Practical Issues On Privacy-Preserving Health Data Mining

Huidong (Warren) Jin[1,2]

[1] NICTA, Locked Bag 8001, Canberra ACT, 2601, Australia.
Huidong.Jin@nicta.com.au
[2] RSISE, the Australian National University, Canberra ACT, 2601, Australia.

Abstract. Privacy-preserving data mining techniques could encourage health data custodians to provide accurate information for mining by ensuring that the data mining procedures and results cannot, with any reasonable degree of certainty, violate data privacy. We outline privacy-preserving data mining techniques/systems in the literature and in industry. They range from privacy-preserving data publishing, privacy-preserving (distributed) computation to privacy-preserving data mining result release. We discuss their strength and weaknesses respectively, and indicate there is no perfect technical solution yet. We also provide and discuss a possible development framework for privacy-preserving health data mining systems.
Keywords: Data anonymisation, secure multiparty computation, encryption, privacy inference, health data privacy

1 Introduction

Health information, according to the Australian Commonwealth Privacy Act [1], is defined to be

1. *information or an opinion about:*
   (a) *the health or a disability (at any time) of an individual; or*
   (b) *an individual's expressed wishes about the future provision of health services to him or her; or*
   (c) *a health service provided, or to be provided, to an individual;*
   that is also personal information; or
2. *other personal information collected to provide, or in providing, a health service; or*
3. *other personal information about an individual collected in connection with the donation, or intended donation, by the individual of his or her body parts, organs or body substances.*

As important *personal information*, health information is classified as being one type of *sensitive information* [1].

With the development of powerful data mining tools/systems, we are facing the dilemma that a health data mining system should satisfy user requests

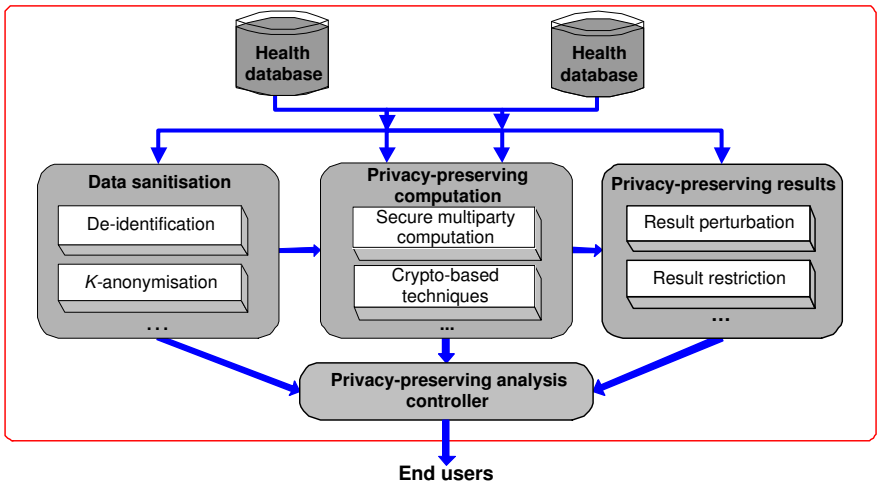

Fig. 1. Illustration of privacy-preserving health data mining system.

for discovering valuable knowledge from databases [2,3,4,5,6], while guarding against the ability to infer any privacy about individuals. The identification of an individual person or organisation (by the third party) should not be able to be made from mining procedures or results that we release. Furthermore, information attributable to an individual person or organisation should not be disclosed. We should develop policy, procedures as well as new techniques for privacy confidentiality with the aim of meeting legislative obligations. We only concentrate on technical issues in this paper.

