Statistical Science
2002, Vol. 17, No. 3, 235–255

Statistical Fraud Detection: A Review

Richard J. Bolton and David J. Hand

Abstract. Fraud is increasing dramatically with the expansion of modern technology and the global superhighways of communication, resulting in the loss of billions of dollars worldwide each year. Although prevention technologies are the best way to reduce fraud, fraudsters are adaptive and, given time, will usually find ways to circumvent such measures. Methodologies for the detection of fraud are essential if we are to catch fraudsters once fraud prevention has failed. Statistics and machine learning provide effective technologies for fraud detection and have been applied successfully to detect activities such as money laundering, e-commerce credit card fraud, telecommunications fraud and computer intrusion, to name but a few. We describe the tools available for statistical fraud detection and the areas in which fraud detection technologies are most used.

Key words and phrases: Fraud detection, fraud prevention, statistics, machine learning, money laundering, computer intrusion, e-commerce, credit cards, telecommunications.

1. INTRODUCTION

The *Concise Oxford Dictionary* defines fraud as "criminal deception; the use of false representations to gain an unjust advantage." Fraud is as old as humanity itself and can take an unlimited variety of different forms. However, in recent years, the development of new technologies (which have made it easier for us to communicate and helped increase our spending power) has also provided yet further ways in which criminals may commit fraud. Traditional forms of fraudulent behavior such as money laundering have become easier to perpetrate and have been joined by new kinds of fraud such as mobile telecommunications fraud and computer intrusion.

We begin by distinguishing between fraud prevention and fraud detection. Fraud *prevention* describes measures to stop fraud from occurring in the first place. These include elaborate designs, fluorescent fibers, multitone drawings, watermarks, laminated metal strips and holographs on banknotes, personal identification numbers for bankcards, Internet security systems for credit card transactions, Subscriber Identity Module (SIM) cards for mobile phones, and passwords on computer systems and telephone bank accounts. Of course, none of these methods is perfect and, in general, a compromise has to be struck between expense and inconvenience (e.g., to a customer) on the one hand, and effectiveness on the other.

In contrast, fraud *detection* involves identifying fraud as quickly as possible once it has been perpetrated. Fraud detection comes into play once fraud prevention has failed. In practice, of course fraud detection must be used continuously, as one will typically be unaware that fraud prevention has failed. We can try to prevent credit card fraud by guarding our cards assiduously, but if nevertheless the card's details are stolen, then we need to be able to detect, as soon as possible, that fraud is being perpetrated.

Fraud detection is a continuously evolving discipline. Whenever it becomes known that one detection method is in place, criminals will adapt their strategies and try others. Of course, new criminals are also constantly entering the field. Many of them will not be aware of the fraud detection methods which have been successful in the past and will adopt strategies which lead to identifiable frauds. This means that the earlier detection tools need to be applied as well as the latest developments.

Richard J. Bolton is Research Associate in the Statistics Section of the Department of Mathematics at Imperial College. David J. Hand is Professor of Statistics in the Department of Mathematics at Imperial College, London SW7 2BZ, United Kingdom (e-mail: r.bolton, d.j.hand@ic.ac.uk).

The development of new fraud detection methods is made more difficult by the fact that the exchange of ideas in fraud detection is severely limited. It does not make sense to describe fraud detection techniques in great detail in the public domain, as this gives criminals the information that they require to evade detection. Data sets are not made available and results are often censored, making them difficult to assess (e.g., Leonard, 1993).

Many fraud detection problems involve huge data sets that are constantly evolving. For example, the credit card company Barclaycard carries approximately 350 million transactions a year in the United Kingdom alone (Hand, Blunt, Kelly and Adams, 2000), The Royal Bank of Scotland, which has the largest credit card merchant acquiring business in Europe, carries over a billion transactions a year and AT&T carries around 275 million calls each weekday (Cortes and Pregibon, 1998). Processing these data sets in a search for fraudulent transactions or calls requires more than mere novelty of statistical model, and also needs fast and efficient algorithms: data mining techniques are relevant. These numbers also indicate the potential value of fraud detection: if 0.1% of a 100 million transactions are fraudulent, each losing the company just £10, then overall the company loses £1 million.

Statistical tools for fraud detection are many and varied, since data from different applications can be diverse in both size and type, but there are common themes. Such tools are essentially based on comparing the observed data with expected values, but expected values can be derived in various ways, depending on the context. They may be single numerical summaries of some aspect of behavior and they are often simple graphical summaries in which an anomaly is readily apparent, but they are also often more complex (multivariate) behavior profiles. Such behavior profiles may be based on past behavior of the system being studied (e.g., the way a bank account has been previously used) or be extrapolated from other similar systems. Things are often further complicated by the fact that, in some domains (e.g., trading on the stock market) a given actor may behave in a fraudulent manner some of the time and not at other times.

Statistical fraud detection methods may be *supervised* or *unsupervised*. In supervised methods, samples of both fraudulent and nonfraudulent records are used to construct models which allow one to assign new observations into one of the two classes. Of course, this requires one to be confident about the true classes of the original data used to build the models. It also requires that one has examples of both classes. Furthermore, it can only be used to detect frauds of a type which have previously occurred.

In contrast, unsupervised methods simply seek those accounts, customers and so forth which are most dissimilar from the norm. These can then be examined more closely. Outliers are a basic form of nonstandard observation. Tools used for checking data quality can be used, but the detection of accidental errors is a rather different problem from the detection of deliberately falsified data or data which accurately describe a fraudulent pattern.

This leads us to note the fundamental point that we can seldom be certain, by statistical analysis alone, that a fraud has been perpetrated. Rather, the analysis should be regarded as alerting us to the fact that an observation is anomalous, or more likely to be fraudulent than others, so that it can then be investigated in more detail. One can think of the objective of the statistical analysis as being to return a *suspicion score* (where we will regard a higher score as more suspicious than a lower one). The higher the score is, then the more unusual is the observation or the more like previously fraudulent values it is. The fact that there are many different ways in which fraud can be perpetrated and many different scenarios in which it can occur means that there are many different ways to compute suspicion scores.

Suspicion scores can be computed for each record in the database (for each customer with a bank account or credit card, for each owner of a mobile phone, for each desktop computer and so on), and these can be updated as time progresses. These scores can then be rank ordered and investigative attention can be focussed on those with the highest scores or on those which exhibit a sudden increase. Here issues of cost enter: given that it is too expensive to undertake a detailed investigation of all records, one concentrates investigation on those thought most likely to be fraudulent.

One of the difficulties with fraud detection is that typically there are many legitimate records for each fraudulent one. A detection method which correctly identifies 99% of the legitimate records as legitimate and 99% of the fraudulent records as fraudulent might be regarded as a highly effective system. However, if only 1 in 1000 records is fraudulent, then, on average, in every 100 that the system flags as fraudulent, only about 9 will in fact be so. In particular, this means that to identify those 9 requires detailed examination of all 100—at possibly considerable cost. This leads us to

a more general point: fraud can be reduced to as low a level as one likes, but only by virtue of a corresponding level of effort and cost. In practice, some compromise has to be reached, often a commercial compromise, between the cost of detecting a fraud and the savings to be made by detecting it. Sometimes the issues are complicated by, for example, the adverse publicity accompanying fraud detection. At a business level, revealing that a bank is a significant target for fraud, even if much has been detected, does little to inspire confidence, and at a personal level, taking action which implies to an innocent customer that they may be suspected of fraud is obviously detrimental to good customer relations.

The body of this paper is structured according to different areas of fraud detection. Clearly we cannot hope to cover all areas in which statistical methods can be applied. Instead, we have selected a few areas where such methods are used and where there is a body of expertise and of literature describing them. However, before looking at the details of different application areas, Section 2 provides a brief overview of some tools for fraud detection.

2. FRAUD DETECTION TOOLS

As we mentioned above, fraud detection can be supervised or unsupervised. Supervised methods use a database of known fraudulent/legitimate cases from which to construct a model which yields a suspicion score for new cases. Traditional statistical classification methods (Hand, 1981; McLachlan, 1992), such as linear discriminant analysis and logistic discrimination, have proved to be effective tools for many applications, but more powerful tools (Ripley, 1996; Hand, 1997; Webb, 1999), especially neural networks, have also been extensively applied. Rule-based methods are supervised learning algorithms that produce classifiers using rules of the form *If* {certain conditions}, *Then* {a consequent}. Examples of such algorithms include BAYES (Clark and Niblett, 1989), FOIL (Quinlan, 1990) and RIPPER (Cohen, 1995). Tree-based algorithms such as CART (Breiman, Friedman, Olshen and Stone, 1984) and C4.5 (Quinlan, 1993) produce classifiers of a similar form. Combinations of some or all of these algorithms can be created using meta-learning algorithms to improve prediction in fraud detection (e.g., Chan, Fan, Prodromidis and Stolfo, 1999).

Major considerations when building a supervised tool for fraud detection include those of uneven class sizes and different costs of different types of misclassification. We must also take into consideration the costs of investigating observations and the benefits of identifying fraud. Moreover, often class membership is uncertain. For example, credit transactions may be labelled incorrectly: a fraudulent transaction may remain unobserved and thus be labeled legitimate (and the extent of this may remain unknown) or a legitimate transaction may be misreported as fraudulent. Some work has addressed misclassification of training samples (e.g., Lachenbruch, 1966, 1974; Chhikara and McKeon, 1984), but not in the context of fraud detection as far as we are aware. Issues such as these were discussed by Chan and Stolfo (1998) and Provost and Fawcett (2001).

Link analysis relates known fraudsters to other individuals using record linkage and social network methods (Wasserman and Faust, 1994). For example, in telecommunications networks, security investigators have found that fraudsters seldom work in isolation from each other. Also, after an account has been disconnected for fraud, the fraudster will often call the same numbers from another account (Cortes, Pregibon and Volinsky, 2001). Telephone calls from an account can thus be linked to fraudulent accounts to indicate intrusion. A similar approach has been taken in money laundering (Goldberg and Senator, 1995, 1998; Senator et al., 1995).

