Jamie Yoo
CPSC 457
May 1, 2006

Controlling Sensitive Information

"Today we stand at another computational crossroads. We are moving past the 1960s vision of computers that hold important financial, education, and credit information. We are moving into an integrated future in which computers will track the most mundane and the most intimate aspects of our lives."

—Simson Garfinkel, Database Nation

INTRODUCTION

During the past decade, there has been a dramatic growth in the quantity of personal information that is being collected and sold, and the expansion of the Internet has significantly facilitated this development. Unfortunately, the data collection industry has gone largely unregulated, allowing databanks like ChoicePoint and Acxiom to collect massive amounts of information and sell them to third parties, without any express permission from the data subjects. Insufficient regulation of this industry has not only led to blatant violations of the privacy rights of the individuals whose information is being sold but also to security risks. This paper will discuss a standard evaluating the process of personal information collection, the ways in which the current system fail to meet these standards, and technical and legal suggestions for addressing these failures.

FAIR INFORMATION PRACTICES

"Over the past quarter century, government agencies in the United States, Canada, and Europe have studied the manner in which entities collect and use personal information – their 'information practices' – and the safeguards required to assure those practices are fair and provide adequate privacy protection. The result has been a series of reports, guidelines, and

model codes that represent widely-accepted principles concerning fair information practices. Common to all of these documents are five core principles of privacy protection: (1) Notice/Awareness; (2) Choice/Consent; (3) Access/Participation; (4) Integrity/Security; and (5) Enforcement/Redress."

Over three decades ago in 1973, a task force at the U.S. Department of Health Education and Welfare, or HEW, set out to analyze the impact of computerization of information on medical records privacy. At the end of the investigation, the task force presented the Code of Fair Information Practices, which consisted of five basic principles: openness, disclosure, secondary use, correction, and security. In the following years, various countries adopted the Principles as law, and then in 1980, the Organization of Economic Cooperation and Development, or the OECD, adopted an expanded set of eight principles as part of the "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data" (OECD). The OECD is an international body comprised of 24 countries throughout the world, including the United States. However, while most other industrialized countries have codified the principles into omnibus privacy laws, the United States has yet to pass such a law at the federal level, although the Principles have been used as a reference for sector-specific laws, such as the Fair Credit Reporting Act, the Right of Financial Privacy Act, the Electronic Communications Privacy Act, and the Video Privacy Act (Privacy Rights Clearinghouse). Thus, even given the United States' failure to codify the Principles, it is well established that the principles provide for an effective framework for discussing the standards by which the process of personal data collection should be executed. Consequently, the following section will briefly review the Principles so that both current common practices as well as suggested practices can be evaluated using these standards.

The Principles

The following are the eight fair information practice principles as described by the OECD's

"Guidelines on the Protection of Privacy and Transborder Flows of Personal Data":

(1) Collection limitation: There should be limits to the collection of personal data and any such data should be obtain by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

(2) Data quality: Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete, relevant and kept up-to-date.

(3) Security safeguards: Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

(4) Openness: There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available for establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

(5) Purpose specification: The purpose for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

(6) Use limitation: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified as described above, except with the consent of the data subject or by the authority of law.

(7) Individual participation: An individual should have the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him; c) to be given reasons if a request is denied and to be able to challenge such denial; and d) to challenge data relating to him and, if the challenge is successful, to have the data erased, rectified, completed or amended.

(8) Accountability: Individuals controlling the collection or use of personal information should be accountable for taking steps to ensure the implementation of the Fair Information Practices.

INFORMATION RESELLERS

In the United States, these principles are generally considered to be guidelines and do not have the force of law. Thus, for instance, by 1983, 182 American companies had claimed to have adopted the guidelines, but the application of the principles varied and very few every implemented practices that mapped directly to the guidelines (Schwartz). Today, the limitations of having principles and not laws are exemplified by the practices of the data brokering industry.

Data brokers, also called data or information resellers, are businesses that collect and aggregate information from multiple sources and make it available to their customers (GAO 2). Not all resellers focus exclusively on aggregating and reselling personal information. For example, Dun & Bradstreet primarily provides information on commercial enterprises (D&B), but in doing so, even these types of resellers also deal with the personal information of the individuals who are associated with those enterprises. When it comes to personal information, the data resellers generally work with three types of information:

(1) Public records such as birth and death records, property records, motor vehicle and voter registrations, criminal records, and civil case files.

(2) Publicly available information not found in public records but nevertheless publicly available through other sources, such as telephone directories, business directories, classified ads or magazines, Internet sites, and other sources accessible by the general public.

