Review on Privacy Preservation by Applying Scalable MapReduce BottomUp Generalization (MRBUG) Technique

Umar Y. Solanki
Department Of Computer Engineering, Mescoe,
SavitribaiPhule Pune University,
Maharashtra India

Rubeena A. Khan
Department Of Computer Engineering, Mescoe,
SavitribaiPhule Pune University,
Maharashtra India

ABSTRACT
Privacy is one of the most concerned issues in data publishing. Personal data like electronic health records and financial transaction records are usually deemed extremely sensitive although these data can offer significant human benefits if they are analyzed and mined by organizations such as disease research Centre. The emerging research field in data mining, Privacy Preserving Data Publishing (PPDP) [11], targets these challenges. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. It aims at developing techniques that enable publishing data while minimizing data distortion for maintaining utility and ensuring that privacy is preserved.

Keywords
privacy, privacy preserving data mining(ppdm), k-anonymity, suppression, generalization ,mapreduce.

1. INTRODUCTION
Consider a data holder, such as a Life Insurance Corporation (LIC) or a medical institution, that has a privately held collection of person-specific, field structured microdata(each record in the collection is of a different entity). Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? These data are analyzed and mined by organizations. This can bring considerable economic loss or severe social reputation impairment to data owners. Hence, data privacy issues need to be addressed urgently before data sets are analyzed or shared.

2. THREATS TO PRIVACY
Example: Re-identification by linking multiple datasets
Healthcare research organization (HRO) recommends the government collect data having attributes which include the patient's ZIP code, birth date, gender etc. Insurance Commission (IC) is responsible for purchasing medical datasets from data publishers. Thus Insurance related information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals.

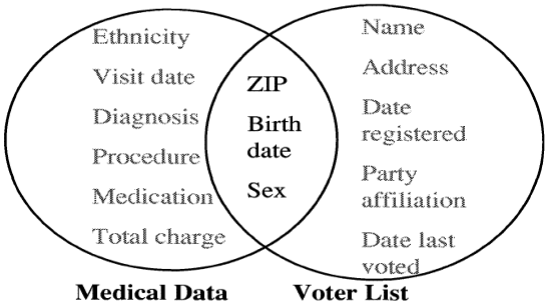

Figure 1. Linking to re-identify data.

The example above provides a demonstration of re-identification by directly linking (or "matching") on shared attributes.

Table 1. Voter list Data

Name	BirthdDate	Gender	Zipcode
Ajay	**21/01/76**	**Male**	**53715**
Amit	01/01/81	Male	55410
Rajashri	10/01/84	Female	90210
Abdul	19/04/72	Male	02174
Tom	21/02/96	Male	02237

Table 2. Hospital patient Data

BirthDate	Gender	Zipcode	Disease
21/01/76	**Male**	**53715**	Flu
11/02/98	Female	65423	Diabetes
14/09/87	Male	90213	Headache
25/06/76	Male	55312	Fever
22/09/87	Female	02174	Fever

From voters data we know “Ajay” is the male born on 21/01/76 staying at zip code 53715. Therefore even though Ajay’s name is not mentioned in the Hospital data we can infer it from voters data and conclude he is having “Flu” which leads to breach in his privacy.

3. BASIC DEFINATIONS

The attributes in a database table comes under three categories namely Key attribute, Quasi attribute, Sensitive attribute.

3.1 Key attribute

An attribute denoted by 'K' consists of values which is the most unique value for to identify the individual from dataset 'S'. Key attributes are used as primary key to identify a record, such as employee id and Social Security Number.

3.2 Quasi attribute

A set of non-sensitive attributes $\{Q_1,Q_2, \ldots,Qp\}$ of a dataset 'S' is called a quasi-identifier, if these attributes can be linked with external data to uniquely identify (can be called as candidate key) at least one individual in the general population. Quasi-identifier (Qi) attributes are those, such as age and zip code, that in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in the micro data belong. Unlike key attributes, quasi attributes cannot be removed from the micro data, because any attribute is potentially a quasi attribute.

