Chapter 15

A STUDY OF INFERENCE PROBLEMS IN DISTRIBUTED DATABASES

LiWu Chang and Ira Moskowitz

Abstract The database inference problem is a challenging and important issue in data privacy. Database inference for distributed systems is an emerging research area. This paper describes a framework and approach to address the inference problem for distributed databases.

Keywords: Database inference, multiple databases, Bayesian networks, cooperative environment

1. Introduction

Often, database managers must decide which non-sensitive data to release. This is referred to as data *sanitization* or data *downgrading*. Issues surrounding downgrading are of particular importance to network architectures which utilize a multiple single level [14] approach for keeping sensitive data away from the generic user. In a distributed environment, data may be distributed among different data sites (e.g., [22]). Therefore, before data is downgraded, database managers must take into account other data that users may have access to.

Let us call the authorized users who either manage the entire database (e.g., database managers), or who are allowed access to the entirety of the data *High*, and the generic user, for whom access is restricted *Low*. Of course we are tacitly assuming that there are two types of data: low data which is available to all, and sensitive data to which only High users are allowed access. Thus, High data (High's data) may have both sensitive and non-sensitive components. High's concern is to keep the sensitive information away from Low. Therefore, High is allowed to downgrade the non-sensitive parts to Low.

For information sharing needs, High decides which data to release to Low. Obviously, High will not release sensitive data. However, *database inference* occurs when Low is able to infer the sensitive information from the data that is

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35697-6_26

E. Gudes et al. (eds.), *Research Directions in Data and Applications Security*

released. To prevent database inference, non-sensitive data which is related to sensitive data must be examined and perhaps modified, thus requiring further data sanitization. The problem of preventing database inference in a stand-alone database is quite challenging and has recently been under intensive study from diverse aspects (e.g., [2,3,5-9,11,13,15-19,29,30]). Database inference in distributed databases is an area in which very little work has been done. In [11], the authors showed that sensitive information can be revealed if users link information from several databases in a deterministic manner. However, the deterministic approach does not concern itself with the equally important problem of probabilistic relationships embedded in the data. We analyze the inference problem under a *probabilistic framework.*

Let us consider the following scenario which we will use throughout the paper. Given a database D consisting of categorical attributes, the most common technique for mitigating inference is that of modifying the non-sensitive data (i.e., values of attributes) in the database. However, even with appropriate data modification, sensitive data can still be compromised when data from other databases is incorporated. For instance, an AIDS diagnosis is often considered sensitive and is not disclosed. Given a second database containing information related to drug abuse, however, one may discover from the two databases that drug abusers' intravenous injections may cause these individuals to contract AIDS. Therefore, knowing a patient's history of drug injection could allow one to infer that the patient has a higher chance of contracting AIDS than does the general population, even if the diagnosis has not been revealed. Information about patients' drug abuse should therefore be treated as sensitive. As another example, the occurrence of non-Hodgkin lymphomas (NHL) is higher in AIDS patients than it is in the general population. Hence, the diagnosis of NHL should also be treated as sensitive, since there is a high correlation between AIDS and NHL. On the other hand, while, an AIDS patient may show signs of mental depression, depression is a common symptom of many diseases, such as low thyroid function. Thus, the symptom of mental depression may not be indicative of an AIDS diagnosis. Of concern, however, is that fact that, in a distributed environment, Low may obtain additional attribute data as discussed above which may allow Low to infer sensitive data—in this case, a diagnosis of AIDS.

2. Conceptual Model

We proposed a stand-alone model, the Rational Downgrader [20], for downgrading (sanitization) using inferential analysis. In a distributed environment, the conceptual model of the Rational Downgrader must be modified to include "knowledge" from both external data items and their rules.

