Privacy and Security Policies in Big Data

Sharvari Tamane
MGM's Jawaharlal Nehru Engineering College, India

Vijender Kumar Solanki
Institute of Technology and Science Ghaziabad, India

Nilanjan Dey
Techno India College of Technology, India

A volume in the Advances in Information Security, Privacy, and Ethics (AISPE) Book Series

IGI GLOBAL
DISSEMINATOR OF KNOWLEDGE
www.igi-global.com

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Tamane, Sharvari, 1973- editor. | Solanki, Vijender Kumar 1980- editor. | Dey, Nilanjan, 1984- editor.
Title: Privacy and security policies in big data / Sharvari Tamane, Vijender Kumar Solanki, and Nilanjan Dey, editors.
Description: Hershey, PA : Information Science Reference, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2017003838| ISBN 9781522524861 (hardcover) | ISBN 9781522524878 (ebook)
Subjects: LCSH: Big data--Security measures. | Data protection. | Privacy, Right of. | Telecommunication policy.
Classification: LCC QA76.9.B45 P75 2017 | DDC 005.8--dc23 LC record available at https://lccn.loc.gov/2017003838

This book is published in the IGI Global book series Advances in Information Security, Privacy, and Ethics (AISPE) (ISSN: 1948-9730; eISSN: 1948-9749)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 9
Differential Privacy Approach for Big Data Privacy in Healthcare

Marmar Moussa
University of Connecticut, USA

Steven A. Demurjian
University of Connecticut, USA

ABSTRACT

This chapter presents a survey of the most important security and privacy issues related to large-scale data sharing and mining in big data with focus on differential privacy as a promising approach for achieving privacy especially in statistical databases often used in healthcare. A case study is presented utilizing differential privacy in healthcare domain, the chapter analyzes and compares the major differentially private data release strategies and noise mechanisms such as the Laplace and the exponential mechanisms. The background section discusses several security and privacy approaches in big data including authentication and encryption protocols, and privacy preserving techniques such as k-anonymity. Next, the chapter introduces the differential privacy concepts used in the interactive and non-interactive data sharing models and the various noise mechanisms used. An instrumental case study is then presented to examine the effect of applying differential privacy in analytics. The chapter then explores the future trends and finally, provides a conclusion.

INTRODUCTION

Big Data analysis influences most aspects of our modern society, such as mobile services, retail, manufacturing, financial services, medicine and life sciences, as well as physical sciences to name a few (Bertino et al., 2011). Scientific research is being revolutionized by Big Data everyday, for instance in bioinformatics with Next Generation Sequencing increasing the size and number of experimental data sets exponentially. In healthcare, Big Data with transforming patient care towards prevention with substantial home-based and continuous form of monitoring available to patients is definitely personalizing

DOI: 10.4018/978-1-5225-2486-1.ch009

healthcare to the benefit of patients. While the potential benefits of Big Data are real and significant, there remain several considerable technical challenges. However, in this broad range of application areas, data is being collected at an unprecedented scale. The emergence and ever increasing emphasis on the big data era means that more and more information on an individual's health, financials, location, and online activity are continuously being harvested, collected, and processed in the cloud and stored in big data repositories. This results in increased concerns regarding the privacy of these large sets of personal data and the loss of an individual's control over his/her sensitive data (Boyd & Crawford, 2012).

The impact of privacy concerns on a big data application is particularly evident in the healthcare domain which has a long established history in requiring that health information technology must comply with the Health Insurance Portability and Accountability Act (HIPAA) for most importantly release of a patient's medical information as well as security and availability as well. HIPAA must also apply to big-data applications for healthcare. This is strongly tied to a movement towards patient controlled access to their medical information with patients able to define the privacy to determine who can see what information at which times. This is evidenced by work that has emphasized granularity and patient control (Sujansky et al., 2010) and a lifetime electronic health record with complete information available anywhere (Caine, 2013). In healthcare there is a need to distinguish levels of security based on the confidentiality and privacy of the data itself and the way that a patient would seek to make such data available to stakeholders. All of these security and privacy concerns must be addressed within big data applications for healthcare as well as in other domains.

This chapter explores the issues related to the security in general and privacy in specific for big data applications, particularly given that the usage of state-of-the-art analytics has explicitly led to growing privacy concerns. As a result, protecting privacy becomes quite harder as information is processed multiple times and shared among multiple diverse entities in the cloud. One example of this problem involves de-identification and anonymization techniques that have been utilized under the false assumption that they allow organizations to reap the benefits of analytics while preserving individuals' privacy. This relies on the assumption that removing certain personal information from a data set would ensure the identity of the users participating in that data set to remain anonymous. However, this has proved to be a misconception as demonstrated by several re-identification and linkage attacks that different data sources harmfully leak private information when combined and when adversaries are able to use some background knowledge, this will be further discussed in the section "Big Data Security and Privacy Issues".

The first focus of this chapter is to explore the utilization of differential privacy to addresses the aforementioned problems in privacy in order to provide confidence to users that their data is carefully controlled. Differential privacy (DWork, 2006) is defined as the application of noise functions of certain characteristics to datasets or query results so that no specifics of individual records present in the original dataset are revealed, while simultaneously allowing the dataset to provide typical big data analytical insights. This constraint allows the various big data analytics mechanisms to behave almost identically on any two datasets that are sufficiently close but only differ by the applied noise mechanism. A formal differential privacy model (DWork, 2006) defined differential privacy as: "the risk to one's privacy should not substantially increase as a result of participating in a statistical database." Differential privacy has recently received increased attention as a general pipeline for the protection of personal information, especially in the fields of big data analytics. The appeal of differential privacy is that there are usually little or no pre-assumptions about a potential attackers pre-existing background knowledge and offers a solid mathematical formulation of the notion of privacy. In contrast to the aforementioned anonymization techniques, the privacy guarantees of differential privacy are rather strong, but can come

at the expense of accuracy. This degradation in accuracy would be problematic in a big data application for healthcare if the underlying patient and/or genomic data is incorrect. In addition, there is increased complexity for designing and implementing a differentially private version of nearly every algorithm utilized for a complex task (e.g., data mining) that overshadows the wide application of differential privacy in practice, it is hence of utmost importance to carefully study the gains and costs of applying differential privacy in healthcare.

The second focus of this chapter highlights the issues related to security and privacy for big data applications by presenting a survey and analysis of the most important security and privacy issues in large-scale data processing associated with big data as well as presenting an case study utilizing differential privacy in the healthcare domain. This chapter will summarize several security challenges for big data from four different perspectives (Inukollu et al., 2014): architecture and network related issues such as network protocol security and node validations; authentication and authorization related issues such as node authentication and access control protocols; data related issues such as encryption, key management and data privacy issues; and, general issues such as logging (Peleg et al., 2008). To understand the case study of differential privacy in healthcare, the chapter includes a comprehensive survey of privacy challenges when sharing or releasing big data for analytics. Our work in this chapter will further present the most promising technologies for preserving privacy in big data applications, such as various k-anonymity (Clifton, 2013) and differential privacy techniques. This chapter will also present a theoretical and empirical comparison with respect to the two major differential privacy settings (DWork, 2010): interactive settings where a dataset owner provides a set of differentially private data querying algorithms for a data requester to interact with vs. non-interactive settings where a differentially private data set is released once and the data requester interactions are directly focused on that released privacy preserving dataset. As part of the discussion, this chapter analyzes and compares the major differentially private data release strategies and noise mechanisms such as the Laplace mechanism and the exponential mechanism (DWork, 2014).

This chapter has five sections additional to the introduction. The *Background* section provides general background on characteristics of big data applications, big data challenges, big data processing technologies, and big data analysis techniques. The *Big Data Security and Data Privacy Issue* section discusses the security and privacy challenges and techniques for big data including: architecture-level node authentication protocols, data-encryption protocols, and privacy preserving techniques such as k-anonymity and differential privacy it also introduces and surveys the differential privacy concepts by explaining the interactive and non-interactive models for differential privacy and the various noise mechanisms used in releasing differentially private data. Then, the *Differential Privacy Case Study in Healthcare Data Mining* section presents a case study of applying an approach for differential privacy in big data analytics that can be particularly useful for a domain such as healthcare. The *Future Trends* section explores the potential directions of differential privacy in applying privacy-by-design principles to different data domains and ensuring privacy-aware data usage. Finally, the *Conclusion* section summarizes the contributions of the chapter.

