PRIVACY ISSUES IN KNOWLEDGE DISCOVERY AND DATA MINING

Ljiljana Brankovic[1] and Vladimir Estivill-Castro[2]

Abstract

Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. However, there is an increasing public concern about the individuals' privacy. Privacy is commonly seen as the right of individuals to control information about themselves. The appearance of technology for Knowledge Discovery and Data Mining (KDDM) has revitalized concern about the following general privacy issues:

- secondary use of the personal information,
- handling misinformation, and
- granulated access to personal information.

They demonstrate that existing privacy laws and policies are well behind the developments in technology, and no longer offer adequate protection.

We also discuss new privacy threats posed KDDM, which includes massive data collection, data warehouses, statistical analysis and deductive learning techniques. KDDM uses vast amounts of data to generate hypotheses and discover general patterns. KDDM poses the following new challenges to privacy.

- stereotypes,
- guarding personal data from KDDM researchers,
- individuals from training sets, and
- combination of patterns.

We discuss the possible solutions and their impact on the quality of discovered patterns.

1 Introduction

Not surprisingly, data is treated today as one of the most important corporate assets of companies, governments and research institutions supporting fact-based decision making. It is possible to have fast access, to correlate information stored in independent and distant databases, to analyze and visualize data on-line and use data mining tools for automatic and semi-automatic exploration and pattern discovery [3, 4, 11]. *Knowledge Discovery and Data Mining* (KDDM) is an umbrella term

[1] Lecturer, Department of Computer Science and Software Engineering, The University of Newcastle, NSW 2308, Australia

[2] Senior Lecturer, Department of Computer Science and Software Engineering, The University of Newcastle, NSW 2308, Australia.

describing several activities and techniques for extracting information from data and suggesting patterns in very large databases. Marketing applications have adopted and expanded KDDM techniques [3, 23].

Now, KDDM is moving to other domains where privacy issues are very delicate. Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. For the analysis of crime data, KDDM techniques have been applied by the FBI in the US as a part of the investigation of the Oklahoma City bombing, the Unabomber case, and many lower-profile crimes [3]. Another example is the application of KDDM to analyzing medical data [14].

Despite its benefits to social goals, KDDM applications inspire reservations. Individuals easily imagine the potential misuses from unauthorized tapping into financial transactions or medical records. There is an increasing public concern about the individuals' privacy. Surveys in the US [8, Sur96] reveal growing concern about the use of personal information. The newest Equifax/Harris Consumer Privacy Survey shows that over 70% of respondents are against unrestricted usage of their medical data for research purposes. As many as 78% believe that computer technology represents a threat to personal privacy and that the use of computers must be restricted sharply in the future, if privacy is to be preserved. Up to 76% believe they have lost control over their personal information and that their privacy is threatened. Time/CNN [15] and other recent studies [8] reveal that at least 93\% of respondents believe companies selling personal data should be required to gain permission from individuals. In another study [8], 96% of respondents believe that private information should never be used for another purposes without permission, and over 20% had personally experienced a privacy invasion. By contrast, in 1970 the same Equifax/Harris Survey found only 33\% considered computer technology a threat to their privacy [15].

Privacy is commonly seen as the right of individuals to control information about themselves. In 1990, Lotus Development Corporation announced the release of a CD-ROM with the data on 100 million households in the US. The data was so detailed that it generated strong public opposition and Lotus abandoned the project [8]. However, this mostly affected small business, as large companies already had access and continued to use Lotus data sets [8]. At least 400 million credit records, 700 million annual drug records, 100 million medical records and 600 million personal records are sold yearly in the US by 200 superbureaus [19]. Among the records sold are bank balances, rental histories, retail purchases, criminal records, unlisted phone numbers and recent phone calls [19]. Combined, the information helps to develop *data images* of individuals that are sold to direct marketers, private individuals, investigators, and government organizations [19,20]. These data images are now the subject of analysis by automatic and semiautomatic Knowledge Discovery and Mining tools[20].

