Privacy Preserving Data Publishing in Cloud

M. P. Karthikeyan
MCA
Assistant Professor in department of Maths (CA)
Sri Ramakrishna College of Arts & Science

***Abstract*:- The numbers of organizations which aggregate and assign in-collect important special data for an assortment of clashing need which in addition to numerical and common aspect research. In these position the data dealer is usually allow with difficulty. On the one side it is more wanted to preserve the anonymity and important message of existence. On another side it is also adequacy of the data for research.**

This paper covered the initial on the approach of anonymity that are explained with good manner to original identification or with good value to the delicate quality. A considerable valuation that shows to pass the data to the available of high quality data that respects different meaningful concept of privacy that is it is possible to do this efficiently for large data sets. Cloud computing is changing the way that organizations supervise the data because of physique and minimum cost and pervasive process. Privacy preserving has commenced as an essential entanglement with the presence of active cloud computing. In this concept we describes the numerous characteristic anonymization privacy preserving method recycled in the cloud computing.

1. INTRODUCTION

Personal information is collected, stored, analyzed, and distributed in the course of everyday life. In the medical domain, the US Department of Health and Human Services has announced a major initiative toward digitizing the patient records maintained by hospitals, pharmacies, etc. [2]. In the United States, three independent credit reporting agencies maintain databases of personal finance information that are widely used in credit evaluation [3,4] Enterprises supported their business by procuring information technology infrastructure and developing their software on top of that infrastructure.

Supermarkets and other retailers maintain and analyze large databases of customer purchase information, collected by way of various affinity and discount programs. For example, when a customer makes a purchase using its “Club Card,” the Safeway supermarket chain records data about the transaction, including “the amount and content of your purchases and the time and place these purchases are made” [5]. On the surface, this appears harmless, yet there is the potential for abuse. For example, in a Los Angeles court case, Robert Rivera sued Vons grocery store (owned by Safeway) after a slip-and-fall incident. During negotiations, Mr. Rivera’s attorney claimed that Vons had accessed his client’s shopping records, and planned to introduce at trial information regarding Rivera’s frequent purchases of alcohol, implying that he was drunk at the time of the accident [6].

Cloud computing presents a model in which information technology infrastructure is leased and used according to the need of the enterprise. The benefit of this model is that it converts capital expenditure of an enterprise into operational expenditure. Cloud is described as a convenient model using efficient computing resources stressing on four deployment models. Private cloud is solely operated for an organization by either itself or a third party. Public cloud is available for general public use and is owned by an organization selling cloud services. Community cloud provides an infrastructure that is shared by several organizations, also called federation of clouds.

Hybrid cloud is a composition of two, more clouds or multi-clouds (community, private, public).Cloud computing is fast becoming a popular option for renting of computing and storage infrastructure services (called Infrastructure as a Service or IaaS) for remote platform building and customization for business processes (called Platform as a Service or PaaS) and for renting of business applications as a whole (called Software as a Service or SaaS). Privacy concerns arise whenever sensitive data is outsourced to the cloud. By using encryption, the cloud server (i.e. its administrator) is prevented from learning content in the outsourced databases.

1.1 PRIVACY-PRESERVING DATA PUBLISHING

A typical scenario for data collection and publishing is described in Figure 1. In the data collection phase, the data publisher collects data from record owners (e.g., Alice and Bob). In the data publishing phase, the data publisher releases the collected data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data. In this survey, data mining has a broad sense, not necessarily restricted to pattern mining or model building. For example, a hospital collects data from patients and publishes the patient records to an external medical center. In this example, the hospital is the data publisher, patients are record owners, and the medical center is the data recipient. The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis.

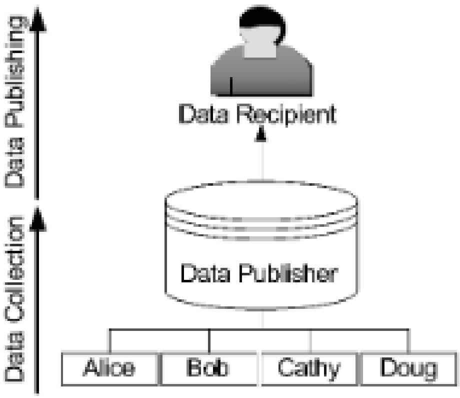


Figure 1.1 Data collection and data publishing

There are two models of data publishers [Gehrke 2006]. In the untrusted model, the data publisher is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions [Yang et al. 2005]; anonymous communications [Chaum 1981; Jakobsson et al. 2002]; and statistical methods [Warner 1965] were proposed to collect records anonymously from their owners without revealing the owners' identity. In the trusted model, the data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher; how-ever, the trust is not transitive to the data recipient. In this survey, we assume the trusted model of data publishers and consider privacy issues in the data publishing phase. In practice, every data publishing scenario has its own assumptions and requirements of the data publisher, the data recipients, and the data publishing purpose. The following are several desirable assumptions and properties in practical data publishing:

