J. Neville, Ö, Şimşek, D. Jensen, J. Komoroske, K. Palmer and H. Goldberg (2005). Using Relational Knowledge Discovery to Prevent Securities Fraud. University of Massachusetts, Technical Report 05-23.

Using Relational Knowledge Discovery to Prevent Securities Fraud

Jennifer Neville, Özgür Şimşek, David Jensen
Department of Computer Science
University of Massachusetts Amherst
Amherst MA 01003-9264 USA
{jneville, ozgur, jensen}@cs.umass.edu

John Komoroske, Kelly Palmer, Henry Goldberg
National Association of Securities Dealers
1735 K Street, NW
Washington, DC 20006-1516 USA
{komorosj, palmerk, goldberh}@nasd.com

ABSTRACT

We describe an application of relational knowledge discovery to a key regulatory mission of the National Association of Securities Dealers (NASD). NASD is the world's largest private-sector securities regulator, with responsibility for preventing and discovering misconduct among securities brokers. Our goal was to help focus NASD's limited regulatory resources on the brokers who are most likely to engage in securities violations. Using statistical relational learning algorithms, we developed models that rank brokers with respect to the probability that they would commit a serious violation of securities regulations in the near future. Our models incorporate organizational relationships among brokers (e.g., past coworker), which domain experts consider important but cannot easily use otherwise. The learned models were subjected to an extensive evaluation using more than 18 months of data unseen by the model developers and comprising over two person weeks of effort by NASD staff. Model predictions were found to correlate highly with the subjective evaluations of experienced NASD examiners. Furthermore, in all performance measures, our models performed as well as or better than the handcrafted rules that are currently in use at NASD.

1. INTRODUCTION

National Association of Securities Dealers (NASD) is the world's largest private-sector securities regulator, with responsibility for preventing and discovering misconduct among securities brokers, such as fraud and other violations of securities regulations. In accomplishing this regulatory mission, it is critical for NASD to target its limited resources on those brokers who are most likely to be engaged in fraudulent behavior. This paper describes an application of relational knowledge discovery methods to identify such brokers, which was a joint effort between NASD and researchers at the University of Massachusetts (UMass) Amherst.

Using publicly available data, we learned statistical relational models of broker behavior that provide a ranking of active brokers with respect to their probability of committing a serious securities violation in the near future. The intention is to use this ranking to improve NASD's assignment of field examinations—brokers who are ranked higher would be more likely to receive additional examinations by NASD staff. This approach limits the effect of false positives as human analysts will further evaluate the brokers identified by the model.

NASD currently identifies high-risk brokers using a set of handcrafted rules. These rules are based on information intrinsic to the brokers such as the number and type of past violations. They do not exploit social, professional, and organizational relationships among brokers even though NASD experts believe this information is central to the task. Indeed, fraud and malfeasance are usually social phenomena, communicated and encouraged by the presence of other individuals who also wish to commit fraud [3]. It is, however, difficult to accurately specify these patterns manually. As such, relational learning methods have the potential to improve current techniques.

Our approach to modeling in this domain exploits recent work on learning accurate, interpretable models of relational data [9][11]. We learned relational probability tree (RPT) models, an extension of probability estimation trees for relational domains [12]. These models have three attractive characteristics. First, they provide a ranking of brokers (with respect to estimated probability of misconduct), rather than the binary classification provided by the handcrafted rules. Second, they are able to represent and reason with the relational context information analysts believe to be important. And third, due to their selectivity and intuitive representation, tree models are usually easily interpretable—a quality that is often important for domain experts to trust, and make regular use of, the rules.

The learned models were subjected to an extensive evaluation by NASD staff that took over two person weeks of effort. This evaluation showed that the models ranked brokers in a manner consistent with the subjective ratings of experienced examiners. Furthermore, in all performance measures, our models performed as well as or better than the handcrafted rules that are currently in use at NASD. Most notably, our models identified high-risk brokers not previously detected with the handcrafted rules and combined with the current NASD process to significantly increase the accuracy of predicting high-risk brokers.

In the remainder of this paper, we relate our experience developing statistical relational models for this task. We start with a description of the regulatory mission of NASD and the data used to train the models. We then outline the prediction task and our modeling approach. We continue with an empirical evaluation of the models and conclude with implications and future research directions.

