ijecs
open access
www.ijecs.in
International Journal of Engineering and Computer Science
Volume 12 Issue 04, April 2023, Page No.25691-25699
ISSN: 2319-7242 DOI: 10.18535/ijecs/v12i01.4671

Differential Privacy

"Working Towards Differential Privacy for Sensitive Text"

[1]Mohammad Naeem Kanyar, [2]Dr.Suchitra Nair R
[1]PG student, [2]Head of School
[1] School of CS & IT, [2]Department of MSc CSIT
[1] Jain (Deemed-to-be University), Bangalore, India
[1]naeem.kanyar@gmail.com, [2]r.suchithra@jainuniversity.ac.in

Abstract
The differential-privacy idea states that maintaining privacy often includes adding noise to a data set to make it more challenging to identify data that corresponds to specific Individuals. The accuracy of data analysis is typically decreased when noise is added, and Differential privacy provides a technique to evaluate the accuracy-privacy trade-off. Although it may be more difficult to discern between analyses performed on somewhat dissimilar datasets, injecting random noise can also reduce the usefulness of the analysis. If not, enough noise is supplied to a very tiny data collection, analyses could become Practically useless. The trade-off between value and privacy should, however, become more manageable as the size of the data set increase. Along these lines, in this paper, the Fundamental ideas of sensitivity and privacy budget in differential privacy, the noise mechanisms utilized as a part of differential privacy, the composition properties, the ways through which it can be achieved and the developments in this field to date have been presented.

Keywords: Differential, Privacy, Exponential, Laplace, Noises

1. Introduction
By using a technique called differential privacy, scientists and database analysts can access datasets that include personal data about individuals without revealing the individuals' identities. This is achieved by sparingly obscuring the information provided by the database. While causing just enough interruption to protect privacy, there is still enough room for researchers to use the data.[1] Privacy has a monetary value. Even better, it can evaluate privacy tactics and determine the most successful ones are the most successful. Even better, it can create defenses against hackers who have additional knowledge. And as if that weren't impressive enough, this can accomplish everything at once. Differential privacy is a probabilistic theory that explains these answers as well as others. For instance, as part of a contest to see if anyone could surpass its shared filtering algorithm, Netflix launched a database of its users' ratings in 2007.[2] The researchers were nevertheless able to violate privacy even though this dataset did not contain any personally identifiable information. They successfully retrieved 95% of all deleted personal data from the dataset. The researchers in this instance breached privacy by using supplemental data from IMDB. The two largest concerns are preventing unauthorized access to data and minimizing data leaks during data analysis. For a given computing job T and a specific value of, there will be a huge number of privacy-preserving methods for completing T in a -differentially private way. Some people will have a better sense of accuracy than others. Similar to finding a numerically stable method, establishing a highly accurate -differentially private approach for T at the position where is small can be difficult. One condition for defining privacy concerning reference to data analysis is that the researcher must not know anything about any person in the data set than he did before the study began. Privacy uses a technique to give consumers privacy; specifically, An early version of privacy by a randomized procedure is a randomized response, a technique created by social scientists to collect statistical data regarding humiliating or illegal behavior. Having a property

allows for the capture of this behavior.[3] Differentiated privacy is not always a guarantee that what someone thinks to be their mystery will remain that way. In other words, it guarantees that no one's involvement in research nor the contents of what one provided to the study will be revealed.

