability density function:

$$(\mid) = \frac{1}{2} \; ^{-\perp\!\!\!\perp}$$

The variance of the Laplace distribution is $\sigma^2=2b^2$.

$\Delta = \max_{1,\,2} \| (1) - (2) \|$

Differential Privacy has two vital properties:

- Sequential Composition: A set of mechanism M_m is provided by differential privacy on an input set D is $\sum_m \varepsilon_m$.
- Parallel composition: If mechanism M_m acts on every disjoint subset $D_m \in D$, the privacy provided is (max (ε_m)) - Differential Privacy for all M_m [16].

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

4.1 Achieving Differential Privacy:

Among many approaches to differential privacy, we have discussed the one which is concerned to the output. For example, if we want to get a view that how many people are doing something embarrassing like biting their nails (Onychophagia).So, we can set up a server with the question "Are you a nail biter?" with the 'Yes' and 'No' options below it. The answer can be collected somewhere in the server and instead of sending real answers we can introduce some noise. For example, a person named John has a Yes answer. Before sending his real response a differential privacy algorithm will flip a coin. If it comes 'Heads' the algorithm sends the real answer to the server and if comes 'Tails' the algorithm flips the second coin. After flipping the second coin it sends 'No' i f its Heads and sends 'Yes' if its Tails. The data is received by the server but due to the added noise, the results cannot be really trusted. The record for John might say that he is a nail biter but there is at least one-fourth chance that he is not a nail biter.

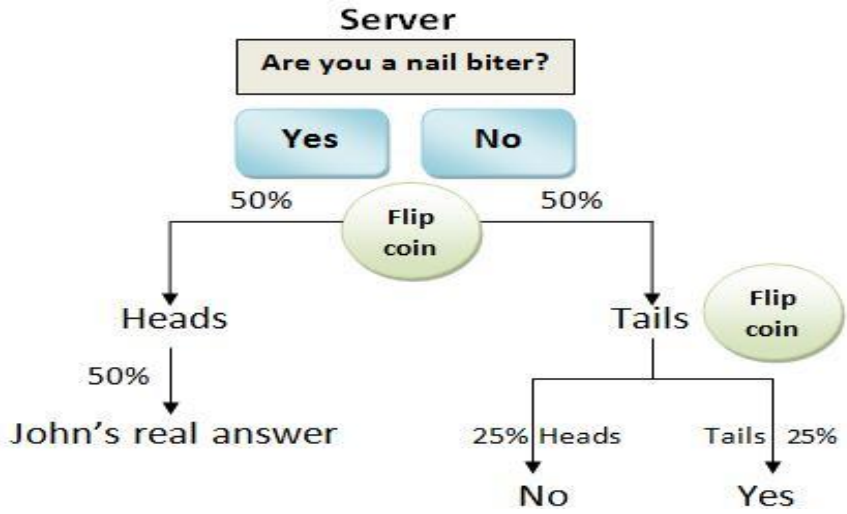


Figure 4: Achieving Differential Privacy

This was a noise added output preserving data privacy and provides equilibrium between privacy and data utility. Differential privacy is suitable with all types of datasets and the speed of processing also remains unaffected. It results in lesser data loss and guarantees true privacy.

5 Related Work

In this paper, we have analysed and observed many areas of research in the field of privacy preserving data mining which helped us to present our new hybrid concept regarding the privacy of large microdata being published every day i.e. big data. Pierangela Samarati used anonymization technique called k-anonymity for privacy preserving data publishing without effecting the truthfulness of sanitized data by using generalization and suppression. Such methods preserve the 'truthfulness' of information but make information less precise. Also, it preserves privacy against record linkage attack alone and fails to protect attribute linkage. Another method, l-diversity occasionally unableto handle identity disclosure attack and attribute disclosure attack. *t*-closeness method is not successful in dealing with issues of the privacy against identity disclosure attack while it is capable of handling privacy in case of attribute disclosure attack. Although there is a huge computational complexity.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

R. Mahesh and T. Meyyappan also used generalization and suppression techniques by splitting the dataset into n number of classes. Some general values were used as a replacement of Quasi-identifiers which were being suppressed by using some character (e.g. *). Duplicate records were eliminated for reducing information loss but the method works well only in case of numeric data. Also, the exact values can never be gained when required.