The non expert data publisher. The data publisher is not required to have the knowledge to perform data mining on behalf of the data recipient. Any data mining activities have to be performed by the data recipient after receiving the data from the data publisher. Sometimes, the data publisher does not even know who the recipients are at the time of publication, or has no interest in data mining. For example, the hospitals in California publish patient records on the Web [Carlisle et al. 2007]. The hospitals do not know who the recipients are and how the recipients will use the data. The hospital publishes patient records because it is required by regulations [Carlisle et al. 2007] or because it supports general medical research, not because the hospital needs the result of data mining. Therefore, it is not reasonable to expect the data publisher to do more than anonymize the data for publication in such a scenario. In other scenarios, the data publisher is interested in the data mining result, but lacks the in-house expertise to conduct the analysis, and hence outsources the data mining activities to some external data miners. In this case, the data mining task performed by the recipient is known in advance. In the effort to improve the quality of the data mining result, the data publisher could release a customized data set that preserves specific types of patterns for such a data mining task. Still, the actual data mining activities are performed by the data recipient, not by the data publisher.

The data recipient could be an attacker. In PPDP, one assumption is that the data recipient could also be an attacker. For example, the data recipient, say drug research companies, is a trustworthy entity; however, it is difficult to guarantee that all staff in the company is trustworthy as well. This assumption makes the PPDP problems and solutions very different from the encryption and cryptographic approaches, in which only authorized and trustworthy recipients are given the private key for accessing the clear text. A major challenge in PPDP is to simultaneously preserve both the privacy and information usefulness in the anonymous data.

Publish data, not the data mining result. PPDP emphasizes publishing data records about individuals (i.e., micro data). Clearly, this requirement is more stringent than publishing data mining results, such as classifiers, association rules, or statistics about groups of individuals. For example, in the case of the Netflix data release, useful information may be some type of associations of movie ratings. However, Netflix decided to publish data records instead of such associations because the participants, with data records, have greater flexibility in performing the required analysis and data exploration, such as mining patterns in one partition but not in other partitions; visualizing the transactions containing a specific pattern; trying different modeling methods and parameters, and so forth. The assumption for publishing data and not the data mining results is also closely related to the assumption of a non expert data publisher. For example, Netflix does not know in advance how the interested parties might analyze the data. In this case, some basic "information nuggets" should be retained in the published data, but the nuggets cannot replace the data.

Truthfulness at the record level. In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life. Consider the example of patient records. The pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug [Emam 2006]. If a published record does not correspond to an existing patient in real life, it is difficult to deploy data mining results in the real world. Randomized and synthetic data do not meet this requirement. Although an encrypted record corresponds to a real life patient, the encryption hides the semantics required for acting on the patient represented.

1.2 THE ANONYMIZATION APPROACH

In the most basic form of PPDP, the data publisher has a table of the form

D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes),

where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; Quasi Identifier (QID) is a set of attributes that could potentially identify record owners; Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories

[Burnett et al. 2003]. The four sets of attributes are disjoint. Most works assume that each record in the table represents a distinct record owner.

Anonymization [Cox 1980; Dalenius 1986] refers to the PPDP approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed. Even with all explicit identifiers being removed, Sweeney [2002a] showed a real-life privacy threat to William Weld, former governor of the state of Massachusetts. In Sweeney's example, an individual's name in a public voter list was linked with his record in a published medical database through the combination of zip code, date of birth, and sex, as shown in Figure. Each of these attributes does not uniquely identify a record owner, but their combination, called the quasi identifier [Dalenius 1986], often singles out a unique or a small number of record owners. According to Sweeney [2002a], 87% of the U.S. population had reported characteristics that likely made them unique based on only such quasi-identifiers. In the above example, the owner of a record is re identified by linking his quasi identifier. To perform such linking attacks, the attacker needs two pieces of prior knowledge: the victim's record in the released data and the quasi-identifier of the victim. Such knowledge can be obtained by observation. For example, the attacker noticed that his boss was hospitalized, and therefore knew that his boss's medical record would appear in the released patient database. Also, it was not difficult for the attacker to obtain his boss's zip code, date of birth, and sex, which could serve as the quasi-identifier in linking attacks.

