Privacy–Preserving Data Mining

P

Alexandre Evfimievski
IBM Almaden Research Center, USA

Tyrone Grandison
IBM Almaden Research Center, USA

INTRODUCTION

Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Most traditional data mining techniques analyze and model the data set statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure individual data records. This domain separation points to the technical feasibility of PPDM.

Historically, issues related to PPDM were first studied by the national statistical agencies interested in collecting private social and economical data, such as census and tax records, and making it available for analysis by public servants, companies, and researchers. Building accurate socioeconomical models is vital for business planning and public policy. Yet, there is no way of knowing in advance what models may be needed, nor is it feasible for the statistical agency to perform all data processing for everyone, playing the role of a trusted third party. Instead, the agency provides the data in a sanitized form that allows statistical processing and protects the privacy of individual records, solving a problem known as privacy-preserving data publishing. For a survey of work in statistical databases, see Adam and Wortmann (1989) and Willenborg and de Waal (2001).

The term privacy-preserving data mining was introduced in the papers Agrawal and Srikant (2000) and Lindell and Pinkas (2000). These papers considered two fundamental problems of PPDM: privacy-preserving data collection and mining a data set partitioned across several private enterprises. Agrawal and Srikant devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values; Lindell and Pinkas invented a cryptographic protocol for decision tree construction over a data set horizontally partitioned between two parties. These methods were subsequently refined and extended by many researchers worldwide.

Other areas that influence the development of PPDM include cryptography and secure multiparty computation (Goldreich, 2004; Stinson, 2006), database query auditing for disclosure detection and prevention (Dinur & Nissim, 2003; Kenthapadi, Mishra, & Nissim, 2005; Kleinberg, Papadimitriou, & Raghavan, 2000), database privacy and policy enforcement (Aggarwal et al., 2004; Agrawal, Kiernan, Srikant, & Xu 2002), database security (Castano, Fugini, Martella, & Samarati, 1995), and of course, specific application domains.

SURVEY OF APPROACHES

The naïve approach to PPDM is "security by obscurity," where algorithms have no proven privacy guarantees. By its nature, privacy preservation is claimed for all data sets and attacks of a certain class, a claim that cannot be proven by examples or informal considerations (Chawla, Dwork, McSherry, Smith, & Wee, 2005). We will avoid further discussion of this approach in this forum. Recently, however, a number of principled approaches have been developed to enable PPDM, some listed below according to their method of defining and enforcing privacy.

Suppression

Privacy can be preserved by simply suppressing all sensitive data before any disclosure or computation occurs. Given a database, we can suppress specific attributes in particular records as dictated by our privacy policy.

For a partial suppression, an exact attribute value can be replaced with a less informative value by rounding (e.g., $23.45 to $20.00), top coding (e.g., age above 70 is set to 70), generalization (e.g., address to zip code), using intervals (e.g., age 23 to 20-25, name Johnson to J-K), and so forth. Often the privacy guarantee trivially follows from the suppression policy. However, the analysis may be difficult if the choice of alternative suppressions depends on the data being suppressed, or if there is dependency between disclosed and suppressed data. Suppression cannot be used if data mining requires full access to the sensitive values.

Rather than protecting the sensitive values of individual records, we may be interested in suppressing the identity (of a person) linked to a specific record. The process of altering the data set to limit identity linkage is called de-identification. One popular definition for de-identification privacy is *k*-anonymity, formulated in Samarati and Sweeney (1998). A set of personal records is said to be *k*-anonymous if every record is indistinguishable from at least $k − 1$ other records over given quasi-identifier subsets of attributes. A subset of attributes is a quasi-identifier if its value combination may help link some record to other personal information available to an attacker, for example, the combination of age, sex, and address.