(3) Nonpublic information derived from proprietary or nonpublic sources, such as credit header data, product warranty registrations, and other application information provided to private businesses directly by consumers.

Types of customers that data resellers market to vary significantly and include commercial enterprises, non-profit organizations, individuals, as well as government agencies. For example, according to the GAO, the Departments of Justice, Homeland Security and State

and the Social Security Administration "spent approximately $30 million on contractual arrangements with resellers" (2) in 2005.

Unfortunately, upon inspection, it becomes obvious that many times these data resellers do not follow the guidelines for fair information practices that were listed in the previous section. To begin with, when considering the first principles, collection limitation, although data resellers do follow data collection restrictions that are expressly regulated by laws, such as the Fair Credit Reporting Act and the Gramm-Leach-Bliley Act, beyond these specific legal restrictions, information resellers generally tend to maximize the quantity and variety of personal information that they aggregate in order to provide data valuable to a broad range of customers. Resellers also collect information from a wide variety of sources, including state motor vehicle records, local government records on births, real property, and voter registrations, and various court records, as well as from telephone directories, Internet sites, and consumer applications for products or services. Thus the extensive diversity of sources and types of information illustrate the broad nature of the collection of personal information done by data resellers, which completely undermines the first principle of fair information practices.

Regarding the principle second principle, data quality, data resellers fail to comply in several ways. First, the personal data that resellers acquire are rarely relevant to the purposes for which they were collected, since data resellers tend to maximize the amount of data they aggregate in order to maximize potential uses for potential customers. In a general sense, information resellers find it sufficient to specify their purpose by indicating the business categories of the customers for whom they collect information. Oftentimes, it is difficult for resellers to provide greater specificity because they make their data available to many customers for a wide range of legitimate purposes.

Fortunately, since the data that resellers aggregate and store in their databases is their product, they do tend to make efforts to implement security measures. At the same time, however, there have been known breeches of these systems, the most infamous being the 2005 ChoicePoint incident where "scammers culled the personal information of tens of thousands of Americans in a recent attack on [ChoicePoint's] consumer database, resulting in 750 individual cases of identity theft" (Hines).

The fourth principle of openness is in a large sense ignored by the data resellers. Many times data subjects are completely unaware that these companies have collected any information about them. Additionally, resellers, particularly marketing firms, very rarely disclose the individual sources of each piece of information.

The fifth principle is perhaps the most blatantly violated. At the time of collection, the data subject is probably informed of a purpose for which personal data is being collected. However, this purpose rarely includes any of the dozens of purposes for which the data is sold and resold.

In terms of consent, data subject very rarely are informed about the acquisition of their personal information by data brokers, let alone asked for their consent. Resellers have argued that it may not be appropriate or practical for them to provide notice or obtain consent from individuals. They contend that since in many instances the company does not have a direct relationship with the data subject, they are therefore not in a position to interact with the consumer for purposes such as providing notice. Additionally, resellers often argue that requiring resellers to notify and obtain consent from each individual about whom they obtain information would result in consumers being overwhelmed with notices and negate the value of notice. Some information resellers offer consumers an "opt-out" option where individuals can

request that their information be suppressed from selected databases. However, resellers generally offer this option only with respect to selected types of information and under limited circumstances, and many times they will still reserve the right to deny requests.

An individual should have the right to be informed that a data controller has data relating to him and to be able to request correction of inaccurate information. While the data resellers do have interest in maintaining accurate information, the costs of checking each pierce of information seems to be higher than the price that they are willing to pay. In many cases, it is difficult for a data subject to request that inaccurate information be corrected or removed because the resellers often require a considerable amount of paperwork and patience.

Individuals controlling the collection or use of personal information should be accountable for taking steps to ensure the implementation of the Fair Information Practices since they are simply guidelines. When a company is found liable, it is usually because it failed to follow policies that it defined for itself. However, the company is not obligated to implement all the Principles in those policies.

FIGHTING THE LEAKING OF PERSONAL INFORMATION

Realizing that the practice of over-collecting of personal data was becoming common practice without the proper application of many of these fair information practices principles, technology experts have made efforts to develop tools that could help individuals retain some of privacy. The following sections will discuss two specific examples of these tools: P3P and EPAL.

P3P

"At its most basic level, P3P is a standardized set of multiple-choice questions, covering all the major aspects of a Web site's privacy policies. Taken together, they present a clear snapshot of how a site handles personal information about its users."

