Legal Memorandum

Published by The Heritage Foundation

No. 11
May 26, 2004

Technologies That Can Protect Privacy as Information Is Shared to Combat Terrorism

James X. Dempsey and Paul Rosenzweig

Information technology has much to offer in achieving the compelling national goal of preventing terrorism. At the same time, government access to and use of personal information raises concerns about the protection of privacy and due process. Given the limited applicability of current privacy laws to the modern digital data environment, resolving this conflict will require the adoption of new policies for collection and access, use, disclosure and retention of information, and for redress and oversight.

But technology itself should also be part of the solution. The same technology that permits the accumulation, sharing, and analysis of huge databases also allows for the incorporation into information sharing systems of features that protect information from abuse or misuse. In this paper, we focus on three technologies that hold great promise: anonymization of data, permissioning rules built into the data and search engines to regulate access, and immutable audit trails that can identify abuse (while also assisting in linking people into ad hoc collaborative teams).

Too much of the debate over the role of information technology in counterterrorism—particularly around the role of commercial databases—has consisted of unsubstantiated claims of utility or non-specific fears of abuse. In order to make progress in improving the nation's response to terrorism and preserving civil liberties, it is necessary to assess carefully what uses of information technologies will be effective in fighting terrorism, what their impact on civil liberties will be, and how any adverse impacts can be avoided. To

Talking Points

- Government access to and use of personal information raises concerns about the protection of privacy and due process as information technology is used to combat terrorism.
- Resolving this conflict will require the adoption of new policies for collection and access, use, disclosure and retention of information, and for redress and oversight.
- The same technology that permits the accumulation, sharing, and analysis of huge databases also allows for features that protect information from abuse or misuse.
- Among the promising technologies are: anonymization of data, permissioning rules built into the data and search engines to regulate access, and immutable audit trails that can identify abuse.

This paper, in its entirety, can be found at:
www.heritage.org/research/homelanddefense/lm11.cfm

Produced by the Center for Legal and Judicial Studies

Published by The Heritage Foundation
214 Massachusetts Ave., NE
Washington, DC 20002–4999
(202) 546-4400 heritage.org

Nothing written here is to be construed as legal advice on any matter, as an attempt to create an attorney-client relationship, or as an attempt to aid or hinder the passage of any bill before Congress.

achieve this, we need dialogue and consultation between those designing the technologies and those framing the policies for their use. We also need to examine the information practices of the private sector, which has experience operating under constraints on the use of information that are in some respects more stringent than those applicable to law enforcement and intelligence agencies.[1] And we need to draw on academic research to identify emerging technologies that can overcome privacy concerns.

In considering whether to implement new technologies that share, analyze, and correlate disparate data (sometimes called "data mining" or "knowledge discovery") in aid of efforts to prevent terrorism, are there ways to design and implement the technology to affirmatively protect privacy? In particular, we will examine three solutions that have been put forth: (1) anonymization techniques that allow data to be usefully shared or searched without disclosing identity; (2) permissioning systems that build privacy rules and authorization standards into databases and search engines; and (3) immutable audit trails that will make it possible to identify misuse or inappropriate access to or disclosure of sensitive data. We think that each has promise and that further careful research regarding them might contribute to the development of tools for enhanced information analysis that simultaneously protect individual privacy. In other words, security and privacy need not be traded off. The current crisis facing America might not require a zero-sum response.

The Heritage Foundation and the Center for Democracy and Technology have been pursuing these issues independently and jointly, conducting a series of consultations with industry, government, academic researchers, and privacy advocates. [2] Our inquiries overlap with and benefit from other efforts. Among these, by far the most in-depth is the work of the Markle Foundation Task Force on National Security in the Information Age.[3] The Center for Strategic and International Studies has conducted a very useful series of seminars from which we have drawn insight.[4]

The Interests at Stake—More Than Anonymity

At the outset it is useful to describe exactly what are the concerns posed by government use of personally identifiable information. The word "privacy" as commonly used in debates about information sharing and analysis is somewhat misleading. In the context of personally identifiable information, privacy is not just about keeping information hidden or secret. Rather, the modern concept of privacy as protected under U.S. civil law extends to information that an individual has disclosed to another in the course of a commercial or governmental transaction and even to data that are publicly available. In this sense, privacy is about notice, fairness, and consequences rather than about what is withheld or hidden.

Data privacy is based on the premise that individuals should retain some control over the use of information about themselves and should be able to manage the consequences of others' use of that information. A set of commonly accepted "fair information practices" captures this conception of privacy and is reflected, albeit in piecemeal fashion, in various U.S. privacy laws and in the practices of commercial entities and government agencies. Under these rules, data

[1] *See* James X. Dempsey, "Privacy's Gap: The Largely Non-Existent Legal Framework for Government Mining of Commercial Data" (May 19, 2003), available at *http://www.cdt.org/security/usapatriot/030528cdt.pdf*. *See also* "Matrices of Laws Governing Governmental and Commercial Access to Privately Held Data," accompanying the report of the Markle Task Force on National Security in the Information Age, *http://www.markletaskforce.org/privacyrules.html*.