To achieve *k*-anonymity, quasi-identifier attributes are completely or partially suppressed. A particular suppression policy is chosen to maximize the utility of the *k*-anonymized data set (Bayardo & Agrawal, 2005; Iyengar, 2002). The attributes that are not among quasi-identifiers, even if sensitive (e.g., diagnosis), are not suppressed and may get linked to an identity (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006). Utility maximization may create an exploitable dependence between the suppressed data and the suppression policy. Finally, *k*-anonymity is difficult to enforce before all data are collected in one trusted place; however, a cryptographic solution is proposed in Zhong, Yang, and Wright (2005) based on Shamir's secret sharing scheme.

Suppression can also be used to protect from the discovery of certain statistical characteristics, such as sensitive association rules, while minimizing the distortion of other data mining results. Many related optimization problems are computationally intractable, but some heuristic algorithms were studied (Atallah, Elmagarmid, Ibrahim, Bertino, & Verykios, 1999; Oliveira & Zaïane, 2003).

Randomization

Suppose there is one central server, for example, of a company, and many customers, each having a small piece of information. The server collects the information and performs data mining to build an aggregate data model. The randomization approach (Warner, 1965) protects the customers' data by letting them randomly perturb their records before sending them to the server, taking away some true information and introducing some noise. At the server's side, statistical estimation over noisy data is employed to recover the aggregates needed for data mining. Noise can be introduced, for example, by adding or multiplying random values to numerical attributes (Agrawal & Srikant, 2000) or by deleting real items and adding bogus items to set-valued records (Evfimievski, Srikant, Agrawal, & Gehrke, 2002; Rizvi & Haritsa, 2002). Given the right choice of the method and the amount of randomization, it is sometimes possible to protect individual values while estimating the aggregate model with relatively high accuracy.

Privacy protection by data perturbation has been extensively studied in the statistical databases community (Adam & Wortmann, 1989; Willenborg & de Waal, 2001). In contrast to the above scenario, this research focuses mainly on the protection of published views once all original data are collected in a single trusted repository. Many more perturbation techniques are available in this case, including attribute swapping across records and data resampling by imputation.

A popular privacy definition to characterize randomization has its roots in the classical secrecy framework (Shannon, 1949) and in the work on disclosure risk and harm measures for statistical databases (Lambert, 1993), but received its current formulation only recently (Blum, Dwork, McSherry, & Nissim, 2005; Dinur & Nissim, 2003; Evfimievski, Gehrke, & Srikant, 2003). To deal with the uncertainty arising from randomization, the data miner's knowledge (belief) is modeled as a probability distribution. A simplified version of the definition is given in the next paragraphs.

Suppose Alice is a customer and Bob is a company employee interested in mining customers' data. Alice has a private record x and a randomization algorithm R. To allow Bob to do the mining while protecting her own privacy, Alice sends Bob a randomized record x' $\sim R(x)$. Let us denote by $p_R(x' \mid x)$ the probability that

algorithm R outputs x' on input x. We say that algorithm R achieves ε-leakage (also called ε-privacy or at most γ-amplification) at output x' if for every pair of private records x_1 and x_2 we have

$$p_R(x'|x_1) \;/\; p_R(x'|x_2) \;\leq\; \gamma, \quad \text{where} \;\; \gamma \;=\; \exp(\varepsilon).$$

We assume that Bob has some a priori belief about Alice's record, defined as the probability distribution $p(x)$ over all possible private records. Once Bob receives a randomized record, his belief changes to some a posteriori distribution. If randomization R achieves ε-leakage at output x', then randomized record x' gives Bob only a bounded amount of knowledge of Alice's unknown private record x. In fact, for every question Q about Alice's record, Bob's a posteriori belief $p(Q|x')$ that the answer to Q is yes is bounded with respect to his a priori belief $p(Q)$ as follows.

$$\frac{\dfrac{p(Q|x')}{1 - p(Q|x')}}{} \;\leq\; \gamma \;\; \frac{p(Q)}{1 - p(Q)}$$

If R achieves ε-leakage at every output, Bob's knowledge gain about Alice's record is always bounded; if R achieves ε-leakage at some outputs but not others, Bob's knowledge gain is bounded only with a certain probability.