2. Literature Review

Both domestically and globally, related research and applications of differential data protection for personalized recommendations are in their infancy. Data analysis and publication are the two parts of a complete personalized recommendation system. The functioning of these two factors will all reveal users' private data. There are an increasing number of research findings on the training dataset and data gathering of differential online privacy as a result of researchers' keen interest in the technologies of differentiated privacy protection. A catalog of attacks has been published that retrieves sensitive characteristics or portions of text data from text embedding created by the well Language Models without making assumptions about structure or trends in the input text.[4] *Blum et al* proposed the *Su LQ-based ID3 approach*.[5] in line with the varying privacy protections offered by data mining categorization algorithms. This method linked the Laplacian noisy mechanism and gain ratio to select the suitable segmentation features. It is vital to calculate the information received for each characteristic, however, if there are numerous comparable segmentation attributes, a lot of privacy budgets will be wasted.[6] Numerous surveys have been conducted on differential privacy. Dwork's initial review contained an overview of principles, techniques, and particular differentially private data publication algorithms. The summary of compelling uses and probable future advancements in data release and data processing was then given by Dwork et al. A book by Dwork provides detailed explanations of algorithms supporting differential privacy against threats and privacy-preserving-preserving approaches for mechanism construction and machine learning.[7] Earlier, a number of privacy-protection techniques were developed, but they were unsuccessful. The Commonwealth of New England Group Insurance Commission (GIC) withheld some data, such as name, house number, and other personal details, to protect the clients' privacy when GIC disclosed the anonymous data medical files of its clients for study to help society in the middle of the 1990s.[8] Latanya Sweeney, a Ph.D. candidate at MIT at the time, identified the health record simply by comparing and matching the voting machine and the GIC database. So, hiding some information won't necessarily safeguard someone's identity.[9] Lots of sensitive training data are needed for natural language processing (NLP). According to Latanya Sweeney.[10] Because seemingly harmless fields might be related to certain other sources of data to facilitate re-identification, traditional redaction techniques (such as removing common personal identifying information) frequently fall short. Due to the various language models (LMs) recent success, security researchers have created sophisticated privacy threats.[11] Retrieving text from (a written document of) the training data makes use of a Language Model that has already been developed using the training data. The integration of federated learning with local differential privacy was suggested by Yang Zhao as a way to support crowd-sourcing applications and the development of machine learning general models.[12] For important data, use the Laplace mechanism, while for non-significant data, use the compression method with the inputs as a sparse vector. It is desirable to determine the quantity of noise and the number of positive integers by applying the Haar transform to the wavelet matrices in the compression mechanism. Additionally, demonstrate theoretically that our suggested strategy accomplishes -differential privacy.[13] a fresh approach that carefully chooses a played crucial factor for the Laplace distributed in order to maintain the differential privacy guarantee. In order to take data-dependent normalization variables into account and to study the privacy guarantee for various classes of ranging constraint configurations, the privacy promise in the framework of the Laplace distributions is modified.[14] to comprehend its use and how it might impact data analysis. To evaluate the accuracy of four classification techniques (Logistic Regression, Naive Bayes, MLP, and SVM), conduct trials with different privacy levels.[15] These brand-new, emergent issues in the technical domains of computer vision, supervised learning, and multi-agent networks can be resolved through differential privacy techniques. Applications including robots, natural language processing, and computer vision have benefited from advances in machine learning, machine learning, and multi-agent systems.[16]

3. Challenges with the Application of DP

The common services that people use on a daily basis, such as search results, mobile services, internet community activity, and so on, hold a vast amount of private information. This massive amount of statistically sensitive personal information has significant social worth and can be used to improve economic utility, understand disease transmission, allocate resources, and other things. Differential privacy in text sanitization has the main drawback of adding noise to the text, which makes it challenging for readers to grasp the original content. Differential privacy can also be computationally expensive, which makes it challenging to use in real-time applications. Additionally, it is open to assault from bad actors who might exploit the noise to extrapolate private information from a text. Since the assurance of data privacy includes limiting access to information, regulating the way information is used and making efforts to protect privacy, none of these measures are sufficient to provide the necessary level of data privacy. As a result, the destruction of privacy protection is imminent. Thus, differential privacy becomes necessary in the pursuit of improved and more reliable data privacy. With differential privacy given to the data, such analysis is not feasible. It stops an analyst from acquiring knowledge specific to particular people. For instance, differential privacy is inappropriate for a bank looking to identify instances of fraud. The inaccuracy introduced is comparable to sampling errors.[17] It is getting more and harder to keep sensitive information safe from being accessed or misused as big data usage increases. In addition, when working with text sanitization, there are numerous legal and regulatory criteria for protecting the data that must be considered. Since the assurance of data privacy includes limiting access to information, regulating the way information is used, and making efforts to protect privacy, none of these measures are sufficient to provide the necessary level of data privacy. As a result, the loss of privacy protection is imminent. Thus, differential privacy becomes necessary in the pursuit of improved and more reliable data privacy. It is not appropriate for all issues. Analysis at the individual level: With differential privacy given to the data, such research is not possible. It avoids.[18]

4. Research Methodology

The algorithm, going to be used in this project is the application of Laplace and Exponential mechanisms to add random noise to data for differential privacy using Python programming language. The dataset is a general dataset of random student information. However, the mechanisms could be applied to any existing or newly generated dataset. Laplace Mechanism (LM) and Exponential Mechanism (EM) are the two main noise mechanisms in DP. The volume of noise alludes to worldwide sensitivity and privacy budget. Thus, using both methods of adding random noises in this project. As a simple definition, differential privacy forms data anonymously via injecting noise into the dataset studiously. It allows data experts to execute all possible (useful) statistical analyses without identifying any personal information. Laplace Mechanism (LM) and Exponential Mechanism (EM) are the two main noise mechanisms in DP. Sensitivity and privacy budget.[19] The volume of noise alludes to worldwide

