INFORMATION OF THE WORLD, UNITE!

Mashing everyone's personal data, from credit-card bills to cell phone logs, into one all-encompassing digital dossier is the stuff of Orwellian nightmares. But it is not as easy as most people assume

By Simson L. Garfinkel

KEY CONCEPTS

- The idea of linking together databases, known as data fusion, is the bête noire of privacy advocates. So far, however, it seems to be limited to specific contexts, such as gambling casinos and child-support enforcement.
- Data fusion is challenging because databases are riddled with errors and meaningless coincidences. New algorithms overcome some of these hurdles, but do they shift the overall ratio of cost and benefit?

—The Editors

A few years ago I bought a latte at Starbucks on the way to the airport, parked my car and got on a flight for the U.K. Eight hours later I got off at Heathrow, bought a prepay chip for my cell phone and went to buy a ticket for the train into London, when my credit card gave up the ghost and refused to work anymore. Not until I got back to the U.S. did I find out what had happened. Apparently, the small purchase at Starbucks, followed by the overseas purchase of the cell phone card, had tripped some kind of antifraud data-mining algorithm in my credit-card company's computer. It tried to call me, got my voice mail and proceeded to blacklist my credit card.

What I found so exasperating about the entire experience was that the computer should have known that the person using my card in England was *me*. After all, I had bought my plane ticket with that same card and had flown with a major U.S. carrier. Aren't all those databases supposed to be tied together?

Most people probably assume they are. We have come to expect from Hollywood films such as *Enemy of the State* and the Jason Bourne trilogy that shadowy organizations have instant access to all the databases we rely on and, with a few keystrokes, can spy on our every movement. The process of collecting information from multiple sources and merging it, known as data fusion, is supposed to create an information resource that is more powerful, more flexible and more accurate than any of the original sources. Proponents of data fusion say that their systems let organizations make better use of the data they already have; critics say that fusion threatens civil liberties by using information in ways that were never envisioned when it was first collected. Both sides assume that data-fusion systems actually work. The reality is that the systems are nowhere nearly as omniscient, as reliable or as well developed as many people think.

Out of Many, One

The technology of data fusion can trace its heritage back to the computerized matching programs of the 1970s. When Congress passed the Privacy Act in 1974, it also authorized the creation of the Federal Parent Locator Service, which now operates a giant blacklist, denying a wide range of federal benefits such as passports to noncustodial parents who are behind on their child support. Those data are fused with the National Directory of New Hires to find recently employed parents who are not up to date on their payments so that their wages can be garnished.

The term "data fusion" entered the technical vernacular in 1984, when researchers at Lock-

MULTITUDE OF DATA SOURCES can be merged into a single profile through the process of data fusion.

MELISSA THOMAS (*photoillustration*); PATRIK STOLLARZ *Getty Images* (*cell phone*); THINKSTOCK/CORBIS (*checkout card in book*); JIM CRAIGMYLE *Corbis* (*wand in laptop*); ELISE AMENDOLA *AP Photo* (*driver's license*); PURESTOCK (*passport*); DAVID MACK *Photo Researchers, Inc.* (*DNA*); PHILIP JAMES CORWIN *Corbis* (*bank statement*); SHEILA TERRY *Photo Researchers, Inc.* (*computer hard disk*)

heed Martin's Advanced Technology Center published two articles about a "tactical data fusion" system that would meld battlefield information from sensors, databases and other sources in real time for human analysts. Since then, the idea has blossomed. Bioinformatics investigators speak of genomic data fusion. The Department of Homeland Security has spent more than $250 million setting up some 58 state or local fusion centers. Nielsen, the consumer marketing company, has developed data-fusion products for identifying and targeting potential customers with specific characteristics, rather than wasting effort on the traditional scattershot approach to marketing.

