A TAXONOMY TO GUIDE RESEARCH ON THE APPLICATION OF DATA MINING TO FRAUD DETECTION IN FINANCIAL STATEMENT AUDITS

ABSTRACT

This paper explores the application of data mining techniques to fraud detection in the audit of financial statements and proposes a taxonomy to support and guide future research. Currently, the application of data mining to auditing is at an early stage of development and researchers take a scatter-shot approach, investigating patterns in financial statement disclosures, text in annual reports and MD&As, and the nature of journal entries without appropriate guidance being drawn from lessons in known fraud patterns. To develop structure to research in data mining, we create a taxonomy that combines research on patterns of observed fraud schemes with an appreciation of areas that benefit from productive application of data mining. We encapsulate traditional views of data mining that operates primarily on quantitative data, such as financial statement and journal entry data. In addition, we draw on other forms of data mining, notably text and email mining.

Keywords: Auditing, fraud, data mining
A TAXONOMY TO GUIDE RESEARCH ON THE APPLICATION OF DATA MINING TO FRAUD DETECTION IN FINANCIAL STATEMENT AUDITS

I. INTRODUCTION

This study explores the application of data mining techniques to fraud detection as a core component of financial statement audits\(^1\). Although data mining is used in specialized audits (e.g., fraud audits or forensic audits) by accounting firms; data mining is seldom used in traditional financial statement audits. There are a variety of possible reasons for this lack of use. For example, data mining software can have a steep learning curve and, if used improperly, data mining can produce an overwhelming number of false positives and spurious patterns that will require auditors to expend substantial time to subsequently investigate. The primary contribution of this paper is identifying specific fraud and evidence combinations where data mining would be the most effective in traditional audits as well as those combinations where data mining would be least effective. Identifying the more effective use of data mining could encourage auditors to include data mining as a regular element of their analytical procedures. Future researchers can build on our exploratory findings to further refine the application of data mining in financial statement audits.

Specifically, the paper proposes a taxonomy that includes three components, namely, account schemes and evidence schemes (as defined by Gao and Srivastava (2011))\(^2\), and data mining functionality to identify the most effective combinations of those three components. Data

---

\(^1\) While the paper focuses on the detection of fraud within the financial statement audit conducted by external auditors, most of the key messages are also relevant for internal auditors.

\(^2\) Gao and Srivastava (2011) divided fraud schemes into two components: account schemes reflecting the accounts impacted by the fraud (e.g., fictitious revenue) and evidence schemes reflecting how the fraudster implemented the fraud (e.g., fake documents).
mining has the potential to enhance the efficiency and effectiveness of the audit. Productive data mining toolsets are now more widely available and auditors have access to a cornucopia of audit-relevant information both internal and external to the client organization. Internal data can include financial data, non-financial data, and email archives. Externally, an array of quantitative information on organizations is now available on the Internet and in commercial financial and textual databases. These include news reports, blog postings, and Twitter feeds. For public companies, regulatory filings such as the filings made on the SEC’s EDGAR database in XBRL format are available.

There has been an increased interest in data mining for fraud detection in the regulatory and professional domain. For example, the SEC has developed an “Accounting Quality Model,” designed to identify anomalous financial statement filings to the Commission. The tool mines the XBRL data repository along with other datasets (Lewis 2012; Rohman and Berg 2013). The Advisory Committee on the Auditing Profession (ACAP) to the U. S. Treasury recommended that the PCAOB establish a fraud center that would, in part, facilitate “sharing of practices, and data and innovation in fraud prevention and detection methodologies and technologies” (ACAP 2008, VII:1). To assist auditors with data availability, the AICPA is promoting an “audit data standard” (Titera 2013; Zhang et al. 2012). The objective of the standard is to facilitate ready extraction of entity-level transactional data for audit interrogation. This standardization of data elements that can be requested from different audit clients will increase the economy of scale of using data mining software, which in turn can help justify the cost of learning the software and developing data mining routines that can be applied to multiple clients.

---

3 The Audit Data Standards are available at:
http://www.aicpa.org/InterestAreas/FRC/AssuranceAdvisoryServices/Pages/AuditDataStandardWorkingGroup.aspx
In the academic arena, there has been a small but growing literature that applies data mining techniques to auditing in general and fraud detection in particular (Alden et al. 2012; Janse et al. 2010; Perols 2011; Ravisankar et al. 2011). This comes after a hiatus of more than a decade, and results in part from the increased availability of data sets, such as the EDGAR database, and from improved data mining tool. Despite this recent spurt in research, much of the existing research does not achieve the greatest possible return on research investment. Researchers take a scatter-shot approach, investigating patterns in financial statement disclosures, text in annual reports and MD&As, and the nature of journal entries without appropriate guidance drawn from lessons in known fraud patterns. The taxonomy we present in this paper combines research on where patterns of observed fraud with an appreciation of those areas where data mining can be productively applied in the audit process.

The organization of the remainder of the paper is as follows. Section II explores the nature of data mining and recent research on the application of data mining to the financial statement audit. Section III discusses the application of data mining to each phase of the audit to illustrate that data mining can have wide applicability to an audit. Section IV describes the development of our taxonomy of fraud schemes (account schemes and evidence schemes) and data mining techniques. Section V provides concluding comments and sets out future research directions.

II. LITERATURE REVIEW

In this section, we introduce key techniques of data analysis, data extraction, and data mining that auditors can apply in their conduct of the audit. We apply a number of examples from the audit domain. We then introduce recent research that applies a variety of data mining techniques to the audit and fraud context.
Data Analysis and Extraction versus Data Mining

Practicing auditors are rather imprecise when considering the nature of data mining, confounding data extraction, data analysis, and data mining. Figure 1 demonstrates the conceptual relationship between data extraction, data analysis, and data mining. As the auditor moves from data extraction to data analysis to data mining, the software becomes more sophisticated in terms of functionality and provides a greater amount of automated diagnostic and predictive power. The triangle in Figure 1 indicates the relative frequency of the use of these different techniques by auditors based on our review of professional literature. The following paragraphs provide an overview of the different categories of data examination tools included in Figure 1.

Inset Figure 1 about here

Data Extraction and Query

For financial statement audits, Excel, ACL, and CaseWare IDEA are illustrative of the most frequently used tools to examine client data. These tools are essentially used as data query and extraction tools that perform data analysis with a variety of descriptive statistics and a limited range of statistical techniques. ACL and CaseWare IDEA also include tools to prepare samples of populations for further audit procedures. These and similar tools traditionally have been called computer assisted audit tools and techniques (CAATTs) (Coderre 2009). The auditor subsequently performs any actual analysis, in the investigative sense of the word, of the data. With these tools, the auditor has already made the decision as to what data he/she wants to extract. For example, if a company has an internal control that requires the controller sign checks to vendors greater than $5,000, the auditor will use one of these tools to extract check numbers and related data for all checks between, say, $4,900 and $4,999. This extraction is used because
the auditor believes that there is a risk that a fraudster will write fraudulent checks just below the $5,000 threshold. The auditor will investigate those checks and supporting documentation to determine if any checks violate the control or are fraudulent.