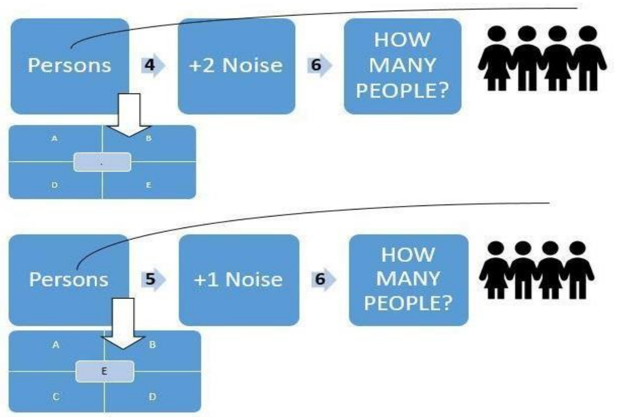

Figure 1 Differential Privacy through the noise

The example in *Figure. 1* to further grasp the idea of differential privacy. As seen in the above graphic, differential privacy prevents one from learning more about a person (Person C) regardless of whether or not she is included in the database.

Differential Privacy

DP is a definition, not a calculation. It was initially created by Dwork, Nissim, McSherry, and Smith, with real commitments by numerous others throughout the year [20][21]. Generally, DP works by embedding a go-between bit of programming between the examiner and the database [22].

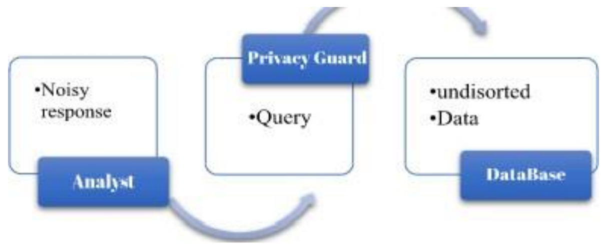

Figure. 2 Differential privacy mechanism

Figure 2 shows the differential privacy mechanism [23]. The analyst asks a question of the Privacy guard, a middle-man piece of software. Using a unique methodology, the guard evaluates the query's impact on privacy. The guard then sends the query to the database, which returns a clear response based on information that has not been altered in any way. In order to protect the privacy of the people whose data is in the system; the guard then modifies the response and gives it back to the analyzer. The quantity of "noise" added is scaled to the security impact. Let's assume that D1 and D2 are two datasets. If D1 and D2 in Eq. 1 only differ by one value, they are considered to be neighbors.

$$Pr[M(D1) = x] \leq exp(\varepsilon)\, Pr[M(D2) = x] \qquad (1)$$

In other words, based on the computation's output, it is impossible to determine which input piece of data was used because the likelihood of getting this result would be the same with or without that item. It is impossible to infer anything helpful about the object from the computation's output alone since it is impossible to determine whether the object was used at all. To achieve anonymity, the calculation of the formula must be randomized.

Mechanisms Used in DP

The two primary noise mechanisms in DP are the Laplace mechanism (LM) and the exponential mechanism (EM). The magnitude of noise alludes to privacy budget and global sensitivity [24] This project will employ the usage of Python programming to apply Laplace and exponential techniques to add randomness to data for different datasets. The data set consists of a general collection of random student data. The processes, however, could be used with any dataset, whether it was already created or not. The two primary noise mechanisms in DP are the Laplace Mechanism (LM) and the Exponential Mechanism (EM). [25] The level of noise suggests sensitivities and privacy budgets on a global scale. Consequently, this project uses both techniques for including random noises. The two primary noise mechanisms in DP are the Laplace Mechanism (LM) and the Exponential Mechanism (EM). The level of noise suggests sensitivity and privacy budget on a global scale.

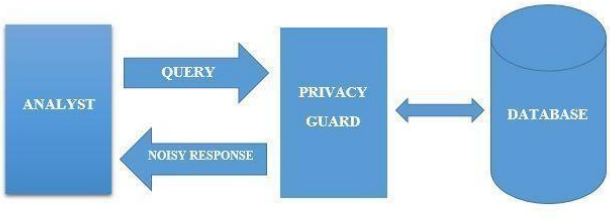

Figure.3 Differential Privacy Mechanism

The differential privacy technique is depicted in *Figure. 3*.[26] An intermediary piece of software called Privacy guard is questioned by the analyzer. The guard assesses the query's effect on privacy using a special methodology. The guard then queries the database, and the database responds with a clear answer based on data that hasn't been changed in any manner. The guard then alters the answer and gives it back to the analysis in order to preserve the privacy of the individuals whose data is in the database. According to the privacy impact, more "noise" is injected in varying amounts. To put it another way, it is impossible to tell if a certain piece of data was used an obtain an in-depth on the computation's result because of the probability that the result would have been produced without it.