The above definition assumes that Bob cannot gain any knowledge of Alice's record by collecting data from other customers, that is, that all customers are independent. The parameter ε is chosen to attain the right balance between privacy and the accuracy of the aggregate estimators used by the data miner (Dwork, McSherry, Nissim, & Smith, 2006). One advantage of randomization is that privacy guarantees can be proven by just studying the randomization algorithm, not the data mining operations. One disadvantage is that the results are always approximate; high-enough accuracy often requires a lot of randomized data (Evfimievski et al., 2002).

Cryptography

The cryptographic approach to PPDM assumes that the data are stored at several private parties who agree to disclose the result of a certain data mining computation performed jointly over their data. The parties engage in a cryptographic protocol; that is, they exchange messages encrypted to make some operations efficient while others computationally intractable. In effect, they blindly run their data mining algorithm. Classical works in secure multiparty computation such as Yao (1986) and Goldreich, Micali, and Wigderson (1987) show that any function $F(x_1, x_2, \ldots, x_n)$ computable in polynomial time is also securely computable in polynomial time by n parties, each holding one argument, under quite broad assumptions regarding how much the parties trust each other. However, this generic methodology can only be scaled to database-sized arguments with significant additional research effort.

The first adaptation of cryptographic techniques to data mining is done by Lindell and Pinkas (2000) for the problem of decision tree construction over horizontally partitioned data; it was followed by many papers covering different data mining techniques and assumptions. The assumptions include restrictions on the input data and permitted disclosure, the computational hardness of certain mathematical operations such as factoring a large integer, and the adversarial potential of the parties involved: The parties may be passive (honest but curious, running the protocol correctly but taking advantage of all incoming messages) or malicious (running a different protocol), some parties may be allowed to collude (represent a single adversary), and so forth. In addition to the generic methodology such as oblivious transfer and secure Boolean circuit evaluation, the key cryptographic constructs often used in PPDM include homomorphic and commutative encryption functions, secure multiparty scalar product, and polynomial computation. The use of randomness is essential for all protocols.

The privacy guarantee used in this approach is based on the notion of computational indistinguishability between random variables. Let X_k and Y_k be two random variables that output Boolean vectors of length polynomial in k; they are called computationally indistinguishable if for all polynomial algorithms A_k (alternatively, for any sequence of circuits of size polynomial in k), for all $c > 0$, and for all sufficiently large integers k,

$$| \text{Prob}\,[A_k(X_k) = 1] \;-\; \text{Prob}\,[A_k(Y_k) = 1] | \;<\; 1/k^c.$$

The above essentially says that no polynomial algorithm can tell apart X_k from Y_k. To prove that a cryptographic protocol is secure, we show that each party's view of the protocol (all its incoming messages and random choices) is computationally indistinguishable from a simulation of this view by this party alone. When simulating the view of the protocol, the party is given everything it is allowed to learn, including

P

the final data mining output. The exact formulation of the privacy guarantee depends on the adversarial assumptions. Goldreich (2004) and Stinson (2006) provide a thorough introduction into the cryptographic framework.

Scalability is the main stumbling block for the cryptographic PPDM; the approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

Summarization

This approach to PPDM consists of releasing the data in the form of a summary that allows the (approximate) evaluation of certain classes of aggregate queries while hiding the individual records. In a sense, summarization extends randomization, but a summary is often expected to be much shorter, ideally of sublinear size with respect to the original data set. The idea goes back to statistical databases, where two summarization techniques were studied and widely applied: sampling and tabular data representation (Adam & Wortmann, 1989; Willenborg & de Waal, 2001). Sampling corresponds to replacing the private data set with a small sample of its records, often combined with suppression or perturbation of their values to prevent re-identification. Tabular representation summarizes data in a collection of aggregate quantities such as sums, averages, or counts aggregated over the range of some attributes while other attributes are fixed, similarly to OLAP (online analytical processing) cubes. Verifying privacy guarantees for tabular data is challenging because of the potential for disclosure by inference. Privacy in OLAP cubes was also studied by Wang, Jajodia, and Wijesekera (2004) and Agrawal, Srikant, and Thomas (2005).

