Data Surveillance

Can you find a terrorist in your database? Do the register receipts from discount drug stores hold the secret to stopping avian flu outbreaks before they become epidemics? Can anonymized data be "re-identified," compromising privacy and possibly jeopardizing personal safety?

Simson Garfinkel and Michael D. Smith
Harvard University

Government and industry are increasingly turning to data mining with the hope that advanced statistical techniques will connect the dots and uncover important patterns in massive databases. Proponents hope that this so-called data surveillance technology will be able to anticipate and prevent terrorist attacks, detect disease outbreaks, and allow for detailed social science research—all without the corresponding risks to personal privacy because machines, not people, perform the surveillance.

Emergence of data surveillance

The US public got its first look at data surveillance in 2002 when it learned about the Pentagon's Total Information Awareness (TIA) project. The project, one of several sponsored by the Information Awareness Office (IAO) at DARPA, quickly became a lightening rod for attacks by privacy activists, critics of the Bush administration, and even conspiracy theorists. Some of these attacks were motivated by what the IAO proposed; others were based on the fact that the IAO was headed by Admiral John Poindexter, who had a pivotal role in the 1980s Iran-Contra Affair; and still others were based on the IAO's logo—the all-seeing Eye of Providence floating above an unfinished pyramid (similar to the seal on the back of the US$1 bill), carefully watching over the Earth.

Critics charge that data surveillance is fraught with problems and hidden danger. One school of thought says that surveillance is surveillance: whether the surveillance is done by a person or a computer, some kind of violation to personal privacy or liberty has occurred. The database, once created, must be protected with extraordinary measures and is subject to abuse. Moreover, if data surveillance technology says a person is a potential terrorist, that person could then be subject to additional surveillance, questioning, or detention—even if they haven't done anything wrong. After extensive media coverage and congressional hearings, Congress terminated funding for the IAO in 2003.

Data surveillance jumped again to the front pages of US newspapers in May 2006, when *USA Today* published a story alleging that the nation's telephone companies had shared the telephone records of millions of Americans with the National Security Agency (NSA). According to *USA Today*, the NSA used this information to create a database of detailed information for every telephone call made within the nation's borders. The spy agency then mined this database to uncover hidden terrorist networks.

A database of every phone call within the US extending years back through time could certainly prove useful for fighting domestic terrorists. If a French male of Moroccan descent were arrested in Minnesota after receiving 50 hours of flight training and a US$14,000 wire transfer from a known terrorist, such a database could provide a report of every person who called or received a telephone call from that individual. This kind of *social network analysis* would prove invaluable in finding the would-be pilot's comrades-in-arms, but it might also identify his flight teacher, his pizza delivery service, and even the teenager who mows his lawn.

Yet in all of the media coverage of the TIA, the NSA database, and similar projects, many questions seem to

be left unasked and unanswered: Does the technology really work—can you find the terrorist in the database? Can you find the terrorist without destroying everybody else's privacy in the process? Is it possible to perform privacy-protecting data mining—at least in theory? And is it possible to turn that theory into practice, or do too many real-world concerns get in the way?

Does the technology really work—can you find the terrorist in the database … without destroying everybody else's privacy in the process?

The workshop

To explore these questions, as well as to improve the public's general understanding of these questions, Harvard's Center for Research on Computation and Society held a day-long workshop in June 2006 on data surveillance and privacy protection. Robert Popp, who served as deputy of the IAO under Poindexter and was a driving force behind the TIA program, gave the keynote address. Popp presented his and Poindexter's vision for using data surveillance for countering terrorism; an article based on that presentation appears on p. 18 of this special issue. (For more information about the workshop, please visit http://crcs.deas.harvard.edu/workshop/2006/.)

Whether data surveillance can find real terrorists and stop actual attacks before they happen is still unproven in the academic literature. Although there's no denying that data surveillance has resulted in arrests, it's not clear if those individuals were actually terrorists or merely people writing novels about terrorists. On the other hand, attendees learned that data surveillance is good for a lot more activities than hunting terrorists.

For example, it might be able to detect and help public health officials contain an outbreak of avian flu. Kenneth Mandl, a researcher at the Harvard Medical School Center for Biomedical Informatics, showed how records of emergency rooms' admissions could anticipate the deaths associated with pneumonia and influenza reported to the US Centers for Disease Control. This isn't tremendously surprising, of course—many deaths result from individuals who didn't seek treatment until it was too late, then went to the emergency room. But what's exciting, Mandl reported, is that pediatric admissions peak roughly a month before adult emergency room admissions. With this knowledge, public health officials could build a system that predicted adult outbreaks by monitoring admissions of children. If outbreaks can be predicted, it might be possible to nip them in the bud.

The Realtime Outbreak and Disease Surveillance (RODS) project at the University of Pittsburgh might one day be able to provide health officials with even better advance warning. The RODS project monitors the sale of over-the-counter cold remedies and other healthcare products in more than 20,000 stores throughout the US. The theory here is that people will attempt to treat themselves with over-the-counter cold medications before they get so sick that they report to the hospital emergency room. Their work so far indicates that sales of these medications peak two weeks before hospital admissions do.

RODS researchers stress that this so-called *biosurveillance* (the continuous collation and analysis of medically related statistical data) doesn't violate privacy because no personally identifiable information is ever assembled or reported. The system collects only aggregate sales from a sampling of the nation's largest drugstores and mass-merchandise chains. Of course, in the case of an actual avian flu outbreak or bioterrorism attack, it would be helpful to know the names of those infected. But collecting this information isn't necessary to achieve the project's primary goals and would create unacceptable risks to those involved because the data would be so easily subject to abuse.

In his presentation, Mandl discussed five techniques that organizations collecting data could adopt to protect privacy in large-scale data mining efforts. Policies must be set in place to limit access to sensitive data, and then organizations must self-police themselves to make sure that those policies are enforced. When possible, data subjects should be given the ability to exert personal control over their own information. Data should be de-identified whenever possible. Finally, says Mandl, all data must be stored with encryption, so that it's protected in the event of a breach.

All of Mandl's techniques assume that sensitive information will be collected and ultimately used in a highly controlled environment. Indeed, this is a model familiar to most people today. Law enforcement, national intelligence organizations, businesses, and even journalists collect a lot of sensitive information during the day-to-day course of their work and then carefully control who can access the information and how they can use it. Big fences and background investigations are an unfortunate but necessary part of the security model for these organizations.

But another approach on the horizon is to use advanced algorithms and cryptographic theory to avoid the problem in the first place. Algorithms and systems now under development make it possible to collect information in a kind of predigested form that allows some queries (but not others) and that makes it impossible to recover the original, undigested data. The US National

Science Foundation-funded project on Privacy, Obligations, and Rights in Technologies of Information Assessment is developing a host of tools and technologies to enable this kind of privacy-preserving data mining. Other work is being done at the University of California, Los Angeles', Center for Information and Computer Security.

In this special issue of *IEEE Security & Privacy*, we present two articles based on presentations at our workshop. In addition to the article from Popp and Poindexter, we have an article by Jeff Jonas, founder of IBM's Entity Analytic Solutions division, who explains the process of entity resolution—a technique by which names in different databases are determined to represent the same person.

One of the fundamental technical disagreements between these two articles pertains to the kinds of queries performed. Popp and Poindexter advocate pattern-based queries—for example, a standing query could scan for anyone who buys a large quantity of fertilizer, fuel oil, 55-gallon drums, and then rents a truck. Such queries might find new, spontaneous terrorist groups, but they're also more likely to pick up people who have no evil intent—for example, farmers. Jonas, meanwhile, argues that we should focus our limited resources using relationship information—for example, looking for terrorists by looking for individuals who have nonobvious relationships with other terrorists.

Although many of the intelligence techniques and operational details of the global war on terrorism must necessarily remain secret for them to be effective, we believe that it's both possible and necessary to have an informed academic debate on both the tools and appropriateness of data surveillance. If these techniques are effective, they could be tremendously beneficial to our society in the fight against terrorism, disease, and even economic inefficiency. But if they don't work, we need to know that, too—so that we can spend our limited resources on other approaches. □

Simson L. Garfinkel *is a postdoctoral fellow at Harvard's Center for Research on Computation and Society. His research interests include computer security, computer forensics, and privacy. Contact him at simsong@acm.org.*

Michael D. Smith *is the Gordon McKay Professor of Computer Science and Electrical Engineering and the associate dean for computer science and engineering at Harvard's Division of Engineering and Applied Sciences. His research interests include dynamic optimization, machine-specific and profile-driven compilation, high-performance computer architecture, and practical applications of computer security.*

Countering Terrorism through Information and Privacy Protection Technologies

Security and privacy aren't dichotomous or conflicting concerns—the solution lies in developing and integrating advanced information technologies for counterterrorism along with privacy-protection technologies to safeguard civil liberties. Coordinated policies can help bind the two to their intended use.

Robert Popp
National Security Innovations

John Poindexter
JMP Consulting

The terrorist attacks of September 11, 2001 transformed America like no other event since Pearl Harbor. The resulting battle against terrorism has become a national focus, and "connecting the dots" has become the watchword for using information and intelligence to protect the US from future attacks.

Advanced and emerging information technologies offer key assets in confronting a secretive, asymmetric, and networked enemy. Yet, in a free and open society, policies must ensure that these powerful technologies are used responsibly and that privacy and civil liberties remain protected. In short, Americans want the government to protect them from terrorist attacks, but fear the privacy implications of the government's use of powerful technology inadequately controlled by regulation and oversight. Some people believe the dual objectives of greater security and greater privacy present competing needs and require a trade-off; others disagree.[1–3]

This article describes a vision for countering terrorism through information and privacy-protection technologies. This vision was initially imagined as part of a research and development (R&D) agenda sponsored by DARPA in 2002 in the form of the Information Awareness Office (IAO) and the Total Information Awareness (TIA) program. It includes a critical focus and commitment to delicately balancing national security objectives with privacy and civil liberties. We strongly believe that the two don't conflict and that the ultimate solution lies in utilizing information technologies for counterterrorism along with privacy-protection technologies to safeguard civil liberties, and twining them together with coordinated policies that bind them to their intended use.