Laplace Mechanism

As the name suggests, the Laplace mechanism will just compute function and perturb each coordinate with noise drawn from the LM distribution. The scale of the noise will be adjusted to the sensitivity of the function (divided by ε). LM is used when the output is numerical.

Given a dataset D and the function f: D $\rightarrow R$^d, global sensitivity is Δf; random algorithm in Eq. 2,

$$A(D) = f(D) + noise \qquad (2)$$

Satisfies ε-differential privacy if the noise complies with the Laplace distribution; that is, noise~Lap($\Delta f/\varepsilon$); there, the location parameter (LP) is zero and the scale parameter (SP) is $\Delta f/\varepsilon$. Let Lap (b) signify the Laplace distribution when LP is 0 and SP is b, and its probability density function is $(\chi) = \exp(-|\chi|/b)/2b$. The larger noise added to the output is, the larger b is and, in the meanwhile, the smaller ε becomes. Let $\sigma(\chi)$ denote standard deviation; $D(\chi)$ denotes variance, and *noise*~Lap(b) in Eq. 3,

$$\sigma(\chi) = \sqrt{D(\chi)},\ D(\chi) = 2b2,\ and\ \ b = \Delta f / \varepsilon \qquad (3)$$

Then the results obtained are in Eqs. 4 and 5,

$$D(x) = 2(\Delta f / \varepsilon)2 = 2\Delta f\, 2 / \varepsilon 2 \qquad (4)$$

$$\sigma(\chi) = \sqrt{D(\chi)} = \sqrt{2\Delta f\, 2/\varepsilon\, 2} = \sqrt{2\Delta f / \varepsilon} \qquad (5)$$

Consider numeric queries *f: D → R* k that map databases to k real numbers. The Laplace mechanism adds noise to the query answer. An important parameter that determines the amount of noise to

ensure differential privacy is the l1-sensitivity of the query.[27]
The sensitivity of a function $f : D \rightarrow R$ k is Δ

$$\Delta 1f = \max_{(D1,\ D2) \in Nd(D)} \| f(D1) - f(D2) \|$$

(*Theorem. 1*)

Theorem 1. Given a numeric query f: D → R k, the Laplace mechanism adds to the query answer f(D) with a vector (η1, · · ·, ηk), where ηi is i.d. random variables drawn from the Laplace distribution centered at 0 with scale b = Δ1f /ϵ, denoted by Lap(b). The Laplace mechanism preserves (ϵ, 0)-differential privacy. This parameter of query measures the largest possible change to the query answer between any pairs of neighboring databases

Exponential Mechanism

The exponential mechanism is another security-controlled plan to fulfill differential privacy when the outputs are non-numerical. Intuitionally, the exponential mechanism still guarantees that this change of a single DB tuple does not influence the outcome of the score function. The exponential mechanism was designed for circumstances in which it was wished to pick the best response. Let D denote the input dataset; $r \in R$ denotes one of the potential answers, given a score function u: $D \times R \rightarrow R$; if a random algorithm A selects an answer based on the probability as follows, then the algorithm A is said to satisfy ε-differential privacy in Eq. 6:

$$A(D,u) = r{:}|Pr[r \in R] \infty\ exp(\varepsilon u(D,r)\,/\,2\Delta u) \quad (6)$$

where Δu denotes the sensitivity of score function u and is defined as in Eq. 7:

$$\Delta u = max(r \in R)\, max((\ \| D\Delta D' \|\ ) = 1)\, |u(D,r) - u(D',r)| \quad (7)$$

The exponential mechanism can yield non-numerical results as indicated by their values of the score function. The output probability refers to the privacy budget from the given definition, and the highest scored result is given as output with higher probability when ε is larger; in the interim, when the difference between the output probabilities grows, the security turns out to be less; vice versa, the smaller ε is, the higher the security will be. The exponential mechanism can characterize a complex distribution over a large arbitrary domain; thus, it may not be conceivable to implement the exponential mechanism proficiently when the range of u is super polynomials large in the natural parameters of the issue. Given some arbitrary range R, the exponential mechanism [28] is defined with respect to some utility function u: D × R → R, which maps database and output pairs to utility scores. For a fixed database D, a better output from R should have a larger score. The sensitivity of the utility score is defined as

$$\Delta u = \max_{r \in R} \max_{(D1, D2) \in Nd(D)} |u(D1,\ r) - u(D2,\ r)|$$

(Theorem. 2)

