INTERSECTIONS OF LAW AND TECHNOLOGY IN BALANCING PRIVACY RIGHTS WITH FREE INFORMATION FLOW

Christopher Johnson
IBM Almaden Research Center
650 Harry Road, San Jose CA 95120
johnsocm@us.ibm.com

Rakesh Agrawal
Microsoft Search Labs
1065 La Avenida, Mountain View, CA 94043
rakesh.agrawal@microsoft.com

ABSTRACT

Technological advances in information collection and analysis have created a conflict between individual privacy rights and society's interest in free information flow. The increased availability of information offers great benefits to scientific research, business intelligence, and national security. However, individuals are losing control of their personal information, as their medical, financial, and consumer records are often disseminated without their knowledge or consent. Current data protection laws do not strike the proper balance between these compelling interests, while existing information systems are not designed to provide adequate privacy safeguards. Thus, technologists and legislators must work together to create an effective modern privacy regime. In this paper, we demonstrate how technologists can design systems to protect individual privacy throughout the data lifecycle. We also describe how technology-aware legislators can structure laws to limit privacy abuses without obstructing beneficial uses of information.

KEY WORDS

Privacy, access controls, auditing, information flow

1. Introduction

The availability of information has increased dramatically in recent years. Innovations in collection, aggregation, and storage technologies allow private companies and government entities to gather and maintain vast amounts of personal information. Enhanced analytical capabilities enable data collectors to derive significant commercial and scientific value from this information. Although these advances have brought great benefits, they have also introduced serious privacy threats. Many organizations maintain massive databases of personal information, often without the knowledge or consent of the data subjects. This information can be transferred and disseminated easily via modern communication technologies. As a consequence, privacy breaches are increasing [1] and identity theft is growing at an alarming rate [2].

Governments around the world have responded to these privacy threats by enacting various data protection laws. The European Union [3], Canada [4], Japan [5], and Australia [6] have adopted cross-industry laws that limit collection, processing, and disclosure of personal data. The United States has enacted laws that restrict personal information disclosure in select industries such as health care [7] and financial services [8]. However, current laws are not sufficient on their own to safeguard privacy in the current technological environment. Some data protection laws are not broad enough in their application [9] or are not effective in practice [10], while others have failed to keep pace with technology [11]. Conversely, overly-restrictive privacy regulation may unnecessarily obstruct valuable scientific research [12], commercial activity [13], or free speech rights [14].

The technical community has also not sufficiently responded to emerging privacy concerns. Commercial vendors have concentrated development efforts more on securing information from external threat than restricting the use and disclosure of properly obtained information. Researchers have not considered privacy in the design of many new technologies. Most privacy controls that are available operate at a coarse level of granularity and do not accommodate complex disclosure policies. Moreover, such controls are typically offered as separate components rather than standard features of information systems.

We suggest that law and technology should function together to maintain an optimal balance between individual privacy rights and free information flow. In the following sections, we describe how information systems can be architected, and legal regulations can be narrowly tailored, to protect individual privacy rights without unduly restricting legitimate uses of information.

Paper Organization The remainder of this paper is organized as follows. Section 2 discusses methods of privacy-preserving data collection. Section 3 proposes using fine-grained access controls and auditing to govern access and disclosure of personal information. Section 4 emphasizes the importance of mitigating privacy risks in the aggregation and sharing of personal information. Section 5 discusses techniques of de-identifying information prior to publication to preserve individual privacy. Section 6 suggests establishing policies and security requirements for storage of personal information. Section 7 offers our concluding thoughts.

2. Collection

The first step in the data lifecycle is to collect personal information in a manner that respects individual privacy. Businesses use customer loyalty cards, surveys, product registrations, web search tracking, and many other methods to collect large volumes of personal information about consumer preferences for analytic purposes. They compile large databases of this personal information that can be traded and mined for useful commercial insights. In the U.S., the ability to maintain such personal information in private databases is largely unregulated. Nonetheless, disclosing this information may expose customers to embarrassment, surveillance, and other invasions of privacy. Fortunately, businesses can employ various technical solutions to garner the same benefits from data mining without inserting any personally identifiable information into their databases.