2. BACKGROUND

2.1 NASD's Regulatory Mission

NASD is the world's largest private-sector securities regulator. It regulates every firm in the United States that conducts securities business with the public (called *broker-dealers*), and it is subject to oversight by the U.S. Securities and

Exchange Commission (SEC). Established in 1939, NASD has a nationwide staff of more than 2,000, and its regulatory responsibility now includes 5,200 securities firms that operate more that 99,000 branch offices and employ 660,000 individual securities brokers.

NASD rules regulate every aspect of the brokerage business for NASD members. NASD responsibilities include examination, licensing, testing and registration; enforcement; market surveillance; rule writing; professional training; dispute resolution; and investor education. NASD examines broker-dealer firms for compliance with NASD rules, Municipal Securities Rulemaking Board rules, and the federal securities laws. NASD also disciplines those who fail to comply, and in 2004 filed about 1,400 enforcement actions, barred or suspended 830 brokers from the securities industry, and collected $104 million in fines. In addition, NASD monitors all trading on the NASDAQ Stock Market, which covers more than 70 million orders, quotes, and trades per day.

NASD examines firms both on a periodic basis (called cycle examinations) and also in response to complaints or other reasons (called cause examinations). In 2004, NASD's Member Regulation Department conducted 2,275 cycle examinations and 5,967 cause examinations, which required more than 500 field examiners, as well as headquarters staff. Properly targeted examinations are critical to protecting investors and the integrity of securities markets. Early discovery of securities violations can prevent serious harm, recover fraudulently obtained funds, and lead to swift punishment of perpetrators. They can also prevent future violations through increased regulatory scrutiny.

It is critically important for NASD to identify firms and brokers who have a higher probability of committing serious violations in the future because this allows efficient allocation of the limited resources of examiners and other NASD staff. Currently, NASD uses a variety of methods to identify high-risk brokers and firms, particularly highlighting broker-dealer firms and individual brokers who have had regulatory or financial problems in the past. Because of the difficulty of this task, NASD continually seeks methods to both predict violations and assign its examinations more precisely.

2.2 The Central Registration Depository

One key tool for accomplishing NASD's regulatory mission is its Central Registration Depository (CRD®) system. CRD was established to aid in the licensing and registration of its broker-dealers and the brokers who work for them. CRD maintains information on all federally registered broker-dealers and brokers for the SEC, NASD, the states, and other federally authorized private sector regulators, such as the New York Stock Exchange.

Originally implemented in June 1981, CRD has grown to include data on approximately 3.4 million brokers, 360,000 branches, and 25,000 firms. For firms, CRD information includes data such as ownership and business locations. For individual brokers, CRD includes qualification and employment information. Information in CRD is self-reported by the registered firms and brokers, although incorrect or missing reports can trigger regulatory action by NASD.

Figure 1 shows a relational schema for the NASD data, indicating entities and relationships that were used in our analysis. Although the CRD database employs a much more complex schema, Figure 1 provides a guide to the major types of objects and links provided to the relational learning algorithms. The frequency counts in Figure 1 refer to a subset of the CRD used for this analysis, which was restricted to firms and brokers who have had an approved NASD registration.

One of the most important categories of data in CRD captures disciplinary information from a number of sources, including state regulators, SEC, NASD, New York Stock Exchange, American Stock Exchange, and FBI, as well as from the registered brokers and the brokerage firms themselves. This disciplinary information, generally referred to as *disclosures*, includes information on criminal, regulatory, and civil judicial actions, customer complaints, and termination actions. Other disclosure types report financial problems such as bankruptcies, bond denials, and liens. Disclosure information on individual brokers is provided free of charge to the public through NASD's BrokerCheck system (www.nasdbrokercheck.com).

Because one indicator of future problematic behavior is past behavior, NASD uses disclosure counts of individual brokers from the CRD to assist it in targeting its examinations toward those who are at higher risk to commit future violations.