BACKGROUND

This section provides background information on four areas: characteristics of big data applications, big data challenges, big data processing technologies, and big data analysis techniques. To begin, the term

Big Data started as a nebulous term, used by Computer Science researches to describe the exponential rate in data acquisition and recording in the internet age. Big Data is considered a Framework of utilities and characteristics common to all NoSQL platforms. Gartner Research's definition of Big Data is widely adopted; the three Vs of Big Data consists of Volume, Variety and Velocity. A 4th V was also added to make it: variety, volume, velocity and value. Big Data differs from a data warehouse in architecture in that it follows a distributed approach, a data warehouse on the other hand follows a centralized one (Lane, 2013). The major characteristics of big data are: very large data sets (Volume), extremely fast insertion (Velocity), and multiple data types (Variety). Corresponding characteristics (Lane, 2013) include: distributed parallel processing, clustered deployments, providing data analysis capabilities, distributed and redundant data storage, modular design, inexpensive, hardware agnostic, easy to use (relatively), available (commercial or open source), and extensible can be augmented or altered In big data applications, the time from data acquisition to meaningful information realization is critical to extract value from various data sources, including mobile devices, the web and a growing list of automated sensory technologies. Application that can realize this goal would have a huge advantage to organizations that would benefit from speed, capacity, and scalability of cloud storage (Cheung, 2013). Organizations that in addition to benefiting from these big data characteristics also combine predictive analytics with big data have opportunity to explore further benefits in application areas including: digital marketing optimization such as web analytics for online advertisement, context-based recommendations etc.; data exploration and discovery such as statistics, data science, exploring new markets, etc.; fraud detection and prevention and network monitoring and security analysis; social network and relationship analysis with the ultimate goal to influence relevant markets; machine generated analysis for instance remote sensing; data retention and archiving to insure survivability; and, data visualization to present information suitable for users (Arpaia, 2013).

The next background area is Big Data significant challenges further than the analysis phase that occurs in multiple phases (Brown, 2011). In *Data Acquisition and Recording* phase, the *data volume* challenge puts pressure on capacity. The use of data reduction mechanisms can smartly process raw data while defining filters that help to not accidently discard useful information in the process. The general *correct metadata challenge* can alleviate the overhead imposed by the necessity of recording metadata. For example, a processing error at a prerequisite step can render depending analysis erroneous. However, with suitable provenance, one can easily identify all depending subsequent processing steps. This is reflected in the *data preparing and cleaning challenge* in order to effectively extract meaningful information from often noisy data and expressing the data in a form suitable for analysis is often application dependent and is a continuous technical challenge. This is true in healthcare which must combine: electronic health records (EHR) from databases in hospitals (Kendall, 2015); transcribed dictations from several physicians; data structured and collected from biosensors and other modern fitness devices and various –sometime uncertain- measurements; and, medical imaging data such as x-ray, CT, MRI, etc. *Data Integration, Aggregation, and Representation challenges* have led to novel strategies emerged that involve storing unstructured data in distributed NoSQL databases such as the Apache Hadoop Distributed File System (HDFS) (Borthakur, 2007), a single logical file system distributed across many data servers and it is able to scale on demand based on required capacity and was designed to run on commodity hardware and hence be highly scalable and available. HDFS contains MapReduce, a programming model and an associated implementation for effectively processing and querying large data sets with a parallel, distributed algorithm on a cluster of nodes across the disparate data servers. The *Data Modeling and Analysis challenge* must deal with data that is often noisy and dynamic, almost always heterogeneous,

and could also sometimes be untrustworthy. "Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments" (Cyril, 2015). The final *result interpretation challenge* involves the need to provide a rich palette of visualizations for results of data analysis. A user needs to be able to not only view and understand the results from the analysis phase, but be able to test the data model and deploy it into the real world and conclude predictive as well as prescriptive results from it for final decision support (Tene, 2012).

The third background area reviews big data processing technologies that are needed to address, volume, variety and velocity utilizing a divide and conquer approach to provide for these characteristics and handle semi-structured and sometimes unstructured data in a distributed environment. NoSQL commercially available systems (e.g., MongoDB, Apache Casandra, etc.) can be leveraged. Existing NoSQL implementations can be classified as: a Key-Value Store where the content of the data is represented as a collection of individual key and value pairs; a Graph Database that represents the data in graph objects utilizing Graph Theory; or a Document Stores that organizes the data in a container object per document (e.g., XML or JSON) to encapsulate all attributes for a given object. These implementations do not necessarily satisfy core concepts (e.g., atomicity, consistency, isolation or durability (ACID)); the very same set of properties that are present in almost all relational database management systems today to guarantee that relational database transactions are processed reliably. One dominant approach as previously mentioned is Hadoop Distributed File System(HDFS). HDFS provides built-in support for data fault-tolerance via data replication and load-balancing using MapReduce which can benefit of the locality of data. MapReduce has three functions to manage the local data, its writes to temporary storage, moving data for downstream processing, and redistributing data based on outputs.

The final background area is big data analysis techniques that are utilized used to generate numerous and also more insightful results than when applied to smaller less diverse sets. Advantages and issues are present for each of the techniques based on its individual characteristics. Analysis techniques include: crowd sourcing, data mining, genetic algorithms, data fusion and integration, machine learning, NLP, neural networks and simulation, pattern recognition, predictive modeling, semantic and sentiment analysis, and statistics. Two techniques of increasing interest are predictive analysis and descriptive analysis. Predictive Analytics applies mathematics, statistics and probability theory in conjunction with the overarching computer science discipline of machine learning, data modeling and algorithm development. From clinical analytics in Clinical Decision Support System (CDSS) to business analytics in Operations Research (OR), Predictive Analytics aids decision makers to make choices and solve problems that have long lasting impacts. Predictive Analytics relies on Descriptive Analytics to provide the descriptive information as well as a foundational framework for such applications. Descriptive analytics, however, only describes the present conditions, whereas predictive analysis is a model-driven and data-driven approach for generating what-if scenarios exploiting the meanings of underlying data.

BIG DATA SECURITY AND DATA PRIVACY ISSUES

This section explores issues in big data security and privacy organized into a five-part discussion to set the context for the chapter in order to support the presentation of the healthcare case study in the next section. In part one, big data security is reviewed with the objective to provide a general background about the current security challenges to which differential privacy is one of the potential answers. In

part two, the concepts of differential privacy are reviewed including different approaches, models, and algorithms. Part three of this section reviews the critical properties of differentially private algorithms that are utilized to design differentially private data sharing models. Part four of this section explains a select subset of the noise mechanisms that are the core part of implementing sound differentially private algorithms. The fifth and final part of the section details models of releasing sensitive data with differential privacy that can be leveraged to sharing sensitive datasets in domains like healthcare or other domains with potentially sensitive individual's data.

In the first part of this section, we explore security. Big data have similar vulnerabilities similar to traditional web applications and most data warehouses including: vetting of nodes and client applications before they join into a cluster, protecting data at rest, ensuring network privacy and communications, and also node management; most of these are lacking in the NoSQL platforms utilized for big data. Security controls for big data can be considered from four different levels: architecture/network, authentication and authorization, and data. In the architecture/network level security the challenges are network protocols and network security for distributed nodes validation and internode communication to verify security consistency across a highly distributed cluster of heterogenous platforms. A set of pre-configuration tools has emerged as one possible solution that validate and 'fix' node issues before adding them back to the cluster to ensure a form of baseline security. For authentication and authorization level, the challenges are on authentication methods to manage administrative rights for nodes and authentication of applications and nodes and to ensure that secure administrative passwords are being used correctly and that application users also are being correctly and securely authenticated before gaining access to the cluster. For the data level security the challenges are data encryption to protect the integrity of data at rest as we need to ensure administrators or other unauthorized application processes cannot gain direct access to files while at the same time preventing information leakage or exposure (Chaudhuri, 2012).

In the second part of this section, we explore differential privacy. In the past, various ad-hoc approaches to anonymizing public records have failed when researchers successfully identified personal information by linking several seemingly separate databases (Barbaro, 2006). Two well-known instances of successful "Linkage Attacks" have been the Netflix Database (Bennett, 2007) and the Massachusetts Group Insurance Commission (GIC) medical encounter database (Dankar, 2012). The objective in the general case is for a statistical database where: "a trusted party holds a dataset of sensitive information (e.g. medical records, voter registration information, email usage) with the goal of providing global, statistical information about the data publicly available, while preserving the privacy of the users whose information the data set contains" (Dwork, 2004). The notion of indistinguishability, later termed Differential Privacy (Dwork, 2006), formalizes the exact notion of "privacy" in statistical databases. Informally, differential privacy can be defined to stipulate that any individual has a very small influence on the distribution of the outcome of the computation. As a result, an attacker cannot learn anything about an individual's report to the database, even in the presence of any auxiliary information she may have.