Because external data items and their rules were not in the original model, the Rational Downgrader is composed of three components: the GUARD, the Decision Maker, and the Parsimonious Downgrader (or Filter). Initially, High inputs its candidate for the data that it would like to release to Low. The Decision Maker generates rules from the available data set, and uses the external data items and their associated rules, to form its output rules. The GUARD determines whether there is inference, and if it is "excessive" based on the Decision Maker's rules. If the inference is excessive, then the Parsimonious Downgrader will implement a protection plan to lessen the inference (i.e., decides to modify by deleting certain data from the database). The inference mechanism is based on a decision theoretical framework (e.g., [10,21,23,27,28]). It is the Bayesian network framework that will be used for our inference analysis in this paper. The output of the Rational Downgrader is the database to be released to Low. Our goal is to make modifications as parsimoniously as possible and thus avoid imposing unnecessary changes which lessen functionality.

A Bayesian network describes the probabilistic dependency relationships among the attributes of a database. A Bayesian network B may be generated from empirical data or can be constructed from *a priori* knowledge.

In a distributed system, it is misleading to evaluate downgrading in one database only. Publicly released data from one database may cause the inference of sensitive data in another database. If to-be-released data causes additional inference concerns due to another database, the Parsimonious Downgrader will incorporate the new requirements into its protection plan. As a result, only a mutually agreed upon data set will be released.

3. Database Inference

We consider the case for which sensitive data is associated with one particular attribute. In a medical database, AIDS diagnoses are the sensitive information. We use High database and Low database to indicate, respectively, the portion of a database viewed by a database manager (the High user) and a generic (Low) user. We are interested in studying probabilistic influences on the sensitive information from attributes that are related to the medical diagnosis only. (See, e.g., [1,3]) for disclosure protection of background attributes (e.g., age, address).) A sample of those relevant attributes is given in Table 1. Table 1 is the medical database for AIDS diagnoses which contains 20 data records (i.e., patients), which are uniquely identified by their key, and four attributes (excluding the key) (i.e., "hepatitis," "depression," "AIDS" and "transfusion") where each attribute has two values: a 'y' indicating the occurrence of the (diagnosis) result and an 'n' otherwise. In addition, Table 1 shows the High view (denoted here as D_H) in our discussion. The diagnosis of one disease (e.g., "AIDS") often causes the occurrence of another physical disorder (e.g., "mental depression"). Consequently, knowing the diagnosis of a physical disorder may lead to the inference of sensitive information (i.e., AIDS) about a patient. Thus, to protect sensitive information about one disease may require the protection of other probabilistically-related records. In this paper, we use a Bayesian network representation to describe the probabilistic relationship. A corresponding Bayesian network representation is given in Figure 1H[1] (see [10,21] for details on how to construct a Bayesian network), which shows that "AIDS" may affect the consequence of both "hepatitis" and "mental depression" and a cause of "AIDS" is a (blood) transfusion.[2]

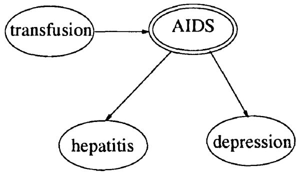

Figure 1H. B-net of the High.

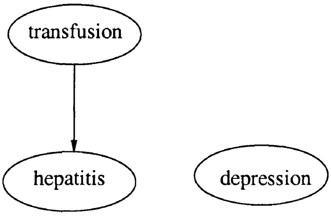

Figure 2L. B-net of the Low.

Table 2 shows the database after being initially downgraded (denoted here as D_L). Table 2 is what High would input into the Rational Downgrader, a patient is identified by its key. The threat with which we are concerned is that of Low inferring sensitive relations about the ith data item (or record) in the database. The dashes represent data that is considered sensitive and, thus, is not downgraded. Note that Table 2 is not in the form in which it will be released; it has not yet undergone the procedure that determines whether excessive inference may exist.