We address revitalized general privacy issues and new threats to privacy by the application of KDDM. We distinguish them from threats to privacy or security resulting from the expansion of computer networks and on-line distributed information systems. The rest of this paper is organized as follows. Section 2 reviews revitalized privacy threats, illustrating how analysis of large volumes of personal data may result in loss of privacy. Section 3 highlights new privacy threats directly associated with KDDM. Section 4 Section 5

2 Revitalized Privacy Threats.

2.1 Secondary Use of the Personal Information.

As we pointed out, recent surveys on privacy show a great concern about the use of personal data for purposes other than the one for which data has been collected. An extreme case occurred in 1989. Despite collecting over $16 million USD by selling the driver-license data from 19.5 million Californian residents, the Department of Motor Vehicles in California revised its data selling policy after Robert Brado used their services to obtain the address of actress Rebecca Schaeffer and later killed her in her apartment. While it is very unlikely that KDDM tools will reveal directly precise confidential data, the exploratory KDDM tools may correlate or disclose confidential, sensitive facts about individuals resulting in a significant reduction of possibilities[9]. In fact, this is how they were applied in the investigation of the Unabomber case and other criminal investigations. They facilitated filtering large volumes of reports from informants so resources could be concentrated on much fewer promising leads and suspects. Thus, we would not expect that detailed personal addresses would be disclosed by a KDDM analysis; however, with enough data about the patterns of behavior of young actresses in Los Angeles, the potential addresses may be reduced to a few possibilities making it feasible to visit them all.

A simple application of Link Analysis [3] can correlate phone records and banking records to determine, with a certain degree of accuracy, if bank customers have a fax machine at home and how this impacts the likelihood of accepting offers on equity loans. Most individuals consider the use of information beyond its initial collection for secondary analysis a direct invasion of privacy, and perhaps even more if this reveals aspects like what does a person have inside its home. Individuals understand that phone companies need to monitor length of phone calls for billing purposes, and that their bank must keep track of transactions in their accounts, but when this data is used for secondary purposes for which the individual has nor provided authorization, a serious issue in privacy appears.

2.2 Handling Misinformation.

Misinformation can cause serious and long-term damage, so individuals should be able challenge the correctness of data about themselves. For example, District Cablevision in Washington fired James Russell Wiggings, on the basis of information obtained from Equifax, Atlanta, about Wiggings' conviction for cocaine possession; the information was actually about James Ray Wiggings, and the case ended up in court.

This illustrates a serious issue in defining property of the data containing personal records. While individuals and legislators supporting the right of privacy favor the view that a person's data is the person property, data collectors favor the view that the data collector owns the data. This ethical issues has been illustrated by some cases of celebrities and other public figures who have been able to obtain rights on reproduction of their photographed image. However, for the average citizen, the horizon is not so promising.

The existing market of personal data postulates that the gathering institution owns the data. However, the attitude of data collectors and marketers towards privacy is more moderate than 20 years ago when marketers believed that there was "too much privacy already" [19].
The reason for this change, apart from the fact that privacy is under much bigger threat now, is probably the fear of losing the trust of customers and massive public opposition [15].
Many data owners acknowledge that there is a "Big Brother" aspect in the exploitation of personal data sets, and data collectors now take some measures to preserve the customers trust [12]. Others imply that the "sinister purpose" of data mining is the "product of junk science and journalistic excess" [23], but nevertheless believe that "marketers should take a proactive stance and work to diffuse the issue before it becomes a major problem" [23].

2.3 Granulated Access to Personal Information.

The access to personal data should be on a need-to-know basis, and limited to relevant information only. For example, employers are obliged to perform a background check when hiring a worker but it is widely accepted that information about diet and exercise habits should not affect hiring decisions.