We do not discount data extraction tools for enhancing audit efficiency. Indeed, data extraction software provides powerful audit tools. Queries that fully exploit the functions built into ACL and CaseWare IDEA can be effective in identifying suspicious data patterns in client datasets. These tools, however, are essentially in use as data query and extraction tools to reduce the client’s full population of data to smaller, more-manageable population of data. After extraction, the auditor performs an analysis or investigation (frequently manually) of the extracted sample from the population. It is up to the auditor’s training and experience to identify any suspicious extracted items that require additional investigation.

**Data Analysis**

Data analysis toolsets available to auditors provide a range of analytical techniques from simple to relatively sophisticated. For example, at the low end, data analysis includes basic descriptive statistics such as counts, minimums, maximums, dispersion, and means. Low-end analysis also may include ratio analysis. At the higher end, data analysis can include inferential statistics such as univariate and multivariate regression, canonical correlation analysis and other outputs of statistical software. Like data extraction, the principal limiting factor on the use of data analysis tools comes from the collective knowledge of the auditors. For large organizations, it is easy for auditors to be overwhelmed by significant volumes of data (Eppler and Mengis 2004).
Data Mining

As illustrated in Figure 1, data analysis and data mining definitions and examples overlap. A term frequently associated with data mining is knowledge discovery (Delen and Al-Hawamdeh 2009). Data mining is about discovering patterns, rules, or models based on one or more populations of data. The results (patterns, rules, or models) that can predict future outcomes that fall outside the predicted ranges are red flags that the auditor should investigate. Data mining techniques can be divided into two broad categories—directed (or top-down approach) and undirected (or bottom-up approach). With directed data mining, there is identification of a specific variable of audit interest. Data mining technology finds relationships between that variable and a selected population of other variables. With undirected data mining, there is no specific targeted variable (dependent variable); instead the objective is to find any relationship between any variables in a population of data. Another way to characterize these two broad categories is to say that the top-down approaches test specific hypotheses and the bottom-up approach generates new hypotheses.

Text mining is a growing form of data mining that understands statistical and linguistic patterns in bodies of text (Berry and Kogan 2010; Srivastava and Sahami 2009; Weiss 2010; Witten 2005). With enhanced availability of large corpi of text from the Internet and elsewhere, text mining has become an increasingly important and widely used family of techniques. When text is communicated within a community of interest, for example by emails, social network analysis can be applied to the corpus (Debreceny and Gray 2011; Scott 2013; Worrell et al. 2013).
Application of Data Mining to Auditing and Fraud

In recent years, there has been an increase in the application of data mining techniques to financial statement audits. In this subsection, we review research that uses data mining to understand potential fraudulent patterns in financial statement disclosures, business processes within the corporation, journal entries and in text generated within and by the corporation.

The most long standing application of data mining has been in mining financial statement disclosures (Fanning and Cogger 1998; Feroz et al. 2000; Green and Choi 1997). After a hiatus of a decade, there again has been an increasing attention to mining of financial statement data for identification of potentially fraudulent corporate reports. For example, Ravisankar et al. (2011) data mine core financial statement data points (e.g. Net and Gross Profit) and ratios (e.g. Long term debt/Total capital and reserves) to identify fraudulent reports. Their data set is 101 Chinese corporations with known frauds, matched with non-fraud corporations. Ravisankar et al. (2011) employ a range of data mining techniques including support vector machines, neural networks and genetic programming. One of the neural network techniques, Probabilistic Neural Network, was the most productive, with a surprising greater than 90% ability to identify correctly the corporations in the data set. Alden et al. (2012) employ Evolutionary algorithms (EAs) to identify fraud in financial statements, with relatively high discriminant capability.

Perols (2011) addresses issues that arise with the identification of fraud in financial statements. These include low levels of observed fraud in a given population of audited entities, variations in cost between, for example, false positives, and false negatives (with the latter being dramatically more expensive to the auditor and capital markets than the former) and the messy nature of financial statement data. After accounting for these characteristics in the design of his data set, Perols finds that logistic regression and the widely used support vector machines (SVM)
classification technique both perform well in identifying fraudulent firms and outperform other techniques such as different forms of neural networks. As Perols notes, logistic regression and SVM are both well understood and relatively efficient to deploy in production environments.

Another growing strand in the application of data mining to auditing is process mining (Jans et al. 2013; Van der Aalst 2011). This technique extracts business process knowledge from event logs generated by corporate information systems, typically ERP systems. Process mining research has been carried out on mining corporate decision making processes (van der Aalst et al. 2011), compliance and risk management (Caron et al. 2013) and control structures (Elsas 2008). While process mining could involve a variety of data sources, such as workflow and role analysis in ERP systems, most recent research concentrates on mining the process knowledge inherent in activity logs generated by application systems, such as ERP applications. Alles et al. (2006) and Jans et al. (2010) both develop systems to capture and then mine, processes from large scale log files. In the area of employee fraud, Jans et al. (2010) employ univariate and multivariate clustering techniques to identify potentially fraudulent transactions in a large data set of purchase requisitions within one corporation. As is typical in this class of data mining, the analysis cannot categorically identify actual frauds, and the number of questionable purchase orders identified is too large for human investigation. A sample must be taken and subject to further inspection.

In establishing the opportunity to commit fraud, understanding the roles that users can play within enterprise systems is an important input to audit of internal controls and of particular transactions. Understanding of roles closely aligns to analysis of segregation of duties. In one of the first papers on role mining, Colantonio et al. (2011) establish a methodology for role modeling and apply this to a large set of role data from a major corporation.
Arguably, the area where auditors currently conduct the most analysis of large-scale datasets is the analysis of client journal entries required under SAS No. 99. Debreceny and Gray (2010) take the first step in data mining journal entries. Journal entries have unusual characteristics that derive from the underlying business processes that give rise to the entries. Debreceny and Gray (2010) mine a corpus of 29 sets of journal entries from practice. They show that some techniques, such as Benford’s Law, do not apply to this corpus. Unfortunately, the authors did not have access to known fraudulent journal entries and as a result do not employ more sophisticated data mining techniques. Argyrou (2012) employs Self Organizing Map (SOM) to identify suspicious transactions in a corpus of journal entries for a single corporation. Argyrou seeds errors in the corpus. The SOM technique identifies the errors with a high degree of accuracy and analyzes the cost of misclassification of entries between Type I and II errors.

An increasing area of interest is the application of a range of text data mining techniques to the audit. One of the most important sources of internal corporate communications is email, which is an all-pervasive method for communication within entities and between staff in those entities and the ecosystem of suppliers, customers, advisors and consultants. Text is at the heart of the email conversation. The semi-structured nature of emails with date, recipient and subject fields add important temporal and social network dimensions to the data mining of emails. Further, given the importance of maintaining emails for discovery in future litigation and regulatory requirements has meant that organizations increasingly keep well-structured email archives. Debreceny and Gray (2011) introduce and analyze the text mining and social network analysis that apply in the data mining of emails from a fraud and audit perspective. They provide a practical example of social network analysis of emails, based on an audit-driven analysis of the
publicly available Enron archive of emails. They provide a structured evaluation of potential research in the application of data mining of emails to audit and fraud detection.