A target attribute $\mathcal{T}$ is an attribute that has dashes (meaning missing) in it (from Low's viewpoint). Thus, $\mathcal{T}$ represents sensitive information. We wish to lessen any inference that Low may attempt to draw about the target node. Since data is not completely revealed, the corresponding Bayesian network structure[3] for D_L differs from that of D_H and is shown in Figure 1L. The challenge for Low who is attempting to discern sensitive information is to restore the missing information in Table 2. Note that Table 2 still contains the "AIDS" attribute, even though the values are all missing. This is because we take "paranoid" view that Low knows what sensitive attribute High is con-

[1] Figure 1H is a Bayesian Network for High B_H. An attribute is denoted by a node. An arrow indicates the probabilistic dependency between two attributes. A double circle denotes that the attribute is sensitive.

[2] Our sample database is more representative of data taken at the beginning of the AIDS epidemic, rather than today.

[3] There are many ways to construct a Bayesian net. Figures 1H and Figure 1L are constructed as in [2] using a greedy model search.

cerned with, and because, in general, sensitive information may be distributed across many attributes and all the values may not be missing.

Initial downgrading may not be sufficient to protect the sensitive data. Again, we take the "paranoid" view that the Low obtains the prior knowledge (e..g., from previous studies) about the dependency relationship between AIDS and the three attributes "mental depression," "hepatitis," and "transfusion." (The dependency relationship is described in Figure 1H.) Certainly, it is not a surprise that Low could have the "prior" information it desires concerning an important medical condition such as AIDS. With information concerning the dependency, together with data in the Low database of Table 2, Low may be able to restore the hidden sensitive data. A sample restored Low database is shown in Table 3.

Compared with the original values in Table 1, the restored values of Table 3 differ in just four places. The probability of making a correct determination is $16/20 = 0.8$. This is unacceptable. The threat of potential restoration highlights the inadequacy of initial downgrading. We shall mitigate the inference by not downgrading certain non-sensitive information that can lead to probabilistic inferences about the sensitive information [2].

4. Distributed Databases

In the real world, there may be several databases [22] that have an impact on the sensitive information contained in the original downgraded database. The inference problem should take this into account.

These multiple databases may have exactly the same structure and/or they may have overlapping content. The possible interactions between two databases (in the form of relational tables, with schemes $R_1(a_1, ...a_k)$, and $R_2(b_1, ..., b_l)$ are the following. *R_2 augments R_1 with data records, R_2 augments R_1 with different attributes, or a combination.*

What we consider here is when two databases are in different contexts (or, applications), but have attributes which overlap (i.e., the third type of interaction). Also, we assume that data records of the two databases come from the same sample population, but the attribute values of some objects may be unknown. We shall use the structure of a Bayesian network and non-sensitive micro-data in our discussion.

Data transferred from the second database may or may not have direct impact on the sensitive information of the first database. High will integrate some, but not all, publicly released information from different databases that may cause the disclosure of sensitive data. Combinations of all data may render inference analysis an impossible task due to high volumes of data. We shall analyze the impact based on network dependency properties [21] and our practical sanitization policies with the following three databases.

Table 4 shows the diagnosis of non-Hodgkin lymphomas (NHL) disease — a NHL patient is highly likely to be a AIDS patient. Thus, data in Table 4 cannot be released[4] if the database manager of the NHL database also agrees with the sanitization/downgrading principle that AIDS data must be kept private. Based on a Bayesian network model of Table 5, low thyroid function causes mental depression, which in turn causes high blood pressure. For a mentally depressed patient, information concerning the patient's low thyroid function would have a negative impact on a possible AIDS diagnosis. The degree of impact depends on the correlation between AIDS and mental depression. It can be tested with available data. However, knowing the state of mental depression would block the impact of information concerning blood pressure. Table 6 is an database including information about illegal drug use, which shows the frequency with which an illegal drug user either takes intravenous injections or smokes. Data indicates an individual who injects illegal drugs is likely to have hepatitis. The relationship between AIDS and drug abuse is not shown in Table 6. However, for an intravenous drug abuser, intravenous injection is basically a form of blood transfusion.[5] Thus, the probability that an illegal intravenous drug abuser is also an AIDS patient is high. Table 7 shows the combination of the original High data with records of illegal intravenous drug abusers and records of the thyroid function, where the "*" denotes attribute values that are unknown because data records of these databases are not completely overlapped.[6] (Here, the assumption is that the database manager of the AIDS database is able to identify and select patients from the other two databases.) The dependency relationship between attributes of the combined database is given by the Bayesian network $\oplus B_H$ of Figure 3. Note that Figure 3 resulted from composing dependency relationships derived from these three databases, together with the knowledge about the relationship between intravenous injection and blood transfusion, and is not generated from combined data.[7] In Figure 3, the probabilistic dependency of the inverted fork (e.g., the child node "hepatitis" and parent nodes "injection" and "AIDS") is described by the or-ing operation (i.e., either "AIDS" or "injection" causes "hepatitis") as Pr(hepatitis=y—injection=y,AIDS=y) =

