August, 2011

Analysis of Digital Financial Data

Robert L Kardell
Over the past 15 to 20 years, the amount of digital financial data that investigators receive in response to subpoenas, seize during the execution of a search warrant, or collect during an arrest has risen dramatically. Financial data quickly can overwhelm investigators as the number of transactions may run into the thousands for individuals or even into the millions for businesses. Culling through the data and identifying relevant transactions can become a daunting task, especially for law enforcement professionals investigating allegations of fraud on an unfamiliar proprietary bookkeeping system.

Fraud investigators inside and outside the government have developed methods to quickly identify suspicious financial transactions. Determining the most useful techniques for a particular investigation depends on two factors: the amount and type of data received. Based on these two characteristics, the range of successful techniques will differ for each case. The author reviews some of the tools and methods an investigator can use to conduct investigative analysis and identify suspicious financial transactions.¹
DIGITAL ANALYSIS 
TOOLS

While many computers come equipped with standard office software, numerous commercial products allow investigators to analyze financial data with the techniques presented below. Each has limitations of some sort, either with price or usability, but, ultimately, the choice depends on the user’s preferences and knowledge.

Spreadsheets
The most common method to cull through data and identify suspicious transactions is to load data into spreadsheet programs, for which numerous commercial products exist. Legacy versions of popular commercial software were limited to approximately 65,000 rows of data for each spreadsheet in a workbook, but the newest versions can handle many more rows of data. In the past, for data sets with more than 65,000 rows of data, investigators had to split the information into multiple sheets or enter it into a database. Although databases certainly can conduct more complex and faster analysis than spreadsheets, the same techniques can yield equally effective results in either type of program.

Databases
Working with databases may prove much more complicated than working with spreadsheets. However, using databases also can be extremely fast, even with large data sets. Many reasonably priced or free servers exist, and these systems provide a solution for investigators who have a large data set or multiple users who need to access it simultaneously. Each system can process several million rows of transactions and other data.

Hybrid Applications
Some applications used for accounting/forensic analysis combine aspects of database and spreadsheet programs. One such commercial hybrid application, though originally designed for financial audits, can conduct accounting/forensic or fraud analysis very efficiently. The application fuses characteristics of a spreadsheet and a database; its format resembles a spreadsheet, but it can handle an amount of data comparable to a database. It can create, import, edit, and even link multiple data tables to each other. Subsets of tables easily can be created from the results of searches and queries. The program comes with a number of built-in formulae to assist with analysis, and it can process unique scripting language to automate routine analytical tasks.

METHODS TO IDENTIFY UNUSUAL TRANSACTIONS

Data Assessment and Technique Identification
To identify unusual transactions, investigators should begin by reviewing all of the data and the transaction records that they possess because the characteristics of the data completely

“While many computers come equipped with standard office software, numerous commercial products allow investigators to analyze financial data....”

Special Agent Kardell, a certified public accountant, serves in the Terrorist Financing Operations Section, FBI’s Counterterrorism Division.
determine the analyses that investigators can perform. First, investigators should brainstorm potential clues and other insights that they might glean from each of the data fields.

In one case, for example, investigators wanted to examine a company’s personnel records for fictitious employees. Fictitious employees provide a great tactic for fraudsters or companies to hide embezzlement, payments to foreign officials, or otherwise nonreimbursable expenses.

The company provided investigators a list of its employees’ social security numbers (SSNs). Initially, however, the investigators did not see how they could analyze SSNs alone to expose any potentially fictitious employees. After researching the issue, discussing it with others, and reviewing possible fraud techniques, the team developed a method to determine if any of the SSNs were invalid or previously assigned. The Social Security Administration (SSA) provides guides to assess the validity of SSNs, as well as access to the Master Death Index to determine if a number belongs to a now-deceased individual. Using these tools, investigators discovered that many of the alleged employees’ SSNs were fraudulent. Currently, this method is used routinely with large lists of SSNs to determine if the numbers are legitimate. Unfortunately, when the SSA begins to issue random identification numbers in 2011 to deter fraud, this technique’s effectiveness may be compromised.

Data Standardization

Some data inputs follow a tightly controlled, standardized process. But, when entries are free text, input operators can enter whatever information they want in any number of formats. Standardized versus free-text inputs can affect how investigators analyze and draw conclusions from the data. Generally, standardized inputs lead to easy analysis and comparisons, whereas nonstandardized data sets can lead to incorrect or incomplete conclusions and analysis. This issue becomes more complicated when comparing multiple data sets in different formats.

Financial data quickly can overwhelm investigators as the number of transactions may run into the thousands for individuals or even into the millions for businesses.

Standardization refers to the process of manipulating a data set so that it conforms to a certain standard. This helps the investigator draw more accurate comparisons and more complete conclusions. A comparison of the following names provides an example:

- Robert L. Kardell
- Rob Kardell
- Robert Kardell
- Bob Kardell

If an investigator searched “Robert Kardell” across a large data set, those entries would not show as duplicates; but, after standardizing the data set (such as by separating the first and last names into different data fields), a search would identify identical last names. To standardize the data and improve the comparison even further, the investigator could highlight nicknames, such as “Bob” or “Rob,” and replace them with a formal name, such as “Robert.”