In this survey, we review recent work on anonymization approaches to privacy- preserving data publishing (PPDP) and provide our own insights into this topic. There are several fundamental differences between the recent work on PPDP and the previous work proposed by the official statistics community. Recent work on PPDP considers background attacks; inference of sensitive attributes, generalization, and various notions of data utility measures, but the work of the official statistics community does not. The term "privacy-preserving data publishing" has been widely adopted by the computer science community to refer to the recent work discussed in this survey article. In fact, the official statistics community seldom uses the term "privacy-preserving data publishing" to refer to their work. In this survey, we do not intend to provide a detailed coverage of the official statistics methods because some decent surveys already exist [Adam and Wortman 1989; Domingo-Ferrer 2001; Moore 1996; Zayatz 2007].

Typically, the original table does not satisfy a specified privacy requirement and the table must be modified before being published. The modification is done by applying a sequence of anonymization operations to the table. An anonymization operation comes in several flavors: generalization, suppression, anatomization, permutation, and perturbation. Generalization and suppression replace values of specific description, typically the QID attributes, with less specific description. Anatomization and permutation dissociate the correlation between QID and sensitive attributes by grouping and shuffling sensitive values in aid group. Perturbation distorts the data by adding noise, aggregating values, swapping values, or generating synthetic data based on some statistical properties of the original data. Below, we discuss these anonymization operations in detail.

1.3 DE-IDENTIFICATION IN DATA PUBLISHING

Numerous organizations collect and distribute micro data (personal data in its raw, non-aggregate form) for purposes including demographic and public health research. In most cases, attributes that are known to uniquely identify individuals (e.g., Name or Social Security Number) are removed from the released data. However, this fails to account for the possibility of combining other, seemingly innocuous, attributes with external data to uniquely identify individuals. For example, according to one study, 87% of the population of the United States can be uniquely identified on the basis of their 5-digit zip code, sex, and date of birth [6]. The uniqueness of such attribute combinations leads to a class of "linking" attacks, where individuals are "re-identified" by combining multiple (frequently publicly-available) data sets. This type of attack was demonstrated by Sweeney, who was able to combine a public voter registration list and the de-identified patient data of Massachusetts's state employees to determine the medical history of the state's governor [6]. Concern over this type of attack has mounted in recent years due to the ease with which data is distributed over the World Wide Web.

1.4 THE HIPAA PRIVACY RULE

In the medical domain, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) included a number of provisions related to personal privacy. In response to this legislation, the U.S. Department of Health and Human Services issued the regulation "Standards for Privacy of Individually Identifiable Health Information," commonly known as the HIPAA Privacy Rule, with which covered entities were required to demonstrate compliance by 2003. The HIPAA Privacy Rule specifically addresses de-identification, and provides two distinct sets of requirements [7]. By satisfying one of these two provisions, data may be exempt from many of the regulations concerning personally-identifiable health information.

The first provision is deliberately vague, stating that a covered entity may determine that health information is not individually identifiable if [7]: A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

(i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and

(ii) Documents the methods and results of the analysis that justify such determination

In contrast, the second provision (the so-called Safe Harbor) is very specific, and quite restrictive, requiring that eighteen specific types of information, including names and geographic information, be removed entirely for any person (e.g., patients, doctors, etc.) before the data can be considered de-identified [7]. From a technical perspective, neither of these provisions is entirely satisfactory. The first provision should be considered a statistical expert.

The second provision is more precise, but necessitates removing much of the information that is most useful in public health studies (e.g., geography and dates). Throughout this thesis, we seek to provide anonymization techniques that balance rigorous standards of privacy with the often competing goal of releasing useful and informative data. is not explicit about what information is sensitive, what constitutes a “low” risk, or who should be considered a statistical expert. The second provision is more precise, but necessitates removing much of the information that is most useful in public health studies (e.g., geography and dates). Throughout this thesis, we seek to provide anonymization techniques that balance rigorous standards of privacy with the often competing goal of releasing useful and informative data.