[4] The AIDS and the NHL databases can certainly exchange data through a special channel. But, again, the NHL database may not be released to Low and it is not considered part of the distributed data.

[5] Clearly, "drug injection" is not identified with blood transfusion in either database. Nonetheless, it is not a secret that intravenous injection of a drug involves the drawing of blood and drug abusers often share needles.

[6] The position of a data item in either thyroid or drug database (the 1st number) and its corresponding position in the combined table (the second number) are as follows:
thyroid: 1-1, 2-2 3-3 4-6 5-7 7-9 9-11 11-13 12-14 13-15 15-17 17-19 18-20
drug: 1-1 3-2 4-3 6-6 8-8 9-10 12-12 13-13 17-17 18-18 19-19 20-20

[7] If the combined network is generated from the combined data, the unknown mark "*" can be treated as a new value.

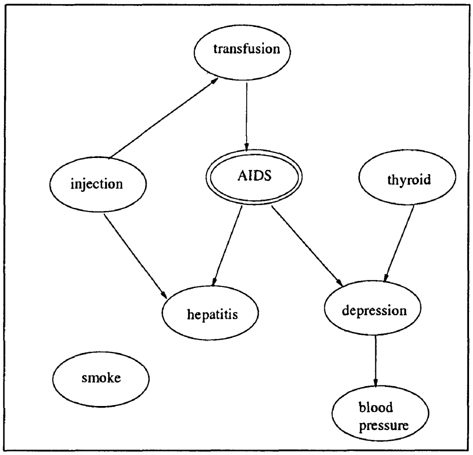

Figure 3. The combined Bayesian network $\oplus B_H$.

1-Pr(hepatitis=n—injection=y) · Pr(hepatitis=n—AIDS=y), where Pr(hepatitis=n—injection=y) and Pr(hepatitis=n—AIDS=y) are obtained from the released data of the drug abuse and AIDS databases, respectively. The probabilistic dependency between "depression," "Thyroid" and "AIDS" will also follow the or-ing relationship. The probabilistic information between "depression" and "blood pressure" will remain unchanged. The relationship between "injection" and "transfusion" is of the form *is_a*. It is known that the generation of a reliable complex network model, in general, demands large volumes of data. Here, we assume that the dependency relationship derived from each individual database is preserved in the combined database. For our current example, this assumption (referred to as *dependency inheritance under combination*) seems to be valid. It is useful in handling the combination of multiple large databases, yet its validity has not been formally proved.

5. Information Reduction

Recall that we showed the inadequacy of initial downgrading, because in our example 80% of the sensitive data was restored. The inference problem worsens if Low gleans information that is causally correlated to sensitive data from other publicly-released databases. This result suggests that High must adopt strategies for mitigating the inference problem based on data of the *combined* database. Therefore, since certain non-sensitive information can lead to probabilistic inferences about the sensitive information, we approach the problem of lessening inference by not downgrading all of the distributed non-sensitive information.