Unsupervised methods are used when there are no prior sets of legitimate and fraudulent observations. Techniques employed here are usually a combination of profiling and outlier detection methods. We model a baseline distribution that represents normal behavior and then attempt to detect observations that show the greatest departure from this norm. There are similarities to author identification in text analysis. Digit analysis using Benford's law is an example of such a method. Benford's law (Hill, 1995) says that the distribution of the first significant digits of numbers drawn from a wide variety of random distributions will have (asymptotically) a certain form. Until recently, this law was regarded as merely a mathematical curiosity with no apparent useful application. However, Nigrini and Mittermaier (1997) and Nigrini (1999) showed that Benford's law can be used to detect fraud in accounting data. The premise behind fraud detection using tools such as Benford's law is that fabricating data which conform to Benford's law is difficult.

Fraudsters adapt to new prevention and detection measures, so fraud detection needs to be adaptive and evolve over time. However, legitimate account users may gradually change their behavior over a longer period of time and it is important to avoid spurious

alarms. Models can be updated at fixed time points or continuously over time; see, for example, Burge and Shawe-Taylor (1997), Fawcett and Provost (1997a), Cortes, Pregibon and Volinsky (2001) and Senator (2000).

Although the basic statistical models for fraud detection can be categorized as supervised or unsupervised, the application areas of fraud detection cannot be described so conveniently. Their diversity is reflected in their particular operational characteristics and the variety and quantity of data available, both features that drive the choice of a suitable fraud detection tool.

3. CREDIT CARD FRAUD

The extent of credit card fraud is difficult to quantify, partly because companies are often loath to release fraud figures in case they frighten the spending public and partly because the figures change (probably grow) over time. Various estimates have been given. For example, Leonard (1993) suggested the cost of Visa/Mastercard fraud in Canada in 1989, 1990 and 1991 was $19, 29 and 46 million (Canadian), respectively. Ghosh and Reilly (1994) suggested a figure of $850 million (U.S.) per year for all types of credit card fraud in the United States, and Aleskerov, Freisleben and Rao (1997) cited estimates of $700 million in the United States each year for Visa/Mastercard and $10 billion worldwide in 1996. Microsoft's Expedia set aside $6 million for credit card fraud in 1999 (Patient, 2000). Total losses through credit card fraud in the United Kingdom have been growing rapidly over the last 4 years [1997, £122 million; 1998, £135 million; 1999, £188 million; 2000, £293 million. *Source*: Association for Payment Clearing Services, London (APACS)] and recently APACS reported £373.7 million losses in the 12 months ending August 2001. Jenkins (2000) says "for every £100 you spend on a card in the UK, 13p is lost to fraudsters." Matters are complicated by issues of exactly what one includes in the fraud figures. For example, bankruptcy fraud arises when the cardholder makes purchases for which he/she has no intention of paying and then files for personal bankruptcy, leaving the bank to cover the losses. Since these are generally regarded as charge-off losses, they often are not included in fraud figures. However, they can be substantial: Ghosh and Reilly (1994) cited one estimate of $2.65 billion for bankruptcy fraud in 1992.

It is in a company and card issuer's interests to prevent fraud or, failing this, to detect fraud as soon as possible. Otherwise consumer trust in both the card and the company decreases and revenue is lost, in addition to the direct losses made through fraudulent sales. Because of the potential for loss of sales due to loss of confidence, in general, the merchants assume responsibility for fraud losses, even when the vendor has obtained authorization from the card issuer.

Credit card fraud may be perpetrated in various ways (a description of the credit card industry and how it functions is given in Blunt and Hand, 2000), including simple theft, application fraud and counterfeit cards. In all of these, the fraudster uses a physical card, but physical possession is not essential to perpetrate credit card fraud: one of the major fraud areas is "cardholder-not-present" fraud, where only the card's details are given (e.g., over the phone).

Use of a stolen card is perhaps the most straightforward type of credit card fraud. In this case, the fraudster typically spends as much as possible in as short a space of time as possible, before the theft is detected and the card is stopped; hence, detecting the theft early can prevent large losses.

Application fraud arises when individuals obtain new credit cards from issuing companies using false personal information. Traditional credit scorecards (Hand and Henley, 1997) are used to detect customers who are likely to default, and the reasons for this may include fraud. Such scorecards are based on the details given on the application forms and perhaps also on other details such as bureau information. Statistical models which monitor behavior over time can be used to detect cards which have been obtained from a fraudulent application (e.g., a first time card holder who runs out and rapidly makes many purchases should arouse suspicion). With application fraud, however, urgency is not as important to the fraudster and it might not be until accounts are sent out or repayment dates begin to pass that fraud is suspected.

Cardholder-not-present fraud occurs when the transaction is made remotely, so that only the card's details are needed, and a manual signature and card imprint are not required at the time of purchase. Such transactions include telephone sales and on-line transactions, and this type of fraud accounts for a high proportion of losses. To undertake such fraud it is necessary to obtain the details of the card without the cardholder's knowledge. This is done in various ways, including "skimming," where employees illegally copy the magnetic strip on a credit card by swiping it through a small handheld card reader, "shoulder surfers," who enter card details into a mobile phone while standing behind a purchaser in a queue, and people posing as credit

card company employees taking details of credit card transactions from companies over the phone. Counterfeit cards, currently the largest source of credit card fraud in the United Kingdom (*source*: APACS), can also be created using this information. Transactions made by fraudsters using counterfeit cards and making cardholder-not-present purchases can be detected through methods which seek changes in transaction patterns, as well as checking for particular patterns which are known to be indicative of counterfeiting.

Credit card databases contain information on each transaction. This information includes such things as merchant code, account number, type of credit card, type of purchase, client name, size of transaction and date of transaction. Some of these data are numerical (e.g., transaction size) and others are nominal categorical (e.g., merchant code, which can have hundreds of thousands of categories) or symbolic. The mixed data types have led to the application of a wide variety of statistical, machine learning and data mining tools.

Suspicion scores to detect whether an account has been compromised can be based on models of individual customers' previous usage patterns, standard expected usage patterns, particular patterns which are known to be often associated with fraud, and on supervised models. A simple example of the patterns exhibited by individual customers is given in Figure 16 of Hand and Blunt (2001), which shows how the slopes of cumulative credit card spending over time are remarkably linear. Sudden jumps in these curves or sudden changes of slope (transaction or expenditure rate suddenly exceeding some threshold) merit investigation. Likewise, some customers practice "jam jarring"—restricting particular cards to particular types of purchases (e.g., using a given card for petrol purchases only and a different one for supermarket purchases), so that usage of a card to make an unusual type of purchase can trigger an alarm for such customers. At a more general level, suspicion scores can also be based on expected overall usage profiles. For example, first time credit card users are typically initially fairly tentative in their usage, whereas those transferring loans from another card are generally not so reticent. Finally, examples of overall transaction patterns known to be intrinsically suspicious are the sudden purchase of many small electrical items or jewelry (goods which permit easy black market resale) and the immediate use of a new card in a wide range of different locations.

We commented above that, for obvious reasons, there is a dearth of published literature on fraud detection. Much of that which has been published appears in the methodological data analytic literature, where the aim is to illustrate new data analytic tools by applying them to the detection of fraud, rather than to describe methods of fraud detection per se. Furthermore, since anomaly detection methods are very context dependent, much of the published literature in the area concentrates on supervised classification methods. In particular, rule-based systems and neural networks have attracted interest. Researchers who have used neural networks for credit card fraud detection include Ghosh and Reilly (1994), Aleskerov et al. (1997), Dorronsoro, Ginel, Sanchez and Cruz (1997) and Brause, Langsdorf and Hepp (1999), mainly in the context of supervised classification. HNC Software has developed *Falcon*, a software package that relies heavily on neural network technology to detect credit card fraud.

Supervised methods, using samples from the fraudulent/nonfraudulent classes as the basis to construct classification rules to detect future cases of fraud, suffer from the problem of unbalanced class sizes mentioned above: the legitimate transactions generally far outnumber the fraudulent ones. Brause, Langsdorf and Hepp (1999) said that, in their database of credit card transactions, "the probability of fraud is very low (0.2%) and has been lowered in a preprocessing step by a conventional fraud detecting system down to 0.1%." Hassibi (2000) remarked that "out of some 12 billion transactions made annually, approximately 10 million—or one out of every 1200 transactions—turn out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts are fraudulent." It follows from this sort of figure that simple misclassification rate cannot be used as a performance measure: with a bad rate of 0.1%, simply classifying every transaction as legitimate will yield an error rate of only 0.001. Instead, one must either minimize an appropriate cost-weighted loss or fix some parameter (such as the number of cases one can afford to investigate in detail) and then try to maximize the number of fraudulent cases detected subject to the constraints.

Stolfo et al. (1997a, b) outlined a meta-classifier system for detecting credit card fraud that is based on the idea of using different local fraud detection tools within each different corporate environment and merging the results to yield a more accurate global tool. This work was elaborated in Chan and Stolfo (1998), Chan, Fan, Prodromidis and Stolfo (1999) and Stolfo et al. (1999), who described a more realistic cost model to accompany the different classification outcomes. Wheeler and Aitken (2000) also explored the combination of multiple classification rules.

4. MONEY LAUNDERING

Money laundering is the process of obscuring the source, ownership or use of funds, usually cash, that are the profits of illicit activity. The size of the problem is indicated in a 1995 U.S. Office of Technology Assessment (OTA) report (U.S. Congress, 1995): "Federal agencies estimate that as much as $300 billion is laundered annually, worldwide. From $40 billion to $80 billion of this may be drug profits made in the United States." Prevention is attempted by means of legal constraints and requirements—the burden of which is gradually increasing—and there has been much debate recently about the use of encryption. However, no prevention strategy is foolproof and detection is essential. In particular, the September 11th terrorist attacks on New York City and the Pentagon have focused attention on the detection of money laundering in an attempt to starve terrorist networks of funds.

Wire transfers provide a natural domain for laundering: according to the OTA report, each day in 1995 about half a million wire transfers, valued at more than $2 trillion (U.S.), were carried out using the Fedwire and CHIPS systems, along with almost a quarter of a million transfers using the SWIFT system. It is estimated that around 0.05–0.1% of these transactions involved laundering. Sophisticated statistical and other on-line data analytic procedures are needed to detect such laundering activity. Since it is now becoming a legal requirement to show that all reasonable means have been used to detect fraud, we may expect to see even greater application of such tools.

Wire transfers contain items such as date of transfer, identity of sender, routing number of originating bank, identity of recipient, routing number of recipient bank and amount transferred. Sometimes those fields not needed for transfer are left blank, free text fields may be completed in different ways and, worse still, but inevitable, sometimes the data have errors. Automatic error detection (and correction) software has been developed, based on semantic and syntactic constraints on possible content, but, of course, this can never be a complete solution. Matters are also complicated by the fact that banks do not share their data. Of course, banks are not the only bodies that transfer money electronically, and other businesses have been established precisely for this purpose [the OTA report (U.S. Congress, 1995) estimates the number of such businesses as 200,000].

