PRIVACY AND TECHNOLOGY: EXPLORING GOVERNMENT USE OF COMMERCIAL DATA FOR HOMELAND SECURITY – A ROADMAP FOR THE FUTURE

Highlights of the DHS Privacy Office Workshop

September 8 and 9, 2005

Introduction

On September 8 and 9, 2005, the Department of Homeland Security Privacy Office hosted its first public workshop, *Privacy and Technology: Exploring Government Use of Commercial Data for Homeland Security* (Workshop). The objective of the Workshop was to look at the policy, legal, and technology issues associated with the government's use of personally identifiable commercial data in protecting the homeland. The Workshop panelists represented a broad range of expertise and perspectives, including representatives from academia, business leaders, privacy advocates, legal experts, technologists, and policy leaders. This is a summary of the highlights of the Workshop. The Workshop agenda and a full transcript is available on the DHS Privacy Office website: www.dhs.gov/privacy.

What is "commercial data" and how is it used?

The panelists began by defining terms. "Commercial data," as used at the Workshop, is data sold by companies that are often referred to as data providers, data aggregators, or information brokers. These data providers collect personal information about individuals from a wide variety of sources. The data ranges from directory information, such as individual names, addresses, and telephone numbers, to records of retail purchases, including travel, insurance, and financial data, and public record information obtained from federal, state, and local offices, including court documents, professional licenses, and property records. Although it may be a common belief that data providers have huge data bases full of information that they can tap for multiple purposes at any time, their general practice is to build specific information products for specific applications with specific uses.

The commercial data providers described four basic uses of commercial data: (1) insurance carriers use it to underwrite insurance policies for consumers; (2) businesses use it to screen their employees; (3) businesses also use it to authenticate existing or potential customers; and (4) governments use it for a wide variety of purposes, including to conduct law enforcement investigations, homeland security operations, and entitlement programs. The government's use of commercial data was the focus of the Workshop.

Why does the government use commercial data?

The panelists explained that the government uses commercial data for a wide variety of purposes, including identity verification, screening, fraud detection, research, and law enforcement. One of the leading uses of commercial data is to increase accuracy and integrity of government data. In some instances, commercial companies have done a better job than the government of organizing and cleaning data to eliminate errors. Commercial data is used extensively to assist in disaster relief and for risk-based analysis of threats and natural disasters. It is also used for public health services and for verification and implementation of government entitlement programs.

The most common use of commercial data by government involves one person accessing one commercial data record at a time and with a specific reason for doing so. Law enforcement, for example, is one place where commercial data is particularly important and is used proactively to augment information that has been retrieved from government databases.

What are some examples of the government's use of commercial data?

The government panelists offered some examples of how their agencies use commercial data. For instance, the Treasury Department uses commercial data to investigate terrorist financing and learn about a suspected terrorist's associations, movement of assets, and wealth. Unexplained wealth was suggested as a possible barometer to identify a target -- for example, where an individual has very limited economic means but flies first class. Also, associational relationships, asset transfers, and ownership of assets can be used when the agency has targeted an individual for investigation. Commercial data is used to augment the data obtained from the reports of financial institutions' pursuant to the Bank Secrecy Act.[1] The data includes reports of suspicious activity, large base currency transactions, and inbound and outbound currency movements to extrapolate criminal activity. Access to this data, however, is subject to protections provided by the Right to Financial Privacy Act.[2]

The FBI uses commercial data to assist in conducting threat assessment. Following 9/11, for example, they received thousands of phone calls reporting suspicious activities. Some calls were very general in nature, but agents were required to follow up on every report.

[1] The Financial Recordkeeping and Reporting of Currency and Foreign Transactions Act of 1970 (31 USC 5311 et seq.) is referred to as the Bank Secrecy Act (BSA). The purpose of the BSA is to require United States financial institutions to maintain appropriate records and file certain reports involving currency transactions and customer relationships as a tool to fight drug trafficking, money laundering, and other crimes. Congress enacted the BSA to prevent banks and other financial service providers from being used as intermediaries for, or to hide the transfer of, money derived from criminal activity.

