Journal of Big Data

SURVEY PAPER

Open Access

Differential privacy: its technological prescriptive using big data

Priyank Jain*, Manasi Gyanchandani and Nilay Khare

*Correspondence:
priyankjain1984@gmail.com
Computer Science & Engineering, MANIT, Bhopal, MP, India

Abstract

Data is being produced in large amounts and in rapid pace which is diverse in quality, hence, the term big data used. Now, big data has started to influence modern day life in almost every sphere, be it business, education or healthcare. Data being a part and parcel of everyday life, privacy has become a topic requiring emphasis. Privacy can be defined as the capacity of a person or group to seclude themselves or information about themselves, and thereby express them selectively. Privacy in big data can be achieved through various means but here the focus is on differential privacy. Differential privacy is one such field with one of the strongest mathematical guarantee and with a large scope of future development. 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 till date has been presented. The research gap and future directions have also been mentioned as part of this paper.

Keywords: Differential privacy, Big data, Big data privacy, Airavat, PINQ, Geo-indistinguishability, GUPT, Privacy budget, Sensitivity, Laplace, Exponential

Introduction

Differential privacy [1] is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals. This is done by introducing a minimum distraction in the information provided by the database system. The distraction introduced is large enough so that they protect the privacy and at the same time small enough so that the information provided to analyst is still useful. Earlier some techniques have been used to protect the privacy, but proved to be unsuccessful. In mid-90s when the Commonwealth of Massachusetts Group Insurance Commission (GIC) released the anonymous health record of its clients for research to benefit the society [2]. GIC hides some information like name, street address etc. so as to protect their privacy. Latanya Sweeney (then a Ph.D. student in MIT) using the publicly available voter database and database released by GIC, successfully identified the health record by just comparing and co-relating them. Thus hiding some information cannot assures the protection of individual identity. With the advent of big data, privacy

Springer Open

has been a topic of utmost concern. Data is being collected from all sources and traded and there are few effective controls over how it is used or secured. Privacy is a term that incorporates varied components including physical privacy, communications privacy, and information privacy.

This paper presents the basics of differential privacy as a privacy preserving mechanism [3, 4] for big data. Differential privacy (DP) was considered to manage protection dangers to avert undesirable re-distinguishing proof and other security dangers to people whose individual data is available in big datasets, while giving helpful access to information. DP tends to the puzzle of adapting nothing around an individual while learning profitable data about a people. Based on the survey of privacy aspects in big data [5], few of the privacy preserving mechanisms can be listed down as in Table 1.

"Differential privacy" section gives the definition and an overview of differential privacy. "Basic terms of differentially privacy" section represents basic terms of differential privacy. "Mechanisms used in differential privacy" section consists of the noise mechanisms and properties of DP. "Approaches to achieve differential privacy" section provides the approaches to achieving DP. "Approaches to achieve differential privacy" section also puts forward a perspective of big data in terms DP and summarizes the implementations different methods of DP. "Differential privacy and big data" section represents differential privacy and big data with comparative study of different differential privacy mechanism. "Research gap, Challenges, Conclusion and future work" section consists of the research gap, also provides the challenges and future work prospects & ends the paper with the conclusion.

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 years [6, 7]. Generally, DP works by embedding's a go-between bit of programming between the examiner and the database [1].

Figure 1 shows the differential privacy mechanism [8]. The analyst sends a query to an intermediate piece of software, the Privacy guard. The guard assesses the privacy

Table 1 Privacy mechanisms in big data

Name	Approach	Protects from adversial	Focus
HybrEx [44]	Data separation	Cloud	Only does partitioning as a measure to handle privacy Does not deal with key generation in map phase Does not deal with adversarial users
EPiC [45]	Homomorphic encryption	Cloud	User does not trust cloud infrastructure Supports counting operation only Purpose specific
Secret-shared [46, 47]	Secret sharing	User and cloud	Cost overhead
Airavat [22]	MAC + differential privacy	User	Full trust on cloud providers Cannot guarantee privacy for calculations which output keys created by untrusted mappers

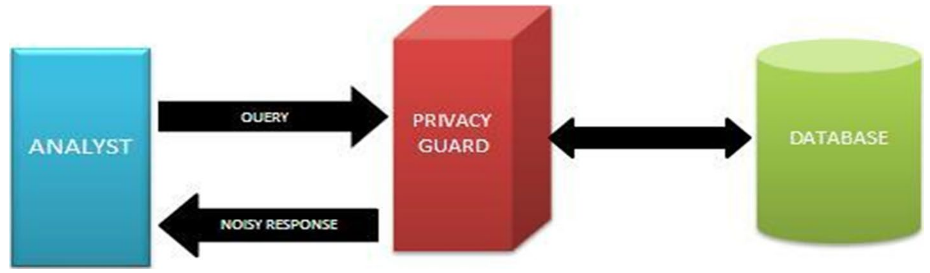

Fig. 1 Differential privacy mechanism [1]

impact of the query by making use of a special algorithm. Then, the query is sent to the database by the guard getting back a clean answer based on data that has not been distorted in any way. The guard then adds the appropriate amount of "noise," scaled to the privacy impact, accordingly making the answer uncertain to ensure the confidentiality of the individuals whose information is in the database, and sends the modified response back to the analyst.

Formally speaking let D_1, D_2 be two datasets. D_1 and D_2 in Eq. 1 are said to be neighbors if they differ in at most one data entry. An algorithm M is ε-differentially private if for all pairs of neighboring datasets D_1, D_2 and all outputs x,

$$\Pr[M(D_1) = x] \leq \exp(\varepsilon) \Pr[M(D_2) = x] \tag{1}$$

i.e. given the output of the computation, one cannot tell if any specific data item was used as part of the input because the probability of producing this output would have been the same even without that item. Not being able to tell whether the item was used at all in the computation precludes learning any useful information about it from the computation's output alone. The computation of the function must be randomized to achieve privacy.

To better understand the concept of differential privacy, we can take an example as shown in Fig. 2.

As we can observe from the figure above, with differential privacy, one cannot learn more about an individual (Person C) whether she is in or not in the database.

One of the major challenges is protecting data from unauthorized access and minimizing data leakage during analysis of data. For a given computational task T and a given value of ε there will be numerous differentially private algorithms for achieving T in an ε-differentially private way. Some will have better accuracy than others. At the point when ε is small, finding a highly accurate ε-differentially private algorithm for T can be troublesome, much as finding a numerically stable algorithm for a specific computational task can require effort.

One way to deal with characterizing privacy with regards to data analysis is to require that the analyst does not know not any more about any individual in the data set after the analysis is completed than she knew before the analysis was begun.

Differential privacy [9, 10] will provide privacy by process; specifically, it will present randomness. An early case of privacy by randomized process is randomized

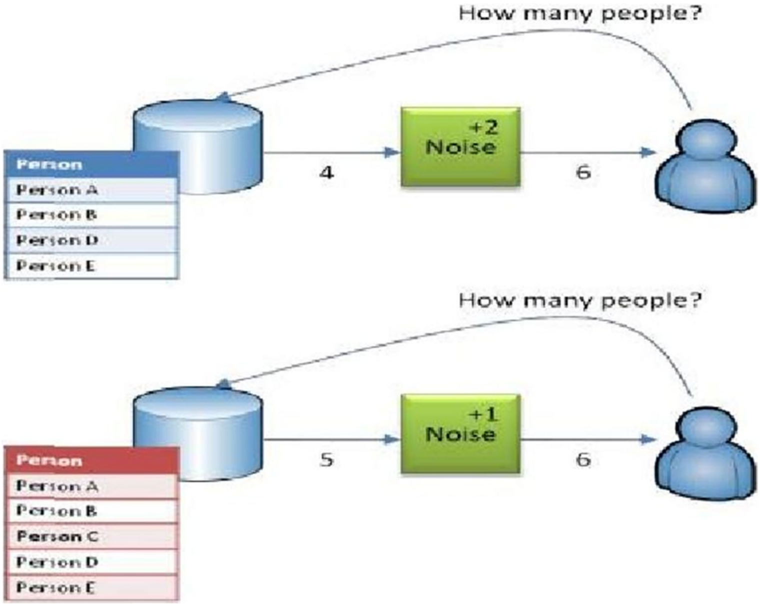

Fig. 2 Differential privacy through noise [17]

response, a strategy created in the social sciences to collect statistical information about embarrassing or unlawful conduct, caught by having a property P.