Unfortunately, *privacy* faces changes over time, and a common understanding of what is meant by "privacy" is still missing [7]. For example, it is not directly defined in the Australian Commonwealth Privacy Act [1]. This fact has led to the proliferation of a wide variety of different techniques for privacy-preserving data mining. They range from privacy-preserving data publishing, privacy-preserving (distributed) computation, and privacy-preserving result release, as illustrated in the middle of Fig. 1. These techniques are crucial to develop privacy-preserving health data mining systems. We review their typical techniques and discuss their pros and cons in Sections 2, 3 and-4, respectively. We try to answer the question whether they are sufficient for a privacy-preserving health data mining system in practice. In last section, we suggest a system development framework in order to make use of their strength to protect health data privacy thoroughly.

2 Privacy-preserving Data Publishing

The first category of privacy-preserving data mining techniques are for privacy-preserving data publishing. These techniques mainly de-identify, perturb, swap,

re-code, anonymise, or even simulate raw data before conducting data mining or publishing [8,9].

A typical real-world example is Public Use Microdata Sample (PUMS) files, provided by U.S. Census Bureau [10]. These data files have been edited to protect the confidentiality of all individuals. For example, they may have been de-identified by

- removing name, address and any other information that might uniquely identify any individual;
- controlling the amount of detail (say, data items that are most likely to be used for identifying unit records are only released in broad categories);
- changing a small number of values - particularly unusual values - and removing very unusual records;
- changing a small number of values or their combinations, adding noise to continuous attributes and suppressing cell counts in tabulated data;
- controlling the modes of access to restrict access to more detailed data.

Similar examples may be found from other statistical agencies including Confidentialised Unit Record Files (CURF) provided by the Australian Bureau of Statistics (ABS) [11] and Statistics New Zealand [12]. The CURF microdata are used widely by universities, government and private sector researchers and analysts.

Latanya Sweeney [13] argued whether such kind of de-identification is enough for data privacy protection. She showed that such kind of protection of individual sources does not guarantee protection when sources are cross-examined: a sensitive medical record, for instance, can be uniquely linked to a named voter record in a publicly available voter list through some shared attributes. To eliminate such opportunities of inferring private information through link threats, Latanya Sweeney [13] introduced a very important model for protecting individual privacy, k-anonymity, a notion that establishes that the cardinality of the answer to a quasi-identifier will be at least k. The larger the k, the more difficult it is to identify an individual using the quasi-identifier. There are several techniques to randomise data to satisfy the k-anonymity. For example, starting from a very generalised data set, Bayardo and Agrawal [14] proposed an efficient technique to specialise the data set to satisfy the k-anonymity. This top-down specialisation is natural and efficient for handling both categorical and continuous attributes. Fung *et al.* [15] presented an efficient algorithm for determining a generalised version of data that masks sensitive information and remains useful for modelling classification. Compared to [14], this approach is greedy, thus does not guarantee optimality. However, they shew that the optimality is not needed for some problems such as classification. Actually the optimal data is the raw data without any generalisation, but such data is overfitting. The greedy algorithm is significantly faster than the one in [14]. However, optimal k-anonymisation is computationally expensive in general [14]. Data mining researchers have also established other randomisation techniques for specific data mining functionality, such as association analysis [16].

Along the direction of k-anonymity [17], data mining researchers have proposed k-anonymity model for continuous attributes [18], templates to limit sensitive inferences [19,20], l-diversity model [21] and (α, k)-anonymous model [22]. The latter two models mainly aim to handle the uniformity for individuals with the same quasi-identifier value, which can bring privacy threats. Template-based privacy-preservation [19,20] used 'confidence' to handicap the dominance of a sensitive value in an equivalence class (with the same quasi-identifier value). These models can increase privacy protection level.

2.1 Strength

Privacy-preserving data publishing techniques are found to be efficient, useful and appropriate for a data custodian (such as statistical agencies) who may make the same dataset available to many thousands of different users. The data custodian only needs to prepare one anonymised version of raw data, and doesn't have to consider which analyses will be conducted. There is no fear of litigation. For example, in Australia, under the *Census and Statistics Act 1905*, the ABS is authorised to release unit record data provided that it is done in a manner that is not likely to enable identification of a particular person or organisation to which it relates.