3.3 Sensitive attribute

A dataset 'S' consists of values which the user selects as most sensitive attribute. These attributes is what the researchers need, so they are always released directly. Examples of Sensitive attributes are diagnostic report, salary and disease.

3.4 K-Anonymization

Table T is k-anonymous with respect to attributes $X_1, ...,X_d$ if every unique tuple $(x_1, ..., x_d)$ in the (multiset) projection of T on $X_1, ...,X_d$ occurs at least k times. That is, the size of each equivalence class in T with respect to $X_1, ...,X_d$ is at least k.

4 EXISTING WORK

Previous researchers have done the following work in the field of privacy preservation. Many techniques have been proposed previously to preserve privacy.

4.1 K-Anonymity

Latanya sweeney [1] presented a model named kanonymity,It is a popular approach for data anonymization. With k-anonymity an original data set containing personal information can be transformed so that it is difficult for an intruder to determine the identity of the individuals in that data set. A k-anonymized data set has the property that each record is similar to at least another k-1 other records on the potentially identifying variables. Table 3 shows a 2-anonymised table where there are at least 2 records for each value combination of Race,Birth,Gender. While there are several -anonymization algorithm proposals in the literature only a few are suitable for use in practice. Iyengar shows how to attack a very flexible (and highly combinatorial) formulation of -anonymity using a genetic algorithm. The algorithm may run for hours, and because it is an incomplete stochastic search method, it cannot provide any guarantees on solution quality. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly[13], µ-Argus[13] and k-Similar provide guarantees of privacy protection.

Table 3: Example of k-Anonymity where k=2 and Quasi identifiers are (Race,Birth,Gender).

Race	Birth	Gender	ZIP	Disease
Black	**1965**	**M**	02143	short breadth
Black	**1965**	**M**	02146	chest pain
Black	**1965**	**F**	02137	hypertension
Black	**1965**	**F**	02133	obesity
White	**1963**	**M**	02133	diabetes
White	**1963**	**M**	02134	Fever
Black	**1964**	**M**	02131	chest pain
Black	**1964**	**M**	02139	cholera
Black	**1968**	**M**	02142	Flu
Black	**1968**	**M**	02145	Diabetes
White	**1976**	**F**	02176	Jaundice
White	**1976**	**F**	02175	hypertension

Datafly is a System for Providing Anonymity in Medical

Data by automatically generalizing, substituting, inserting and removing information as appropriate with-out losing many of the details found within the data.

The µ-Argus system, like the Datafly System, makes decisions based on bin sizes, generalizes values within fields as needed, and removes extreme outlier information from the released data. µ-Argus program is written in C++ and runs under Windows on a PC.

4.2 Top-Down Specialization for Information and Privacy Preservation

The top-down approach is used to iteratively specialize the data from a general state into a special state. A generalization taxonomy tree (see figure 2) is specified for each categorical and continuous attribute in a virtual identifier. The top-down specialization starts from the top most solution cut and pushes down the solution cut iteratively by specializing some value in the current solution cut until violating the anonymity requirement . The top down approach serves a natural and efficient structure for handling categorical and continuous attributes. The anonymity requirements are guided by maximizing the information utility and minimizing the privacy specificity. TDS generalizes a given table to satisfy a broad range of anonymity requirements without sacrificing significantly the usefulness to classification.

Number of features makes TDS practical such as handling both categorical and continuous attributes, handling multiple virtual identifiers, Scalable computation. [2]

4.3 Bottom-Up Generalization: A Data Mining Solution to Privacy Protection

The bottom-up generalization approach works iteratively to generalize the data. A hierarchy tree (see figure 2) with leaf nodes representing domain values and parent nodes representing less specific values is constructed. Records are generalized by a sequence of generalizations, where each generalization replaces all child values "c" with their parent value "p" in a hierarchy tree, thus leading to an anonymized dataset. An iterative bottom-up generalization approach is used to achieve the required K-anonymity while preserving the usefulness of the generalized data to classification. It's a scalable solution that examines at most one

