Chapter 1

Introduction

This chapter presents the basic idea, and provides the overview scope of the privacy preservation. Then, the privacy preservation model that this thesis is based on, is briefly presented. Next, an incremental problem for privacy preservation is discussed. Finally the literature review, research design, scopes, and method of the thesis are presented.

1.1 Background

Data privacy is one of the most important issues these days. When the collaboration is to be taking place among partners for obtaining the useful knowledge to achieve a good strategic move, the privacy preservation is a necessity for prevent the privacy breach at all cost. For Thailand, Thai people become aware of privacy of individual information since Act of Computer Prohibition has been enforced from B.E 2550 [30]. The organizations involving the individual information of Thai citizen should focus on two main issues as follows.

Firstly, the infrastructure of information system should be concerned about security for the thorough view of management. The information access control [39, 40] for management privilege of users must be designed carefully. One of main solutions of this topic is the authentication system in database management system technology. Then, the network security that protects the information transfer via the network must be considered. The cryptography [41], firewall technology [40], and etc. can be used effectively to overcome this issue.

In the second issue, since data sharing between business collaborators becomes a common practice currently. The shared data can be used to analyze various purposes, such as knowledge discovery in database, data mining, and aggregation query, for developing the latency of business collaborators. When collaborators share data to the others, the privacy issue must be addressed effectively. Typically all identifiers in the

data such as IDs or names must be removed [6, 7, 8]. Unfortunately, there could be the other datasets, which can overlap to the shared data. The overlap can be established by the common attributes between the two datasets. For example, consider the datasets in Table 1.1, and 1.2. Suppose that the dataset in Table 1.1 is released from the social security office to be used to build a classifier by a data analysis agency. For considering this dataset alone could misjudge that the privacy of the individuals, which contains in this dataset, has been already preserved due to the removal of the identifiers. However, there is another dataset, which is released publicly for voting purpose as shown in Table 1.2. Suppose that an adversary wants to fetch private information about a man named "Somchai" whose age is 27 years old. The adversary can overlap two datasets together using Age, and Sex attributes, subsequently, his salary (30,000) is revealed. The Age and Sex attributes are called quasi-identifier attributes. Salary attribute is called sensitive attribute.

Table 1.1: The released dataset

Age	Sex	Salary
27	Female	15000
25	Female	25000
25	Male	18000
25	Female	20000
25	Male	8000
27	Male	30000

Table 1.2: The public dataset

Name	Age	Sex
Manee	27	Female
Wanjai	25	Female
Tawan	25	Male
Somsee	25	Female
Sutee	25	Male
Somchai	27	Male

Over the past decade, the various privacy preservation models and techniques have been proposed [1-10]. When data are to be shared, such techniques can be applied prior to sharing, and the privacy-preserved data can be used for such proposes. However, shared data are often changed all the time. Applying the privacy preservation techniques to the data each time can result in different versions of privacy-preserved data. Comparing them can lead to a privacy breach, which is called an incremental privacy breach [11].

In this research, we present an algorithm for preserving the privacy of the data when the data are not static, i.e., when the data are appended continuously. The focused privacy preservation model is based on (*k*, *e*)-Anonymous [10], which is one of the most prominent models.

1.1.1 (*k*, *e*)-anonymous model

Based on the (*k*, *e*)-anonymous privacy preservation model, the data to be shared are composed of identifier attributes such as the ID or name, quasi-identifier attributes such as gender or age, and sensitive attributes such as salary or disease [6, 7, 8]. To preserve the privacy, all of the identifier attributes in the data must be removed first. This removal typically leaves only the quasi-identifiers and the sensitive data remain.

Thus, the task is to prevent a linkage between the quasi-identifier attributes and the sensitive attributes [1, 2, 5, 7, 8]. The (*k*, *e*)-anonymous model [10] breaks or un-links the associations by first partitioning the data. Then, the sensitive values are shuffled within each partition. The conditions for the partitioning of the (*k*, *e*)-anonymous model are that the number of distinct sensitive values of each partition must be at least *k*, and the error of each partition must be at least *e*. The error is the difference between the maximum and the minimum of the sensitive values in each partition.

An algorithm that generates optimal partitioning, i.e., minimizing the sum of the errors, has been proposed in [10]. The complexity of the data partitioning algorithm is $O(n^2)$, where n is the number of data tuples.