One solution is to randomize sensitive information at the point of collection, such that it can be used for aggregate mining purposes, but not identified back to the individual. Privacy-preserving data mining techniques [15] enable analysts to reconstruct the original distribution of the data without revealing personally identifiable values. Algorithms for building classification models and discovering association rules on top of the privacy-preserved data can then be applied with only a small loss of accuracy. Privacy breaches can be further limited using methods proposed in [16].

Another solution, proposed in [17], allows data subjects to retain control over their personal information at all times, choosing the degree to which this information will be exposed. To accomplish this, companies could use automated privacy agents to replace customer identifiers such as telephone number or social security number with unique, local identifiers. Customers would then rely on these agents to conduct all interactions with the company. This method preserves the customer's privacy, but allows the company to conduct data mining operations over its local database by associating each customer's interactions with a local identifier.

These technologies demonstrate that, for many types of data, businesses can employ privacy-preserving collection techniques without sacrificing the commercial value of the data. Therefore, rather than categorically prohibiting the processing of personal data, lawmakers can allow companies to process aggregate data sets as long as there is not a reasonable probability that the data can be attributed to unique individuals.

3. Access and Disclosure

When enterprises collect information that is personally identifiable, such as medical or financial records, they must ensure that the information is accessed and disclosed in strict accordance with privacy policies and individual preferences. Most data protection laws require companies to notify individuals of their privacy policies prior to collecting any personal data. These polices specify who is entitled to access the data, the purposes for which the data may be used, and any third parties to whom the data may be disclosed. Some laws allow individuals to opt-out of disclosing certain personal information to non-affiliated third parties [8], while others require companies to obtain opt-in consent prior to such disclosures [3] [7] [18]. Many laws also obligate companies to account for past disclosures of information upon request of an administrative agency or individual data subject [7] [19] [20]. In an electronic records infrastructure, it is impractical to manage these access, disclosure, and auditing obligations without the assistance of technology.

Information systems should therefore enable companies to control access to personal data in accordance with fine-grained policies and individual preferences. Most current systems instead implement coarse controls that account only for the user's access privileges, and restrict access to particular tables or records, rather than individual data items. Also, these access controls usually operate at the application level, burdening system performance and requiring each application to be separately recoded to incorporate any policy changes.

Hippocratic databases [21] avoid these typical problems by enforcing fine-grained disclosure policies at the level of the data, without requiring changes to applications. These policies account for the privileges of the user, purpose of access, intended recipient, and individual opt-in and opt-out choices. The active enforcement engine intercepts and rewrites application queries according to installed disclosure policies and choices, returning only compliant information [22]. Hippocratic databases also provide efficiency and flexibility benefits that are unavailable in application-level solutions. Further, they can be designed to allow seamless policy changes and to apply multiple policies to the same query.

This fine-grained policy enforcement capability allows regulators to be less restrictive in their protection of personal data without compromising privacy. Rather than restricting access to an entire category of records or the complete record of a particular individual, laws can restrict access to narrow categories of information within each particular record. Hippocratic databases can also facilitate providing consumers with access to their database records, as required by many laws [7] [23] [24]. Regulators should be aware of these technological capabilities in creating a data protection regime.

3.1 Auditing Data Access and Modification

Enterprises should also be able to audit the access and modification history of all personal information stored in their databases. Some laws entitle individuals to request an accounting of past disclosures of their personal information [7] or verification that the enterprise has complied with it policies [3]. An accurate and efficient auditing capability is necessary to comply with these requirements in an electronic records environment.

Auditing also helps to investigate past database access to uncover any abusive practices, assess the effectiveness of access controls, and deter improper disclosures.

Most conventional auditing systems store the results of every database query, including read queries, which do not modify any data. This causes impractically high storage overhead and poor performance, requiring audit logs to be purged on a regular basis. Audit systems should instead be designed in a way that they are practical to operate and can trace access histories for a number of years. The U.S. President's Information Technology Advisory Committee has identified efficient data access tracking (i.e., auditing) as an essential component of an electronic health records infrastructure [25].