The detection of money laundering presents difficulties not encountered in areas such as, for example, the credit card industry. Whereas credit card fraud comes to light fairly early on, in money laundering it may be years before individual transfers or accounts are definitively and legally identified as part of a laundering process. While, in principle (assuming records have been kept), one could go back and trace the relevant transactions, in practice not all of them would be identified, so detracting from their use in supervised detection methods. Furthermore, there is typically less extensive information available for the account holders in investment banks than there is in retail banking operations. Developing more detailed customer record systems might be a good way forward.

As with other areas of fraud, money laundering detection works hand in hand with prevention. In 1970, for example, in the United States the Bank Secrecy Act required that banks report all currency transactions of over $10,000 to the authorities. However, also as in other areas of fraud, the perpetrators adapt their modus operandi to match the changing tactics of the authorities. So, following the requirement of banks to report currency transactions of over $10,000, the obvious strategy was developed to divide larger sums into multiple amounts of less than $10,000 and deposit them in different banks (a practice termed *smurfing* or *structuring*). In the United States, this is now illegal, but the way the money launderers adapt to the prevailing detection methods can lead one to the pessimistic perspective that only the incompetent money launderers are detected. This, clearly, also limits the value of supervised detection methods: the patterns detected will be those patterns which were characteristic of fraud in the past, but which may no longer be so. Other strategies used by money launderers which limit the value of supervised methods include switching between wire and physical cash movements, the creation of shell businesses, false invoicing and, of course, the fact that a single transfer, in itself, is unlikely to appear to be a laundering transaction. Furthermore, because of the large sums involved, money launderers are highly professional and often have contacts in the banks who can feed back details of the detection strategies being applied.

The number of currency transactions over $10,000 in value increased dramatically after the mid-1980s, to the extent that the number of reports filed is huge (over 10 million in 1994, with total worth of around $500 billion), and this in itself can cause difficulties. In an attempt to cope with this, the Financial Crimes Enforcement Network (FinCEN) of the U.S. Department of the Treasury processes all such reports using

the FinCEN artificial intelligence system (FAIS) described below. More generally, banks are also required to report any suspicious transactions, and about 0.5% of currency transaction reports are so flagged.

Money laundering involves three steps:

1. *Placement*: the introduction of the cash into the banking system or legitimate business (e.g., transferring the banknotes obtained from retail drugs transactions into a cashier's cheque). One way to do this is to pay vastly inflated amounts for goods imported across international frontiers. Pak and Zdanowicz (1994) described statistical analysis of trade databases to detect anomalies in government trade data such as charging $1694 a gram for imports of the drug erythromycin compared with $0.08 a gram for exports.
2. *Layering*: carrying out multiple transactions through multiple accounts with different owners at different financial institutions in the legitimate financial system.
3. *Integration*: merging the funds with money obtained from legitimate activities.

Detection strategies can be targeted at various levels. In general (and in common with some other areas in which fraud is perpetrated), it is very difficult or impossible to characterize an individual transaction as fraudulent. Rather transaction *patterns* must be identified as fraudulent or suspicious. A single deposit of just under $10,000 is not suspicious, but multiple such deposits are; a large sum being deposited is not suspicious, but a large sum being deposited and instantly withdrawn is. In fact, one can distinguish several levels of (potential) analysis: the individual transaction level, the account level, the business level (and, indeed, individuals may have multiple accounts) and the "ring" of businesses level. Analyses can be targeted at particular levels, but more complex approaches can examine several levels simultaneously. (There is an analogy here with speech recognition systems: simple systems focused at the individual phoneme and word levels are not as effective as those which try to recognize these elements in a higher level context of the way words are put together when used.) In general, link analysis, which identifies groups of participants involved in transactions, plays a key role in most money laundering detection strategies. Senator et al. (1995) said "Money laundering typically involves a multitude of transactions, perhaps by distinct individuals, into multiple accounts with different owners at different banks and other financial institutions. Detection of large-scale money laundering schemes requires the ability to reconstruct these patterns of transactions by linking potentially related transactions and then to distinguish the legitimate sets of transactions from the illegitimate ones. This technique of finding relationships between elements of information, called *link analysis*, is the primary analytic technique used in law enforcement intelligence (Andrews and Peterson, 1990)." An obvious and simplistic illustration is the fact that a transaction with a known criminal may rouse suspicion. More subtle methods are based on recognition of the sort of businesses with which money laundering operations transact. Of course, these are all supervised methods and are subject to the weaknesses that those responsible may evolve their strategies. Similar tools are used to detect telecom fraud, as outlined in the following section.

Rule-based systems have been developed, often with the rules based on experience ("flag transactions from countries X and Y"; "flag accounts showing a large deposit followed immediately by a similar sized withdrawal"). Structuring can be detected by computing the cumulative sum of amounts entering an account over a short window, such as a day. Other methods have been developed based on straightforward descriptive statistics, such as rate of transactions and proportion of transactions which are suspicious. The use of the Benford distribution is an extension of this idea. Although one may not usually be interested in detecting changes in an account's behavior, methods such as peer group analysis (Bolton and Hand, 2001) and break detection (Goldberg and Senator, 1997) can be applied to detect money laundering.

One of the most elaborate money laundering detection systems is the U.S. Financial Crimes Enforcement Network AI system (FAIS) described in Senator et al. (1995) and Goldberg and Senator (1998). This system allows users to follow trails of linked transactions. It is built around a "blackboard" architecture, in which program modules can read and write to a central database that contains details of transactions, subjects and accounts. A key component of the system is its suspicion score. This is a rule-based system based on an earlier system developed by the U.S. Customs Service in the mid-1980s. The system computes suspicion scores for various different types of transaction and activity. Simple Bayesian updating is used to combine evidence that suggests that a transaction or activity is illicit to yield an overall suspicion score. Senator et al. (1995) included a brief but interesting discussion of an investigation of whether case-based reasoning (cf. nearest

neighbor methods) and classification tree techniques could usefully be added to the system.

The American National Association of Securities Dealers, Inc., uses an *advanced detection system* (ADS; Kirkland et al., 1998; Senator, 2000) to flag "patterns or practices of regulatory concern." ADS uses a rule pattern matcher and a time-sequence pattern matcher, and (like FAIS) places great emphasis on visualization tools. Also as with FAIS, data mining techniques are used to identify new patterns of potential interest.

A different approach to detecting similar fraudulent behavior is taken by SearchSpace Ltd. (www.searchspace.com), which has developed a system for the London Stock Exchange called MonITARS (monitoring insider trading and regulatory surveillance) that combines genetic algorithms, fuzzy logic and neural network technology to detect insider dealing and market manipulation. Chartier and Spillane (2000) also described an application of neural networks to detect money laundering.

5. TELECOMMUNICATIONS FRAUD

The telecommunications industry has expanded dramatically in the last few years with the development of affordable mobile phone technology. With the increasing number of mobile phone users, global mobile phone fraud is also set to rise. Various estimates have been presented for the cost of this fraud. For example, Cox, Eick, Wills and Brachman (1997) gave a figure of $1 billion a year. *Telecom and Network Security Review* [**4**(5) April 1997] gave a figure of between 4 and 6% of U.S. telecom revenue lost due to fraud. Cahill, Lambert, Pinheiro and Sun (2002) suggested that international figures are worse, with "several new service providers reporting losses over 20%." Moreau et al. (1996) gave a value of "several million ECUs per year." Presumably this refers to within the European Union and, given the size of the other estimates, we wonder if this should be billions. According to a recent report (Neural Technologies, 2000), "the industry already reports a loss of £13 billion each year due to fraud." Mobile Europe (2000) gave a figure of $13 billion (U.S.). The latter article also claimed that it is estimated that fraudsters can steal up to 5% of some operators' revenues, and that some expect telecom fraud as a whole to reach $28 billion per year within 3 years.

Despite the variety in these figures, it is clear that they are all very large. Apart from the fact that they are simply estimates, and hence subject to expected inaccuracies and variability based on the information used to derive them, there are other reasons for the differences. One is the distinction between hard and soft currency. Hard currency is real money, paid by someone other than the perpetrator for the service the perpetrator has stolen. Hynninen (2000) gave the example of the sum one mobile phone operator will pay another for the use of their network. Soft currency is the value of the service the perpetrator has stolen. At least part of this is only a loss if one assumes that the thief would have used the same service even if he or she had had to pay for it. Another reason for the differences derives from the fact that such estimates may be used for different purposes. Hynninen (2000) gave the examples of operators giving estimates on the high side, hoping for more stringent antifraud legislation, and operators giving estimates on the low side to encourage customer confidence.

We need to distinguish between fraud aimed *at* the service provider and fraud enabled *by* the service provider. An example of the former is the resale of stolen call time and an example of the latter is interfering with telephone banking instructions. (It is the possibility of the latter sort of fraud which makes the public wary of using their credit cards over the Internet.) We can also distinguish between revenue fraud and nonrevenue fraud. The aim of the former is to make money for the perpetrator, while the aim of the latter is simply to obtain a service free of charge (or, as with computer hackers, e.g., the simple challenge represented by the system).

There are many different types of telecom fraud (see, e.g., Shawe-Taylor et al., 2000) and these can occur at various levels. The two most prevalent types are subscription fraud and superimposed or "surfing" fraud. Subscription fraud occurs when the fraudster obtains a subscription to a service, often with false identity details, with no intention of paying. This is thus at the level of a phone number—all transactions from this number will be fraudulent. Superimposed fraud is the use of a service without having the necessary authority and is usually detected by the appearance of phantom calls on a bill. There are several ways to carry out superimposed fraud, including mobile phone cloning and obtaining calling card authorization details. Superimposed fraud will generally occur at the level of individual calls—the fraudulent calls will be mixed in with the legitimate ones. Subscription fraud will generally be detected at some point through the billing process—although the aim is to detect it well before that, since large costs can quickly be run up. Superimposed fraud can remain undetected for a long time. The distinction

between these two types of fraud follows a similar distinction in credit card fraud.