Some of the more recent summarization methods are based on pseudorandom sketches, a concept borrowed from limited-memory data-stream processing. Here is an illustration of one such method. Suppose Alice has a small private set S of her favorite book titles and wants to send to Bob a randomized version of this set. Alice splits S into two disjoint subsets, $S = S_0 \cup S_1$, then constructs her randomized record S_R by including S_1, excluding S_0, and for every book not in S including it into S_R at random with probability 1/2. If there are 1,000,000 possible book titles, S_R will contain around 500,000 items, most of them purely random.

Luckily, however, S_R can be shortened. Let $G(\xi, i)$ be a pseudorandom generator that takes a short random seed ξ and a book number i and computes a bit b_i. Now Alice has a better strategy: Once she selects S_0 and S_1 as before, she sends to Bob a randomly chosen seed ξ such that $G(\xi, \#book) = 0$ for all books in S_0 and $G(\xi, \#book) = 1$ for all books in S_1. Bob can use G and ξ to reconstruct the entire randomized record; and if G is sufficiently well mixing, every book not in S still satisfies $G(\xi, \#book) = 1$ with probability 1/2. Thus, the short seed ξ serves as the summary of a randomized record. For complete analysis, see Evfimievski et al. (2003) and Mishra and Sandler (2006).

The summarization approach is still in its infancy, so more results are likely to come in the future. There has also been some work on combining sketches and approximation techniques with the cryptographic approach (Feigenbaum, Ishai, Malkin, Nissim, Strauss, & Wright, 2001; Halevi, Krauthgamer, Kushilevitz, & Nissim, 2001). Feigenbaum et al. observe that the disclosure of an approximate function $f_{appr}(x) \approx f(x)$ over private data x may be unacceptable even if the exact result $f(x)$ is permitted to disclose; indeed, just by learning whether $f_{appr}(x) \leq f(x)$ or $f_{appr}(x) > f(x)$, the adversary may already get an extra bit of information about x. This issue is important to keep in mind when designing sketch-based PPDM protocols.

APPLICATION SCENARIOS

- **Surveys and Data Collection.** Companies collect personal preferences of their customers for targeted product recommendations, or conduct surveys for business planning; political parties conduct opinion polls to adjust their strategy. The coverage of such data collection may significantly increase if all respondents are aware that their privacy is provably protected, also eliminating the bias associated with evasive answers. The randomization approach has been considered as a solution in this domain (Agrawal & Srikant, 2000; Evfimievski et al., 2002; Warner, 1965).
- **Monitoring for Emergencies.** Early detection of large-scale abnormalities with potential implications for public safety or national security is important in protecting our well-being. Disease outbreaks, environmental disasters, terrorist acts, and manufacturing accidents can often be detected

and contained before they endanger a large population. The first indication of an impending disaster can be difficult to notice by looking at any individual case, but is easy to see using data mining: an unusual increase in certain health symptoms or nonprescription drug purchases, a surge in car accidents, a change in online traffic pattern, and so forth. To be effective, an early-detection system would have to collect personal, commercial, and sensor data from a variety of sources, making privacy issues paramount (DeRosa, 2004; Perez-Pena, 2003).

- **Product Traceability.** Before a product (e.g., a car or a drug) reaches its end user, it usually passes through a long chain of processing steps, such as manufacturing, packaging, transportation, storage, and sale. In the near future, many products and package units will carry a radio frequency identification (RFID) tag and will be automatically registered at every processing step (Finkenzeller, 2003; Garfinkel & Rosenberg, 2005). This will create a vast distributed collection of RFID traces, which can be mined to detect business patterns, market trends, inefficiencies and bottlenecks, criminal activity such as theft and counterfeiting, and so on. However, such extremely detailed business process data are a highly valuable and sensitive asset to the companies involved. Privacy safeguards will be very important to enable cooperative RFID data mining.