Study participants report whether they have property P as follows:

1. Flip a coin.
2. On the event of tails, then react truthfully.
3. On the event of heads occur then flip a second coin and react “Yes” if heads and “No” if tails. “Privacy” comes from the plausible deniability of any result [11].

For the most part, differential privacy does not ensure that what one believes to be one’s secrets will remain secret. It simply guarantees that one’s participation in a survey will not be disclosed, nor will the participation lead to disclosure of any specifics that one has contributed to the survey.

Basic terms of differential privacy

Privacy budget (*ε*)

It tells quantity of queries should be answered in given data due to privacy restriction. A cost for each query is deducted depending on how private it is. The parameter *ε* is called privacy budget, which is utilized to control the proportion of output of function *A* in Eq. 2 in neighboring datasets D_1 and D_2. The smaller ε is, all the closer the proportion is to 1, i.e. the output probability distributions of function *A* in neighboring datasets D_1 and D_2 are roughly the same. The higher the security becomes, the lower the utility gets. The value of *ε* is small normally, such as 0.01, 0.1, 0.5, and 0.8 [12].

$$\Pr[A(D_1 \in S] \leq e^{\in} \times \Pr[A(D_2) \in S], \tag{2}$$

Sensitivity

It tells us how much noise is to be added to the results. It depends on how much the output can change if addition or removal of a single row is done. Given a sequence of counting queries Q, D_1 and D_2 are neighboring datasets, the sensitivity of Q, denoted ΔQ in Eq. 3, is:

$$\Delta \mathrm{Q} = \max\|\mathrm{Q}(\mathrm{D}_1) - \mathrm{Q}(\mathrm{D}_2)\| \tag{3}$$

Mechanisms used in differential privacy

Noise

The two primary noise mechanisms in DP are Laplace mechanism (LM) and exponential mechanism (EM). The magnitude of noise alludes to privacy budget and global sensitivity [13–15].

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 [11].

Given a dataset D and the function f: D→R^d, global sensitivity is Δf; random algorithm in Eq. 4,

$$\mathrm{A}(D) = f(D) + noise \tag{4}$$

Satisfies ε-differential privacy if the noise complies with the Laplace distribution; that is, *noise*~Lap($\Delta f/\varepsilon$); there into, location parameter (LP) is zero and 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 [16]. Let $\sigma(\chi)$ denote standard deviation; $\mathrm{D}(\chi)$ denotes variance, and *noise*~Lap(b) in Eq. 5,

$$\sigma(\chi) = \sqrt{\mathrm{D}(\chi)},\ D(\chi) = 2b^2, \quad \text{and} \quad b = \Delta f/\varepsilon; \tag{5}$$

Then the results obtained are in Eqs. 6 and 7,

$$\mathrm{D(x)} = 2(\Delta f/\varepsilon)^2 = 2\Delta f^2/\varepsilon^2, \tag{6}$$

$$\sigma(\chi) = \sqrt{\mathrm{D}(\chi)} = \sqrt{2\Delta f^2/\varepsilon^2} = \sqrt{2\Delta f/\varepsilon} \tag{7}$$

Exponential mechanism

Exponential mechanism is another security-controlled plan to fulfil differential privacy when the outputs are non-numerical. Intuitionally, exponential mechanism still guarantees that this change of a single DB tuple does not influence the outcome of the score function [16]. The exponential mechanism was designed for circumstances in which it was wished to pick the best response. Let D denote the input dataset; r € R denotes one of the potential answers, given a score function u: $D \times R \rightarrow \mathrm{R}$; 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. 8:

$$A(D,\ u) = \{r : |\Pr[r \in R] \infty \exp(\varepsilon u(D, r)/2\Delta u)\}, \tag{8}$$

where Δu denotes the sensitivity of score function u and is defined as in Eq. 9:

$$\Delta u = \max(r \in R)\ \max((\|D\Delta D'\|) = 1)|u(D, r) - u(D', r)| \tag{9}$$

The exponential mechanism can yield non-numerical results as indicated by their values of score function. The output probability refers to 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 polynomially large in the natural parameters of the issue [11].

Combination properties

Differential privacy has two important properties:

Sequential composition (SC)

SC refers to a sequence of calculations, each providing DP in isolation, providing also DP in sequence. If there are n independent algorithm: A_1,..., A_n, whose privacy budgets are ε_1,...,ε_n, respectively, then any function K of them:

$$\mathrm{K}(\mathrm{A}_1, \ldots \mathrm{A}_n) \text{ is } \varepsilon_i - \sum_{i=1}^{n} \text{differentially private.}$$

SC demonstrates that protection spending plan and error are combined directly when various DP is used to release information for the same dataset. SC says that by social occasion differentially private information about an arrangement of people, DP isn't broken, yet rather the level of privacy diminishes.

Parallel composition (PC)

PC can be alluded to a few ε-differentially private calculations, each on information from a disjoint arrangement of subjects giving ε-differentially private yield on the information from the pooled set of subjects. In the event that the past systems are registered on disjoint subsets of the private dataset then the capacity K would be max {i}-differentially private. PC exhibits that the level of protection ensure relies on the estimation of ε_i; when the estimation of ε_i grows, the security diminishes step by step. Parallel piece can be repeated as: The arrival of ε-differentially private informational indexes D_i, where D_i alluding to disjoint arrangements of people, for some i ∈ I, is ε-differentially private. The said organization properties hold for differentially private inquiry replies and differentially private datasets. Along these lines, to adjust the security and the utility of the information, it is required to be considered as on the best way to meet the greatest number of questions utilizing the littlest spending plan before depleting ε or how to enhance the exactness of inquiry under given ε.

DP example on dataset application

In order to see the effect of differential privacy on a dataset, consider its use against a publicly available healthcare dataset, in this case a dataset containing the intraocular pressure for 238 patients visiting an ocular disease clinic. A common way to view the dataset is in terms of intraocular pressure (IOP) ranges as noted in Table 2 [17]. When applying differential privacy with an epsilon value of $\varepsilon = 0.1$, an example of the result set that might be returned is shown in Table 3. Table 3 is acquired by applying differential privacy according to PINQ [18], an implementation of DP.

Table 2 shows the actual values of the number of patients in the column Number of patients (real value) whereas in number of patients (differential privacy applied) column the values that would be answered in case a query about the number of patients in made. This is only one of many possible results; since differential privacy by PINQ is the application of random noise, the results can be different every time. The application of this random noise can result in wide ranges of variation. The higher the epsilon (ε) value, the smaller the degree of variation and the more accurate the values are to the real value.

Approaches to achieve differential privacy

There are two principle ways to produce differentially private data sets [19]:

(i) Create a manufactured informational collection from a ε-differentially private model for the information (typically from a differentially private histogram), or

(ii) Add noise to cover the estimations of the first records (most likely in mix with some earlier conglomeration capacity to decrease the measure of required commotion).