1.5 PRIVACY LEGISLATIONS IN CLOUD COMPUTING

Privacy is an essential human right, enshrined in the Universal Declaration of Human Rights and International treaty of Political and Civil Rights[8]. Article 12 of the Universal Declaration of Human Rights states that “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks.” In Europe, the Charter of Fundamental Rights of the European Union (2000) became legally binding in European Union law as part of the Lisbon Treaty (in force since December 2009). EU Directive 95/46/EC, and e-privacy and electronic communications Directive 2002/58/EC covering also data retention, are the main legal instruments in Europe covering privacy and the processing of personal data.

According to a report of the Business Software Alliance in February 2012, it was observed that there is a huge divide between developed and developing countries in terms of adequate legislation for protection of personal data. The report also highlighted that even among developed countries there are conflicting data protection regulations which could hamper transfer of personal data across borders. For example, some developed countries are considering restricting provision of cloud services only to local companies within the country. An international approach is necessary to bring together the multiple data protection regimes and harmonize the business rules for providers and protection rules for the citizen. Without greater coordination at international level on government policies the main advantage and efficiency of cloud computing which is to be able to move data and software services freely across borders will not be achieved.

1.6 ANONYMITY FRAMEWORK & DEFINITIONS

This section gives an overview of the problems considered throughout this thesis. We begin with a single input relation R, containing non-aggregate personal data collected by a centralized organization. As in the majority of previous work on this topic, we assume that each attribute in R can be uniquely characterized by at most one of the following types based on knowledge of the application domain:

• **Identifier** Unique identifiers (denoted *ID*), such as *Name* and *Social Security Number* are
Removed entirely from the published data.1

• **Quasi-Identifier** The quasi-identifier is a set of attributes {Q1, ...,Qd} that is externally
available in combination (either in a single table, or through joins) to the data recipient.2
Examples include the combination of *Birth Date*, *Sex*, and *Zip Code*.

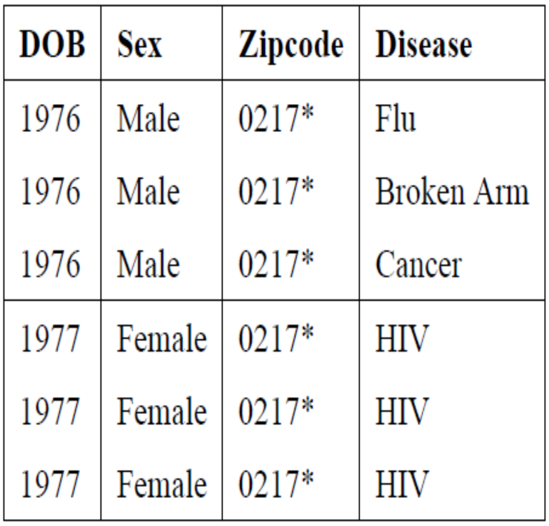
• **Sensitive Attribute** An attribute *S* is considered sensitive if an adversary should not be
Permitted to uniquely associate its value with an identifier. An example is a patient’s *Disease*
Attribute.

We consider the problem of producing a sanitized *snapshot* of R [Q1, ...,Qd, S], which we denote R*[Q1, ...,Qd, S], that is intended to limit the risk of a linking attack.3 Throughout the remainder of this section, we will describe R∗ in terms of an abstract *bucketization*. Specifically, it is convenient to think of R* as horizontally partitioning R[Q1, ...,Qd, S] into a set of non-overlapping *equivalence classes* R1, ...,Rm, each with identical quasi-identifier values. A more thorough discussion of generalization techniques is in Section 1.4, but notice that it is possible to represent generalized tables in this way. For example, the generalization in Table 1.2 divides the records from Table 1.1 into two equivalence classes. Throughout the thesis, we will assume bag semantics unless otherwise noted.

