techworld

A powerful grasp of the nonobvious

Door: Simson Garfinkel

Back when I used to sell security software for a living, one of the things that drove me crazy was the U.S. Treasury Department's Office of Foreign Assets Control and its Specially Designated Nationals List.

The OFAC SDN is a list of foreign individuals and organizations with which it is illegal for U.S. companies and individuals to do business. Say a salesman at my company got a phone call from the Army of the Republic of Ilirida. Instead of taking a credit card number and sending the army its software, we were supposed to call the OFAC hotline and file a report. That's because ARI is actually a little-known terrorist group in the Balkins.

All U.S. persons must comply with the OFAC regulations. In practice, this means that any person or organization that does business overseas is supposed to download a copy of the SDN on a regular basis and use it to cross-check new customers and business relations. While it's possible to manually search the SDN for every overseas transaction, many businesses now use software to automatically scan their transaction stream and compare it against the U.S. government's blacklist

But scanning the SDN is really just the beginning of what entity resolution systems can offer. Using systems that are now on the market, it's possible to scour your existing databases to detect potential fraud and insider abuse. In the future, these techniques might even be used for performing large-scale medical or social research.

Complying with blacklists like the SDN can be difficult because the data quality is often quite poor. While the SDN contains roughly 3,200 names of individuals, many of them appear multiple times with variant spellings. This may be because U.S. intelligence has been poor, or because a person has used several spellings in an effort to hide his identity. Some people are identified by name, city of residence, date of birth and passport number; others appear on the blacklist as just a name, with no additional identifying information. One such name is Foday Sankoh, the former leader of the rebel groups in Sierra Leone. Presumably, if someone named Foday Sankoh tried to buy a copy of my company's software, we would have to refuse the sale. On the other hand, Mr. Sankoh is listed on the SDN as being deceased (although no date is given). So perhaps we could have allowed the sale, assuming that Mr. Sankoh could prove that he was actually alive--and thus, presumably, a different Foday Sankoh than the deceased person on the government's list.

In addition to the SDN, many industries have their own lists that they too have to monitor. The gaming industry, for example, maintains a list of people who have been "banned for life" because of cheating or criminal involvement. Airlines have to match names against the government's "no-fly list." Many hotels, restaurants and even universities have their own lists of individuals who are no longer welcome because of some prior offense.

One of the best-known entity resolution systems was prototyped in the late 1990s by Jeff Jonas to help Las Vegas casinos find cheaters and criminals, especially those with inside connections. Originally called Non-Obvious Relationship Analysis, or NORA, the company was rebranded Entity Analytic Solutions in 2001 and sold to IBM in January 2005. I recently had a chance to speak with Jonas and also John Bliss, the company's privacy strategist. Together the two entrepreneurs do an excellent job explaining how their technology works and how it is possible for a company to perform these sophisticated data mining operations without compromising the privacy of its customers or employees.

Today the IBM Entity Analytic Solutions platform consists of three major modules. The Identity Resolution module determines if data elements in different databases actually represent the

same person or organization--that is, the same entity. The Relationship Resolution module looks for obvious and nonobvious relationships between different entities across all of the databases. Finally, an Anonymous Resolution module allows different organizations to see if they have an entity in common without compromising the names or other identifying information of anyone in their database. All of these systems perform their magic using data already in an organization's possession--no third-party data banks are needed.

One way that a casino in Las Vegas used this software was to look for suspicious or inappropriate relationships between the casino's players and its employees. If a particular player consistently wins when playing blackjack with a particular dealer, there's probably reason to be a little suspicious. If the player and the dealer live in the same apartment building, it's almost certainly time to investigate. This kind of relationship can normally be hard to deduce, because the player and employee addresses are almost certainly stored in different data banks--and probably in different forms. To get around that problem, the Relationship Resolution module builds a new database that summarizes the previously identified entity information from all of the organization's various sources.

The casino in question loaded up the software with its list of employees, vendors, slot club members, table games players, in-house arrests and known cheaters. It found 24 active players that were known cheaters, 23 players who had previously been arrested or involved in some kind of incident, 12 cases in which employees were themselves the player, 192 employees who had possible relationships with a vendor, and seven cases of employees who were the vendor--definitely a no-no.

In a similar test for a major banking organization, IBM loaded the Entity Analytic Solutions software with 100,000 customer records and 20,000 "bad guys" from World-Check, a company that tracks individuals and organizations that might pose risks to the financial industry. The database was then seeded with 1,387 fictional records containing data from 572 bad guys. The system found 97 percent of the bad guys--but it also found 127 previously unknown relationships that the bank then investigated. Talk about an effective demo!
IBM's Anonymous Resolution is based on the use of hash functions. A typical application might be to see if the same person is receiving aid from multiple organizations at the same time--presumably something that would be in violation of those organizations' rules. Instead of exchanging the actual names of the people receiving aid, the IBM system lets organizations exchange one-way cryptographic hashes of the names. The system preprocesses the names so that minor variations in spelling won't prevent a match, and it allows the organizations exchanging data to further protect the information with a cryptographic key. In theory, such a system could be used to perform a large-scale medical study based on "anonymized" data from hospitals, pharmacies and insurance companies.

But while this kind of anonymous resolution is relatively easy to understand, it has a potential problem: When there is a match, it's possible for the organizations involved to learn the identity of the matching individual by tracing back the matching hash. This may represent an unacceptable opportunity for personal information to leak in some cases.

An entity resolution system that doesn't suffer from this problem was developed by Carnegie Mellon University professor Latanya Sweeney to track clients of domestic violence homeless shelters. This system (which Dr. Sweeney presented at a recent workshop in Cambridge that I organized) uses a special encryption cipher that allows each homeless shelter to contribute information to an encrypted value without being able to decode the information contributed by the other shelters. As a result, it's possible for all of the shelters in a network to determine the number of individuals that have visited more than one shelter, but it isn't possible to tell who visited there.

Properly implemented, entity resolution systems are a powerful tool for organizations to manage the flow and use of information about people involved with an enterprise. But these systems need to be carefully designed and audited.

For example, the existence of a possible relationship in a data bank does not imply the existence of intent--in fact, it may not even be a real relationship. Instead, it may be the result of an error in one of the source data banks. That's why it's necessary to keep pedigree of every single information element within the system and have provisions for automatically updating the entire index whenever a data source is refreshed.

Likewise--as with so many things in the security world--it is important to avoid using these systems as a substitute for human judgment. "We produce a list," says Jonas, who is now an IBM Distinguished Engineer, but it's up to human beings to review that list and determine whether the relationship actually deserves further attention, or if it is simply a chance occurrence.

Simson Garfinkel is researching computer forensics and human thought at Harvard University. Send feedback to machineshop@cxo.com