An Experimental Technique on Potential Issues and Prospective Solution for Preserving Privacy in Big data

Pooja Choudhary, Kanwal Garg

Abstract: Big Data is extremely a large amount of unstructured data coming from different sources along with high speed and is highly defined by 4 V’s that are volume, velocity, variety and value. Big data cannot be handled by conventional methods as they are meant for small structured datasets which are incapable in storing and processing large datasets. In present scenario, Hadoop, Storm, Spark, Flink etc. are certain frameworks which are proposed for storing and processing the data speedily. Big data contains variety of data including person-specific information. This personal information needs to be preserved otherwise publishing data may put the individual’s privacy at risk. Keeping this in view, various anonymity principles, privacy preserving techniques and metrics had been reviewed. Therefore, the premise of the present review work is to elaborate potential issues and prospective solutions for privacy preservation in person-specific information in big data environment. Taking privacy into consideration, this paper reviews various anonymity principles, its techniques and metrics. The objective of this paper is to provide some privacy issues and its perspectivesolutions.

Index Terms: Big Data, Anonymity, Privacy Preserving Data Publishing(PPDP), Privacy Preserving Data Mining(PPDM).

I. INTRODUCTION

Data in world’s database is increasing tremendously. Expansion in network connectivity and data communication technology, makes data sharing among people is common. In India alone, 258.27 million people are sharing data using social networks[2]. The most popular social networks like Facebook, Instagram, WhatsApp, Twitter etc. are generating unstructured data in huge volume. In addition to this, Industrial Development Corporation(IDC)[1] also estimates that the digital data volume will grow 40% to 50% per year. In 2020, it is expected to reach 40ZB.

This massive growth of unstructured data create challenges for traditional methods to process query faster for which they are not meant. Therefore, an efficient framework is required to handle and process this pile of big data. Apache Hadoop, Spark, Storm, Flink are some of such frameworks which work on big dataheaps.

In the recent year, it has been observed that big data frameworks are not efficient enough to secure person-specific confidential information in big data. Data mining techniques, if applied, on such big data, has the capability to reveal such confidential and sensitive knowledge about data. Thus, it can be a cause of threat to the security of individual. In order to get the benefits of big data without attacking the individual’s private territory, it is essential to implementdata protection safeguards in data policies and guidelines for controlled data access at the beginning is not sufficientand a strong mechanism for later phases is also required.

The privacy preserving problem can be better understood by user based methodology[35]. In this methodology, four different types of users, namely, data provider, data collector, data miner and decision maker are identified and their privacy concerns and its respective methods are provided. Understanding the responsibility differentiation among the users two mechanism for preserving privacy are provided, that are, Privacy Preserving Data Publishing (PPDP) and Privacy Preserving Data Mining(PPDM)[30]. In PPDP, new methods and tools are used for publishing proficient information. Generalisation and Suppression are two examples used for making data anonymous at publishing time. In PPDM, data mining technique functions are expanded to work with perturbed data[30]. This mechanism includes modification of data by adding noise, swapping, randomisation, etc. In the upcoming paragraph, different anonymity approaches are reviewed and the issues which leads to affect the privacy are explored.