Other types of telecom fraud include "ghosting" (technology that tricks the network so as to obtain free calls) and insider fraud, where telecom company employees sell information to criminals that can be exploited for fraudulent gain. This, of course, is a universal cause of fraud, whatever the domain. "Tumbling" is a type of superimposed fraud in which rolling fake serial numbers are used on cloned handsets, so that successive calls are attributed to different legitimate phones. The chance of detection by spotting unusual patterns is small and the illicit phone will operate until all of the assumed identities have been spotted. The term "spoofing" is sometimes used to describe users pretending to be someone else.

Telecommunications networks generate vast quantities of data, sometimes on the order of several gigabytes per day, so that data mining techniques are of particular importance. The 1998 database of AT&T, for example, contained 350 million profiles and processed 275 million call records per day (Cortes and Pregibon, 1998).

As with other fraud domains, apart from some domain specific tools, methods for detection hinge around outlier detection and supervised classification, either using rule-based methods or based on comparing statistically derived suspicion scores with some threshold. At a low level, simple rule-based detection systems use rules such as the apparent use of the same phone in two very distant geographical locations in quick succession, calls which appear to overlap in time, and very high value and very long calls. At a higher level, statistical summaries of call distributions (often called *profiles* or *signatures* at the user level) are compared with thresholds determined either by experts or by application of supervised learning methods to known fraud/nonfraud cases. Murad and Pinkas (1999) and Rosset et al. (1999) distinguished between profiling at the levels of individual calls, daily call patterns and overall call patterns, and described what are effectively outlier detection methods for detecting anomalous behavior. A particularly interesting description of profiling methods was given by Cortes and Pregibon (1998). Cortes, Fisher, Pregibon and Rogers (2000) described the *Hancock* language for writing programs for processing profiles, basing the signatures on such quantities as average call duration, longest call duration, number of calls to particular regions in the last day and so on. Profiling and classification techniques also were described by Fawcett and Provost (1997a, b, 1999) and Moreau, Verrelst and Vandewalle (1997). Some work (see, e.g., Fawcett and Provost, 1997a) has focused on detecting changes in behavior.

A general complication is that signatures and thresholds may need to depend on time of day, type of account and so on, and that they will probably need to be updated over time. Cahill et al. (2002) suggested excluding the very suspicious scores in this updating process, although more work is needed in this area.

Once again, neural networks have been widely used. The main fraud detection software of the Fraud Solutions Unit of Nortel Networks (Nortel, 2000) uses a combination of profiling and neural networks. Likewise, ASPeCT (Moreau et al., 1996; Shawe-Taylor et al., 2000), a project of the European Commission, Vodaphone, other European telecom companies and academics, developed a combined rule-based profiling and neural network approach. Taniguchi, Haft, Hollmén and Tresp (1998) described neural networks, mixture models and Bayesian networks in telecom fraud detection based on call records stored for billing.

Link analysis, with links updated over time, establishes the "communities of interest" (Cortes, Pregibon and Volinsky, 2001) that can indicate networks of fraudsters. These methods are based on the observation that fraudsters seldom change their calling habits, but are often closely linked to other fraudsters. Using similar patterns of transactions to infer the presence of a particular fraudster is in the spirit of phenomenal data mining (McCarthy, 2000).

Visualization methods (Cox et al., 1997), developed for mining very large data sets, have also been developed for use in telecom fraud detection. Here human pattern recognition skills interact with graphical computer display of quantities of calls between different subscribers in various geographical locations. A possible future scenario would be to code into software the patterns which humans detect.

The telecom market will become even more complicated over time—with more opportunity for fraud. At present the extent of fraud is measured by considering factors such as call lengths and tariffs. The third generation of mobile phone technology will also need to take into account such things as the content of the calls (because of the packet switching technology used, equally long data transmissions may contain very different numbers of data packets) and the priority of the call.

6. COMPUTER INTRUSION

On Thursday, September 21, 2000, a 16-year-old boy was jailed for hacking into both the Pentagon and NASA computer systems. Between the 14th and 25th of October 2000 Microsoft security tracked the illegal activity of a hacker on the Microsoft Corporate Network. These examples illustrate that even exceptionally well protected domains can have their computer security compromised.

Computer intrusion fraud is big business and computer intrusion detection is a hugely intensive area of research. Hackers can find passwords, read and change files, alter source code, read e-mails and so on. Denning (1997) listed eight kinds of computer intrusion. If the hackers can be prevented from penetrating the computer system or can be detected early enough, then such crime can be virtually eliminated. However, as with all fraud when the prizes are high, the attacks are adaptive and once one kind of intrusion has been recognized the hacker will try a different route. Because of its importance, a great deal of effort has been put into developing intrusion detection methods, and there are several commercial products available, including Cisco secure intrusion detection system (CSIDS, 1999) and next-generation intrusion detection expert system (NIDES; Anderson, Frivold and Valdes, 1995).

Since the only record of a hacker's activities is the sequence of commands that is used when compromising the system, analysts of computer intrusion data predominantly use sequence analysis techniques. As with other fraud situations, both supervised and unsupervised methods are used. In the context of intrusion detection, supervised methods are sometimes called *misuse detection*, while the unsupervised methods used are generally methods of anomaly detection, based on profiles of usage patterns for each legitimate user. Supervised methods have the problem described in other contexts, that they can, of course, only work on intrusion patterns which have already occurred (or partial matches to these). Lee and Stolfo (1998) applied classification techniques to data from a user or program that has been identified as either normal or abnormal. Lippmann et al. (2000) concluded that emphasis should be placed on developing methods for detecting new patterns of intrusion rather than old patterns, but Kumar and Spafford (1994) remarked that "a majority of break-ins ... are the result of a small number of known attacks, as evidenced by reports from response teams (e.g., CERT). Automating detection of these attacks should therefore result in the detection of a significant number of break-in attempts." Shieh and Gligor (1991, 1997) described a pattern-matching method and argued that it is more effective than statistical methods at detecting known types of intrusion, but is unable to detect novel kinds of intrusion patterns, which could be detected by statistical methods.

Since intrusion represents behavior and the aim is to distinguish between intrusion behavior and usual behavior in sequences, Markov models have naturally been applied (e.g., Ju and Vardi, 2001). Qu et al. (1998) also used probabilities of events to define the profile. Forrest, Hofmeyr, Somayaji and Longstaff (1996) described a method based on how natural immune systems distinguish between self and alien patterns. As with telecom data, both individual user patterns and overall network behavior change over time, so that a detection system must be able to adapt to changes, but not adapt so rapidly that it also accepts intrusions as legitimate changes. Lane and Brodley (1998) and Kosoresow and Hofmeyr (1997) also used similarity of sequences that can be interpreted in a probabilistic framework.

Inevitably, neural networks have been used: Ryan, Lin and Miikkulainen (1997) performed profiling by training a neural network on the process data and also referenced other neural approaches. In one of the more careful studies in the area, Schonlau et al. (2001) described a comparative study of six statistical approaches for detecting impersonation of other users (masquerading), where they took real usage data from 50 users and planted contaminating data from other users to serve as the masquerade targets to be detected. A nice overview of statistical issues in computer intrusion detection was given by Marchette (2001), and the October 2000 edition of *Computer Networks* [**34**(4)] is a special issue on (relatively) recent advances in intrusion detection systems, including several examples of new approaches to computer intrusion detection.

7. MEDICAL AND SCIENTIFIC FRAUD

Medical fraud can occur at various levels. It can occur in clinical trials (see, e.g., Buyse et al., 1999). It can also occur in a more commercial context: for example, prescription fraud, submitting claims for patients who are dead or who do not exist, and upcoding, where a doctor performs a medical procedure, but charges the insurer for one that is more expensive, or perhaps does not even perform one at all. Allen (2000) gave an example of bills submitted for more than 24 hours in a working day. He, Wang, Graco and Hawkins (1997)

and He, Graco and Yao (1999) described the use of neural networks, genetic algorithms and nearest neighbor methods to classify the practice profiles of general practitioners in Australia into classes from normal to abnormal.

Medical fraud is often linked to insurance fraud: Terry Allen, a statistician with the Utah Bureau of Medicaid Fraud, estimated that up to 10% of the $800 million annual claims may be stolen (Allen, 2000). Major and Riedinger (1992) created a knowledge/statistical-based system to detect healthcare fraud by comparing observations with those with which they should be most similar (e.g., having similar geodemographics). Brockett, Xia and Derrig (1998) used neural networks to classify fraudulent and nonfraudulent claims for automobile bodily injury in healthcare insurance claims. Glasgow (1997) gave a short discussion of risk and fraud in the insurance industry. A glossary of several of the different types of medical fraud is available at http://www.motherjones.com/mother_jones/MA95/davis2.html.

Of course, medicine is not the only scientific area where data have sometimes been fabricated, falsified or carefully selected to support a pet theory. Problems of fraud in science are attracting increased attention, but they have always been with us: errant scientists have been known to massage figures from experiments to push through development of a product or reach a magical significance level for a publication. Dmitriy Yuryev described such a case on his webpages at http://www.orc.ru/∼yur77/statfr.htm. Moreover, there are many classical cases in which the data have been suspected of being massaged (including the work of Galileo, Newton, Babbage, Kepler, Mendel, Millikan and Burt). Press and Tanur (2001) presented a fascinating discussion of the role of subjectivity in the scientific process, illustrating with many examples. The borderline between subconscious selection of data and out-and-out distortion is a fine one.

8. CONCLUSIONS

The areas we have outlined are perhaps those in which statistical and other data analytic tools have made the most impact on fraud detection. This is typically because there are large quantities of information, and this information is numerical or can easily be converted into the numerical in the form of counts and proportions. However, other areas, not mentioned above, have also used statistical tools for fraud detection. Irregularities in financial statements can be used to detect accounting and management fraud in contexts broader than those of money laundering. Digit analysis tools have found favor in accountancy (e.g., Nigrini and Mittermaier, 1997; Nigrini, 1999). Statistical sampling methods are important in financial audit, and screening tools are applied to decide which tax returns merit detailed investigation. We mentioned insurance fraud in the context of medicine, but it clearly occurs more widely. Artís, Ayuso and Guillén (1999) described an approach to modelling fraud behavior in car insurance, and Fanning, Cogger and Srivastava (1995) and Green and Choi (1997) examined neural network classification methods for detecting management fraud. Statistical tools for fraud detection have also been applied to sporting events. For example, Robinson and Tawn (1995), Smith (1997) and Barao and Tawn (1999) examined the results of running events to see if some exceptional times were out of line with what might be expected.