Table 2 Intraocular pressure ranges ranges based on the Woolson dataset (Woolson) with the differential privacy applied at ε = 0.1 [17]

IOP	Number of patients (real value)	Number of patients (differential privacy applied)
08–09	1	− 1.089174486
10–11	2	0.332053694
12–13	17	16.71256706
14–15	20	18.67770913
16–17	43	62.75784853
18–19	63	79.93521335
20–21	57	44.60007283
22–23	23	28.04627043
24–25	7	− 14.62243625
26–27	2	15.67088707
28–29	0	19.38616653
30–31	2	9.491168864

Table 3 Advantages and drawbacks of DP mechanisms

Mechanism	Advantages	Drawbacks
PINQ [18]	First platform providing differential privacy guarantees Expands the set of capable users of sensitive data, increases the portability of privacy-preserving algorithms across data sets and domains, and broadens the scope of the analysis of sensitive data	Does not consider the application developer to be an adversary Subjected to a weaker privacy constraint. Hence, vulnerable to state attack, privacy budget attack and timing attacks It further requires the developers to rewrite the application to make use of the PINQ primitives
Airavat [22]	First system that integrates mandatory access control with differential privacy, enabling many privacy-preserving MapReduce computations without the need to audit untrusted code Can be deployed in large scale distribution, without the need of rewriting existing MapReduce applications	Cannot confine every computation performed by untrusted code. Only considers the map program to be an "untrusted" computation while the reduce program is "trusted" to be implemented in a differentially private manner Supports only limited Reducer functions Vulnerable to state attack and timing attacks
GUPT [26]	Uses the aging model of data sensitivity, to allow analysts to describe the abstract 'privacy budget' in terms of expected accuracy of the final output GUPT automatically allocates a privacy budget to each query in order to match the data analysts' accuracy requirements Defends against side channel attacks such as the privacy budget attacks, state attacks and timing attacks	GUPT assumes that the output dimensions are known in advance. This may however not always be true in practice Inherits limitations of differential privacy regarding splitting og privacy budget
Geo-indistinguishability [25]	Proposes a generalized notion of differential privacy instantiated with the Euclidean metric which can be naturally applied to location privacy Offers the best privacy guarantees for the same utility, among all those which do not depend on the prior knowledge of the adversary, i.e., the mechanism is designed once and for all and it is applicable also when we do not know the prior	Linear degradation of the user's privacy that limits use of the mechanism over time The level of noise of the Laplacian mechanism has to be fixed in advance independently of the movements of the user Despite achieving the flexible behavior, the tiled mechanism would not satisfy geo-indistinguishability to its full potential
Telco big data [12]	First attempt to implement three basic DP architectures in the deployed telecommunication (telco) big data platform for data mining applications Proposed with the observation that the accuracy loss increases by increasing the variety of features, but decreases by increasing the volume of training data	The privacy of people in the training data is protected, but the privacy of people in the prediction data (that is, the data which will be applied to the trained model to) is not Design of adjustable privacy budget assignment strategies is required for better accuracy along with privacy guarantee
e-Health data release [21]	Improves the performance of the previous work by designing a new private partition algorithm of histogram and also proposing a heuristic hierarchical query method Real experiments were conducted and the schemes compared with the existing one to show that the proposal is more efficient in terms of data processing and updating Increase of the accuracy of data release through consistency and gives a proof of privacy to show that the proposed algorithm is under differential privacy	Data release issues under differential privacy, such as real time monitoring and publishing of e-health data is proposed

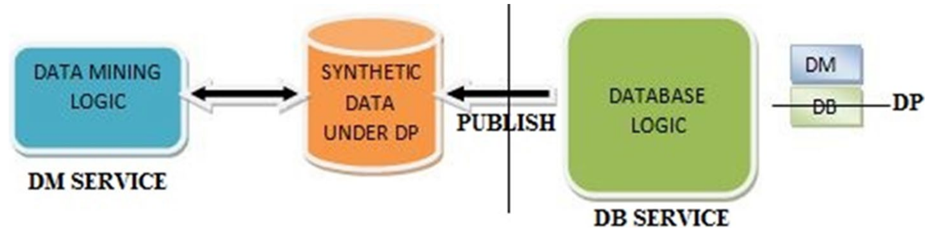

Fig. 3 Data publication architecture [12]

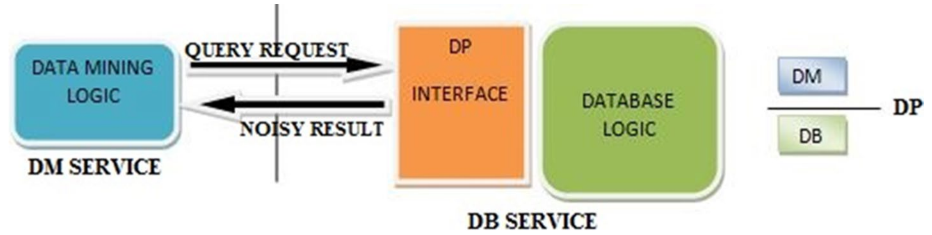

Fig. 4 Separated architecture [12]

On the basis of above two principles there are following approaches to achieve differential privacy:

Differential privacy in Telco big data platform [12]

Differential privacy is a privacy definition which guarantees the result of any authenticated query/calculation to be insensitive to any individual record in the database. Differential privacy is favored because of its strong mathematical boundary of the leaked privacy. The implementations of DP can be broadly categorized into three fundamental architectures:

1. Data publication architecture;
2. Separated architecture; and
3. Hybridized architecture [12].

(i) *Data publication architecture* In the data publication architecture as appeared in Fig. 3, the Database service utilizes a schema to publish a synthetic dataset with the DP guarantee from the real original dataset. Along these lines, the differential privacy interface is executed inside the database amongst original and synthetic datasets. Since the synthetic dataset is privacy insensitive, any data mining [20] service can be directly applied on top of the published and protected synthetic dataset.

The advantage of this architecture is that all data mining algorithms can be utilized without privacy concerns.

The weakness is that data mining service runs on the synthetic dataset instead of the real/original one, so that the mining quality is truly limited by the schema of generating synthetic dataset under the differential privacy guarantee.

(ii) *Separated architecture* [12] The separated architecture as in Fig. 4 is an implementation that isolates the database service from the data mining service through a differentially private interface. The database provides the query interface, which supports the traditional aggregation queries (like the counting queries) with the differential privacy guarantee. The database service has no clue about how the data mining service will use the results of these queries.

The advantage of this framework is that the traditional database structure does not require any change to support specific data mining services. Since the data mining services are particularly intended to utilize these query results, the system accuracy is expected to be higher than the data publication architecture.

The data mining services being on the top of aggregation queries, cannot be implemented and optimized beyond the scope of traditional queries. This may bring about some design limitations of the data mining services and lead to some accuracy loss.

(iii) *Hybridized architecture* [12] Figure 5 shows the Hybridized architecture. It adapts only the differential privacy interface into data mining services. In this situation, the database service is designed to support some specific queries for specific DM services.

The benefit of this architecture is the differential privacy implementation is optimized for a specific data mining method. So, the accuracy of the data mining is expected to be the highest among the three basic architectures.

The shortcoming is that the logics of both data mining and database services depend closely. The database developers must handle extra types of queries for specific data mining services, which are different from the traditional ones supported by database services.

Efficient e-health data release with consistency guarantee under differential privacy [21]

Another mechanism to implement DP is to apply differential privacy after data release. Data can be released in the form of histogram, tree structure, time series, graph or pattern mining. So, one of the mechanisms presented here can be to apply DP on histogram data release [21]. The steps include:

Step 1: User generates a query sequence Q and submits it to the private database D.

Step 2: Private database receives Q from the user. The query Q could be a set of range queries that asks statistical information of the dataset. Database firstly maps

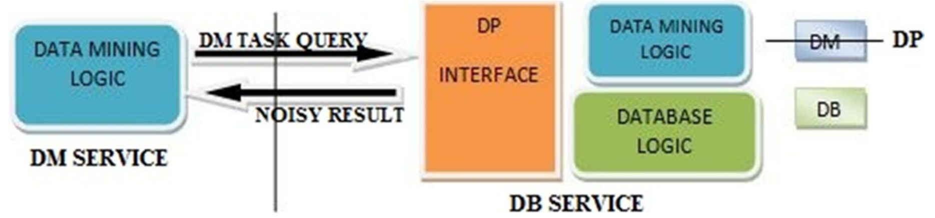

Fig. 5 Hybridized architecture [12]

its dataset into a histogram and sends it to differential privacy mechanism together with Q.