2.2 Weaknesses

For data privacy protection, there are several weak points if we would like to anonymise raw health data for publishing.

1. The quasi-identifiers required in the k-anonymity, the l-diversity and the (α, k)-anonymity models are quite difficult to specify beforehand, especially when we don't know what kind of information adversaries may have. This renders privacy-preserving data publishing techniques difficult for verification in practice.
2. Data mining results based on these published data can be rather different from the true, and we may have to develop special or complicated data mining techniques to compensate for anonymised or perturbed data.
3. In general, optimal anonymisation, say, k-anonymisation, is computationally expensive [14] in order to maintain health data accuracy as high as possible.

3 Privacy-preserving (Distributed) Computation

In this paper, our definition of privacy-preserving (distributed) computation implies that nothing other than the final computation result is revealed during the whole computation procedure. In other words, intermediate computation results in a data mining procedure don't disclose privacy of the data.

This definition is equivalent to the "security" definition used in the Secure Multiparty Computation (SMC) literature. Developing privacy-preserving (distributed) computation techniques is the mainstream of privacy-preserving data

mining in the literature. Vaidya and Clifton [23] gave a summary of these SMC-based techniques for vertically partitioned or horizontally partitioned data. For these SMC-based techniques, there exists a general solution based on circuit evaluation: Take a Boolean circuit representing the given functionality and produce a protocol for evaluating this circuit [24]. Circuit evaluation protocol scans the circuit from input wires to output wires, processing a single gate in each basic step. When entering each basic step, the parties hold shares of the values of the input wires, and when the step is completed they hold shares of the output wire. Thus evaluating the circuit "reduces" to evaluating single gates on values shared by both parties. This general solution is elegant in its simplicity and generality and proves the existence of a solution, but, is highly inefficient! The reason for this inefficiency is the all-to-all communication operations required during the protocol. Such operations are very costly in large-scale distributed networks [25]. Typically, specific solutions for specific problems can be much more efficient [23]. A key insight is to trade off computation and communication cost for accuracy, i.e., improve efficiency over the generic SMC method. Extending the SMC protocol, Gilburd *et al.* [25] have recently proposed the *k-privacy* definition and k-TTP concepts in order to scale-up SMC to hundreds of parties, where *k-privacy* is defined as the privacy attained when no party learns statistics of a group of less than k parties.

Some privacy-preserving (distributed) computation approaches lies on canonical encryption techniques. Comparing with other techniques, a stronger encryption scheme can be more effective and acceptable in protecting data privacy. For example, homomorphic encryption is a powerful cryptographic tool where certain computation operators are allowed to performed on encrypted data without prior decryption. Wright and Yang [26] used a homomorphic encryption technique to construct Bayesian networks from two parties but without revealing anything about their data to each other. O'Keefe *et al.* [27] used it to establish several protocols for privacy-preserving data linkage and extraction across databases of sensitive information about individuals, in an environment of constraints on organisation's ability to share data and a need to protect individuals' privacy and confidentiality.

A third possible strategy is to only release aggregate data from a party to the others rather than unit records. For example, only data summaries of at least k unit records will be released for final clustering [18]. There are a couple of techniques available to generate clusters based on summary statistics of groups of unit records, such as [28,4].

3.1 Strength

This category of techniques are easy for the data mining community to accept since they lie on well-developed secure computation and encryption techniques.

3.2 Weaknesses

There are several issues should bear in mind when we deploy privacy-preserving (distributed) computation as a solution.

1. Most of the privacy-preserving computation protocols lie on the semi-honest model, which assumes that participating parties follow the prescribed protocol but try to infer private information using the messages they receive during the protocol. Although the semi-honest model is realistic in many settings, there are cases where it may be better to use the "malicious model" in which we try to prevent any malicious behaviour by using more expensive cryptographic techniques.
2. Encryption/decryption operations are usually very costly in large-scale distributed networks [25], especially for the malicious model.
3. Existing efficient SMC protocols are mainly suitable for a proportion of data mining functionalities. A complete but concise set of privacy-preserving computation primitives are still open for further research [23].
4. Last but not the least, these privacy-preserving (distributed) computation techniques alone cannot provide a solution for protecting data privacy even for distributed databases. For example, as illustrated in Example 1, what if data mining results themselves disclose data privacy? More examples about frequent sequential patterns containing identification and sensitive attribute values can be found in [29]. We turn to this problem in the next section.