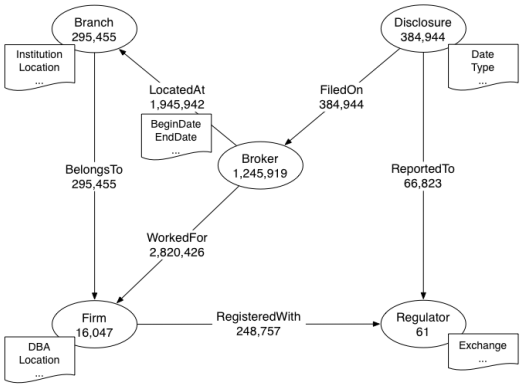

Figure 1: CRD data schema.

3. TASK DESCRIPTION

Our goal was to develop a statistical model to identify which brokers warrant additional attention from NASD examiners. There are two reasons to instigate reviews: (1) to uncover broker violations, and (2) to prevent future violations by increasing supervision on those brokers who are believed to be most likely to commit them. Unfortunately, there is no attribute in the data that records, in retrospect, whether examiners should have reviewed particular brokers. Instead, we use the existence of serious violations as a surrogate measure.

To quantify broker misconduct, we use a ranking of disclosure severity provided by NASD experts. We regard disclosures of type *investigation* or *regulatory-action* as "serious violations" and label the brokers who have had a serious violation in a given time period as positive examples. In other words, the surrogate measure we use is whether a broker will have an investigation or regulatory-action disclosure in the near future, under the assumption that examiners would have wanted to review these brokers before they committed these actions.

We restricted our analysis to small and moderate sized firms with fewer than 15 brokers. These firms account for 10-20% of the brokers under NASD jurisdiction. There were two reasons for this restriction. First, the patterns of behavior differ between small and large firms. Second, large firms typically have more extensive compliance mechanisms in place.

Currently, NASD generates a list of *higher-risk brokers* (HRB) using a set of handcrafted rules they have formed using their domain knowledge and experience. This approach has two weaknesses we aim to address.

First, the handcrafted rules simply categorize the brokers as "higher-risk" and "lower-risk," rather than providing a risk-ordered ranking of brokers. A ranking would be more useful to examiners, as it would allow them to focus their attention on brokers considered to have the highest risk.

Second, NASD's handcrafted rules use only information intrinsic to the brokers. In other words, they do not utilize relational context information such as the conduct of past and current coworkers. NASD experts believe that organizational relationships can play an important role in predicting serious violations. For example, brokers that have had serious violations in the past may influence their coworkers to participate in future schemes. Furthermore, some firms tend to be associated with continuous misconduct (i.e., they do not regulate their own employees and may even encourage violations). Lastly, high-risk brokers can move from one firm to another collectively, operating in clusters, which heightens the chance of regulatory problems. A model that is able to use relational context information has the potential to capture these types of behavior and provide more accurate predictions.

4. MODELING APPROACH

NASD's task of ranking brokers for examination has three characteristics that are common to many knowledge discovery tasks, but that are rarely addressed in combination. Accurate ranking of brokers is inherently probabilistic, relational, and temporal.

- *Probabilistic* — Any attempt to predict the future behavior of brokers is inherently probabilistic. There can be many underlying reasons for a particular pattern of behavior, and CRD data can never fully capture the complex motivations of, and influences on, a particular broker. Instead, the goal of the statistical model is to focus the attention of NASD examiners on brokers whose past behavior indicates that they are at greater risk for particular future behaviors. Probabilistic predictions particularly aid this goal because they facilitate the assessment of both absolute and relative risk.
- *Relational* — The majority of the patterns discussed by expert NASD examiners reflect aspects of the social, professional, and institutional networks within which brokers operate. Fraud and malfeasance are usually social phenomena, communicated and encouraged by the presence of other individuals who also wish to commit fraud. Yet the existing methods used by NASD to automatically filter brokers for analysts do little to reflect these networks. Fortunately, recent developments in relational knowledge discovery (e.g., [7][16]) offer the potential to develop statistical models that incorporate aspects of these networks into predictive models.
- *Temporal* — NASD wishes to predict behavior in the relatively near future, so our analysis focused on predicting the probability of at least one serious disclosure in the next calendar year. Ideally, a model might predict a probability distribution of serious disclosures across *all* future years, allowing for more informative reasoning of the type outlined by Provost & Domingos [13] in their discussion of "activity monitoring." However, we focused on predicting disclosures in the next year as a reasonable approximation to this task that provided the most immediate value to NASD.