There also seem to be two directions in this issue[7]. Some advocate the removal of fields and even prohibiting their collection. Others support the release for very detailed data so research and analysis can advance. The new privacy laws in Germany illustrate the first approach. These regulations have dramatically reduced the number of variables in the census and the micro census [18]. The second approach is illustrated by personal data placed on large on-line networked databases, like the Physician Computer Network in the US, with the intention to build and expand knowledge. Another example of this approach are more precise geo-referenced data sets in Geographical Information Systems and their databases that include cadastral data, aerial photography of individual properties and detailed features of private properties.

Scholars from diverse backgrounds in history, sociology, business and political science have concluded that the existing privacy laws are far behind the developments in information technology and do not protect privacy well even for the issues mentioned above [19].
We now present new privacy threats that confirm that existing privacy laws and policies are well behind the developments in technology, and no longer offer adequate protection.

3 New Privacy Threats

We discuss new privacy threats posed by Knowledge Discovery and Data Mining (KDDM), which includes massive data collection, data warehouses, statistical analysis and deductive learning techniques. KDDM uses vast amounts of data to generate hypotheses and discover general patterns. Besides revitalizing the issues in Section 2, KDDM poses the following new challenges to privacy.

3. 1 Stereotypes

General patterns may be used for guessing confidential properties. Also, they may lead to stereotypes and prejudices. If the patterns are based on properties such as race, gender or

nationality, this issue can be very sensitive and controversial. Examples are debates over studies about intelligence across different races [24] and crime rates in urban groups [Taylor91].
The issue raises debate because KDDM tools may allow the application of different commercial standards based on race or ethnic group. Banks may use KDDM tools to find a different pattern of behavior between two racial or ethnic groups and then deny credit or apply a different policy based on this attribute[13]. An example of different commercialization based on ethnic group occurs in marketing of petrol in Canada because it is well know that French-Canadians are more likely to buy premium petrol than Anglo-Canadians. While these examples seem harmless, wide application of KDDM patterns can spread stereotypes. Moreover, analysis of data regarding commercial transaction or behavior is necessary for business efficiency and competitiveness. Data analysis has made it possible for car insurance companies to adjust their premiums based on the age of the driver. It has allowed health insurance companies also adjust their premiums based on some patterns of behavior of the policyholder (for example, the practicing of sports like scuba diving or hand-gliding).

This may have serious implications into the type of data mining technology applied to some task. For example, an inductive learner may be used to create a classifier out of previous credit card applications. If the learner is of the type that can explain why it is denying or granting some credit level, the corporation using it may detect that it is using controversial attributes (i.e., gender, nationality) to perform the credit ranking. However, if the tools represents knowledge in implicit forms that are hard to interpret (a la Artificial Neural Networks), it may be possible that credit assignment is been mainly based on controversial attributes without the awareness of the credit-card company and the applicant. Many legal issues may be raised if the applicant is to discover this.

Another problem is that KDDM tools use parametric and non-parametric methods from the fields of statistics and machine learning. Their limitations are difficult to explain to a lay person. However, it may be easy to claim that the pattern was derived by a sophisticated computer program, and thus, accepted as truth about individuals. In particular, the fact that they are obtaining rules of thumb and generalization that work in most cases (and are about groups) but are false of each individual is one of the hardest concepts to grasp. This is true even statistics. One toss of a fair coin will end up in heads or in tail and not in between, but such in between is the "expected value" of such a single event. Similarly, statistics allows us to predict accurately values about populations, like how many smokers will develop cancer this year, but it will not guarantee than a Mr., X will develop cancer, and despite him being a smoker.

3.2 Guarding personal data from KDDM researchers

How could planning decisions be taken, if census data was not collected? How could epidemics be understood if medical records were not analyzed? Individuals benefit from data collection efforts in the process of building knowledge that guides society. The protection of privacy cannot simply be achieved by restricting data collection or restricting the use of computer and networking technology. Researchers feel that privacy regulations will enforce so many restrictions on data, that it would make the data useless.