[2] The Right to Financial Privacy Act (12 U.S.C. 3401 et seq.) provides several different methods that the government can use to obtain financial and credit card records. Several provisions require that the government provide notice to the customer whose records are being sought, and others permit government access only with an administrative subpoena, search warrant, or judicial subpoena. The Act also allows the government to obtain records without notice in certain limited circumstances.

The FBI checked commercial databases as part of their investigations. Panelists described several examples where commercial data could have been used to assist 9/11 investigators, saving time and money, but was not used. In one instance, the FBI spent millions of dollars to send out agents from every office in the country to identify certified scuba training schools after receiving information about the possibility that terrorists were engaged in scuba diving training for the purpose of conducting underwater bombing. In retrospect, the FBI could have purchased that data for much less from a commercial data provider. Another instance cited involved foreign travelers visiting the United States. All foreign travelers must complete a form I-94, which states where the traveler is staying during his or her visit. Post 9/11, the government learned that none of the hijackers told the truth on their form I-94 regarding where they were staying. A panelist suggested that if the government had used commercial data to check the hotel addresses on I-94 forms, it would have noticed that there was no hotel under the name or address the hijackers had listed, raising questions about the veracity of the travelers.

In addition to law enforcement, the government also uses commercial data to assist in disaster relief. During the initial reaction to the Hurricane Katrina disaster, commercial data providers were called upon to supply visualization and mapping technology support to assist in the establishment of relief shelters, as well as to catalog infrastructure damage. Commercial data was also used to identify every doctor, nurse and pharmacy to assist in disaster relief and to validate the identity and status of victims that applied for assistance either online or through call centers. It was also used to supply vital records for the people who lost their homes and to provide screening for the volunteer organizations to ensure that the new volunteers were legitimate and not likely to engage in fraudulent activities.

What are the benefits of commercial data?

The commercial data providers stated that until the government does a better job of sharing and authenticating its own data, it is often easier and more accurate to get data from commercial sources. They identified four major benefits to using commercial data:

1. It saves time. What might take a government entity weeks to acquire by sending resources to investigate subjects, can be done in a matter of minutes by using existing commercial data.

2. It is better quality. Commercial data is more accurate and can be more precise than government data.

3. It is more current. Commercial data is updated more quickly and therefore is more accurate.

4. It can protect and minimize the impingement on civil rights, civil liberties, and privacy because you are not doing broad-based searches that involve hundreds of people – it enables the search to hone in quickly on just the correct individual.

What security safeguards do commercial data providers have in place to protect the data?

According to the data providers, they have policies, practices, and procedures in place with which they, as well as their customers, both private and government, must comply. They stated that have imposed strict contract terms and conditions on their customers in recognition of the privacy concerns raised by the abundance of commercial data and the ease with which it can be obtained.

The data providers reported that they have instituted contractual terms and conditions to implement the legal requirements of the Fair Credit Reporting Act, the Gramm-Leach-Bliley Act, and the Driver's Privacy Protection Act governing state motor vehicle records. Second, industry guidelines or practices have been adopted around the use of certain kinds of information, and commercial data providers have developed their own policies to provide further protections. Contractors also have to adhere to rules that were placed on the data by the original source.

The commercial data providers described some of the contractual safeguards they have implemented to protect data. For example, the person who enters into the agreement with the commercial data provider must be the actual end user of the data and the end user must protect against misuse of the information. The contract requires that the user of the data maintain it properly and take specific steps in response to a breach. Finally, contract terms include audit provisions, which provide that the government or any other client must agree to allow the commercial data providers to audit whether the data is being used for permissible purposes. The data providers did not describe how often such audits are conducted.

Does the Privacy Act of 1974 apply to commercial data and if so, does it provide adequate protections?

