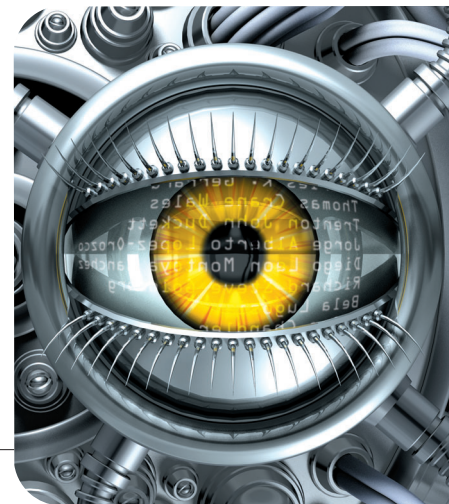


Threat and Fraud Intelligence, Las Vegas Style

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.casino.citytimes.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 promo-

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

tional 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?

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 evi-

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

dence 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.

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 Abderragib” 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.

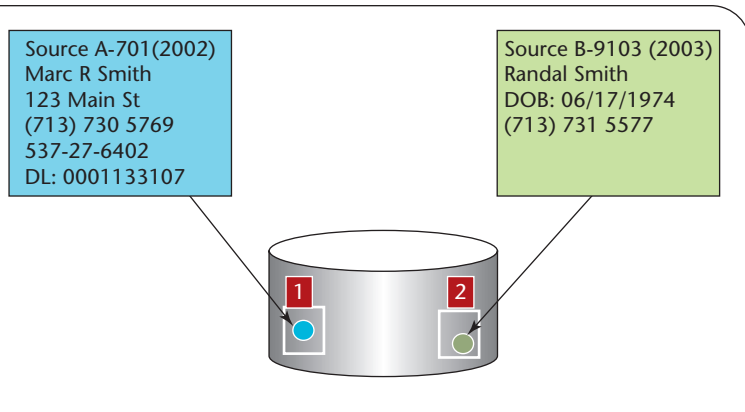


Figure 1. Records in an identity resolution system. These two records are initially determined not to be the same person.

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-

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 count-

less 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.