[2] This paper in particular draws on a joint Center for Democracy & Technology–Heritage Foundation consultation on December 1, 2003, during which Jeff Jonas (SRD), Rebecca Wright (Stevens Institute of Technology), Stewart Baker (Steptoe & Johnson), and Doug Tygar (University of California at Berkeley) gave presentations on privacy-protecting technologies and engaged in dialogue with a diverse group of participants from government, the private sector, and the public interest community. Any errors in this paper, whether technical or non-technical, remain, of course, our own.

[3] "Protecting America's Freedom in the Information Age" (October 2002) and "Creating a Trusted Information Network for Homeland Security" (December 2003), available at *http://www.markle.org*.

[4] *See* Mary DeRosa, "Data Mining and Data Analysis for Counterterrorism," Center for Strategic and International Studies (March 2004).

that are sold or exchanged commercially or that are "publicly available" are nevertheless subject to constraints intended to protect the data subject.[5]

Recognizing that "privacy" involves values in addition to confidentiality, what then are the concerns when the government uses personally identifiable information? Documented problems fall into five categories:

- **Unintentional mistake—mistaken identity.** When personally identifiable information is used to make judgments about people, a person sometimes will be misidentified as a criminal or a suspected terrorist or a risk when in fact he is innocent but shares some identifiers with someone who is of interest to the government. This is not a hypothetical matter; it is a problem that affects, for example, the airline passenger screening system.[6]
- **Unintentional mistake—faulty inference.** In analyzing for counterterrorism purposes information that was generated or collected for non-terrorism purposes, government employees occasionally misinterpret the information and draw the erroneous inference that someone may be connected with terrorism.[7]
- **Intentional abuse.** Government employees have used authorized access to personal information for unauthorized purposes. For example, an employee of an intelligence or law enforcement agency may perform checks for a fee for private investigators, transferring the information to an unauthorized and possibly uncontrolled use.[8]
- **Security breach.** Through poor security practices, information in the hands of government agencies or their contactors has been stolen or carelessly disclosed. This is also not a hypothetical problem.[9]
- **Mission creep.** Government information systems justified in the name of fighting international terrorism may be turned to more ordinary

[5] Medical records are shared among numerous participants in the health care and insurance system, but they are protected by privacy rules that prohibit use of medical information for purposes unrelated to health care. Arrest records are publicly available governmental records, but under anti-discrimination laws they cannot be used for employment purposes unless they include disposition data. Under the Driver's Privacy Protection Act, driver's license data can be obtained for some purposes and not for others. Bankruptcy records are publicly available, but under the Fair Credit Reporting Act (FCRA) they cannot be included in credit reports if they are more than 10 years old. The FCRA imposes a number of data quality requirements on commercial compilations of publicly available data. Under the Act, individuals are legally entitled to access their credit reports and insist upon corrections, even though none of the data in the reports are confidential and some are publicly available.

[6] Ann Davis, "Why a 'No Fly List' Aimed at Terrorists Delays Others," Wall Street Journal (April 22, 2003).

[7] In the frenzied aftermath of 9/11, when the government was looking for associates of the hijackers, it arrested a man who had obtained a driver's license at the same motor vehicle office and within minutes of one of the hijackers obtaining a license. The man was correctly identified, but the inference drawn from his Arab name and proximity in time to one of the hijackers was wrong, albeit in good faith. Tamara Lytle and Jim Leusner, "The Price of Protection: Push for Safety Clouds Individual Rights," Orlando Sentinel, A1 (Aug. 29, 2002).

[8] This is quite a persistent problem. In December 2002, a former DEA agent was convicted of selling to a private investigation firm criminal history and law enforcement data he obtained from law enforcement computer systems while an agent. *Privacy Times*, Vol. 23, No. 1 (January 2, 2003), at 10. In another documented case, between 1994 and 2000, an officer with the Los Angeles Police Department searched the police department's computers for information on celebrities and his ex-girlfriend. He used the information on his ex-girlfriend to stalk her and he allegedly sold the information on celebrities to tabloids. *Privacy Times*, Vol. 23, No. 8 (April 15, 2003), at 5. In May 2002, two FBI agents were indicted on fraud charges for allegedly accessing FBI databases to provide information on companies to manipulate stocks. One of the agents used information from the National Crime Information Center (NCIC) database to discredit a company executive and lower stock prices. The agents also used confidential records from FBI databases to monitor government investigations of other stock manipulators. From the Electronic Privacy Information Center (EPIC), *http://www.epic.org/privacy/publicrecords/*. An FBI agent in Las Vegas working with a member of the state Attorney General's office was accused of selling information from the FBI's NCIC database to organized crime leaders. EPIC, *http://www.epic.org/privacy/publicrecords/*.