We modify non-sensitive data by "blocking," i.e., replacing an attribute value with a "?," indicating no knowledge about the attribute value.[8] Given a database D, we let D^m denote D after at least one of its non-sensitive entries has been blocked. Instead of sending D_L to Low, High blocks some of the non-sensitive information, and sends D_L^m to Low.

Shannon formalized the idea of secrecy [25] in cryptography. We use the similar definition of perfect noninference [20] for database inference:
DEF: If the *a priori* probability distribution of sensitive High data does not change in the presence of Low data, then we have *perfect noninference*.

The ideal downgrading policy for distributed databases is one that ensures perfect noninference. The condition of perfect noninference is unlikely to be achieved in practice because of performance issues. Our pragmatic policy of lessening inference states that modification of non-sensitive information should lessen the inference of sensitive information, while, at the same time, minimizing the loss of functionality.

For a given database, we measure the effect of modification, τ, based on the probabilistic term $Pr(D|B)$, which describes the sample probability, given the probabilistic dependency representation B. In essence, τ is a measure of the loss of functionality of a downgraded database. τ is a metric of the Low view and is measured by High.

$$\tau \overset{\text{def}}{=} \frac{|\log Pr(D_L|B_L) - \log Pr(D_L^m|B_L)|}{|\log Pr(D_L|B_L)|},$$

where $Pr(D_L^m|B_L)$ is computed by averaging over instantiations of the modified values. We evaluate the AIDS database only because we may not have control over other databases. The tolerance τ provides a margin within which the information protection strategies operate. Thus, we often associate an upper bound U to τ, so that $\tau \leq U$. τ measures the percentage of change in a sample probability. τ can be viewed as a sensitivity metric which estimates the rate of change in the output of a model with respect to changes in model inputs [12]. Our emphasis is on the magnitude of the change in probability. The log-scale measure (base 10) is used in this criterion to reflect the fact that the sample probability is a small number as a result of multiplication of the probabilities of each individual data record. Therefore a logarithmic approach somewhat normalizes the probabilities.

What criterion is used for High to select non-sensitive attribute values for modification (blocking)? Such a selection criterion is not unique. We present two selection criteria that are based on a Bayesian network framework.

[8] We do not use perturbation (i.e., replacing an attribute value with another different value), which introduces erroneous data, because of the negative performance side effects.

SC(1) *maximum difference:* Intuitively, attribute values which maximally change the probability of target values, $\mathcal{T} = t_i$, in terms of the probabilistic model $\oplus B_H$ (as shown in Figure 3), should be selected. (The set of attribute values is associated with non-sensitive information, or, in the current example, the non-target attribute.) Since B_H best interprets the data D_H from which it is derived and since the quality of modified database is expected to deteriorate, we want to select the set of attribute values that maximally decrease $Pr(D_H | \oplus B_H)$. This selection criterion is based on:

$$V_1 \overset{\text{def}}{=} (\log Pr(D_H^m | \oplus B_H) - \log Pr(D_H | \oplus B_H)), \tag{1}$$

For a given set of attribute values, N, (ranging from 0 to the number of all available attribute values except sensitive data,) to be modified, $Pr(D_H^m | \oplus B_H)$ is the average of all possible instantiation of this set of values. We use averaging because the value of N in our experiment is small. In case of large N, the value for modification (i.e., "?") may be viewed as a new symbol because instantiation could induce large variation when computing the sample probability and also be very time consuming. Note that we discard those values of $Pr(D_H^m | \oplus B_H)$ that are greater than $Pr(D_H | \oplus B_H)$. We do not think that situation will arise due to our (unproven) observation that modification results in decreasing the likelihood measure. We shall show the result of modification based on this criterion after we present the second selection criterion.

SC(2) *non-informative state:* We do not discuss this method due to pale limits. However, we will consider it in future work.

As discussed, τ lets us measure how the functionality of the database for the Low user, after blocking modifications, has changed with respect to Low. With the definitions of SC(1) in mind, our optimization goal is to *Maximize* V_1, while keeping $\tau \leq U$, if V_1 is chosen.