An active thread is the application of text mining to fraud is deception analysis. This approach to research leverages well known cues of intentional misrepresentation by content deception in language made by senders intending to deceive information recipients (Debreceny and Gray 2011). Known patterns of deceptive language include higher levels of active language and negative emotion in the writing. Foundational theories include Information Manipulation Theory (IMT) (McCornack 1992) and Interpersonal Detection Theory (Buller and Burgoon 1996). Humpherys et al. (2011) employ linguistic cues in Management’s Discussion and Analysis (MD&A) to identify possible deception that may be indicative of fraudulent misrepresentation. Humpherys et al. (2011) take 101 actions by the SEC in the Commission’s Accounting and Auditing Enforcement Releases (AAERs) as being examples of financial statement fraud. They matched 101 fraudulent examples with a similar number of non-fraudulent examples. Using a model that identified characteristics such as active language, affect, word length and sentence complexity, they were able to develop a model that correctly identified 67% of fraud and non-fraud examples. Glancy and Yadav (2011) analyze deception in both the MD&A and notes to the financial statements. They propose a new model for detecting deception, which they term the Computational Fraud Detection Model (CFDM), with generation of Singular Value Decomposition vectors (SVD) that is typical of data mining, underpinning their model. They use a commercial mining product (SAS’s Enterprise Miner) to undertake the mining. As with Humpherys et al. (2011), Glancy and Yadav (2011) use SEC AAERs to identify fraud. Glancy and Yadav are able to identify potentially fraudulent corporations with reasonable
accuracy, using textual disclosures that predate the AAER. In other words, these very preliminary results are able to proactively identify potentially fraudulent corporations.

While this early flowering and rejuvenation of research on data mining is encouraging, there is little direction. Most of the papers seem opportunistic in nature. With the important exception of Perols (2011) and to some extent of (Humpherys et al. 2011), the research is largely devoid of understanding the audit environment, known patterns in fraud or of the implementation of data mining in professional service firms. We address some of these issues in the next sections.

III. DATA MINING AT EACH PHASE OF THE AUDIT

This section introduces the objectives of the audit and its various phases from a fraud detection perspective. The section explores expanding the auditor’s data domain to improve auditor performance in fraud detection. While auditors may feel overwhelmed by the data they currently examine in the audit, such data comprises only a small part of the client’s potentially fraud-relevant information. In addition, there is a wide range of data external to the client that is potentially relevant in assessing possible material misstatements arising from fraud. SAS No. 1 mandates that the auditor “has a responsibility to plan and perform the [financial statement] audit to obtain reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud” (PCAOB 2003). The focus of this study is on the detection of fraud. SAS No. 99 decomposes fraud into fraudulent financial reporting and misstatements arising from the misappropriation of assets. While the latter class of fraud is important, it is not, as we will discuss in more detail later, as significant as fraudulent financial reporting (PCAOB 2002). SAS No. 99 (para. 06) defines fraudulent financial reporting as “intentional misstatements or omissions of amounts or disclosures in financial statements
designed to deceive financial statement users where the effect causes the financial statements not to be presented, in all material respects, in conformity with generally accepted accounting principles (GAAP)” (emphasis added) (PCAOB 2002). SAS No. 99 and the other auditing standards require that the auditor undertake a variety of analytical and planning tasks and substantive audit procedures to support the detection of errors arising from fraudulent financial reporting. A stylized view of the various phases of the audit is shown in Figure 2.

Phases of the Audit

We now address the nature of each of these phases in turn; pointing to the potential role that data mining may play in each phase.

Understanding the client

The assessment of risk of material misstatement whether from error or fraud drives the audit. Developing an understanding of the client is a key aspect of the high-level audit planning. As part of the early stage of the planning process, the auditor must understand a variety of client-related risk factors. These include ownership and organizational structures; the nature of value adding processes; business partner and related party relationships, and the regulatory environment in which the client operates. Given the complexity of the typical audit client in terms of business processes, organizational structures, and business partner relationships, there is potential to exploit data mining tools and techniques to increase the auditor’s analytical power over client performance data, external relationships, and networks. Typically, much of this data mining will focus on readily available data external to the client such as financial statements, press releases and analyses, analyst reports, share prices, regulatory filings etc. Higher quality
performance data is within the client and is open to data mining. The cost of acquiring and interpreting this latter class of data is relatively high, however, and careful attention is necessary to ensure that appropriate data mining techniques are employed for high value outcomes.

**Audit planning**

Detailed risk-based audit planning follows this high-level risk assessment of the client. As part of the planning process, the auditor normally will undertake a variety of analytical procedures to develop expectations of realized account balances such as levels of debt, cash generation, and levels of accruals. Often auditors will undertake relatively simplistic ratio analysis of core financial statement data points to generate these expectations. Data mining that combines analysis of external data of national and industry trends with client-level data may make this part of the audit more effective and efficient. National and industry levels of output, revenue and profitability are averages, and clients will necessarily perform at a different level than these averages. Nonetheless, data mining of this combined data set is likely to deliver superior results in the development of audit plans.

**Planning fraud-related procedures**

SAS No. 99 requires that the auditor systematically address the risks of potential error arising from fraud as part of the planning process. Auditors are required to discuss fraud risks in the audit planning process in a methodical fashion, including undertaking active brainstorming by the audit team. Completing a variety of analytical procedures is also an important component of this phase of the audit. SAS No. 99 appropriately notes that traditional analytical procedures may be at too high a level of aggregation. For example, SAS No. 99 notes in respect of potentially fraudulent reporting of revenue:
In planning the audit, the auditor also should perform analytical procedures relating to revenue with the objective of identifying unusual or unexpected relationships involving revenue accounts that may indicate a material misstatement due to fraudulent financial reporting. An example of such an analytical procedure that addresses this objective is a comparison of sales volume, as determined from recorded revenue amounts, with production capacity. An excess of sales volume over production capacity may be indicative of recording fictitious sales. As another example, a trend analysis of revenues by month and sales returns by month during and shortly after the reporting period may indicate the existence of undisclosed side agreements with customers to return goods that would preclude revenue recognition.(para. .29)

While these examples are relatively naïve trend ratios analyzed temporally, they provide a flavor of the audit risk assessment procedures that can be undertaken with data mining tools. For example, rather than a straightforward ratio analysis of known metrics (e.g. inventory turnover, product line profitability), data mining of several metrics could lead to the development of a more sophisticated model of the client’s value adding processes and risk profile. Another vital part of the planning stage is to consider the risks arising from weaknesses in internal controls. The SAS No. 99 standard (paras. .44 and .45) requires auditors to assess the design and implementation of controls intended to prevent fraud. As part of their commitment to monitor the effectiveness of internal controls, clients often prepare extensive matrices of account assertions and the design and existence of controls. These matrices can be subject to relatively straightforward data analysis for assessment of risks and key controls for subsequent testing.
**Evidence on controls**

Following the planning phase, the collection of audit evidence commences. A key aspect of this process is to assess the operating effectiveness of internal controls. Process mining on the operation of internal controls throughout the operating cycle is a potential application of data mining. Process mining typically operates on the system logs of, for example, ERP systems that underpin the general ledger. Conducting process mining on control processes, and attempted and actual control overrides is an example of evidence collection on the operation of internal controls.