Theorem 2 (The Exponential Mechanism). The exponential mechanism takes in the database D ∈ D and score function u: D × R → R, and outputs an element r ∈ R with probability proportional to exp(ϵu(r, D) 2Δu). This mechanism satisfies (ϵ, 0)-differential privacy. Differential privacy is compositional that is, running a differentially private mechanism twice also satisfies differential privacy, but at an increased privacy cost. The compositionality of differential privacy separates it from a number of other privacy notions, including de-identification and k-anonymity. In both of those cases, two separate releases of data may individually satisfy the desired property but may violate the property when taken together. Two differentially private releases of data, in contrast, may result in increased privacy costs, but will always satisfy differential privacy for some value o.[29] Differential privacy algorithms, such as the Laplace mechanism and the Exponential mechanism, add noise to the output of a dataset to protect the privacy of individual data points. The Laplace mechanism adds noise sampled from a Laplace distribution with a scale parameter that is proportional to the sensitivity of the function being computed.[30] Both the Laplace mechanism and the Exponential mechanism allow for different levels of privacy protection by adjusting the privacy parameter ε. A smaller ε value provides more privacy protection but also more noise in the output, which can decrease the accuracy of the function being computed. The sensitivity of the function being computed also plays a role in the amount of noise added, with higher sensitivity requiring more noise to be added for the same level of privacy protection. The Laplace mechanism provides differential privacy by adding noise that is proportional to the sensitivity of the function, which means that functions with low sensitivity have less noise added to them, and vice versa. However, the Laplace mechanism can result in decreased utility as the amount of noise added can be significant, particularly for datasets with low sensitivity. The Exponential mechanism is a probabilistic algorithm that outputs an item from a database that maximizes the utility of a function while providing differential privacy. The algorithm chooses an item from the database that has a high probability of being the one that maximizes the utility of the function, with a probability that is proportional to the difference between the utility of the item and the maximum possible utility. The algorithm of the Exponential mechanism can be described as follows: The output of the Exponential mechanism is the item d that is selected. The Exponential mechanism provides differential privacy by adding randomness to the selection process of the item, which makes it difficult for an attacker to determine which item was selected. The amount of noise added is proportional to the privacy parameter ε and the sensitivity Δf of the function. The Exponential mechanism can provide high utility for datasets with low sensitivity and can be more efficient than the Laplace mechanism.[31]

5. Result and Discussion

This Application loads a dataset from a CSV file, applies Laplace and exponential mechanisms to the 'first_name', 'Last_name', 'mobile_number', and 'USN' columns of the dataset, and then prints the modified dataset. The Laplace mechanism adds Laplace noise to the length of the string representation of the 'first_name' and 'Last_name' columns. It uses an epsilon value of 1 for both columns. The exponential mechanism adds exponential noise to the 'mobile_number' and 'USN' columns of the dataset. It uses an epsilon value of 0.5 for both columns. The output of the code will be the modified dataset with noisy values for the 'first_name', 'Last_name', 'mobile_number', and 'USN' columns. The degree of noise will depend on the values of epsilon used for each mechanism. The noisy values should be privacy-preserving, i.e., they should not reveal sensitive information about individuals in the dataset while still providing accurate statistical information. in the following table.1, 2, and graphs the result is shown.