- **Medical Research.** Personal health records are one of the most sensitive types of private data; their privacy standards have been codified into law in many countries, for example HIPAA (Health Insurance Portability and Accountability Act) in the United States (Office for Civil Rights [OCR], 2003). On the other hand, data mining over health records is vital for medical, pharmaceutical, and environmental research. For example, a researcher may want to study the effect of a certain gene A on an adverse reaction to drug B (Agrawal, Evfimievski, & Srikant, 2003). However, due to privacy concerns, the DNA sequences and the medical histories are stored at different data repositories and cannot be brought together. Then, PPDM over vertically partitioned data can be used to compute the aggregate counts while preserving the privacy of records.

- **Social Networks.** In business as well as in life, the right connections make a huge difference. Whether it is expertise location, job search, or romance matchmaking, finding new connections is notoriously difficult, not least because the publicly available data are often very scarce and of low quality. Most of the relevant information is personal, copyrighted, or confidential, and therefore kept away from the Web. It is possible that PPDM techniques can be utilized to allow limited disclosure options, prompting more people to engage in productive social networking and guarding against abuse.

FUTURE TRENDS

The main technical challenge for PPDM is to make its algorithms scale and achieve higher accuracy while keeping the privacy guarantees. The known proof techniques and privacy definitions are not yet flexible enough to take full advantage of existing PPDM approaches. Adding a minor assumption (from the practical viewpoint) may slash the computation cost or allow much better accuracy if the PPDM methodology is augmented to leverage this assumption. On the other hand, proving complexity lower bounds and accuracy upper bounds will expose the theoretical limits of PPDM.

One particularly interesting minor assumption is the existence of a computationally limited trusted third party. Computer manufacturers such as IBM produce special devices called secure coprocessors (Dyer et al. 2001) that contain an entire computer within a sealed tamper-proof box. Secure coprocessors are able to withstand most hardware and software attacks, or destroy all data if opened. For practical purposes, these devices can be considered trusted parties, albeit very restricted in the speed of computation, in the volume of storage, and in communication with the untrusted components. It is known that secure coprocessors can be leveraged to enable privacy-preserving operations over data sets much larger than their storage capacity (Agrawal, Asonov, Kantarcioglu, & Li, 2006; Smith & Safford, 2000). Thus, applying them to PPDM looks natural.

If a data mining party cannot get accurate results because of privacy constraints enforced by the data contributors, it may be willing to pay for more data.

Kleinberg, Papadimitriou, and Raghavan (2001) suggest measuring the amount of private information in terms of its monetary value as a form of intellectual property. The cost of each piece of data must be determined in a fair way so as to reflect the contribution of this piece in the overall profit. The paper borrows the notions of fairness from the theory of coalitional games: the core and the Shapley value. Bridging game theory and PPDM could lay the theoretical foundation for a market of private data, where all participants receive appropriate compensations for their business contribution.

Among potential future applications for PPDM, we would like to emphasize data mining in health care and medical research. During the last few years, the attention of the U.S. government has been focused on transitioning the national health care system to an infrastructure based upon information technology (President's Information Technology Advisory Committee [PITAC], 2004); a similar trend occurs or is expected in countries around the world. Within a short time, millions of medical records will be available for mining, and their privacy protection will be required by law, potentially creating an urgent demand for PPDM. In addition to the traditional data mining tasks, new health-care-specific tasks will likely become important, such as record linkage or mining over ontology-based and semistructured data, for example, annotated images.

CONCLUSION

Privacy-preserving data mining emerged in response to two equally important (and seemingly disparate) needs: data analysis in order to deliver better services and ensuring the privacy rights of the data owners. Difficult as the task of addressing these needs may seem, several tangible efforts have been accomplished. In this article, an overview of the popular approaches for doing PPDM was presented, namely, suppression, randomization, cryptography, and summarization. The privacy guarantees, advantages, and disadvantages of each approach were stated in order to provide a balanced view of the state of the art. Finally, the scenarios where PPDM may be used and some directions for future work were outlined.