Differential privacy applied to datasets or query results is making a promise to not reveal the specifics of individual records present in the original dataset and achieves this constraint by requiring the mechanism that is to be considered differentially private to behave almost in an identical manner on any two given datasets that are considered sufficiently close. Based on Dwork's work (Dwork, 2004): "imagine a dataset A whose records are members of some abstract domain D, and which can be described as a function from D to the natural numbers N, with $F(x)$ indicating the frequency (number of occurrences) for x in the dataset. $|| A - B ||$ is used to indicate the sum of the absolute values of difference

in frequencies (i.e., the total number of records that would have to be added and removed to change A into another dataset B)". This leads to a definition:

Definition 1: Differential Privacy: "A mechanism M mapping datasets to distributions over an output space R provides (ε, δ) - differential privacy if for every S $\subseteq R$ and for all data sets A, B where $|| A - B ||\leq 1$,

$$Pr[M(A) \in S] \leq e^{\varepsilon} Pr[M(B) \in S] + \delta.\text{"} \quad (1)$$

For $\delta = 0$ in (1) M is said to provide ε -differential privacy. Prior to the field of differential privacy being defined by Dwork between 2004 and 2006, the privacy protection schemes proposed in research mainly included: data distortion methods, data encryption techniques, and restrictive release of only partial or selected group of records. However, these methods failed to offer quantifiable guarantee for user's privacy and did not clearly discuss the extent of an adversary's ability for which they offer protection (De Montjoye, 2015).

To substitute for these failed methods, de-identification and anonymization privacy algorithms were used under the false assumption that removing certain personal information from a data set would ensure the identity of the users participating in that data set to remain anonymous. However, several re-identification and linkage attacks demonstrated that different data sources are harmfully leaking private information when combined, especially given some background knowledge of adversaries. This lack of actual privacy guarantee is one of the reasons anonymization techniques do not satisfy most of the privacy requirements of sensitive data releases. Some of the most interesting examples for such shortcomings were the Netflix Prize related attack presented in (Bennett, 2007) as mentioned before the most recent attack of the credit card metadata re-identification in (De Montojoye, 2015). Anonymity models based on restrictive release of sensitive data were also proposed in part to guarantee user privacy in healthcare application, among such techniques is k-anonymity introduced by Sweeney in 2002 and some of its variations like l-diversity (Sweeney, 2002). It is simply suggesting a property that each record is considered indistinguishable from at least l other records to insure privacy. Although k anonymity and its variants do provide stronger privacy guarantees, there are several points that hinder its wide use for privacy like the high computational cost as k-anonymity is considered np-hard.

Another important facet in differential privacy is global sensitivity and its relation to noise-based privacy to provide protection for sensitive data sharing. For example, using Definition 1, one can derive that when two datasets A and B differ only in the data of one individual, then the gap where $|| A - B ||$ is maximum can be utilized obscure in order to make it difficult to an attacker to infer whether or not a specific individual information is in fact present in one of the dataset versions under consideration. Differential privacy proposes adding noise to the original data set A to cover this gap. In other words, the mechanism for differential privacy $M(f, A) = f(A) + noise$. The difference that this noise must obscure can be calculated as follows: given that A and B are two data sets that differ in exactly one individual's data, and $F(A) = x$ is a deterministic, non-privatized function over data set A , which returns a vector X of k real number results, then, the global sensitivity of F is then defined as:

$$\Delta f = max_{A,B} \mid f(A) - f(B) \mid \tag{2}$$

Intuitively, the global sensitivity represents the sum of the worst case difference in answers that can be caused by adding or removing an individual's information from a data set. An example of the noise that can be added to the results of F so as to cover the sensitivity represented in (2) are the random values taken from a Laplacian Distribution with standard deviation that is large enough to cover this gap.

The final aspect of differential privacy is the privacy budget which is utilized to control and regulate the loss in privacy when querying the altered data sets. As noted in (Dwork, 2010), setting a value for ε is not always an easy task and has not been adequately covered in the differential privacy literature. Non-specialist data holders have difficulty measuring the privacy protection of a dataset provided from a specific ε value. What exactly it means to have an information gain of $\varepsilon = 0.01$ is not always intuitive to the data owner. To our knowledge there is still no generalized experimental evaluation to guide the user on choosing an appropriate ε value. In (Dwork, 2010), the recommended values of 0.01, or 0.1, and sometimes ln2, and ln3 can be used as starting values for tuning the best parameter value for each data set. One effort (Dankar, 2013) has suggested that ε cannot be defined in general but will always depend on the dataset under consideration.

In the third part of this section, we review the critical properties of differentially private algorithms and results, namely: data types' invariance, parallel and sequential composition, post-processing invariance, and quantifiable privacy. Data types invariance relies on the assumption that records form our data sets are invariant (Dwork, 2014) and works best when there are few records for each participant. Data types' invariance requires no assumptions about the type of data of the data sets' records. Different from other privacy methods, the privacy guarantees provided by differential privacy do not rely on classifying attributes as sensitive or not, nor perturbing the source data, nor suppressing values that are scarce or sensitive. Independence of data type is an important property that removes the need to customize privacy guarantees for different domains, misclassifying attributes as insensitive, or overlooking sensitive combinations of insensitive attributes. Meaningful guarantees for unstructured data can be provided, like free text and binary data that have previously vexed sensitivity classification. Even mutable records can be supported, replacing each record with a time-line of its contents. Furthermore, by ignoring entirely the records semantics one can provide guarantees for arbitrary functions of them. This property allows analysts to write their own tailored analyses, rather than choose from a set of predefined computations over limited declassified attributes.

The next property, parallel and sequential composition (McSherry, 2009) can be performed over structurally disjoint subsets of the data, where the same sequence of analyses provides $\max_i \varepsilon_i$-differential privacy. An example of such a sequence of analyses is the grouping of results for horizontally distributed datasets analysis, where each record is guaranteed to participate in at most one aggregation. This means that while general sequences of queries accumulate privacy costs additively, when the queries are applied to disjoint subsets of the data, the bound can be improved. Specifically, if the domain of input records is partitioned into disjoint sets, independent of the actual data, and the restrictions of the input data to each part are subjected to differentially private analysis, the ultimate privacy guarantee depends only on the worst of the guarantees of each analysis, not the sum. The Sequential Composition Theorem for differential privacy states that the sequence of $M_i(X)$ ($\sum_i$)-differential privacy provides, if each M_i ε_i-differential privacy provides. This defines the privacy guarantees degrade as more informa-

tion is exposed. Sequential composition is crucial for any privacy platform that expects to process more than one query. Privacy definitions that are not robust to sequential composition are usually hard to implement in practice.

The third property, post-processing invariance follows the sequential composition property of differentially private algorithms, but it can run independently so that subsequent computations can consider and incorporate the resulting outcomes of any preceding computations. The final property, quantifiable privacy differs from the previous properties which allow a designer of a differentially private mechanism to bind the privacy implications of arbitrary sequences of arbitrary queries composed of permitted transformations and aggregations leading to quantifiable privacy whose value can be calculated as the needed operations are designed or executed. Queries which arrive in sequence have their epsilon values accumulate; queries applied in parallel require us to track only the maximum (McSherry, 2009). Note that the privacy guarantees degrade as more information is exposed and more accuracy is required, however this decrease in privacy guarantee is somewhat well-controlled and not as drastically deteriorating as in the case of k-anonymity for instance (McSherry, 2009).

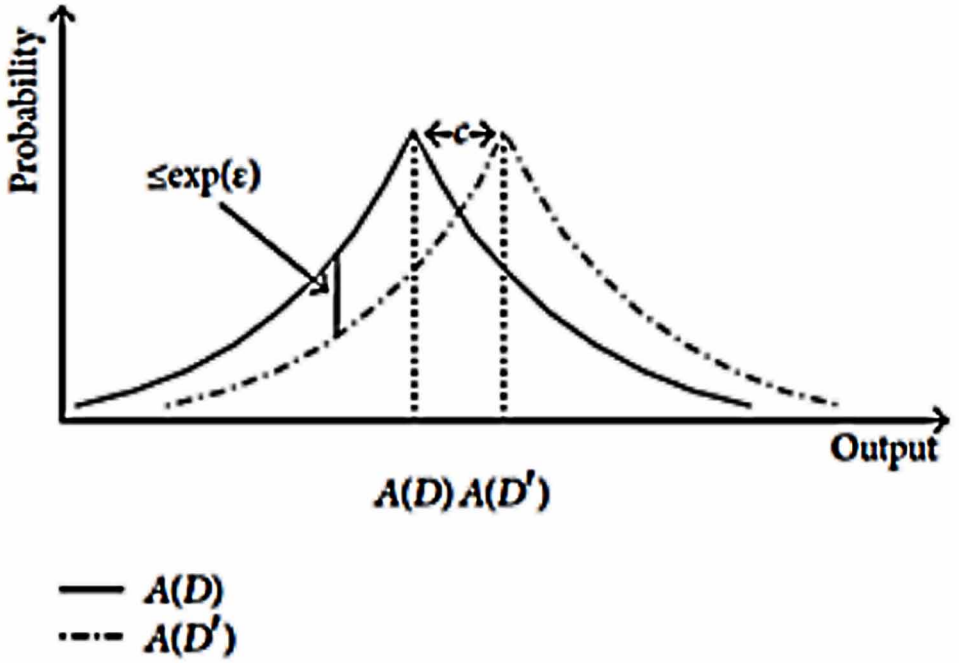
In the fourth part of this section, we explore differential privacy noise mechanisms which are utilized to support the analysis when a worst-case of data sets is present and to produce a similar distribution of privatized results, noise is added to span over the sensitivity gap. Adding Laplacian Noise is not the only way, but as suggested in (Dwork, 2008), was proposed that differential privacy can be achieved by adding random noise drawn from the Laplace distribution to the result of an algorithm. Following from the definition of global sensitivity and given a dataset D and the function $F : D \rightarrow R^d$, global sensitivity is ΔF; random algorithm $A(D) = F(D) + noise$ satisfies ε-differential privacy if the noise obeys the Laplace distribution; that is, $noise : Lap(\Delta F \,/\, \varepsilon)$; note that the location parameter of the Laplace Distribution in this case is 0 and the scale parameter is $(\Delta F \,/\, \varepsilon)$. While the Laplace mechanism is used when the output is numerical, the exponential mechanism presents another possible scheme to control security and achieve differential privacy when the outputs are non-numerical. The exponential mechanism satisfies the constraint that the change of a single database record does not affect the outcome of a predefined score function. The exponential mechanism can output non-numerical results according to their values of that score function. The output probability as shown in Figure 1, refers to privacy budget. The highest scored result is shown with higher probability when ε is larger ; meanwhile, when the difference between the output probabilities grows, the offered privacy becomes less and vice versa, the smaller ε is, the higher the privacy will be; this can be inferred from equation (3) below. A formal definition of the mechanism is given in (Dwork, 2006) as follows: Let D denote the input dataset, $r \in R$ denotes a potential result, given a score function $u : DxR \rightarrow R$; if a random algorithm A selects an answer based on the following probability:

$$A(D,u) = r :| \; Pr[r \in R]\alpha \exp(\frac{\varepsilon u(D,r)}{2\Delta u}) \tag{3}$$

where Δu defines the sensitivity of the score function u, then algorithm A is said to satisfy ε-differential privacy.

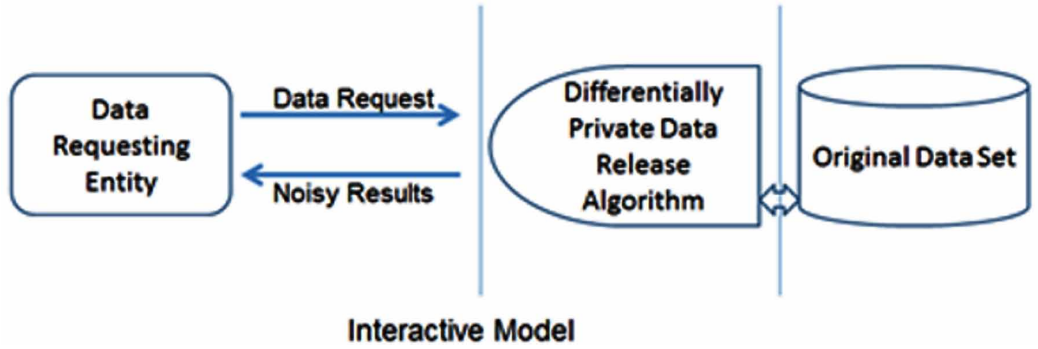
In the fifth and final part of this section, we review alternative models of releasing sensitive data with differential privacy with a focus on the two strategies for incorporating differential privacy mechanisms when releasing sensitive information or data sets: the interactive and the non-interactive strategies

Figure 1. Probabilities of dp-mechanism A for D and D'



(El Emam, 2011). In the interactive model as shown in Figure 2; a data owner provides a data querying algorithm or tool based on the concepts of differential privacy. Then, the application or user requesting the data sends their query request to that tool/algorithm. When the query algorithm receives this request, the un-sanitized data is recovered from the original database and a privatizing process is performed over the raw data with the sanitized data finally submitted to the requesting party. In this model, the permitted number of queries is restricted by privacy budget ε; so more queries essentially leads to a potentially smaller budget for each query if total budget ε is a constant and a larger noise is added to the query result; this could render the query results to becomes unusable. As a result, the key to the model, is to design the query algorithm to provide the maximum number of queries permitted under limited budget ε that makes sense.

Figure 2. Interactive data release model

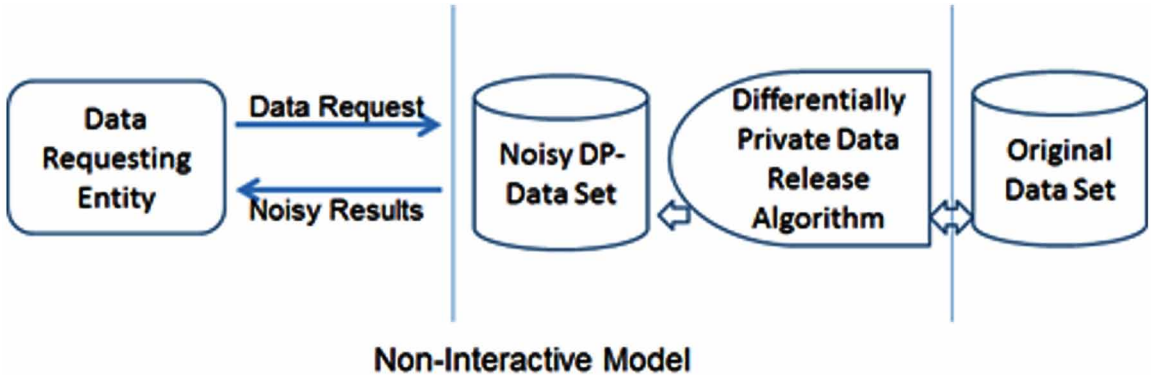


In the non-interactive model shown in Figure 3, the data owner is a trusted party and releases an already differentially private dataset, and data requesting party sends a query request. When the sanitized dataset receives the query request, the noisy result is returned to the requesting party. In this model, the number of permitted queries is unrestricted by privacy budget ε, so how to design the release algorithm with high efficiency to enhance the accuracy of a possible query over the data set at hand is the key to this model.

The interactive and the non-interactive models require histogram release, tree-structure release, and/or time series data algorithms to ensure privacy guarantee and simultaneously provide high utilization of the data and results. In support of the *histogram release algorithm*, a histogram is constructed by splitting the input dataset into mutually disjoint subsets named buckets or sometimes bins that depend on a set of properties. The only access to the original data set is performed through the differential privacy interface when users send their data queries, and the differential privacy histogram is directly utilized to answer those queries. The most straightforward and simple method is to add Laplace noise to each of the histogram buckets. In order to reduce the potential query errors, multiple buckets are possibly merged into one partition; this can be achieved when the number of tuples that fall into each partition or bin is the average value of the number of tuples in these buckets (Sarathy et al., 2011). The noise should be added to each bucket before merging, and the total noise after merging the relevant bins becomes smaller than the value before the merging. However, merge operation can introduce an error caused by approximation as the number of tuples in the partition is the average value of the number of tuples in multiple buckets. Therefore, we need to make the smallest possible number of partitions to minimize the noise error and to make the number of tuples in a partition the same as much as possible in order to reduce the approximation error. In general, a finer-grained partitioning mechanism introduces smaller approximation error but could cause larger noise error, so finding the right balance between approximation error and noise error is an important task.

The *tree-structure data release algorithms* were proposed (Wang, 2015) to support a differential privacy budgeting strategy and reduce the query error, a series of methods based on tree-structure data split. As private spatial decompositions, these algorithms divide the geospatial data into smaller regions and for each of these regions, statistics are obtained on the points within. The approach is called data-dependent decomposition if the partition discloses the sensitive information during spatial decomposition, otherwise, the approach is called data-independent decomposition. For data-dependent decomposition,

Figure 3. Non-interactive data release models



the noise is added to the node in order to hide the real values when a node is disclosed during splitting. For data-independent decomposition, algorithms based on quadtree were proposed which recursively divide the data space into equal quadrants without disclosure of node data information.

Lastly, the *time series data release algorithm* is an example of methods used for applications with data such as MHR or GPS data. As explained in (Wang 2015): "The real time data with higher correlation between time stamps has a timescale, if the length of the time series is T and the noise $noise_t$ is added to the data x_t at time k, $noise : Lap(T / \varepsilon)$, when T is large, the added noise gets large and leads to poor utility of the data." As for time series data, and for the purpose of reducing the error, the algorithm DFT was proposed based on discrete Fourier transform. In (Mohammed, 2011): " For time series D, DFT first executes the discrete Fourier transform, that is, $F = DFT(D)$, and retains only the first k coefficients; then the Laplace noise is added to those coefficients and the inverse Fourier transform is executed on the noisy coefficients F', that is $D^* = IDET(F')$. Finally, the perturbed data D^* is released."