USN	first_name	Last_name	Age	gender	mobile_number	course	year	GPA	City	country
1001	naeem	kanyar	26	male	6936952884	MSCIT	2021	7	bengluru	India
1002	immanuel	jones	29	male	6936952884	MSCIT	2019	8	Thimpu	bhutan
1003	manasvi	kunt	35	female	3498563456	MCA	2015	6	tehran	iran
1004	hasti	gohel	39	female	4556677834	BCA	2009	8	london	UK
1005	krunal	patoliya	25	male	3445677854	BBA	2018	6	los Angele	USA
1006	nafas	saeedi	27	female	4556677898	B.COM	2014	4	berlin	Germany
1007	ishita	gohel	24	female	7654789065	M.COM	2022	8	Beijing	China
1008	ahmad	ahmadi	23	male	3456874534	CSE	2021	7	Herat	Afghanistan
1009	hastu	gohel	37	female	5678894576	BBA	2008	5	jammu	India
1010	ali	ahmadi	31	male	3456678899	M.COM	2011	8	mashhad	iran
1011	sara	sarwari	30	female	4512567857	CSE	2015	4	Beijing	China
1012	riddhi	kaur	26	female	5435678945	MSCIT	2017	7	kabul	Afghanistan
1013	shaheed	salami	38	male	3456789056	BCA	2001	9	Thimpu	bhutan
1014	unnati	singh	32	female	4346789675	MSCIT	2005	5	london	UK
1015	ishan	omari	23	male	6745789056	B.COM	2020	8	los Angele	USA
1016	ciyona	bevin	21	female	4556677845	BCA	2022	7	jammu	India
1017	navdeep	kaur	36	female	4509306783	CSE	2012	6	tehran	iran
1018	owen	singh	20	male	4457217637	BBA	2021	8	kabul	Afghanistan
1019	naeem	kanyar	26	male	4405128492	MSCIT	2021	7	bengluru	India
1020	immanuel	jones	29	male	4353039346	MSCIT	2019	8	Thimpu	bhutan
1021	manasvi	kunt	35	female	4300950201	MCA	2015	6	tehran	iran
1022	hasti	gohel	39	female	4248861055	BCA	2009	8	london	UK
1023	krunal	patoliya	25	male	4196771910	BBA	2018	6	los Angele	USA
1024	nafas	saeedi	27	female	4144682764	B.COM	2014	4	berlin	Germany
1025	ishita	gohel	24	female	[illegible]	M.COM	[illegible]	8	Beijing	China

Table.1 Original Dataset

According to the table.1 the original Dataset consists of the student information that is used in the implementation of the actual application program, the dataset includes two types of data information, sensitive information and non-sensitive

USN	first_name	Last_name	Age	gender	mobile_number	course	year	GPA	City	country
1007	nae	kany	26	male	-6116878907	MSCIT	2021	7	bengluru	India
1039	immanue	jo	29	male	1383958051	MSCIT	2019	8	Thimpu	bhutan
1113	manasvi	kun	35	female	3988415327	MCA	2015	6	tehran	iran
1193	h	goh	39	female	342175139.7	BCA	2009	8	london	UK
1209	kruna	patoliy	25	male	-6012700616	BBA	2018	6	los Angeles	USA
1057	nafas00000000	saeedi	27	female	-2835262738	B.COM	2014	4	berlin	Germany
1058	ish	go	24	female	-2574817010	M.COM	2022	8	Beijing	China
1032	ahm	ahmadi	23	male	2165295234	CSE	2021	7	Herat	Afghanistan
1107	has	gohel00	37	female	-1428855809	BBA	2008	5	jammu	India
1163	a	ahmad	31	male	3498563456	M.COM	2011	8	mashhad	iran
1237	sa	sarwari00	30	female	3456874534	CSE	2015	4	Beijing	China
1099	riddhi	kaur000	26	female	2686186689	MSCIT	2017	7	kabul	Afghanistan
1108	shaheed00000	sa	38	male	-1741390682	BCA	2001	9	Thimpu	bhutan
1221	unnati00	singh	32	female	-8096266437	MSCIT	2005	5	london	UK
1232		omari	23	male	-334983752.3	B.COM	2020	8	los Angeles	USA
1028	ciyo	be	21	female	-5387630869	BCA	2022	7	jammu	India
1176	navdeep00	kaur0000	36	female	4509306783	CSE	2012	6	tehran	iran
1253	ow	si	20	male	-2470638719	BBA	2021	8	kabul	Afghanistan
1125	nae	kanyar000	26	male	-9138049348	MSCIT	2021	7	bengluru	India
1166	immanuel	jones0	29	male	3988415327	MSCIT	2019	8	Thimpu	bhutan
1219	manasvi0	kunt	35	female	-7367018400	MCA	2015	6	tehran	iran
1108	has	gohel000	39	female	-387072897.8	BCA	2009	8	london	UK
1255	krunal00	patol	25	male	-9346405930	BBA	2018	6	los Angeles	USA
1181	nafas	saeedi000	27	female	1592314633	B.COM	2014	4	berlin	Germany
1083	ishita	gohel00	24	female	6936952884	M.COM	2022	8	Beijing	China

Table.2 Noisy Dataset.

According to the table.2, it is showing the output of the program differential Privacy adding noise, that as it is showing the sensitive column in the above dataset which are the USN number, First Name, Last Name, and Contact number.

