Chapter 6

Conclusion

To sum up the contribution of this thesis, in brief, a practical algorithm was proposed to resolve an incremental data privacy breach on (*k*, *e*)-anonymous model [10]. The algorithm was proven as being efficient and effective. The following describes the overview of the study.

In Chapter 1, the (*k*, *e*)-anonymous model was discussed on the room for improvement, i.e. the incremental privacy breach can occur. A few examples are presented to illustrate the cases. To be more specifically, an instance related to the incremental problem was discussed i.e. when the dataset is increased, the new version of the preserved dataset may cause the breach with the previous version of preserved dataset by comparing between them.

In Chapter 2, we first presented the statistic values and news with regards to the privacy breach in the real world. Next, some privacy preservation models were discussed thoroughly for solving the privacy breach problems. Also, a few related incremental algorithms for privacy preservation were discussed. At the last of this chapter, some ideas and methods to solve the incremental issue were introduced.

In Chapter 3, the basic definitions and the problem statement of this thesis were defined. Here, two types of the incremental privacy breach for the (*k*, *e*)-anonymous model are introduced, i.e. different privacy breach and intersection privacy breach. The graphical concepts of the breach are provided to help understanding the problem.

In Chapter 4, we observed the impacts of each case of data increment. Some lemmas and the theorem from the observations were proposed. Then, the proposed algorithm then was conducted from the theorem. Its complexity was analyzed to provide the outlook for its efficiency.

In Chapter 5, first, the experiment procedure, the dataset, and the environment were described. Then, the experiment results were presented. Last, the experimental outcomes were discussed to confirm the theoretical studies.

6.1 Research Contribution

In this work, we address the incremental privacy breach problem based on the (*k*, *e*)-Anonymous model. The characteristics of the effects of multiple dataset released are observed theoretically. We propose that an incremental privacy breach occurs when there is overlap between a partition range of a new dataset and any partition of any existing dataset. According to such studies, we propose a polynomial-time algorithm with $O(pn^3)$ in which *n* is the number of tuples. With the proposed algorithm, only the most recent previously released dataset is required to being considered rather than all of the previous datasets.

In addition, not all of the partitions in the previous dataset must be considered for an incremental privacy breach determination; some of them can be discarded from further computing by our proposed observation. Based on the fact that our proposed work always guarantees the optimal result, we can conclude that our algorithm is both effective and highly efficient, as confirmed by the experiment results.

6.2 Future work

In the future work, the systematic problem, in which the subset of the datasets being queried, and then the attackers can attempt to accumulate the knowledge from each query for the attack, should be investigated.

Also, an additional data modification type such as deletion will be considered. For example, 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. 6.1a) and Fig. 6.1b) respectively. Suppose that the attacker knows that Mary's data, i.e., a 30 years old female, are collected and exists in the dataset at time t_1. Obviously, her data are in partition p_1 of the t_1-released dataset.

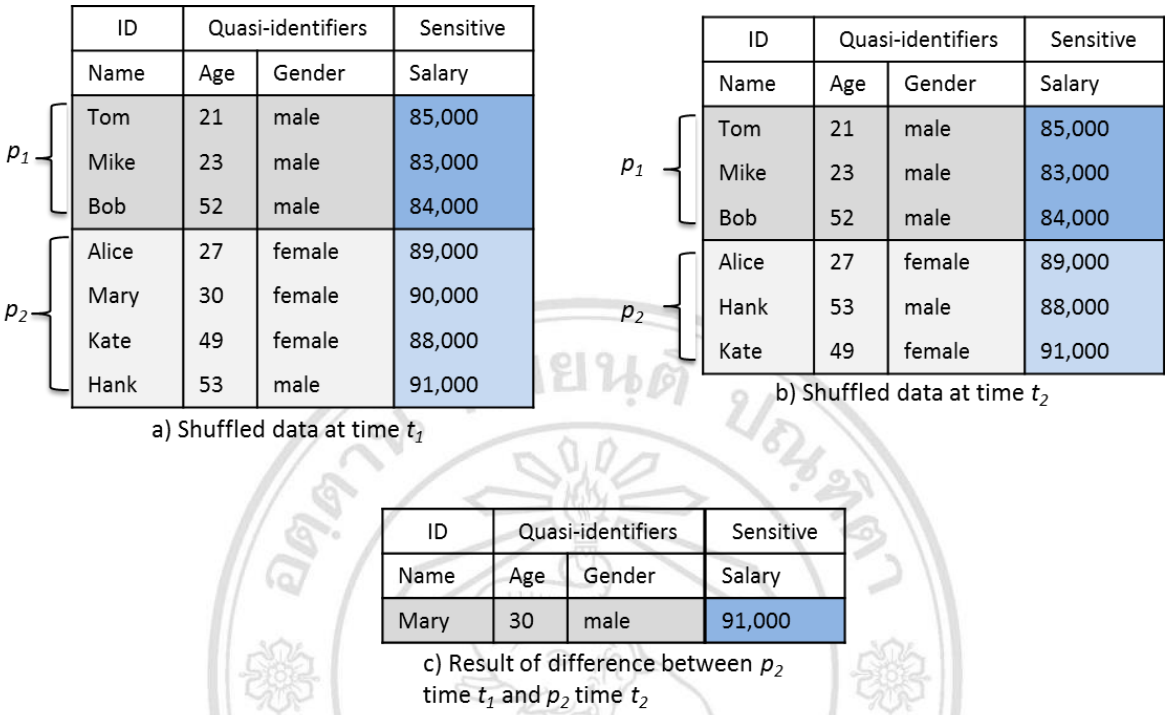

Figure 6.1: A Deleting Privacy Breach Example

Subsequently, when the new version of this dataset is released at time t_2, the attacker can observe that Mary's data are not in partition p_2 at this time. The attacker can investigate the partitions in both versions of the dataset and can identify the difference between partition p_2 at time t_1 and t_2. Unfortunately, the difference reveals a sensitive value, the salary, of "Mary", as shown in Fig. 6.1c).

Another interesting issue is as follows. To apply (k, e) anonymous model with sensitive attribute type is categorical such as "Disease" attribute, defining the total order relationship might be the approach for resolving. In addition, to define the proper values of k and e values, data pre-process might be the method for overcoming this issue.

Last but not least, the concurrency control issue, in which multiple users might add, delete, or query the dataset at the same time, can be considered.