On the other hand, researchers that apply KDDM tools to datasets containing personal information should not be given access to individual data. But, such a restricted access can make KDDM tasks very difficult, or even impossible. Thus an appropriate balance between a need for privacy and a need for knowledge discovery should be found.

3.3 Individuals from training sets

The classification task in KDDM takes as input a set of cases and their classes (training set); the output is a classifier, that is, an operator that assigns classes to new, unclassified cases. For example, the cases may correspond to patients and classes to diagnoses.

The first problem is how to provide the analyst with KDDM tools a training set. If such a set is provided from real data, then each record of a training set is a disclosure of the information of the individual corresponding to the record. The second problem is how to protect privacy if somebody has a classifier and a record that knows belongs to the training set that built the classifier, but does not have the class. The KDDM classifiers are typically very accurate when applied to cases from the training set. Thus a classifier and knowledge that case A is in the training set reveals the class of case A. In this paper we argue that a classifier should be modified in such a way so as to have similar accuracy when applied to the cases from the training set, as when applied to the new cases.

3.4 Combination of patterns

Combining two or more general patterns may lead to a disclosure of individual information, either with certainty, or with a high probability[16]. Also, knowledge of totals and other similar facts about the training data may be correlated to facilitate compromising individual values either with certainty or with a high probability. For example, consider a data set where:

- there are 10 people, 2 females and 8 males,
- there are 8 cases of disease A,
- none of the females has disease A.

If it is known that Mr. *X*'s information is part of the data, it is possible to infer that Mr. *X* has disease A.

This is a difficult problem. A simpler problem is how to protect individual values while preserving values of parameters that are calculated by statistical inference. This parameters usually describe a probabilistic model and their accurate inference hardly reveals anything about an individuals record. Thus, perturbation to the data that preserve the estimators of statistical inference technique could potentially be sufficient for the simpler problem.

How to hide what can be inferred by all possible methods? This implies in a sense that we discover first everything that is to be discovered about a data set in order to be sure that we conceal it [27]. Or in other words, to understand how to protect values in individual records, we first must understand very well the capabilities of KDDM tools and technology. The we know what we are up against.

4 Discussion

The OECD Principles on Data Collection addresses some of the concerns raised by the issues in Section 2. However, only 24 countries have adopted them, thus far with varying degrees. Twelve nations have adopted all OECD's principles in statutory law. Australia, Canada, New Zealand and the US do not offer protection to personal data handled by private corporations [22]. In Australia, the Privacy Act 1988 predates the growing on-line purchasing and other massive networked data-collection mechanisms. Australia's Privacy Commissioner, Moira Scollay, has taken steps to simplify privacy regulations and provide a single national framework for data-matching systems such as FLY-BUYS cards. However, at present, she has only proposed for discussion a draft of principles for fair handling of personal information.

We hope that we have clearly illustrated that tools for Knowledge Discovery poses a threat to privacy, in the sense that discovered general patterns classify individuals into categories, revealing in that way confidential personal information with certain probability.

Other researchers in KDDM seem to have contradictory views [24]. Some believe that KDDM is not a threat to privacy, since the derived knowledge is only about and from groups [28]. Clearly, this is only true if the identified group is large, but otherwise, information about an individual is partially disclosed. Also, performing inferences about groups does not prevent the creation of stereotypes.

Others clearly support our views saying that KDDM deals mainly with huge amounts of microdata [18]. Some fear there would be different academic standards: "Statutory limitations that vary from country to country … suggests that the practice …varies from country to country" [22]. Europe has adopted the OECD directives and investigators across all fields of scholarly research now require "the subject's written consent or data may not be processed" [5]. Others think that subject approval may not be sufficient for Data Miners to refer or disclose incidentally discovered patterns or relationships [17]. That is, how can one ask approval from an individual about something yet unknown?