Background and motivation

Terrorists are typically indistinguishable from the local civilian population. They aren't part of an organized, conventional military force—rather, they form highly adaptive organizational webs based on tribal or religious affinities. They conduct quasi-military operations using instruments of legitimate activity found in any open or modern society, making extensive use of the Internet, cell phones, the press, schools, houses of worship, prisons, hospitals, commercial vehicles, and financial systems. Terrorists deliberately attack civilian populations with the objective to kill as many people as possible and create chaos and destruction. They see weapons of mass destruction not as an option of last resort but as an equalizer—a weapon of choice.

Of the numerous challenges to countering terrorism, none are more significant then being able to detect, identify, and preempt terrorists and terrorist cells whose identities and whereabouts are unknown a priori. (Alan Dershowitz's *Preemption: A Knife that Cuts Both Ways* [W.W. Norton & Company, 2006] offers an extensive discussion of preemption and the need for a legal structure.) In our judgment, if preemption is the goal, the key to detecting terrorists is to look for patterns of activity indicative of terrorist plots based on observations of current plots and past terrorist attacks, including estimates about how terrorists will adapt to avoid detection. Our fundamental hypothesis is if terrorists plan to launch an attack, the plot must involve people (the terrorists, their financiers, and so forth). The transactions all these people

conduct will manifest in databases owned by public, commercial, and government sectors and will leave a signature—detectable clues—in the information space. Because terrorists operate worldwide, data associated with their activities will be mixed with data about people who aren't terrorists. If the government wants access to this activity data, then it must also have some way to protect the privacy of those who aren't involved in terrorism.

This hypothesis has several inherent critical challenges. First, can counterterrorism analysts imagine and understand the numerous signatures that terrorist plans, plots, and activities will create? Second, if they do understand these signatures, can analysts detect them when they're embedded in a world of information noise before the attacks happen (in this context, *noise* refers to transactions corresponding to nonterrorists)? Finally, can analysts detect these signatures without adversely violating the privacy or civil liberties of nonterrorists? Ultimately, the goal should be to understand the level of improvement possible in our counterterrorism capabilities if the government could use advanced information technologies and access a greater portion of the information space; but also consider the impact—if any—on policies such as privacy, and then mitigate this impact with privacy-protection technology and corresponding policy.[2,3]

Countering terrorism

Information technology plays a crucial role—and is a major tenet—of our counterterrorism strategy because it ultimately has to make sense out of and connect the relatively few and sparse dots embedded within the massive amounts of information potentially available to, and already flowing into, the government's intelligence and counterterrorism agencies.

Numerous information technologies can help intelligence analysts detect and understand the clues terrorists leave behind when plotting their next move. In the simplest terms, these technologies fall into one of two broad categories: collections and analytics. Figure 1 provides a simple illustration of this counterterrorism framework. For collections, we won't discuss the vast array of sensor technologies that fall within this category here; instead, see Table 1 (p. 27), which provides a sample of the authorization provided to the US intelligence community for its foreign and domestic intelligence and counterintelligence data collections.

For analytics, key intelligence tools include collaboration; text analysis and decision aides; natural language processing (in particular, speech-to-text transcription and foreign-to-English translation); pattern analysis; and predictive (anticipatory) modeling. These technologies help analysts create models (and discover instances of new models) of terrorist activity patterns; search and exploit vast amounts of multimedia, multiformat, and multilingual speech and text; extract entities and entity relationships from massive amounts of data; collaborate, reason, and share information and analyses so that analysts can hypothesize, test, and propose theories and mitigating strategies about plausible futures; and advise decision- and policy-makers on the impact of current or future policies and prospective courses of action. We don't discuss these technologies in detail here, but more information appears elsewhere.[1,4,5]

In our view, modeling tools play a crucial role in countering terrorism. The analytical community first creates scenarios of terrorist plots and attacks using previous attacks, intelligence reports, red teams, war games, table-top exercises, and the like. These terrorism scenarios would consist of a range of transactions and steps that terrorists must perform in support of their plot to attack a specific target type using a specific mode of attack. Analysts then codify these scenarios in a set of quantitative and computational models based on a wide range of nonlinear mathematical and nondeterministic stochastic computational approaches for capturing social phenomena and pathological behavior. These models are essentially hypotheses about terrorist plots and would be translated into a series of questions about the types of transactions terrorists would need to execute, the types of evidence analysts would need to accrue, the keywords and patterns analysts would need to associate, and the like.

Terrorist activity isn't easily reduced or amenable to classical analytical methods; moreover, the associated data can be incredibly poor due to ambiguous, erroneous, and conflicting reports. No single theory or modeling approach is sufficient, so we must integrate an ensemble of models that have more information than any single model has to estimate a range of plausible futures and provide competing explanations as to what the information means. Robust adaptive strategies that hedge across these plausible futures will provide practical actionable options for the decision-maker to consider.[1,4,5]

Early results show promise

The importance (and promise) of these information technologies has already emerged through experiments conducted with several entities in the intelligence community. Experiments let us assess these technologies for utility and merit in the context of real-world problems before large amounts of funds are expended to fully implement them. Moreover, to push the envelope of what's

Numerous information technologies can help intelligence analysts detect and understand the clues terrorists leave behind.

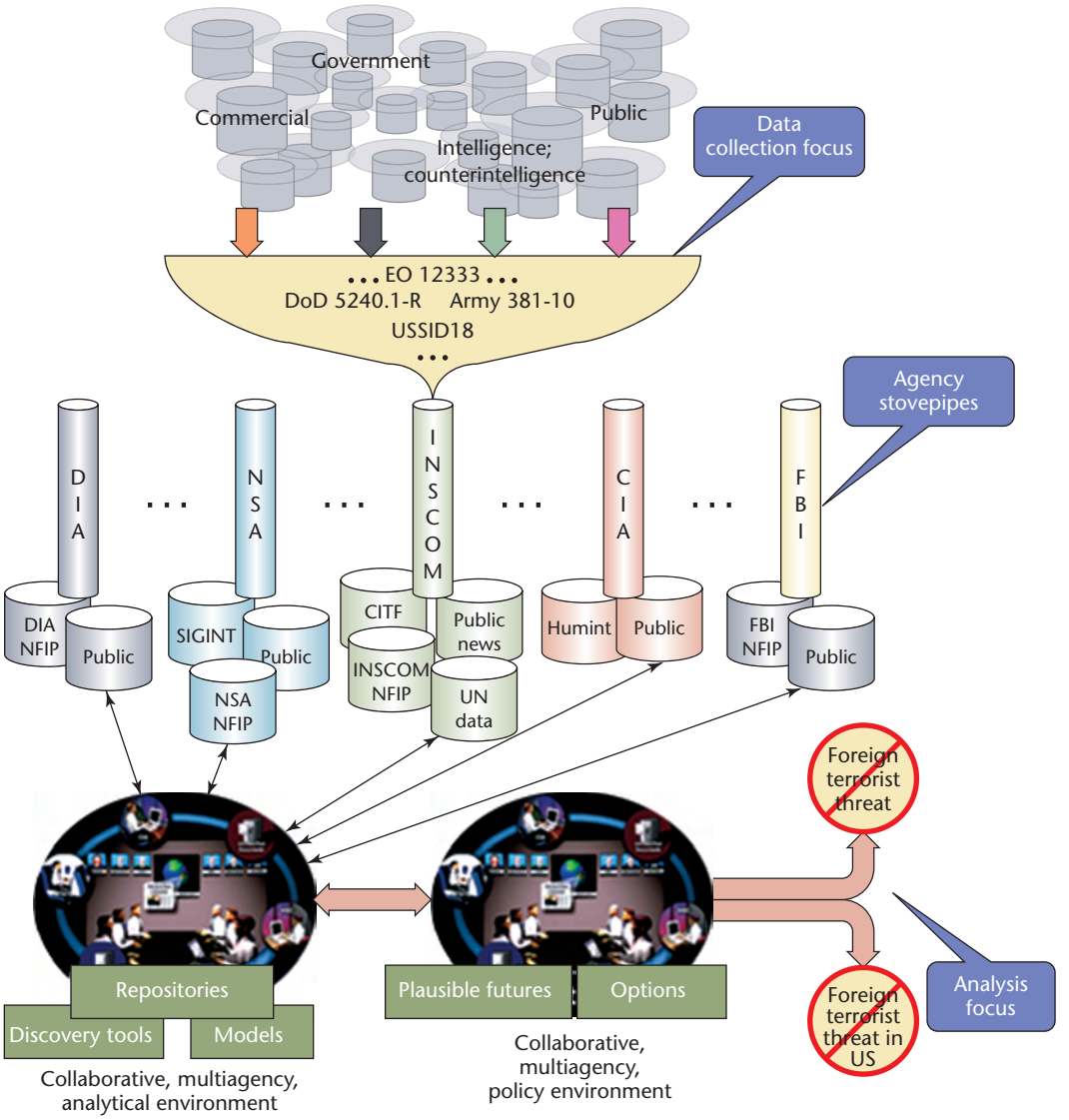

Figure 1. Counterterrorism framework. Information technologies fall into one of two broad categories: collections and analytics.

possible, failure in the experimental environment is an acceptable outcome for a particular technology.