All three of these problem characteristics indicate the potential for a statistical relational model to provide better indicators for examiners than a broker's actual disclosures. Specifically, a relational model can capture dependencies among broker characteristics, past behavior, and future behavior that go beyond what can be captured in simple filtering rules. In addition, it can capture dependencies that go beyond an individual broker to consider the behavior of the broker's past and present coworkers, branches, and firms. Finally, a statistical relational model might be able to identify and represent complex temporal trends of behavior that suggest particularly high risk for serious disclosures in the next year, even though past behavior has been relatively benign.

4.1 Relational Probability Trees

We use relational probability trees (RPTs) [12] for this task. RPTs extend probability estimation trees [13] to a relational setting. Due to their selectivity and intuitive representation of knowledge, tree models are often easily interpretable. This makes RPTs an attractive modeling approach for NASD examiners. The RPT learning algorithm adjusts for biases towards particular features due to the unique characteristics of relational data. Specifically, three characteristics—concentrated linkage, degree disparity, and relational autocorrelation—can complicate efforts to construct good statistical models, leading to feature selection bias and discovery of spurious correlations [9][11]. By adjusting for these biases, the RPT algorithm is able to learn relatively compact and parsimonious tree models.

RPT models estimate probability distributions over class labels in a manner similar to conventional tree models. However, the learning algorithm looks beyond the attributes of the object for which the class label is defined and considers the effects of attributes in the relational neighborhood of the object being classified. The RPT learning algorithm uses subgraphs as training examples. Each subgraph includes different types of objects (e.g., firms, disclosures), links that represent relationships between these objects (e.g., employment links between a broker and a branch), and attributes on these objects and links. In each subgraph, there is a single target object to be classified; the other objects and links in the subgraph form the target's relational neighborhood. To classify brokers, we constructed subgraphs around brokers, including information about their current and past employment, and their disclosures. A hypothetical subgraph for this task is shown in Figure 2.

The RPT algorithm automatically constructs and searches over aggregated relational features to model the distribution of the class label. For example, to predict the value of an attribute (e.g., *broker-will-have-serious-violation*) based on the attributes of related objects (e.g., characteristics of the broker's coworkers), a relational feature may ask whether the average employment length of the coworkers is less than 12 months. The algorithm constructs features from the attributes of

different object/link types in the subgraphs and multiple methods of aggregating the values of those attributes. The algorithm considers aggregation functions of mode, average, count, proportion, and degree. Count, proportion and degree features consider a number of different thresholds (e.g., *proportion>10%*). The algorithm searches for the best binary discretization of continuous attributes for features (e.g., *count(disclosure.year>2004)*). For the experiments reported in this paper, we considered 10 thresholds and 10 discretizations per feature. The algorithm uses pre-pruning in the form of a p-value cutoff and a depth cutoff to limit tree size. All experiments reported in this paper used α=0.2/|attributes| and a depth cutoff of seven.

Given an RPT model learned from a set of training examples, the model can be applied to unseen subgraphs for prediction. The chosen feature tests are applied to each subgraph and the example travels down the tree to a leaf node. The model then uses the probability distribution estimated for that leaf node to make a prediction about the class label of the example.

Alternatively, an ensemble of RPT models can be used to improve the probability estimates for each instance. *Bagging* is an ensemble method that reduces variance without increasing bias [7]. The bagging procedure involves learning multiple trees, each from a different bootstrapped *pseudosample* (i.e., sample *N* instances with replacement from the original sample), and then computing probability estimates by averaging the predictions of the trees on the test set.

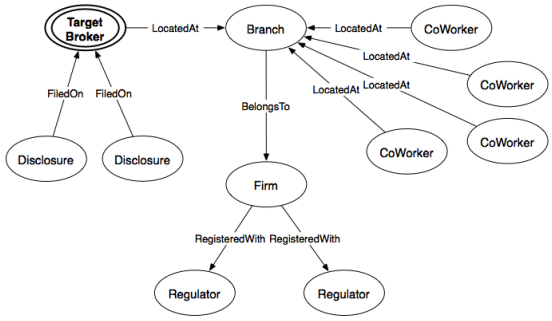

Figure 2: Sample subgraph.