To illustrate the effectiveness of the (*k*, *e*)-Anonymous model, let us consider the following example. The data in Fig.1a) has the attributes "Age" and "Gender" as the quasi-identifier attributes and the attribute "Salary" as the sensitive attribute. Suppose that the *k* value is set to 3 and the *e* value is set to 2000. Then, the data in Fig. 1.1a) can be partitioned with respect to the *k* and the *e* values, as shown in Fig. 1.1b). As a matter of fact, the number of distinct sensitive values in each partition is at least 3, and the difference between the maximum and minimum value of each partition is at least 2000. After the shuffling, it can be seen that the associations between the sensitive attribute and the quasi-identifier attributes are un-linked, and thus the privacy can be preserved.

D_0

Quasi-identifiers		Sensitive
Age	Gender	Salary
21	male	83,000
23	male	84,000
52	male	85,000
27	female	88,000
53	male	89,000
37	male	90,000
49	female	91,000

a) The Dataset

$P[D_0]$

Quasi-identifiers		Sensitive
Age	Gender	Salary
21	male	83,000
23	male	84,000
52	male	85,000
27	female	88,000
53	male	89,000
37	male	90,000
49	female	91,000

b) The Partitioned Dataset

$P'[D_0]$

Quasi-identifiers		Sensitive
Age	Gender	Salary
21	male	85,000
23	male	83,000
52	male	84,000
27	female	89,000
53	male	88,000
37	male	90,000
49	female	91,000

c) The Shuffled Dataset

Figure 1.1: An Example Dataset for (*k*, *e*)-anonymous Model (*k* = 3, and *e* = 2000)

Once the data are to be queried or shared, the privacy-preserved data can be utilized safely. For example, suppose that a query to determine the sum of the salaries of the males is issued. The result of the query from the non privacy-preserved data in Fig. 1.1a) is 431,000. At the same time, the result from the privacy-preserved data in Fig. 1.1c) is 429,000-433,000. It can be seen that there is only a slight difference between the exact answer and the privacy-preserved answer. This type of query can be considered to be an aggregation query, which is fundamental for any further data analysis, e.g., OLAP or machine learning algorithms. Retaining the ability to answer this type of query means that a high utility of the privacy-preserved data can be obtained. In [10], it is reported that the optimal solutions from the (*k*, *e*)-Anonymous model can achieve less than 10% of the relative error of the aggregation query.

1.1.2 An incremental privacy breach

If we consider only an individual privacy-preserved dataset, privacy can be preserved with regard to the *k* and *e* values. However, considering two or more versions of the datasets from different points in time could lead to the incremental privacy breach. Let us consider the following example. Let *k* be 3 and let *e* be 2000. Suppose that the two versions of privacy-preserved datasets at time t_1 and time t_2 are given as in Fig. 1.2a) and Fig. 1.2b) respectively. (Note that the name attribute in the figure is only for referencing and that both datasets have been privacy-preserved optimally.) Suppose that the attacker knows that Bob's data, i.e., for a 52 years old male, are collected and exist in the dataset at time t_1. Obviously, his data are in partition p_1 of the t_1-released dataset.