The second panel of the Workshop discussed the adequacy of the Privacy Act to address government use of commercial data. The Privacy Act provides the fundamental privacy protections that govern personal information held by the federal government. It establishes certain rules for the collection, maintenance, use, and disclosure of personal information by federal executive branch agencies. It requires notice to and consent from individuals when the government collects and shares information about them. It also gives citizens and legal permanent residents the right to see information the government has about them in certain databases, requires agencies to maintain an accounting of disclosures of personal information, and holds government databases to certain accuracy standards. Most law enforcement, intelligence, and homeland security databases, however, are exempt from a number of Privacy Act requirements, including access.

The Privacy Act was promulgated in 1974, at a time when America was experiencing great distrust of the federal government due to recent historical events and the awareness that computers were quickly revolutionizing the way information was being used and stored. As a result, the Privacy Act reflected a consensus that some restraints were needed on the government's information collection and that the government should disclose its information practices both to individuals and to the public.

The Workshop panelists, however, identified a number of shortcomings and limitations of the current implementation of the Privacy Act. When the Privacy Act was enacted, data mining was very different from what it is today. A number of panelists expressed concern that the Act, by failing to keep up with the changes in information practices and technology, is, in effect, "broken." One panelist said the Act as interpreted has become more about disclosure than about privacy and that there is a lack of substantive standards for determining when a government agency should be authorized to collect information from whatever source and use if for homeland security purposes. This is critical, according to the panelist, because it is at the collection point that most privacy interests are at stake.

In general, the Privacy Act calls for the individual to consent to the disclosure of personal information held in a System of Records;[3] however, this fundamental principle is subject to a number of general and specific exemptions. Although many of the exceptions in the Privacy Act are permissive and not mandatory, the panelists argued that law enforcement and intelligence agencies too often exempt their own records from various provisions of the Privacy Act.6[4] In addition, an agency can share its records with any other agency if the sharing is a "routine use" and has been announced in the Federal Register.[5] A "routine use" is defined in the Privacy Act as any use that is compatible with the purpose for which the information was collected. Several panelists observed that the routine uses for a database are often stated so broadly that its uses can gradually increase until its scope becomes far greater than its originally stated goals. This amounts to "mission creep." Panelists urged that new life could be given to the Act if agencies stopped abusing these exemptions.

The panelists also noted that the Privacy Act currently provides little restraint on data mining of commercial databases because its protections apply only where the government is creating a "system of records" and not to the use of private sector databases. Much of the government's use of commercial data does not require that the data leave the hands of the data providers. Instead, government agencies contract with data providers to scan or query their data without pulling the data directly into a government database. If the government is simply accessing external databases created by third parties for their own reasons, some would argue that the government is not creating a system of records subject to Privacy Act requirements.

[3] The Privacy Act defines the term "system of records" to mean "a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol or other identifying particular assigned to the individual." 5 U.S.C. § 552a (a)(5).

[4] There are two general exemptions under the Privacy Act. The first permits the Central Intelligence Agency to exempt its records from certain subsections of the Act. 5 U.S.C. § 552a(j)(l). The second applies to selected records maintained by an agency or component whose principal function is any activity pertaining to criminal law enforcement. 5 U.S.C. § 552a(j)(2). In addition, there are seven specific Privacy Act exemptions that can be applied to systems of records. 5 U.S.C. § 552a (k).

[5] 5 U.S.C. § 552a (a)(7).

The Privacy Act does include a provision that extends its coverage to databases created by government contractors, but the panelists questioned whether agencies are interpreting the Privacy Act to cover instances where contractors are providing access to their own pre-existing databases. Subsection (m) of the Privacy Act states:

> "When an agency provides by a contract for the operation by or on behalf of the agency of a system of records to accomplish an agency function, the agency shall, consistent with its authority, cause the requirements of [the Privacy Act] to be applied to such system."

Some of the panelists stated that Subsection (m) was intended to prevent government agencies from avoiding the Privacy Act by outsourcing their systems of records to private contractors.