4 Privacy-preserving Results Release

This category of techniques attempt to disseminate data mining results without undue risk of disclosure of individual information. Put it in other words, the disclosure of discovered knowledge do not open up the risk of privacy breaches, especially *indirect privacy divulgence*. Indirect privacy divulgence can take place by performing *inference* on pieces of superficially "insensitive" information (say, rules or patterns). In other words, inference is the process of deducing sensitive/private information from the legitimate responses received to user queries or data mining results [30]. A *privacy inference channel* indicates a series of released information from which it is probable to infer sensitive information such as identification or sensitive attribute values. Huang *et al.* discussed how to infer private information from randomised data by considering data correlation. In this section, we only discuss the privacy inference from data mining results.

A real-world privacy-preserving result release scenario is to protect the patients' privacy, such as identification and health status, in the healthcare sector. In Australia, for example, the government agency Medicare Australia holds data on drug prescriptions, while each state government holds local hospitalisation data including diagnoses [31,6]. To enhance healthcare quality, government agencies could analyse the health events and release knowledge discovered, e.g., frequent itemsets.

Example 1. Bob gets 2 itemsets from the above healthcare databases:

1. $\{a, b, c\}$ with support 1000, i.e., 1000 patients having a, b and c. a and b indicate, say, drugs while c one condition;
2. $\{a, b\}$ with support 1001;

These frequent itemsets represent a number of individuals as required by the minimum support threshold, and seemingly do not compromise privacy. However, these released frequent itemsets alone can indirectly divulge privacy. For example, Bob can easily infer that if a patient took Drugs a and b, he/she most likely suffered Condition c, which can be sensitive like HIV. In addition, Bob knows one and only one patient in the databases suffering c but not taking Drugs a and b. Through linkage with other data sources, this patient can be re-identified. This results in privacy leakage via linking attacks [22].

There are several strategies we may use in order to present results in a privacy-preserving way. One is **result perturbation** or **sanitisation**, where results are perturbed or sanitised to remove possible inference channels for privacy disclosure before releasing. They can be rounded up, recoded, dropped, suppressed, or perturbed. The second one is **result restriction**, where certain sensitive analyses are restricted if the disclosure of their results casting high risk on privacy. There are a volume of literature in statistical databases about how to restrict statistical analyses or queries [32]. Normally, a privacy-preserving health data mining system needs to use both strategies.

The Remote Data Access Laboratory (RADL) [33] provided by the ABS gives a good industrial example. It provides secure online access to a range of basic CURFs that have previously been released on CD-ROM, as well as new expanded datasets that contain more CURF data than those can be made available on CD-ROM. Using the RADL, at any time from their desktops, researchers are able to interrogate/query CURF data that they are approved to access. The ABS imposes a number of privacy and confidentiality preserving restrictions on the nature of queries and the nature and size of the outputs returned. For example, outputs will only be returned automatically if they pass *print/table limit checks* and *the total output size check* [33]. A similar system, Privacy-Preserving Analytics$^{\text{TM}}$ [34,35], was developed by CSIRO. Different from query support in the RADL, Privacy-Preserving Analytics$^{\text{TM}}$ provides a set of data analytic tools (e.g., exploratory analysis and regression) with the objective of ensuring that no unit record is released and also ensuring that information released about unit records is insufficient to make identification of any individual likely. The system also contains some data mining functionalities like sequential pattern mining [29].