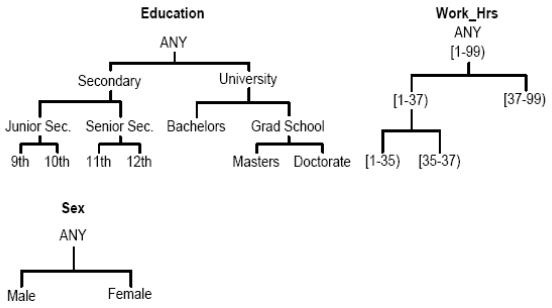

Figure 2. Taxanomy tree based on the quasi attributes(education,work hours, gender).

generalization per attribute in the virtual identifier in each iteration, where the work for examining one generalization is proportional to the number of (distinct) virtual identifier values that are actually generalized.[3]

4.4 Sensitive Attribute based Non-Homogeneous Anonymization for Privacy Preserving Data Mining

suppression: a well-studied technique for masking sensitive information, this technique performs non-homogeneous anonymization, which reduces information loss and provides high degree of data utility which might be more adequate for data mining purposes. Based on the sensitive attribute, non-homogeneous anonymization technique (generalization and suppression) is applied to the identified quasi attributes and the non sensitive attributes are directly published[5]

4.5 Anatomy based privacy preservation

Anatomy is the process of releasing all the quasi-identifier and sensitive data items directly in two separate tables. This approach protects the privacy and capture large amount of correlation in micro data by Combining with a grouping mechanism. A linear-time algorithm for calculating anatomized tables that obey the diversity privacy requirement is developed which minimizes the error of reconstructing micro data [6].

5 RELATED TERMINOLOGIES

5.1 Generalization

In this method, individual values of attributes are replaced by with a broader category. This operation replaces some values with a parent value in the taxonomy of an attribute. For example, the age value of '23' can be generalized by '20 < =Age = 30'.

5.2 Suppression

This operation replaces some values with a special value (e.g. a asterisk `*'), indicating that the replaced values are not disclosed. Typical suppression schemes include record suppression, value suppression, cell suppression, etc.

5.3 Atomization

This operation does not modify the quasi-identifier (QID) or the sensitive attribute (SA), but de-associates the relationship between the two. Anatomization-based method releases the data on QID and the data on SA in two separate tables.

5.4 Permutation

This operation de-associates the relationship between a quasi-identifier and a numerical sensitive attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.

5.5 Perturbation

This operation replaces the original data values with some synthetic data values, so that the statistical information computed from the perturbed data does not differ significantly from the statistical information computed from the original data. Typical perturbation methods include adding noise, swapping data, and generating synthetic data.

5.6 Bottom-Up Generalization

is one of the efficient k-anonymization approach. K-Anonymity, where the attributes are suppressed or generalized until each row is identical with at least k-1 other rows. Bottom-Up Generalization (BUG) approach of anonymization is the process of starting from the lowest anonymization level which is iteratively performed. Consider a generalization G : {c} → p. Let Rc denote the set of records containing c, and let Rp denote the set of records containing p after applying G.generalization is represented as

GEN: Child(q) -> q

where

q is a domain value

&

Child(q) consists of all child domain values of q.

The concept of anonymity was proposed in [1]. Bottomup

generalization was used to achieve anonymity in Datafly

system [13]. It assumed a single virtual identifier that includes all attributes that could potentially be used to link the data to an external source.

Algorithm : The bottom-up generalization
1: **while** *R* does not satisfy the anonymity requirement **do**
2: **for all** generalization *G* **do**
3: compute *IP*(*G*);
4: **end for**;
5: find the best generalization *Gbest*;
6: generalize *R* by *Gbest*;
7: **end while**;
8: output *R*;

5.7 MapReduce Bottom Up Generalization (MRBUG)

The traditional approach specified by the previous researchers works well where we have less volume of data that can be accommodated by standard database servers, or up to the limit of the processor which is processing the data. But when it comes to dealing with huge amounts of data, it is really a tedious task to process such data through a traditional database server.