The panelists recommended a number of ways in which to work within the Privacy Act to better address privacy protections. First, agencies need to be better educated about the Fair Information Practice Principles so that they do not exempt themselves from the Act or simply think that providing notice in the Federal Register is sufficient. Second, agencies and contractors should be assessed penalties when they do not comply or seek to do an end run around the Act. Some also recommended legislative fixes for the Privacy Act, including government regulation of data providers and liability for those that fail to provide accurate data, secure it, or audit it.

In addition to discussing the Privacy Act, panelists stressed the value of applying the Fair Information Practice Principles to the government's collection and use of commercial data. These principles, which underlie the Privacy Act and numerous international privacy frameworks, call for protecting privacy by providing transparency, accountability, and redress regarding the collection, use, disclosure, access, retention, integrity, and security of personal information, whether the data is obtained by the government or the private sector. Several panelists noted that perhaps too much emphasis has been put on the notice element, rather than on the other principles, but one panelist said notice is particularly important given that it is impossible to begin to enforce the Privacy Act where the data usage is done in secret. The panelists agreed that the Privacy Act should be fully implemented to protect individual privacy.

Panelists also discussed the implementation of the E-Government requirement that agencies conduct a Privacy Impact Assessment (PIA) when introducing new technologies. A number of specific concerns were raised regarding the implementation of the PIA requirement. First, panelists suggested that too often some agencies leave the "A" out of the PIA and simply engage in checking off boxes rather than conducting a rigorous risk evaluation. Some agencies see it as just another administrative requirement and only describe what they are doing rather than engaging in any real analysis of the privacy risks, alternatives, and practical means to mitigate the risks. Second, PIAs for national security purposes do not need to be made public. Panelists suggested that some agencies take advantage of the exemption in the Act to avoid doing the assessment even for internal purposes. A panelist commended DHS and the U.S. Postal Service for their excellent work in issuing PIAs, but recommended that OMB provide more guidance for agencies to improve the quality of PIAs.

How can technology help government analyze commercial data to help protect homeland security?

The technology panels addressed both data mining and data analysis and discussed what technology can and cannot do to help protect homeland security. Data analysis was described by the panelists as less sophisticated than data mining and requiring less technology. The good news, according to the technologists, is that there is a lot of technology out there to do basic data analysis where data, including commercial data, is pulled from so many data sources to an end user's desktop with a minimum of training. Data mining, on the other hand, is a more sophisticated type of analysis.

Panelists described two types of data analysis – subject-based searches and pattern-based data mining. In subject based searches, a suspect has been identified and data is searched and analyzed to find out more about that suspect. It provides for efficiency in investigations for law enforcement. Pattern-based data mining works by gathering everyone's information and then seeing what patterns emerge to create a profile that can then be applied to data about others. Since the public are not allowed to view the resulting profiles or lists, panelists felt that this type of data mining could be problematic as questions of accuracy and redress could not be addressed. They noted that there was a great potential for false positives.

Panelists also acknowledged that some forms of data mining may raise constitutional questions. They cited the difference between a retailer that conducts data mining for marketing purposes and the government's use of data mining to determine whether a person is placed on a terrorist list or is prohibited from flying on a plane. Clearly the potential consequences associated with government data mining are much more significant than private sector uses. It was recommended that whenever an agency wishes to conduct data mining, including the use of commercial data, there should be a discussion about the purpose of the data mining and whether it is cost effective and constitutionally and legally appropriate.

One panelist said that, at its core, information technology is a "force-multiplier" to our ability to make meaning out of raw data. Data that would otherwise seem random and disconnected can be brought together and made meaningful by the use of technology, but getting from raw to meaningful data requires taking a few steps. Step one is to make sure that you have a clean set of data to work with. Step two is to look at the most basic level of analysis to try to refine the hypotheses. The third step is to try to limit the number of variables, although not artificially, because the more variables you have the more difficult it becomes to solve a problem.