Plagiarism is also a type of fraud. We briefly referred to the use of statistical tools for author verification and such methods can be applied here. However, statistical tools can also be applied more widely. For example, with the evolution of the Internet it is extremely easy for students to plagiarize articles and pass them off as their own in school or university coursework. The website http://www.plagiarism.org describes a system that can take a manuscript and compare it against their "substantial database" of articles from the Web. A statistical measure of the originality of the manuscript is returned.

As we commented in the Introduction, fraud detection is a post hoc strategy, being applied after fraud prevention has failed. Statistical tools are also applied in some fraud prevention methods. For example, so-called *biometric* methods of fraud detection are slowly becoming more widespread. These include computerized fingerprint and retinal identification, and also face recognition (although this has received most publicity in the context of recognizing football hooligans).

In many of the applications we have discussed, speed of processing is of the essence. This is particularly the case in transaction processing, especially with telecom and intrusion data, where vast numbers of records are processed every day, but also applies in credit card, banking and retail sectors.

A key issue in all of this work is how effective the statistical tools are in detecting fraud and a fundamental problem is that one typically does not know how many fraudulent cases slip through the net. In applications such as banking fraud and telecom fraud, where

speed of detection matters, measures such as average time to detection after fraud starts (in minutes, numbers of transactions, etc.) should also be reported. Measures of this aspect interact with measures of final detection rate: in many situations an account, telephone and so forth, will have to be used for several fraudulent transactions before it is detected as fraudulent, so that several false negative classifications will necessarily be made.

An appropriate overall strategy is to use a graded system of investigation. Accounts with very high suspicion scores merit immediate and intensive (and expensive) investigation, while those with large but less dramatic scores merit closer (but not expensive) observation. Once again, it is a matter of choosing a suitable compromise.

Finally, it is worth repeating the conclusions reached by Schonlau et al. (2001), in the context of statistical tools for computer intrusion detection: "statistical methods can detect intrusions, even in difficult circumstances," but also "many challenges and opportunities for statistics and statisticians remain." We believe this positive conclusion holds more generally. Fraud detection is an important area, one in many ways ideal for the application of statistical and data analytic tools, and one where statisticians can make a very substantial and important contribution.

ACKNOWLEDGMENT

The work of Richard Bolton was supported by a ROPA award from the Engineering and Physical Sciences Research Council of the United Kingdom.

REFERENCES

ALESKEROV, E., FREISLEBEN, B. and RAO, B. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. In *Computational Intelligence for Financial Engineering. Proceedings of the IEEE/IAFE* 220–226. IEEE, Piscataway, NJ.

ALLEN, T. (2000). A day in the life of a Medicaid fraud statistician. *Stats* **29** 20–22.

ANDERSON, D., FRIVOLD, T. and VALDES, A. (1995). Next-generation intrusion detection expert system (NIDES): A summary. Technical Report SRI-CSL-95-07, Computer Science Laboratory, SRI International, Menlo Park, CA.

ANDREWS, P. P. and PETERSON, M. B., eds. (1990). *Criminal Intelligence Analysis*. Palmer Enterprises, Loomis, CA.

ARTÍS, M., AYUSO, M. and GUILLÉN, M. (1999). Modelling different types of automobile insurance fraud behaviour in the Spanish market. *Insurance Mathematics and Economics* **24** 67–81.

BARAO, M. I. and TAWN, J. A. (1999). Extremal analysis of short series with outliers: Sea-levels and athletics records. *Appl. Statist.* **48** 469–487.

BLUNT, G. and HAND, D. J. (2000). The UK credit card market. Technical report, Dept. Mathematics, Imperial College, London.

BOLTON, R. J. and HAND, D. J. (2001). Unsupervised profiling methods for fraud detection. In *Conference on Credit Scoring and Credit Control* **7**, Edinburgh, UK, 5–7 Sept.

BRAUSE, R., LANGSDORF, T. and HEPP, M. (1999). Neural data mining for credit card fraud detection. In *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence* 103–106. IEEE Computer Society Press, Silver Spring, MD.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BROCKETT, P. L., XIA, X. and DERRIG, R. A. (1998). Using Kohonen's self-organising feature map to uncover automobile bodily injury claims fraud. *The Journal of Risk and Insurance* **65** 245–274.

BURGE, P. and SHAWE-TAYLOR, J. (1997). Detecting cellular fraud using adaptive prototypes. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 9–13. AAAI Press, Menlo Park, CA.

BUYSE, M., GEORGE, S. L., EVANS, S., GELLER, N. L., RANSTAM, J., SCHERRER, B., LESAFFRE, E., MURRAY, G., EDLER, L., HUTTON, J., COLTON, T., LACHENBRUCH, P. and VERMA, B. L. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* **18** 3435–3451.

CAHILL, M. H., LAMBERT, D., PINHEIRO, J. C. and SUN, D. X. (2002). Detecting fraud in the real world. In *Handbook of Massive Datasets* (J. Abello, P. M. Pardalos and M. G. C. Resende, eds.). Kluwer, Dordrecht.

CHAN, P. K., FAN, W., PRODROMIDIS, A. L. and STOLFO, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems* **14**(6) 67–74.

CHAN, P. and STOLFO, S. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* 164–168. AAAI Press, Menlo Park, CA.

CHARTIER, B. and SPILLANE, T. (2000). Money laundering detection with a neural network. In *Business Applications of Neural Networks* (P. J. G. Lisboa, A. Vellido and B. Edisbury, eds.) 159–172. World Scientific, Singapore.

CHHIKARA, R. S. and MCKEON, J. (1984). Linear discriminant analysis with misallocation in training samples. *J. Amer. Statist. Assoc.* **79** 899–906.

CLARK, P. and NIBLETT, T. (1989). The CN2 induction algorithm. *Machine Learning* **3** 261–285.

COHEN, W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning* 115–123. Morgan Kaufmann, Palo Alto, CA.

CORTES, C., FISHER, K., PREGIBON, D. and ROGERS, A. (2000). Hancock: A language for extracting signatures from data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 9–17. ACM Press, New York.

CORTES, C. and PREGIBON, D. (1998). Giga-mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* 174–178. AAAI Press, Menlo Park, CA.

CORTES, C, PREGIBON, D. and VOLINSKY, C. (2001). Communities of interest. *Lecture Notes in Comput. Sci.* **2189** 105–114.

COX, K. C., EICK, S. G. and WILLS, G. J. (1997). Visual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery* **1** 225–231.

CSIDS (1999). Cisco secure intrusion detection system technical overview. Available at http://www.wheelgroup.com/warp/public/cc/cisco/mkt/security/nranger/tech/ntran_tc.htm.

DENNING, D. E. (1997). Cyberspace attacks and countermeasures. In *Internet Besieged* (D. E. Denning and P. J. Denning, eds.) 29–55. ACM Press, New York.

DORRONSORO, J. R., GINEL, F., SANCHEZ, C. and CRUZ, C. S. (1997). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks* **8** 827–834.

FANNING, K., COGGER, K. O. and SRIVASTAVA, R. (1995). Detection of management fraud: A neural network approach. *International Journal of Intelligent Systems in Accounting, Finance and Management* **4** 113–126.

FAWCETT, T. and PROVOST, F. (1997a). Adaptive fraud detection. *Data Mining and Knowledge Discovery* **1** 291–316.

FAWCETT, T. and PROVOST, F. (1997b). Combining data mining and machine learning for effective fraud detection. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 14–19. AAAI Press, Menlo Park, CA.

FAWCETT, T. and PROVOST, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 53–62. ACM Press, New York.

FORREST, S., HOFMEYR, S., SOMAYAJI, A. and LONGSTAFF, T. (1996). A sense of self for UNIX processes. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy* 120–128. IEEE Computer Society Press, Silver Spring, MD.

GHOSH, S. and REILLY, D. L. (1994). Credit card fraud detection with a neural network. In *Proceedings of the 27th Hawaii International Conference on System Sciences* (J. F. Nunamaker and R. H. Sprague, eds.) **3** 621–630. IEEE Computer Society Press, Los Alamitos, CA.

GLASGOW, B. (1997). Risk and fraud in the insurance industry. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 20–21. AAAI Press, Menlo Park, CA.

GOLDBERG, H. and SENATOR, T. E. (1995). Restructuring databases for knowledge discovery by consolidation and link formation. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* 136–141. AAAI Press, Menlo Park, CA.

GOLDBERG, H. and SENATOR, T. E. (1997). Break detection systems. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 22–28. AAAI Press, Menlo Park, CA.

GOLDBERG, H. and SENATOR, T. E. (1998). The FinCEN AI system: Finding financial crimes in a large database of cash transactions. In *Agent Technology*: *Foundations, Applications, and Markets* (N. Jennings and M. Wooldridge, eds.) 283–302. Springer, Berlin.

GREEN, B. P. and CHOI, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing* **16** 14–28.

HAND, D. J. (1981). *Discrimination and Classification*. Wiley, Chichester.

HAND, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester.

HAND, D. J. and BLUNT, G. (2001). Prospecting for gems in credit card data. *IMA Journal of Management Mathematics* **12** 173–200.

HAND, D. J., BLUNT, G., KELLY, M. G. and ADAMS, N. M. (2000). Data mining for fun and profit (with discussion). *Statist. Sci.* **15** 111–131.

HAND, D. J. and HENLEY, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *J. Roy. Statist. Soc. Ser. A* **160** 523–541.

HASSIBI, K. (2000). Detecting payment card fraud with neural networks. In *Business Applications of Neural Networks* (P. J. G. Lisboa, A. Vellido and B. Edisbury, eds.). World Scientific, Singapore.

HE, H., GRACO, W. and YAO, X. (1999). Application of genetic algorithm and k-nearest neighbour method in medical fraud detection. *Lecture Notes in Comput. Sci.* **1585** 74–81. Springer, Berlin.

HE, H. X., WANG, J. C., GRACO, W. and HAWKINS, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications* **13** 329–336.

HILL, T. P. (1995). A statistical derivation of the significant-digit law. *Statist. Sci.* **10** 354–363.

HYNNINEN, J. (2000). Experiences in mobile phone fraud. Seminar on Network Security. Report Tik-110.501, Helsinki Univ. Technology.

JENKINS, P. (2000). Getting smart with fraudsters. *Financial Times*, September 23.

JENSEN, D. (1997). Prospective assessment of AI technologies for fraud detection: a case study. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 34–38. AAAI Press, Menlo Park, CA.