The auditing system described in [26] addresses these shortcomings by constructing an audit application largely over existing database infrastructure. It stores all data updates, insertions, and deletions in backlog tables, which are populated using database triggers or replication features. It also records all queries and contextual information in query logs. Upon receiving an audit request, the system generates a list of suspicious queries. Using the query logs and backlog tables, the system can identify the user, time, purpose, recipient, and exact information disclosed for each query that has accessed the specified data. Because it does not store read accesses, which constitute the bulk of database queries in any installation, this audit system requires dramatically less storage overhead. Thus, it offers a practical solution that does not disrupt existing production systems.

The integrity of any auditing system relies on the assurance that no administrators or privileged users will be able to modify the audit logs (query logs and backlog tables in the case of [26]). Cryptographic techniques, such as one-way hashes, partial result authentication codes, and off-site digital notarization services have been proposed to detect tampering in audit logs [27]. Serialized audit records with digital signatures have also been suggested to create "tamper evident" logs, while distributed log storage with write-once, read many hardware may deter or prevent audit log alterations [28].

As database systems acquire the capability to audit past access efficiently and securely, the law should require more accountability for enterprises that store personal information. Currently, the Health Insurance Portability and Accountability Act (HIPAA) [7] requires health care organizations to account for only unlawful disclosures of personal information. The Gramm-Leach-Bliley Act [8] does not specifically impose an auditing requirement on financial institutions. The European Union Directive on Data Protection [3] requires member states to provide individuals with legal recourse to ensure compliance with privacy regulations, but does not outline specific auditing requirements. By mandating increased accountability for access to personal information, data protection laws would further deter improper disclosures and encourage more effective means to investigate privacy breaches.

3.2 Tracking Data Disclosures

The audit systems described above track the queries that accessed a particular data item, but do not always identify who actually disclosed it. Determining who accessed a specific data item may be very helpful in identifying potential sources of a breach, but the size of audit output may be very large for records that are frequently accessed. Hence, the audit may fail to isolate a user who properly accessed a sensitive data item, but wrongfully disclosed it. For example, suppose a doctor accesses a patient record under legitimate authorization and subsequently discloses the patient's confidential test results in violation of policy. In this case, if many users have properly accessed the same test results, the system cannot determine the actual source of disclosure. Thus, it would be desirable to rank audit output based upon the relative probability that each query resulted in the improper disclosure.

Several novel methods are proposed in [29] to rank potential disclosure sources (i.e., queries) based upon their proximity to a suspicious database table. These include a method based on information retrieval, another that uses statistical record linkage, and a third based on a probabilistic derivational model. This research is still in progress, but it demonstrates useful approaches to ranking audit results. Although query ranking may not provide conclusive evidence of who is responsible for a wrongful disclosure, it can greatly assist in auditors in prioritizing an investigation of hundreds of potential sources.

4. Processing

4.1 Aggregation and Identification

"Aggregation" is the gathering together of bits and pieces of information to form a more complete representation of a person [30], while "identification" involves attributing certain pieces of information to a particular person [31]. Aggregation and identification provide many benefits, including personalization of consumer offerings, more accurate credit reporting [31], consolidation of electronic patient records [32], and improved intelligence gathering [33]. However, they also threaten individual privacy by making a composite record easily available to people that would not otherwise be able to locate and combine all of the information themselves. In combination, these disparate pieces of information may reveal intimate details of one's personal life, facilitate surveillance, provide incomplete or distorted identities, or carry stigmas [31]. Publicizing this information may even chill important political or social activity by revealing information that would better be kept anonymous. Thus, creators of legal protections and technological solutions should endeavour to leverage the benefits of aggregation and identification without the attendant privacy burdens.