Figure 2 shows an approach to understanding how to measure the operational payoff of information technologies for counterterrorism. As the graphic shows, when doing traditional analysis, an analyst spends much of his or her time on the major processes broadly defined as research, analysis, and production. This *bathtub curve* shows that analysts spend much time doing research and production but too little time doing analysis. An objective of conducting experiments with this curve is to determine whether we can improve analyses via information technology by reversing this trend and inverting the curve.[4,5]

Specifically, Figure 2 shows the results of an experiment in which the intelligence question posed to analysts was, "What is the threat posed by Al Qaeda's weapons of mass destruction capabilities to several cities in the US?" The data were drawn from various classified intelligence sources, foreign news reports, and the Associated Press (AP) and other wire services.[4,5] The information technologies used in the experiments included a peer-to-peer collaboration tool, a structured argumentation decision aide, a multilingual processing tool for audio phonetic searching/indexing as well as text filtering/categorization, and several graph-based link analysis tools. The results of the experiment show an inverted bathtub curve, allowing for more and better analysis in a shorter period of time, as a result of analysts using information technologies. The obvious significance is that analysts spend a greater percentage of their

Table 1. Sample of the US intelligence community's legal authority for data collection.

AUTHORITY	DESCRIPTION
Executive Order (EO) 12333	Authorizes US intelligence activities
Foreign Intelligence Surveillance Act (FISA) of 1978	Prescribes procedures for physical and electronic surveillance and collection of intelligence information between or among foreign powers
USA Patriot Act	Dramatically expands the authority of American law enforcement for fighting terrorism in the US and abroad
US Department of Defense (DoD) Directive 5240.1-R	Provides the DoD with implementation guidance for EO 12333
Army regulation 381-10	Provides the Army with implementation guidance for DoD Directive 5240.1-R
US Signals Intelligence Directive (USSID) 18	Governs signal intelligence (SIGINT) for the National Security Agency (NSA)

time doing what is most important in our view—namely, the critical-thinking tasks instead of the more mundane research and production tasks. The results also included an impressive savings in analyst labor (that is, half as many analysts participated in the IT-enhanced analysis) and an increase in the number of reports produced (that is, analysts created five reports in the time it took to create one manually).

Our explanation for the bathtub curve's inversion for the intelligence question at hand includes

- The time spent in the research phase shrank dramatically by using the collaboration tool (Groove) across multiple agencies to harvest and share "all" pertinent data.
- The structured argumentation modeling tool (SEAS, for Structured Evidentiary Argumentation System) let analysts explicitly represent their hypotheses for comparison and assessment, and identify evidentiary data gaps for which data must be searched and harvested.
- The multilingual processing tool (FastTalk) let analysts phonetically index and search vast quantities of foreign audio streams and thereby reduce the time required to find pertinent data.
- The link analysis tools (Analyst Notebook) let analysts automatically capture portions of their analysis in an easy-to-understand visual format.[4,5]

Figure 3 shows the utility of various information technologies in detainee operations support. In this scenario, actual government interrogators questioned actual detainees at the US military facility at Guantanamo Bay, Cuba, and wanted analytical support to make sense of the stacks of real reports from hundreds of interrogation sessions. The analysts used a link analysis tool to find nonobvious relationships between different entities (people, places, and things), a group detection tool to find nonobvious groupings among entities, an entity resolution tool to resolve entities and aliases in the interrogation reports, a Bayesian classification tool to classify detainees of unknown status as either statistically more likely to resemble known terrorists or nonterrorists, and a link chart visualization tool to pull everything together. These tools showed the interrogators web-like diagrams of connections (or relationships) among different entities that weren't readily apparent, inconsistencies in detainee stories, salient relationships across detainees, useless data to disregard, and data that could be most informative for follow-up interrogations. The tools' output also included a rank-ordered list of detainees with the likelihood that each had attributes resembling known terrorists or nonterrorists.[4–6] (It should be noted that officials at Guantanamo Bay established the "ground truth" in terms of which detainees were terrorists and which ones weren't.) Based on conversations with the intelligence analysts who performed this work, anecdotal evidence suggests that the detainees classified as "likely a terrorist" were in fact terrorists, and no cases

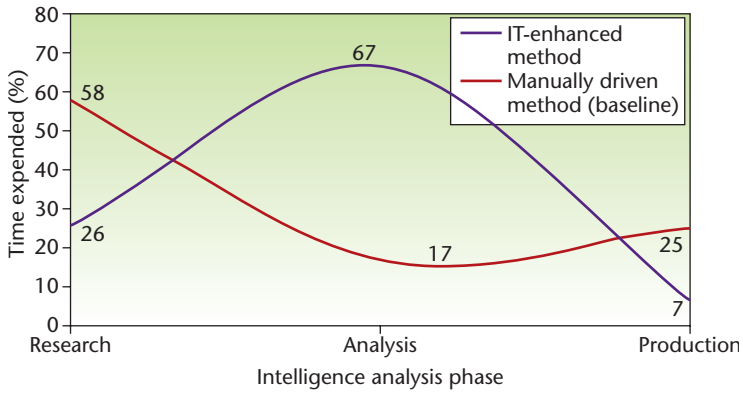

Figure 2. The analyst "bathtub" curve. The red curve represents the baseline distribution of time an analyst manually spends on research, analysis, and production; the blue curve represents the improvement due to information technology enhancements.

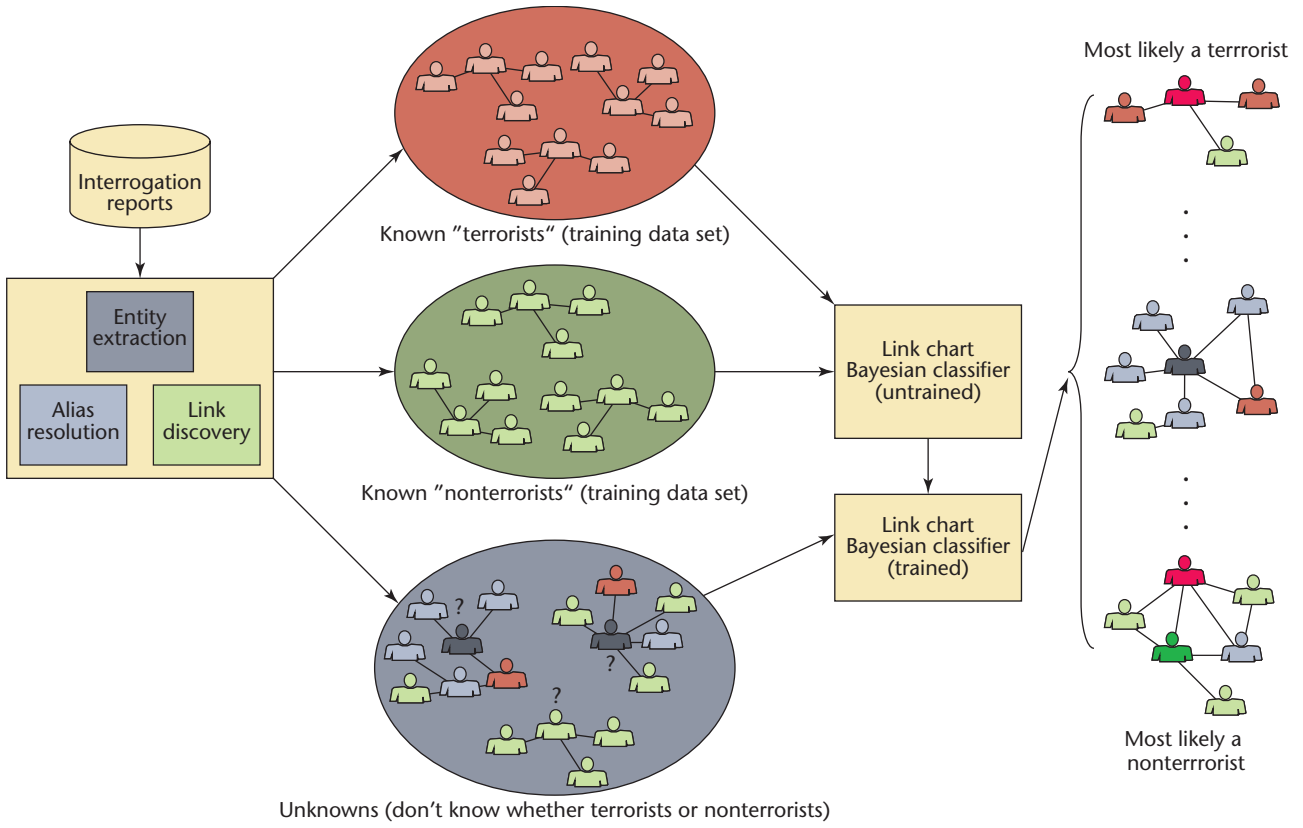

Figure 3. Detainee operations support at Guantanamo Bay. The analysts used information technologies to build web-like diagrams of relationships between entities that weren't immediately apparent.

were found in which detainees who weren't terrorists were classified as "likely a terrorist."

Figure 4 shows an experiment in which a novel multilingual IT front-end system automatically ingests, transforms, extracts, and autopopulates in near real time the back-end analytical models from massive amounts of text data. In this experiment, the problem concerned understanding and forecasting the preconditions and root causes that give rise to instability in nation states. Failed states are important because they offer a safe haven and potential breeding ground for terrorists. The challenge posed to analysts here was to assess and forecast the level of instability in two specific countries in Southeast Asia. The data came from a variety of open sources and included more than 1 million English documents and 2,300 non-English documents. The information technologies used included a back-end rebel activity model (RAM) based on a Bayesian network and hidden Markov models (HMMs) that measured the amount of rebel activity (on the part of separatists, insurgents, terrorists, Islamic extremists, and so forth); a front-end language-independent text-based transformation and categorization tool based on a Hilbert engine (a technology that numerically encodes ASCII text into vectors in Hilbert space); and a linguistic pattern analyzer (LPA) that automatically populates the HMMs in the RAM model.

The experiment's results were impressive—given a corpus of 1,236,300 documents, a human would need 117 man years to read it all (assuming it took 12 minutes to read each document), or 280 humans to read the documents in six months. The automated front-end system based on LPA, the Hilbert engine, and RAM would take a mere 0.05 man years with a one-time cost of 0.76 man years to configure LPA with the numerous multilingual scripts. Assuming it cost US$100K per man year, the automated front-end would provide a savings of US$11,695,141 over the human method.