Step 3: Differential privacy mechanism first divides the histogram into adjacent but disjoint buckets. Each bucket consists of several bins. This process is entitled as partition stage or stage 1. The algorithm designed for stage 1 should comply with ε_1-differential privacy. After stage 1, each bin's count is replaced by the average count of its corresponding bucket. Then, differential privacy mechanism answers Q through the new divided histogram and puts results into a non-negative integral vector $q = Q(D)$. Let Y denote a vector, where each element is sampled from a Laplace distribution with scale $\Delta Q/\varepsilon 2$. Random vector Y has the same length as that of Q. As can be seen in Fig. 6, add Y to q, and have the noisy output denoted as p. The above operations are entitled as stage 2, where ε_2 is the privacy budget of this stage. Moreover, p is submitted to post processing mechanism.

Step 4: Stage 2 leaves an issue of query result inconsistency. For example, the result of a larger range query should be equal to its sub-intervals. However, this numerical relationship may be broken when Y is added. Post-processing mechanism is aimed to solve the problem and transforms noisy answer pinto consistent answer p′. Stage 3 is used to represent the above mechanism. Without processing the original data q, Stage 3 consumes no privacy budget. User receives query result p′ from the post-processing mechanism. Per the composition property of differential privacy, the total privacy budget of the scheme is $(\varepsilon_1 + \varepsilon_2)$.

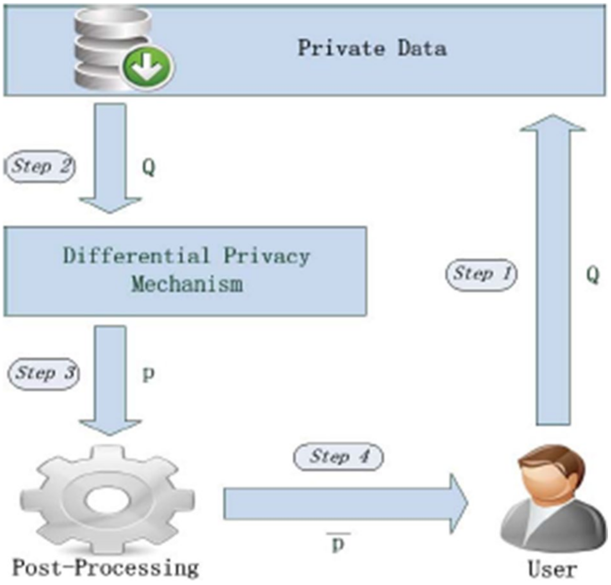

Fig. 6 Efficient e-health data release system model [21]

Airavat model: security and privacy for Map Reduce [22]

One more mechanism to give an end to end privacy and security guarantee can be to implement DP with access control mechanism such as mandatory access control (MAC) [22].

It basically consists of three components:

1. *Data provider* Data provider sets several privacy parameters (e.g. privacy budget, etc.) as well as different access control labels for their own data.
2. *Computation provider* This entity might attempt to access input values, intermediate values, or even the final output through malicious data mining algorithms written in the framework.
3. Computation framework as in Fig. 7.

Such kind of framework ensures access restriction along with privacy guarantee through differential privacy.

Location based privacy based on differential privacy [23, 24]

Location based services are the ones that are used to find places. In terms of differential privacy, it would mean that the exact location is not revealed but assuming that revealing an approximate location is fine. The principle of geo-indistinguishability is the same being a formal notion of privacy that protects the user's exact location, while allowing approximate information.

(i) The notion of geo-indistinguishability, which is a property similar to that of differential privacy [25] is based on the idea to obfuscate the real location by reporting an approximate one, using some random noise. The idea is that from the reported location, the attacker may be able to make a good guess of the area where the user is actually located, but it should not be able to make a good guess of the exact location of the user within this area.

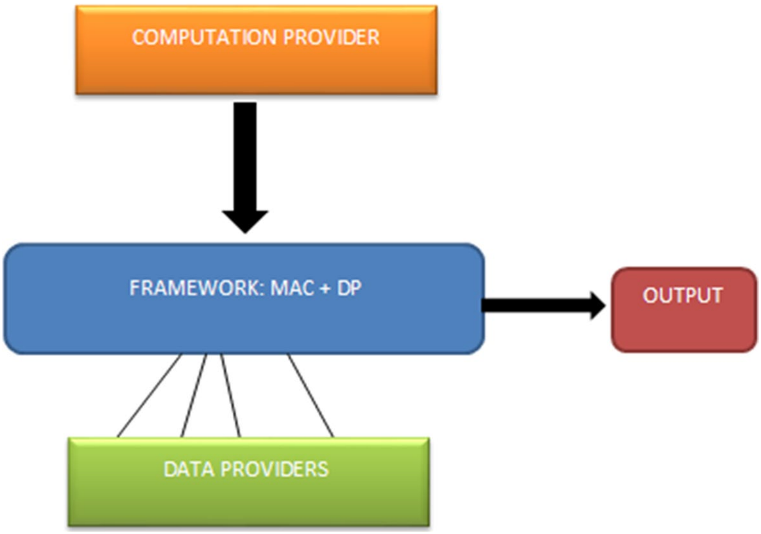

Fig. 7 Framework for access control and DP mechanism

This mechanism can be implemented by using a noise with a Laplacian distribution that is a negative exponential with respect to the distance from the real location, like in the case of differential privacy. This method provides a good level of robustness with respect to composition of attacks, in that the level of privacy decreases in a controlled way (linearly). Geo-indistinguishability, which guarantees that the user's location is protected, within a radius r, with a level of noise that decreases with r, at a rate that depends on the desired level of privacy. Intuitively, this means that the real location is highly indistinguishable from the locations that are close, and gradually more distinguishable from those that are far away. It is a particular instance of d-privacy, an extension of differential privacy to arbitrary metric domains, obtained by replacing the Hamming distance, implicit in the definition of differential privacy with the intended distance namely the geographical distance in our case. Like differential privacy, geo-indistinguishability is independent from the side knowledge of the adversary and robust with respect to composition of attacks. Location Guard is an open source web browser extension based on geo-indistinguishability that provides location privacy. Implementation of geo-indistinguishability by adding random noise drawn from a planar Laplace distribution is shown in Fig. 8.

(ii) Geo-indistinguishability and its current implementation Location Guard are just a preliminary approach to location privacy, and they present two main limitations:

First, when used repeatedly, there is a linear degradation of the user's privacy that limits the use of the mechanism over time.

Second, the level of noise of the Laplacian mechanism has to be fixed in advance independently of the movements of the user, providing the same protection in areas with very different privacy characteristic, like a dense city or a sparse countryside. This limits the flexibility of the mechanism over space.

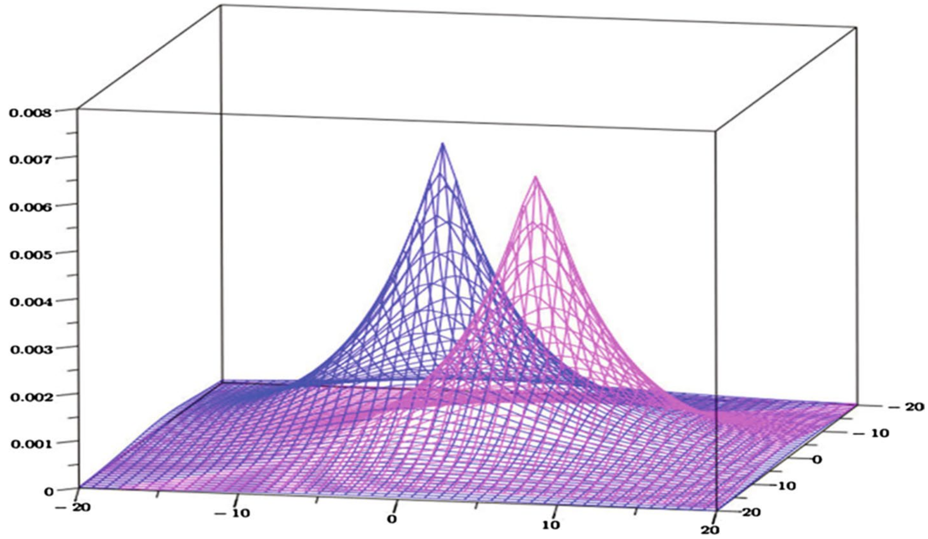

Fig. 8 Probability density functions of two planar Laplacians, centered on the (real) locations (− 2, − 4) and (5, 3), respectively [23]

As such, extensions were proposed along with a scope for future work:

Geographic fences

Areas around highly recurrent locations where the mechanism reports uniformly, effectively stopping the privacy erosion. On one side the user has to release publicly the position of her fences but on the other the budget cost when reporting from inside them is zero, leading to a practical solution that can be used in combination with the predictive mechanism.