[9] In December 2002, the personal health care information and Social Security numbers of more than 500,000 military personnel, retirees, and family members were stolen from the Phoenix office of TriWest Healthcare Alliance, a contractor administering the Department of Defense's health plan. *Privacy Times*, Vol. 23, No. 1 (January 2, 2003), at 2.

> criminal justice or administrative purposes, further expanding the government's control over individuals. Already, we have seen mission creep at the Transportation Security Administration's passenger screening system (CAPPS II).[10]

All of these dangers are inherent in police and intelligence work, but it is necessary to give them extra attention in light of the power of computer technology. The potential today for abuse far exceeds anything that was possible in the era of paper records.

A sixth area of concern may seem more ephemeral: the risk that citizens will feel under generalized surveillance, thus diminishing their trust in government and inhibiting their participation in lawful activities, whether it be taking firearms training or enrolling in pilot school, if they feel the result would be adverse inferences being drawn against them. [11]

Of these potential abuses and problems, the technologies we discuss in this paper can mitigate four: intentional abuse, security breach, mission creep, and the uneasiness about data aggregation. The errors that flow from unintentional mistake require different responses that we mention briefly: improving data quality (especially of government watch lists), redress mechanisms for those wrongly suspected of involvement in terrorism, and oversight.

The Threshold Issue: Effectiveness

In considering the application of information technologies to counterterrorism, efficacy should be a threshold issue. If it cannot be shown that a particular use of commercial databases will yield improvements in national security, then the application should not be deployed and there should be no need to reach the civil liberties questions.[12]

It is beyond the scope of this paper to examine the questions of effectiveness. In general, however, it seems indubitable that there are uses of commercial data, and combinations of government and commercial data, that could aid in the fight against terrorism. Using commercially compiled data to quickly determine where a suspect may be residing, for example, seems clearly effective. The value of others, including some of the pattern-based searches sometimes referred to as "data mining," remains speculative and unproven.

While the effectiveness of specific applications is being assessed, it is essential to simultaneously consider the privacy issues associated with counterterrorism uses of data before any implementations go forward. If privacy is taken into account in the research and development phase, protections can be built into the design of any applications that are shown to be useful.

Context: The Uses of Commercial and Governmental Databases

We focus here on the uses for counterterrorism purposes of data collected by commercial entities or government agencies for purposes other than counterterrorism. It is important to put this focus in context. As agencies research, and policymakers debate, the use of commercial data to prevent terrorism,

[10] James Carafano, Paul Rosenzweig & Ha Nguyen, "Passenger Screening Program is Vital—and Vital to Get Right," *Web Memo* No. 428 (The Heritage Foundation, February 2003). The dangers of mission creep are significant. *See* Paul Rosenzweig, "Can the Use of Factual Data Analysis Strengthen National Security?" Testimony Before the United States House of Representatives Committee on Government Reform Subcommittee on Technology Information Policy, Intergovernmental Relations, and the Census (May 20, 2003); Paul Rosenzweig & Michael Scardaville, "The Need to Protect Civil Liberties While Combating Terrorism: Legal Principles and the Total Information Awareness Program," Heritage Foundation *Legal Memorandum* No. 6, Feb. 5, 2003, at pp. 10–11.

[11] The public is clearly concerned about the commingling of data. An uproar arose in February 2000 when DoubleClick announced plans to combine consumer information it collected from Web users with information collected about offline activity by a subsidiary. It is clear that while consumers might in some cases choose to disclose personal information, they do not want the information they disclose combined into massive dossiers. The Supreme Court has noted that there is a "distinction, in terms of personal privacy, between scattered disclosure of the bits of information . . . and revelation of the [information] as a whole." *Department of Justice v. Reporters Committee*, 489 U.S 780, 764 (1989). The court went on: "Plainly there is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information."

[12] Paul Rosenzweig, "Principles for Safeguarding Civil Liberties in an Age of Terrorism," *Executive Memorandum* No. 854, The Heritage Foundation (January 31, 2003); Paul Rosenzweig, "A Watchful America," *The Responsive Community,* Fall 2002, p. 89.

counterterrorism agencies still cannot readily access information in their own databases and cannot share with each other information specifically collected for counterterrorism purposes. The government only recently consolidated watch list information through the Terrorist Screening Center, but the process is far from complete and a significant percentage of that information almost certainly is still incomplete or inaccurate.[13] Efforts to make better use of information specifically collected for counterterrorism purposes are every bit as important as—if not more important than—accessing commercial data.