For modification, our approach is to evaluate attribute values of the AIDS database and analyze the potential impact from combined data. Let N denote the total number of non-targeted attribute values to be modified. Assume that $N = 4$, $U = 5\%$. Now we use the selection criterion SC(1) for our example. Consider the original Low database of Table 2. The modified Low database set is given in Table 8 by using SC(1). The choice that maximizes V_1 is that of blocking the "hepatitis" value for data item three, the "depression" value for data item four, and items three and eight from"transfusion." Modification is restricted to attribute values of the original Low database. ¿From Figure 3, it can be seen that inference occurs if Low obtains information about injection. This information can be used to restore the modified values of "transfusion," if the relationship between blood transfusion and injection is known *a priori.* This result will render sanitization a failure.

To remedy this, we consider two cases: that of downgrading in a *cooperative* environment, and in a *non-cooperative* environment. In a cooperative environment, an effective approach is to mutually examine the data sets to be released. In our example, the to-be-released data from the AIDS site is sent to the drug abuse site for examination. (Of course, one party needs to initiate the move.) Based on the received data, the drug abuse database manager makes modifications to its own data in order to ensure the safety of those values that have been modified from the AIDS database. The result is then sent back to the AIDS database. In our example, the modified combined database is given in Table 9, where values of "injection" are replaced by "?" in the third and the eighth data records, because "injection" causally affects "transfusion" as described in Figure 3. Note that since "hepatitis" is also causally correlated with "injection", modification of "injection" in the third record also minimizes restoration of a "?" with respect to "hepatitis" of the same record. Attributes that are causally related to "depression" are "thyroid" and "blood pressure." Since both values of "thyroid" and "blood pressure" of the fourth record are unknown, no further modification is needed. Table 9 is the downgraded Low database. (One may choose to replace symbols "-", "?" and "*" with a blank space and the outcome of the replacement is the Low database of Table 10.) Of course, in a non-cooperative environment, if the drug abuse database manager has already released his or her data, the manager of the AIDS database may take the defensive measure of increasing N. Such modification will, of course, occur at the expense of lowering the performance of the released database.

We do not impose a definite limit on the amount of data to be modified. The rule of thumb is that the process of modification stops when the modified database can no longer support the Bayesian network structure of the original database. (See also [4] for the estimation of lower bound.) The change usually undermines legitimate usage of a database.

6. Conclusions

The database inference problem with respect to distributed databases is essentially a new research area. In this paper, we analyzed the inference problem arising from distributed databases and we presented our approach which is based on the framework of Bayesian networks.

Acknowledgments

This research was supported by the Office of Naval Research.

References

[1] Bethlehem, J., Keller, W. and Pannekoek, J. (1990) Disclosure control of microdata, *Journal of the American Statistical Association*, Vol. 85, pp. 38-45.

[2] Chang, L. and Moskowitz, I.S. (1998) Bayesian methods applied to the database inference problem, *Database Security XII* (ed. S. Jajodia), Kluwer, pp. 237-251.

[3] Chang, L and Moskowitz, I.S. (2000) An integrated framework for database inference and privacy protection, *Data And Applications Security* (eds. Thuraisingham, van de Riet, Dittrich & Tari), Kluwer, pp. 161-172.

[4] Clifton, C. (1999) Protecting against data mining through samples, *Advances in Database And Information Systems Security* (eds. Atluri & Hale), Kluwer, pp. 193-207.

[5] Cox, L. (1987) A constructive procedure for unbiased controlled rounding, *Journal of the American Statistical Association*, Vol. 82, pp. 520-524.

[6] Denning, D. (1980) Secure statistical database with random sample queries, *ACM Transactions on Database Systems*, Vol. 5(3), pp. 291-315.

[7] Dobra, A. and Fienberg, S.E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, *NAS Proceedings*, Vol. 97(22), pp. 11885-11892.

[8] Duncan, G. and Roehrig P. (2002) Cyclic perturbation: Protecting confidentiality in tabular data (manuscript).