In Fig. 9 we can see an example of fence introduced in an elastic metric. On the left there is the distribution of reported locations inside the fence that is perfectly uniform, covering a few blocks and proving an adequate level of privacy while costing zero on the budget. On the right the distribution of reported locations can be seen of a point right outside, the fence is clearly visible and the mechanism reports right around it.

Tiled mechanism

An elastic mechanism is the one that adapts the level of noise to the semantic characteristics of each location, such as population and presence of position of interests. A lighter version of the elastic mechanism is proposed wherein instead of adapting the noise differently in locations tens of meters apart, it can only adapt to large areas of a city, covering tens of square kilometers. These areas, that are called tiles, are small enough to distinguish a park from a residential area, but still easily computable. In order to build the set of tiles, two online geographical services are queried, overpass-turbo and dbpedia to obtain a set of polygons together with a quantitative description of the amount of privacy they provide. This dataset should cover an area large enough to contain most of the user usual movement and it can easily reach a few tens of kilometers while retaining a small size. Once this small dataset is build, a mapping from tiles to their privacy mass is got which is used to define a function l that, for each location, finds the containing polygon and returns a privacy level adapted to the privacy mass provided by the tile. Examples of the kind of maps that we aim at obtaining with this method are shown in Fig. 10.

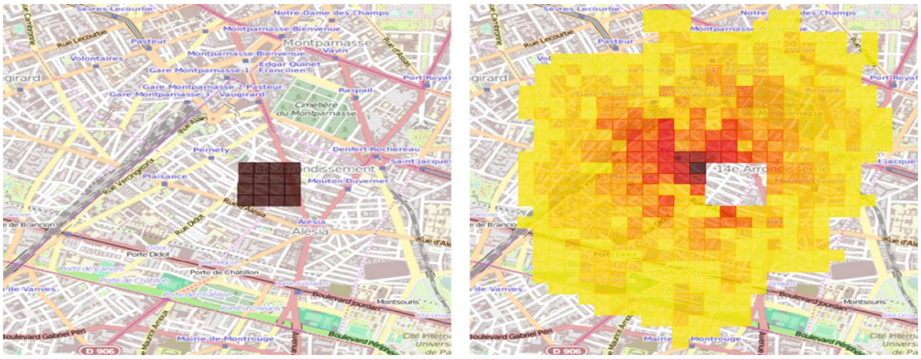

Fig. 9 Probability distribution of reported location inside and outside the fence [23]

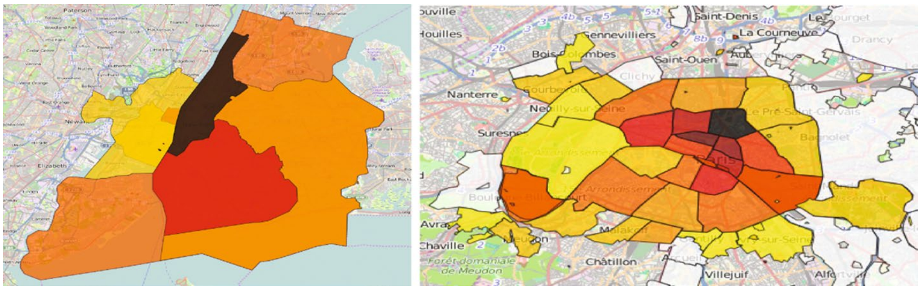

Fig. 10 Polygons computed for New York and Paris [23]

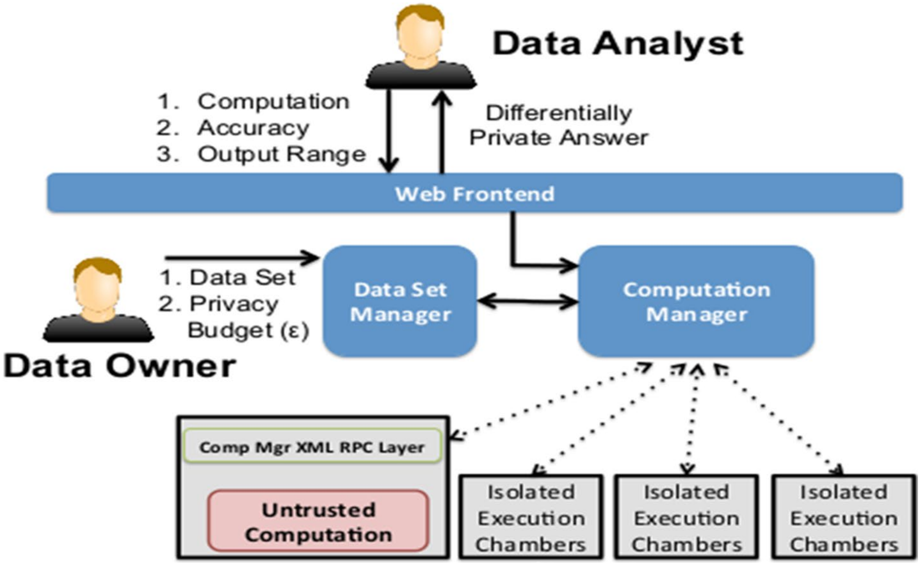

Fig. 11 Overview of GUPT's architecture [26]

GUPT mechanism for differential privacy [26]

GUPT is a system that guarantees differential privacy to programs not developed with privacy in mind, makes no trust assumptions about the analysis program, and is secure to all known classes of side-channel attacks. GUPT uses a new model of data sensitivity that degrades privacy of data over time. This enables efficient allocation of different levels of privacy for different user applications while guaranteeing an overall constant level of privacy and maximizing the utility of each application. GUPT also introduces techniques that improve the accuracy of output while achieving the same level of privacy. These approaches enable GUPT to easily execute a wide variety of data analysis programs while providing both utility and privacy.

There are three logical parties:

1. The analyst/programmer, who wishes to perform aggregate data analytics over sensitive datasets.

2. The data owner, who owns one or more datasets, and would like to allow analysts to perform data analytics over the datasets without compromising the privacy of users in the dataset.
3. The service provider, who hosts the GUPT service.

The data owner and the service provider are assumed to be trusted whereas the analyst is untrusted. The approach is as shown in Fig. 11.

The dataset manager is a database that registers instances of the available datasets and maintains the available privacy budget.

The computation manager instantiates computations and seamlessly pipes data from the dataset to the appropriate instances. The isolated execution chambers isolate and prevent any malicious behavior by the computation instances.

PINQ [18]

"Privacy Integrated Queries" is a trustworthy platform for privacy-preserving data analysis. PINQ provides private access to arbitrarily sensitive data, without requiring privacy expertise of analysts or providers. The interface and behavior are very much like that of Language Integrated Queries (LINQ), and the privacy guarantees are the unconditional guarantees of differential privacy.