**Evidence on potential fraud considerations**

When undertaking audit fieldwork, the auditor conducts substantive procedures, some of which relate specifically to the detection of fraud. For example, SAS No. 99 notes that substantive analytical procedures in respect of revenue may exploit “disaggregated data, for example, comparing revenue reported by month and by product line or business segment.” Data mining on corporate performance data and revenue databases and other client databases will be potentially important as examples of fraud-related analytical procedures. Apart from substantive analytical procedures, SAS No. 99 notes that changes in the nature, timing, and extent of substantive audit procedures are required as the auditor responds to fraud-specific risks assessments.

**Concluding the audit, review and reporting**

SAS No. 99 requires the auditor to evaluate the likelihood of material misstatement due to fraud “at or near the completion of fieldwork” (para. .74) and respond appropriately. Such responses may include re-evaluating the use of data mining procedures rejected as cost-ineffective earlier in the planning process or opening up new data mining procedures because of
red flags observed during the fieldwork. For example, substantive analytical procedures or other tests on revenue may indicate need for mining of, for example, key executive emails as we discuss in more detail later in this paper. Following this process, final review of audit evidence is undertaken, leading to decisions on the nature of the final audit report. It is unlikely that data mining will play a role in this review and reporting phase.

Planning the Use of Data Mining

The discussion in the previous paragraphs shows how data mining might be employed at each phase of the audit. When considering the use of data mining the auditor must undertake a data examination process, as illustrated in Figure 3.

First, the audit team must decide what variables to examine to meet their audit objectives—and to define what constitutes red flags. These client-related internal and external variables would be selected based on audit firm’s policies and procedures; audit templates that are developed by the firm for specific audit objectives; the fraud-related brainstorming session of the audit team, and the variables that might be suggested by the individual auditors based on their own experiences or something they have seen at the client during the current audit or past audits.

Second, those internal and external variables must map to the variables in the data analysis and mining software that the auditor employs. For example, if the auditor wanted to run a query that included vendor ID, check ID, check amount, and check date of the client’s database, the auditor would have to determine the appropriate field names and corresponding tables that are in a client’s database. Third, the auditor would have to determine which
procedures and tools they would use to examine the collected data. They could create queries to extract data, determine specific analysis to perform on the extracted data, and/or decide on which data mining techniques are most appropriate. Fourth, the auditor must then investigate the resulting red flags and determine what actions to take next based on the investigation results.

In terms of the investigation, if the auditors used simple data queries that have little or no built in diagnostic value. In these circumstances, the auditor is essentially conducting the complete investigation. With data analysis tools the tools can help pinpoint transactions that are more suspicious and, thereby, increasing the auditor’s productivity. Finally, data mining tools would have the highest level of diagnostic ability and, as such, would reduce the amount of subsequent investigation that the auditor team must undertake.

IV. FRAUD PATTERNS AND DATA MINING TAXONOMY

The application of data mining techniques to fraud detection requires 1) analysis of fraud risks; 2) identification of potential methods undertaken to commit the potential fraud; 3) determination of the availability of indicators of the fraud and associated methods, and 4) selection of appropriate data mining techniques most likely to discover these indicators. The previous sections pointed to possible use of data mining techniques and sources of data in the context of the financial statement audit. It would be a major undertaking if auditors attempted to evaluate all data mining techniques and all data elements in all data sources. The objective of this section is to provide guidance as to the most effective and least effective application of data mining for fraud detection. We develop a taxonomy with three components. The first two components are the two components of fraud schemes, namely, account schemes (accounts impacted by the fraud) and associated evidence schemes that fraudsters have adopted to undertake the fraud drawing upon the work of Gao and Srivastava (2011). The third component
is the applicability of data mining techniques to each combination of the account scheme and evidence scheme, but based primarily on the characteristics of the evidence scheme. The most fundamental question is: Does the evidence scheme include something that can be data mined? Some of the evidence schemes identified by Gao and Srivastava can cost effectively be subjected to direct use of data mining techniques. Others require data mining and assessment of indirect signals. We draw this discussion together in a set of recommendations for areas of research and data mining projects that the researchers and the auditing profession might productively invest resources.

**Classes of Account Schemes and Associated Evidence Schemes**

While theft or misappropriation of assets represents a significant area of concern, earnings management has represented the most significant frauds in recent decades. Earnings management is placed in a continuum from genuine and justifiable choices on accruals to deliberate actions that are entirely without merit and are fraudulent (POB 2000). Gao and Srivastava (2011) develop a fraud taxonomy that considers both the type of fraud and the method employed to implement and hide the fraud. Gao and Srivastava describe this latter aspect as the *evidence scheme*. They studied SEC Enforcement Actions from 1997 until 2002 as the foundation for the instantiation of their fraud taxonomy. They found 100 enforcement actions in that period for which information existed on evidence schemes, representing 148 individual fraudulent events. Table 1 lists the account schemes (classes of fraud), in descending order of occurrence. Unsurprisingly, given the traditional focus on fraudulent revenues and revenue recognition, Gao and Srivastava identify significantly more than half of the fraudulent actions as involving revenues.
Gao and Srivastava also studied the evidence schemes involved in each of these frauds. Fraudsters use evidence schemes for dissimulation (“hiding the real”) or simulation (“showing the false”). The taxonomy of evidence schemes include fake documents (e.g. “fictional shipping documents” or “fabricated stock certificates”); hidden documents (e.g. “keeping secret collection memoranda to track collections of contingent transactions”), and collusion with third parties (e.g. “requesting customers to provide false audit confirmations”). Table 2 lists the evidence schemes, again in descending order of occurrence.

Gao and Srivastava provide data which correlates occurrence of frauds and evidence schemes. Figure 4 provides a visual representation of this correlation. The three black cells represent the highest combination of account scheme and evidence scheme frequencies. These cells represent areas where data mining could potentially provide significant returns. The many white cells represent no intersection between the particular account scheme and the particular evidence scheme. There would be little reason to invest significant research or other resources, at least in the short term, in building data mining techniques or skills for these null cells.

Fraud Scheme Indicators

The relative frequencies of combination of account schemes and evidence schemes shown in Figure 4 provide guidance for auditors in both the planning and fieldwork phases of the audit. The key question is how can data mining identify fraud indicators for these combinations?

---

4 Gao and Srivastava draws this formulation from Bowyer (1982).
As discussed below, some of the evidence schemes are subject to direct and cost-effective identification by data mining techniques. Other evidence schemes will require identification of secondary evidence for indicators that may lead the auditor to the direct evidence within that scheme.