But although data fusion has many faces, its use in identifying potential terrorists has stirred the greatest public debate. "The key to detecting terrorists is to look for patterns of activity indicative of terrorist plots based on observation of current plots and past terrorist attacks," wrote Rear Admiral John Poindexter and Robert L. Popp of the Defense Advanced Research Projects Agency (DARPA) in 2006. They argued that the World Trade Center bombing of 1993 and the Oklahoma City bombing of 1995 might have been prevented if the government could have scanned commercial databases for large purchases of fertilizer by nonfarmers. But getting those purchase records and combining them with a database of farm ownership and employment records would have required unprecedented government access to private computer systems. Every transaction—and thus every person—in the country would have been monitored without probable cause. For these reasons, among others, Congress killed Poindexter and Popp's research program, the Total Information Awareness project, in 2003.

Do Not Fold, Spindle or Mash

A wall of government secrecy does nothing to allay civil libertarians' fears. Agencies have revealed little about the data-fusion systems that they may or may not have deployed to protect national security: they argue that the bad guys would have an easier time evading fusion programs if they knew how they work. But enough information is publicly available to indicate that data fusion poses more than just ethical and legal problems; it also raises technical issues.

Data quality is one. Much of the information in databases was originally collected for purely statistical purposes and may not be accurate enough to make automated judgments with potentially punitive outcomes. In 1994 Roger Clarke of the Australian National University in Canberra studied computerized matching pro-

FUSION AND CONFUSION

To see how much information is out there, a *Scientific American* editor ordered an $80 report from an online consolidator of personal data, including criminal, real-estate and bankruptcy records. It was riddled with errors such as misspellings and confusion with namesakes elsewhere in the country—many of whom had liens on their property, though, thankfully, there were no criminal records. The report showed no signs of identity theft. Many people are not so fortunate.

[CASE STUDY]

Games People Play

Las Vegas casinos have been pioneers in fusing data from various sources because they face so many schemes to rip them off. Here are several examples based on true stories.

grams maintained by federal and state governments in the U.S. and Australia. These systems scanned millions of records and flagged thousands of potential "hits." But most turned out to be false positives. For example, one program for finding welfare cheats matched the employment records of the Department of Health and Human Services against the welfare rolls of the counties surrounding Washington, D.C. It generated roughly 1,000 hits, but further investigation showed that three quarters of the people identified were innocent. The benefits did not justify the costs of collecting data, training personnel and chasing down the false positives.

Many people feel that if a data-fusion program could anticipate and stop a major terrorist attack, it would be worth whatever it cost. Poindexter, a career naval officer, compared the technical problems to finding an enemy submarine in the vastness of the ocean. But finding the signatures of terrorist preparations in an ocean of data is much harder than finding subs in an ocean of water. The world's oceans may be huge, but every spot can be uniquely identified by a latitude, longitude and depth. The data oceans are not so easily categorized. Moreover, the world's seas are not doubling in size every few years, as the data oceans are. Much of information space is unmapped; data are spread across millions of individual computer systems, many hidden or otherwise unknown to the authorities.

Fusion is hard because we are drowning in data from a multitude of sources, all with different levels of detail and uncertainty. The real challenge in data fusion is not getting the data but making sense of them.

What's on Your Hard Drive?

A good way to understand the data-fusion problem is to start with the information on the hard drive of your computer. Between 1998 and 2005 I did just that: I purchased more than 1,000 used hard drives on eBay, at small computer stores and at swap meets; I even scavenged some from computers left abandoned on street corners. In January 2003 Abhi Shelat, now a computer sci-

HIDDEN DATA

Word-processing and other computer files typically contain "metadata" (such as the date of creation and your name and type of computer) and even deleted passages, such as those snide remarks you wrote in the first draft of a memo to your boss. A godsend to detectives and investigative journalists, such information becomes especially incriminating when merged with other data.

The only trouble is, sometimes the metadata are wrong. SCIENTIFIC AMERICAN ran earlier drafts of this article through two freeware metadata analyzers. They said the author had used OpenOffice on a Windows XP machine. But Garfinkel tells us he actually wrote them with Microsoft Office 2008 on a Mac. Oops. We did, however, enjoy seeing that one draft was revision number 139—reassuring us that he had indeed worked hard.

