Preface: Proceedings of the ICDM 2002 Workshop on Privacy, Security, and Data Mining

Vladimir Estivill-Castro **Chris Clifton**

School of Computing and Information Technology
Nathan and Logan Campuses
Griffith University
Brisbane 4111, Queensland, Australia
Email: V.Estivill-Castro@cit.gu.edu.au

Department of Computer Sciences
Purdue University
West Lafayette, IN, 47907-2606 USA
Email: clifton@cs.purdue.edu

These volume contains papers selected for presentation at the Workshop on Privacy, Security, and Data Mining. The workshop was held in conjunction with the 2002 IEEE International Conference on Data Mining, in Maebashi Terrsa, Maebashi City, Japan, on December 9, 2002

In the light of developments in technology to analyze personal data, public concerns regarding privacy are rising (Nash 1998). Clarke (Clarke 1988, Clarke 1999) succinctly defined two fundamental notions:

- *Information Privacy* as "the interest individuals have in controlling, or at least significantly influencing, the handling of data concerning themselves" and
- *Dataveillance (Data Surveillance)* as "the systematic use of personal data systems in the investigation or monitoring of the actions or communications of one or more persons".

The recent emergence of *data mining* technology to analyze vast amounts of data opens new threats to information privacy and facilitates data surveillance (Brankovic & Estivill-Castro 1999, Clifton & Marks 1996, Estivill-Castro, Brankovik & Dowe 1999). It is now possible to have fast access, to correlate information stored in independent and distant databases, to analyze and visualize data on-line and to use data mining tools for automatic and semi-automatic exploration and pattern discovery (Berry & Linoff 1997, Berson & Smith 1998, Fayyad & Uthurusamy 1996).

The motto for the workshop was "How do we mine data when we aren't allowed to see it?". One of the key requirements of a data mining project is *access* to the relevant data. Privacy and security concerns can constrain such access, threatening to derail data mining projects.

However, huge volumes of detailed personal data are regularly collected and analyzed by marketing applications using data mining technology (Bigus 1996, Berry & Linoff 1997, Peacock 1998). Commercial applications in which individuals may be unaware of "behind the scenes" use of Data Mining are now documented (John 1999). Existing laws are behind developments in information technology and do not protect privacy well (Brankovic & Estivill-Castro 1999, Laudon 1996, O'Leary 1995). Privacy advocates face limitations to push legislation restricting the secondary use of personal data, since analyzing data brings collective benefit in many contexts (Gordon & Williams 1997). Even from the first conferences on data mining, central figures in the community, such as O'Leary (O'Leary 1991) Fayyad, Piatesky-Shapiro and Smyth (Fayyad, Piatesky-Shapiro & Smyth 1996) as well as Klösgen (Klösgen 1995), wrote on some of the evident initial privacy concerns. However, the fascination with the promise of interpretation of large volumes of raw data pushed aside privacy issues.

Many data miners believed data mining did not represent a threat to privacy (O'Leary 1995, Bonorris 1995, Khaw & Lee 1995, Piatetsky-Shapiro 1995). The organizers of this workshop have had a continuous interest on the topic and it could be said that they maintained the interest (Brankovic & Estivill-Castro 1999, Clifton & Marks 1996, Clifton 2000, Estivill-Castro et al. 1999). Estivill-Castro and Brankovic's work indicated renewed and new treats to privacy. Clifton's work also highlighted challenges and proposed controversial small samples methods (Clifton 1999, Clifton 2000). Estivill-Castro and Brankovic (Brankovic & Estivill-Castro 1999, Estivill-Castro et al. 1999) indicated the potential of data perturbation methods. The approach was enriched and brought to the main core of KDDM under the title of "Privacy Preserving Data Mining" by Agrawal and Srikant (Agrawal & Srikant 2000).

Following this, Broder (Broder 2000) reported on the active battle between web miners (extremely hungry for personalized data) and privacy advocates (resentful of the facilitation of monitoring and tracking technologies for visitation of web sites) after the recent hype for web mining. The conflict is in need of technology that can achieve a balance. In 2001, Estivill-Castro and Brankovic organized a special session on Privacy in Data Mining in the Fifth Multi-Conference on Systemics, Cybernetics and Informatics, which was held in Orlando, Florida, in July 23, 2001.

While some continue to believe that statistical and knowledge discovery and data mining (KDDM) research is detached from this issue, we can certainly see that the debate is gaining momentum as KDDM and statistical tools are more widely adopted by public and private organizations hosting large databases of personal records. Today, the interest is apparent by the appearance in major conferences of research in these topics (Agrawal & Aggarwal 2001, Catlett 2002, Lindell & B. 2000, Vaidya & C. Clifton 2002).

The workshop brought together researchers and practitioners to identify problems and solutions where data mining interferes with privacy and security.