Epsilon is a parameter used in differential privacy to control the amount of privacy protection that is provided to individuals in a dataset. In the code, epsilon is used in two different mechanisms: The Laplace mechanism and the exponential mechanism. In the Laplace mechanism, epsilon is used to determine the amount of noise added to each string's length. A larger value of epsilon results in more noise being added, which provides stronger privacy protection and reduces the data's accuracy. In the exponential mechanism, epsilon is used to determine the level of differential privacy that is provided to each individual record in the dataset. A larger epsilon value provides less privacy protection and results in more accurate data

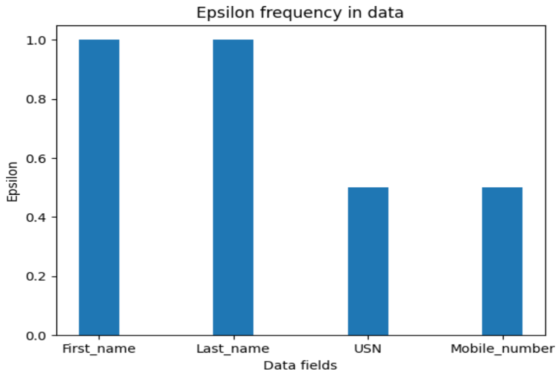

Figure. 4 Epsilon Frequency in Dataset

According to *Figure.4* overall, epsilon is a key parameter in differential privacy that balances the trade-off between privacy protection and data accuracy. A smaller value of epsilon provides stronger privacy protection but may result in less accurate data, while a larger value of epsilon provides less privacy protection but more accurate data.

6. Conclusion

This work is focused on solving the issue of privacy protection in the customized recommendation. The conflict between the rising demand for privacy protection and the damage that personalized recommendation technology does to people's personal privacy data continues to be a difficult problem for modern society. This work makes two contributions. First, this describes a sanitization methodology that adheres to an open-world concept in which the sanitizer may not be aware of the relationships that the attacker will use. Therefore, presuming the attacker can access information outside of the data set. Second, this work does not claim that there are always effective sanitization techniques. Instead, it provides information that the sanitizer can utilize to evaluate risk by recording those relationships that lead to the sanitization being reversed. In other words, the possibility that an adversary may desensitize the data is reduced to a question about the likelihood of the adversary in the estimation of the sanitizer (and other interested parties). Prior to choosing how to sanitize data, it is crucial to decide what data needs to be cleaned up. Sanitization aims to stop an adversary from drawing undesirable conclusions from the sanitized data, including the ability to extract the original, raw data that corresponds to the sanitized data. Desensitization approaches, as demonstrated above, take advantage of connections between data, either inside the sanitized data set or between the sanitized data set and outside sources of knowledge. Our study clarifies these connections and incorporates them into the choice of what should be sanitized.

7. References

[1] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. M. Chow, “Differential Privacy for Text Analytics via Natural Text Sanitization,” *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3853–3866, Jun. 2021, doi: 10.48550/arxiv.2106.01221.

[2] A. N and C. Obimbo, “Privacy-Preserving Data Publishing: A Classification Perspective,” *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 9, 2014, doi: 10.14569/IJACSA.2014.050919.

[3] P. Jain, M. Gyanchandani, and N. Khare, “Differential privacy: it is technological prescriptive using big data,” *J Big Data*, vol. 5, no. 1, pp. 1–24, Dec. 2018, doi: 10.1186/S40537-018-0124-9/TABLES/4.

[4] C. Yin, L. Shi, R. Sun, and J. Wang, “Improved collaborative filtering recommendation algorithm based on differential privacy protection,” *Journal of Supercomputing*, vol. 76, no. 7, pp. 5161–5174, Jul. 2020, doi: 10.1007/S11227-019-02751-7/METRICS.

[5] M. Hardt, K. Ligett, *T* Caltech, and F. McSherry, “A Simple and Practical Algorithm for Differentially Private Data Release.”

[6] T. Zhang, T. Zhu, R. Liu, and W. Zhou, “Correlated data in differential privacy: Definition and analysis,” *Concurr Comput*, vol. 34, no. 16, p. e6015, Jul. 2022, doi: 10.1002/CPE.6015.

[7] S. K. Sood, “A combined approach to ensure data security in cloud computing,” *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 1831–1838, Nov. 2012, doi: 10.1016/J.JNCA.2012.07.007.

[8] M. Bishop *et al.*, “Relationships and data sanitization: A study in scarlet,” *Proceedings New Security Paradigms Workshop*, pp. 151–163, 2010, doi: 10.1145/1900546.1900567.