KENN BROWN *Mondolithic Studios*

entist at the University of Virginia, and I published a paper detailing what we found.

About a third of the drives were no longer functional, and another third had been properly wiped before being discarded. But the remaining third were a jackpot of personal information: e-mail messages, memoranda, financial records. One drive had previously been part of an automatic teller machine and recorded thousands of credit-card numbers. Another had been used by a supermarket to submit credit-card payments to its bank. Neither drive had been properly wiped before being resold on the open market.

The tools that enabled me to search the drives are widely available and not particularly sophisticated. Police departments around the world use the same kinds of tools to recover files from computers and cell phones. Sometimes users are unaware of the digital bread crumbs they leave. Consider the case of the so-called BTK killer, who committed eight murders in Wichita, Kan., in the 1970s and 1980s, then went underground. The killer resurfaced in March 2004, sending a letter to the *Wichita Eagle* detailing his earlier crimes and a floppy disk with a Microsoft Word document on it to a local television station. The file contained "metadata" that linked it to a computer at a local church. Police discovered that the person who had used it was president of the congregation council—and the killer.

Making a Hash of the Files

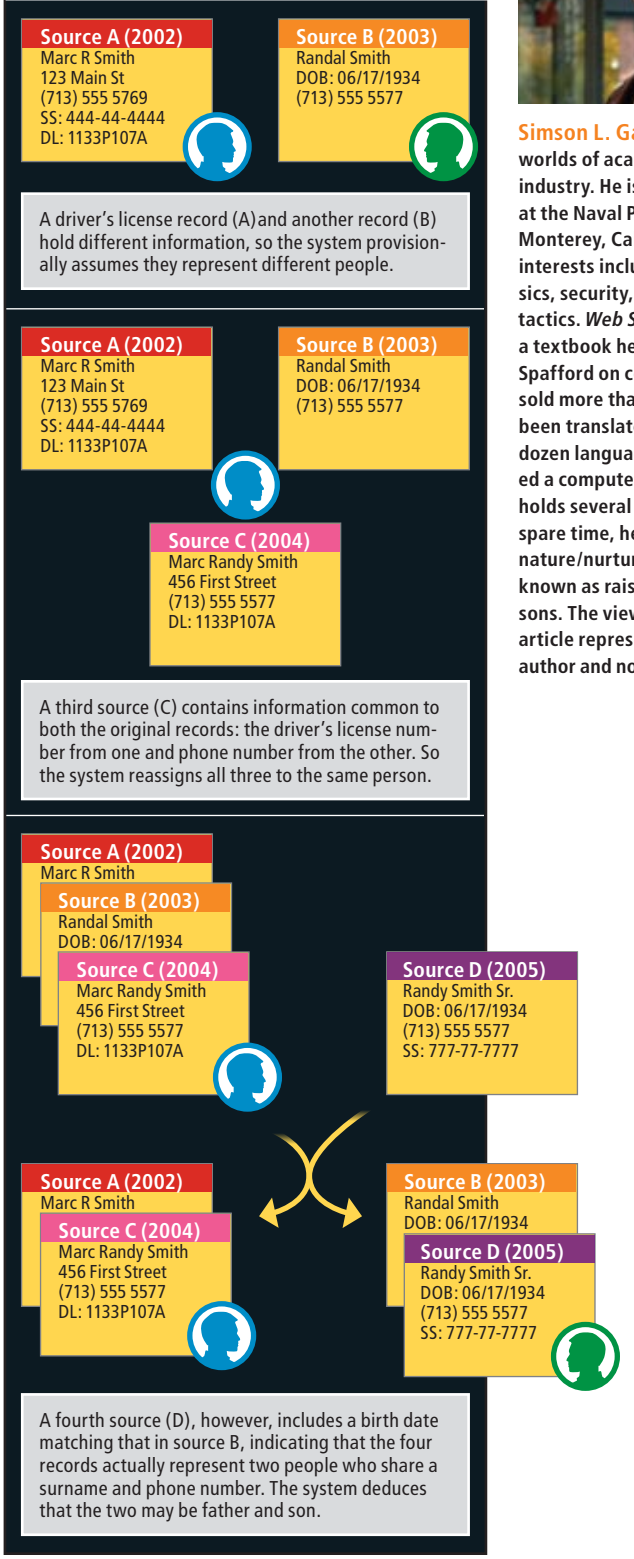
But figuring out which documents are important and which are worthless is difficult and requires fusing outside knowledge with the information on the hard drive. For example, when I started analyzing hard drives back in the 1990s, many of them contained copies of the *Island Hopper News*. It seemed highly suspicious. Then I learned that this electronic newspaper was actually a demo file distributed by Microsoft with a product called Visual Studio 6.0. Had I been unaware, I might have drawn spurious conclusions about the drive's previous owners.