Evidence schemes that directly influence information within the accounting information system, broadly defined, are tractable with data mining approaches. For example, frauds involving the “spreading of fraudulent items among accounts” and “reversal accounting entries” are likely to be highly tractable with mining of journal entries or transactions that underlie the journal entries. Fraud indicators for evidence schemes that do not directly affect the accounting information system are more difficult and require analysis of secondary evidence. For example, it is difficult but not impossible to mine application transaction databases for “hidden” or “altered” documents. In the case of revenue frauds, for example, there must be journal entries in the sales systems or general ledger to book revenue for fabricated sales transactions. This class of transaction might well be highlighted as an outlier in a database of journal entries because of multiple indicators (e.g. some unusual combination of the date of the transaction(s), atypical pattern of transactions for a single client, journal entry description etc.).

Collusion with third parties and fraudulent related party transactions are particularly difficult to detect directly. However, it is possible that prima facie evidence of collusion will come from the accounting information system. For example, in a fraudulent revenue scenario, there may be indicators of unusual relationships such as dramatically increased sales to a client or higher profit margins in a product group or client, perhaps coupled with higher days outstanding in receivables. Supplementary mining of emails with the name(s) of the product group or client and of those managers associated with approval of sales may bring evidence of
collusion to light. Of course, if executives are determined to hide collusion, they will ensure that not all interactions with the relevant third parties or with other executives within the corporation use the corporate email system. The reality, however, is that these systems are so much part of the corporate DNA that executives fall back on them even for communications that are subsequently clear evidence of wrongdoing.

SAS No. 99 provides some guidance on these types of secondary evidential matter. The standard suggests that auditors analyze key performance metrics longitudinally and cross-sectionally in comparison to peers for evidence of unusual and potentially fraudulent financial reporting patterns. At a much more detailed level, as we discussed earlier, the standard suggests that evidence of potentially fraudulent revenue may rest in a variety of detailed analytical procedures. These procedures may require a drill down into the nature, type and timing of sales transactions. The auditor should also assess revenue-associated events affecting accounts such as inventory, accounts receivable, and cash. Related indicators may also include analysis of sales levels with production volume. The procedures suggested by SAS No. 99 imply that not only does the auditor have to maintain a detailed understanding of the value-adding processes of the client, but also makes an in-depth analysis of key performance metrics and financial statement relationships. These metrics and relationships will need to be assessed for the client as a whole as well as core subsets such as geographic and product categories.

Paragraph .68 of SAS No. 99 provides guidance to the auditor on the assessment of fraud risks. Some of the risks identified in Paragraph .68 are identified during the conduct of the audit as a by-product of other procedures. For example, SAS No. 99 alerts auditors to “inconsistent, vague, or implausible responses from management or employees arising from inquiries or analytical procedures,” “complaints by management about the conduct of the audit or
management intimidation of audit team members, particularly in connection with the auditor’s critical assessment of audit evidence or in the resolution of potential disagreements with management” and “undue time pressures imposed by management to resolve complex or contentious issues.” These risk factors provide color in the conduct of the audit, but are not subject to testing directly by substantive procedures involving data mining or otherwise.

Conversely, guidance in Paragraph .68 to auditors to assess the following risks all relate in some way to potential data mining:

- Last-minute adjustments that significantly affect financial results
- Unsupported or unauthorized balances or transactions
- Evidence of employees’ access to systems and records inconsistent with that necessary to perform their authorized duties
- Significant unexplained items on reconciliations

The accounting DNA is represented in the journal entries in the client’s general ledger. Sections .58 through .61 of SAS No. 99 mandates a set of procedures to test the accounting information system and the journal entries therein for potential misstatements arising from fraud. As SAS No. 99 notes, a number of financial statement frauds have involved “inappropriate or unauthorized” journal entries as well as adjustments made outside the formal accounting system, such in the consolidation process leading to a set of consolidated financial statements or even by manual entries to the draft of financial statements. These latter adjustments are beyond the scope of data mining. Evidence of such frauds are seen directly within, for example, specialist

---

5 SAS No. 99 followed the report of the then Public Oversight Board’s Panel on Audit Effectiveness (POB 2000). The Panel’s report provides further elucidation of the factors involved in financial statement fraud involving journal entries and otherwise.
computerized consolidation systems or by a process of reconciliation from the final adjusted trial balance to the final reported financial statements.⁶

**Overlaying Data Mining onto the Fraud Scheme Taxonomy**

At the beginning of this section, we set out a taxonomy of fraud schemes, which includes account schemes and evidence schemes. We first set up a categorization of data mining targets that is fine-tuned for the external auditing domain, illustrated in Figure 5. The critical concept underlying Figure 5 is that it provides an analysis of where data mining can be productively applied within the context of the external audit. The likelihood that data mining can be applied successfully in the audit setting is scored. A variety of data mining techniques can be applied to each of the data mining targets. However, not all targets are equally productive targets. Factors include the connection of the information source to the key concerns of the financial statement audit, the availability of data to the audit and the quality of the semantic representation of the information source. As a result, we rank the productivity of each of the rows in Figure 5, and return to the method for ranking shortly. First, we explain each of the columns in Figure 5.

---

*Insert Figure 5 about here*

---

The first dimension represents source of information that exists within the client’s information system and those that exist outside the client (column A). We then further sub-divide the information sources within each of these two principal classes (column B). The first two classes of data mining targets within the client’s information system are those that connect in some ways to the client’s accounting information system. The “AIS Core” represents the heart of the accounting information system, represented primarily by the General Ledger or equivalent

---

⁶ A closely related class of financial statement fraud is inappropriate disclosure in footnotes or misclassification within the financial statements.
system or module in enterprise systems. There are two targets for data mining within the AIS core – journal entries that represent the totality of transactions linking all accounts within the accounting period. The second class of data mining target within the core is analysis of transactions over time within a particular account or class of accounts.

AIS “connected” are transaction systems that feed journal entries to the AIS core. The first component of AIS connected is the set of applications that exist within the enterprise system that are tightly coupled with the General Ledger. Examples include customer relationship management, inventory management, sales and payable applications. The focus of this data mining target is on aspects of the early part of the lifecycle of the transaction. The second data mining target in this category is a variety of application systems that pass summary journal entries to the General Ledger but which are not necessarily tightly coupled. Examples include maintenance and production systems, in the manufacturing environment. For both “AIS core” and “AIS connections,” we recognize that there will be considerable variation between clients depending on factors such as the client’s adoption of enterprise systems, geographical distribution, and number of divisions. The next category of data mining targets is the class of client information systems that do not connect to the General Ledger, but which have audit-relevant information. This category has two classes of data mining targets: corporate email and other document and information repositories.

The next category of data mining targets are the formal disclosures upon which the auditor either is directly responsible for auditing or which are closely associated with the audit objective. These targets include the financial statements and notes and the MD&A as well as press releases and regulatory filings, such as 8-K filings with the SEC, closely associated with financial reporting process. The financial statements and MD&A will include both prior
quarterly and annual statements, as appropriate, as well as the draft of the current reports. The press releases and regulatory filings will run throughout the fiscal year.