DIFFERENTIAL PRIVACY CASE STUDY IN HEALTHCARE DATA MINING

This section discusses privacy-preserving healthcare data sharing and the way to transform raw healthcare data or related querying results into a version that is immunized against privacy attacks to support healthcare data mining. The intent is to illustrate the way that differential privacy can support effective big data typical analytics and mining tasks like for instance k-means clustering (classification) or logistic regression of healthcare data. In order to achieve this, this section discuss a simple design for secure healthcare data releasing and sharing using differential privacy concepts and conduct comprehensive experiments of applying differentially private algorithms to healthcare data sets released for data mining and study the impact of enforcing differential privacy on the results quality. In the process, we evaluate the impact of differential privacy on the data mining tasks by comparing the performance metrics of k-means clustering (classification) with and without the application of differential privacy. In the remainder of this section, a case study of a specific problem in healthcare is presented in three parts. In part one, background on healthcare data sharing and the challenges of the usage of differential privacy is motivated. In part two, we explore the design and implementation of the case study for healthcare data mining that also includes a description of the data set utilized. Part three reviews and discusses the results of the case study. It is important to note that this case study is an instrumental case study used to accomplish a better understanding of the differential privacy in data mining rather than understanding the particular outcome of the classification process used. It provides insight into and helps to refine the application of the differential privacy theory. The classification process itself is of secondary interest; it plays a supportive role in facilitating our understanding of the differential privacy parameters.

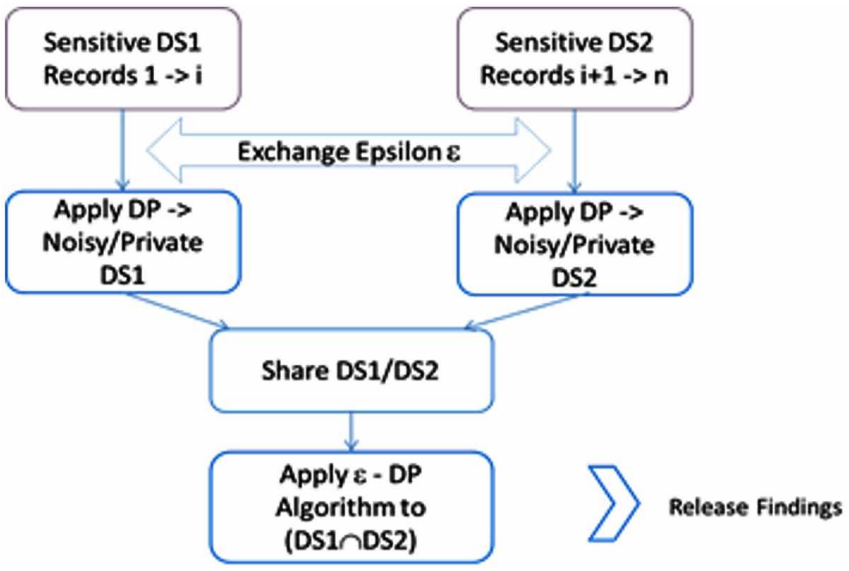
In the first part of this section, we explore the way that healthcare data sharing with appropriate privacy protection can be achieved to enable health research is one of the most critical challenges in health and medical informatics. Current de-identification approaches or microdata of original records minus basic personal attributes release are subject to various re-identification and disclosure risks and do not provide sufficient protection for patients and discourages most of them to participate in clinical trials (El Emam, 2011). "A complementary approach is to release statistical macrodata (i.e. derived statistics), which can also be used to construct synthetic data that mimic the original data" (Francis, 2015).

While differential privacy has emerged as one of the stronger provable privacy guarantees for statistical data release, there is still and open question as to whether or not differential privacy can efficiently and effectively release a dataset while guaranteeing practicality and data usefulness for health applications. "Applying differential privacy to healthcare data presents additional new challenges due to the high dimensionality, high correlation, and cross-institution distribution in healthcare datasets that are necessary to support cross-sectional, longitudinal, and cross-institutional studies." (El Emam, 2013). Healthcare data often contains categorical data (e.g., diagnosis, procedure codes, drugs dispensed, laboratory tests, etc.) as well as numeric data (e.g., age, length of stay in hospital, and time since last visit). As a result, both types, numeric and categorical, are addressed by the use of privacy mechanisms similar to both the Laplace and the Exponential mechanisms at the same time to the same data set release (Cormode et al., 2012).

Note that a non-interactive mechanism of differential privacy allows the computation of statistics without having to publish the original data set would be quite suitable for several healthcare data publishing contexts, such as in public health. In public health, ongoing monitoring of some data variables for a epidemic, product recall, etc., often relies on the computation of a well defined and previously known set of statistics at almost regular intervals. Another example also from the public health domain is performance or safety reporting (e.g., the number of eligible patients that received certain screening and the occurrence rate of surgical site infections) which also involves the computation of well-defined statistics. Therefore, for such reporting purposes, differentially private statistic analyses are a good match to the process. A basic and simple design for the way to share data among horizontally divided healthcare data sets is shown in Figure 4. These distributed data sets, in the case of healthcare, would be gathered from different health information technology systems. For example, surgical site infections would be gathered from hospitals and surgical centers for the infections documented on patients.

The goal of differential privacy in this problem domain is to provide enough accuracy in the shared result set between research groups or institutions to help retain statistical significance of potential analyses

Figure 4. Sharing data using differential privacy in horizontally distributed data sets



while still ensuring enough noise has been added to provide the aspired privacy guarantees; the solution has to also provide a mechanism that can prevent researchers from possibly exploiting some weaknesses of the system by asking too many high accuracy questions. The querying researcher could ask a mixture of questions both with high accuracy (low privacy) and low accuracy (high privacy) epsilon values to meet the desired goals of statistical significance and patient privacy within an allowed total value of accuracy (a 'privacy budget') (Lee, 2008). This challenge forces us to take a closer look at several design considerations like calculating the sensitivity function for each algorithm, the way to determine and control the allowed accuracy, and identifying possible optimization options for the applied techniques. Other challenges also emerge from the nature of the data set, so appropriate values of epsilon specifically for healthcare datasets need to be evaluated to balance this tradeoff between the desired privacy guarantees and the required statistical accuracy.

In the second part of this section, we explore the design and implementation of the case study for healthcare data mining. To serve as a basis for the case study, the medical data set from the Breast Cancer Wisconsin data sets from the Machine Learning Repository of University of California Irvine Lichman (2013) has been utilized. The data set consists of several groups each with multiple instances of breast tissue cells and the related attributes. The authors use the data as one group for the purpose of this case study. Attributes have been used to describe cell instances such as: radius which is the mean of distances from center to points on the perimeter; texture which is the standard deviation of gray-scale values; perimeter which is the perimeter of the cell; area which is the surface area of the cell; smoothness which indicates the local variation in radius lengths; clump thickness which measures the thickness of the clump formed by the cancer cells since these tend to group in multilayers; uniformity of cell size and shape which indicates whether or not the size and shape are uniform; etc. Each instance has one of two possible classes: benign or malignant given as ground truth. There are 699 cell instances from the Breast Cancer Wisconsin data set from UCI where 458 are benign and 241 malignant used as ground truth.

Given the data, we proceed to a discussion of the design considerations and steps in our case study. The first design consideration is to decide on the location to add the noise when implementing a differentially private version of our k-means clustering algorithm. A basic and simple design for the way to share data among horizontally divided data sets was shown in Figure 4. In the non-interactive scenario, the process is fairly straightforward to transform data before releasing the perturbed values to a standard k-means implementation of the algorithm. The results are expected to differ in accuracy from the original data used as input and the results are evaluated using various metrics as presented in the results and discussion sections. The interactive approach requires a second design consideration. For this approach, one can utilize a variation from the standard k-means that works as follows: for a given set of k points in space and an accuracy parameter epsilon, we repeatedly apply an update rule, replacing each center with the calculated noisy average of those data points closer to it than to any other. The privacy budget then depends not only on epsilon, but also on the dimensionality of the data d as well as the number of centers k. Our approach has identified a chance for optimization by utilizing the Parallel Composition property of differential privacy, which lead to dividing the data into disjoint sets before applying the perturbation using Laplacian noise. Note that the individual in charge of this design process must be careful with adding noise to averages since the count used in denominator cannot be changed, but only the summed values.

Differential privacy protects a patient's healthcare data by adding exponentially distributed random noise to the results of a query against a data set that perturbs the data in order to preserve anonymity. Exponentially distributed random noise has some interesting properties that provide privacy guarantees.

For the case study, the differential privacy noise mechanisms can be applied to specify the amount of accuracy (ε) desired, and translate this to the privacy 'budget' that it can guarantee. For example, in the aforementioned breast cancer data set, noise can be added to every numeric variable. The concern, especially within healthcare datasets, is that the addition of random noise, while providing privacy guarantees, will significantly reduce statistical accuracy. In the case of the breast cancer data set, there must be guarantees that the noise doesn't impact the information on the cells. By applying differential privacy against the used dataset, one should be able to either alleviate or confirm the concern. The expectation given the current popularity of differential privacy in research is that one will be able to determine the range of candidate epsilon values to achieve practicality and guarantee an acceptable level of privacy.