Signatures in silos

One of the major criticisms leveled against an approach such as ours is that what we're describing is mass dataveillance—warehousing massive amounts of data in a megadatabase and using data mining techniques that will lead to multiple false positives and a massive invasion of Americans' privacy. We disagree. Although we appreciate the significant information policy challenges concerning data analysis in actual transaction spaces, we

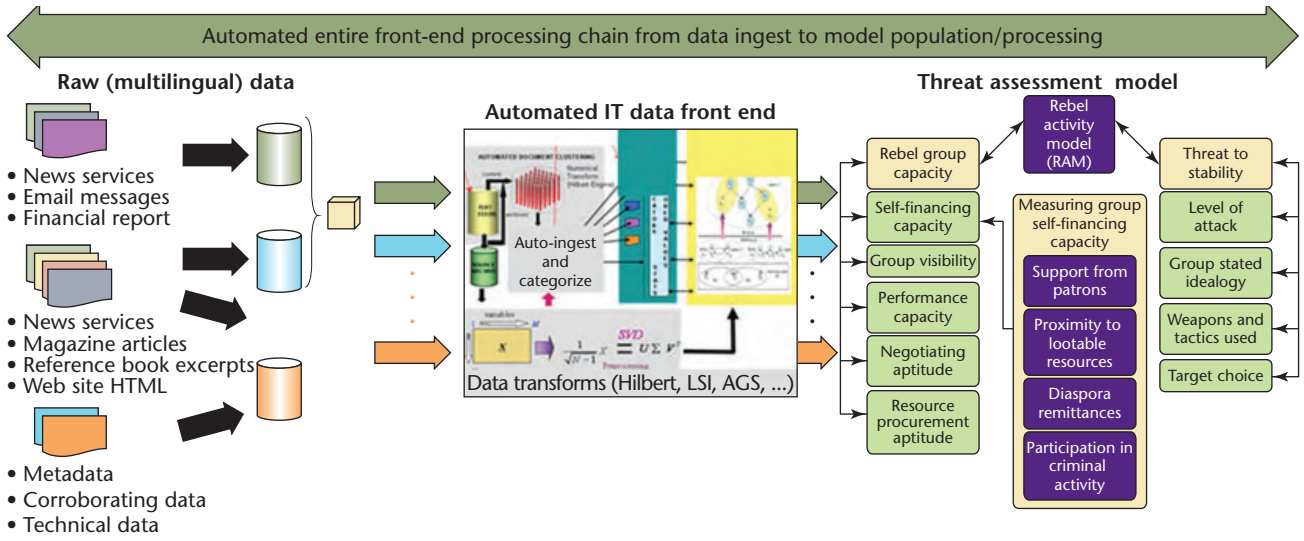

Figure 4. A multilingual IT front-end system. This tool automatically ingests, transforms, extracts, and autopopulates in near real time the back-end analytical models from massive amounts of text data.

Table 2. Data mining vs. terrorism detection.

COMMERCIAL DATA MINING	TERRORISM DETECTION
Discover comprehensive models of databases to develop statistically valid patterns	Detect connected instances of rare patterns
No starting points	Known starting points or matches with patterns estimated by analysts
Apply models over entire data	Reduce search space; results are starting points for human analysis
Independent instances (records)	Linked transactions (networks)
No correlation between instances	Significant autocorrelation
Minimal consolidation needed	Consolidation is key
Dense attributes	Sparse attributes
Sampling okay	Sampling destroys connections
Homogenous data	Heterogeneous data
Uniform privacy policy	Nonuniform privacy policy

believe technology and enabling policies can help preserve civil liberties and protect the privacy of those people who aren't terrorists while keeping us all safer from attack.

Data mining commonly refers to using techniques rooted in statistics, rule-based logic, or artificial intelligence to comb through large amounts of data to discover previously unknown but statistically significant patterns. However, the general counterterrorism problem is much harder because unlike commercial data mining applications, we must find extremely rare instances of patterns across an extremely wide variety of activities and hidden relationships among individuals. Table 2 gives a series of reasons for why commercial data mining isn't the same as terrorism detection in this context. We call our technique for counterterrorism activity *data analysis*, not data mining.[7]

Instead of warehousing data in one megadatabase, we believe data must be left distributed over the large number of heterogeneous databases residing with their data owners. In an intelligence context, agency silos and stovepipes aren't necessarily bad—they allow analysts from different agencies to create alternative competing hypotheses, and they also protect agency-specific sources and methods. In our judgment, the goal shouldn't be to tear down these silos, but to punch holes in them and enable collaboration across agencies when appropriate and advantageous.

Advanced search and discovery tools should be used to search and query relevant databases—under rigorous access control and privacy protections—with the results of the search/query added to the analytical models. Because finding evidence of a suspicious terrorist plot isn't

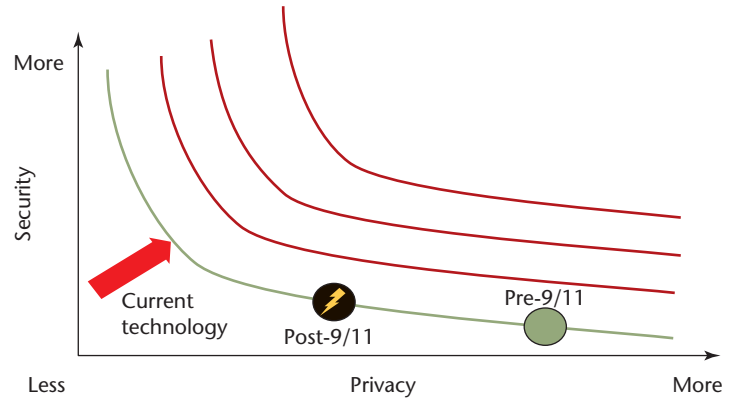

Figure 5. Security vs. privacy curves. Laws and policies dictate where we are on a curve; new privacy technology can create new curves.

easy, we believe two basic types of queries are necessary: subject-based queries (sometimes referred to as *particularized suspicion*) and pattern-based queries (sometimes referred to as *nonparticularized suspicion*).

Subject-based queries let analysts start with known suspects, look for links to other suspects, people, places, things, or suspicious activities, and do so within well-defined and practiced sets of legal and regulatory protocols. Law enforcement personnel have used this technique successfully for years as part of their background investigations and as a forensic tool. In the previous section, we gave examples of how subject-based queries can be beneficial for counterterrorism purposes (such as in the Guantanamo Bay detainee example), but this might not be enough. To get ahead of the terrorism problem, we need to consider pattern-based queries that don't require a subject's prior identification.

Pattern-based queries let analysts take a predictive model and create specific patterns that correspond to anticipated terrorist plots, and use (largely existing) discovery tools and advanced search methods to find instances of these patterns in the information space. This latter approach becomes essential because it can provide clues about terrorist sleeper cells made up of people who have never engaged in activity that would link them to known terrorists. Nonparticularized suspicion raises even higher the question of civil liberties, though—currently, no well-defined or practiced legal or regulatory protocols govern its operation, so a new privacy policy framework for management and oversight is needed (we'll briefly discuss this later).

With respect to false positives, some of our critics have stated that pattern-based queries create more false positives than they help resolve. Dealing with false positives—which are a legitimate concern given that the government might get it wrong and stigmatize or inconvenience nonterrorists—requires pattern-based queries to be issued iteratively in a privacy-sensitive manner (specifically, via anonymization and selective revelation techniques). Handling them also requires multiple stages of human-driven analysis in which analysts can't act on the results of such queries until a third-party legal authority has established sufficient probable cause. Analysts would refine queries in stages, seeking to gain more confirmation while invoking numerous privacy-protection techniques in the process. This isn't unlike the tried and proven signal-processing analysis techniques found in antisubmarine warfare, in which human-driven analysis addresses false positives at various stages in a similar manner.[8]

Safeguarding civil liberties

Americans expect their government to protect them from enemy attack as well as safeguard (or at least not violate) their civil liberties and privacy. We believe these two ideals aren't mutually exclusive: Figure 5 shows how our goal (and challenge) is to maximize security at an acceptable level of privacy. In other words, we can pick acceptable levels of privacy and through the development and use of technology, create new level of privacy versus security curves, thus increasing security. A full discussion of what privacy means from a legal and regulatory context is beyond this article's scope, but for a working definition, we would argue that personal privacy is only violated if the violated party suffers some tangible loss, such as unwarranted arrest or detention, for example. The right balance between the two must be understood, as well as the corresponding social costs, benefits, and roles played by the public, government, and private sectors.

As discussed earlier, analysts must systematically use information technologies to detect and discover instances of known or emerging terrorist signatures, but they must also be able to exploit the permitted information sources they need to access and do so while protecting the privacy of nonterrorists. Privacy-protection technology is a key part of the solution not only to protect privacy but also to encourage the intelligence, law enforcement, and counterterrorism communities to share data without fear of compromising sources and methods. However, the American public has legitimate concerns about whether protections for privacy are adequate to address the potential negative consequences of increased government use of permitted information sources. These concerns are heightened because there is little understanding or knowledge about how the government might use this data.