[9] Hale, J. and Shenoi, S. (1996) Analyzing FD inference in relational databases, *Data and Knowledge Engineering Journal*, Vol. 18, pp. 167-183.

[10] Heckerman, D. (1996) Bayesian networks for knowledge discovery, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, pp. 273-305.

[11] Hinke, T., Delugach, H. and Wolf, R. (1997) Protecting databases from inference attack, *Computers & Security*, Vol. 16(8), pp. 687-708.

[12] Isukapalli, S. (1999) Uncertainty analysis of transport-transformation models, Ph.D. Dissertation, CCI Department, State University of New Jersey at Rutgers.

[13] Johnsten, T. and Raghavan, V. (1999) Impact of decision-region based classification mining algorithms on database security, *Advances in Database And Information Systems Security* (eds. Atluri & Hale), Kluwer, pp. 177-191.

[14] Kang, M.H., Froscher, J.N. and Moskowitz, I.S. (1997) An architecture for multilevel secure interoperability, *Proceedings of the 13th ACSAC Conference*.

[15] Lin, T.Y. (1993) Rough patterns in data-rough sets and intrusion detection systems, *Foundations of Computer Science and Decision Support*, Vol. 18(3-4), pp. 225-241.

[16] Marks, D. (1996) Inference in MLS database systems, *IEEE Transactions of Knowledge and Data Engineering*, Vol 8(1), pp. 46-55.

[17] Matloff, N. (1988) Inference control via query restriction vs. data modification: A perspective, *Database Security: Status and Prospects*, North-Holland, pp. 159-166.

[18] Morgenstern, M. (1988) Controlling logical inference in multilevel database systems, *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pp. 245-255.

[19] Moskowitz, I.S. and Chang, L. (2000) An entropy-based framework for database inference, *LNCS 1768* (ed. A. Pfitzmann), Springer, pp. 405-418.

[20] Moskowitz, I.S. and Chang, L. (2000) A computational intelligence approach to the database inference problem, *Advances in Intelligent Systems: Theory and Applications* (ed. M. Mohammadian), IOS Press, pp. 377-387.

[21] Pearl, J. (2000) *Causality*, Cambridge.

[22] Prodromidis, A., Chan, P. and Stolfo, S. (2000) Meta-learning in distributed data mining systems: Issues and approaches, *Advances in Distributed and Parallel Knowledge Discovery* (eds. Kargupta & Chan), Chapter 3, AAAI/MIT.

[23] Quinlan, R. (1992) *C4.5*, Morgan Kaufmann.

[24] Saygin, Y., Verykios, V. and Clifton, C. (2001) Using unknowns to prevent discovery of association rules, *SIGMOD Record*, Vol. 30(4), pp. 45-54.

[25] Shannon, C. (1949) Communication theory of secrecy systems, *Bell Systems Technical Journal*, Vol. 28, pp. 656-715.

[26] Spiegelhalter, D. and Lauritzen, S. (1990) Sequential updating of conditional probabilities on directed graphical structures, *Networks*, Vol. 20, pp. 579-605.

[27] Spirtes, P., Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search.* Springer-Verlag.

[28] Thuraisingham, B. (1998) *Data Mining: Technologies, Tools and Trends*, CRC Press.

[29] Yip, R. and Levitt, K. The design and implementation of a data level database inference detection system, *Database Security XII* (ed. S. Jajodia), Kluwer, pp. 253-266.

[30] Zayatz, L. and Rowland, S. (1999) Disclosure limitation for American Factfinder, Census Bureau Report (manuscript).

A short-handed notation is used for each attribute of each table.
H: hepatitis; D: depression; A: AIDS; T: blood transfusion; N: NHL cancer;
P: blood pressure; Y: low thyroid; I: intravenous injection; S: smoke

Table 1. D_H — sample medical records (the 1st databse)
(The superscript i of the key refers to data records that come from the ith database.
In this table, the superscript i is 1.)