Among the many data mining situations where these privacy and security issues arise some examples are:

- Identifying public health problem outbreaks (e.g., epidemics, biological warfare instances) (Meaney 2001). There are many data collectors (insurance companies, HMOs, public

health agencies). Individual privacy concerns limit the willingness of the data custodians to share data, even with government agencies such as the U.S. Centers for Disease Control. Can we accomplish the desired results while still preserving privacy of individual entities?

- Collaborative corporations or entities. Ford and Firestone shared a problem with a jointly produced product: Ford Explorers with Firestone tires. Ford and Firestone may have been able to use association rule techniques to detect problems earlier. This would have required extensive data sharing. Factors such as trade secrets and agreements with other manufacturers stand in the way of the necessary sharing. Could we obtain the same results, while still preserving the secrecy of each side's data?

  Government entities face similar problems, such as limitations on sharing between law enforcement, intelligence agencies, and tax collection.

- Multi-national corporations. An individual country's legal system may prevent sharing of customer data between a subsidiary and its parent.

The workshop's aim was to bring participants up to speed on the issues and solutions in this area, outline key research problems, and encourage collaborations to address these problems. To this end, a strong program committee reviewed and assessed the quality of submissions. It considered in its assessment the potential of the submission to open discussion and stimulate research on privacy in data mining as well as the quality of the solution in novelty and originality.

The panel selected 7 out 11 submissions. In this process, each paper was reviewed by at least three members of the program committee. Three of the submissions were accepted as full papers and one paper was regarded as the best. Wenliang Du and Zhijun Zhan's paper titled "Building Decision Tree Classifier on Private Data" was selected to compete with regular papers from the IEEE 2002 International Conference on Data Mining to have extended versions considered for possible publication in the Journal of Knowledge and Information Systems. The other two regular papers are Tom Johnsten and Vijay V. Raghavan "A Methodology for Hiding Knowledge in Databases" and Stanley R. M. Oliveira and Osmar R. Zaïane "Foundations for an Access Control Model for Privacy Preservation in Multi-Relational Association Rule Mining".

Another four papers were accepted in a "discussion format". These papers were regarded as contributions that do generate discussion of privacy issues in data mining. But, the expert panel and the reviewers noted questions that were still open and needed further discussion. The authors of these papers were invited to produce rejoinders to the most challenging comments from reviewers. As a result, the editors of these proceedings believe that the workshop has indeed achieved its goal to encourage discussion. Also, because some of these challenges indicate limitations of the current technology, they also pose open problems and new research directions. In this aspect as well, we believe the workshop met its objectives.

Thus, the publications here have followed a rigorous process of peer review. We could not have achieved these delicate selection task without the assistance of the program committee.

We look forward for advances in Privacy, Security and Data Mining. We hope that the technology will evolve into a safe practice. However, the areas that need further exploration and advancement include:

- Privacy and security policies and their implications on data mining, including issues of data collection and ownership.
- Learning from perturbed / obscured data.
- Techniques for protecting confidentiality of sensitive information, including work on statistical databases, and obscuring or restricting data access to prevent violation of privacy and security policies.
- Learning from distributed data sets with limits on sharing of information.
- Algorithms for balancing privacy and knowledge discovery in data mining.
- Use of data mining results to reconstruct private information, and corporate security in the face of analysis by KDDM and statistical tools of public data by competitors.
- Case studies of security and privacy policies and their impact on data mining, e.g., privacy issues in medical databases or analysis of personal records for customer relationship management.
- Controversial applications of knowledge discovery and data mining, including secondary use of personal data, fraud detection, credit record checking, knowledge discovery of competitors' (suppliers') strengths by transaction analysis.

There are many people who made this workshop possible. We thank Dr. Einoshin Suzuki of the organizing committee for ICDM'02. We also thank Prof. John Roddick for assistance in producing the workshop proceedings as Volume 14 of the Conferences in Research and Practice in Information Technology series.

References

Agrawal, D. & Aggarwal, C. (2001), On the design and quantification of privacy preserving data mining algorithms, *in* 'Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems', ACM Press, Santa Barbara, CA.

Agrawal, R. & Srikant, R. (2000), Privacy-preserving data mining, *in* W. Chen, J. Naughton & P. A. Bernstein, eds, 'Proceedings of SIGMOD', SIGMOD RECORD 0163-5808; 2000; VOL 29; NO 2, ACM Press, Dallas, TX, pp. 439–450.

Berry, M. & Linoff, G. (1997), *Data Mining Techniques — for Marketing, Sales and Customer Support*, John Wiley & Sons, NY, USA.

Berson, A. & Smith, S. (1998), *Data Warehousing, Data Mining, & OLAP*, Series on Data Warehousing and Data Management, McGraw-Hill, NY, USA.

Bigus, J. (1996), *Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*, McGraw-Hill, NY.

Bonorris, S. (1995), 'Cautionary notes for the automated processing of data', *IEEE Expert* **10**(2), 53–54.