The R&D community has explored several promising privacy-protection technologies, especially those that are most relevant to the pattern-based query approach. We briefly describe some of them here, but more detailed information appears elsewhere.[9–11]

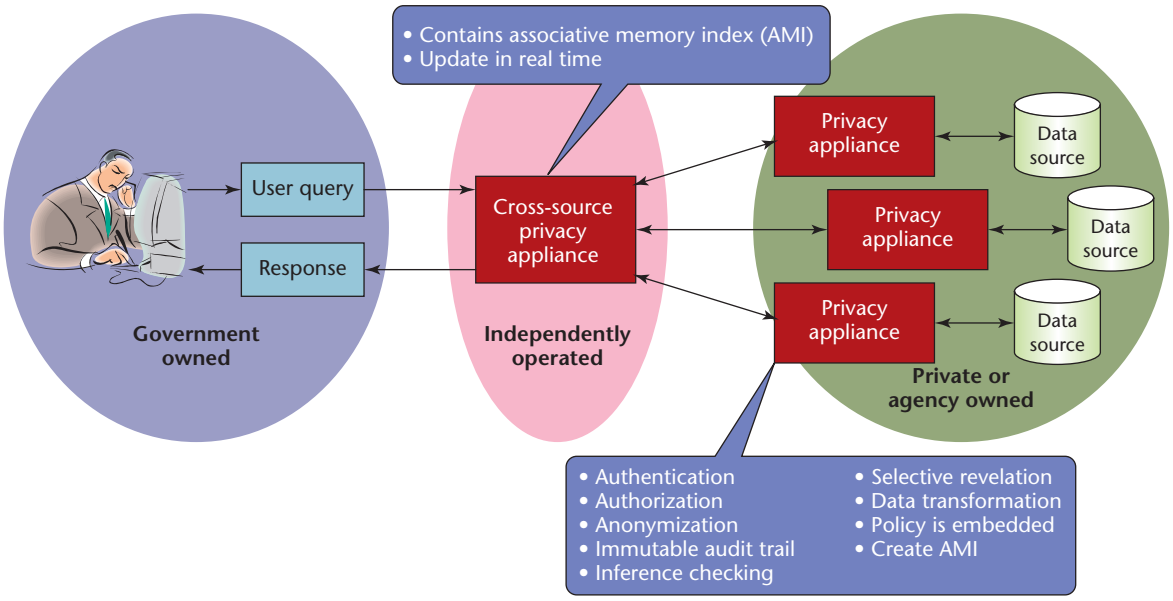

Figure 6. Privacy appliance concept. A tamper-resistant, cryptographically protected device serves as a trusted privacy-enforcing guard between the user and the database.

Privacy appliance

Our privacy appliance concept involves the use of a separate tamper-resistant, cryptographically protected device placed on top of databases. The appliance would be a trusted, guarded interface between the user and the database analogous to a firewall, smart proxy, or a Web accelerator. It would implement several privacy functions and accounting policies to enforce access rules established between the database owner and the user. It would also explicitly publish the details of its technology, verify the user's access permissions and credentials (packaged with the query in terms of specific legal and policy authorities), and filter out queries not permitted or that illegally violate privacy. Finally, it would create an immutable audit log that captures the user's activity and transmits it to an appropriate trusted third-party oversight authority to ensure that abuses are detected, stopped, and reported. (Granted, our privacy appliance concept assumes the third party is trusted, which is often the hardest problem to solve.) The privacy appliance's operation must be automated to respond to the dynamic, time-sensitive nature and scale of the problem and to ensure the privacy policy's implementation. Figure 6 illustrates the privacy appliance concept in terms of some of its key privacy functions as well as how it would work operationally.

Data transformation

Used within the privacy appliance, data transformation employs well-known mathematical encoding techniques to transform data from a plaintext representation to cipher, thus making computer processing more efficient and the data unintelligible to humans. Once transformed, analysts could apply a plethora of data analysis functions to understand the data's significance, keeping the identities of subjects hidden from analysts but still allowing the detection of terrorist activity patterns, such as data searching, alias and entity resolution, and pattern-query matching. Because the data is represented in unintelligible cipher, no personally identifiable data is disclosed to the analyst, thus privacy protection is maintained.

Anonymization

Similar to data transformation, anonymization is a technique used within the privacy appliance: it generalizes or obfuscates data, providing the system with a guarantee that any personally identifiable information in the released data can't be determined, yet the data still remains useful from an analytical viewpoint. As an example, instead of releasing to an analyst a database record consisting of [name(first, last); telephone #(area code, exchange, line number); address(street, town, state, zip code)], an anonymized version of this database record could be [name(first); telephone #(area code); address(state)]. For this approach to work, analysts will have to make connections between queries and thus will require some sort of anonymized unique identifier as well. Much more thorough treatment of various anonymization techniques and applications for privacy appears elsewhere.[10,11]

Selective revelation

Another technique to employ in the privacy appliance is selective revelation, which gives incremental access to and analysis of increasingly personally identifiable data. In this approach, what an analyst gets back in response to a pattern-based query varies in depth and specificity depending on the analyst, the investigation's status, and other criteria. The analyst's knowledge of an individual's identity would occur only after a sufficient level of suspicion and appropriate legal threshold were met. The approach proceeds incrementally by requiring data owners to release subsets of data—anonymized, filtered, or statistically characterized—to an analyst's pattern-based query. Initially, no personally identifiable data is provided in response to the query. If the results turn out to be meaningful after iteration and refined patterns or queries—say, only an acceptably few individuals match the query, or the level of suspicion or probable cause has been heightened—then additional permissions and authorization through an appropriate (yet currently nonexistent) legal framework would need to be secured to release personally identifiable data of individuals under suspicion.

Immutable audit

Another technique to be used in the privacy appliance is an immutable audit, which automatically and permanently records all accesses to data, with no possibility of undetected alteration or tampering. To prevent potential abuses by malicious agents, audit logs would be designed so that any misdeeds or corruption are detected with the highest probability. Audit logs would be cryptographically protected and transmitted to a trusted third-party oversight authority. Privacy tools to query and analyze audit logs are also critical. The contents of the audit log could contain fields such as the analyst's identity and credentials, the authorizations and permissions allowed, the date and time of the data access, the data requested, and the data returned.

Self-reporting data

An important technology that isn't directly related to the privacy appliance but is important from a civil liberties perspective is self-reporting data. This is a method for truth maintenance as well as for reporting on the data's distribution. Data used in analysis should be active (that is, it should report back to a central authority about where it is and for what it's being used); this point is essential to correct any information that's later proved to be false.

Privacy laws

Government access to personally identifiable data raises legitimate concerns about the protection of civil liberties, privacy, and due process. Given the limited applicability of current privacy laws to the modern digital era, practical policies for new information technology use, redress, and oversight are vital. New privacy policy can help ensure that controls and protections accompany the use of the information technologies we discussed earlier. Here, we list several basic principles as examples of the types of policy to consider:[12]

- *Neutrality*. New information technology should build in existing legal and policy limitations about access to personally identifiable or third-party data.
- *Minimize intrusiveness*. Personally identifiable data is voluntary but might be required as a condition of service (such as driver's licenses), thus it should be anonymized or rendered pseudononymous and disaggregated (when possible).
- *Intermediate not ultimate consequence*. Personal identification by a new information technology shouldn't directly lead to ultimate consequence (such as arrest); instead, analysts should view it as cause for additional investigation.
- *Audits and oversight*. New information technology should have strong built-in technological safeguards such as audit and oversight mechanisms to detect and deter abuse.
- *Accountability*. New information technology should be used in a manner that ensures accountability of the executive branch to the legislative branch for its use.
- *Necessity of redress mechanisms*. Robust legal mechanisms for the correction of false positives should be in place.
- *People and policy*. Internal policy controls, training, administrative oversight, enhanced congressional oversight, and civil and criminal penalties for abuse should all be in place.

We hope these considerations will be taken into account along with legal protocols for pattern-based searches; technology does and can play a key role in the careful balance of security with privacy.

Information and privacy-protection technologies are powerful tools for counterterrorism, but it's a mistake to view technology as the complete solution to the problem. Rather, the solution is a product of the whole system—the people, culture, policy, process, and technology. Technological tools can help analysts do their jobs better, automate some functions that analysts would otherwise have to perform manually, and even do some early sorting of masses of data. But in the complex world of counterterrorism, the technologies alone aren't likely to be the only source for a conclusion or decision.

Ultimately, the goal should be to understand the level of improvement possible in our counterterrorism operations using advanced tools such as those described here but also to consider their impact—if any—on privacy. If research shows that a significant improvement to detect

and preempt terrorism is possible while still protecting the privacy of nonterrorists, then it's up to the government and the public to decide whether to change existing laws and policies. However, research is critical to prove the value (and limits) of this work, so it's unrealistic to draw conclusions about its outcomes prior to R&D completion. As has been reported,[6] research and development continues on information technologies to improve national security; encouragingly, the Office of the Director of National Intelligence (ODNI) is embarking on an R&D program to address many of the concerns raised about potential privacy infringements. □

Acknowledgments

The views expressed herein are the authors' alone and don't reflect the views of any private-sector or governmental entity.

References

1. R. Popp and J. Yen, eds., *Emergent Information Technologies and Enabling Policies for Counter-Terrorism*, Wiley & Sons/IEEE Press, 2006.
2. *Report to Congress Regarding the Terrorism Information Awareness Program*, DARPA, May 2003; [response to Consolidated Appropriations Resolution, Pub. L. no.108-7, div. M, sec. 111(b), 2003].
3. J. Poindexter, "Overview of the Information Awareness Office," *DARPATech 2002*, DARPA, 2002; www.fas.org/irp/agency/dod/poindexter.html.
4. R. Popp et al., "Countering Terrorism through Information Technology," *Comm. ACM*, vol. 47, no. 3, 2004, pp. 36–43.
5. E. Jonietz, "Total Information Overload," *MIT Tech. Rev.*, vol. 106, no. 6, 2003, p. 68.
6. S. Harris, "Signals and Noise," *Nat'l J.*, vol. 38, no. 24, 2006, pp. 50–58.
7. M. DeRosa, *Data Mining and Data Analysis for Counterterrorism*, CSIS Press, 2004.
8. T. Senator, "Multi-Stage Classification," *Proc. 5th IEEE Int'l Conf. Data Mining* (ICDM 05), IEEE CS Press, 2005, pp. 386–393.
9. "Security with Privacy," DARPA's Information Systems Advanced Technology (ISAT) study, Dec. 2002; www.cs.berkeley.edu/~tygar/papers/ISAT-final-briefing.pdf.
10. L. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," *J. Law, Medicine and Ethics*, vol. 25, nos. 2-3, 1997, pp. 98–110.
11. L. Sweeney, "*k*-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, 2002, pp. 557–570.
12. P. Rosenzweig, "Privacy and Consequences: Legal and Policy Structures for Implementing New Counter-Terrorism Technologies and Protecting Civil Liberty," *Emergent Information Technologies and Enabling Policies for Counter-Terrorism*, R. Popp and J. Yen, eds., Wiley & Sons/IEEE Press, 2006, pp. 421–438.