4 Possible Solutions

We discuss the possible solutions and their impact on the quality of discovered patterns. The issue of how data analysis may result in interpretations that may create or reinforce stereotypes is more an issue of the application of the technology than a problem of the technology. Is similar to attributing the impact on large audiences with political propaganda and biased media as a disadvantage of TV. This is not a disadvantage of the technology, but more a very complex issue of the social structures that allow the production and broadcasting of materials, later distributed through the powerful technology. Thus, the issue of created or reinforced stereotypes with KDDM tools falls in the sociological, anthropological and legal domain. The interpretation of patterns from KDDM is as difficult as the interpretation of statistics. For example, the times/distances/etc achieved by female athletes Vs male athletes in Olympic competitions constitute data that statistically distinguishes two groups. When does finding two groups justify creating different policies is not an issue of the technology that discovered the two clusters.

In what follows, we concentrate in technological solutions that could alleviate some of the risk to privacy associated with the other issues listed in Section 2 and Section 3. Those treats to privacy listed in Section 2 are not totally new and there is some understanding of some of them from previous attempts to assure privacy in the face of the need to perform data analysis. In particular, the field of Statistical Databases has done much progress in this direction and we believe it can provide much insight into new issues like those listed in Section 3.

The field of statistical databases [2] has developed methods to guard against the disclosure of individual data while satisfying requests for aggregate statistical information. In particular, the experience form this field indicates that removing identifiers such as names, addresses, telephone numbers and social security numbers is a minimum requirement for privacy but it is not enough [18]. Reidentification based on remaining fields may still be possible [18]. In fact, removing identifiers is considered the weakest approach, and should never be used on its own [2, 18]. For a simple example, note that early geographical analysis of medical records, replaced personal addresses by latitude and longitude coordinates. Today, Geographical Information Systems are equipped with precise city maps where individual homes can be identified, forcing researchers to use other methods to conceal individual homes while still observing the geographical spread of a disease [RushtonA97].

In fact, the information models used in KDDM and statistical databases are quite similar [10, 26]. The main objectives of a security control mechanism in a statistical database are to provide statistical users with a sufficient amount of high quality statistics (statistical quality is measured by the consistency, bias and precision), and at the same time, to prevent disclosure of confidential individual information. Similarly to the context of statistical databases, the objective of a security control mechanism for KDDM is to prevent the disclosure of confidential individual values. However, unlike in statistical databases, another objective is not to minimize the distortion of all statistics, but rather to preserve general patterns. In other words, the objective is to preserve those statistics based on attributes, which have a high impact on the patterns and rules to be discovered.

Any statistic involving the confidential attribute reveals some information about individuals values. For example, if the answer to the statistical query "What is the sum of gross income of all small business of kind *A* in town *X*?" is $50 million, one learns that all small business in *X* have gross income less than $50 million; this information may or may not be significant. It is the matter of security policy to define what significant information is.

If disclosure of a confidential individual value occurs, the database is said to be *compromised* [2]. *Positive compromise* occurs if a user finds the exact value of a confidential attribute, and *partial compromise* occurs if a user is able to obtain substantial information about a confidential value, without disclosing it exactly [2].

Various techniques have been proposed for security protection in statistical databases, but none of them is both effective and efficient. All methods trade-off privacy of individual values for statistics and/or pattern distortion [2].

Techniques in statistical databases can be classified into two main categories,

- *query restriction* and
- *noise addition.*

Query restriction techniques provide exact answers to some queries, and reject others that may lead to a database compromise. Hence, the statistical quality of released information is high, but these techniques are typically overly restrictive [1]. They are inadequate against skilled users who have previous knowledge about information stored in the database, and/or require high initial implementation effort.

When applied to KDDM, query restriction techniques may deny some particularly important information and obscure general patterns. A recent proposal has been based on this [6]. The idea here is to supply a subset of the data so restricted that is not useful for the data miner. This has been criticized since,

- first, why would a miner will acquire or investigate data guaranteed not to have anything useful,
- second, the only way to guarantee that the data set is really contains no patterns is to find them all (which requires infinite computational time [27] or to provide a very small set, and
- third, for this scheme to work, it is assumed that each miner will not cooperate with other miners (and in particular, that nobody gains access to more data).