Brankovic, L. & Estivill-Castro, V. (1999), Privacy issues in knowledge discovery and data mining, *in* C. Simpson, ed., 'AICEC99 Conference Proceedings', Swinburne University of Technology, Australian Institute of Computer Ethics, Melbourne, Australia, pp. 89–99.

Broder, A. (2000), Data mining, the Internet and privacy, *in* B. Masand & M. Spiliopoulou, eds, 'Proceedings of WEBKDD-99, International Workshop on Web usage Analysis and user Profiling', Springer-Verlag Lecture Notes in Artificial Intelligence 1836, pp. 56–73.

Catlett, J. (2002), Among those dark electronic mills: Privacy and data mining, *in* R. Ramakrishnan, ed., 'ACM SIGKDD International Conference ON KNOWLEDGE DISCOVERY AND DATA MINING; 2000; 6TH', ACM, Association for Computing Machinery, Boston, MA, p. 4.

Clarke, R. (1988), 'Information technology and dataveillance', *Communications of the ACM* **31**(5), 498–512.

Clarke, R. (1999), 'Person-location and person-tracking: Technologies, risks and policy implications', Introduction to Dataveillance and Information Privacy, and Definitions of Terms. www.anu.edu.au/people/Roger.Clarke.

Clifton, C. (1999), Protecting against data mining through samples, *in* 'Thirteenth Annual IFIP WG 11.3 Working Conference on Database Security', Seatle, WA.

Clifton, C. (2000), 'Using sample size to limit exposure to data mining', *Journal of Computer Security* **8**(4), 281–307.

Clifton, C. & Marks, D. (1996), Security and privacy implications of data mining, *in* 'SIGMOD Workshop on Data Mining and Knowledge Discovery', ACM, Montreal, Canada.

Estivill-Castro, V., Brankovik, L. & Dowe, D. (1999), 'Privacy in data mining', *Privacy - Law & Policy Reporter* **6**(3), 33–35.

Fayyad, U., Piatesky-Shapiro, G. & Smyth, P. (1996), From data mining to knowledge discovery: An overview, *in* U. Fayyad, G. Piatesky-Shapiro, P. Smyth & R. Uthurusamy, eds, 'Advances in Knowledge Discovery and Data Mining', AAAI Press / MIT Press, Menlo Park, CA, pp. 1–36.

Fayyad, U. & Uthurusamy, R. (1996), 'Data mining and knowledge discovery in databases', *Communications of the ACM* **39**(11), 24–26. Special issue on Data Mining.

Gordon, M. & Williams, M. (1997), Spatial data mining for health research, planning and education, *in* C. Waegemann, ed., 'Proceedings of TEPR-97: Towards an Electronic Patient', Medical Records Institute, Newton, MA. USA, pp. 212–218.

John, G. (1999), 'Behind-the-scenes data mining', *Newsletter of ACM SIG on KDDM* **1**(1), 9–11.

Khaw, Y.-T. & Lee, H.-Y. (1995), 'Privacy & knowledge discovery', *IEEE Expert* **10**(2), 58.

Klösgen, W. (1995), Anonymization techniques for knowledge discovery in databases, *in* U. Fayyad & R. Uthurusamy, eds, 'Proceedings of the First International Conference on Knowledge Discovery and Data Mining', AAAI Press, Menlo Park, pp. 186–191.

Laudon, K. C. (1996), 'Markets and privacy', *Communications of the ACM* **39**(9), 92–104.

Lindell, Y. & B., P. (2000), Privacy preserving data mining, *in* M. Bellare, ed., 'Proceedings of CRYPTO-00 Advances in Cyptology', Springer-Verlag Lecture Notes in Computer Science 1880, Santa Barbara, California, USA, pp. 36–54.

Meaney, M. E. (2001), 'Data mining, dataveillance, and medical information privacy', *BIOMEDICAL ETHICS REVIEWS* pp. 145–164. Conference 1999.

Nash, K. (1998), 'Electronic profiling — critics fear systems may trample civil rights', *Computerworld* **32**(6), 1,28.

O'Leary, D. (1991), Knowledge discovery as a threat to database security, *in* G. Piatetsky-Shapiro & W. Frawley, eds, 'Knowledge Discovery in Databases', AAAI Press, Menlo Park, CA, pp. 507–516.

O'Leary, D. (1995), 'Some privacy issues in knowledge discovery: the OECD personal privacy guidelines', *IEEE Expert* **10**(2), 48–52.

Peacock, P. R. (1998), 'Data mining in marketing: Part 2', *Marketing Management* **7**(1), 15–25.

Piatetsky-Shapiro, G. (1995), 'Knowledge discovery in personal data vs privacy: a mini-symposium', *IEEE Expert* **10**(2), 46–47.

Vaidya, J. & C. Clifton, C. (2002), Privacy preserving association rule mining in vertically partitioned data, *in* 'The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', SIGKDD, ACM Press.