More important, some uses of non-terrorism information are more clearly useful and pose fewer privacy concerns than others. For example, the government may use commercial and governmental databases to learn more about individuals it has reason to believe are terrorists. These suspected terrorists may have been identified, for example, based on an informant's tip or from a list seized from the apartment of a known terrorist. Once it has particularized suspicion through traditional investigative means, the government will then seek to learn more about these people: where do they live, where do they bank, what communications facilities do they use, what property do they own, what licenses do they have? The government's specific interest may be in quickly locating a particular suspected terrorist. Commercial databases can answer these questions, and such uses are to be fostered. It is now widely recognized that recourse to commercial databases and government databases of information compiled for reasons other than counterterrorism would have located two of the 9/11 hijackers when the government was desperately seeking them.[14]

Commercial databases can also help improve the amount and quality of identifying information in watch lists. Too many watch list entries have limited utility because they contain minimal identifying information; other databases may provide additional information such as date of birth, an address, or some other identifier that can enhance the fidelity of the watch list and make identification possible or more accurate.[15] Commercial databases also can help when intelligence information suggests that terrorists may be planning a particular mode of attack. Inquiries may include, for example, checking lists of airline pilots or lists of those who have taken scuba diving lessons, to identify persons on terrorist watch lists. Another simple illustration of this use of commercial data is the matching of watch lists with the names of airline passengers.

In all of the foregoing examples, the government is seeking to learn more about a person whom traditional investigative or intelligence methods have indicated might be a terrorist. An extension of these "subject-based" queries is when the government also seeks to find out who is associated with these known or suspected terrorists, in order to identify previously unknown persons who may be involved in terrorism. This involves questions like: Who else lives or has lived with a suspected terrorist? It has also been shown that many of the other 9/11 hijackers could have been identified based on their associations with the two known al-Qaeda members.[16]

[13] See Statement of Donna A. Bucella, Director, Terrorist Screening Center, before the House Judiciary Committee and the House Select Committee on Homeland Security, March 25, 2004; GAO, *Information Technology: Terrorist Watch Lists Should Be Consolidated To Promote Better Integration and Sharing*, GAO–03–322 (April 2003), available at *http://www.gao.gov/new.items.d03322.pdf*.

[14] *See Report of the Joint Inquiry Into the Terrorist Attacks of September 11, 2001*, House Permanent Select Committee on Intelligence and Senate Select Committee on Intelligence, 107th Cong., 2nd Sess., S. Rept. No. 107–351 and H. Rept. No. 107–792, Dec. 2002, additional views of Sen. Richard C. Shelby, at 43 (available at *http://www.fas.org/irp/congress/2002_rpt/911rept.pdf*); Markle Foundation Task Force, "Protecting America's Freedom in the Information Age" (October 2002), p. 28 (available at *http://www.markle.org*).

[15] If a watch list entry consists of name alone, it is not viable to use secondary sources to enhance the fidelity (e.g., enhancing the name with a date of birth, address, etc.). Viable watch list entries must include at least one other identifier to be useful (e.g., date of birth, passport, etc.) for the purpose of using secondary sources for enhancing fidelity.

[16] Markle Foundation Task Force, "Protecting America's Freedom in the Information Age" (October 2002), p. 28 (available at *http://www.markle.org*).

In all of these scenarios—finding out more about a known terrorist or identifying previously unknown terrorists by identifying non-obvious relationships between known terrorists and others—the government should be allowed, subject to appropriate standards, to make extensive use of databases of information collected for non-terrorism purposes.

A third use of information remains highly speculative: the use of pattern analysis. A pattern-based query is not focused on a specific, uniquely identifiable individual or individuals. Rather, using existing intelligence data, intelligence analysts will develop detailed models of potential terrorist activities. The models can be developed in an iterative process: First, a group of analysts intending to replicate the activities of potential terrorists may conduct operations in a virtual (i.e., artificial) world of cyberspace, creating data transactions (by securing driver's licenses, boarding airplanes, purchasing goods, etc.). They will repeat these operations for as many different terrorist scenarios as their imaginations will support. Then a separate team of analysts, using the same intelligence data and their own imaginations will try to develop database search inquiries that are capable of identifying the terrorist operation patterns created in virtual data space with a high degree of accuracy and distinguishing them from patterns of innocent activity. Thus, the utility of the effort will turn on its ability to accurately predict terrorist patterns or "signatures" and to sift those patterns from the ocean of transactional information about innocent conduct. In this paper we offer no conclusions about the efficacy of such searches or the legal framework that should apply to them.[17]

Current Data Sharing Obstacles

Currently, the question of government access to commercial databases poses a seemingly intractable dilemma: The government is reluctant to query commercial databases because, in the course of doing so, it identifies to the commercial data aggregator the individuals of interest. Employees of commercial vendors could "mine" the government's queries and determine whom the government is investigating, possibly even tipping off targets or otherwise jeopardizing investigations. Consequently, the government is not likely to share its most sensitive watch lists with corporate America

On the other hand, corporations are reluctant to send their databases to the government or give the government access. These databases contain information on the lawful transactions of innocent people. For one, the knowledge that a corporation is sharing customer information with the government is likely to generate criticism and could undermine customer trust.[18] Indeed, there is a risk that customers might start giving false information, degrading the quality of the commercial data. Second, the wholesale disclosure of databases to the government opens the door to intentional abuses, security breaches, and mission creep. For these reasons, various commentators have concluded that the private-sector data should remain in the hands of the private sector.[19]