In the data mining community, there are mainly two kinds of research efforts to release data mining results without compromising privacy, based on whether or not system designers have *a priori* knowledge of what is private (or sensitive). (1) If private information is given before hand, new technologies are developed to perturb original data in order to protect these sensitive information [36,37]. Unfortunately, these techniques may inevitably introduce some fake knowledge, e.g., infrequent itemsets may become frequent. (2) The other efforts focus on individual's *privacy*, which is mainly concerned with the *anonymity* of individuals.

That is, without any background knowledge of what is sensitive, we have to protect the anonymity of individuals in the analysed data [13]. This is a concrete objective. For example, Atzori *et al.* [38] proposed the k-anonymous patterns concept, where a pattern P is said to be k-anonymous if its support is 0 or larger than an anonymity threshold k for a given binary database. A pattern, such as $(a \vee b) \wedge \neg c$, is defined as a logical sentence built by AND ($\wedge$), OR($\vee$), and NOT ($\neg$) logical connectives on binary attributes in the database. For instance, $supp((a \wedge b) \wedge \neg c) = 1$ in Example 1. Atzori *et al.* [38] studied privacy inference channels that involve only frequent itemsets. They further developed two methods for blocking these privacy inference channels by distorting released frequent itemsets [39,38]. However, there exists more complicated inference channels. For example, Fienberg and Slavkovic [40] discussed possible inferential disclosure following the release of information on one or more association rules. Jin *et al.* [29] proposed *k-anonymous sequential patterns* and *α-dissociative sequential patterns* as two concrete objectives for privacy protection in order to release frequent sequential patterns.

4.1 Strength

Privacy-preserving result release is an essential component of a privacy-preserving health data mining system. As we discussed above, privacy-preserving (distributed) computation alone cannot construct a privacy-preserving data mining system.

Compared with privacy-preserving data publishing techniques in Section 2, there is no need for special and complex analysis techniques to compensate for anonymised data. Standard statistical analysis or data mining tools are generally applicable, therefore the analysis and the interpretation of the results are made easier. For example, there is no need to adjust data mining techniques just because noise has been added to one or more attributes.

The released results can be generated from original raw data, and they can be quite accurate even after sanitisation [29].

Sometimes it is easier to sanitise or restrict data mining results rather than anonymise data for privacy protection.

4.2 Weaknesses

Some possible weaknesses of privacy-preserving result release techniques are listed as follows.

1. The result sanitisation procedures can be computationally expensive, though there exist some efficient sanitisation techniques for specific data mining functionalities, say, privacy-preserving sequential pattern release [29].
2. The sanitised results may still disclose some sensitive information even after deliberate computation. For example, there exist privacy inference channels for frequent itemsets besides those discussed and blocked by Atzori *et al.* [39,38].

3. At this stage, the sanitisation procedures are mostly developed for a specific form of knowledge, for example, privacy-preserving sequential pattern release [29]. There is little investigation for privacy-preserving result release when data mining results from different functionalities are released.
4. Sometimes, a denial of service in a privacy-preserving system can be informative for privacy divulgence too.

5 Discussions and Possible Solutions

There are a number of research and development efforts on privacy-preserving data mining techniques/systems, ranging from privacy-preserving data publishing, privacy-preserving (distributed) computation, and privacy-preserving result release. We have discussed that they are suitable for different scenarios, e.g., most of privacy-preserving result release techniques have been developed for specific patterns or rules. They need to be incorporated with different legislation and policy. Usually such a single technique alone cannot provide an ideal technical solution [41] for privacy-preserving health data mining systems.

One possible privacy-preserving health data mining system development framework is illustrated in Fig. 1. The main motivation is to integrate various existing techniques together in order to take use of their strength at the appropriate time. The basic idea is to add a privacy-preserving data mining controller to determine which kinds of techniques should be used for a coming analysis query. An original data set may be first de-identified and/or k-anonymised, then the analysis is computed in a privacy-preserving way if data are distributed, and final data mining results are examined and sanitised before releasing. It must assess risk, especially taking into account interactions between different forms of releases (data or data mining results). It also can assess utility, accounting for data or analyses that become unreleasable. Such a health data mining system may as well be developed from a single data mining functionality first, and then add more functionalities. It stops when it is becomes impossible to include more data mining functionalities. For a set of users with different data mining requirements, a privacy-preserving health data mining system may provide modes of access appropriate to the level of data mining available.