Currently, there is much more effort toward developing aggregation and identification technologies than in designing legal protections to prevent abuses. For instance, casinos use sophisticated identity resolution software to find potential cheaters and embezzlers [33]. This software combines numerous value attributes to establish unique identities across data silos and attribute various personal information to those identities [34] [35]. Identity resolution has also been used to create master indexes of medical records among distributed data sources [36]. Anonymous entity resolution technologies, which can link owner-maintained information through anonymous indices, are also available [37]. However, there have not been significant legal or market drivers for privacy-preserving entity resolution technologies to date.

The U.S. legal system in particular has failed to keep pace with technological innovations in the aggregation and identification of personal data. Private data brokers use using these technologies to construct digital dossiers of consumers using various pieces of privately and publicly available information [30]. At least five data brokers maintain data on almost all households in the U.S. [38]. ChoicePoint, the largest U.S. data broker, holds more that 19 billion records, including information on criminal histories, insurance claims, and DNA data [39]. Data brokers often make these dossiers available to businesses and government entities without providing notice to the data subjects or an opportunity to access the records.

Unfortunately, data brokers in the United States are not regulated by the provisions of the Fair Credit Reporting Act (FCRA) [23] or the Privacy Act [24] that require notice to consumers, access to compiled records, and an opportunity to correct inaccuracies. They are not subject to the FCRA because their offerings do not meet its definition of a “consumer report,” and they are not regulated by the Privacy Act because the aggregated data does not originate from the government and data brokers are not considered government contractors under the Act.

A pending bill in the U.S. House of Representatives, entitled the Data Accountability and Trust Act (DATA) [40], would protect individuals from some of the threats of personal data aggregation. Specifically, DATA would require private data brokers to (i) submit to Federal Trade Commission (FTC) audits of their information security practices following any privacy breach, (ii) provide consumers with at least annual access to their information and an opportunity to correct or dispute inaccuracies, and (iii) immediately notify the FTC and affected consumers of any breach of the security of their information.

In addition, the U.S. Supreme Court has recognized that aggregation of public information may violate individual privacy rights. The Court denied a request for disclosure of FBI rap sheets under the Freedom of Information Act (FOIA), notwithstanding that public information was used to compile the rap sheets [41]. Federal appellate courts have not applied this same standard to challenges of state Megan’s Laws, which entail aggregation and publication of information about convicted sex offenders, because they did not involve FOIA requests [42]. Thus, there is a conflict in U.S. courts regarding whether disclosure of aggregated information violates individual privacy rights.

If enacted, the DATA bill would be a promising first step toward safeguarding Americans from the privacy threats of data aggregation and identification. Still, legislators should take even further action to empower individual control over personal data, including limitations on the third parties to whom the data may be transferred, restrictions on the purposes for which the data may be disclosed, and outer bounds on personal records retention.

Technology can play an important role by developing aggregation and identification methods that are resistant to malicious attack and preserve the privacy of data subjects. Examples of such methods include anonymous entity resolution [37] and anonymous master person indices [32]. Research should also continue efforts to develop secure and efficient techniques of querying encrypted data [e.g., 43] and indexing, searching, and linking information over distributed, encrypted databases.

4.2 Information Sharing

Another source of privacy breach is information sharing. Enterprises may leak personal information by failing to enforce privacy policies and preferences after legitimate data transfers. They may also inadvertently release personal information as a result of performing aggregate information sharing operations over multiple databases. The following scenarios provide examples of these problems and propose technological solutions.

Enforcement After Data Transfer Suppose a family practice doctor would like to share electronic health records with a cardiac specialist to whom she has referred a patient. The HIPAA Privacy Rule allows the doctor to share these records for treatment purposes, but requires the specialist to comply with the original disclosure obligations of the doctor, including specific patient preferences. Even assuming the specialist is trustworthy, she may later disclose certain information inadvertently if she is not aware of the patient’s unique preferences. Her disclosure obligations become even more complex if the primary doctor resides in another state with privacy laws that are stricter than the baseline established by HIPAA.

Some commentators have suggested that Congress address this problem by adopting a uniform set of healthcare privacy regulations that pre-empt even more restrictive state laws [44]. Ignoring the potential Constitutional issues with this approach, it is possible that technical innovations in policy enforcement could make federal pre-emption unnecessary.