Both prongs of the dilemma also affect government-to-government data sharing. The FBI would be reluctant to share its watch list with the Department of Education, for example, and the Department of Education would be reluctant to share its student database with the FBI. Furthermore, given the global nature of travel and commerce, it is necessary to share data transnationally in support of counterterrorism missions such as cargo container inspection and the screening of international air travelers to the United States. Efforts to share data internationally must comply with privacy rules more stringent in some ways than those of the United States.[20]

[17] Paul Rosenzweig has offered answers to those questions in "Proposals for Implementing the Terrorist Information Awareness System," Heritage Foundation *Legal Memorandum* No. 8 (August 7, 2003).

[18] For one example of this phenomenon, consider the public outcry when it became known that jetBlue had shared passenger information with the Defense Department. *See* DHS Privacy Office, "Report to the Public on Events Surrounding jetBlue Data Transfer" (Feb. 20, 2004), available at *http://www.dhs.gov/interweb/assetlibrary/PrivacyOffice_jetBlueFINAL.pdf*.

[19] *See* "Creating a Trusted Information Network for Homeland Security" (December 2003), available at *http://www.markle.org*.

[20] *See, e.g.*, Stewart Baker, *et al.*, "Anonymization, Data-Matching, and Privacy: A Case Study," (Dec. 2003) (examining EU privacy rules and their inter-relation with the US CAPPS II program), available at *http://www.9-11commission.gov/hearings/hearing6/witness_baker.pdf*.

Anonymized Data Analysis Offers a Solution to This Dilemma

Anonymizing technology is available that, if properly applied, would allow multiple data holders to collaborate to analyze information while protecting the privacy and security of the information. If both the privacy of personal information and the operational sensitivity of the information the government has on known or suspected terrorists can be assured, the reluctance to share data would be minimized. This would enable analysis of data from diverse sources, without requiring data to be gathered in a single place in a form that could be read or used for other purposes. This would limit abuses, including mission creep.

The most immediately promising technology is based on "one-way hashing," a well-established technique of modern cryptography that can scramble any piece of information into a representative digital signature. In simplest terms, hashing uses a mathematical formula (an algorithm, referred to as the "hash function") to scramble data. If two people apply the same algorithm to the same data, they will produce identical outputs, known as the hash value. For our purposes, the significance of the technology lies in the fact that the process, if properly designed and applied, can be irreversible: Someone who has the hash value and the algorithm still cannot unscramble the hash value and produce plaintext (the input value). Because two people applying the same algorithm to the same data produce identical hash values, two people can match their data without either one having to disclose his entire databases to the other. If two pieces of data produce the same hash value, that means the data are exactly the same; if there is no match, the data are not revealed.

Illustration

Take the situation where an agency has the following watch list entry:

Record #100031
Khalid Al-Midhar
Saudi Arabia
DOB: 07/12/76

When a one-way hash is applied to the personally identifiable data in this record, it produces the following:

Source:Agency #101
Record:#100031
Name:cbd034409c22929518fa494f99dc9964
Citizen:b835b521c29f399c78124c4b59341691
DOB: 799709b2e5f26f796078fd815bebf724

Meanwhile, airline reservation data could be hashed and compared with the hash values of the watch list entries. If any two hash values were the same, that would mean the original data were the same. If a number of hash values for two records matched, that would indicate that the two records related to the same person and then, with appropriate legal and procedural safeguards, the underlying plaintext data for just the two relevant persons could be shared between the airline and the government agency.

However, what if the airline reservation data used a slightly different spelling, and expressed some data elements in a different format and had different data fields:

#VX1RU9
Khaleed Al-midhar
San Francisco
DOB: 12/07/76[21]
ID: 33000102334

Given even small variations, the hash value produced would be different because the input value would be different. The hash process therefore needs to take into account the fact that there could be many variant spellings of names and variations in the way dates and addresses are expressed. 123 Main Street, 123 Main St., and 123 Main each will produce completely different hashes.

Actually, it is necessary to deal with variability whether data are hashed or not. Before matching of personally identifiable data can be effective, steps must be taken to ensure that Bob Jones and Robert Jones are matched but that Robert Jones Jr. and

[21] Note that in the hypothetical record from the government agency, the date was expressed in the form "7/12/76." Without the use of variants, the two dates would not match.

Robert Jones Sr. are not. The problem is difficult enough in English, but when names are transliterated from Arabic or another character set, it becomes even more difficult. The problem is not insoluble, however. Data analyzers in the private sector have developed sound techniques to prepare data for correlation—whether the data are anonymized or not. These techniques include both standardization and the creation of variants.