The technology panelists agreed that choosing the right technology for the right purpose with the right discrete functionality is crucial. The panelists described two different approaches to developing information technology -- one is proactive and the other reactive. When being proactive, you are designing information technology to help you anticipate some undesirable event and to prevent it from occurring. When you are being reactive, it is because your proactive process did not successfully predict what was going to happen. When operating in the reactive mode, you have fewer choices and have to use technology to engineer a solution to a problem that was not foreseen, forcing you to spend a lot more time and money in the process.

The technology panelists stressed that good information technology requires good data and that can be context-dependent. Some tools are available to try to overcome bad data; but technology cannot replace or cure faulty data. The cliché -- garbage in, garbage out -- applies here, so it is important to spend time with the end user to understand what data to collect, how it is being used, and what restrictions there may be on the data. That said, the panelists agreed that sometimes requiring that data matches be 100% accurate could result in false negatives, causing some people who should be identified as a suspect to be missed. Ultimately, good data analysis requires the addition of the human element to evaluate the results and determine its meaning and importance.

How can technology help protect individual privacy while enabling government agencies to analyze data?

The technology panelists observed that traditional methods for protecting privacy may be inadequate in today’s world, but new techniques such as biometrics, encryption, and anonymization offer great potential. Encrypting or anonymizing data before you analyze it enables you essentially to engage in knowledge discovery without disclosure of personal information. You are able to determine the key identity attributes necessary for knowledge discovery without revealing the identity of the subject of the inquiry.

A panelist gave the following example of this technique. A cruise line has a passenger manifest that the government would like to match against a watch list of known and suspected terrorists. The cruise line may not be inclined to share its manifest with the government, and the government may not want to share its terrorist list with the cruise line. How can the government protect national security while not making the corporation give up a valuable asset and anger its customers? One alternative is to anonymize the data -- take the names and identifying information and apply a one-way hash to the information, which is essentially a digital signature, thereby making the data indecipherable in the reverse. If both the terrorist list and the passenger list are anonymized using the same hash, then they can be compared against each other without compromising the privacy of those on the passenger list. This was described as a “surgical approach” to matching data that promotes information sharing in a privacy-enhanced manner.

In addition to examining privacy-enhancing technologies to analyze data, technologists are also working on technology tools to do identity management and to govern how data is accessed, used, and shared. The technologists expressed concern that currently no strong, consistent rules exist across the federal government on how to model metadata and, in particular, identity metadata. The federal government is currently working to develop a Federal Enterprise Architecture Data Reference Model to provide consistency and standardization for describing, categorizing, and sharing data to support government programs. A common data model will streamline information exchange processes within the federal government and between government and external stakeholders. The panelists noted, however, that this standardization must take privacy into consideration.

The government must first take a disciplined approach to information modeling and identify the privacy ramifications associated with a particular model or program. Business rules can be set within the data system to perform some of the necessary checks and audits. One panelist described the need to create formal privacy decision trees and rules to apply to data across the federal government. A privacy decision tree, for example, could be developed for the government's use of commercial data. Ultimately agency information systems will have policy engines that validate that data is being released appropriately. The technologists cautioned, however, that technology is not a panacea, but rather just one way of helping to solve part of a problem. In the absence of business rules, processes, controls, oversight, and audits, technology could make a problem even worse. Although technology has made it more difficult to protect individual privacy by making is so easy to collect and share personal information, technology may also offer the ability to protect privacy by using such tools as encryption, anonymization, and metadata to achieve more security and more privacy rather than viewing them as tradeoffs.

Roadmap for the Future

The final panel of the Workshop sought to construct a roadmap for DHS regarding its use of commercial data. Like the earlier legal panel, it began by discussing the importance of the Privacy Act and the Fair Information Practice Principles described above. The panelists pointed to these principles as the overarching framework for the government's use of personal information, but they shared the same concerns expressed by the earlier panel about the adequacy of the Privacy Act to address current information technologies and the government's collection and use of commercial data. Some of the panelists noted the particular challenge of applying them to the Department's homeland security mission, especially the need to reexamine whether the traditional principles of notice and choice even apply in this context. Others noted that often these principles are too broad and too vague to give meaningful guidance in particular situations.