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y	y	y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y	y	y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A	n	n	y	y	n	n	n	y	y	n	y	n	n	y	n	n	n	y	n	n
T	n	n	y	n	n	n	n	y	n	n	y	n	n	n	n	n	n	y	n	y

Table 2. D_L — medical records of Low database

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y	y	y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y	y	y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
T	n	n	y	n	n	n	n	y	n	n	y	n	n	n	n	n	n	y	n	y

Table 3. a restored Low database

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y	y	y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y	y	y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A	n	n	y	**N**	n	n	n	y	**N**	n	y	n	n	**N**	n	n	n	y	n	**Y**
T	n	n	y	n	n	n	n	y	n	n	y	n	n	n	n	n	n	y	n	y

Table 4. NHL cancer database (the 2nd database)

key^2 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
N	n	n	y	y	y	n	y	y	n	y	y	y	n	n	y	y	n	y	n
A	n	n	y	y	n	n	y	y	y	y	y	y	n	y	y	y	n	y	n

Table 5. thyroid database (the 3rd database)

key^3 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P	n	y	y	y	n	n	n	n	n	y	y	n	y	y	n	n	y	y
D	n	y	y	n	n	y	y	y	y	y	n	y	y	y	y	n	y	n
Y	n	y	n	n	n	y	n	y	n	y	n	y	y	n	y	n	y	n

Table 6. drug abuse database (the 4th database)

key^4 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
I	n	y	n	y	y	n	y	y	n	y	y	n	n	y	y	y	n	y	n	y
S	y	n	y	y	y	y	y	y	y	n	y	y	y	y	y	n	y	y	y	n
H	n	y	y	y	y	n	y	y	n	y	y	y	n	y	y	y	n	y	n	y

Table 7. combined High database

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y	y	y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y	y	y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A	n	n	y	y	n	n	n	y	y	n	y	n	n	y	n	n	n	y	n	n
T	n	n	y	n	n	n	n	y	n	n	y	n	n	n	n	n	n	y	n	y
I^4	n	n	y	*	*	n	*	y	*	n	y	n	n	*	*	*	n	y	n	y
S^4	y	y	y	*	*	y	*	y	*	y	n	y	y	*	*	*	y	y	y	n
P^3	n	y	y	*	*	y	n	*	n	*	n	*	y	n	y	*	n	*	y	y
Y^3	n	y	n	*	*	n	n	*	n	*	n	*	n	y	y	*	y	*	y	n

Table 8. D_L^m———modified medical records

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y	?	y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y	?	y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
T	n	n	?	n	n	n	n	?	n	n	y	n	n	n	n	n	n	y	n	y

Table 9. modified combined Low database

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y	?	y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y	?	y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
T	n	n	?	n	n	n	n	?	n	n	y	n	n	n	n	n	n	y	n	y
I^4	n	n	?	*	*	n	*	?	*	n	y	n	n	*	*	*	n	y	n	y
S^4	y	y	y	*	*	y	*	y	*	y	n	y	y	*	*	*	y	y	y	n
P^3	n	y	y	*	*	y	n	*	n	*	n	*	y	n	y	*	n	*	y	y
Y^3	n	y	n	*	*	n	n	*	n	*	n	*	n	y	y	*	y	*	y	n

Table 10. (alternative) modified combined Low database

key^1 →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	n	y		y	n	n	y	y	y	n	n	y	n	y	n	y	n	y	n	y
D	n	y	y		y	n	n	n	y	n	y	n	n	y	y	n	y	y	y	n
A																				
T	n	n		n	n	n	n		n	n	y	n	n	n	n	n	n	y	n	y
I^4	n	n				n				n	y	n	n				n	y	n	y
S^4	y	y	y			y		y		y	n	y	y				y	y	y	n
P^3	n	y	y			y	n		n		n		y	n	y		n		y	y
Y^3	n	y	n			n	n		n		n		n	y	y		y		y	n