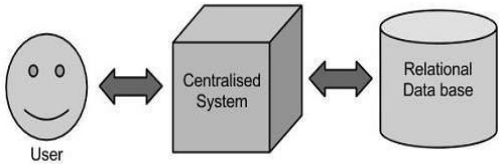

Figure 3: Traditional approach for data processing

To overcome the drawback of traditional system Google proposed a programming model called MapReduce. MaprReduce divides the task into small parts and assigns those parts to many computers connected over the physical/virtual network, and collects the results to form the final result dataset

MapReduce job consists of two primitive functions, Map and Reduce, defined over a data structure named key-value pair (key, value). Map is a function which parcels out task to other different nodes in distributed cluster. Reduce is a function that consolidate the task and resolves results into single value.

MRBUG which is an extension of centralized BUG is a highly developed Bottom-Up Generalization approach which improves the scalability and performance of traditional centralized BUG. Time required to process a large dataset also reduces linearly because MapReduce provides task level parallelization which means that multiple mapper or reducer tasks in a MR job are executed simultaneously on data partitions, wherein data partition run MRBUG Driver on data set input splits, it combines all anonymization levels of the partitioned data items and then apply generalization to original data set without violating the k-anonymity.

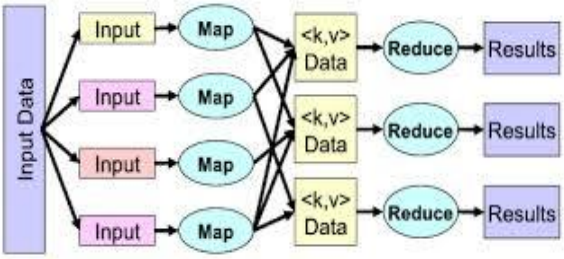

Figure 4. MapReduce Framework

The Hadoop MR framework is fault-tolerant since each node in cluster had to report back with status updates and completed work periodically. The Master Node(name node) send periodical heart beat signals to the data notes, if a data node remains static for longer interval than the expected ,then a master node notes it and re-assigns that task to other nodes. A multiple MR jobs are required to accomplish task. So, a group of MR jobs are orchestrated in one MR driver to achieve the task. MR framework consists of MR Driver and two types of jobs.

One is Information Loss per Privacy Gain (ILPG) Initialization and second one is ILPG Update. The MR driver arranges the execution of jobs. Hadoop which provides the mechanism to set global variables for the Mappers and the Reducers. The best Generalization which is passed into Map function of ILPG Update job. In Bottom-Up Approach, the data is initialized first to its current state. Then the generalizations process is followed so that k -anonymity is not violated. That is, to climb the Taxonomy Tree of the attribute from bottom to up till required Anonymity is achieved. [10][11].

6 OUR ANALYSIS

Sr. No	Author	anonymization operation used	Benefits	Limitations
1	Irit Dinur Kobbi Nissim	Perturbation	Reduces information loss ,achieves high degree of Data Integrity	Reduction in data utility.
2	P.Usha, R.Shriram, S.Sathishkumar	Suppression	non-homogeneous anonymization can be extended to several sensitive attributes	problem of large-scale data anonymization
3	Benjamin C. M. Fung, ,Ke Wang, Yu	Specialization	Scalable, high information gain, classifiers can be applied to compress dataset.	Privacy loss
4	Ke Wang, Yu, P.S,Chakraborty, S	Generalization	Scalable, high privacy gain, high degree of data utility	Information loss
5	Xiao and Y. Tao	Anatomization	Minimizes error of reconstructing micro data	Not scalable, reduction in data utility does not support classifiers.

7 CONCLUSION & FUTURE WORK

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining by comparing various methodologies used by earlier researchers. For data provider, his privacy-preserving objective is to effectively control the amount of sensitive data revealed

to others. To achieve this goal, he can utilize security tools to limit other's access to his data.To achieve the privacy-preserving various methods from different research fields are required. We have reviewed recent progress in related studies, and discussed problems awaiting to be further investigated. We hope that the review presented in this paper can offer researchers different insights into the issue of privacy-preserving data mining, and promote the exploration of new solutions to the security of sensitive information.

To achieve this goal, we propose a novel technique called MRBUG which is scalable and at the same time provides a high degree of privacy gain. The algorithm helps quantify the effects of various parameter settings and data preparation methods on performance as well as anonymization quality.