One alternative is to implement policy-based access controls, such as those described in [21] [22] above the primary care physician’s database. The primary care doctor could provide the specialist with authorization to access the patient’s existing medical records and upload additional records via a web interface. The specialist then accesses all information in accordance with enterprise policies and patient preferences. Prior to uploading any of patient data, the specialist uploads his own privacy policy into the primary care database, including any more restrictive provisions of state law. The policy engine then applies simple, conservative conflict resolution rules in simultaneously enforcing the policies of the specialist and the primary care doctor for any records generated by the specialist.

An even better, but more complex, alternative would be to generate "sticky policies" [45] that transfer with the patient records as metadata. The transferee must then be capable of applying the source disclosure policies to any information in its database. Assuming that the entities have interoperable enforcement systems, sticky policies would be much more effective than traditional contractual provisions in ensuring that personal data is always processed in accordance with the patient's expectations.

Querying Multiple Data Sources Another information sharing problem involves processing aggregate queries over autonomous data sources. Suppose that a medical researcher would like to test a hypothesis concerning correlations between certain genetic expressions and adverse reactions to a new drug. The researcher wants to query the databases of a local hospital and a gene bank to test his theory. However, both entities are subject to strict privacy regulations and may not release any personally identifiable health information without patient consent.

Under HIPAA, the entities would have to obtain patient consent or approval of their respective Institutional Review Boards (IRBs) to share the data sets necessary to conduct the research protocol. To conduct this research in the European Union, the entities would likely need to obtain patient consent or approval of a national ethics committee prior to proceeding with the research. Accordingly, these laws may impede important medical research in the interest of protecting patient privacy.

Minimal sharing technologies can promote the type of information sharing necessary for this research protocol without disclosing any personally identifiable data. A commutative encryption methodology is demonstrated in [46] [47] that allows two or more autonomous entities to run queries across their databases in such a way that the results of the query are revealed, but no other data is exposed among the databases. This methodology can be deployed on a web services infrastructure to compute a secure join over two autonomous data sets (in this case, the hospital and the gene bank) without revealing any patient identities or compromising the security of either data set. Secure coprocessors can be used to compute secure joins across multiple sovereign databases [48]. Encrypted data servers have also been employed to process such secure information sharing operations [49].

These information sharing methodologies show how technology can reduce the need regulation in some areas. By implementing distributed access controls and sticky policies, affiliated enterprises can more easily enforce conflicting policy rules and laws across domains. And by using secure co-processors, sovereign entities can compute queries securely across multiple databases, enabling epidemiological research, selective document sharing, and other information sharing operations. Hence, technology can be employed to promote information flow and scientific discovery without increasing privacy risk.

5. Publication

There are many situations in which enterprises seek to publish sets or subsets of personal data. For instance, the census bureau may publish aggregate statistics derived from questionnaire responses without disclosing about the identities of specific citizens. Financial institutions may publish customer data to credit bureaus or affiliates that perform services on their behalf. Medical institutions may provide patient health records to research institutions under an obligation of confidentiality or with specified identifiers removed. Election bureaus may disclose voter registries or aggregate voting results within each district.

In each case, there is a substantial public interest in disclosing the information, but disclosure poses some risk that private information may be revealed. The increased availability of various public data sets and advanced analytic tools amplify the risk that private information may be revealed. Data linkage attacks involve combining publicly known facts with publicly available data sets to re-identify data subjects. For instance, a researcher in [50] combined the birth date, zip code, and gender of the governor of Massachusetts with a public voter registry and a de-identified medical database to reveal the governor's private health records. The potential to mount these types of linkage attacks calls for improved legal and technological protections for published information.

Laws that allow publication of naively de-identified data are insufficient. In a multi-dimensional world, subjects may be re-identified based on a variety of unique characteristics. Data protection laws must therefore require increased standards for de-identification to remove any reasonable probability that sensitive records can be re-identified using publicly available information.