Since the objective of query restriction is not to maximize the number of answerable queries, but rather to answer all, or most of, the important queries, partitioning is a potential approach. This requires the development of methods for grouping record on the basis of their values for important attributes. These attributes may be identified using rough set theory [SanJoseprivate coom].

Noise addition techniques prevent compromise by introducing an error either to data or to results of queries. These techniques are robust (resistant to users' supplementary knowledge) and provide answers to all queries, but at the expense of allowing partial disclosure and/or providing information with less statistical quality. Oleary has suggested that noise addition by itself could protect from all those KDDM techniques that are very sensitive to noise [21].

In particular, a probability distribution data perturbation technique referred to as *data swapping* [Denn82], seems suitable for privacy protection in KDDM [10].

Data swapping refers to interchanging the values in the records of the database in such a way that low-order statistics are preserved [25]. For example, *k*-order statistics are those that employ exactly *k* attributes. A database *D* is said to be *d*-transformable if there exists a database *D'* that has no records in common with *D*, but has the same *k*-order Counts as *D*, for *k Î{0,…,d}.* As an example, consider databases *D* and *D'* in Figure 1, where *D'* is 2-transformation of *D*. Note that all 2-order Counts are preserved and, since *D* and *D'* have no records in common, the released data is protected from positive compromise. However, finding a general data swap is a computationally intractable problem [Denn82].

D			D'		
Gender	Age	Drug	Gender	Age	Drug
Female	20	Yes	Female	20	No
Female	30	No	Female	30	Yes
Male	20	No	Male	20	Yes
Male	30	Yes	Male	30	No

Figure 1

Our recent investigations have shown that it is possible to transform the data by data swapping so the rules produced by decision trees (a ubiquitous KDDM technique) on the new set correspond to finding a *d*-transformation of a given database, but relaxing the condition that *D* and *D'* have no records in common. The exchange of values of the confidential attribute consists of randomly shuffling within heterogeneous leaves of the decision tree. Thus, all the statistics, which do not involve this attribute, are always preserved. Similarly, the statistics that involve the confidential attribute and whose query sets are defined by nodes of a decision tree will also be preserved. All the perturbed statistics are those whose query sets could be obtained by the repeating process of splitting heterogeneous leaves, until homogeneous leaves are reached. Since the heterogeneous leaves have a vast majority of records belonging to a single class and no straightforward way for further splitting, we can argue that the most seriously disturbed statistics will be those that involve a small number of records. Furthermore, we can balance the statistical precision against the security level by choosing to perform the swapping in the internal nodes, rather than in the leaves of the decision tree: the closer to the root, the higher the security but lower the precision.

5 Conclusion

Knowledge Discovery and Data Mining revitalizes some issues and posses new threats to privacy. Some of these can be directly attributed to the fact that this powerful techniques may enable the correlation of separate data sets in other to significantly reduce the possible values of private information. Other can be more attributed to the interpretation, application and actions taken from the inferences obtain with the tools. While this raises concerns, there is a body of knowledge in the field of statistical databases that could potentially be extended and adapted to develop new techniques to balance the rights to privacy and the needs for knowledge and analysis of large volumes of information. Some of this new privacy protection methods are emerging as the application of KDD tools moves to more controversial datasets.

6 References

[1] N.R. Adam and D.H. Jones. "Security of statistical databases with an output perturbation technique". *Journal of Management and Information Systems*, 6, (1): 101--110, 1989.

[2] N.R. Adam and J.C. Wortmann. "Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4): 515--556, 1989.

[3] M.J.A. Berry and G.Linoff. "Data Mining Techniques --- for Marketing, Sales and Customer Support. John Wiley & Sons, NY, USA, 1997.

[4] A. Berson and S.J. Smith. "Data Warehousing, Data Mining, & OLAP". Series on Data Warehousing and Data Management. McGraw-Hill, NY, USA, 1998.