REFERENCES

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

Aggarwal, G., Bawa, M., Ganesan, P., Garcia-Molina, H., Kenthapadi, K., Mishra, N., et al. (2004). Vision paper: Enabling privacy for the paranoids. *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB '04)* (pp. 708-719).

Agrawal, R., Asonov, D., Kantarcioglu, M., & Li, Y. (2006). *Sovereign joins.* Proceedings of the 22nd International Conference on Data Engineering (ICDE'06).

Agrawal, R., Evfimievski, A., & Srikant, R. (2003). Information sharing across private databases. *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD '03)* (pp. 86-97).

Agrawal, R., Kiernan, J., Srikant, R., & Xu, Y. (2002). *Hippocratic databases.* Proceedings of the 28th International Conference on Very Large Data Bases (VLDB'02), Hong Kong, China.

Agrawal, R., & Srikant, R. (2000). *Privacy preserving data mining.* Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, TX.

Agrawal, R., Srikant, R., & Thomas, D. (2005). Privacy preserving OLAP. *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD'05)* (pp. 251-262).

Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., & Verykios, V. (1999). Disclosure limitation of sensitive rules. *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange* (pp. 45-52).

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal *k*-anonymization. *Proceedings of the 21st International Conference on Data Engineering (ICDE'05)* (pp. 217-228).

Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). *Practical privacy: The SuLQ framework.* Proceedings of ACM Symposium on Principles of Database Systems (PODS'05), Baltimore.

Castano, S., Fugini, M., Martella, G., & Samarati, P. (1995). *Database security.* Addison Wesley.

Chawla, S., Dwork, C., McSherry, F., Smith, A., & Wee, H. (2005). Towards privacy in public databases. *Proceedings of 2nd Theory of Cryptography Conference* (pp. 363-385).

DeRosa, M. (2004). *Data mining and data analysis for counterterrorism.* Center for Strategic and International Studies.

Dinur, I., & Nissim, K. (2003). *Revealing information while preserving privacy.* Proceedings of ACM Symposium on Principles of Database Systems (PODS'03), San Diego, CA.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). *Calibrating noise to sensitivity in private data analysis.* Proceedings of Theory of Cryptography (TCC'06), New York.

Dyer, J. G., Lindemann, M., Perez, R., Sailer, R., Smith, S. W., van Doorn, L., et al. (2001). The IBM secure coprocessor: Overview and retrospective. *IEEE Computer*, pp. 57-66.

Evfimievski, A., Gehrke, J., & Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. *Proceedings of ACM Symposium on Principles of Database Systems (PODS'03)* (pp. 211-222).

Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2002). Privacy preserving mining of association rules. *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)* (pp. 217-228).

Feigenbaum, J., Ishai, Y., Malkin, T., Nissim, K., Strauss, M., & Wright R. (2001). Secure multiparty computation of approximations. *Proceedings of 28th International Colloquium on Automata, Languages and Programming* (pp. 927-938).

Finkenzeller, K. (2003). *RFID handbook: Fundamentals and applications in contactless smart cards and identification* (2nd ed.). John Wiley & Sons.

Garfinkel, S., & Rosenberg, B. (2005). *RFID: Applications, security, and privacy.* Addison-Wesley Professional.

Goldreich, O. (2004). *Foundations of cryptography* (Vol. I, II). Cambridge University Press.

Goldreich, O., Micali, S., & Wigderson, A. (1987). How to play any mental game. *Proceedings of 19th Annual ACM Conference on Theory of Computing (STOC'87)* (pp. 218-229).

Halevi, S., Krauthgamer, R., Kushilevitz, E., & Nissim, K. (2001). Private approximation of NP-hard functions. *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC'01)* (pp. 550-559).