[9] J. He, L. Cai, and X. Guan, “Differential Private Noise Adding Mechanism and Its Application on Consensus Algorithm,” in *IEEE Transactions on Signal Processing*,

2020, vol. 68, pp. 4069–4082. doi: 10.1109/TSP.2020.3006760.

[10] L. Sweeney, "Only You, Your Doctor, and Many Others May Know," *Technol Sci*, Accessed: Feb. 22, 2023. [Online]. Available: /a/2015092903/

[11] N. Carlini *et al.*, *Extracting Training Data from Large Language Models*. 2021. Accessed: Feb. 22, 2023. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/theofanos

[12] Y. Zhao *et al.*, "Local Differential Privacy based Federated Learning for Internet of Things," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.08856

[13] H. Liu, Z. Wu, Y. Zhou, C. Peng, F. Tian, and L. Lu, "Privacy-preserving monotonicity of differential privacy mechanisms," *Applied Sciences (Switzerland)*, vol. 8, no. 11, Oct. 2018, doi: 10.3390/app8112081.

[14] W. L. Croft, J.-R. Sack, and W. Shi, "Differential Privacy Via a Truncated and Normalized Laplace Mechanism," *J Comput Sci Technol*, vol. 37, no. 2, pp. 369–388, Nov. 2019, doi: 10.1007/s11390-020-0193-z.

[15] D. G. Hasuda and J. de Melo Bezerra, "Exploring Differential Privacy in Practice," in *International Conference on Enterprise Information Systems, ICEIS - Proceedings*, 2021, vol. 1, pp. 877–884. doi: 10.5220/0010440408770884.

[16] T. Zhu, D. Ye, W. Wang, W. Zhou, and P. S. Yu, "More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence," *IEEE Trans Knowl Data Eng*, vol. 34, no. 6, pp. 2824–2843, Jun. 2022, doi: 10.1109/TKDE.2020.3014246.

[17] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity," *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479, Nov. 2018, doi: 10.48550/arxiv.1811.12469.

[18] C. Wright and K. Rumsey, "The Strengths, Weaknesses, and Promise of Differential Privacy as a Privacy-Protection Framework".

[19] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014, doi: 10.1561/0400000042.

[20] "What is Differential Privacy? – A Few Thoughts on Cryptographic Engineering." https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/ (accessed Feb. 25, 2023).

[21] P. Jain, M. Gyanchandani, and N. Khare, "Differential privacy: it is technological prescriptive using big data," *J Big Data*, vol. 5, no. 1, pp. 1–24, Dec. 2018, doi: 10.1186/S40537-018-0124-9/TABLES/4.

[22] "Microsoft Corporation | Differential Privacy for Everyone Differential Privacy for Everyone," 2012. [Online]. Available: http://paulohm.com/

[23] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *J Big Data*, vol. 3, no. 1, pp. 1–25, Dec. 2016, doi: 10.1186/S40537-016-0059-Y/TABLES/5.

[24] J. M. Abowd and I. M. Schmutte, "An economic analysis of privacy protection and statistical accuracy as social choices†," *American Economic Review*, vol. 109, no. 1, pp. 171–202, Jan. 2019, doi: 10.1257/aer.20170627.

[25] "Differential Privacy Defined - Privacy and Ethics Coursera." https://www.coursera.org/lecture/data-results/differential-privacy-defined-phj4C (accessed Feb. 25, 2023).

[26] P. Jain, M. Gyanchandani, and N. Khare, "Differential privacy: its technological prescriptive using big data," *J Big Data*, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40537-018-0124-9.

[27] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in

private data analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3876 LNCS, pp. 265–284, 2006, doi: 10.1007/11681878_14/COVER.

[28] F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," pp. 94–103, Apr. 2008, doi: 10.1109/FOCS.2007.66.

[29] M. Abadi *et al.*, "Deep learning with differential privacy," *Proceedings of the ACM Conference on Computer and Communications Security*, vol. 24-28-October-2016, pp. 308–318, Oct. 2016, doi: 10.1145/2976749.2978318.

[30] A. Ghosh, T. Roughgarden, and M. Sundararajan, "UNIVERSALLY UTILITY-MAXIMIZING PRIVACY MECHANISMS *," vol. 41, no. 6, pp. 1673–1693, doi: 10.1137/09076828X.

[31] C. Dwork, A. Roth, C. Dwork, and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends R in Theoretical Computer Science*, vol. 9, pp. 211–407, 2014, doi: 10.1561/0400000042.