Himanshu Taneja suggested the technique in which thecombinationof the k-anonymity, l-diversity and t-closeness approach are meant for privacy preservation. He suggested the method by using ARX anonymization tool. For the purpose of achieving k-anoymity efforts was done to create different hierachiesfor quasi-identifiers while sensitive attributes were considered for l-diversity on the sensitive attributes and helped in reducing the average re-identification risks from 100% to 2.33%.

P.Usha, et. al. suggested a heterogeneous anonymization technique by using generalization was being used on the quasi-identifiers and non-sensitive data got published. In this, the fields were categorized and divided into tables.Furthermore, the table got divided into tables ofreactive attributes, quasi identifiers and rest of attributes. The heterogeneous anonymization is used on the starting of the table whereas it was required to combine both the two tables to get the results.

Unil Yun et. al. brought into light a technique of fast perturbation applied on larger size database. The researcher showed the data set in the structure of tree form in which every node shouldin the tree consists to a particular field such as name, parent, set, utility information list, node link and count. In this procedure authorsuses a tree structure where each node contains a separate path thus reduced the computation access time.

Krishna Mohan Pd Shrivastva et. al. have very well analyzed Differential Privacy along with suggestions how it would be suitable for privacy preserving big data. Another method, introduced a concept of providing privacy over a non-interactive framework using differential privacy and generalization. It proved to have better performance for classification purposes and scalability after anonymizing but didn't show the impact of this approach in terms of information loss.

6 Proposed Method

All the privacy preserving techniques discussed above are used for privacy protection but have certain limitations due to one or the other reason. Techniques like k-anonymity and perturbation results in huge information loss in many cases. Also, various cryptographic and encryption algorithms are used to achieve stronger privacy but there working is complex and time consuming. The proposed concept introduces a hybrid privacy preserving approach using anonymization and differential privacy. It will not only help to reduce the issue of information loss but will also preserve the data quality for further data mining analysis purposes.

The proposed concept focuses on collecting a high dimensional dataset from which it can categories the quasi-identifiers and the sensitive attributes. Quasi-identifiers like age, gender, job or any other identifier can also be hidden by using suppression and generalization techniques of anonymization and by using noisy values for them introduced by the mechanisms of differential privacy like Laplasian noise mechanism.

After the generation of a new sanitized dataset resulted by using the proposed algorithm, comparisons can be done with other privacy preserving techniques and the performance of proposed method can be evaluated by various parameters like information entropy, information loss, efficiency in terms of execution time and space and also calculating the classification accuracy of sanitized dataset. Besides reducing high information loss, the proposed method preserves the data utility for classification purposes.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

7 Conclusion

Big data generating large volumes of data calls for its privacy and security. In this paper, we presented an analysis of various threats to an individual's privacy emerging at an exponential rate. We analysed the existing techniques of privacy like anonymization (k-anonymity, l-diversity, t-closeness) and differential privacy. When using only anonymization technique on the whole dataset we loss many information contents and have minimum data utility as compared to differential privacy. Differential privacy guarantees privacy because in case of large data sets, the noise will be needed to protect privacy which leads to lesser data loss. So, we combined anonymization and differential privacy which resulted in minimum information loss along with balancing the data utility.

References

[1] Min Chen, Shiwen Mao, Yunhao Liu: Big Data: A Survey,*Journal of Mobile Networks and Applications: Volume: 19, Issue: 2, April 2014: DOI 10.1007/s11036-013-0489-0*

[2] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham, Fellow, IEEE: Preventing Private Information Inference Attacks on Social Networks, *IEEE Transactions on Knowledge and Data Engineering*: Volume: 25, Issue: 8, Aug. 2013.