In order to observe the effect of the level of noise introduced on the learning performance of the privacy preserving learning algorithm, the case study in this paper varied ε and performed several runs of k-means classification algorithm on the data set. This results in a number of privacy settings that are explained as follows: The first setting utilizes the original data with no differential privacy applied, where the results of this run are considered to be the new ground truth or base line that is utilized in comparing the results before and after applying differential privacy settings. Several program runs were then performed with both interactive and non-interactive settings and with different DP algorithms applied. The interactive implementation utilized C# as programming language and explicit implementation of k-means was utilized to include the noise adding functions in different steps of the algorithm. This implementation also utilized the PINQ package for querying the data set. Note that since our clustering results are binary in terms of benign or malignant designations of cancer cells, one can view the clustering in this case as a binary classifier for the sake of evaluation. The non-interactive setting was implemented in the R programming language and was applied by adding the noise adding functions in compact, well defined-steps before 'releasing' the data set for evaluation. The aforementioned implementations were run with a number of ε _i values applied to the data set, analyzed each result set and averaged errors, and relied on two comparison techniques to judge the algorithms comparisons before and after applying differential privacy noise. This is accomplished by first defining the similarity Jaccard index of clustering results, and second, using the original classifier results as the ground truth/baseline in order to compare with the new version of the dp-classifier.

In the last part of this section, we review and discuss our results. Figure 5 shows the relation between different values of ε and the average query error calculated for an intermediate counting step of the algorithm. To study the classifier performance with and without differential privacy applied, we review the silhouette plots of Figure 6. Note the decrease in the average silhouette width, which the actual mislabeled records are 3 out of 699 data records, which can be considered acceptable for clustering or classification applications such as the case in our dataset, especially considering that $\varepsilon = 0.1$ for these results. In addition, for the non-interactive setting, we evaluated for $\varepsilon = 0.1$ in the similarity of the clustering algorithm before and after applying DP in terms of the contingency table with the co-membership of the observations, using Jaccard coefficient similarity statistic as shown in Figure 7. The similarity coefficient with a value of Cluster Similarity = 0.9306204 indicated high similarity which translates to high accuracy. This result is quite encouraging in general and for this data set, achieved with an acceptable privacy loss of $\varepsilon = 0.1$. Figure 8 shows the silhouette plot for $\varepsilon = 0.05$, which translates to better privacy guarantee than $\varepsilon = 0.1$ of the previous plot, a cluster similarity index of only 0.52 and a decreased accuracy in label prediction. These results can be further fine-tuned to project the desired tradeoff between privacy loss and accuracy based on the application and the end goal of the data

Figure 5. Results of 33 runs using incremental values of epsilon

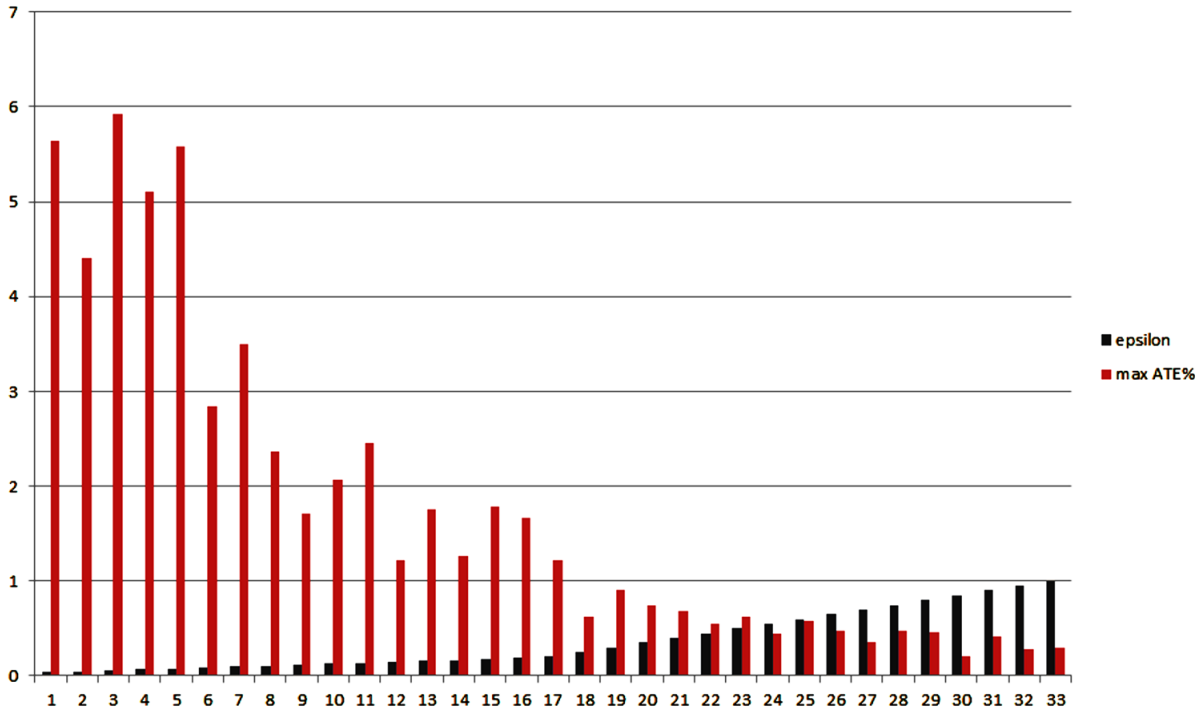
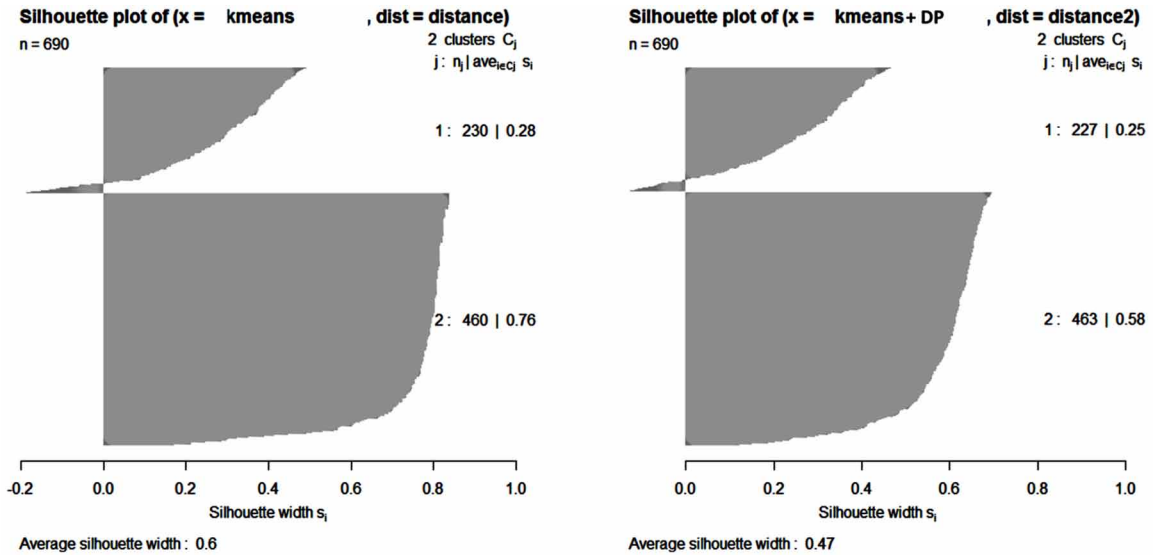


Figure 6. Silhouette plot with no DP applied on the left and with ε -DP applied in non-interactive setting



sharing process. In particular, if the breast cancer data set is shared for the benign-malignant classifier model, then fitting is recommended to ensure a high privacy loss in favor of enhanced accuracy. In this case, a more strict privacy parameter could be applied when utilizing this data set as a test set rather than a training set for such a model.