Besides developing privacy-preserving data mining techniques, we still need to placing restrictions on how the data mining results or released data are used. For example, every organisation and individual user may sign a legal undertaking with the data mining service provider. A user should

- not attempt to identify particular persons or organisations implied in the released data or data mining results or both;
- not disclose, directly or indirectly, the private information to individuals or organisations who have not signed an appropriate legal undertakings with the service provider;
- not attempt to match, with or without using identifiers, the information with any other list of persons or organisations;

– comply with any other direction and requirements specified by the data mining service provider.

Acknowledgements

The author thanks the useful discussions and suggestions from previous colleagues Christine M. O'Keefe, Ross Sparks, Jie Chen, Hongxing He and Damien McAullay in CSIRO; He also thank Dr. Jiuyong Li at the University of Southern Australia for discussions on privacy-preserving data publishing issues. Partial financial support for this project from the Australian Research Council is gratefully acknowledged. NICTA is funded by the Australian Governments Department of Communications, Information Technology, and the Arts and the Australian Research Council through Backing Australias Ability and the ICT Research Centre of Excellence programs.

References

1. The Office of Legislative Drafting: Privacy Act 1988 (Cth). Attorney-general's Department, Canberra, Australia (2004) http://www.privacy.gov.au/act/privacyact.
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. 2nd edn. Morgan Kaufmann Publishers (2006)
3. Jin, H.D., Shum, W., Leung, K.S., Wong, M.L.: Expanding self-organizing map for data visualization and cluster analysis. Information Sciences **163** (2004) 157–173
4. Jin, H., Wong, M.L., Leung, K.S.: Scalable model-based clustering for large databases based on data summarization. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(11) (2005) 1710–1719
5. Jin, H., Chen, J., He, H., Williams, G.J., Kelman, C., O'Keefe, C.M.: Mining unexpected temporal associations: Applications in detecting adverse drug reactions. Transactions on Information Technology in Biomedicine (2007) To appear.
6. Jin, H., Chen, J., Kelman, C., He, H., McAullay, D., O'Keefe, C.M.: Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases. In: Proceedings of PAKDD'06, Singapore (2006) 867–876
7. Crompton, M.: What is privacy? In: Privacy and Security in the Information Age Conference, Melbourne (2001) http://www.privacy.gov.au/news/speeches/sp51note1.html.
8. Oliveira, S.R.M., Zaane, O.R.: Protecting sensitive knowledge by data sanitization. In: ICDM'03: Proceedings of the Third IEEE International Conference on Data Mining, IEEE Computer Society (2003) 613–616
9. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the Third IEEE International Conference on Data Mining, IEEE Computer Society (2003) 1–9
10. U.S. Census Bureau: Public-use microdata samples (PUMS). http://www.census.gov/main/www/pums.html (2007) Accessed on 21 Jan 2007.
11. Australian Bureau of Statistics: Confidentialised unit record file (CURF). http://www.abs.gov.au (2007) Accessed on 20 Jan 2007.