PINQ is a software application framework as shown in Fig. 12, that applies the mathematics of differential privacy. It is built on top of LINQ which is a .NET framework that allows developers and IT professionals to query any data source using the same methods. This allows users to apply differential privacy without understanding the mathematics behind it. For example, with LINQ .NET framework method can be written to ask a question from a database and it will provide with the answer. With PINQ, the same question is asked but it will give an answer with differential privacy applied, meaning that there will be some exponential noise added to results. How much noise is used depends on the epsilon value, ε, which the user provides to PINQ. The larger the epsilon value, the less noise is applied to the result set, which means it is more accurate but potentially more privacy revealing. Conversely, the smaller the epsilon value, the more noise is applied to the result set which means there is more privacy but less accuracy [17].

PINQ provides a restricted programming language with a small number of trusted, primitive data operations in the LINQ framework. PINQ employs a request/reply model, which avoids adding noise to the intermediate results of the computation by keeping them on a trusted data server or an abstraction of a trusted data server provided by a distributed system. PINQ provides a thin protective layer in front of existing data sources, presenting an interface that appears to be that of the raw data itself. Figure 11 shows the control/data flow in PINQ. Here, an analyst initiates a request to a PINQ object, whose agent (A) confirms, recursively, differentially private access. Once approved by the providers' agents, data (D) flows back through trusted code ensuring the appropriate level of differential privacy.

PINQ provides language level guarantees. PINQ's contribution is not only that one can write private programs, but that one can write only private programs.

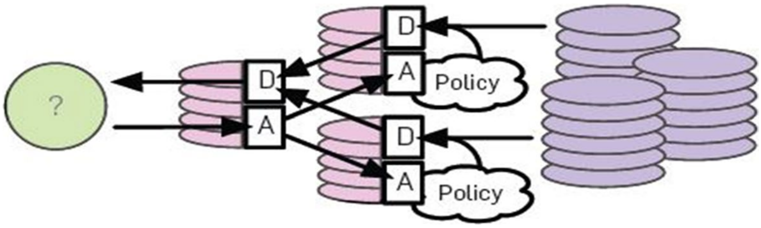

Fig. 12 PINQ request/reply model [18]

Advantages and drawbacks of different differential privacy mechanisms

The advantages along with the drawbacks of the different differential privacy mechanisms above can be formulated as in Table 3.

Differential privacy and big data

DP is a worldview that concerns nothing about the sensitive information except for the component of the presence of the information. The primary reason for DP stays to give security saving responses to inquiries performed on the original information. Any user who adds a computation to a larger pool of aggregate data should not be revealed as the source, let alone what data they contributed. In terms of big data [27–29], differential privacy has been implemented in various scenarios. Airavat [22] uses the MapReduce paradigm for DP implementation. However, airavat itself consists of potential drawbacks and not properly secure. So, airavat can be taken as the first step in implementing differential privacy through mapreduce for big data keeping its considerations and principles in view. But MapReduce implementation of DP requires more work to be done. In the GUPT [26] mechanism, points regarding the block size that impacts the noise and accuracy were mentioned. It was observed that increase in block size leads to increase in noise and decrease in estimation error. So, the block size needs to be optimally chosen. The block size can be compared to the mapper in case of MapReduce. Using the observations, the number of mappers can be optimally chosen to see the impact it has on the overall output. As in [16], the data release mechanisms fulfilling differential privacy needs to be evaluated according to the dataset, its purpose in providing privacy and the phase of big data in which it is to be implemented. The basic query operations and the noise added to guarantee privacy with the optimal parameters being chosen are to be referred from the implementations of DP [18] till date as almost every mechanism has presented a whole new perspective of the DP scenario along with research gaps in the concerned area. Location based services is also an emerging field in differential privacy implementation. Semantic data is required for accurate results but also the mechanism needs to be flexible in the presence of large amount of data. Position and size of an area in accordance to the user and their point of interests must be fulfilling privacy [23]. Gathering this large scale of data and then processing on it to give the best results is still work in progress.

Apple's case

Apple has implemented differential privacy in its iOS 10. Apple can collect and store its users' data in a format that lets it glean useful notions about what people do, say, like and want. But it can't extract anything about a single, specific one of those people that might represent a privacy violation. And neither, in theory, could hackers or intelligence agencies. Apple is actually sending more of data than ever off of an individual's device to its servers for analysis, just as Google and Facebook and every other data-hungry tech firm does but apple is only transmitting that data in a transformed, differentially private form. The method used has been mentioned as three transformations: Hashing, a cryptographic function that irreversibly turns data into a unique string of random looking characters; subsampling, or taking only a portion of the data; and noise injection, adding random data that obscures the real, sensitive personal information [30]. The points below describes the mechanism in detail:

- Hashing takes a string of text and turns it into a shorter value with a fixed length and mixes these keys up into irreversibly random strings of unique characters or "hash". This obscures data so the device isn't storing any of it in its original form.
- Subsampling means that instead of collecting every word a person types, Apple will only use a smaller sample of them. For example, if a person has a long text conversation with a friend liberally using emoji, instead of collecting that entire conversation, subsampling might instead use only the parts Apple is interested in, such as the emoji.
- Finally, the device injects noise, adding random data into the original dataset in order to make it vaguer. This means that Apple gets a result that has been masked ever so slightly and therefore isn't quite exact.
- All this happens on the device, so it has already been shortened, mixed up, sampled, and blurred before it is even sent to the cloud for Apple to analyze.

Apple is only using differential privacy in four specific areas:

- When enough people replace a word with a particular emoji, it will become a suggestion for everyone.
- When new words are added to enough local dictionaries to be considered commonplace, Apple will add it to everyone else's dictionary too.
- A search term can be used in Spotlight, and it will then provide app suggestions and open that link in said app or allow to install it from the App Store.
- It will provide more accurate results for Lookup Hints in Notes.

So, the power of differential privacy relies on Apple being able to examine large amounts of aggregate data, all the while ensuring that it is none the wiser about who is sending them that data. Since DP is a whole new paradigm on such a large scale scenario, people who don't trust one the mechanism can opt out right from the device's settings [31].

Therefore, as it can be seen, differential privacy is being adopted in the world of big data with many research prospective further ahead.

Table 4 Comparative study of different differential privacy mechanism

S. no	Years	Paper/work	Focus
1	2008	US Census Bureau [17]	Protecting patient data confidentiality and indicating driving examples
2	2009	PINQ [18]	Interactive DP which guarantees, at runtime, that inquiries adhere to a worldwide security spending plan
3	2010	Airavat model [22]	MAC + differential privacy, i.e. access control mechanisms in integration with DP
4	2012	GUPT [26]	Makes protection saving information investigation simple for security non-specialists, the expert can transfer subjective information mining projects and GUPT ensures the security of the yields
5	2014	Google's *Rappor*: randomized aggregatable privacy-preserving ordinal response	For telemetry, for example, learning insights about undesirable programming commandeering clients' settings
6	2014	Location privacy—*geo-indistinguishability* [25]	Ensures the client's correct area, while permitting surmised data—normally expected to acquire a specific wanted administration—to be discharged
7	2015	Google	For sharing historical traffic statistics
8	2015	DP in telecommunication big data platform, VLDB 2015 [12]	Implemented three basic DP architectures in the deployed telecommunication big data platform
9	2015	Efficient e-health data release with consistency guarantee under differential privacy, 2015 [21]	Investigated e-wellbeing information discharge issue and proposed an effective and secure e-wellbeing information discharge conspire with consistency ensure under DP
10	2016	Apple's iOS 10 [30]	DP implemented in the messaging app and search recommendations

Comparative study of different differential privacy mechanism

Table 4 lists down the comparative study of different differential privacy mechanism chronologically.

Research gap, challenges, conclusion and future prospects

Research gap

1. Differential privacy has been termed as a promising technique that can achieve mathematically precise guarantee of privacy. The future work that could be explored is the other data release issues under DP, such as real time monitoring and publishing of e-health data [21].
2. Differential privacy has been a topic that is widely explored by academia and the research community but less in industry due to its strong privacy guarantee. To make DP practically usable, three possible research directions can be made:
    a. Relaxing privacy guarantee and studying its effectiveness on specific industrial applications.
    b. Designing specific privacy scheme for specific data mining algorithms.
    c. Using large volumes of data but with low variety for training the classification models [12].
3. Differential privacy, unlike previous schemes, is currently the strongest privacy protection, which does not need any background information assumption of attackers.