***Robert Popp**, now CEO of National Security Innovations, recently served as a senior government executive within the US Department of Defense as deputy director of the Information Awareness Office (IAO) at DARPA and assistant deputy undersecretary of defense for advanced systems and concepts in the Office of the Secretary of Defense. He serves on the Defense Science Board, is a senior associate for the Center for Strategic and International Studies, and is an associate editor of* IEEE Transactions on Systems, Man, and Cybernetics. *Popp holds two patents, authored numerous journal and conference papers, and edited* Emergent Information Technologies and Enabling Policies for Counter-Terrorism *(Wiley & Sons/IEEE Press, 2006). Popp has a PhD in electrical engineering from the University of Connecticut, and a BA/MA in computer science from Boston University. Contact him at rpopp@nationalsecurityinnovations.com.*

***John Poindexter**, now a private consultant, most recently served as director of the Information Awareness Office (IAO) at DARPA. He also serves on the board of directors for Saffron Technology, a computer software company that produces associative memory applications. Prior to working at DARPA, Poindexter served as National Security Advisor and Deputy National Security Advisor under President Ronald Reagan from 1983 to 1986. He has a PhD and an MS in physics, both from the California Institute of Technology. Contact him at john@jmpconsultant.com.*

Threat and Fraud Intelligence, Las Vegas Style

Matching and relating identities is of the utmost importance for Las Vegas casinos. The author describes a specific matching technique known as identity resolution. This approach provides superior results over traditional identity matching systems.

Jeff Jonas
IBM

Las Vegas, Nevada, is possibly the most interesting real-world setting for a high-stakes game of data surveillance. Most of the 38 million people who visit the city annually are attracted by the gambling, entertainment, shopping, architecture, dining, and shows.[1] However, among them are a few thousand "opportunists" who converge on Las Vegas solely to exploit its vulnerabilities. Some have become so infamous that gaming regulators have banned them from ever again stepping foot in a Nevada casino. In fact, if a casino gets caught doing business with such a person, it can be heavily fined or, worse, lose its gaming license.

If you're a casino operator, knowing with whom you're doing business isn't just good business in terms of protecting corporate assets—it's a matter of legal responsibility. Finding a few bad actors, while minimizing the disruption, inconvenience, and privacy invasions to tens of millions of innocent tourists, has by necessity grown from an art mastered by a few practitioners into a teachable discipline. Elements of that discipline include regulatory policy, industry best practice procedures, staff development, and information technology.

This article presents the general problem domain of matching and relating identities, examines traditional approaches to the problem, and introduces identity resolution and relationship awareness. This combination offers improved accuracy, scalability, and sustainability over traditional methods.

Enterprise surveillance

Before the age of electronic surveillance, casino security personnel peered out from behind one-way mirrors with binoculars while standing high above the casino floor on specially constructed catwalks.

Casinos began replacing these catwalks with electronic surveillance cameras in the early 1980s, and camera usage and placement were soon codified by state law.[2] Prior to the 1990s, casinos also hired people to compare casino and hotel guest lists against watch lists containing *subjects of interest*—that is, both crooks and highly desired customers the casinos wanted to pamper. But as visitor volume grew, everything else did as well. Las Vegas hotels soon had 3,000 or more rooms—at times, more than 100,000 people a day make their way through a mega-resort; currently, 18 of the 20 largest hotels in the US are in Las Vegas (www.airhighways.com/las_vegas.htm).

Today, casino operators rely on automated systems to focus the casino's finite surveillance and investigatory resources. Just as casinos use perimeter surveillance systems to ensure that no one makes his or her way into the Mirage hotel's volcano fire spectacular, information-based systems monitor for subjects of interest who might be engaged in inappropriate transactions on the casino's premises. These technologies also watch for fraud and "insider threats" (when employees secretly work against the enterprise's interests)—an especially dangerous scenario in a business where a single corrupt dealer can cost the casino US$250,000 in 15 minutes (if a dealer lets a player use a "pre-ordered" deck in a table game, for example).

For the gaming industry, subjects of interest come from several sources:

- *Gaming regulatory compliance.* Nevada gaming regulators publish an exclusionary list of the individuals banned

via statute from transacting with casinos under penalty of license revocation or fines.[3]

- *Federal regulatory compliance*. The US Department of State Office of Foreign Assets Control (OFAC) publishes a list of specially designated nationals that describes the countries, individuals, and organizations banned by various federal statutes from transacting with US businesses (www.ustreas.gov/offices/enforcement/ofac/).
- *Legally barred*. Casinos can "formally trespass" a party, after which the person will be arrested if he or she returns.
- *Convicted cheaters*. Many casinos consider those people previously arrested for felonious acts against the gaming industry to be a potential risk that warrants additional levels of scrutiny. Sometimes a patron is identified as a former gaming felon and permitted to play, as any decision to act involves human oversight and consideration of all available facts (for example, the crime occurred many years ago and involved slots, but this individual is playing blackjack today).
- *Suspected cheaters and card counters*. Although card counting isn't illegal, casinos have the right to prevent anyone from engaging in gaming activity, which is likely if the player is determined to have a technique that materially changes the natural odds of the game. Many casinos subscribe to one or more subscription services that facilitate information sharing (within the gaming industry) of gaming arrests, individuals suspected of illegally manipulating games, and card counting.
- *Self-declared problem gamblers*. Casino patrons can place themselves on a voluntary self-exclusion list as a "problem gambler," after which the casino inherits a degree of responsibility to neither market to nor allow the person to engage in casino activity. Problem gamblers have sued casinos when they inadvertently allowed such people to accrue additional losses (www.americangaming.org/publications/rglu_detail.cfv?id=229 and www.casinocitytimes.com/news/article.cfm?contentId=160709), although such suits are rarely successful.

Like any large organization, casinos have many disparate information systems, each with their own data sets. These include systems concerned with hotel reservations, hotel property management, customer loyalty, credit, point of sale, human resource job applicants, human resource employment, and vendors, to name a few. Arguably, system data that has a nexus with a subject of interest would constitute a degree of risk to the enterprise and might warrant additional scrutiny. Whether a known cheater has just rented a room, joined the loyalty club, or applied for a job, management appreciates being notified.

In many cases, however, relationships are nonobvious. Traditional practices, for example, help casinos catch and detain roulette cheaters (in this case, the surveillance-room operator simply observes an illegal activity while spot-checking a game). Although the dealer can claim embarrassment for missing a blatant scam, gaming regulators take the fact that the dealer lives in the same apartment unit as the cheater as evidence of collusion, so both are arrested. Or consider a promotions manager who "randomly" selects a ticket for a prize drawing and congratulates the winner of a new car; the recipient has a different last name, but she is, in fact, the manager's sister. This is evidenced again by common information between the manager's employment data and the winner's self-provided information. Traditional analysis might not discover these connections.

Detecting fraud becomes harder as casinos and their information systems grow in complexity. Let's look at three hard-to-detect scenarios from the gaming industry that illustrate this.

Las Vegas problem scenario #1

An individual barred by gaming regulators from transacting with casinos has just enrolled in your slot club, using a slightly different name and a date of birth in which the month and day are transposed. He's now playing in your casino, which places your gaming license at risk. How would you know?

Las Vegas problem scenario #2

An employee who works in surveillance has just put in for an address change in your payroll system. This same address is consistent with that of an individual arrested early last year for a $375,000 baccarat scam and now serving time in jail. You had always suspected help from the inside but had no evidence. This latest piece of data in the payroll system would be an important lead, but how would you ever discover this important new fact?

Las Vegas problem scenario #3

Your marketing team buys a list for a new direct mail campaign. It has been scrubbed of problem gamblers who have voluntarily placed themselves on the self-exclusionary list. You send the remaining people on the list a promotional offer for the upcoming New Year's Eve event. In the months between when the promotional offer mails and New Year's Eve, two recipients place themselves on the self-exclusionary list. Can you detect them before they arrive? Can you detect them when they arrive?

Detecting fraud becomes harder as casino systems grow in complexity. Traditional analysis might not discover nonobvious relationships.

Problems with scenarios

These scenarios are very difficult to discover, especially manually, by humans, even though the enterprise contains all the necessary evidence. That's because this evidence is trapped across isolated operational systems, and although these three problem scenarios clearly involve identity matching, traditional matching algorithms address more mundane missions such as cleaning up customer mailing lists, detecting duplicate enrollment in loyalty-club programs, or detecting the arrival of a high roller who didn't contact his host. Traditional algorithms aren't well suited to these scenarios—especially problem scenarios two and three.

Data from operational business systems is plagued by both intentional errors and legitimate natural variability.

Matching is further hampered by the poor quality of the underlying data. Lists containing subjects of interest commonly have typographical errors. Data from operational business systems is plagued by both intentional errors (those who intentionally misspell their names to frustrate data matching efforts), and legitimate natural variability (Bob versus Robert and 123 Main Street versus 123 S. Maine Street).

International data complicates matters further still. In 2005, 12 percent of tourists who visited Las Vegas came from abroad.[1] However, data entry operators (and programmers) might not know how to handle international names—for example, the name حاج محمد عثمان عبد الرقيب might be entered as "Haj Imhemed Otmane Abderaqib" in West Africa or "Hajj Mohamed Uthman Abd Al Ragib" in Iraq: both English spellings signify the same individual.

Dates are often a problem as well. Months and days are sometimes transposed, especially in international settings. Numbers often have transposition errors or might have been entered with a different number of leading zeros.