5. KNOWLEDGE DISCOVERY PROCESS

This work was conducted as a joint project between NASD and the Knowledge Discovery Laboratory at UMass Amherst Department of Computer Science. The project proceeded in two iterations of a four-stage process of task specification, data p and evaluation rough time estimates in parentheses):

5.1 First Iteration

- *Scoping and task selection* (one month) — We discussed the basic needs of NASD and analysis capabilities of statistical relational data mining tools in a series of conference calls, email communication, and a visit by NASD staff to UMass. We decided to focus on predicting future disclosures of brokers in small firms and jointly developed a dataset specification that identified entities (e.g., brokers, firms, branches, and disclosures), relations (e.g., worked-for), and attributes (e.g., disclosure type, broker qualification).
- *Data preparation* (three months) — NASD staff prepared an initial data set in a UMass-supplied format. UMass researchers then imported the data and constructed more than 20 variables from the supplied data, and produced seven subsets corresponding to individual years. Training sets were then constructed that corresponded to contiguous periods of years (e.g., 1997-1999). In addition, a class label was constructed that indicated whether a broker had received a disclosure of type regulatory-action, civil/judicial-action, investigation, criminal, termination-for-cause, or arbitration-award.
- *Data mining* (one month) — UMass researchers constructed relational probability trees (RPTs) for each training set. These RPTs estimated the probability that a broker would have a positive class label in the following year. An additional set of trees was constructed that estimated whether a broker would have such a label for the first time.
- *Evaluation* (one month) — UMass researchers evaluated the constructed models using conventional measures such as accuracy, precision, and recall. The models and the evaluation results were presented to a wide selection of NASD staff. While the models met with general approval, a variety of new issues were raised about the class label and the task.

5.2 Second Iteration

- *Task refinement* (two months) — The task specification was criticized by some NASD staff with particular knowledge about examinations and the CRD data. These staff had not been involved in the initial task selection, and pointed out several misinterpretations of the categories of disclosures. Based on the new interpretation, a revised class label was derived, in which brokers with positive labels had at least one of the two most serious categories of disclosures (regulatory-action or investigation)
- *Data refinement* (two months) — In addition to revising the class label, a day-long meeting of UMass researchers and NASD examiners in Boston resulted in suggestions for several new categories of attributes that examiners believed would be predictive of the new class label. These attributes attempted to characterize the movements of groups of brokers from firm to firm, and distinguish "problem" firm environments.
- *Data mining* (one month) — Based on the new class label and some additional attributes, new models were constructed. In contrast to previous results, ensembles of RPTs, learned through bagging, outperformed single trees, and bagged ensembles of RPTs became our default model.
- *Evaluation* (one month) — As before, the RPTs were analyzed using conventional evaluation metrics such as area under the ROC curve (AUC) and accuracy. However, the trees appeared accurate enough to subject them to a far more extensive evaluation in a more realistic setting. An evaluation protocol was developed jointly with NASD examiners, and was then conducted double-blind with four examiners over a one-week period (see Section 6 for details). Also, during this second iteration, it was discovered that an approach (the "Higher-Risk Broker" list) was currently in use at NASD for a very similar task, so the evaluation compared the RPTs to this approach as well.