The second principal source of information for data mining exists outside the client’s information system, divided between those that are related to finance and broader sources. Finance-related data mining targets include the XBRL filings within the SEC’s EDGAR system and other EDGAR filings, which a has large set of highly disaggregated, entity-specific disclosures. We differentiate the XBRL filings to the SEC from analysis of the financial statements discussed above, as the difference between peer analyses, utilizing all filings to EDGAR, from client-level analysis using only the client’s financial statements. There are also a variety of financial databases (e.g. S&P Capital IQ) that can be mined for fraud detection. In addition, stock prices and volumes can provide pointers to potential fraud. Finally, in this group of data mining targets is LEXIS/NEXIS, which can provide analysis of news items, court filings and a variety of other information sources. Information sources with potential for data mining beyond finance include social media, such as Twitter and Facebook.

For each data mining target, a variety of data mining techniques will be available. For example, data mining of journal entries may include account analysis; temporal scrutiny; exploration of internal controls over journal entries and assessment of text embedded in journal entries (Debreceny and Gray 2010).

We now introduce our scoring system, for rating each of the data mining targets. The first factor we title “signaling.” In the context of the financial statement audit, some requests for additional information will be seen by the client as a relatively normal part of the audit. Other requests would be seen by the client as highly unusual. These requests may trigger a significant level of concern from the client as to the direction that the audit was undertaking. As a result,
these requests for data would only be undertaken when the benefit from mining this data source was expected to be especially rewarding. An example of a request for additional data that might raise only relatively low concern from the client would be data from maintenance systems (relevant to questions of asset valuation) or sales systems (relevant to revenue recognition). These sources of data for mining are closely aligned to established flows of accounting transactions in the General Ledger. An example of requests for additional data that is likely to raise many questions or concerns from the client would be a request for the content or headers (metadata) of senior management emails. At the same time, the likely payoff from mining these emails is likely to be high. A secondary factor is the cost of acquiring and understanding data sources. It has taken the major audit firms some years to set up business processes to acquire and understand client journal entries. Signaling is scored as 1) nil because either the auditor receives the data from the client in the normal course of the audit (e.g. journal entries, financial statements) (scored as three), 2) moderate as a clear connection to standard audit practices can be established (e.g. maintenance data) (scored as two) and 3) high, as (e.g. emails) (scored as one).

The next attribute of these data mining targets is the range of data types available. We classify this into numeric, textual and abstract representations. Different data mining techniques apply to each of these data types. For example, when considering journal entries, the economic values in the journal entries are the most obvious subject of data mining. At the same time, however, the text in comment fields and elsewhere can be mined. Each data type contributes a score of one (maximum of three). The next dimension of data mining attributes is the level of semantic representation of the information source. Higher semantic representation will assist the efficacy of data mining. Again, using journal entries as an example, data mining can leverage the internal logic of the double entry bookkeeping system and the chart of accounts. We score this
as: low-one, medium-two and high-three. The maximum score is, then, nine. In our analysis, we take the highest score for each class of data mining targets. For example, as shown in Figure 5, we identify five finance-related sources of information sources outside the client. The score for the first of these targets, the SEC’s EDGAR XBRL repository, has a score of eight. We apply this score to this group of data mining targets.

**Applying scoring system to fraud and evidence schemes**

We then apply the scoring system described above to each cell in the fraud and evidence scheme matrix. We score the likelihood that a particular class of data mining targets will apply to a particular combination of fraud and evidence scheme. Where we see a high degree of application of the particular category of data mining targets to the given combination of fraud and evidence scheme, the score shown in Column F of Figure 5 is applied in full. If the data mining targets are believed to be only of moderate utility, a weight of 50% is set. Finally, if data mining targets are considered to be only of low utility, the weight is set at zero. The score for all classes of data mining targets are summed for each combination of fraud and evidence scheme. The scores are then trifurcated into low, medium and high levels of application of data mining to each combination of fraud and evidence scheme.

Figure 6 is a “heatmap” of the fraud and evidence schemes and the possible contribution of data mining in identifying that fraud scheme (a specific account scheme and evidence scheme combination). The figure sets out a matrix of fraud schemes (rows) and evidence schemes (columns). The diagonal separates intensity of occurrences (lower-left) and likely benefits to be gained from data mining (upper-right). For example, frauds that involved fictitious revenue that employed fake documents (upper left cell) were one of the more common types of account/evidence combinations. The likely contribution of data mining is moderate in the
identification of this fraud scheme. In the remainder of this sub-section, we analyze the ratings of the application of data mining to the particular fraud schemes.

Insert Figure 6 about here

Consider, for example, “omitted disclosures” as a particular class of fraud, categorized as having low applicability of data mining, for all but one evidence scheme (“Altered Documents”). The rationale for this categorization rests on the nature of the fraud, which involves either failure to disclose material information in the financial statements or misrepresentation of the real state of affairs. The underlying information is available in the accounting information system. The fraud arises from the way the financial statements reports on this information. For this class of fraud, several classes of data mining targets have relatively low utility. Then consider the evidence scheme of “client misrepresentation.” This evidence scheme involves the client’s responses to auditor enquiry being, at worst, a direct falsehood or, at best, dissembling. The There is, then, no direct transactional evidence upon which data mining can operate that we might classify as client misrepresentation. As a result, shows low application of data mining for this evidence scheme for all but two account schemes: “Overvalued Assets/Understated Assets,” and “Fictitious Assets.” For these three fraud schemes, data mining would operate primarily on indirect signals rather than directly on the client misrepresentation. For example, we indicate that frauds involving overvalued assets and understated expenses that employ posting of reversal entries would be tractable with data mining. Not only is most or all of the information needed for data mining within a database of journal entries, but there is also triangulation between two classes of accounts and across time (prior to period-end and the reversal in the next period). Direct mining of the evidence is feasible and cost-effective. The assessment of the potential for data mining in other cells draws on the same two factors.
This assessment of contribution of data mining illustrated in Figure 6 is tentative and is based upon our understanding of the conduct of the audit and the tools and data sets likely to be available to auditors over the next few years. The likely contribution of data mining is not independent of the incidence of frauds. When frauds are very rare, it is more difficult for data mining to be successful in identifying fraud indicators (red flags). Further, merely because data mining is likely to be productive does not necessarily mean that there will be a payoff in areas of high productivity. The cost-benefit relationship of data mining techniques to evidence schemes falls along a continuum with implications for choice of mining targets.

Where frauds involve data triangulation with data sources that are more readily available to the auditor, data mining is of increased interest. The SEC database has a high proportion of frauds associated with revenue, including premature recognition and fictitious revenue. This class of fraud provides an interesting case study of where either direct or indirect data mining techniques for fraud markers are feasible. Typically, underlying entries for these sales are in supporting sales systems (“AIS connected”) and in some cases evidence appearing in subsequent journal entries in the general ledger (“AIS core”). Direct data mining of sales systems seems to be an area of particular interest. The objective would be to identify unusual transactions for subsequent audit enquiries. Automated analysis of patterns of sales and receivables databases should discover outliers (rapid increase in sales volume, high margins, high levels of outstanding receivables etc.). Final determination of the evidence of fraud will require understanding of the value-adding processes of the client and making enquiries of the client and inspection of underlying documentation.