Naïve identity matching

Organizations typically employ three general types of identity matching systems:

- *Merge/purge* and *match/merge*. Direct marketing organizations developed these systems to eliminate duplicate customer records in mailing lists. These systems generally operate on data in batches; when organizations need a new de-duplicated list, they run the process again from scratch.
- *"Binary" matching engines*. This system tests an identity in one data set for its presence in a second data set. These matching engines are also sometimes used to compare one identity with another single identity (versus a list of possibilities), with the output often expected to be a confidence value pertaining to the likelihood that the two identity records are the same. These systems were designed to help organizations recognize individuals with whom they had previously done business (the recognition becomes apparent during certain transactions, like checking into the hotel) or, alternatively, recognize that the identity under evaluation is known as a subject of interest—that is, on a watch list—thus warranting special handling. This type of identity matching system can be batch-handled or conducted in real time, although real time is typically preferred.
- *Centralized identity catalogues*. These systems collect identity data from disparate and heterogeneous data sources and assemble it into unique identities, while retaining pointers to the original data source and record with the purpose of creating an index. Such systems help users locate enterprise content much in the same way the library's card catalog helps people locate books.

Each of the three types of identity matching systems uses either probabilistic or deterministic matching algorithms. *Probabilistic techniques* rely on training data sets to compute attribute distribution and frequency. Mark is a common first name, for example, but Rody is rare. These statistics are stored and used later to determine confidence levels in record matching. As a result, any record containing simply the name Rody and a residence in Maine might be considered the same person with a high degree of probability. These systems lose accuracy when the underlying data's statistics deviate from the original training set. To remedy this situation, such systems must be retrained from time to time and then all the data reprocessed.

Deterministic techniques rely on pre-coded expert rules to define when records should be matched. One rule might be that if the names are close (Robert versus Rob) and the social security numbers are the same, the system should consider the records as matching identities. These systems fail—sometimes spectacularly—when the rules are no longer appropriate for the data being collected.

Nonobvious relationship awareness

Nonobvious relationship awareness (NORA) is a system that Systems Research and Development of Nevada (which I founded) developed specifically to solve Las Vegas casinos' identity matching problems. It ran on a single server, accepted data feeds from numerous enterprise information systems, and built a model of identities and relationships between identities (such

as shared addresses or phone numbers) in real time. If a new identity matched or related to another identity in a manner that warranted human scrutiny (based on basic rules, such as good guy connected to very bad guy), the system would immediately generate an intelligence alert.

Requirements for the system became ambitious:

- *Sequence neutrality*. It needed to react to new data as that data loaded. Matches and nonmatches had to be automatically re-evaluated to see if the matches were still probable as the new data loaded. This capability was designed to eliminate the necessity of database reloads. (See http://jeffjonas.typepad.com/jeff_jonas/2006/01/sequence_neutra.html for more on sequence neutrality).
- *Relationship aware*. Relationship awareness was designed into the identity resolution process so that newly discovered relationships could generate realtime intelligence. Discovered relationships also persisted in the database, which is essential to generate alerts to beyond one degree of separation.
- *Perpetual analytics*. When the system discovered something of relevance during the identity matching process, it had to publish an alert in real time to secondary systems or users before the opportunity to act was lost.
- *Context accumulation*. Identity resolution algorithms evaluate incoming records against fully constructed identities, which are made up of the accumulated attributes of all prior records. This technique enabled new records to match to known identities *in toto*, rather than relying on binary matching that could only match records in pairs. Context accumulation improved accuracy and greatly improved the handling of low-fidelity data that might otherwise have been left as a large collection of unmatched orphan records.
- *Extensible*. The system needed to accept new data sources and new attributes through the modification of configuration files, without requiring that the system be taken offline.
- *Knowledge-based name evaluations*. The system needed detailed name evaluation algorithms for high-accuracy name matching. Ideally, the algorithms would be based on actual names taken from all over the world and developed into statistical models to determine how and how often each name occurred in its variant form. This empirical approach required that the system be able to automatically determine the culture that the name most likely came from because names vary in predictable ways depending on their cultural origin.
- *Real time*. The system had to handle additions, changes, and deletions from real-time operational business systems. Processing times are so fast that matching results and accompanying intelligence (such as if the person is on a watch list or the address is missing an apartment number based on prior observations) could be returned to the operational systems in sub-seconds.
- *Scalable*. The system had to be able to process records on a standard transaction server, adding information to a repository that holds tens of millions of identities.

Following IBM's acquisition of the company, NORA's underlying code base was improved and the technology renamed IBM Identity Resolution and Relationship Resolution. Today, the system is implemented in C++ on top of an industry-standard SQL database with a Web services interface and optional plug-ins (knowledge-based name libraries, postal base files, and so on). Data integration services transform operational data into prescribed and uniform XML documents. Configuration files specify the deterministic matching rules and probabilistic-emulating thresholds, relationship scoring, and conditions under which to issue intelligence alerts.

Although the gaming industry has relatively low daily transactional volumes, the identity resolution engine is capable of performing in real time against extraordinary data volumes. The gaming industry's requirements of less than 1 million affected records a day means that a typical installation might involve a single Intel-based server and any one of several leading SQL database engines. But performance testing has demonstrated that the system can handle multibillion-row databases consisting of hundreds of millions of constructed identities and ingest new identities at a rate of more than 2,000 identity resolutions per second; such ultra-large deployments require 64 or more CPUs and multiple terabytes of storage, and move the performance bottleneck from the analytic engine to the database engine itself.

Identity resolution

Identity resolution is designed to assemble i identity records from j data sources into k constructed, persistent identities. The term "persistent" indicates that matching outcomes are physically stored in a database at the moment a match is computed. Think of each constructed identity as a bag filled with all the observed attributes from specific source identity records from which that identity was originally derived. Thus, newly presented identity records are never evaluated against another single identity record (binary matching), but rather are evaluated against fully preconstructed identities (each one built from i previously resolved identity records).

Accurately evaluating the similarity of proper names is undoubtedly one of the most complex (and most important) elements of any identity matching system. Dictionary-based approaches—for example, mapping Bob and Robert to the same database key—fail to handle the complex ways in which names from different cultures can appear.

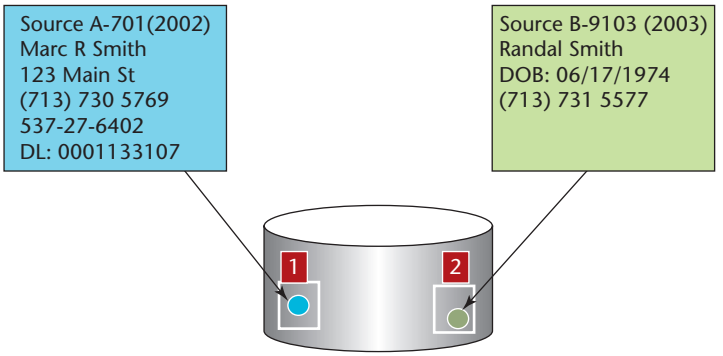

Figure 1. Records in an identity resolution system. These two records are initially determined not to be the same person.

The same is true for systems like Soundex, which is a phonetic algorithm for indexing names by their sound when pronounced in English. The basic aim is for names with the same pronunciation to be encoded to the same string so that matching can occur despite minor differences in spelling. Such systems' attempts to neutralize slight variations in name spelling by assigning some form of reduced "key" to a name (by eliminating vowels or eliminating double consonants) frequently fail because of external factors—for example, different fuzzy matching rules are needed for names from different cultures.

We found that the deterministic method is essential for eliminating dependence on training data sets. As such, the system no longer needed periodic reloads to account for statistical changes to the underlying universe of data. But we found many common conditions in which deterministic techniques failed—specifically, certain attributes were so overused that it made more sense to ignore them than to use them for identity matching and detecting relationships. For example, two people with the first name of "Rick" who share the same social security number are probably the same person—unless the number is 111-11-1111. Two people who have the same phone number probably live at the same address—unless that phone number is a travel agency's phone number. We refer to such values as *generic* because the overuse diminishes the usefulness of the value itself. It's impossible to know all of these generic values a priori—for one reason, they keep changing—thus probabilistic-like techniques are used to automatically detect and remember them.

Thus, our deployed identity resolution system uses a hybrid matching approach that combines deterministic expert rules with a probabilistic-like component to detect generics in real time (to avoid the drawback of training data sets). The result is expert rules that look something like this:

```
If the name is similar
  AND there is a matching unique
   identifier
  THEN match
  UNLESS this unique identifier is
generic
```

A unique identifier might include social security or credit-card numbers, or a passport number, but wouldn't include such values as phone number or date of birth. The term "generic" here means the value has become so widely used (across a predefined number of discreet identities) that we can no longer use this same value to disambiguate one identity from another.

The actual deterministic matching rules are much more elaborate in practice because they must explicitly address fuzzy matching (such as transposition errors in numbers, malformed addresses, and month and day reversal in dates of birth), contain rules to deal with conflicting attributes (such as the name, address, and phone number are the same, but the difference is the junior versus senior designations), and so on.

As an example of sequence-neutral behavior, let's consider an identity resolution system that encounters a record in 2002 for Marc R. Smith at 123 Main Street, phone number 713-730-5769, and driver's license number 0001133107. The following year, it encounters a second record for Randal Smith born 6/17/1974 phone number 713-731-5577. With no other information, the system concludes that these are two different people (see Figure 1).

Then, in 2004, the system encounters a new record for Marc Randy Smith, phone number 713-731-5577 and driver's license number 1133107. Because this 2004 record matches both the 2002 and the 2003 records, the system would collapse all three records into a single identity (see Figure 2).

In 2005, the system encounters a fourth record: Randy Smith Sr., born 6/17/1934, with the phone number 713-731-5577. The system now splits the records into two distinct identities: the 2003 and 2005 records are for Randy Smith Sr., the father, whereas the 2002 and 2004 records are for Mark Randy Smith, the son (see Figure 3).