II. LITERATURE REVIEW

Privacy is a claim of an individual upto what extent his personal contents can be communicated to others[24]. To specify a level of protection against privacy breaches, some well known principles, anonymisation operations and their metrics are reviewed.
reviewed. To make datasets anonymous, Samarati and Sweeney[28] proposed k-anonymity principle where the
author defined the classification of attributes and stated that
if each record is indistinguishable from at least k-1 records,
then it is k-anonymous. But this model have homogeneity
attack on sensitive attributes which was addressed by
Machanavajjhala et al.[25] and proposed l- diversity
principle where sensitive attributes must be “diverse”
within each quasi-identifier equivalence class.
Afterwards, Wang and Fung[31] and Dwork and Lei[15]
proposed (X,Y)-privacy and ε-differential privacy
respectively. Later, Xiao and Tao[33] introduced guarding
nodes replacing k-anonymity and l-diversity. Li et al.[23]
came with t-closeness principle where the distance between
distribution in indistinguishable group and distribution in
whole data of sensitive attribute was not more than a
threshold t. As l-diversity considered only categorical
sensitive attribute Zhang et al.[36] proposed (k, ε)-
anonymity model for protecting numerical sensitive
attributes. To prevent the proximity attack on numerical sensitive identifiers, Li et al.[22] modified (k, ε)-anonymity
principle and named it as ( , m)-anonymity. This principle
demanded that all sensitive values should be isolated in the
whole range. Nergiz et al.[26] proposed -presence where
attacker was not able to identify any individual as being in
the anonymised database with certainty greater than
.
To get the anonymised datasets, the anonymised
techniques were categorised under generalisation,
suppression and perturbation[24]. Generalisation meant
replacing child values with parent values. Typical
generalisation schemes were full domain generalisation[28],
full subtree generalisation[18], single dimension
partitioning[7][18], multi-dimensional generalisation[21]
and bucketisation[32]. Suppression meant deleting values or
replacing it with special values(for example, Asterik ‘*’).
Typical suppression schemes[30] included record
suppression, value suppression, cell suppression, etc. In
perturbation, the original data values were replaced with
some randomised data values in such a way that the result
did not differ significantly. It was based on randomisation.
Some perturbation were permutation[36] to disassociate the
relationship between quasi-identifier and sensitive
attributes, additive random noise[5][14][17] to replace the
original sensitive values using some distribution, data
swapping[16] for exchanging sensitive attribute values
among individual records and condensation[4].
Lastly, a metric was required to measure the information
loss with respect to privacy. It was used for guiding the
anonymisation algorithms to retain the information quality
with minimum distortion. Data utility metric was categorised
under general-purpose, specific-purpose and tradeoff
purpose[30]. General-purpose metric included Generalised
Height[28], LossMetric[18][7] and Discriminbility Metric[7].
Specific-purpose metric included Classification
Metric[7][18] and KL-Divergence[19]. Tradeoff-purpose
metric determined the optimality between privacy and
information requirements at every anonymisation
operations[30].
After reviewing the anonymity principles, techniques and
data utility metrics, it can be concluded that privacy
protection is a complex social issue in big data era. People
have spend a lot of money to preserve their private data
with intention to stop the abuses. To preserve private data,
data need to be anonymised. Use of anonymisation before
data mining techniques cut down the privacy breaches. In
such way, sensitive data mining patterns can be prevented
from being generated during access. But it was observed
that while anonymising data, data quality became inferior.
To make a balance between anonymity of person-specific
data and information loss, certain metrics comes into
picture and optimal metrics can be used.

III. POTENTIAL ISSUES
There are some issues that are faced by privacy
preservation approaches from beginning to end. Some of
them are
Secondary Purpose[9]: In routine, data are collected when
we shop, use public transport, access service sites using cell
phones or any other electronic devices. This data is
collected for primary purposes like determining time for
suitable train from home to college, searching areas nearby.
But other than primary, it can be used for secondary
purposes as well, for example, business use of customer
data to make promotional offers. But this data contain
highly sensitive information which if revealed put
individuals at risk. So, the extent to which anonymisation is
required to be applied for protecting useful data is a
challenging task to be accomplished.
Misinformation[9]: The openness of people over social
media become the most effective channel of misinformation.
Misinformation is fake or inaccurate information which is
spread unintentionally or intentionally over media and tend
to spread by people to their friends informng aboutthe
underlying issue. Misinformation lead users with serious
and destructive impacts. For example, the celebrities and
other public figures’ respect get turmoil in few seconds for
which they took years to earn. It is difficult to decode what
information is valuable and what is not. From the privacy
perspective, it is a serious issue in defining characteristics
of data containing personal records.

Multiple Release[30]: A database is useful for different
users with different purposes. Suppose there is person
specific information in Table T(Sex, Age, Political Party
Affiliation, Past arrests, Race, Acquitted charges). One
recipient(CBI Officer) is interested in classification
modelling of target attribute “Past arrests” with attributes
(Political Party Affiliation, Sex, Age). Another user (such
as a social service department) is interested in clustering
analysis on (Political Party Affiliation, Age, Race). If
single release (Sex, Age, Race, Political Party

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

Retrieval Number: H10360688S319/19@BEIESP

IJITEE
Affiliation) is made for both purposes then information is released unnecessarily. And if both are released after applying anonymity techniques then it is difficult to prevent them from join attack by attacker.

Sequential Release[30]: Sequential Release means when new information become available, data is released in same sequence as before continuously. This suffers from the problem of unsorted matching attack where the tuples are in sameorderineveryrelease. Under this, data publisher has released tables T1, T2, …, Ti−1 previously and now releases the next table Ti , where all Ti are projections of same underlying table T. And the join ofTi−1 previously and now releases attacker an easy access to information. Unlike the multiple release publishing, all previous released table cannot be modified, only attempt of privacy violation protection is dependent on next anonymising table Ti.

Continuous Data[30]: Continuous data releases table Ti when there is any insertion or deletion of records and publisher already released tables T1, T2, …, Ti−1. All releases share same databaseschema and suffers from complementary release attacks. If adversary knows about the timestamp and quasi-identifiers of a victim, then victim’s privacy is at risk as attacker precisely come to know about victim’s record in released table.