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)* (pp. 279-288).

Kenthapadi, K., Mishra, N., & Nissim, K. (2005). Simulatable auditing. *Proceedings of ACM Symposium on Principles of Database Systems (PODS'05)* (pp. 118-127).

Kleinberg, J. M., Papadimitriou, C. H., & Raghavan, P. (2000). Auditing Boolean attributes. *Proceedings of ACM Symposium on Principles of Database Systems (PODS'00)* (pp. 86-91).

Kleinberg, J. M., Papadimitriou, C. H., & Raghavan, P. (2001). On the value of private information. *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge.*

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics, 9*(2), 313-331.

Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. In *Lecture notes in computer science: Vol. 1880. Proceedings of Advances in Cryptology: Crypto'00* (pp. 20-24). Springer-Verlag.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). *l-diversity: Privacy beyond k-anonymity.* Proceedings of 22nd IEEE International Conference on Data Engineering (ICDE'06), Atlanta, GA.

Mishra, N., & Sandler, M. (2006). Privacy via pseudorandom sketches. *Proceedings of ACM Symposium on Principles of Database Systems (PODS'06)* (pp. 143-152).

Office for Civil Rights (OCR). (2003). *Summary of the HIPAA privacy rule.* U.S. Department of Health and Human Services.

Oliveira, S. R. M., & Zaïane, O. R. (2003). Protecting sensitive knowledge by data sanitization. *Proceedings*

P

of the 3ʳᵈ IEEE International Conference on Data Mining (ICDM'03) (pp. 613-616).

Perez-Pena, R. (2003, April 4). An early warning system for diseases in New York. *New York Times*.

President's Information Technology Advisory Committee (PITAC). (2004). *Revolutionizing health care through information technology.* Author.

Rizvi, S. J., & Haritsa, J. R. (2002). *Maintaining data privacy in association rule mining.* Proceedings of the 28ᵗʰ International Conference on Very Large Data Bases (VLDB'02), Hong Kong, China.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA.

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal, 28*(4), 656-715.

Smith, S. W., & Safford, D. (2000). *Practical private information retrieval with secure coprocessors* (Research Rep. No. RC-21806). IBM T. J. Watson Research Center.

Stinson, D. R. (2006). *Cryptography theory and practice* (3ʳᵈ ed.). Chapman & Hall/CRC.

Wang, L., Jajodia, S., & Wijesekera, D. (2004). Securing OLAP data cubes against privacy breaches. *Proceedings of 2004 IEEE Symposium on Security and Privacy* (pp. 161-175).

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*(309), 63-69.

Willenborg, L., & de Waal, T. (2001). Elements of statistical disclosure control. In *Lecture notes in statistics* (Vol. 155). Springer-Verlag.

Yao, A. C. (1986). How to generate and exchange secrets. *Proceedings of the 27ᵗʰ Annual Symposium on Foundations of Computer Science (FOCS'86)* (pp. 162-167).

Zhong, S., Yang, Z., & Wright, R. N. (2005). Privacy-enhancing *k*-anonymization of customer data. *Proceedings of the 24ᵗʰ ACM Symposium on Principles of Database Systems (PODS'05)* (pp. 139-147).

KEY TERMS

Data Mining: The process of automatically searching large volumes of data for patterns.

De-Identification: Altering a data set to limit identity linkage.

Privacy: Individuals or groups' right to control the flow of sensitive information about themselves.

Pseudorandom Generator: An algorithm that takes a short random seed and outputs a long sequence of bits that look independent random under a certain class of tests.

Randomization: The act of haphazardly perturbing data before disclosure.

Secure Multiparty Computation: A method for two or more parties to perform a joint computation without revealing their inputs.

Statistical Database: A database with social or economical data used for aggregate statistical analysis rather than for the retrieval of individual records.

Summarization: Transforming a data set into a short summary to allow statistical analysis while losing the individual records.

Suppression: Withholding information due to disclosure constraints.