Data Standardization

Names standardization is a technique for converting all variants into a single standard that can be used for matching purposes. For example, Robert, Bob, and Bobby would all be standardized to Robert before being hashed:

“Robert”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“ROBERT”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“Rob”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“Bob”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“Bobby”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29

One technique is to use name standardization tables. Commercial companies specializing in analysis of personally identifiable information have compiled name standardization cross reference tables for tens of thousands of names in dozens of languages. The same standardization process must be applied to dates and addresses.

Variations

Standardization alone is not enough for many data sets. While Rob and Bobby can be standardized to Robert, what about a record in which the first name is listed as just “R”? For precise identity recognition or “entity resolution,” a record for Bobby Jones should be matched on the standardized basis of “Robert Jones” and as the variant “R Jones.” The same is true of dates. While “Dec.,” “December,” and “12” can all be standardized, it is not clear whether the date 07/12/76 refers to December 7, 1976 or July 12, 1976. Both variants will have to be used to avoid false negatives, and a 1976 variation may be introduced to account for records containing only a year (although this risks introducing more false positives).

07/12/76	07/12/76	799709b2e5f26f796078fd815bebf724
	12/07/76	8ceb0fe202b794c27694a83a5ad91df4
	1976	dd055f53a45702fe05e449c30ac80df9

These methodologies end up translating a single set of identifiers into a standard value and multiple options and then hashing each one, producing a full set of possibilities such as that in Table 1 (see below).

Techniques to Improve Security—Addressing “Dictionary Attacks”

Hashing is not entirely secure. If two people exchange hashed data, they will readily be able to identify overlaps—that is, A will be able to identify any items in his database that match those in B’s database. Moreover, A can create millions of data elements not in his database (common names, all 10 digit phone numbers, all 999,999,999 possible Social Security numbers, etc.) and compute a hash value from each of them. He can then compare all of the hash values he has received from B to see what matches. This is called a “dictionary attack” and exposes a potentially insecure aspect of the hash methodology. The dictionary attack can also be used by someone who has stolen a copy of the database, especially if the hashing program used is commercially available.

Table 1

Agency Name	Record # 100031	Subject Name	Standardized Value	cdb034409c22929518fa494f99dc9964
Agency Name	Record # 100031		Variation 1	9269bb3bc60366245144cdb5e960cfd8
Agency Name	Record # 100031		Variation 2	4ffe35db90d94c6041fb8ddf7b44df29
Agency Name	Record # 100031	Citizenship	Standardized Value	b835b521c29f399c78124c4b59341691
Agency Name	Record # 100031	Date of Birth	Standardized Value	799709b2e5f26f796078fd815bebf724
Agency Name	Record # 100031		Variation 1	40ddba83c22acc2acaddff12c66d7adf
Agency Name	Record # 100031		Variation 2	e4310b75f2fa9595f8154411924b19b1

To make the dictionary attack by a hacker more difficult, "salt" can be used. Salt is a random string of data that is concatenated with information before being operated on by the hash function. This is an additional input to the one-way hash algorithm. It would, for example, use the hash algorithm to compute the value not of "Robert" but of "Robert_And_The_Added_Salt." Using salt makes dictionary attacks practically impossible for someone who does not know the salt value, such as a hacker or other outsider, who would have to compute the hash values for all possible salt values.

However, salting alone does not prevent abuse by the original holders of the data if they know both the algorithm and the salt value. A user of the system who pushed all possible values into the system (using the agreed-upon salt value) would be capable of discerning all potential hash values.

One defense against this "Salt + Dictionary" attack would be to limit queries (e.g., the system only supports 500 queries a day). Another approach would be to transfer the anonymized data set to a third party who would not have access to the salt value and therefore could not do a dictionary attack. [22] Yet another technique that has been discussed is to use a secure coprocessor that can receive data and perform the hashing function (with salt) without disclosing one party's input to another party.[23]

How Would This Work?

Anonymization can be used in a context that focuses on subject-based queries meeting the "particularized suspicion" standard. Consider the real case in which the government was seeking to locate Khalid al-Midhar and Nawaq (sometimes Nawaf) al-Hamzi. The government would standardize the names of the suspected terrorists and any other identifying information it had, create variants, and then anonymize the data by a one-way hash. Hash values would be sent to a recipient party for direct use or to a third-party repository that would draw upon the best available commercial and government databases, which would provide similarly hashed data. Only if there was a match would information be returned to the government (perhaps pursuant to a separate legal authorization). The government agencies would never have to acquire or gain access to the entire commercial database.

This could be especially useful in the case of intelligence agencies that are barred from collecting information about "U.S. persons" (citizens and permanent resident aliens), a stricture that might currently limit such agencies' use of public records data held by commercial vendors, since most of the data in those systems pertain to U.S. persons. If these intelligence agencies could conduct searches for information about non-U.S. persons using anonymized data, they would never get access to U.S. person data and would never have to expose their searches to a commercial data aggregator. For example, because al-Midhar and al-Hamzi were known to have visas, intelligence analysts knew also that they were not U.S. persons. Therefore, if searches were conducted on anonymized data, any identifying information obtained about them would never have to include U.S. person data.