—W3C

The Platform for Privacy Preferences Project was aimed at providing a standardized, XML-based policy specification language for the purpose of being used to specify an organization's privacy practices in a machine readable language that can be parsed and used by policy-checking agents on the user's behalf. For instance, a Web browser can compare an organizations P3P policies with the user's set preferences and then decide whether it will allow the page to load, prevent the page from loading, or warn the user that the site does not comply. While at first P3P was praised as a much needed step towards privacy-enhancing technology, it has also been criticized a variety of reasons, such as the limited vocabulary that it recognizes as well as the fact that it does not provide any mechanism for enforcement or monitoring of the organization's activities. Thus there is no way for the data subject to know whether the policy preferences that he has specified are actually being respected.

EPAL

IBM's Enterprise Privacy Authorization Language was designed as a tool for making privacy policies into machine-enforceable. EPAL is also an XML-based privacy policy specification language, but unlike P3P, EPAL is intended to specify internal privacy policies for an organization. In other words, EPAL is a formal language that organizations can use to automate and enforce privacy policies across IT applications and systems. EPAL policies, unlike P3P are enforceable, since they are expressed in a manner similar to access control policies.

WHAT IS MISSING?

P3P and EPAL are reasonable starting points for the development of privacy-enhancing systems but they are still short of ideal, particularly since they lack any mechanisms for accountability. Although the two languages are different in that P3P is data-centric while EPAL is access-centric, both are limited by the fact that they are simply specification languages and thus provide no real mechanism for accountability. Thus, an enterprise is still relatively free to define whatever policies it feels, interpreting the fair information practices as liberally as it chooses. Not only that, but these standards have also been criticized for their "opt-out" approach to privacy, instead of a user-enterprise negotiation approach. The following section will propose a system in which these issues, as well as the other fair information practice principles, will be addressed.

RECOMMENDATION

The system that I will now recommend is based on three primary concepts: (1) Trusted Auditing Authority, (2) Identity Based Encryption, and (3) Trusted Computing. At the most basic level, in other words, the system that is recommended for enhancing privacy protections in the data brokering industry will be based on a trusted auditing authority who serves both as a certifying as well as an auditing authority. More specifically, when a user discloses personal information, it is encrypted and bound to a set of privacy preferences, which may be specified using P3P-like technology. The trusted authority will then interpret those privacy preferences and compare them to the privacy environment of the receiving organization. Once it has confirmed that the receiving organization has an approved privacy environment, the trusted authority will issue the IBE decryption key. The trusted authority will also maintain logs of this

disclosure in order to establish a type of a paper trail. Similarly, this process can be used when data is disclosed by one organization to another. Since the trusted authority is thus maintaining a log of disclosures, it is then able to provide the data subject option of inquiring who has been given his personal information. In order to fully enhance the privacy environment of these organizations that will be receiving personal data, the system should ideally be based on Trusted Computing technology. Only then can you have truly "sticky" privacy policies that are permanently linked to the personal information that was disclosed.

Trusted Auditing Authority

The concept of a trusted third party is not uncommon in cryptographic protocols. A conventional certificate authority acts as a trusted third party when it issues digital certificates for use by each party. However, instead of simply issuing certificates to verify credentials, the third party in the recommended system should also act as an auditing and tracking service. More specifically, the auditor should only issue certificates to organizations that can guarantee to a reasonable extent that it has a privacy environment that will respect the user's privacy preferences.

Having an auditing authority will undoubtedly help to guarantee the application of several of the fair information practice principles. For one, users will have a more active role in the entire process of data sharing and selling. As a result, most of the fair information practice principles, including data collection, data quality, openness, and purpose specification, can all be addressed by the auditor, who will have the ability to approve or reject an organizations privacy environment. Additionally, the auditing process will also help dramatically with the accountability concern, since establishing a paper trail will allow both data subjects and law

enforcement agencies to better investigate where leakages happen and other privacy violations occur.

Trusted Computing

In particular, we are interested in the "sealed storage" concept of Trusted Computing, so that the personal information stored by data collectors can be protected and can only be read by the a specific combination of software and hardware. For instance, the Trusted Computing environment will allow "sticky" privacy policies, which means that the polices preferred by the data subject are permanently bound to the data itself, perhaps in the form of metadata. If the data recipient attempts to use the data in ways that do not comply with the privacy policy, or if the recipient attempts to tamper with the privacy policy itself, the Trusted Computing environment can then either render the data useless or delete it altogether. For instance, if the data subject expects the information to be deleted after five years, the system based on Trusted Computing be programmed to automatically enforce this rule and allow the data to "self destruct." As a result, Trusted Computing is a convenient mechanism that the Trusted Authority can use to confirm that the receiving organization has a privacy friendly computing environment. Thus in this fashion, Trusted Computing helps a great deal with enforcement and accountability.