Investigators can standardize addresses (e.g., BLVD versus Boulevard and ST versus Street), dates (e.g., MM/DD/YYYY versus DD/MM/YYYY), quantitative amounts (e.g., $10,000.00 versus 10000.00 USD), and phone numbers. Because most data can be presented in several alternate forms, standard formats allow for more accurate analysis. The exact standardization process depends completely on the
format of the data. Additionally, investigators should document each manipulation so that they can determine if final matches, comparisons, and conclusions are based on original or standardized data.

**Unusual Amounts**

To identify unusual amounts, an investigator should follow three steps: 1) sort the numbers, 2) identify the highest amounts and their payees, and 3) assess the reasonableness of those transactions. However, unusual amounts can vary for each respective payee and, thus, must be viewed in relation to other amounts received by the payee. It is difficult, then, to deem certain amounts unusual until they are separated by vendor. For instance, sorting a data set by amounts quickly will identify any payments in excess of $3,000, but this amount only should rouse suspicion for certain vendors; rent payments regularly exceed $3,000, but a $5,000 payment to a cellular phone company certainly demands greater attention. If possible, investigators first should sort the transactions by payee and then by amount to determine unusually high or low payments to a particular type of vendor. Investigators may have to standardize the data before they conduct this analysis so that they can make proper comparisons.

**Statistical Outliers**

Statistical analysis presents other tactics to determine unusual transaction amounts. Though statistics may seem irrelevant to fraud investigations, in actuality, such analysis can reveal suspicious data that otherwise might go unnoticed.

Most of the statistical analysis investigators perform relates to means, standard deviations, and bell curve distributions. A quick review of the bell curve shows that the large majority (87 percent) of quantities in a distribution should fall within one standard deviation of the mean, while two and three standard deviations should encompass 95 and 99 percent of the quantities, respectively. Most spreadsheet programs easily can calculate the mean to determine which transactions fall on either end of the bell curve.

Statistical outliers are especially useful when the number of transactions reaches the hundreds of thousands or even millions—the more transactions involved, the more useful such analysis will be. Large data sets, such as bank transactions, wire transfers, and payroll records, quickly can be analyzed for questionable transactions. Investigators successfully used this type of analysis in an embezzlement case when statistical outliers helped them uncover falsified checks.

**Round-Number Transactions**

Analyzing round-number transactions can prove useful for reviewing retail business transactions. Consumers can think about how often they have gone to the grocery store and their purchases totaled an even dollar amount; this rarely occurs because most businesses price their products just below the
next whole dollar amount (e.g., $2.99) and because the price fails to reflect tax. Most spreadsheet and database programs can process a formula to quickly highlight transactions with round dollar amounts. Then, the investigator can examine this subset of round-number transactions to determine if any particular payer name, credit card number, or Electronic Benefit Transfer (EBT) card number repeats throughout the list. This technique proved valuable for a possible food stamp fraud case in which certain patrons received cash in exchange for charges to their EBT card. Searching the transactions for rounded dollar amounts quickly revealed a set of individuals whose cards showed a large number of such transactions.

Timing

The timing of financial transactions can be as important as the amounts and payees. Often, transactional information includes a date and time stamp. This date and time information indicates transactions that occur on weekends, after normal business hours, or on a regular basis. In the previously mentioned food stamp fraud case, the timing of the transactions proved important when several transactions occurred within a short time span. The quick time frame exposed the sham transactions because these larger dollar transactions could not have been processed in such short periods of time.

When working with banks and other businesses that operate during standard business hours, transactions that occur outside the normal business day may indicate illegitimate activity. Investigators have uncovered instances of loan manipulation by noting changes in the master loan file that occurred after close of business, on holidays, or on weekends.

Again, this technique may require data standardization. Spreadsheets and databases each process formulae that convert numerical dates into days of the week and, thus, quickly can isolate transactions that occur on a weekend. Investigators can use similar formulae to note the time lapse between two transactions, which reveals transactions that occur in rapid succession.

Counting

One of the easiest, most useful methods to discover unusual financial transactions is simply counting the number of transactions to individual vendors. Most personal and business expenses are paid monthly, and, therefore, most vendors receive 12 payments per year. As such, if an individual pays a landlord or a cellular phone provider 13 or 14 times per year, it might indicate a hidden payment to another vendor, an illicit payment, or a questionable transaction. By contrast, a vendor, employee, or person who receives just one payment in a 12-month period should draw suspicion—a vendor who receives too few payments may indicate money diverted to another purpose or person.