De-identification techniques should be resistant to data linkage attacks, but robust for valuable data mining operations. Methods that rely on data condensation, scrambling, or swapping may not provide accurate mining results because the data is not entirely truthful. A method known as ***k***-anonymity [50] [51] attempts to strike a workable balance between these two objectives. In a ***k***-anonymized data set, each record is indistinguishable from at least ***k*** - 1 other records. The process of ***k***-anonymization involves data suppression (deleting cell values or entire records) and cell-value generalization (replacing specific values with more general ones). Larger values of ***k*** provide greater privacy protection. A simple greedy approximation algorithm is proposed in [52] to ***k***-anonymize data along multiple dimensions. In addition, a process called ***l***-diversity is introduced in [53] to protect ***k***-anonymized data sets against attacks that use background knowledge to re-identify data subjects.

Because even simple restrictions of optimized ***k***-anonymity are computationally intractable, an optimal method is proposed in [54] to handle the combinatorics of the problem. "The resulting algorithm finds optimal ***k***-anonymizations under two representative cost measures and wide range of ***k***. . . . The algorithm can produce good anonymizations in circumstances where the input data or input parameters preclude finding an optimal solution in a reasonable amount of time" [54]. This de-identification method produces useful data sets that preserve individual privacy, but retain their integrity for analytical purposes.

Analyzing published data sets provides immense benefits to research, so simply disallowing their publication is not desirable. Governments should instead enact laws that control for the risk of linkage attacks by requiring robust de-identification techniques, rather than naïve scrubbing of identifiers

6. Retention

Another element of privacy protection concerns the retention of personal information. As electronic data storage becomes more available and less expensive, enterprises can store increasing amounts of information for longer periods of time. Sensitive data is also moving to inexpensive personal devices such as memory keys, portable disks, and smart cards. This growing amount of stored data in distributed locations increases the risk of improper information disclosure. Therefore, it is vital to secure these storage devices and retain data pursuant to legal requirements and data subject preferences.

Law and technology should play critical roles in preserving the privacy of retained data. The law should define how long particular data must be retained, the purposes for which it may be retained, and the required security of the information during retention. On the technology side, data retention policies should be designed to govern the storage of all personal information. These policies should resolve conflicts between legal jurisdictions, retention purposes, and data subject preferences. At the end of the retention period, storage systems should be able to remove expired data and forget any persistent data that would allow recreation. Research has explored data lifecycle management and enforcement of retention policies in storage systems [55] [56]. Storage devices must also be secure from data contamination, unintended loss, and leakage. Methods of embedding encryption into microprocessors have recently been proposed to secure data on mobile devices [57].

7. Conclusion

We have demonstrated how law and technology can work in concert to protect individual privacy rights, while maximizing the value gained from available information.

Laws can prevent technology-enabled privacy abuses, such as intrusive surveillance or attribution of sensitive information to uniquely identifiable individuals. Technologists can develop innovations, such as privacy-preserving data mining, that deliver significant value with much lower privacy risks. Technology can even obviate the need for overly-restrictive laws, such as categorical prohibitions of data processing, if information systems are built with sufficient privacy protections. Advances in technology, such as disclosure auditing and remote data access capabilities, can also make feasible certain regulations that may strengthen privacy protections.

Therefore, lawmakers must stay informed about the state of technology as they endeavour to strike a balance between protecting individual privacy and fostering valuable and productive uses of available data. Technologists, on the other hand, must understand the implications of their inventions and take responsibility for their consequence. They should also engage in dialogue with policy makers and develop minimally intrusive, privacy-preserving methods of collecting, analyzing, sharing, and storing personal information.

References

[1] www.privacyrights.org/ar/ChronDataBreaches.htm

[2] U.S. Federal Trade Commission, Identity Theft Victim Compliant Data: Figures and Trends, January 2006.

[3] E.U. Directive on Data Protection, ***Official Journal of the European Communities,*** 23 November 1995.

[4] Personal Information Protection and Electronic Documents Act, Statutes of Canada 2000.

[5] Law on the Protection of Personal Information, promulgated by the Diet of Japan on May 30, 2003.

[6] Australian Privacy Act of 1988, as amended in 2000.