Name	DOB	Sex	Zipcode	Disease
Anand	1/5/76	Male	02173	Cancer
Rajesh	2/18/76	Male	02173	Broken
Suresh	2/24/76	Male	02174	Arm
Anu	5/8/77	Female	02177	Flu
Banu	11/10/77	Female	02174	HIV
Stella	12/1/77	Female	02175	HIV
				HIV

Table 1.1 Original Data (R[ID,Q1, ...,Qd, S])

The first anonymity requirement we consider is k-anonymity, an intuitive means of protecting
individual *identity* that was originally proposed by Samarati and Sweeney. Originally,
k-anonymity was motivated by the idea that each individual in the released data should blend into a crowd. That is, no individual in R* should be uniquely identifiable from a group of size smaller than k on the basis of its quasi-identifier values.



DOB	Sex	Zipcode	Disease
1976	Male	0217*	Flu
1976	Male	0217*	Broken Arm
1976	Male	0217*	Cancer
1977	Female	0217*	HIV
1977	Female	0217*	HIV
1977	Female	0217*	HIV

Table 1.2 Generalized View (R*[Q1, ...,Qd, S])

Definition (k**-Anonymity)** Sanitized view R* is said to be k-anonymous if each unique tuple in the projection of R∗ on Q1, ...,Qd occurs at least k times.

1.7 RECODING TECHNIQUES

In their seminal work, Samarati and Sweeney proposed techniques for generating sanitized view R* using *generalization* and *suppression*. In the Statistics literature, this approach is often called *recoding*.) Informally, the idea is to replace quasi-identifier values with more general ("semantically consistent") values. For example, a Zipcode value could be generalized by suppressing the least significant digit. Subsequently, these ideas have been refined and extended in the literature. From our perspective, the proposed recoding techniques can be roughly categorized along three main dimensions. Each recoding technique implicitly places a set of constraints on the space of possible anonymizations. As we will show throughout the thesis, the choice of recoding technique can greatly influence the quality of the anonymized data.

CONCLUSION:

Cloud is achieved demand and improvement in now days. Cloud can attempt so many challenges for balanced scheduling with recent technologies and methodologies. This thesis has not, by and large, attempted to address the higher-level policy issues surrounding data privacy. The computer security community has long drawn a distinction between the ideas of policy and mechanism, and a similar distinction can be made here. Unfortunately, as in security, where it is not always clear what it means for a system to be secure, it is sometimes difficult to precisely define what it means to protect privacy. Those reasonable expectations of privacy and policies designed in accordance with these expectations are dictated by the application at hand. For this reason, it seems unlikely that there will ever be a single catch-all framework for reasoning about all types of privacy and disclosure. The future research direction appears to lie in defining policies high-level statements of "privacy" that are appropriate philosophically, legally, and technically to specific application scenarios, and developing mechanisms that rigorously enforce these policies. Many technical mechanisms have been developed over the years, including authorization, encryption, aggregation, generalization, output perturbation, and more. In the future, we expect that these mechanisms will form the building blocks for enforcing emerging classes of policies.

REFERENCES:

[1] Cloud Computing: Special theme, European research consortium for Informatics and mathematics (ERCIM), ISSN 0926-4981.
[2] Reuters. US pushes digital medical records, July 22 2004.
[3] Equifax. http://equifax.com.
[4] Experian. http://www.experian.com.
[5] Safeway Inc. Privacy policy, May 26 2007. http://www.safeway.com/privacy page.asp.
[6] L. Sweeney. K-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5):557–570, 2002.
[7] U.S. Department of Health and Human Services Office for Civil Rights. HIPAA administrative simplification regulation text, February 16 2006.
[8] Enterprise Privacy Group. (2008). Privacy by Design: An Overview of privacy Enhancing Technologies. Retrieved from www.ico.gov.uk/upload/documents/pdb_report_html/pbd_pets_paper.pdf
[9] K. LeFevre, D.DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005.
[10] K. LeFevre and D. DeWitt. Scalable anonymization algorithms for large data sets. Universityof Wisconsin Computer Sciences Technical Report 1590, 2007.
[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE), 2006.
[12] G. Aggarwal, M. Bawa, P. Ganesan, H. Garcia-Molina, K. Kenthapadi, R. Motwani, U. Srivastava, D. Thomas, and Y. Xu. Two can keep a secret: A distributed architecture for secure database services. In Proceedings of the 2nd Conference on Innovative Data Systems Research(CIDR), 2005.
[13] G. Aggarwal, T. Feder, K. Kenthapadi, R. Panigrahy, D. Thomas, and A. Zhu. Achievinganonymity via clustering in a metric space. In Proceedings of the 25th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS), 2006.