Anonymization would also apply to efforts to establish non-obvious relationships between known or suspected terrorists and others not previously known to the government.[24]

Policy questions remain to be resolved: What should be the standard for government use of com-

[22] One advantage of an independent third party is that if an anonymized dataset resides at a third party (who is unaware of the salt value), the third party would be unable to perform any personal queries. To do so would require collusion with one of the anonymized data contributors. By contrast, if the anonymized dataset resides locally at location A or B, unauthorized searches are more of a concern. Additional layers of protection could be added. For example, each party, after anonymizing its data with the agreed-upon salt value, could pass the anonymized data to a third party which could use its own salt value (which is also kept as a secret); then the twice anonymized data can be sent back to any party for anonymous data correlation.

[23] *See* Roberto J. Bayardo and Ramakrishnan Srikant, "Technological Solutions for Protecting Privacy," IBM Almaden Research Center, *http://www.computer.org/computer/homepage/0903/webtech/*.

[24] The technique may also be applicable to some narrow applications of the third, more problematic use of data—for pattern-based searches. Both CDT and the Heritage Foundation have expressed skepticism about pattern-based searches to varying degrees, depending on the context.

mercial data in anonymized form? If the data are otherwise "publicly available" or do not fall into the sensitive categories of medical and financial data, is it sufficient that there be a form of internal approval and accountability? Certainly, some higher standard of authorization should apply to the disclosure of original (non-anonymized) data should there be a match between commercial data and the subject of government interest. Security issues may determine who is the appropriate repository of the hashed data.

Perhaps the most important unresolved questions concern the quality of the government's watch lists, especially, what is the standard for watch-listing a person in the first place. There need to be rules for the creation of a record, including minimum thresholds and requirements on the listing agency to locally enhance the entry. (As noted above, other data analysis techniques can raise the fidelity of watch lists.)

Another major set of issues not addressed by anonymization concerns the due process and redress rules that protect a person against adverse consequences based on an honest but mistaken or careless match. It is also necessary to ensure that disclosure of personally identifiable data occurs only in accordance with a set of agreed upon policy and legal rules. There will, no doubt, be policy judgments that need to be made as to the best repository, but we leave those for another day. From a privacy perspective, the goal is not to prevent the most dedicated attacker, but to protect against routine misuse and mission creep.

Benefits of Anonymous Data Analysis

The benefits of anonymous data analysis are clear: No personally identifiable information or transactional data flows anywhere; source data are held and controlled exclusively by the data owner. [25] The risk of insider threat is greatly reduced as only anonymized values are in the database and the watch-listing-party receives only notice of matches. This approach has the value of fitting within the existing legal framework. Many of the relevant databases are available to the government for purchase. If different salt values are used for different purposes, anonymized data shared for one mission could not be conjoined with anonymized data collected for another.[26]

Immutable Audits

Anonymization will not prevent all abuses. For example, it is no protection against an authorized user running searches for unauthorized purposes. In order to deter such conduct, and to identify it and punish it when it occurs, audit trails must be established that log each query and its justification. As every computer criminal knows, the first (or last) place you go when you are doing something wrong is to the logs, to disable them or change them to cover your tracks. In the escalating battle for computer security, experts are developing immutable logging technologies.[27]

Immutable audits can record when data are accessed, by whom they are accessed and when they are changed. The audit records themselves can be protected from change, such that if the audit records are altered or erased, that fact will be readily available. A well-designed auditing process has four goals: (1) to ensure that everyone is subject to audit; (2) to produce cross-organizational audit; (3) to measure accuracy of auditors by cross-validation; and (4) to produce usage records that are tamper-evident.

Other auditing and logging techniques can foster collaboration among analysts by showing who is

[25] One could also envision a system that moved transactional data in cleartext but only associated it with anonymized identity data. One could then analyze or query the transactional data without any access to an identifiable "who." Unraveling the "who" would still require appropriate legal authority and transparency to the original data holder.

[26] There may be other anonymization techniques in addition to one-way hashing. Roberto J. Bayardo and Ramakrishnan Srikant of IBM have noted that techniques from the private information retrieval (PIR) domain may potentially apply to this particular problem. PIR techniques let authenticated users retrieve information from remote databases while preventing the database owner from identifying the specific information accessed. According to Bayardo and Srikant, significant work remains, however, to extend the current theoretical formulations of the problem to the real-world scenarios that arise on the Web. Bayardo and Srikant, "Technological Solutions for Protecting Privacy." See also Michael J. Freedman, Kobbi Nissim, and Benny Pinkas, "Efficient Private Matching and Set Intersection," to appear in *Advances in Cryptology — EUROCRYPT 2004*, May 2004, available at *http://www.scs.cs.nyu.edu/~mfreed/docs/FNP04-pm.pdf*.