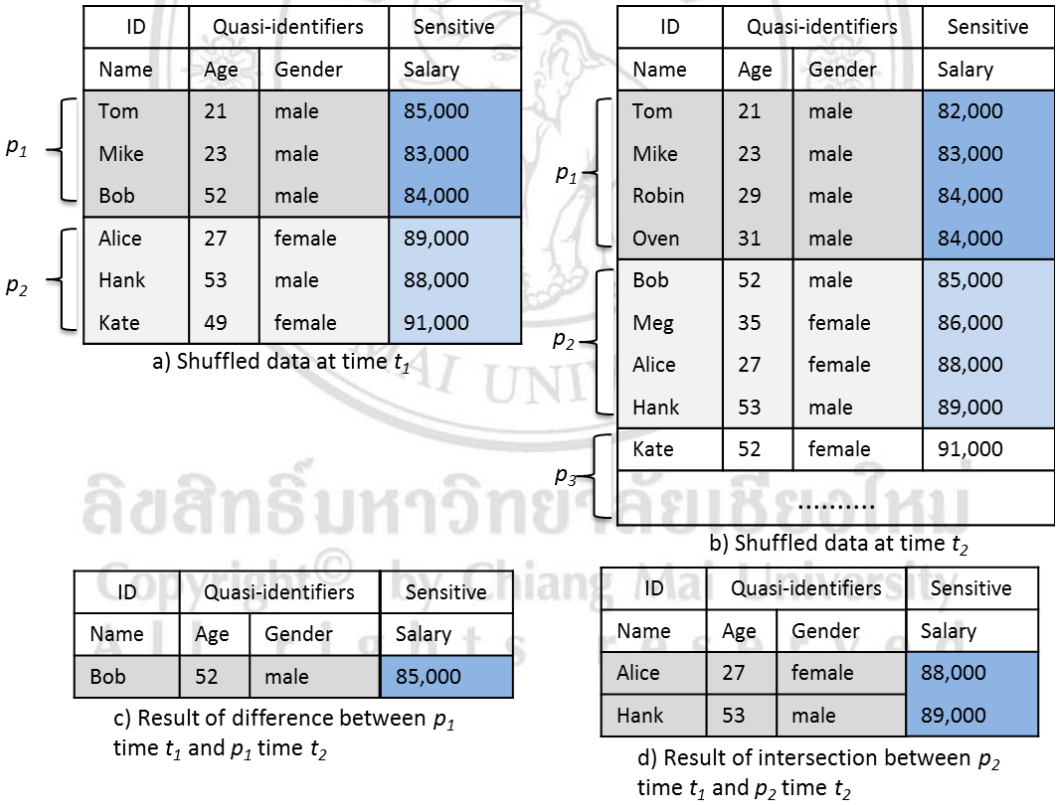

Figure 1.2: An Incremental Privacy Breach Example

Subsequently, when the new version of this dataset is released at time t_2, the attacker can observe that Bob's data are not in partition p_1 at this time. The attacker can investigate the partitions in both versions of the dataset and can identify the difference

between partition p_1 at time t_1 and t_2. Unfortunately, the difference reveals a sensitive value, the salary, of "Bob", as shown in Fig. 1.2c). The other example is when the attacker intersects both versions of the dataset. Fig. 1.2d) shows the result of the intersection between partition p_2 of both datasets. It obviously can reveal that the sensitive attribute value of "Alice" and "Hank" might be 88,000 to 89,000, which causes a privacy breach with regard to the k and e value of the (k, e)-Anonymous model.

In this research, an algorithm that is based on (k, e)-Anonymous to preserve the privacy against an incremental privacy breach is proposed. The proposed algorithm is based on our theoretical studies on the effects of multiple dataset releasing. Such studies can allow us to reduce the computational complexity enormously, especially because it requires considering only the previous version of the dataset against the dataset to be released and not all of the released datasets. At the same time, the privacy is not compromised, i.e., the result of the privacy-preserved dataset is exactly the same as considering all of the released dataset. Thus, the complexity reduces from $O(n^m)$ of the naive approach, where n is the number of tuples and m is the number of released datasets, to $O(pn^3)$ where p is the number of partitions in the previous dataset. We also present the experiment results to show the efficiency of the proposed algorithm on the real-world dataset, as well as the optimal result and the privacy that can always be guaranteed.

1.2 Purposes of the Study

To develop an efficient algorithm for privacy preservation in data incremental scenarios.

1.3 Educational Advantages

The main contribution of this thesis is to develop an algorithm to protect the dataset from incremental privacy breach in incremental data scenarios. Additionally, the efficient algorithms are proposed. These contributions can be applied when the business collaborators want to exchange data to discover the knowledge for business activity improvement.

1.4 Research Design

1.4.1 Define the problems of an incremental privacy breach for each case.

1.4.2 Study the characteristics of (*k, e*)-anonymous model in when data are increased with respect to an incremental privacy breach and minimum summation of error method.

1.4.3 Define the lemma and theorem for development the efficient algorithm to find the optimal solution with respect to minimum summation of error method.

1.4.4 Develop the efficient algorithm to prevent an incremental privacy breach

1.4.5 Evaluate the efficiency (speed) of the incremental algorithm compared with the naïve algorithm.

1.5 Scope

1.5.1 The efficiency of the algorithm is measured by the execution time and the computational complexity.

1.5.2 The experiments are conducted on a real-life data, the Adult Dataset from the UCI Machine Learning Repository [29].

1.6 Method

1.7.1 Review the related literatures about anonymous model, and incremental data processing.

1.7.2 Define the lemmas and the theorems of (*k, e*)-anonymous on incremental data scenarios.

1.7.3 Develop an efficient algorithm from the defined lemmas and theorems.

1.7.4 Design an experiment and define the dataset scenarios.

1.7.5 Conduct the experiments.

1.7.6 Discuss the experiment results.

1.7 Summary

In this chapter, the principle of (*k*, *e*)-anonymous model, the problem of incremental data scenarios of the model, and the methodology to be used in this thesis are presented. In the next chapter, we will present the related works to be used for defining the problem and solution.