[3] Harsh Kupwade Patil and Ravi Seshadri: Big data security and privacy issues in healthcare: *2014 IEEE International Congress on Big Data:* DOI: 10.1109/BigData.Congress.2014.112

[4] Gema Bello-Orgaz, Jason J. Jung, David Camacho: Social big data: Recent achievements and new challenges*: Journal of Computer ScienceDepartment, Universidad Autónoma de Madrid, Spain*, Volume: 28, March 2016. doi:10.1016/j.inffus.2015.08.005

[5] Jeff Sedayao, Rahul Bhardwaj Intel Corporation Santa Clara, United States: Making Big Data, Privacy, and Anonymization work together in the Enterprise: Experiences and Issues*: in the proceedings of 2014 IEEE International Congress on Big Data (Big Data Congress),* June 2014. doi: 10.1109/BigData.Congress.2014.92

[6] Hatem AbdulKader, Emad ElAbd, Waleed Ead: Protecting online social networks profiles by hiding sensitive data attributes: *Symposium onData Mining Applications, SDMA2016*, 30 March 2016. doi: 10.1016/j.procs.2016.04.004

[7] Lin Yao, Dong Liu, Xin Wang, Guowei Wu: Preserving the Relationship Privacy of the Published Social-network Data based on Compressive Sensing: *Quality of Service (IWQoS), 2017 IEEE/ACM 25th International Symposium on.*doi: 10.1109/IWQoS.2017.7969109

[8] Latanya Sweeney: Simple Demographics Often Identify People Uniquely: January 2000.

[9] Bin Zhou, Jian Pei and Wo-Shun Luk: A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data: Volume: 10, Issue: 2, August 2007. doi: 10.1145/1540276.1540279

[10] Daniel C. Barth-Jones: The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now: July 2012.

[11] Philippe Golle: Revisiting the Uniqueness of Simple Demographics in the US Population: *Proceedings of the 5th ACM workshop on Privacyin electronic society*, October 30, 2006. doi: 10.1145/1179601.1179615

[12] Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer: l-Diversity: Privacy Beyond k-Anonymity: *Proceedings of the 22nd InternationalConference on Data Engineering (ICDE'06),* 2006. doi: 10.1145/1217299.1217302

[13] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity: *2007 IEEE 23rdInternational Conference on Data Engineering.*

[14] C.Dwork, F. McSherry, K. Nissim and A. Smith: calibrating noise to sensitivity in private data analusis: Theory of cryptography, pp.265-284, 2006.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

[15] Arik Friedman and Assaf Schuster: Data Mining with Differential Privacy: *in the Proceedings of the 16th ACM SIGKDD internationalconference on Knowledge discovery and data mining*, July 2010. doi: 10.1145/1835804.1835868

[16] Shuo Wang, Richard O. Sinnott: Protecting Personal Trajectories of Social Media Users through Differential Privacy**:***Computers & Security,*Volume: 67, June 2017. doi: http://dx.doi.org/doi: 10.1016/j.cose.2017.02.002

[17] P. Samarati and L. Sweeney: "Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression: p. 19.

[18] P. Samarati: Protecting respondents identities in microdata release: *IEEE Transactions on Knowledge and Data Engineering*, Volume: 13, Issue: 6, pp. 1010–1027, Dec. 2001.

[19] R. Mahesh and T. Meyyappan: Anonymization technique through record elimination to preserve privacy of published data,: 2013, pp. 328–332.

[20] H. Taneja, Kapil, and A. K. Singh: Preserving Privacy of Patients Based on Re-identification Risk: *Procedia Computer Science*, Volume: 70, pp. 448–454, 2015.

[21] P. Usha, R. Shriram, and S. Sathishkumar: Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining: 2014, pp. 1–5

[22] U. Yun and J. Kim: A fast perturbation algorithm using tree structure for privacy preserving utility mining: *Expert Systems with Applications*, Volume: 42, Issue: 3, pp. 1149–1165, Feb. 2015.

[23] K. M. P. Shrivastva, M. A. Rizvi, and S. Singh: Big Data Privacy Based on Differential Privacy a Hope for Big Data: In the proceedings of 2014 Sixth International Conference of Computational Intelligence and Communication Networks: pp. 776–781.

[24] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu: Differentially private data release for data mining: 2011, p. 493.