Users of identity resolution in threat and fraud intelligence missions invariably choose to limit the conditions in which the system generates intelligence "alerts" because they're dealing with a finite number of investigative resources. Through personal communication, I learned that one such Las Vegas casino reported its alert criteria yields two "leads" a day on weekdays and five leads a day on weekends. Notably, a human analyst reviews all intelligence leads before action is taken. (Actions come in many forms; in gaming, for example, an action might involve mild additional scrutiny, more ex-

tensive surveillance, a physical conversation with the subject, a request for identification, removal from the premise, or in certain settings, handing a case over to authorities for an arrest.) Although misidentification is exceedingly rare, further analysis might conclude that the intelligence doesn't warrant any action. For example, a blackjack customer who has a 10-year-old felony slot conviction could be allowed to continue gambling at the casino's discretion.

An unexpected outcome from deployments of identity resolution in the gaming industry was the discovery of insider threats that previously hadn't been considered. The original intention was to discover subjects of interest attempting to transact with the casino, but because the design placed all identities in the same dataspace, the good guys (customers), bad guys (subjects of interest), and most trusted (employees and vendors) commingled as they were matched against each other. This resulted in exciting new kinds of alerts—such as an employee who was also a vendor.

Although matching accuracy is highly dependent on the available data, using the techniques described here achieves the goals of identity resolution, which essentially boil down to accuracy, scalability, and sustainability even in extremely large transactional environments (billions of records representing hundreds of millions of unique identities).

Relationship awareness

Detecting relationships is vastly simplified when a mechanism for doing so is physically embedded into the identity matching algorithm. Stating the obvious, before we can usefully analyze meaningful relationships, we should first resolve unique identities. As such, identity resolution must occur first. We discovered that it was computationally efficient to observe relationships at the moment the identity record is resolved because in-memory residual artifacts (which are required to match an identity) comprise a significant portion of what's needed to determine relevant relationships. Relevant relationships, much like matched identities, were then persisted in the same database.

A database supporting related identities essentially contains pairs of persistent keys—for example: entity ID #22 might be paired with entity ID #309 to denote that these two identities are in some manner related. Notably, some relationships are stronger than others; a relationship score that's assigned with each relationship pair captures this strength. Living at the same address three times over 10 years should yield a higher score than living at the same address once for three months.

As identities are matched and relationships detected, the system evaluates user-configurable rules to determine if any new insight warrants an alert being published as an intelligence alert to a specific system or user. One sim-

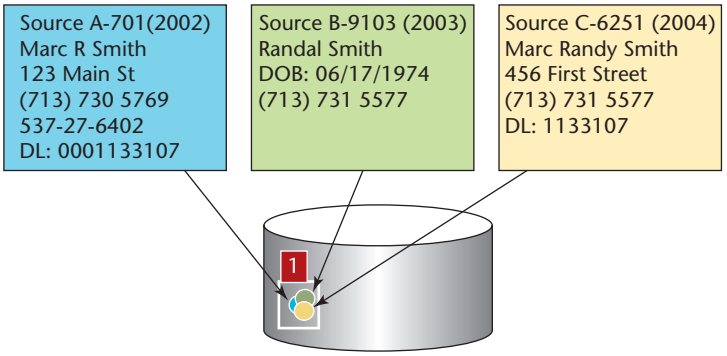

Figure 2. Records in an identity resolution system. The third record arriving in 2004 provides the evidence to conjoin the original two records—resulting in one entity, comprised of three records.

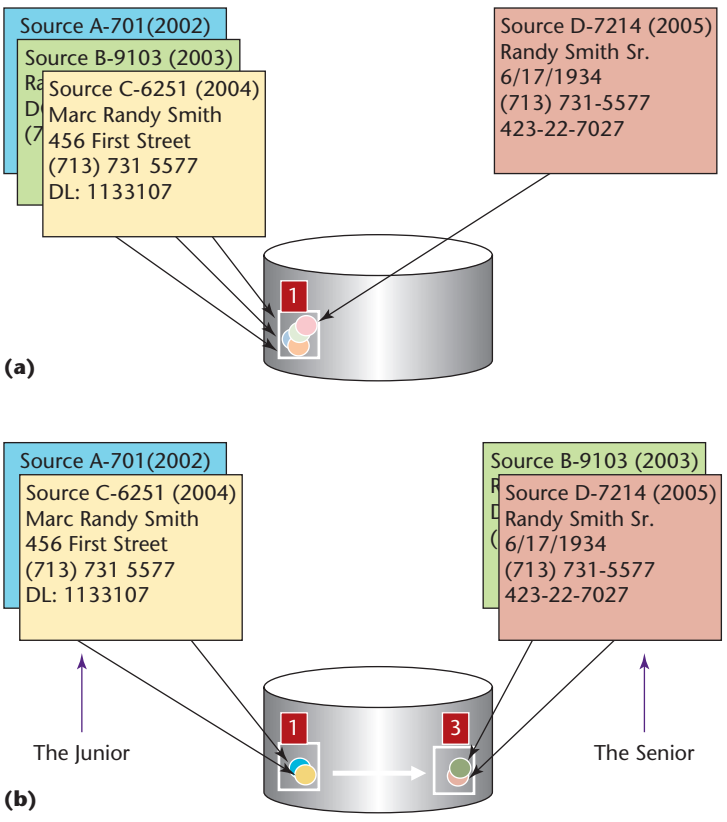

Figure 3. Records in an identity resolution system. (a) The 2005 record appears to match the existing identity. (b) However, the new information learned in the 2005 record contains evidence that there are really two discreet identities (a junior and a senior) and, as such, the records are decoupled to immediately correct for this new information.

plistic way to do this is via conflicting roles. A typical rule might be notification any time a role "employee" is associated to a role "bad guy," for example. In this case, *associated* might mean zero degrees of separation (they're the same person) or one degree of separation (they're roommates). Relationships are maintained in the database to one degree of separation; higher degrees are determined by walking the tree. Although the technology supports searching for any degree of separation between identities, higher orders include many insignificant leads and are thus less useful.

Today, when an employee updates his or her employment record, if this data reveals a relationship to a current or former criminal investigation, such intelligence is detected in real time and published to the appropriate party. Notably, this doesn't mean that any criminal activity has occurred or is imminent; rather, it simply helps focus finite investigatory resources.

Applications outside of gaming

Identity resolution has helped organizations deal with real-time identity awareness in many sectors, including retail, financial services, national security, and disaster response.

In the retail sector, organized retail theft (ORT) is a multibillion-dollar a year fraud problem. Ring leaders hire groups of people to steal select products such as razor blades, batteries, infant formula, and so on. This type of fraud is exceedingly hard to detect because the shoplifting incidents appear as one-se-two-se events, with no real way to see trends of organized activity. Several years ago, using the identity resolution technique with relationship awareness, individual shoplifting incidents from four retailers were processed. The system determined the number of unique people, creating a view of the number of unique parties involved in the shoplifting. During this activity, the system relied on common addresses to create relationships. The result was a report of shared criminal facilities—the first such view of its kind—identifying hundreds of addresses where more than one person had been arrested for stealing at more than one location. One such find was a "Fagan operation" (named after a character in Charles Dickens's *Oliver Twist*), in which an adult employed children to steal a particularly popular brand of jeans.

In 1998, the US government recognized how it could use such technology to help discover nonobvious relationships to identify potential criminal activity from within. Following 9/11, this same technology found its way into several national security missions.

The Hurricane Katrina disaster demonstrated yet another possible use of this technology. After the storm passed, more than 50 Web sites emerged to host the identities of the missing and the found. People identified as missing on one Web site were identified as found on another. Some people registered the same person countless times on a single Web site (sometimes with different name variations and sometimes just in desperation). The question, then, was how many unique people were actually reported missing, how many unique people were reported found, and how many of these people could be reunited if the identities matched? Working in partnership with several agencies within the Louisiana state government and led by the state's Office of Information Technology, many loved ones were reunited.

Achieving real-time enterprise awareness is vital to protecting corporate assets not to mention the integrity of the brand itself. The ability to handle real-time transactional data with sustained accuracy will continue to be of "front and center" importance as organizations seek competitive advantage.

However, as those with criminal intent become more sophisticated, so must organizations raise the bar in their ability to detect and preempt potentially damaging transactional activity. The identity resolution technology described here provides a growing number of industries with enterprise awareness accurate to the split second, to address the most pressing societal problems such as fraud detection and counterterrorism. IBM sells this technology as an off-the-shelf product. In more recent developments such a technique can now be performed using only anonymized data, which greatly reduces the risk of information leakage (unintended disclosure) and when implemented correctly can enhance privacy protections. I hope to publish a similar technical piece on this new technology in the coming year. ▢

References

1. "Top 25 Frequently Asked Questions," Las Vegas Convention and Visitors Authority Research Dept., Mar. 2006; www.lvcva.com/getfile/2005Top25Questions.pdf?fileID=106.
2. "Surveillance Standards for Nonrestricted Licensees," Nevada Gaming Commission and State Gaming Control Board, Nov. 2005; http://gaming.nv.gov/stats_regs/reg5_survel_stnds.pdf.
3. "Regulation 28, List of Excluded Persons," Nevada Gaming Commission and State Gaming Control Board, Feb. 2001; http://gaming.nv.gov/stats_regs/reg28.pdf.

Jeff Jonas *is chief scientist of the IBM Entity Analytic Solutions group and an IBM Distinguished Engineer. He is a member of the Markle Foundation Task Force on National Security in the Information Age and actively contributes his insights on privacy, technology, and homeland security to leading national think tanks, privacy advocacy groups, and policy research organizations, including the Center for Democracy and Technology, Heritage Foundation, and the Office of the Secretary of Defense Highlands Forum. Recently, Jonas was named as a senior advisor to the Center for Strategic and International Studies. Contact him via www.jeffjonas.typepad.com.*