Figure 7. Contingency table with the comembership of the observations

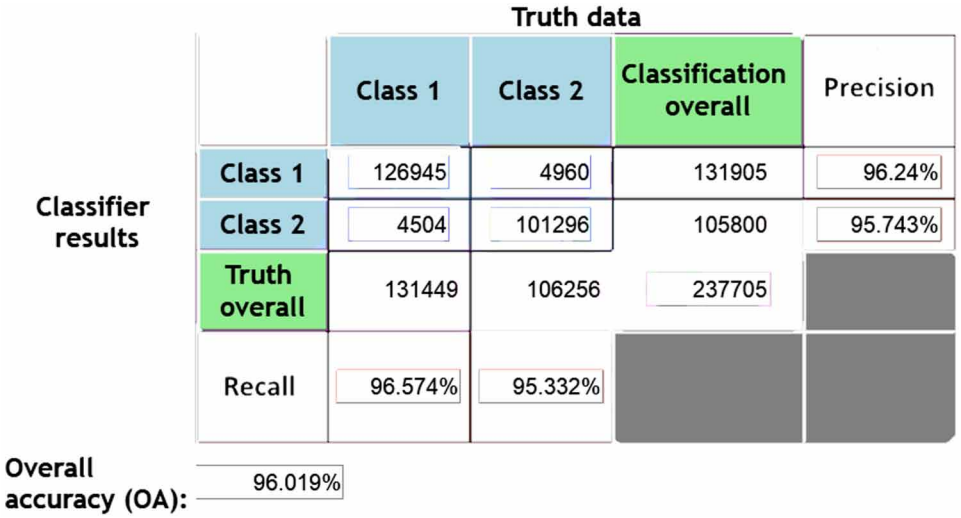
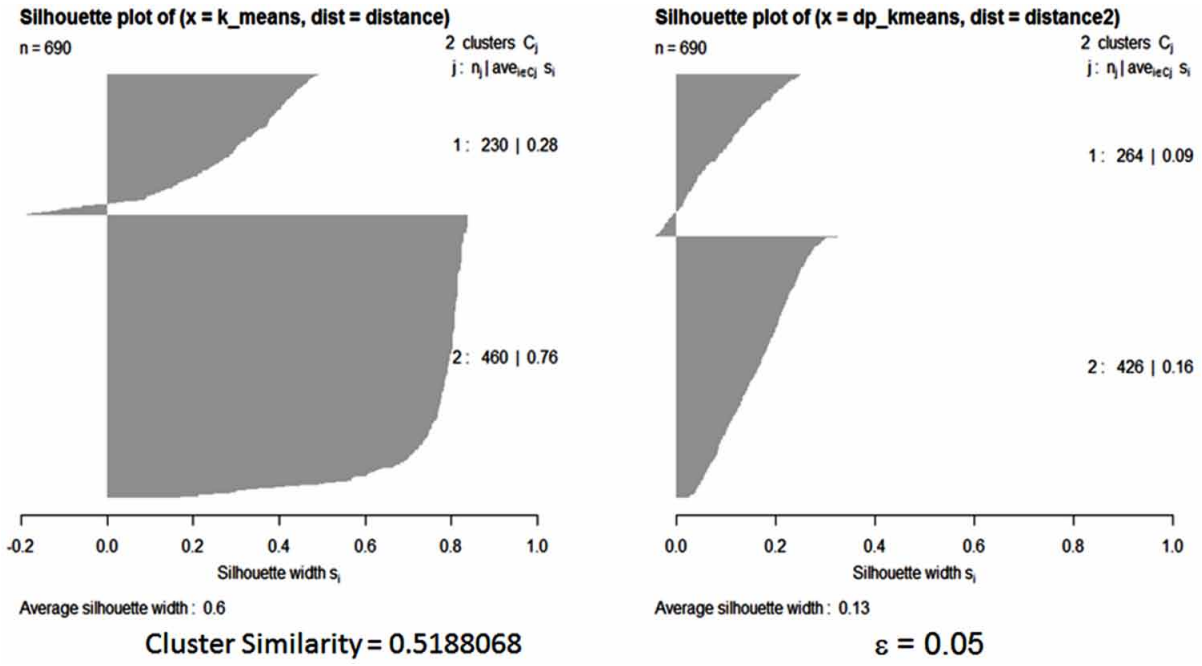


Figure 8. Silhouette plot with $\varepsilon = 0.05$



As for the interactive setting experiment, the results were in general less encouraging. Note that the analysis of the values and optimization ways is a subject of future work. To summarize, our results showed that without much sacrificing privacy $\varepsilon = 0.1$ a 93% similarity was achieved and 96% accuracy relevant to the original clustering performance. In our settings, the non-interactive setting resulted in better metrics than interactive, but this may not occur generally. Specifically, for the breast cancer

data set and similar health care data sets, we observe that the non-interactive setting is in fact the more desirable approach when releasing such sensitive health records for research purposes. Cluster similarity when comparing the ground truth with resulting clusters varied from a ratio of 1 to less than 50% depending on ε values (higher ε means better similarity). Using a synthetic set, created from dp-values of original data enhanced the overall performance and this result needs further evaluation in future work. Finally, as expected, higher ε values lead to a deterioration of privacy. Note that any numeric data can actually be utilized to test, which is one of the strengths of DP and as a result, can applied regardless of what the data actually represents.

FUTURE RESEARCH DIRECTIONS

This section explores future trends in there different areas: guidelines for privacy budgets which explores the way that guidelines can be established to assist in the choice of privacy budgets based on data sets' nature and application domain accuracy and privacy loss tolerance; privacy integrated querying which examines a variation of the data sharing setting models discussed in this chapter that utilizes an interactive design of differentially private mechanisms that can be used in horizontally distributed data; and, private coresets which presents an approach that utilizes private coresets as a spin off to general differential privacy methods.

The first future trend area, guidelines for privacy budgets (Dwork, 2014), are intended to allow the data analyst to understand that while the guarantees of differential privacy are rather strong, they can come at the expense of accuracy. The potentially poor performance in terms of accuracy of such algorithms results directly from the fact that noise must increase with the sensitivity of the query sequence. As a result, more queries means that there will be noisier answers. In addition, any non-interactive solution permitting 'too accurate' answers to 'too many' questions is vulnerable to attack while the 'privacy budget' notion limits the user to a number of allowed queries with low sensitivity. This means that the number of queries where the results are not 'severely' affected is limited. These two observations are also studied in the literature (Dankar et al., 2012) in an effort to assess the practicalities of applying differential privacy methods to real-life problems and data sets.

The second future trend area, privacy integrated querying, such as PINQ (McSherry, 2009), provides an interactive way for data sharing with algorithms to be constructed out of trusted components. These components inherit privacy properties structurally and encapsulate privacy settings rather than require the need of expert analysis and understanding to safely deploy applications in domains like healthcare. This significantly expands the set of possible users and domains of an application. PINQ's implementation focuses on a generic type that supports the same methods as any querying language like SQL, but with an implementation that provides appropriate privacy mechanisms applied before any execution is invoked.

The third future trend area, privacy core sets (Feldman 2009), has found wide-spread usage in a vast host of settings involving very large data sets, with the potential future applicability to differential privacy. "A coreset of a point set P is a small weighted set of points that captures the properties of P." (Feldman 2009). A link is forged between coresets and differentially privacy in the sense that if f a small coreset with low generalized sensitivity in fact does exist (i.e., replacing a single point in the original point set slightly affects the quality of the coreset) this in turn implies the existence of a private coreset for the

same set of queries. This is particularly helpful in settings where the data set to be shared is particularly large and when the main purpose of the data release is to extract general models describing a data set properties.

CONCLUSION

This chapter explored big data privacy and security techniques and policies with a specific focus on differential privacy as utilized for a case study of healthcare data mining. To support the discussion, the *Background* section reviewed big data applications, big data challenges, big data processing technologies, and big data analysis techniques. This was supplemented by the *Big Data Security and Data Privacy Issue* section that focused on differential privacy with a focus on: big data security, differential privacy, properties of differential private algorithms, the impact of noise mechanisms, and models of releasing sensitive data with differential privacy. Using this as a basis, the *Differential Privacy Case Study in Healthcare Data Mining* section presented a case study and: motivated healthcare data sharing and the challenges of the usage of differential privacy; explored the design and implementation; and reviewed and discussed results. To complete the chapter, the *Future Trends* section explored emerging areas including: *guidelines for privacy budgets* for domain accuracy and privacy loss tolerance, *privacy integrated querying* a variant of data sharing, and, *privacy coresets* to augment general differential privacy methods. Overall, this chapter has provided an in-depth examination on big data privacy, big data analytic methods, and recent tools/technologies for privacy preserving big data applications utilizing a real-world scenario in achieving privacy in big data analytics in the domain of healthcare to securely manage the privacy of patient data.

REFERENCES

Arpaia, M. (2013). Leveraging big data to create more secure web applications. *Code as Craft*. Retrieved from http://codeascraft.com/2013/06/04/leveraging-big-data-to-create-more-secure-web-applications/

Barbaro, M., Zeller, T., & Hansell, S. (2006). A face is exposed for aol searcher no. 4417749. *New York Times,* 9.

Bennett, J., & Lanning, S. (2007). The Netflix prize.*Proceedings of KDD cup and workshop*. 35.

Bertino, E., Bernstein, P., Agrawal, D., Davidson, S., Dayal, U., Franklin, M.,... Jadadish, H. V. (2011). *Challenges and Opportunities with Big Data*. Academic Press.