There are a number of promising areas for future work.

In particular protecting privacy through anonymity at the same time preservation of data integrity. Despite its intuitive appeal, it is possible that non-integrity preserving approaches to privacy (such as random perturbation) may produce a more informative result in many circumstances. Indeed, it may be interesting to consider combined approaches, such as k-anonymity over only a subset of potentially identifying columns and randomly perturbing the others. A better

understanding of when and how to apply various privacy-preserving methods deserves further study, optimal algorithms will be useful in this regard since they eliminate the possibility that a poor outcome is the result of a highly sub-optimal solution rather than an inherent limitation of the specific technique.

8 ACKNOWLEDGMENT

I take this opportunity to express my deep sense of gratitude to Principal of my college Dr. A.A. Keste, and our Head of Computer Department Dr. N.F Shaikh for providing me with best facilities, indispensable support, priceless suggestions as well for most valuable time lent as and when required. I also thank my family for their continuous encouragement and support. I thank my friends for their help in collecting information.

9 REFERENCES

[1] latanya sweeney," k-anonymity- a model for protecting privacy", international journal of uncertainty, puzziness and knowledge-Based Systems Vol. 10, No. 5 (2002) 557-570

[2] Benjamin C. M. Fung ,Ke Wang, Yu, "Top-Down Specialization for Information and Privacy Preservation"

[3] Ke Wang, Yu, P.S,Chakraborty, S, " Bottom-up generalization: a data mining solution to privacy protection".

[4] B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACMComput. Surv.,vol. 42, no. 4, pp.1-53, 2010.

[5] P.Usha R.Shriram S.Sathishkumar," Sensitive Attribute based Non-HomogeneousAnonymization for Privacy Preserving Data Mining", ISBN No.978-1-4799-3834-6/14/$31.00©2014 IEEE

[6] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. 32nd Int'l Conf. Very Large Data Bases(VLDB'06), pp. 139-150, 2006.

[7] Syst., vol. 33, no. 3, pp. 1-47, 2008.B. Fung, K. Wang, L. Wang and P.C.K. Hung, "Privacy- Preserving Data Publishing for Cluster Analysis," Data Knowl.Eng., Vol.68,no.6, pp.552-575, 2009.

[8] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol.19, no. 5, pp. 711-725, May 2007.

[9] Hassan Takabi, James B.D. Joshi and Gail-Joon Ahn, "Security and Privacy Challenges in Cloud Computing Environments".

[10] J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," Comm. ACM, vol. 51, no. 1, pp. 107-113,2008.

[11] Dean J, Ghemawat S. "Mapreduce: a flexible data processing tool," Communications of the ACM 2010;53(1):72–77.DOI:10.1145/1629175.1629198.

[12] Lei Xu, Chunxiao Jiang, (member, ieee), Jian Wang, (member, ieee),Jian Yuan, (member, ieee), and Yong Ren, (member, ieee)," information security in big data:

[13] Privacy and Data Mining" , 10.1109/ACCESS.2014.2362522L. Sweeney. Datafly: A system for providing anonymityin medical data. In International Conference on DatabaseSecurity, pages 356–381, 1998.

10 AUTHOR PROFILE

Umar Y. Solanki Department of Computer Engineering,

MES College of Engineering, Pune, Maharashtra, India

Education Details- B.E (CE) from GSM College of Engineering, Pune, Maharashtra, India. Currently pursuing M.E (CE) From MES College of Engineering, Pune, Maharashtra, India

Rubeena A Khan She is Assistant Professor in the Department of Computer Engineering MES College of Engineering, Pune, Maharashtra, India. She has completed her bachelor's degree in Computer Engineering from SavitriBai Phule Pune University and Master's degree in Computer Engineering from Bharti VidyaPeeth Deemed University, Pune, Maharashtra, India. She is currently pursuing her Ph.D in speech synthesis from Shri Jagdishprasad Jhabarmal Tibrewala University (JJTU) in Jhunjhunu, Rajasthan, India.