[7] Health Insurance Portability and Accountability Act of 1996, United States Public Law 104-191.

[8] Financial Modernization Act of 1999, 15 U.S.C. §§ 6801-6809.

[9] D. Solove and C. Hoofnagle, A Model Regime of Privacy Protection, ***University of Illinois Law Review***, Vol. 2006, 364-367.

[10] T. Janger, P. Schwartz, The Gramm-Leach-Bliley Act, Information Privacy, and the Limit of Default Rules, ***Minnesota Law Review***, Vol. 86, pages 1219 et seq, 2002.

[11] D. Solove, Reconstructing Electronic Surveillance Law, ***George Washington Law Review***, Vol. 72, 2004.

[12] S. Pincock, Red Tape Entangles Epidemiology, ***The Scientist***, January 2006.

[13] P. Swire and R. Litan, ***None of Your Business: World Data Flows, Electronic Commerce, and the European Privacy Directive***, Brookings Institution Press, 1998.

[14] E. Volokh, Freedom of Speech and Information Privacy, Stanford Law Review, Volume 52, May 2000.

[15] R. Agrawal and R. Srikant, Privacy-Preserving Data Mining. ***Proc. Of the ACM SIGMOD Conference on Management of Data***, Dallas, Texas, USA, May 2000.

[16] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting Privacy Breaches in Privacy Preserving Data Mining. ***Proc. of the 22nd ACM Symposium on Principles of Database Systems***, San Diego, CA, June 2003.

[17] G. Aggarwal, M. Bawa, P. Ganesan, H. Garcia Molina, K. Kenthapadi, N. Mishra, R. Motwani, U. Srivastava, D. Thomas, and J. Widom, Enabling Privacy for the Paranoids. ***Proc. of the 30th Int'l Conf. on Very Large Databases***, Toronto, Canada, August 2004.

[18] California Financial Information Privacy Act, Financial Code § 4050-4060.

[19] Food and Drug Administration Act, 21 CFR Part 11.

[20] Securities and Exchange Commission Rule 17a-4.

[21] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, Hippocratic Databases. ***Proc. of the 28th Int'l Conf. on Very Large Databases***, Hong Kong, China, August 2002

[22] K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, and D. DeWitt. Limiting Disclosure in Hippocratic Databases. ***Proc. of the 30th Int'l Conf. on Very Large Databases***, Toronto, Canada, August 2004.

[23] Fair Credit Reporting Act, 15 U.S.C. §§ 1681 et seq.

[24] Privacy Act of 1974, 5 U.S.C. § 552a.

[25] U.S. President's Information Technology Advisory Committee, Revolutionizing Health Care through Information Technology, June 2004.

[26] R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzau, and R. Srikant, Auditing Compliance with a Hippocratic Database. ***Proc. of the 30th Int'l Conf. on Very Large Databases***, Toronto, Canada, August 2004.

[27] R. Snodgrass, S. Yao, and C. Collberg, Tamper Detection in Audit Logs. ***Proc. of the 30th Int'l Conf. on Very Large Databases***, Toronto, Canada, August 2004.

[28] Markle Foundation, Implementing a Trusted Information Sharing Environment: Using Immutable Audit Logs to Increase Security, Trust and Accountability, New York, February 2006.

[29] http://www.almaden.ibm.com/software/projects/iis/hdb/Publications/papers/Query_Ranking_Research.pdf**.**

[30] D. Solove, ***The Digital Person*** (New York, NY: NYU Press, 2004).

[31] D. Solove, A Taxonomy of Privacy, ***University of Pennsylvania Law Review.***, January 2006, 507-16.

[32] P. Swire, Research Report: Application of IBM Anonymous Resolution to the Health Care Sector. http://www.peterswire.net/AR_White_Paper.pdf.

[33] D. Kushner, Vegas 911. ***IEEE Spectrum***, April 2006.

[34] IBM DB2 Identity Resolution, http://www-306.ibm.com/software/data/db2/eas/identity/.