[27] Doug Tygar presentation, December 1, 2003.

accessing data, alerting an analyst to the fact that another employee is interested in the same data, thus contributing to effective intelligence analysis.

Permissioning Systems

As the Markle Task Force has illustrated with its SHARE network, it is now possible to build privacy rules directly into databases and search engines. This development draws on technology developed for digital rights management and the expression of privacy preferences in Web browsing. Essentially, the English language rules for access can be translated into machine–readable form and attached to each database and each piece of information within a database. The privacy rules can differ based on the types of data or their contents. Databases can incorporate fundamental privacy principles. For example, the "purpose specification" principle states that the purposes for which information has been collected should be associated with any personal information stored in the database; the "limited use" principle states that the database will run only queries that are consistent with the purposes for which the information has been collected.[28]

Privacy Tool Bar

Privacy rules can be highly complex. In the past, this could inhibit information use. But there is no need to simplify privacy rules if systems can offer users real-time guidance through the maze. This is what Doug Tygar and others have referred to as the privacy toolbar. A well designed permissioning system would:

- Help users produce required documentation to support actions,
- Show privacy status of information,
- Highlight compliance requirements, and
- Support audit functions.

With such a system, an analyst seeking access to information would immediately be told the reason the system denied access, what further information or approval would be needed to change "no" to "yes," further information on what laws/rules apply to the situation; how the rules interact; and how to ask permission for more access. One system that does this is IBM's Enterprise Privacy Authorization Language, which encodes an enterprise's internal privacy-related data-handling policies and practices. EPAL allows privacy-enforcement systems such as IBM's Tivoli Privacy Manager to import and enforce the enterprise's privacy policy.[29]

To illustrate how databases can automatically enforce these principles, consider what happens when queries, tagged with purpose, are submitted to the database. The database first checks whether the user issuing the query is among the users authorized by the privacy policy for that purpose. Next, the database analyzes the query to check whether it accesses any fields not explicitly listed for the query's purpose in the privacy policy. Finally, the database ensures that only records having a purpose attribute that includes the query's purpose will be visible to the query, thereby enforcing any opt-in or opt-out preferences.[30]

Conclusions and Points for Further Research

Technology alone cannot address all the concerns surrounding a complex issue like privacy. The total solution must combine policy, law, and technology. However, there are technologies that can prevent some abuses, especially those associated with the transfer of entire databases. Anonymization makes it possible for the government to get the information relevant to its particularized suspicion without learning anything about the rest of the database.

A major, unresolved issue is the underlying quality of the watch lists. As a practical matter, watch list fidelity is one of the biggest challenges faced by those attempting to identify risks. If a watch list contains inaccurate or incomplete data, it will be very difficult to compare data against that list. In particular, name-only matches are meaningless; more information is necessary to determine whether an individual is, in fact, the person listed. In terms of the government's use of data, this sug-

[28] Bayardo and Srikant, "Technological Solutions for Protecting Privacy."

[29] Id.

[30] Id.

gests that watch lists need to be verified to ensure they are accurate, complete, and up-to-date, and this is particularly important if watch lists become the centerpiece of a system that seeks to identify who has relationships with "known bad guys." Anonymization will not protect innocent people if the government watch list is unreliable. False positives will result in inconvenience or more serious injury to innocent people. And even rigorous attention to data quality must still be combined with due process and redress mechanisms to protect against erroneous adverse consequences.

Another challenge for privacy is adaptation: Bad guys change behavior to avoid getting caught, while good guys change behavior to avoid hassle and protect their privacy. This can engender a vicious cycle leading to loss of privacy and denigration of data. Bad guys learn the rules, share information, adapt behavior, and reduce their signal-to-noise ratio. Analysts may counter these adaptations by making their rules more complex or digging deeper into private data. Doug Tygar has suggested that constructs from game theory and economics can help break that cycle. In any case, it is necessary to understand the limits of privacy policies in the face of adaptation.

Further research is therefore needed. Anonymized data correlation remains in its infancy. Although the commercial sector has been developing and implementing for a number of years techniques for matching data from disparate databases, government researchers have only recently begun working on analyzing data from more than one database for counterterrorism purposes in ways that will comport with applicable privacy and due process rules and preserve operational security. Solving complex knowledge management challenges in ways that enforce high levels of privacy protection should be a focus of government and private research.

In the end, however, we are convinced that information technology, properly designed and implemented with appropriate legal controls and oversight, offers the potential for enabling the government to act in support of vital national security concerns while also serving legitimate and critical privacy and liberty interests.

—James X. Dempsey is Executive Director of the Center for Democracy & Technology. Paul Rosenzweig is Senior Legal Research Fellow in the Center for Legal and Judicial Studies at The Heritage Foundation and Adjunct Professor of Law at George Mason University School of Law.