Putting aside whether the Privacy Act provides sufficient guidance and protection in addressing commercial data, some suggested that OMB and individual agencies, including DHS, could provide for greater privacy protections in the use of commercial data simply by adding contractual protections when purchasing commercial data. Others suggested that DHS and OMB should require immutable audits and adoption of privacy-enhancing technologies.

In addition to using technology to provide greater privacy protections, some panelists urged DHS and other agencies to rethink their use of the Privacy Act exemptions and to understand that the concept of privacy is not at odds with security and achieving the mission. In fact, if an agency walks through the questions posed by the Act and PIAs -- what are you collecting, for what purpose, and who are you going to share it with -- you are asking both operational and privacy questions. So the roadmap for privacy can also be the roadmap for better information management. Too many information programs have failed because they did not answer these questions and clearly define their mission.

The first critical step in any roadmap for the Department's use of commercial data is to know why the data is needed. Only then can the agency begin to answer the questions that the Fair Information Practice Principles ask – is it necessary, is this the least amount of data necessary, is it being used for a limited purpose, and is it retained for no longer than necessary? Defining the purpose at the beginning can help avoid "mission creep," where the data is gathered for one purpose but later used for another. Mission creep ultimately leads to public distrust and often the demise of a program.

The panelists then proceeded to discuss the various privacy principles and their application to using commercial data for data mining to protect the homeland. Some panelists questioned the accuracy of commercial data, concerned that inaccuracies could result in programs that wrongfully identify individuals and produce too many false positives. A related data integrity issue was whether the data is kept current and whether data retention policies are appropriate. One panelist suggested that the time had come for commercial data brokers and federal agencies to be required to give individuals an audit trail of who purchases or obtains their data.

A critical question in building a roadmap for DHS is to first ask how DHS will use the commercial data. If the results are input for further analysis and investigation, then the risks of harm are less. If, hypothetically, DHS were to use the results as a sole or primary basis for depriving someone of their liberty – arrest or deportation, then DHS should consider such serious risks in deciding how to proceed. Another critical question is what protections can be put in place, such as access controls, anonymization of the data where possible, and audits to help govern the use of the data. One panelist pointed to protections provided under the Fair Credit Reporting Act (FCRA), which governs when private sector decisions are based on inferences involving credit reports. Under the FCRA, a bank or insurer must tell individuals when an adverse inference is made about them and give the individual an opportunity to challenge or correct inaccurate or incomplete information. In addition to the FCRA model, panelists urged that the government should establish a redress process to enable individuals to challenge adverse inferences made using government or commercial data.

The panelists agreed that a roadmap for using commercial data begins by defining the purpose and then asking a series of questions: What is the purpose for using the commercial data; What is the information that is needed; Is it accurate enough for the purpose; Is it relevant to the purpose; How are you going to use the information; What kind of analysis are you doing; How are you going to use the results; Is the level of false positives and negatives sufficiently low; Did you test the system before running it on millions of records to see if it yields acceptable results? What privacy protections do you have in place? The panelists urged going beyond implementing the traditional privacy principles to using technology to incorporate formal business rules through metatags on the data itself and requiring “immutable audits,” which make audit trails unalterable, as key privacy protections using new technologies.

Finally, the panelists recommended that the roadmap for DHS require better auditing and enforcement of the existing rules, including the Privacy and E-Government Acts. If there are no consequences for violating these acts, then there is little incentive to obey the rules. Rather than waiting for legislative actions to strengthen the Privacy Act’s governance over commercial data, panelists urged DHS to write privacy guidance, in the spirit of the Privacy Act, to protect privacy when using commercial data to carry out the homeland security mission.