Evidence schemes that involve fake, altered, or hidden documents form a key part of these revenue frauds as well as frauds in areas such as fictitious assets. Examples of documents
are hidden side agreements with clients or altered or fraudulent sales contracts. It is much less likely that direct data mining over sales systems will lead to high quality evidence of fraud markers. Indirect approaches may be more cost-effective and there are two areas of text data mining that are of particular interest that are relevant to frauds that use this class of evidence scheme. First, mining of comment fields and other textual input that support relevant sales and other journal entries for unusual patterns would identify potentially fraudulent transactions. The power of such analyses increase if triangulated with the scrutiny described in the previous bullet points. Textual analysis may be one way to reduce the number of transactions identified for subsequent examination. Textual analysis can use both known stop words as well as changes in text patterns. For example, in frauds involving revenue, when clients make sales regularly to an established client, the descriptive text is unlikely to change dramatically. Conversely, new and unusual sales may give rise to text that does not fit the pattern of sales to this customer. Second, as discussed above and in the following section, text mining of emails for key executives involved in questionable sales transactions would seem to be an area of vital importance for many cells in the heatmap in Figure 6.

Data Mining Issues

In the discussion up to this point, we identified fraud schemes (account scheme/evidence scheme combinations) where data mining has the potential to be the most effective. However, in these closing comments to this section, we recognize that, as with any technology, there are a variety of issues associated with accounting firms expanding their data mining activities that will have to be addressed before data mining can achieve its potential effectiveness, including:
• The availability (or lack thereof) and the quality of client database documentation that the auditors will need to study to determine what data to use in their data mining activities.

• The client’s unwillingness to give the auditors full access to applications and databases so that the auditors can easily expand the data that they data mine during the audit. Currently, clients do not give auditors access to their databases. Instead, the clients prepare copies of the data for the auditors that include only the records and fields that the auditors request. Subsequently requesting additional records and/or fields would cause major delays in the audits. There are a number of solutions to this issue, including creation of an electronic “sandbox” in which the auditor can conduct data mining activities separate from the client’s live data.

• Expanding the data mining to data outside what the auditor requested in past (e.g., non-financial data, email archives, etc.) will signal to the client personnel that “something” is going on. For example, why are the auditors suddenly looking at emails and whose emails are they looking at? As these unusual requests come into the IT departments, rumors will probably spread quickly throughout the client organization.

• Client concerns about proprietary information, which put limits on non-accounting data that the client will allow the auditors to use in their analysis.

• Determining the quality and consistency of the non-accounting (non-GAAP) data that the auditor may want to use. We are suggesting that auditors include non-accounting (non-financial) data in their data mining domain; however, there are no external
standards that apply to non-financial data. As such, the auditors will need their own set of processes to determine the quality of non-financial data and the consistency of that data with the financial statements before that the audit team uses that data.

- Potentially high learning curve costs with respect to understanding the client’s databases and using new data mining software on the audit.

- The incremental computer costs for analyzing 100% of the client’s records and 100% of the fields would be almost zero. However, the labor cost could be very high trying to investigate all of the generated false positives. As such, selecting the appropriate data elements from the client’s databases and interpreting the data mining results are critical skills that need to be developed.

- An issue internal to accounting firms is: How does this data mining impact the risk profile of the accounting firm? When taking traditional small samples, if the “smoking gun” related to a fraud is not in their sample; the firm’s defense is that their audit sampling followed industry practices. However, data mining can be considered the equivalent to taking a 100% sample. If the smoking gun is in that sample, but the auditors missed it, then the auditors no longer not have their traditional industry-practice defense.

V. CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The increasing value of data mining as a financial statement auditing tool is due to the convergence of several factors: (1) increasing emphasis on fraud detection in audits by regulators and standard setters, which provides motivation to use tools to increase auditor productivity; (2) increasing use of data mining tools as a forensic audit tool within accounting firms, which means
there is a growing population of people within accounting firms with data mining experience; and (3) the evolution of more robust and easier to use data mining software tools. In addition, the increasing use of data mining as a *de rigueur* part of e-discovery in law suits have provided many examples of how data mining can be used for forensic investigations and, because of the competition in the e-discovery marketplace, e-discovery has accelerated the development of improved data mining tools.

Whether the auditor employs data extraction, data analysis, or data mining techniques or some mixture of those techniques, these techniques can potentially be applied to each phase of the audit. Depending on the audit phase, the mined data could be a mixture of external and internal data and a mixture of financial and non-financial data as well as textual data (e.g., email data). In fact, problems that are going to be encountered by auditors who attempt to incorporate data mining techniques throughout the audit life cycle include: too many techniques to select from (How to select the best?); too many data sources (How to select the best data sources? How to test quality of non-financial data?); and how prevent overwhelming false-positives and spurious patterns. Regarding the last point, it would be easy for auditors to decide to try many data mining techniques and many data sources because the incremental cost in terms of computer resources would be minimal—but, the labor to investigate the red flags generated by this ad hoc approach could be tremendous.

The challenge facing the auditors is selecting the optimum technique(s) and data source(s) to ameliorate the above problems. Our taxonomy should be an important contributor to that optimization. The heatmap we present in Figure 6 helps identify where data mining could be the most effective (as well as least effective). The taxonomy builds on the real world in that it reflects the frequency of different types of fraud schemes (combinations of account schemes and
evidence schemes). The heatmap also indicates that data mining will not discover every type of fraud scheme and therefore data mining should be judiciously.

**Potential Research Questions**

Because the current use of data mining is relatively ad hoc and new in the financial statement audit domain, there are many potential research questions that could be addressed in future research. Some of those questions include:

- What benefits for audit planning can arise from employing advanced data mining tools on publicly available (external) information? A range of data mining tools could be investigated on a variety of data sources to assess broad and targeted (e.g. revenue) fraud indicators. In Section II, we said that data mining tools fall into two broad categories: direct and indirect. Within each of these categories there are several specific data mining approaches. Determining which specific approach is most effective with specific evidence schemes would be very valuable information for the audit community.

- What will be the impact on audit planning and analytical procedures from the detailed disclosures made to EDGAR in XBRL? From this research, we could more fully understand the likely impact of widespread adoption of XBRL on audit planning and analytical procedures.

- What is the value of textual data as a fraud-detection audit tool? What textual data mining tools seem to work best in the audit environment? Emails and other textual materials have turned out to be incriminating evidence in several highly-visible court cases. Textual data mining software have become more sophisticated. Further, there are products specifically designed for email data mining. Two or three text data mining tools could apply to a two
or three bodies of textual materials (e.g., email). Comparison of these tools can be on various dimensions such as content and readability of output reports and ease of use. Of particular importance are granularity and the number of false positives. As said before, email data mining is *de rigueur* in law suits. Research of legal e-discovery literature might provide valuable insights to the financial audit environment.