12. Statistics New Zealand: Confidentialised unit record file (CURF). http://www.stats.govt.nz/curf-programme (2007) Accessed on 21 Jan 2007.
13. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems **10**(5) (2002) 557–570
14. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: ICDE'05. (2005) 217–228
15. Fung, B., Wang, K., Yu, P.: Top-down specialization for information and privacy preservation. In: Proceedings of 21st International Conference on Data Engineering (ICDE'2005). (2005) 205– 216
16. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of SIGMOD'00, ACM Press (2000) 439–450
17. Li, J., Wang, H., Jin, H., Yong, J.: Current developments of k-anonymous data releasing. In: Proceedings of ehPASS'06, Brisbane, Australia (2006) 109–121
18. Jin, W., Ge, R., Qian, W.: On robust and effective k-anonymity in large databases. In: PAKDD'06. (2006) 621–636
19. Wang, K., Fung, B.C., Yu, P.S.: Template-based privacy preservation in classification problems. In: ICDM'05: Proceedings of the Fifth IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2005) 466–473
20. Wang, K., Fung, B.C.M., Yu, P.S.: Handicapping attacker's confidence: An alternative to k-anonymization. Knowledge and Information Systems: An International Journal (2006)
21. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: ℓ-diversity: Privacy beyond κ-anonymity. In: Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006). (2006) 24
22. Wong, R., Li, J., Fu, A., Wang, K.: (alpha,k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In: KDD'06. (2006) 754–759
23. Vaidya, J., Clifton, C.: Privacy-preserving data mining: Why, how, and when. IEEE Security & Privacy **2**(6) (2004) 19–27
24. Yao, A.: Protocols for secure computations. In: Proceedings of the twenty-third annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society (1982) 160–164
25. Gilburd, B., Schuster, A., Wolff, R.: k-TTP: a new privacy model for large-scale distributed environments. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press (2004) 563–568
26. Wright, R., Yang, Z.: Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press (2004) 713–718
27. O'Keefe, C.M., Yung, M., Gu, L., Baxter, R.: Privacy-preserving data linkage protocols. In: WPES'04. (2004) 94–102
28. Jin, H., Leung, K.S., Wong, M.L., Xu, Z.B.: Scalable model-based cluster analysis using clustering features. Pattern Recognition **38**(5) (2005) 637–649
29. Jin, H., Chen, J., He, H., O'Keefe, C.M.: Privacy-preserving sequential pattern release. In: Proceedings of PAKDD'07, Nanjing, China (2007) 547–554
30. Woodruff, D., Staddon, J.: Private inference control. In: CCS'04: Proceedings of the 11th ACM conference on Computer and communications security, ACM Press (2004) 188–197
31. Li, J., Fu, A.W.C., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R., Kelman, C.: Mining risk patterns in medical data. In: KDD'05. (2005) 770–775

32. Adam, N.R., Worthmann, J.C.: Security-control methods for statistical databases: a comparative study. ACM Comput. Surv. **21**(4) (1989) 515–556
33. Australian Bureau of Statistics: Remote access data laboratory (RADL) – user guide. http://www.abs.gov.au (2006) Accessed on 20 Jan 2007.
34. Sparks, R., Carter, C., Donnelly, J., Duncan, J., O'Keefe, C., Ryan, L.: A framework for performing statistical analyses of unit record health data without violating either privacy or confidentiality of individuals. In: Proceedings of the 55th Session of the International Statistical Institute, Sydney (2005)
35. Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T., McAullay, D., Ryan, L.: Privacy-preserving analytics: remote access methods for exploratory data analysis and statistical modelling. Under review, CSIRO (2006)
36. Fule, P., Roddick, J.F.: Detecting privacy and ethical sensitivity in data mining results. In: Proceedings of ACS'04. (2004) 159–166
37. Oliveira, S.R.M., Zaïane, O.R., Saygin, Y.: Secure association rule sharing. In: PAKDD'04. (2004) 74–85
38. Atzori, M., Bonchi, F., Giannotti, F., Pedreschi, D.: k-anonymous patterns. In: PADD'05. (2005) 10–21
39. Atzori, M., Bonchi, F., Giannotti, F., Pedreschi, D.: Blocking anonymity threats raised by frequent itemset mining. In: ICDM'05. (2005) 561–564
40. Fienberg, S.E., Slavkovic, A.B.: Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery **11**(2) (2005) 155–180
41. Bayardo, R.J., Srikant, R.: Technological solutions for protecting privacy. IEEE Computer **36**(9) (2003) 115–118