While there exist some controversy over Trusted Computing, much of the debate is irrelevant to the proposed system, particularly since we are mostly concerned with Trusted Computing being implemented on commercial database servers. Thus, the "Can you trust your computer?" concern that applies to personal computers is not really an issue in this system (Stallman).

LIMITATIONS

One inherent limitation is that this recommendation still relies on the ability to map a policy defined by a natural language onto one defined by a machine readable language. While there do exist policies that are clear and unambiguous, there are arguably just as many legal policies and principles that are deliberately ambiguous, particularly when the legal principle is based on common law. If there were not ambiguity in our legal code, a significant portion of the burden on our court systems could be removed. In many circumstances, legislators are reluctant to define clear boundaries and rules because there is the understanding that it is difficult to predefine every legitimate exception that may occur. At the same time, however, regardless of this inherent difficulty, the system that was described in the previous sections is still an important first step because it at least provides for a mechanism for implementing a minimal set of standards that can be expanded or condensed as necessary as time progresses.

Another limitation is that secure trusted third party networks can be difficult to build, particularly in terms of establishing a reputation as being both trustworthy and secure. For one thing, there is inherent difficulty in establishing trust in an unknown third party, particularly when trusted third party networks are viewed by some as having inherent security vulnerabilities. However, even while it may be true that finding reliable trusted third parties is difficult, it is not impossible. With the help of appropriate legislation, there is no real reason why a digital auditing system cannot be implemented in the same fashion as the financial auditing system.

Additionally, Trusted Computing environments tend to be more expensive and more difficult to maintain than traditional systems, particularly since much of the Trusted Computing technology is implemented in hardware and not software. Unless this system is required by law,

like the financial auditing system, higher costs will discourage at least a minimum number of companies from using this data auditing system. This will work against maximizing the number of member organizations in this system, which is important because the effectiveness and usefulness of the system is dependent on having a large network of organization who can disclose information through the trusted auditing authority system. However, besides Trusted Computing, there has yet to be any other way of binding data permanently to its policy preferences to the extent where if polices are violated, the data becomes unusable.

This recommended system as a whole is somewhat of a costly process in terms of time and resources. Requiring that each disclosure be essentially approved and logged by a trusted third party will most certainly come at the cost of efficiency. However, this problem is a much larger question that asks what price are we as a society willing to pay to ensure that our privacy is protected. This question is one that in many ways has yet to be answered for our non-digital interactions as well, and goes beyond the scope of this paper.

CONCLUSION

The data brokering industry has flourished in the recent decade or so, and its privacy practices has gone largely unregulated. Even in cases where regulations do exist, there is relatively no mechanism for enforcement or accountability. Thus, it is important to develop privacy-enhancing tools that can be implemented in conjunction with legislative efforts. Otherwise, privacy will most likely be increasingly violated as commercial enterprises become gradually more accustomed to easy access to vast amounts of proprietary consumer data, until eventually privacy is only a myth and no longer a right that can be protected.

Bibliography

Garfinkel, Simson. Database Nation. Sebastopol, CA: O'Reilly & Associates, 2000.

Organization of Economic Cooperation and Development. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Paris: OECD Publications, 2001.

Privacy Rights Clearinghouse. "A Review of the Fair Information Principles: The Foundation of Privacy Public Policy." Feb 2004. <http://www.privacyrights.org/ar/fairinfo.htm>.

Schwartz, Ari. "HR 4049 Privacy Commission Act." Testimony of Ari Schwartz, Policy Analyst, The Center for Democracy and Technology Before the House Committee on Government Reform Subcommittee on Government Management, Information and Technology. 12 Apr 2000. <http://www.cdt.org/testimony/000412schwartz.shtml>.

Government Accountability Office. "Personal Information: Agencies and Resellers Vary in Providing Privacy Protections." GAO Testimony Before the Subcommittee on Commercial and Administrative Law and the Subcommittee on the Constitution, Committee on the Judiciary, House of Representatives. 4 Apr 2006.

Dun & Bradstreet. "About D&B." <www.dnb.com>.

Hines, Matt. "ChoicePoint data theft widens to 145,000 people." CNET News.com. 18 Feb 2005. <http://news.com.com/ChoicePoint+data+theft+widens+to+145%2C000+people/2100-1029_3-5582144.html?tag=nl>.

Stallman, Richard. “Can you trust your computer?” Free Software, Free Society: The Selected Essays of Richard Stallman. 2002. <http://www.gnu.org/philosophy/can-you-trust.html>.