The only way to screen out innocent files is to sample the world of digital documents and build a list of those that are widely available. One fast, automated way to do so is to create a so-called hash set. Cryptographic hash algorithms can assign a unique electronic fingerprint to any digital file. Two of the most popular are MD5, which creates a 128-bit fingerprint, and SHA-1, which generates a fingerprint 160 bits long. Then, instead of comparing two files byte by byte, forensics tools can examine the fingerprints.

[BEHIND THE SCENES OF FUSION]

How It Works

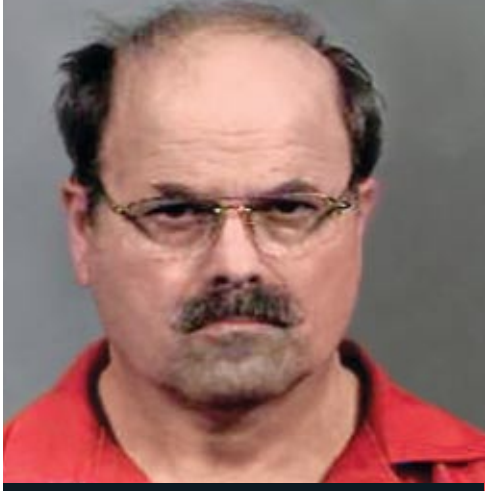
Originally developed for casinos, one data-fusion algorithm illustrates how to deal with partial, ambiguous information.



[THE AUTHOR]

Simson L. Garfinkel bridges the worlds of academia, journalism and industry. He is a computer scientist at the Naval Postgraduate School in Monterey, Calif., where his research interests include computer forensics, security, privacy and terrorist tactics. *Web Security & Commerce*, a textbook he wrote with Gene Spafford on computer security, has sold more than 250,000 copies and been translated into more than a dozen languages. Garfinkel founded a computer security firm and holds several related patents. In his spare time, he is conducting a nature/nurture experiment also known as raising identical twin sons. The views expressed in this article represent the opinion of the author and not the U.S. government.

COURTESY OF SIMSON L. GARFINKEL (*author*); MELISSA THOMAS (*illustration*)



DENNIS RADER, aka the BTK killer, gave himself away through metadata hidden in a Microsoft Word file he had sent to a TV station.

HURRICANE KATRINA evacuees, shown here at the Houston Astrodome, were reunited with relatives by a simple data-fusion system.

Supported by a grant from the Department of Justice, the National Software Reference Library at the National Institute of Standards and Technology (NIST) acquires software from hundreds of publishers and reduces every file to a cryptographic hash. NIST then distributes the database, which now has more than 46 million entries, to give forensic investigators a quick and reliable way of purging files that have been distributed by software publishers—files such as the *Island Hopper News*—and can therefore be safely ignored. Databases available from other federal agencies include e-fingerprints of computer hacker tools and of child pornography.