Borthakur, D. (2007). The Hadoop distributed file system: Architecture and design. *Hadoop Project Website*, *11*, 21.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Brown, B., Michael, C., & Manyika, J. (2011). Are you ready for the era of 'big data'. *The McKinsey Quarterly*, *4*(1), 24–35.

Caine, K., & Hanania, R. (2013). Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association*, *20*(1), 7–15. doi:10.1136/amiajnl-2012-001023 PMID:23184192

Chaudhuri, S. (2012). What next? A half-dozen data management research goals for big data and the cloud. *Proceedings of the31st Symposium on Principles of Database Systems*, 1-4. doi:10.1145/2213556.2213558

Cheung, S. (2013). Developing a big data application for data exploration and discovery. *IBM Developerworks*. Retrieved from http://www.ibm.com/developerworks/library/bd-exploration/

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 88–93. doi:10.1109/ICDEW.2013.6547433

Cormode, G., Procopiuc, C., Srivastava, D., Shen, E., & Yu, T. (2012). Differentially private spatial decompositions. *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, 20–31. doi:10.1109/ICDE.2012.16

Cyril, N., & Soman, A. (2015). *Big Data Analysis using Hadoop*. Academic Press.

Dankar, F. K., & El Emam, K. (2012). The application of differential privacy to health data. *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, 158–166. doi:10.1145/2320765.2320816

Dankar, F. K., & El Emam, K. (2013). Practicing differential privacy in health care: A review. *Transactions on Data Privacy*, *6*(1), 35–67.

De Montjoye, Y. A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539. doi:10.1126/science.1256297 PMID:25635097

Dwork, C. (2006). Differential privacy. *Proceedings of 33rd International Colloquium,* 1-12. doi:10.1007/11787006_1

Dwork, C. (2006). Differential privacy. In *Automata, languages and programming* (pp. 1–12). Springer. doi:10.1007/11787006_1

Dwork, C. (2008). An ad omnia approach to defining and achieving private data analysis. In Privacy, Security, and Trust in KDD, PinKDD 2007, (pp. 1–13). Springer. doi:10.1007/978-3-540-78478-4_1

Dwork, C. (2008). Differential privacy: A survey of results. In *Theory and applications of models of computation* (pp. 1–19). Springer. doi:10.1007/978-3-540-79228-4_1

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. doi:10.1126/science.aaa9375 PMID:26250683

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. L. (2015): Preserving statistical validity in adaptive data analysis.*Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, 117–126. doi:10.1145/2746539.2746580

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography* (pp. 265–284). Springer. doi:10.1007/11681878_14

Dwork, C., & Nissim, K. (2004). Privacy-preserving datamining on vertically partitioned databases. In Advances in Cryptology–CRYPTO 2004 (pp. 528–544). Springer. doi:10.1007/978-3-540-28628-8_32

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3-4), 211–407. doi:10.1561/0400000042

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3-4), 211–407. doi:10.1561/0400000042

Dwork, C., & Smith, A. (2010). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, *1*(2), 2.

El Emam, K., Mercer, J., Moreau, K., Grava-Gubins, I., Buckeridge, D., & Jonker, E. (2011). Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak. *BMC Public Health*, *11*(1), 454. doi:10.1186/1471-2458-11-454 PMID:21658256

Feldman, D., Fiat, A., Kaplan, H., & Nissim, K. (2009). Private coresets. In *Proceedings of the forty-first annual ACM symposium on Theory of computing* (pp. 361-370). ACM.

Francis, T., Madiajagan, M., & Kumar, V. (2015). Privacy issues and techniques in e-health systems. *Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research,* 113–115.

Health Insurance Portability and Accountability Act. (n.d.). Retrieved from http://www.hhs.gov/ocr/hipaa

Inukollu, V., Arsi, S., & Ravuri, R. (2014). Security issues associated with big data in cloud computing. *International Journal of Network Security & Its Applications*, *6*(3), 45–56. doi:10.5121/ijnsa.2014.6304

Kendall, D., & Quill, E. (2015). A Lifetime Electronic Health Record for Every American. Washington, DC: Third Way. Available at http://www.thirdway.org/report/a-lifetime-electronic-health-record-for-every-american

Lane, A. (2012). *Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments.* Securosis, LLC. Retrieved from https://securosis.com/assets/library/reports/SecuringBigData_FINAL.pdf

Lane, A. (2013). *Security Implications Of Big Data Strategies*. Dark Reading's Database Security Tech Center Report. Retrieved from http://www.darkreading.com/risk/security-implications-of-big-data-strategies/d/d-id/1139379

Lee, D. G. Y. (2008). *Protecting patient data confidentiality using differential privacy*. Academic Press.

Lichman, M. (2013). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science.

McSherry, F. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 19–30. doi:10.1145/1559845.1559850

Mohammed, N., Chen, R., Fung, B., & Yu, P. (2011). Differentially private data release for data mining. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 493–501. doi:10.1145/2020408.2020487

Peleg, M., Beimel, D., Dori, D., & Denekamp, Y. (2008, December). Situation-Based Access Control: Privacy management via modeling of patient data access scenarios. *Journal of Biomedical Informatics*, *41*(6), 1028–1040. doi:10.1016/j.jbi.2008.03.014 PMID:18511349

Sarathy, R., & Muralidhar, K. (2011). Evaluating Laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy.*, *4*(1), 1–17.

Sujansky, V., Faus, S. A., Stone, E., & Brennan, P. F. (2010). A method to implement fine-grained access control for personal health records through standard relational database queries. *Journal of Biomedical Informatics*, *43*(5Suppl), S46–S50. doi:10.1016/j.jbi.2010.08.001 PMID:20696276

Sweeney. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570.

Tene, O., & Polonetsky, J. (2012). Response: Privacy in the Age of Big Data: A Time for Big Decisions. *Stanford Law Review*. Retrieved from http://www.stanfordlawreview.org/online/privacy-paradox/big-data

Vu, D., & Slavkovi, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, 138–143.

Wang, J., Liu, S., & Li, Y. (2015). A review of differential privacy in individual data release. *International Journal of Distributed Sensor Networks*, *2015*, 1. doi:10.1155/2015/743160

Yadav, C., Wang, S., & Kumar, M. (2013). *Algorithm and approaches to handle large data-a survey.* arXiv preprint arXiv:1307.5437

ADDITIONAL READING

Kamal, S., Dey, N., Nimmy, S. F., Ripon, S. H., Ali, N. Y., Ashour, A. S., & Shi, F. et al. (2016). Evolutionary framework for coding area selection from cancer data. *Neural Computing & Applications*, 1–23.

Mason, C. (2015). Engineering Kindness: Building a Machine with Compassionate Intelligence. *International Journal of Synthetic Emotions*, *6*(1), 1–23. doi:10.4018/IJSE.2015010101

Odella, F. (2016). Technology Studies and the Sociological Debate on Monitoring of Social Interactions. *International Journal of Ambient Computing and Intelligence*, *7*(1), 1–26. doi:10.4018/IJACI.2016010101

Panda, S. K., Mishra, S., & Das, S. (2017). An Efficient Intra-Server and Inter-Server Load Balancing Algorithm for Internet Distributed Systems. *International Journal of Rough Sets and Data Analysis*, *4*(1), 1–18. doi:10.4018/IJRSDA.2017010101

Ripon, S. H., Kamal, S., Hossain, S., & Dey, N. (2016). Theoretical Analysis of Different Classifiers under Reduction Rough Data Set: A Brief Proposal. *International Journal of Rough Sets and Data Analysis*, *3*(3), 1–20. doi:10.4018/IJRSDA.2016070101

Vallverdú, J., Shah, H., & Casacuberta, D. (2010). Chatterbox challenge as a test-bed for synthetic emotions. Creating Synthetic Emotions through Technological and Robotic Advancements, 118-144.

Zappi, P., Lombriser, C., Benini, L., & Tröster, G. (2012). Collecting datasets from ambient intelligence environments. Innovative Applications of Ambient Intelligence: Advances in Smart Systems: Advances in Smart Systems, 113.

KEY TERMS AND DEFINITIONS

Big Data Applications: Data storage, processing, and analysis technologies that are characterized by high velocity, volume and variety.

Differential Privacy: A privacy mechanism that allows statistical databases to be used for analysis without individual records being vulnerable for privacy risks.

Interactive Data Release Model: A data sharing model, where the user interacts with a differentially private version of an algorithm to query a confidential data set.

K-Means Clustering: A clustering algorithm that aims to partition observations into k clusters where each observation is assigned to the cluster with the nearest mean.

Laplace Noise: Random noise generating functions that follow the Laplace distribution and it is the simplest noise model to be used in differential privacy.

Non-Interactive Data Release Model: A data sharing model, where the data is transformed via differential privacy noise perturbation before being released to users.

Sensitivity Gap: Represents the sum of the worst case difference in answers that can be caused by adding or removing an individual's information from a data set and that needs to be reflected in the global sensitivity setting in a differential privacy algorithm.