- How can data mining techniques combine with a continuous monitoring objective? In general, data mining is batch processing of very large databases. As the name implies, continuous monitoring involves the real time (or at least, very short-term), monitoring of single transactions to spot an anomaly when it happens (or soon after it happens). So, on the surface, it may seem that the two concepts are mismatched. On the other hand, one of the key aspects of data mining is model building. After data mining software has identified a model, that model could then be the basis to monitor subsequent transactions against the model to identify anomalies. Besides testing each transaction against the model, each new transaction would incorporate into the model so that the model becomes a dynamic rolling model to reflect the changes in the business’ environment.

VI. REFERENCES


Figure 1: Relationship of Various Data Examination Tools

Figure 2: Phases of the Audit (adapted from Debreceny and Gray (2011))
Figure 3: Data Examination Processes in the Audit
Figure 4: Occurrence of Fraud Schemes: Account Scheme and Evidence Scheme Combinations.

<table>
<thead>
<tr>
<th>Fraud Scheme</th>
<th>Collusion with third Parties</th>
<th>Altered Documents</th>
<th>Hidden Documents</th>
<th>Client Misrepresentations</th>
<th>Fake Products</th>
<th>Related Parties</th>
<th>Spreading of Fraudulent Items among Accounts</th>
<th>Reversal Accounting Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fictitious Revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premature Revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overstated Assets/Understated Liabilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fictitious Assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Measures to Overstate Revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overvalued Assets/Equity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omitted Disclosure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frequency
- Nil
- Low
- Medium
- High
**Figure 5: Data Mining Targets in the Audit Environment**

<table>
<thead>
<tr>
<th>A Source</th>
<th>B Target</th>
<th>Examples</th>
<th>C Signaling</th>
<th>D Data Types</th>
<th>E Semantic Rep</th>
<th>F Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Numeric</td>
<td>Text</td>
<td>Abstract</td>
<td></td>
</tr>
<tr>
<td>Inside Client Information System</td>
<td>AIS Core</td>
<td>Direct - Journal Entries</td>
<td>Nil</td>
<td>X</td>
<td>X</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Direct - Account Movements</td>
<td>Nil</td>
<td>X</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>AIS Connected</td>
<td>Life cycle - pre-accounting transactions</td>
<td>Early stages of A/R cycle</td>
<td>Moderate</td>
<td>X</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Non-financial and financial</td>
<td>MRP II, Maintenance</td>
<td>Moderate</td>
<td>X</td>
<td>X</td>
<td>L</td>
</tr>
<tr>
<td>AIS Unconnected</td>
<td>Email</td>
<td>High</td>
<td>Moderate</td>
<td>X</td>
<td>X</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>Document archives</td>
<td>Meeting minutes</td>
<td>Moderate</td>
<td>X</td>
<td>X</td>
<td>L</td>
</tr>
<tr>
<td>Formal Disclosures</td>
<td>Financial Statements - Financials</td>
<td>Nil</td>
<td>X</td>
<td></td>
<td>H</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Financial Statements - Notes</td>
<td>Nil</td>
<td>X</td>
<td>X</td>
<td>L</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>MD&amp;A</td>
<td>Nil</td>
<td>X</td>
<td>L</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Press releases, 8K</td>
<td>Nil</td>
<td>X</td>
<td>L</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Outside Client Information System</td>
<td>Finance Related</td>
<td>EDGAR - XBRL</td>
<td>Peer analysis</td>
<td>Nil</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>EDGAR - Other</td>
<td>Nil</td>
<td>X</td>
<td>X</td>
<td>M</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Financial Databases</td>
<td>Nil</td>
<td>X</td>
<td></td>
<td>?</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Stock Prices/Volume</td>
<td>Nil</td>
<td>X</td>
<td></td>
<td>H</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>LEXIS/NEXIS--Financial related &amp; legal, court records, etc.</td>
<td>Nil</td>
<td>X</td>
<td>L</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Finance Unrelated</td>
<td>Social Media</td>
<td>Nil</td>
<td>X</td>
<td>M</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6: Fraud and Evidence Schemes and Application of Data Mining

<table>
<thead>
<tr>
<th></th>
<th>Fake Documents</th>
<th>Collusion with third Parties</th>
<th>Altered Documents</th>
<th>Hidden Documents</th>
<th>Client Misrepresentations</th>
<th>Fake Products</th>
<th>Related Parties</th>
<th>Spreading of Fraudulent Items among Accounts</th>
<th>Reversal Accounting Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fictitious Revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premature Revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overstated Assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underrated Liabilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fictitious Assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Measures to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overstate Revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overvalued Assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omitted Disclosure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Frequency**
- Nil
- Low
- Medium
- High

**Data Mining Applicability**
- Low
- Medium
- High
### Table 1: Occurrence of Account Schemes (Fraud Types) (n=148)
**Source:** Gao and Srivastava (2011)

<table>
<thead>
<tr>
<th>Account Schemes</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premature Revenue Recognition</td>
<td>34%</td>
</tr>
<tr>
<td>Fictitious Revenues</td>
<td>25%</td>
</tr>
<tr>
<td>Overvalued Assets and Understated Expenses</td>
<td>11%</td>
</tr>
<tr>
<td>Omitted or Understated Expenses/Liabilities</td>
<td>8%</td>
</tr>
<tr>
<td>Other Methods to Overstate Revenues</td>
<td>5%</td>
</tr>
<tr>
<td>Omitted or Improper Disclosure</td>
<td>5%</td>
</tr>
<tr>
<td>Fictitious Assets</td>
<td>5%</td>
</tr>
<tr>
<td>Overvalued Assets/Equity</td>
<td>3%</td>
</tr>
<tr>
<td>“Wrong Way” Frauds</td>
<td>1%</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>1%</td>
</tr>
</tbody>
</table>

There are multiple evidence schemes employed in some frauds.

### Table 2: Evidence Schemes Employed (n=185)
**Source:** Gao and Srivastava (2011)

<table>
<thead>
<tr>
<th>Evidence Schemes</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden Documents/Information</td>
<td>25%</td>
</tr>
<tr>
<td>Altered Documents</td>
<td>22%</td>
</tr>
<tr>
<td>Fake Documents</td>
<td>19%</td>
</tr>
<tr>
<td>Collusion with Third Parties</td>
<td>15%</td>
</tr>
<tr>
<td>Client Misrepresentations</td>
<td>8%</td>
</tr>
<tr>
<td>Improper Related Party Transactions</td>
<td>3%</td>
</tr>
<tr>
<td>Shifts and/or Spreading of Fraudulent Items among Accounts</td>
<td>3%</td>
</tr>
<tr>
<td>Fake Products/Equipment</td>
<td>2%</td>
</tr>
<tr>
<td>Reversal Accounting Entries</td>
<td>2%</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>1%</td>
</tr>
</tbody>
</table>