Duplicates

Identifying duplicate dates, amounts, or invoice numbers also can reveal questionable transactions. Obviously, not all transactions with the same amount or date indicate fraud, but the information can help begin investigators’ analysis. Some common types of fraud involve double-paying an invoice and pocketing either the second payment or the refund. Searching for duplicates of supposedly unique invoice numbers

“Though statistics may seem irrelevant to fraud investigations, in actuality, such analysis can reveal suspicious data that otherwise might go unnoticed.”
or transaction amounts highlights these double payments. Some information that investigators can search for duplicates includes voucher numbers, check numbers, payment amounts, unusual payees, invoice numbers, SSNs, names, dates of birth, addresses, and phone numbers. In addition to double payments, duplicate analysis can uncover common addresses between employees and vendors, undocumented or unknown businesses, relationships between employees and vendors, and even collusion among employees.

Spreadsheets and databases do not have built-in formulae to identify duplicate transactions. But, investigators can find applicable formulae, macros, or scripts by searching the Internet.

**Computer Audit Logs**

An audit log or trail comprises an important piece of any accounting software. Most, if not all, accounting software maintains a log of changes to the data contained within the system, yet the majority of users fail to take advantage of these features. Usually, the software stores this information in a separate table and may track changes in such items as date, time, and user. This information can be crucial to prosecution.

For example, in one case involving a lumber company, a salesman committed fraud by changing invoices and collecting the money from the original invoice in cash; the altered invoice effectively erased the original but used the same identification number. The company failed to realize this tactic, and the salesman embezzled large sums over the course of several years. The owner of the company finally became suspicious when a customer mentioned in passing that the salesman collected payments in cash. This information surprised the owner, so he asked to see copies of the invoices. Then, he traced the invoice numbers to sales receipts for other customers, which led him to discover that the lumber company’s accounting software allowed personnel to edit finalized sales orders. The manager continued to search for altered sales orders until he realized that the software maintained audit logs of changes to such information, including the username of the editor and details about the specific changes.

Audit logs greatly can aid an investigator and streamline an investigation. Bookkeepers and accountants may not be aware of the audit features for the software they use, but most commonly used accounting systems maintain some kind of log. If an accountant or auditor is unaware if a system includes such a feature, they should consider contacting the software company directly.

Bank loan software tracks changes in loan amounts, rates, dates, times, users, and much more. The software captures this information in a table sometimes referred to as a “master loan file.” If a bank employee is suspected of changing loan rates, due dates, or manipulating loan files in any respect, a quick review of this master file can reveal all the changes made by that employee. It also may provide a set of leads and a list of loans to which the employee had access.

**Addresses**

Addresses provide another useful data set to analyze. Address analysis can reveal information about relationships between employees and vendors and possibly expose fictitious vendors. Investigators can start by simply mapping the addresses in common mapping software. Most, if not all, mapping software allows the user to import many addresses at once. Some online applications
can perform address mapping as well. Plotting addresses and creating a visual representation of the information allows an investigator to quickly identify suspicious addresses, such as locations in other states or countries and vendor addresses located close to employee addresses.

Also, investigators can compare addresses against known post office box locations. Obviously, investigators quickly can spot the phrase PO Box when reviewing a data set; but, they may not recognize the address of a commercial postal service as easily. Instead of including a PO Box designation, fraudsters might list the address of the commercial postal service building to appear as a legitimate residential or business address. Investigators can consider purchasing a list of commercial mail box locations and compare them to the addresses provided to identify all such addresses.

**Other Methods**

Many of these types of analysis (e.g., addresses, statistics) can extend to more complex investigative techniques. If simple analysis has not yielded any leads in a data set, investigators can consider more advanced techniques, such as geocoding physical addresses, mapping IP addresses, conducting computer forensic analysis, or examining accelerating or decelerating payments, among others. If investigators wish to learn a more advanced method of statistical analysis, they should consider Benford’s Law, which states that in lists of factual, “real life” numbers (e.g., home prices, population sizes, or electricity bills), the leading digit distributes itself in a nonuniform pattern: 1 appears as the first digit about 30 percent of the time, and larger digits occur as the leading digit with decreasing frequency, to the extent that 9 appears as the first digit less than 5 percent of the time. However, most people who compile fraudulent data distribute digits uniformly. Thus, simply comparing the distribution of the leading digits from a data set versus the distribution predicted by Benford’s Law can reveal suspicious numbers.

Also, investigators should consider ways to combine the above techniques to further narrow their search for questionable transactions. For instance, investigators can search a large data set for transactions that are statistical outliers, rounded numbers, and that occurred outside normal working hours. Another example includes identifying abnormal payments among a list of nonissued SSNs. Combining techniques allows investigators to further narrow their data set when the one method of analysis does not suffice.

**CONCLUSION**

Many financial analysis techniques available to investigators were not possible 15 or 20 years ago. The author mentions just a few of the possible methods to uncover unusual or fraudulent transactions. The appropriate techniques for a particular investigation directly depend on the type, quality, and amount of data provided. As companies begin to amass larger financial data sets, these types of analysis have become important to quickly and efficiently identify suspicious transactions.

**Endnotes**

1 The author bases the material in this article on techniques that he has used over the years. He learned to develop and apply these techniques through experience and discussions with other investigators, police officers, special agents, programmers, and accountants.

2 For additional information, see [http://www.intuitior.com/statistics/Benford’s%20Law.html](http://www.intuitior.com/statistics/Benford’s%20Law.html).