CHAPTER 6

CONCLUSION

This thesis was motivated by the desire to create improved method for frequent itemset mining and to provide a solution for privacy preserving frequent itemset mining when the mining task is performed in-house and outsourced. Privacy and security concerns have become an enduring part of society and commerce. It is increasingly necessary to ensure that useful computation does not violate legal or commercial norms for the safety of personal or organizational data.

In the first technique, the frequent itemset mining using the bloom filter data structure showed significant improvement in time as well as space for both sparse and dense datasets with tolerable level of false positive. Being a probabilistic data structure, this technique provides approximate support count only but the error rate is tolerable even for low support values. However, the performance and storage advantage was considerably more compared to the existing algorithms.

The second research work was a simple alternative framework for securing the data at rest before releasing it for in-house data mining. Organizations performing in-house or on-premise data mining cannot implement commercial database security solutions for protecting the privacy of data. In this work, the original dataset was distorted by means of matrix
manipulation and simple authorization mechanisms were used to provide data privacy. This technique is effective for large datasets and is limited only by the computational resource to process higher orders matrices. But with the existing computational power of the computers in organization, this framework can be effectively adopted easily.

In the case of an objective where the requirement is to preserve privacy with the data mining results, when the data mining task is outsourced to third party service providers, the third technique is preferred. Here, bloom filter data structure along with the cryptography is able to provide the necessary data as well as result privacy. The experimental results proved that the third party analyst cannot decipher any private information (data privacy) nor will be able to gain knowledge about the patterns (knowledge privacy) mined to certain degree. In this research work, the transformed data has achieved the desired quality. The accuracy evaluation metric evaluated how closely the frequent itemsets from transformed dataset matches the corresponding frequent itemsets in the original datasets.

Privacy-preserving data mining remains a difficult problem. Real-world solutions will need to rely on a combination of legal, regulatory, and technological components. It is unlikely that the world would ever reach a point where technological solutions alone can completely guarantee the privacy of individuals while allowing for meaningful exploratory data mining. However, algorithms such as those developed in this thesis remain useful because they precisely identify what security guarantees are possible under different scenarios.
6.1 FUTURE WORK

The three techniques developed in this thesis have accomplished good improvements in solving the problems addressed by them. Nevertheless, it requires further investigations in the following aspects.

- Parallelism can be introduced in all the three algorithms at various stages to further improve performance. Bloom filter insertions and look-ups can be carried out in parallel.
- To further improve the privacy, methods can be devised that allow frequent item set mining to be performed on the distorted data matrix itself without loss in data quality.
- It is important not to forget that intruders and data terrorists will try to compromise information by using various data mining algorithms. Therefore, a PPDM algorithm developed against a particular data mining techniques that assures privacy of information may not attain similar protection against all possible data mining algorithms. In order to provide for a complete evaluation of a PPDM algorithm, it is required to measure its hiding failure against data mining techniques which are different from the technique that the PPDM algorithm has been designed for.