JU, W.-H. and VARDI, Y. (2001). A hybrid high-order Markov chain model for computer intrusion detection. *J. Comput. Graph. Statist.* **10** 277–295.

KIRKLAND, J. D., SENATOR, T. E., HAYDEN, J. J., DYBALA, T., GOLDBERG, H. G. and SHYR, P. (1998). The NASD regulation advanced detection system (ADS). In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98) and of the 10th Conference on Innovative Applications of Artificial Intelligence (IAAI-98)* 1055–1062. AAAI Press, Menlo Park, CA.

KOSORESOW, A. P. and HOFMEYR, S. A. (1997). Intrusion detection via system call traces. *IEEE Software* **14** 35–42.

KUMAR, S. and SPAFFORD, E. (1994). A pattern matching model for misuse intrusion detection. In *Proceedings of the 17th National Computer Security Conference* 11–21.

LACHENBRUCH, P. A. (1966). Discriminant analysis when the initial samples are misclassified. *Technometrics* **8** 657–662.

LACHENBRUCH, P. A. (1974). Discriminant analysis when the initial samples are misclassified. II: Non-random misclassification models. *Technometrics* **16** 419–424.

LANE, T. and BRODLEY, C. E. (1998). Temporal sequence learning and data reduction for anomaly detection. In *Proceedings of the 5th ACM Conference on Computer and Communications Security (CCS-98)* 150–158. ACM Press, New York.

LEE, W. and STOLFO, S. (1998). Data mining approaches for intrusion detection. In *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX 79–93. USENIX Association, Berkeley, CA.

LEONARD, K. J. (1993). Detecting credit card fraud using expert systems. *Computers and Industrial Engineering* **25** 103–106.

LIPPMANN, R., FRIED, D., GRAF, I., HAINES, J., KENDALL, K., MCCLUNG, D., WEBER, D., WEBSTER, S., WYSCHOGROD, D., CUNNINGHAM, R. and ZISSMAN, M. (2000). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion-detection evaluation. Unpublished manuscript, MIT Lincoln Laboratory.

MAJOR, J. A. and RIEDINGER, D. R. (1992). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. *International Journal of Intelligent Systems* **7** 687–703.

MARCHETTE, D. J. (2001). *Computer Intrusion Detection and Network Monitoring*: *A Statistical Viewpoint*. Springer, New York.

MCCARTHY, J. (2000). Phenomenal data mining. *Comm. ACM* **43** 75–79.

MCLACHLAN, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

MOBILE EUROPE (2000). New IP world, new dangers. *Mobile Europe*, March.

MOREAU, Y., PRENEEL, B., BURGE, P., SHAWE-TAYLOR, J., STOERMANN, C. and COOKE, C. (1996). Novel techniques for fraud detection in mobile communications. In *ACTS Mobile Summit, Grenada*.

MOREAU, Y., VERRELST, H. and VANDEWALLE, J. (1997). Detection of mobile phone fraud using supervised neural networks: A first prototype. In *Proceedings of 7th International Conference on Artificial Neural Networks (ICANN'97)* 1065–1070. Springer, Berlin.

MURAD, U. and PINKAS, G. (1999). Unsupervised profiling for identifying superimposed fraud. *Principles of Data Mining and Knowledge Discovery*. *Lecture Notes in Artificial Intelligence* **1704** 251–261. Springer, Berlin.

NEURAL TECHNOLOGIES (2000). Reducing telecoms fraud and churn. Report, Neural Technologies, Ltd., Petersfield, U.K.

NIGRINI, M. J. (1999). I've got your number. *Journal of Accountancy* May 79–83.

NIGRINI, M. J. and MITTERMAIER, L. J. (1997). The use of Benford's law as an aid in analytical procedures. *Auditing*: *A Journal of Practice and Theory* **16** 52–67.

NORTEL (2000). Nortel networks fraud solutions. *Fraud Primer, Issue* 2.0. Nortel Networks Corporation.

PAK, S. J. and ZDANOWICZ, J. S. (1994). A statistical analysis of the U.S. Merchandise Trade Database and its uses in transfer pricing compliance and enforcement. *Tax Management*, May 11.

PATIENT, S. (2000). Reducing online credit card fraud. *Web Developer's Journal*. Available at http://www.webdevelopersjournal.com/articles/card_fraud.html

PRESS, S. J. and TANUR, J. M. (2001). *The Subjectivity of Scientists and the Bayesian Approach*. Wiley, New York.

PROVOST, F. and FAWCETT, T. (2001). Robust classification for imprecise environments. *Machine Learning* **42** 203–210.

QU, D., VETTER, B. M., WANG, F., NARAYAN, R., WU, S. F., HOU, Y. F., GONG, F. and SARGOR, C. (1998). Statistical anomaly detection for link-state routing protocols. In *Proceedings of the Sixth International Conference on Network Protocols* 62–70. IEEE Computer Society Press, Los Alamitos, CA.

QUINLAN, J. R. (1990). Learning logical definitions from relations. *Machine Learning* **5** 239–266.

QUINLAN, J. R. (1993). *C4.5*: *Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press.

ROBINSON, M. E. and TAWN, J. A. (1995). Statistics for exceptional athletics records. *Appl. Statist.* **44** 499–511.

ROSSET, S., MURAD, U., NEUMANN, E., IDAN, Y. and PINKAS, G. (1999). Discovery of fraud rules for telecommunications—challenges and solutions. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 409–413. ACM Press, New York.

RYAN, J., LIN, M. and MIIKKULAINEN, R. (1997). Intrusion detection with neural networks. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 72–79. AAAI Press, Menlo Park, CA.

SCHONLAU, M., DUMOUCHEL, W., JU, W.-H., KARR, A. F., THEUS, M. and VARDI, Y. (2001). Computer intrusion: Detecting masquerades. *Statist. Sci.* **16** 58–74.

SENATOR, T. E. (2000). Ongoing management and application of discovered knowledge in a large regulatory organization: A case study of the use and impact of NASD regulation's advanced detection system (ADS). In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 44–53. ACM Press, New York.

SENATOR, T. E., GOLDBERG, H. G., WOOTON, J., COTTINI, M. A., UMAR KHAN, A. F., KLINGER, C. D., LLAMAS, W. M., MARRONE, M. P. and WONG, R. W. H. (1995). The financial crimes enforcement network AI system (FAIS)—Identifying potential money laundering from reports of large cash transactions. *AI Magazine* **16** 21–39.

SHAWE-TAYLOR, J., HOWKER, K., GOSSET, P., HYLAND, M., VERRELST, H., MOREAU, Y., STOERMANN, C. and BURGE, P. (2000). Novel techniques for profiling and fraud detection in mobile telecommunications. In *Business Applications of Neural Networks* (P. J. G. Lisboa, A. Vellido and B.Edisbury, eds.) 113–139. World Scientific, Singapore.

SHIEH, S.-P. W. and GLIGOR, V. D. (1991). A pattern-oriented intrusion-detection model and its applications. In *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy* 327–342. IEEE Computer Society Press, Silver Spring, MD.

SHIEH, S.-P. W. and GLIGOR, V. D. (1997). On a pattern-oriented model for intrusion detection. *IEEE Transactions on Knowledge and Data Engineering* **9** 661–667.

SMITH, R. L. (1997). Comment on "Statistics for exceptional athletics records," by M. E. Robinson and J. A. Tawn. *Appl. Statist.* **46** 123–128.

STOLFO, S. J., FAN, D. W., LEE, W., PRODROMIDIS, A. L. and CHAN, P. K. (1997a). Credit card fraud detection using meta-learning: Issues and initial results. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 83–90. AAAI Press, Menlo Park, CA.

STOLFO, S., FAN, W., LEE, W., PRODROMIDIS, A. L. and CHAN, P. (1999). Cost-based modeling for fraud and intrusion detection: Results from the JAM Project. In *Proceedings of the DARPA Information Survivability Conference and Exposition* **2** 130–144. IEEE Computer Press, New York.

STOLFO, S. J., PRODROMIDIS, A. L., TSELEPIS, S., LEE, W., FAN, D. W. and CHAN, P. K. (1997b). JAM: Java agents for meta-learning over distributed databases. In *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management* 91–98. AAAI Press, Menlo Park, CA.

TANIGUCHI, M., HAFT, M., HOLLMÉN, J. and TRESP, V. (1998). Fraud detection in communication networks using neural and probabilistic methods. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)* **2** 1241–1244. IEEE Computer Society Press, Silver Spring, MD.

U.S. CONGRESS (1995). Information technologies for the control of money laundering. Office of Technology Assessment, Report OTA-ITC-630, U.S. Government Printing Office, Washington, DC.

WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press.

WEBB, A. R. (1999). *Statistical Pattern Recognition*. Arnold, London.

WHEELER, R. and AITKEN, S. (2000). Multiple algorithms for fraud detection. *Knowledge-Based Systems* **13**(2/3) 93–99.

Comment

Foster Provost

The state of research on fraud detection recalls John Godfrey Saxe's 19th-century poem "The Blind Men and the Elephant" (Felleman, 1936, page 521). Based on a Hindu fable, each blind man experiences only a part of the elephant, which shapes his opinion of the nature of the elephant: the leg makes it seem like a tree, the tail a rope, the trunk a snake and so on. In fact, "... though each was partly in the right ... all were in the wrong." Saxe's poem was a criticism of theological debates, and I do not intend such a harsh criticism of research on fraud detection. However, because the problem is so complex, each research project takes a particular angle of attack, which often obscures the view of other parts of the problem. So, some researchers see the problem as one of classification, others of temporal pattern discovery; to some it is a problem perfect for a hidden Markov model and so on.

So why is fraud detection not simply classification or a member of some other already well-understood problem class? Bolton and Hand outline several characteristics of fraud detection problems that differentiate them [as did Tom Fawcett and I in our review of the problems and techniques of fraud detection (Fawcett and Provost, 2002)]. Consider fraud detection as a classification problem. Fraud detection certainly must be "cost-sensitive"—rather than minimizing error rate, some other loss function must be minimized. In addition, usually the marginal class distribution is skewed strongly toward one class (legitimate behavior). Therefore, modeling for fraud detection at least is a difficult problem of estimating class membership probability, rather than simple classification. However, this still is an unsatisfying attempt to transform the true problem into one for which we have existing tools (practical and conceptual). The objective function for fraud detection systems actually is much more complicated. For example, the value of detection is a function of time. Immediate detection is much more valuable than delayed detection. Unfortunately, evidence builds up over time, so detection is easier the longer it is delayed. In cases of self-revealing fraud, eventually, detection is trivial (e.g., a defrauded customer calls to complain about fraudulent transactions on his or her bill).