Research community prefers differential privacy because of its strong mathematical boundary of the leaked privacy. The first, second and third generation of privacy protection techniques may be explored and compared in telco big data platform. First generation of privacy protection technique removes or replaces the explicitly sensitive identifiers of customers, second generation publishes a sanitized database with certain anonymity or diversity requirements and the third generation is DP [12, 32–34].

4. Improvement of the hybridized architecture by adapting DP on the entire data mining system.
    a. Use of exponential mechanism to select a tree structure.
    b. Use Laplace mechanism to assign proper probability values in the final published tree [12].
5. Division of privacy budget.
    a. Instead of equally dividing the privacy budget to each layer of decision tree [35], design an adjustable privacy budget assignment strategy.
    b. A tradeoff DP mechanism for Random Forests can be designed [12].
6. Airavat cannot confine every computation performed by untrusted code. It cannot guarantee privacy [36–39] for computations which output keys produced by untrusted mappers [22].
7. Local sensitivity measures how much the output of the function changes on the neighboring inputs from a subset of the function's domain. It often requires less amount of noise added to the output to achieve DP guarantee. This approach can be investigated in future work [22].
8. Data release techniques under various strategies are one of the main research contents for differential privacy and the theoretical basis for practical applications. Classification and comparison of the differential privacy data release methods prove that there are few achievements in this research area. But there still exists new directions to be explored. For example:
    a. In multiple-dimensional or graph data, whose sensitivity is so high that it may lead to poor utility of the data, how the sensitivity of function can be reduced and the utility enhanced.
    b. When time series data is concerned, how the utility and the security is to be balanced when the time is increasing and privacy guarantee for online data.
    c. Also, the computational complexity of many differential privacy algorithms is relatively high, so one of the areas for research can be on how to ensure privacy while also improving algorithm efficiency.
    d. If the data are stored in multiple parties, problem exists on how to share the data along with guaranteeing user's privacy between multiple parties. So, an efficient and safe algorithm with lower communication overhead is required.

        And yet again, design of an optimal privacy budgeting strategy is a great challenge [16].
9. Challenges relating to the tradeoff between privacy and utility still apply with the application of DP.
    a. New mechanisms: Conventional mechanisms under differential privacy place no restrictions on the scenarios of the analysis and the potential relations among the

datasets. New mechanisms keeping these points in view can be summed up as follows.

(i) *Matrix mechanism*

This mechanism was developed to deal with correlated linear queries with privacy guarantee on top of Laplace mechanism. These are for cases when multiple correlated queries of a structured database are necessary, while conventional differential privacy would only increase the sensitivity and make the noisy output useless.

The matrix mechanism brings to us more accurate answers to queries from the analysts than simple use of Laplacian mechanism. However, as it is presented, the workload of these queries is supposed known in advance.

(ii) *MWEM algorithm*
This algorithm is designed to improve the accuracy or performance of linear queries by some expert learning techniques of the answer given to the queries from the data analyst in such circumstances when conventional mechanisms for achieving ε-differential privacy add unacceptable levels of noise. The algorithm iteratively maintains and adapts an approximating distribution to the true datasets per the difference between the approximating datasets and the true datasets used as the multiplicative weights for updating.
Whilst a good trade-off between the accuracy and the privacy may be yielded using this algorithm, the policy of optimization is still relatively unknown, so requires further research [40].

10. The newly arrived wave should be focused on more fine-grained protocols that mediate stateful mechanisms that deal with not only correlated datasets, but also correlated processes in virtual clouding environments. In these circumstances, differential privacy might not be the only paradigm on which we need to rely on, and other more sophisticated methods that combine with cryptography may be invented [40].
11. For the tiled mechanism of the location based services, it is required to make the function "*l*" itself differentially private. More work required in this mechanism in order to provide both a formal proof of privacy as well as an efficient implementation to include in Location Guard [23].

Challenges and future prospects

(i) Promises of differential privacy are rather robust and direct but can come at the expense of accuracy.

a. For a series of differentially private queries asked, more noise is required to be added to the results.
b. After a series of queries exhausts the privacy budget, the user needs to be killed.
c. If the original guarantee is to be kept across k queries, noise must be injected k times. When k is large, utility of the output is destroyed.

d. Count works well because the presence or absence of a single record only changes the result slightly. Sums and max can be a problem.

(ii) The privacy guarantee could be relaxed. This can include increasing privacy budget parameter and studying its effectiveness on specific industrial applications. It can be interesting to check what privacy leakage is tolerable in real industrial systems.

(iii) Challenges demand for schemes to be used in combination with cryptography.

(iv) Designing privacy scheme for certain data mining algorithm, like adapting differential privacy over the whole data mining service in the hybridized architecture.

(v) Implement DP for privacy protection in industrial scenarios such as the one mentioned in [41] where anonymization [42, 43] is used for privacy in big data.

(vi) Location based services also require evaluating how to automatically configure position and size, like by using user input, to increase the performance in terms of both privacy and utility [23].

(vii) Differential privacy in big data scenario is yet to be explored and implemented in real world datasets on a large scale to make use of its full potential being the one with the strongest mathematical guarantee.

Conclusion

Differential privacy is one of the major topics in present day research on big data privacy. Here, in this paper, a review on differential privacy is presented, the basics of DP and the current practices. The usage of differential privacy have also been listed down to give an idea of the work done in this field till now. Some of the future work perspectives include applying cryptography techniques to DP, designing efficient schemes for DP implementation in practical life or change in privacy budgets to see its effectiveness. Conclusively, differential privacy is still an area requiring in depth research. Many challenges still exist, but further work might tackle them.

Abbreviations

DP: differential privacy; MR: Map Reduce; KVP: key-value pairs; HDFS: hadoop distributed file system; API: application programming interface.

Authors' contributions

PJ performed the primary literature review and analysis for this manuscript work. MG worked with PJ to develop the article framework and focus, and MG also drafted the manuscript. NK introduced this topic to PJ. MG and NK revised the manuscript for important intellectual content and have given final approval of the version to be published. All authors read and approved the final manuscript.

Authors' information

Mr. Priyank Jain is working as a Ph.D. Research Scholar. He is having more than 8 years experience as an Assistant Professor and in the research field. Mr. Priyank Jain has experience from Indian Institute of Management Ahmedabad India (IIMA) in the research field. His Ph.D. is in the Big data area. His educational qualification is M.Tech and B.E in Information Technology. Mr. Priyank Jains areas of specialization are Big data, Big Data Privacy & Security, Data mining, Privacy-Preserving, and Information Retrieval. Mr. Priyank Jain has publications in various International Conference, International Journal, and National Conference. He is a member of HIMSS.

Dr. Manasi Gyanchandani working as Assistant Professor in MANIT Bhopal. She is having more than 20 years experience, Her educational qualification is Ph.D. in Computer Science and Engineering. Dr. Manasi Gyanchandani area of specialization in Big data, Big Data Privacy and Security, Data mining, Privacy-Preserving, Artificial Intelligence, Expert System, Neural Networks, Intrusion Detection and Information Retrieval. Dr. Manasi Gyanchandani, publications in 01 International Conference, 03 International Journal and 04 National Conference. She is a life member of ISTE.

Dr. Nilay Khare working as Associate Professor in MANIT Bhopal. He is having more than 21 years experience. His educational qualification is Ph.D. in Computer Science and Engineering. Dr. Nilay Khare area of Specialization in Big data, Big Data Privacy & Security, Wireless Networks, Theoretical Computer Science. Dr. Nilay Khare, publications in 54 National and International Conference. He is a life member of ISTE.