[35] O. Benjelloun, H. Garcia-Molina, J. Jonas, Q. Su, and J. Widom, Swoosh: A Generic Approach to Entity Resolution. Stanford University Technical Report, 2005.

[36] S. Ellard, System and Method for Indexing Information about Entities from Different Information Sources. U.S. Patent No. 5,991,758. November 23, 1999.

[37] IBM DB2 Anonymous Resolution, http://www-306.ibm.com/software/data/db2/eas/anonymous/.

[38] A. M. Hughes, ***The Complete Database Marketer*** (Chicago, IL, Irwin, 1996).

[39] M. Rotenberg, Testimony before the U.S. House of Representatives, Subcommittee on Commerce, Trade and Consumer Protection, Committee on Energy and Commerce, March 15, 2005.

[40] United States House of Representatives Bill H.R. 4127, 109th Congress, 1st Session, October 25, 2005.

[41] ***U.S. Dept. of Justice v. Reporters Committee for Freedom of the Press,*** 489 U.S. 749, 780 (1989).

[42] ***Russell v. Gregoire***, 24 F.3d 1079 (9th Cir. 1997); ***Paul P. v. Verniero***, 170 F.3d 396 (3d Cir. 1999).

[43] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, Order-Preserving Encryption for Numeric Data. ***Proc. of the ACM SIGMOD Conference on Management of Data***, Paris, France, June 2004.

[44] Commission on Systemic Interoperability, Ending the Document Game. U.S. Government Printing Office, October 2005 (http://www.endingthedocumentgame.gov).

[45] G. Karjoth, M. Schunter, and M. Waidner, The Platform for Enterprise Privacy Practices – Privacy-Enabled Management of Customer Data. ***Proc. of the 2nd Workshop on Privacy Enhancing Technologies***, 2002.

[46] R. Agrawal, A. Evfimievski, and R. Srikant, Information Sharing across Private Databases. ***Proc. of the ACM SIGMOD Conference on Management of Data***, San Diego, California, June 2003.

[47] B. Huberman, M. Franklin, and T. Hogg, Enhancing Privacy and Trust in Electronic Communities. ***Proc. of the 1st ACM Conference on Electronic Commerce***, Denver, Colorado, November 1999, 78–86.

[48] R. Agrawal, D. Asonov, M. Kantarcioglu, and Y. Li, Sovereign Joins. ***Proc. of the 22nd Int'l Conf. on Data Engineering***, Atlanta, Georgia, April 2006.

[49] K. Goldman and E. Valdez, Matchbox: Secure Data Sharing, ***IEEE Internet Computing***, November 2004.

[50] L. Sweeney, ***k***-Anonymity: a Model for Protecting Privacy. ***International Journal on Uncertainty, Fuzziness and Knowledge-based Systems***, 10 (5), 2002; 557-570.

[51] P. Samarati, L. Sweeney, Generalizing Data to Provide Anonymity when Disclosing Information. ***Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems***, 1998, 188.

[52] K. LeFevre, D. DeWitt, and R. Ramakrishnan, Mondrian Multidimensional K-Anonymity. ***Proc. of the 22nd Int'l Conf. on Data Engineering***, Atlanta, Georgia, April 2006.

[53] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, ***l***-Diversity Beyond ***k***-Anonymity. ***Proc. of the 22nd Int'l Conf. on Data Engineering***, Atlanta, Georgia, April 2006.

[54] R. Bayardo and R. Agrawal, Data Privacy through Optimal ***k***-Anonymization. ***Proc. of the 21st Int'l Conf. on Data Engineering***, Tokyo, Japan, April 2005.

[55] Y. Chen, S. Ong, Holistic Information Management Solutions. IBM Research Report, July 11, 2005.

[56] David Reiner, Gil Press, Mike Lenaghan, David Barta, Rich Urmston, Information Lifecycle Management: The EMC Perspective. ***Proc. of the 20th Int'l Conf. on Data Engineering***, April 2004.

[57] M. Hines, IBM Touts Chip-Level Security. ***EWeek: Enterprise News and Reviews***. April 10, 2006.