But despite their utility, hash databases represent only a small sampling of all the documents out there. To augment them, I developed a technique called cross-drive analysis. It can automatically piece together information scattered across thousands of hard drives, USB memory sticks and other data sources. The technique highlights and isolates identifiers such as e-mail addresses and credit-card numbers and weights them according to how frequently they appear: presumably the more common the identifier, the less important it is. Finally, the technique correlates the identifiers across all the individual devices: if an e-mail address or credit-card number appears on only two disk drives among thousands, there is a good chance that those two drives are related.

Who's Who?

Yet another problem for data fusers is identity. In the electronic world there may be dozens of people sharing the same name and dozens of names used by the same person. Some databanks may list Poindexter as John Marlan Poindexter or J. M. Poindexter or even misspell the rear admiral's last name Pointexter. A person's first name may be listed in one database as Robert, in another as Rob and in a third as Bob. A person whose Arabic name is transliterated Haj Imhemed Otmane Abderaqib in West Africa might be known as Hajj Mohamed Uthman Abd Al Ragib in Iraq.

Matching up the various names and account numbers that inhabit the electronic world with physical bodies is called identity resolution. Without it, data fusion is impossible. Curiously, a great deal of innovation in identity-resolution systems has been driven by casinos in Las Vegas. Under Nevada law, casinos are required to bar self-declared problem gamblers from playing their games. These gamblers voluntarily place their names on a list saying, in effect, "Don't let me gamble again!" But gambling can be an illness, and some people on the list still try to sneak in by changing their name or swapping a few numbers in their birth date. Casinos are also determined to exclude suspected or convicted cheaters. And if a guest is winning large sums at the blackjack table, a casino wants to make sure that the dealer and the player are not roommates.

Accordingly, casinos have funded development of a technique called nonobvious relationship analysis (NORA), which combines identity resolution with databases of credit companies, public records and hotel stays. A NORA system, for instance, might discover that the blackjack dealer's wife once lived in the same apartment building as the player who just won $100,000. In the 1990s software engineer Jeff Jonas developed a system that could match the names in a casino's computers with other sources of information in a way that tolerates error, ambiguity and uncertainty. The system works by building hypotheses based on the data and then revising these hypotheses as new information becomes available.

For example, it might receive a driver's license record for a Marc R. Smith, a credit report for a Randal Smith, and a credit application for a Marc Randy Smith. It might guess that the names belong to the same person—particularly if Marc R. Smith and Marc Randy Smith have the same driver's license number and if Randal Smith and Marc Randy Smith share a phone number. But suppose new data show that Randy Smith, Sr., shares the birth date of Randal Smith but that his Social Security number differs from that of Marc R. Smith. Now the system might revise its guess, deciding that Marc R. Smith is Randal Smith, Jr., whereas Randy Smith is Randal Smith, Sr.

IDENTITY THEFT

Many *Scientific American* staffers have suffered mild forms of identity theft. Though disconcerting, the problems remained contained because databases are largely isolated from one another. But as companies increasingly link them together, the theft of one piece of information could infect a person's entire digital identity.

- One staffer's bank recently froze her credit card after detecting some unusual transactions. Several were legitimate, but two were not. The bank sent a new card. Who stole her card number remains a mystery.
- Another person was surprised to receive a change-of-address confirmation request from her brokerage firm. The new address was not hers. The broker, who was new to the firm, played innocent, so the staffer called the police. It turned out the broker was fishing out seemingly inactive accounts and transferring them to a collaborator, who cashed them out.
- One person started receiving delinquency notices from his cell phone provider. Evidently someone had opened an account under his name. It took a year to clear up the problem and restore his credit rating.

THE AUTHOR studied data on abandoned hard drives as a test case of how data fusion can aid police forensics investigations.

JOHN POINDEXTER, former national security adviser, tried in 2002 to set up a master government database to find terrorists.

The key to making all this work is programming the system so that it never confuses original data with a conclusion inferred from those data.