[5] S. Bonorris. "Cautionary notes for the automated processing of data". *IEEE Expert*, 10(2): 53--54, April 1995.

[6] C. Clifton and D. Marks. "Security and privacy implications of data mining". In SIGMOD *workshop on Data Mining and Knowledge Discovery*, Montreal, Canada, June 1996. ACM.

[7] Private lives; public records --- in today's networked world is personal information to easy to access?". *Computerworld*, 31(6): 81-90, September 1997.

[8] M. J. Culnan. "``How did they get my name?' ': An exploratory investigation of consumer attitudes towards secondary information use*". MIS Quarterly*, 17: 341--361, 1993.

[9] A. Einsenberg. "Data mining and privacy invasion on the net". *Scientific American*, 274(3): 120, 1996.

[10] V. Estivill-Castro and L.~Brankovic. "Data swapping: Balancing privacy against precision in mining for general patterns". Submitted,

[11] U. Fayyad and R. Uthurusamy. "Data mining and knowledge discovery in databases". *Communications of the ACM*, 39(11): 24--26, Nov. 1996. Special issue on Data Mining.

[12] Financial Marketing. "Getting to know you". Institutional Investor, 25(7): 71--73, 1991.

[13] S. Fritz and D. Hosemann. "Behavior scoring for Deutsche Bank's German corporates". In G. Nakhaeizadeh and E. Steurer, editors, *Applications of Machine Learning and Data Mining in Finance*, pages 179--189, Germany, April 1998. European Conference on Machine Learning ECML-98 Workshop Notes, Technische Universität Chemnitz, Facultät für Informatik. CSR-98-06.

[14] M.S. Gordon and M.J. Williams. "Spatial data mining for health research, planning and education". In C.P. Waegemann (editor), Proceedings of TEPR-97: Towards an Electronic Patient, pages 212--218, Newton, MA. USA, 1997. Medical Records Institute.

[15] S. Greengard. "Privacy: Entitlement or illusion?". *Personnel Journal*, 75:74--88, 1996.

[16] "Data warehousing: Digging too deep?" *Informationweek*, (583):14,June 1996.

[17] Y.-T. Khaw and H.-Y. Lee. Privacy & knowledge discovery". *IEEE Expert*, 10(2):58, April 1995.

[18] W. Klösgen. "KDD: Public and private concerns". *IEEE Expert*, 10(2):55--57, 1995.

[19] K. C. Laudon. "Markets and privacy". *Communications of the ACM*, 39(9):92--104, 1996.

[20] K.S. Nash. "Electronic profiling --- critics fear systems may trample civil rights". Computerworld, 32(6):1,28, February 1998.

[21] D.E. O'Leary. "Knowledge discovery as a threat to database security". In G. Piatetsky-Shapiro and W.J. Frawley (editors), *Knowledge Discovery in Databases*, 507--516, Menlo Park, CA, 1991. AAAI Press.

[22] D.E. O'Leary. "Some privacy issues in knowledge discovery: the OECD personal privacy guidelines". *IEEE Expert*, 10(2):48--52, April 1995.

[23] P. R. Peacock. "Data mining in marketing: Part 2". *Marketing Management*, 7(1):15--25, 1998.

[24] G. Piatetsky-Shapiro. "Knowledge discovery in personal data Vs privacy: a mini-symposium". *IEEE Expert*, 10(2):46--47, April 1995.

[25] S. Reiss. "Practical data swapping: The first step". *ACM Transaction on Database Systems*, 9(1):20--37, 1984.

[26] A. Shoshani. "OLAP and statistical databases: Similarities and differences*". Proceedings of 16th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems}*, 185--196, Tucson, AZ, US, 1997.

[27] C.S. Wallace and D.L. Dowe. "Minimum message length and Kolmogorov complexity". *Computer Journal*, 1999. in press.

[28] W. Ziarko. Response to O'Leary's article". *IEEE Expert*, 10(2):59, April 1995.