Acknowledgements
We acknowledge the support of the Information retrieval and Machine Learning Lab of MANIT Bhopal MP India.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
We are presenting this article without any kind of funding.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 15 January 2018 Accepted: 30 March 2018
Published online: 13 April 2018

References

1. Microsoft differential privacy for everyone. 2015. http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf. Accessed 18 Dec 2017.
2. Samarati P. Protecting respondent's privacy in micro data release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.
3. Jain P, Gyanchandani M, Khare Direndrapratap singh N, Rajesh L. A survey on big data privacy using hadoop architecture. Int J Comput Sci Netw Secur (IJCSNS). 2017;17:148.
4. Al-Zobbi M, Shahrestani S, Ruan C. Improving MapReduce privacy by implementing multi-dimensional sensitivity-based anonymization. J Big Data. 2017;4:45.
5. Derbeko P, et al. Security and privacy aspects in MapReduce on clouds: a survey. Comput Sci Rev. 2016;20:1–28. https://doi.org/10.1016/j.cosrev.2016.05.001.
6. Dwork C. Differential privacy. In: ICALP. 2006.
7. Apple announced that they will be using a technique called "Differential Privacy" (henceforth: DP) to improve the privacy of their data collection practices 2016. https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/. Accessed 5 Jan 2018.
8. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. J Big Data. 2016;3:25.
9. Mohammed N, Chen R, Fung BCM, Yu PS. Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, 21–24 August 2011. New York: ACM; 2011. p. 493–501.
10. Friedman A, Schuster A. Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, 25–28 July 2010. New York: ACM; 2010. p. 493–502. https://doi.org/10.1145/1835804.1835868.
11. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2014;9(3–4):211–407. https://doi.org/10.1561/0400000042.
12. Hu X, Yuan M, Yao J, Deng Y, Chen L, Yang Q, Guan H, Zeng J. Differential privacy in telco big data platform. Proc VLDB Endow. 2015;8(12):1692–703. https://doi.org/10.14778/2824032.2824067.
13. Differential privacy in the wild: a tutorial on current practices and open challenges 2016. http://vldb2016.persistent.com/differential_privacy_in_the_wild.php. Accessed 12 Jan 2018.
14. Differential privacy defined. https://www.coursera.org/learn/data-results/lecture/phj4C/differential-privacy-defined. Accessed 11 Dec 2017.
15. Differential privacy. https://en.wikipedia.org/wiki/Differential_privacy. Accessed 2 Dec 2017.
16. Wang J, Liu S, Li Y. A Review of differential privacy in individual data release. Int J Distrib Sensor Netw. 2015;11:259682. https://doi.org/10.1155/2015/259682.
17. Lee DG-Y. Protecting patient data confidentiality using differential privacy. 2008. Scholar Archive. Paper 392.
18. McSherry F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of communications of the ACM, vol. 53(9), 2010.
19. Soria-Comas Jordi, Domingo-Ferrer Josep. Big data privacy: challenges to privacy principles and models. Data Sci Eng. 2016;1(1):21–8. https://doi.org/10.1007/s41019-015-0001-x.
20. Han J. Data mining: concepts and techniques. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2005.
21. Li H, Dai Y, Lin X. Efficient e-health data release with consistency guarantee under differential privacy. In: 17th international conference on e-health networking, application & services (HealthCom). IEEE, Boston, MA; 2015. p. 602–8. https://doi.org/10.1109/HealthCom.2015.7454576. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7454576&isnumber=7454459.

22. Roy I, Setty STV, Kilzer A, Shmatikov V, Witchel E. Airavat: security and privacy for MapReduce. In: Proceedings of the 7th USENIX symposium on networked systems design and implementation, NSDI 2010, San Jose, April 28–30. p. 297–12. 2010.
23. Chatzikokolakis K, Palamidessi C, Stronati M. Location privacy via geo-indistinguishability. In: Leucker M et al. (ed.). Switzerland: Springer International Publishing 2015, ICTAC 2015, LNCS, vol. 9399, 2015. p. 28–38. https://doi.org/10.1007/978-3-319-25150-9.
24. Hien To CS, Ghinita G. A framework for protecting worker location privacy in spatial crowdsourcing. Proc VLDB Endow. 2014;10(7):919–30.
25. Andrés ME, Bordenabe NE, Chatzikokolakis K, Palamidessi P. Geo-Indistinguishability: differential privacy for location-based systems. In: ACM. ISBN: 978-1-4503-2477. https://doi.org/10.1145/2508859.2516735. 2014.
26. Mohan P, Thakurta A, Shi E, Song D, Culler DE. GUPT: privacy preserving data analysis made easy. In: ACM SIGMOD'12, Scottsdale, May 20–24 2012. 2012.
27. Sharma S, Toshniwal D. Scalable two-phase co-occurring sensitive pattern hiding using MapReduce. J Big Data. 2017;4:4.
28. Olshannikova E, Olsson T, Huhtamäki J, Kärkkäinen H. Conceptualizing big social data. J Big Data. 2017;4:3.
29. Toga AW, Dinov ID. Sharing big biomedical data. J Big Data. 2015;2:7.
30. Apple's 'differential privacy' is about collecting your data—but not your data 2016.https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/. Accessed 25 Dec 2017.
31. What is "differential privacy," and how does it keep my data anonymous 2017. https://www.howtogeek.com/290298/what-is-differential-privacy-and-how-does-it-keep-my-data-anonymous/. Accessed 10 Jan 2018.
32. Huang Y, Zhu F, Yuan M, Deng K, Li Y, Ni B, Dai W, Yang Q, Zeng J. Telco churn prediction with big data. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data. Melbourne: SIGMOD'15; 2015. p. 607–18.
33. Jagannathan G, Pillaipakkamnatt K, Wright RN. A practical differentially private random decision tree classifier. Trans Data Privacy. 2012;5(1):273–95.
34. Jiang S, Fiore GA, Yang Y, Ferreira Jr J, Frazzoli E, Gonzalez MC. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: KDD workshop on urban computing. New York,: ACM; 2013. p. 2–9.
35. Lemmens A, Croux C. Bagging and boosting classification trees to predict churn. J Mark Res. 2006;43(2):276–86.
36. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. New York: SIGMOD'05; 2005. p. 49–60.
37. Li N, Li T, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. In: ICDE. Piscataway: IEEE; 2007. p. 106–15.
38. Li N, Qardaji W, Su D, Cao J. Privbasis: frequent itemset mining with differential privacy. Proc VLDB Endow. 2012;5(11):1340–51.
39. Lima E, Mues C, Baesens B. Domain knowledge integration in data mining using decision tables: case studies in churn prediction. J Operational Res Soc. 2009;60(8):1096–106.
40. Yao X, Zhou X, Ma J. Differential privacy of big data: an overview 2016. In: IEEE 2nd international conference on big data security on cloud, IEEE international conference on high performance and smart computing, IEEE international conference on intelligent data and security, Washington DC, 2016. 2016.
41. Sedayao J, Bhardwaj R, Gorade N. Making big data, privacy, and anonymization work together in the enterprise:experiences and issues. In: Anchorage: IEEE international congress on big data; 2014.
42. Liu K Terzi E. Towards identity anonymization on graphs. In: SIGMOD'08, New York: ACM; 2008. pp. 93–106.
43. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. In Liu L, Reuter A, Whang K-Y, Zhang J, editors. ICDE, Washington DC: IEEE Computer Society; p. 24. 2006.
44. Ko SY, Jeon K, Morales R, The HybrEx model for confidentiality and privacy in cloud computing. In: 3rd USENIX workshop on hot topics in cloud computing. HotCloud'11, Portland, June 14–15, 2011. 2011.
45. Blass E, Noubir G, Huu TV. EPiC: efficient privacy preserving counting for MapReduce. 2012.
46. Dolev S, Li Y, Sharma S. Private and secure secret shared MapReduce—(extended abstract). In: Data and applications security and privacy XXX. In: Proceedings 30th annual IFIP WG 11.3 working conference, DBSec 2016, Trento, July 18–21, 2016. 2016.
47. Shamir A. How to share a secret. Commun ACM. 1979;22(11):612–3.