Background Information[30]: Background information is what already known to adversary. In this, adversary with known information infers sensitive values from equivalence class of sensitive attribute. The attacker alienate some values from the set of sensitive attribute values, for example, an employee know that today his boss has an appointment with doctor. With this information, the attacker from released database come to know that his boss is HIV positive. Dwork[30][15], in his research paper, has shown that absolute privacy protection is impossible due to the presence of background knowledge.

Granulated Access to Personal Information[9]: The contradict nature between privacy and commercial interest/research is a major issue. Implementing privacy seeks the removal of certain fields whereas research seeks the release of very detailed data. Apparently, choosing an appropriate level of granularity is a challenge of its own. Providing row-level access, column-level access and cell-level access can derive other sources, identify sensitive attributes and support wide range of analytics respectively, all this need to be saved from adversaries.

Multidatabase Query Inferences[9]: Inferencing is a process in which unauthorised user synthesise the sensitive attributes from the responses that he receive or, in other words, from the combination of non-sensitive attributes in different databases user get to know about sensitive values. Detecting and removing the inferences is an exhaustive and complex task. New emerging technologies such as data mining, data ware-house, web, multilevel database have inference problems and can be considered as hot topic for future research work.

IV. PROSPECTIVE SOLUTIONS

Secondary Purpose: A risk mitigation data model is to be used in privacy preservation for making it useful for secondary purposes. This approach controls the access request of data according to the trust level and risk associated with such data exposure. Certain approaches have been proposed, for example, an evolutive approach[11] was proposed by Diaz–Lopez et al. that make use of dynamic counter measure for risk based access control systems, Al Aqeeli et al. proposed risk mitigating data disclosure algorithm[6] which consider the risk measure formula, etc. Misinformation: Big data contains misinformation/noise that needs to be removed otherwise it affect the fame of a per-son. For example, adversary after mining infers some interpretation belonging to a person but that actually does not. To handle misinformation, certain approaches like Right-click Authenticate[27][13], Cognitive Psychology[20] Social Diffusion Model[10] and 3D Simulation[37] were proposed over different social media. Recently, WhatsApp starts a campaign against the misinformation with tagline “Share Joy Not Rumours”[3].

Different Release Attacks: Certain releases of anonymous table suffers from different attacks. For instance, Unsorted matching attacks, complementary release attacks, homogeneity attacks, etc. To solve homogeneity attack Machanavajjhala et al.[25] proposed l-diversity model where sensitive attribute values should be considered as “diverse” in each equivalence class of quasi-identifier. Sweeney[29] pointed out that unsorted matching, temporal and complementary release attacks were found in anonymity principle. Further, in his research paper, unsorted matching attack solution is provided where rows should be shuffled randomly on every release. Xiao and Tao[34] proposed m-invariance idea and an anonymisation method for solving continuous realeaseproblem.

Background Information: As adversary know about the background information of a person personally. Therefore, protection of individual privacy cannot be provided fully. And also, Dwork and Lei[15], in his research paper, has shown that absolute privacy protection is impossible due to the presence of background knowledge.

Multi-database Query Inferences: Inference problem from multi-database was controlled by perturbative and non-perturbative techniques[12][8]. In perturbative category, methods like rounding, micro-aggregation, data-swapping etc. were used to distort the original dataset before publishing. In non-perturbative, data was not distorted but partially suppressed or there is reduction of details in the original dataset. Accurate result is given when query satisfies certain conditions. Methods of table restriction, query restriction and cell suppression can be used for non-perturbative category.

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-8S3, June 2019

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

Retrieval Number: H10360688S319/19@BEIESP
V. CONCLUSION
In this paper, it is concluded that big data contain some person-specific or personal information that need to be preserved from adversaries making it useful for secondary purposes. Different techniques and data utility metric provide balance between information loss and privacy preservation. Generally, preserving privacy is complex and exhaustive task but important to stop the threats. Different models, for instance, k-anonymity, l-diversity, t-closeness, differential privacy, etc. have been introduced so far for privacy preservation but even these models confront numerous issues. However, some issues are combat with solutions but these solutions are not enough. Hence, this area still requires attention of researchers.

REFERENCES


AUTHORS PROFILE

Pooja Choudhary is pursuing Ph.D. in Computer Science & Applications from Department of Computer Science &Applications, Kurukshetra University, Kurukshetra. She completed her MCA from Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. Her research area is Privacy Preservation In DataMining.

Kanwal Garg is an Assistant Professor at Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. He holds an experience of 18 years. He received his Ph.D. from GJU Science &Technology, Hisar. His area of research includes Big data, Data Mining and Warehousing, Web Mining, Data Stream and OLAP cubes. He has published about 80 papers in National and International Journals.