In most research on modeling for fraud detection, a subproblem is extracted (e.g., classifying transactions or accounts as being fraudulent) and techniques are compared for solving this subproblem—without moving on to compare the techniques for the greater problem of detecting fraud. Each particular subproblem naturally will abstract away those parts that are

Foster Provost is Associate Professor, Leonard N. Stern School of Business, New York University, New York, New York 10012 (e-mail: provost@acm.org).

problematic for the technique at hand (e.g., temporal aspects are ignored for research on applying standard classification approaches). However, fraud detection can benefit from classification, regression, time-series analysis, temporal pattern discovery, techniques for combining evidence and others. For example, temporal sequences of particular actions can provide strong clues to the existence of fraud. A common example of such a temporal sequence is a triggering event followed in a day or two by an acceleration of usage. In credit card fraud, bandits purchase small amounts of gasoline (at a safe, automatic pump) to verify that a card is active before selling it. In wireless telephone fraud, bandits call a movie theater information number for verification. In a standard classification framework, temporal patterns must be engineered carefully into the representation. On the other hand, in a framework designed to focus on the discovery of temporal sequences, many facets of advanced classification may be ignored; for example, classifier learners can take advantage (automatically) of mixed-type variables, including numeric, categorical, set-valued and text, and hierarchical background knowledge (Aronis and Provost, 1997) such as geographic hierarchies.

This is just one example of a pair of different views of the problem, each with its advantages and disadvantages. Another is, as Bolton and Hand point out, the supervised/unsupervised duality to modeling for fraud detection: some fraudulent activity can be detected by applying knowledge generalized from past, labeled cases; other activity is better detected by noticing behavior that differs significantly from the norm.

Fraud detection and intervention can have two modes: automatic and mixed initiative (human/computer). Automatic intervention only occurs when there is very strong evidence that fraud exists; otherwise, false alarms would be disastrous. Remember that fraud detection systems consider millions (sometimes tens or hundreds of millions) of accounts. On a customer base of only 1 million accounts, a daily false-alarm rate of even 1% would yield 10,000 false alarms a day; the cost of dealing with these (e.g., if accounts were incorrectly shut down) could be enormous.

Mixed-initiative detection and intervention deals with cases that do not have enough evidence for automatic intervention (or with applications for which automatic intervention does not make sense). Fraud detection systems create "cases" comprising the evidence collected so far that indicates fraud. Fraud analysts process these cases, often going to auxiliary sources of data to augment their analyses. At any time, a case list can be sorted by some score: a probability of fraud, computed from all the evidence collected so far, an expected loss or simply an ad hoc score. The unit of analysis for the production of the score is complicated: it is composed of a series of transactions, which comprises the potentially fraudulent activity and possibly legitimate activity as well. The unit of analysis also could include other information, such as that taken from account applications, background databases, behavior profiles (which may have been compiled from previous transaction activity) and possibly account annotations made by prior analysts (e.g., "this customer often triggers rule X").

A part of the fraud detection elephant that has not received much attention is the peculiar nonstationary nature of the problem. Not only does the phenomenon being modeled change over time—sometimes dramatically—it changes in direct response to the modeling of it. As soon as a model is put into place, it begins to lose effectiveness. For example, after realizing that the appearance of a large volume of transactions on a brand new account is used as an indicator of application/subscription fraud, criminals begin to lie low and even pay initial bills before ramping up spending. After realizing that "calling dens" in certain locations had led to models that detect wireless fraud based on those locations, criminals constructed roving calling dens (where fraudulent wireless service was provided in the back of a van that drove around the city). This adaptation is problematic for the typical information systems development life cycle (analysis $\rightarrow$ design $\rightarrow$ programming $\rightarrow$ deployment $\rightarrow$ maintenance). At the very least it is necessary for models to be able to be changed quickly and frequently. A more satisfying (but perhaps not yet practicable) solution would be to have a learning system, which can modify its own models in the ongoing arms race.

A practical view of the fraud detection elephant shows other issues that make fraud detection problems difficult. They must be kept in mind if one intends results actually to apply to real fraud detection. Systems for fraud detection, in many applications, face tremendous computational demands. Transactions arrive in real time; often only milliseconds (or less) can

be allocated to process each. In this short time, the system must record the transaction in its database, access relevant account-specific data, process the transaction and historical data through the fraud detection model and create a case, update a case or issue an alarm if warranted (and if not, possibly update a customer's profile). Fraud models must be very efficient to apply. Furthermore, the models must be very space efficient. Storing a neural network or a decision tree for each customer is not feasible for millions of customers; it may be possible only to store for each customer a few parameters to a general model. Thus, both time and space constraints argue for simple fraud detection models.

A user perspective of fraud detection (as a mixed-initiative process) argues for the use of models that are comprehensible to the analysts. For example, for many analysts, rule-based models are easier to interpret than are neural network models. The set of rules that apply to a particular case may guide the subsequent (human) investigation. On the other hand, the most commercially successful vendor of fraud detection systems (to my knowledge) uses neural networks extensively for detecting fraud. Of course, commercial success is a dubious measure of technical quality; however, one can get an interesting view into real world fraud detection systems by studying HNC Software's patent (Gopinathan et al., 1998). (As of this writing, a patent search on keywords "fraud detection" yields 80 patents.) In particular, their extensive list of variables, created to summarize past activity so that a neural network can be applied, illustrates the problem engineering necessary to transform the fraud detection problem into one that is amenable to standard modeling techniques.

It would be useful to have a precise definition of a class (or of several classes) of fraud detection problems, which takes into account the variety of characteristics that make statistical fraud detection difficult. If such a characterization exists already in statistics, the machine learning and data mining communities would benefit from its introduction. Not knowing of one, Tom Fawcett and I attempted to define one class of "activity monitoring" problems and illustrate several instances (Fawcett and Provost, 1999). Earlier we defined "superimposition fraud" (Fawcett and Provost, 1997a) to try to unify similar forms of wireless telephone fraud, calling card fraud, credit card fraud, certain computer intrusions and so on, where fraudulent usage is superimposed upon legitimate usage and for which similar solution methods may apply. However, neither of these captures all of the important characteristics.

The characterization of such a class of problems is important for several reasons. First of all, different fraud detection problems are considerably similar—it is important to understand how well success of different techniques generalizes. Is the similarity superficial? Are there deeper characteristics of the problem or data that must be considered? [This seems to be the case, e.g., with classification problems (Perlich, Provost and Simonoff, 2001).] Also, to succeed at detecting fraud, different sorts of modeling techniques must be composed, for example, temporal patterns may become features for a system for estimating class membership probabilities, and estimators of class membership probability could be used in temporal evidence gathering. Furthermore, systems using different solution methods should be on equal footing for comparison. Seeming success on any subproblem does not necessarily imply success on the greater problem. Finally, it would be beneficial to focus researchers from many disciplines, with many complementary techniques, on a common, very important set of problems. The juxtaposition of knowledge and ideas from multiple disciplines will benefit them all and will be facilitated by the precise formulation of a problem of common interest.

Of course I am not arguing that research must address all of these criteria simultaneously (immediately), and I am not being strongly critical of prior work on fraud detection: we all must abstract away parts of such a complicated problem to make progress on others. Nevertheless, it is important that researchers take as an ultimate goal the solution to the full problem. We all should consider carefully whether partial solutions will or will not be extensible. Fraud detection is a real, important problem with many real, interesting subproblems. Bolton and Hand's review of the state of the art shows that there is a lot of room for useful research. However, the research community should make sure that work is progressing toward the solution to the larger problem, whether by the development of techniques that solve larger portions or by facilitating the composition of techniques in a principled manner.

ACKNOWLEDGMENT

Tom Fawcett and I worked very closely on problems of fraud detection, and my views have been influenced considerably by our discussions and collaborative work.

Comment

Leo Breiman

This is an enjoyable and illuminating article. It deals with an area that few statisticians are aware of, but that is of critical importance economically and in terms of security. I am appreciative to the authors for the education in fraud detection this article gave me and to *Statistical Science* for publishing it. There are some interesting aspects that make this class of problems unique and that I comment on, running the risk of repeating points made in the article.

The analysis has to deal with a large number of problems simultaneously. For instance, in credit card fraud, the records of millions of customers have to be analyzed one by one to set up individual alarm settings. It is not a single unsupervised or supervised problem—a multitude of such problems have to be simultaneously addressed and "solved" for diverse data records. Yet the algorithm selected, modulo a few tunable parameters, has to be "one size fits all." Otherwise the on-line computations are not feasible. The alarm bell settings have to be constantly updated. For instance, as customers age and change their economic level and life styles, usage characteristics change. There are also serious database issues—how to structure the large databases so that the incoming streams of data are accessible for the kind of analysis necessary. Collaboration with database experts is essential.

Most of all, these problems require an uninhibited sense of exploration and can be enjoyable adventures in living with data. The goal is predictive accuracy and the tools are algorithmic models (see Breiman, 2001). The class of problems is novel, even in machine learning. No one tool (neural nets, etc.) is instantly applicable to all of these problems. *The algorithms have to be designed to fit the data.* This means that an essential part of the venture is immersion in and exploration of the data. My experience is that good predictive algorithms do not appear by a selection, unguided by the data, from what algorithms are available. Furthermore, the process is one of successive informed revision. If an algorithm, for instance, has too high a false alarm rate, then one has to go back to the data and try to understand why the false alarm rate is high. Understanding will help to lower the false alarm rate.

The process is an alternation between algorithm and data. Personally, if a user reports that an algorithm I have devised gives anomalous results on his data set, the first thing I do is to request that he ship me the data. By running the data myself and trying to understand what it is about the data that causes the poor performance, I can learn a lot about the deficiencies of the algorithm and, possibly, improve it. Granted that with a changing database running to gigabytes and terrabytes, it may be difficult to look at and understand the data. However, this should not deter analysts—in fact, looking for good ways to display and understand the data is an essential foundation for the construction of good algorithms.

There are other difficult boundary conditions in the instances of fraud detection I have looked at. If one tries to design algorithms that use multidimensional information, the problem is that the algorithm may become too wrapped in the individual data and the false alarm rate rises. However, simple and robust algorithms may not utilize enough information to give a satisfactory detection rate.