Jonas sold the system and his company to IBM in 2005. Since then, IBM has added a feature called anonymous resolution: two organizations can determine whether they share the name of one person in their respective databases—without sharing the names of all the people who do not match. The technique works by comparing cryptographic hashes instead of real names.

Privacy advocates still maintain that hashes, cross-drive analysis, anonymous resolution and the like do little to overcome their fundamental objections. After all, these systems still use personal information for purposes other than the ones for which it was originally acquired. They also make it routine to sweep up private data in a dragnet regardless of whether the people involved are suspected of committing a crime. Yet these systems generate significantly fewer false positives than did those developed in the 1980s. At some point the social benefits may come to outweigh the privacy costs of having a computer snoop through people's records.

Putting It All Together

So just how well do fusion systems actually work? Data quality remains a serious problem. Pull your credit report from each of the nation's three major credit-reporting agencies, for instance, and each report will probably contain errors and inconsistencies. Those data can lie dormant for years without causing much trouble. The danger arises when some newfangled algorithm reads too much into the inconsistencies.

Even when data are accurate, relationships brought to light by comparing databases may have real meaning or may be purely coincidental, as inevitable as finding two people in a room who share the same birthday. Maybe the four people who meet once a week to take a long drive are planning a crime. Then again, they may belong to a softball team and travel together to each week's big game.

Society's expectations for data fusion may be unreasonably high. If terrorists blend in with the population, human investigators and computers alike will be hard-pressed to find them. Most systems of data mining and fusion have some kind of sensitivity adjustment: move the slider to the left, and the system fails to find genuine matches; move it to the right, and the system makes too many predictions that turn out to be wrong. Where should the slider be set? If a system flags every third airline passenger, it will be more likely to spot a real terrorist. But it will also bring air traffic to a standstill and overwhelm law enforcement.

If a data-fusion system does not work as desired, its algorithms could be fundamentally flawed. But the problem could also be a dearth of data. Likewise, if the system is performing well, giving it more data might make it perform even better. In other words, the people building and using these systems are naturally inclined to want more and more input data, no matter how well the systems are working. Thus, data-fusion projects have a built-in tendency toward mission creep—to the consternation not only of civil-liberties advocates but also of those footing the bill. In his 1994 article Clarke concluded that trade-offs "between the State's interest in social control and individual citizens' interest in freedom from unreasonable interference [are] being consistently resolved in favor of the State."

What makes the public debate over data fusion so frustrating to me as a scientist is the fact that so little information has been publicly released about data-fusion systems in actual use. It hearkens back to the cryptography debates of the 1990s, when the U.S. government argued that there were good reasons for legally restricting the use of cryptography but that those reasons were so sensitive that discussing them in public would be a threat to national security. I suspect a similar debate is brewing over the government's use of data fusion, not to mention the applications of this powerful technology in business and even in political activities. It is a debate well worth having—and having in public. ■

MORE TO EXPLORE

Computer Matching by Government Agencies: The Failure of Cost/Benefit Analysis as a Control Mechanism. Roger Clarke in *Information Infrastructure & Policy,* Vol. 4, No. 1, pages 29–65; March 1995. Available at **www.anu.edu.au/people/Roger.Clarke/DV/MatchCBA.html**

Database Nation: The Death of Privacy in the 21st Century. Simson Garfinkel. O'Reilly, 2000.

Forensic Feature Extraction and Cross-Drive Analysis. Simson L. Garfinkel in *Digital Investigation,* Vol. 3, Supplement 1, pages 71–81; September 2006. Available at **www.dfrws.org/2006/proceedings/10-Garfinkel.pdf**

Threat and Fraud Intelligence, Las Vegas Style. Jeff Jonas in *IEEE Security & Privacy,* Vol. 4, No. 6, pages 28–34; November/December 2006. Available at **http://jeffjonas.typepad.com/IEEE.Identity.Resolution.pdf**

Simson L. Garfinkel's Web sites are available at **www.simson.net** and **http://faculty.nps.edu/slgarfin**