The choice between supervised and unsupervised learning may be difficult and interesting. Assume that in the database, examples are available of verified fraud and uncontaminated data. As the authors mention, the cases of verified fraud in the data are a tiny fraction of all of the data.

In detecting credit card fraud, for instance, there are two ways to go. The first is to consider one user (G.B.S.) and let his weekly purchases be instances of class 1. Take all records of a week of fraudulent use and assign them to class 2. Then run a classification algorithm on the two class data constructing a method that discriminates between the two classes. Weight the probabilities of class 1 and class 2 assignment so as to keep the false alarm rate down to a preassigned level. Then run the discrimination method on all future weeks of G.B.S.'s purchases.

This, in machine learning, is called supervised learning. It relies on having two labeled classes of instances to discriminate between. Unsupervised learning occurs where there are no class labels or responses attached

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-3860 (e-mail: leo@stat.berkeley.edu)

to the data vectors. Applied to fraud detection, it takes all weekly purchases by G.B.S. in the recent past and summarizes them in a few descriptive statistics. For instance, one could be total average weekly purchases and their standard deviation. If, in the current week, the total purchases exceed the average by many standard deviations, then an alarm bell goes off—that is, a high suspicion score is recorded.

My impression is that, where applicable, supervised learning will give lower false alarm rates. Think of the uncontaminated weekly data for G.B.S. as forming a fuzzy ball in high dimensions. Unsupervised learning puts a boundary around this ball and assigns a high suspicion score to anything outside of the boundary. Supervised learning creates a second fuzzy ball consisting of fraudulent weekly data and assigns a high suspicion score only if the probability of being in class 2 (fraud) is sufficiently higher than being in class 1. Data that are outside of the unsupervised boundary may not be in the direction of class 2. However, the supervised approach makes the assumption that future fraudulent data will have the same characteristics as past fraudulent data and further assumes that fraudulent use of the G.B.S. account will result in characteristics similar to those in the fraudulent use of other accounts.

Fraud detection has some echoes in other areas. For instance, in the 1970s, Los Angeles had metal detectors buried every $\frac{1}{4}$ mile in every lane in a 17 mile triangular section of heavily traveled freeways. Each detector produced a signal as a car passed over it, resulting in estimates of traffic density and average speed. One goal was to use the data from these detectors, channeled into a central computer, to give early warning of accidents that were blocking the traffic flow. However, at the most critical times, when these freeways were operating at near capacity traffic, stoppages in traffic flow could develop spontaneously. Some sections of freeway were more likely to develop stoppages, for example, a slight upgrade or a curve. A false alarm could generate a dispatch of a tow truck, patrol car or helicopter. My mission, as a consultant, was to develop an algorithm, specific to each section of freeway, to detect accident blockages with high accuracy and low false alarm rate.

In astronomy, an important problem is to develop algorithms that can be applied to the finely detailed pictures of millions of stellar objects and locate those that "are unlike anything we're familiar with to date." Here "unlike" does not mean bigger or smaller, but having different physical characteristics than anything seen to date. I have thought about this problem from time to time, but see no satisfactory solution.

In a number of fields a common problem, in both supervised and unsupervised learning, is that the number of data vectors is large, but the number of class 2 cases (i.e., fraudulent data vectors) is an extremely small fraction of the total. Using human judgment to go over a large database and recognize all class 2 data is not feasible. For example, in astronomy, an interesting class of objects are butterfly stars—stars that have a visual picture that resembles a butterfly. A project at the Lawrence Livermore National Laboratory hoped to identify all butterfly stars in a gigabyte database resulting from a sky survey. Working on a small fraction of the data, a team of astronomers identified about 300 butterfly stars.

The goal of the machine learning group working on this project was to identify almost all of the butterfly stars in the survey while requiring minimal further identification work by the astronomers. This required the construction of an optimal incremental strategy. Use the first 300 identifications to find further objects with high probability of being butterflies, ask the astronomers to say "yes" or "no" on these and then repeat using the larger sample.

The challenges in fraud detection are both formidable and intriguing. Many of the problems are nowhere near solution in terms of satisfactory false alarm and detection rates. It is an open field for the exercise of ingenuity, algorithm creation and data snooping. It is also a field worth billions.

The authors titled their paper "Statistical Fraud Detection," implying that this area is within the realm of statistics—would that it were—but the number of statisticians involved is small. The authors write that they are covering a few areas "in which statistical methods can be applied." The list of statistical methods that I extracted from the article are

Neural nets
Rule-based methods
Tree-based algorithms
Genetic algorithms
Fuzzy logic
Mixture models
Bayesian networks
Meta-learning

These were developed in machine learning, not statistics (with the exception of mixture models), and lead to algorithmic modeling. Because of the emphasis on stochastic data modeling in statistics, very few

statisticians are familiar with algorithm modeling, which is sometimes referred to (with a touch of prudishness) as "ad hoc."

We are ceding some of the most interesting of current statistical problems to computer scientists and engineers allied to the machine learning area. Detection of fraud is an example. Young statisticians need to learn about algorithmic modeling and how it applies to a large variety of statistical problems. The Berkeley Statistics Department made a move in this direction a few years ago by making a joint appointment with the Computer Science Department of an excellent scientist in the machine learning area. We will be doing more.

Rejoinder

Richard J. Bolton and David J. Hand

We would like to thank the discussants for their valuable contributions. They have reinforced some of our points and also drawn attention to points which we glossed over or failed to make. Their contributions have significantly enhanced the value of the paper.

We emphasized that many and varied tools would be required to attack the fraud detection problem and this has been echoed by the discussants, who make the additional important point that, whatever subproblems are identified, the tools that are adapted or developed to attack them should do so in combination and to the benefit of the fraud detection process as a whole. The message is that fraud detection is greater than the sum of its parts and that it can be easy to lose sight of this when dissecting the problem. In a similar vein, Provost also rightly draws attention to the fact that there are additional subtleties in applying even standard tools to fraud detection that may not at first be apparent. For example, his observation that the value of detection is greater the sooner it is made, but that detection becomes easier the more time that has passed. In fact, Hand (1996, 1997) suggested that many, if not most, classification problems have such concealed subtleties, and that researchers in statistics and machine learning have typically extracted only the basic form of the problem. So, as tools for classification bump against the ceiling of the greatest classification accuracy that can be achieved in practice, so it becomes more and more important to take note of these other aspects of the problems.

Both discussants comment on the importance of the temporal aspect of fraud. We agree that the incorporation of temporal information into the (commonly) static classification structure is essential in most cases of fraud detection and that further research on tools for tackling this would be of great benefit. Populations evolve as people enter and exit them, but the behavior of individuals who remain in a population can also change. Breiman describes some interesting examples from outside the fraud detection domain which illustrate that there are other applications where statistical research may offer solutions similar to those required for fraud detection. One such domain, which is affected by changing populations, is credit scoring (Kelly, Hand and Adams, 1999). Still on a temporal theme, the adaptability of fraud detection tools to the changing behavior of fraudsters must be addressed so as to ensure the continued effectiveness of a fraud detection system: as new detection strategies are introduced, so fraudsters will change their behavior accordingly. Models of behavior can help with this, although the indicators of fraud that are independent of a particular account may require a different strategy.

We take Breiman's point that many of the methods we described were developed outside the narrow statistical community. However, we had not intended the word "statistical" to refer merely to the stochastic data model-based statistics of his recent article (Breiman, 2001). Rather, we had intended it in the sense of Chambers' "greater statistics" (Chambers, 1993), "everything related to learning from data." Of course, the point that Breiman makes, that the tools we have described have not been developed by conventional statisticians, is something of an indictment of statisticians (Hand, 1998).

We endorse Provost's conclusion about the importance of looking at the full problem. It is all too easy to abstract a component problem and then overrefine the solution to this, way beyond a level which can be useful or relevant in the context of the overall problem. Conversely, it is all too easy to be misled to a focus on a peripheral or irrelevant aspect of the subproblem. Academic researchers have often been criticized for this in other contexts. Of course, the fact is that many of the

subproblems require specialist expertise and specialists in a narrow area may find it difficult to see the broader picture. Moreover, naturally, such specialists will want to apply their specialist tool: to those who have a hammer, everything looks like a nail.

The discussion contributions have emphasized the fact that fraud detection is an important and challenging area for statisticians; indeed, for data analysts in general. Challenging aspects include the large data sets, the fact that one class is typically very small, that the data are dynamic and that speedy decisions may be very important, that the nature of the frauds changes over time, often in response to the very detection strategies that may be put in place, that there may be no training instances and that detecting fraud involves multiple interconnected approaches. All of these and other aspects mean that collaboration with data experts, who can provide human insight into the underlying processes, is essential.

ADDITIONAL REFERENCES

ARONIS, J. and PROVOST, F. (1997). Increasing the efficiency of data mining algorithms with breadth-first marker propagation. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* 119–122. AAAI Press, Menlo Park, CA.

BREIMAN, L. (2001). Statistical modeling: The two cultures (with discussion). *Statist. Sci.* **16** 199–231.

CHAMBERS, J. M. (1993). Greater or lesser statistics: A choice for future research. *Statist. Comput.* **3** 182–184.

FAWCETT, T. and PROVOST, F. (2002). Fraud detection. In *Handbook of Knowledge Discovery and Data Mining* (W. Kloesgen and J. Zytkow, eds.). Oxford Univ. Press.

FELLEMAN, H., ed. (1936). *The Best Loved Poems of the American People*. Doubleday, New York.

GOPINATHAN, K. M., BIAFORE, L. S., FERGUSON, W. M., LAZARUS, M. A., PATHRIA, A. K. and JOST, A. (1998). Fraud detection using predictive modeling. U.S. Patent 5819226, October 6.

HAND, D. J. (1996). Classification and computers: Shifting the focus. In *COMPSTAT-96*: *Proceedings in Computational Statistics* (A. Prat, ed.) 77–88. Physica, Heidelberg.

HAND, D. J. (1998). Breaking misconceptions—statistics and its relationship to mathematics (with discussion). *The Statistician* **47** 245–250, 284–286.

KELLY, M. G., HAND, D. J. and ADAMS, N. M. (1999). The impact of changing populations on classifier performance. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (S. Chaudhuri and D. Madigan, eds.) 367–371. ACM Press, New York.

PERLICH, C., PROVOST, F. and SIMONOFF, J. S. (2001). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*. To appear.