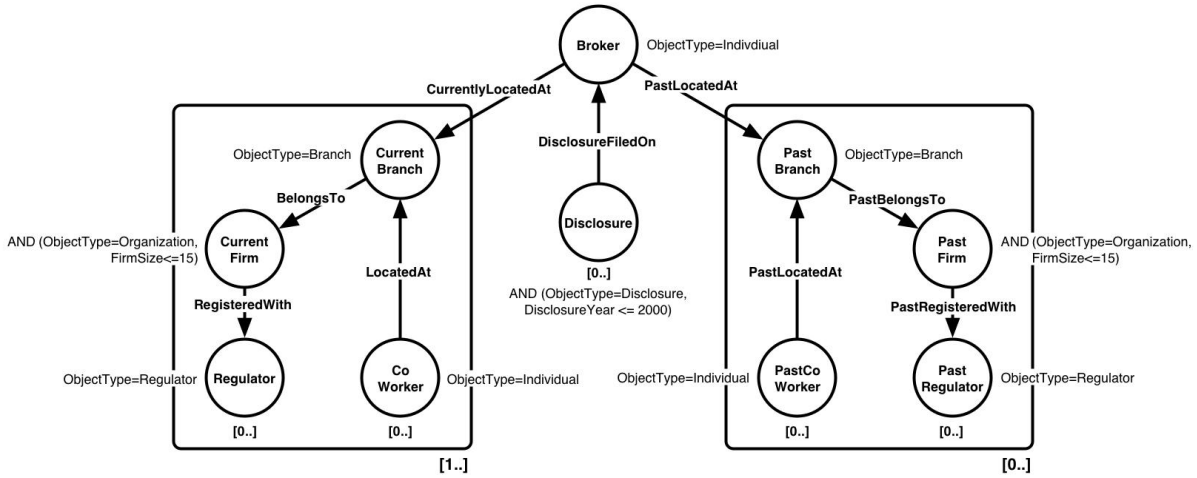

Figure 3: Example query.

In retrospect, several findings of prior work on the knowledge discovery process [5][1] were largely borne out. The analysis process of this project followed a sequence quite similar to the ones described in this prior work. In addition, the vast majority of time was spent on task specification, data preparation, and evaluation, rather than on the data mining step.

6. EMPIRICAL EVALUATION

In this section, we present an empirical evaluation of the claim that RPT models provide a useful ranking of brokers with respect to their likelihood of committing securities fraud in the near future. We examine two surrogate measures of fraudulent behavior. The first is the class label used to train the RPT models, namely whether any serious disclosures were filed on the target broker. The second is a subjective evaluation of brokers by NASD examiners.

Where appropriate, we compare the performance of the RPT models to two baseline models. The first, referred to as *Base*, is an RPT model learned using the same algorithm, but without the attributes in the relational neighborhood surrounding the target broker. *Base* models used only the attributes on the target brokers themselves. The second, referred to as *HRB*, is the binary classification produced by the high-risk broker list.

6.1 Methodology

The training and test instances were subgraphs that centered on a target broker and that included information about the broker's current and past employment. These subgraphs were extracted from the data using the visual query language QGraph [1]. Queries in this language allow for variation in the number and types of objects and links that form the subgraphs and return collections of all matching subgraphs from a database.

Figure 3 shows an example of the type of query used to construct training and test instances[1]. This query is dated December 31, 2000. It returns one subgraph for each broker who, at that date, was working for a firm that employed fewer than 15 brokers. In each subgraph, the relational neighborhood includes the following: 1) any disclosures that have been filed on this broker until the query date, 2) broker's current branch (at query date), all coworkers at this branch, the firm this branch belongs to, and the regulators associated with this firm, 3) broker's past branches, past coworkers at those branches, and firms and regulators associated with these past branches. Figure 2 shows a hypothetical match to this query: a broker who has had two disclosures and who has worked at a single branch.

To address the temporal nature of the prediction task, we created multiple samples, where each sample was a static view of the dataset at a particular point in time. More specifically, the samples we created reflected a static view of the dataset at the end of the calendar years 1996-2001. For example, the 1996 sample was constructed using the data available on December 31, 1996. The samples include a subgraph for each broker active at that date; the relational neighborhood of the target broker reflects what was known about this broker at that date.

The target class label was "has serious disclosure next year", indicating whether at least one disclosure of type *regulatory-action* or *investigation* was filed on the broker within the next calendar year. For example, for the 1996 sample, the target class label was whether there was a serious disclosure filed on the target broker during the calendar year 1997.

One characteristic of the resulting training samples is that there are few positive instances but many negative instances. Table 1 lists the distribution of positive and negative examples in each sample. On average, only 1% of the examples are positive. To increase the absolute number and distribution of positive instances, and to avoid overfitting to the trends of a single year, we constructed training sets by combining samples from three consecutive years. For example, we merged the samples from 1996, 1997, and 1998 into a single training set. Note that if brokers are active during the entire time interval, they will be included as three separate examples, with

[1] For clarity of illustration, we omit temporal constraints from the example query.

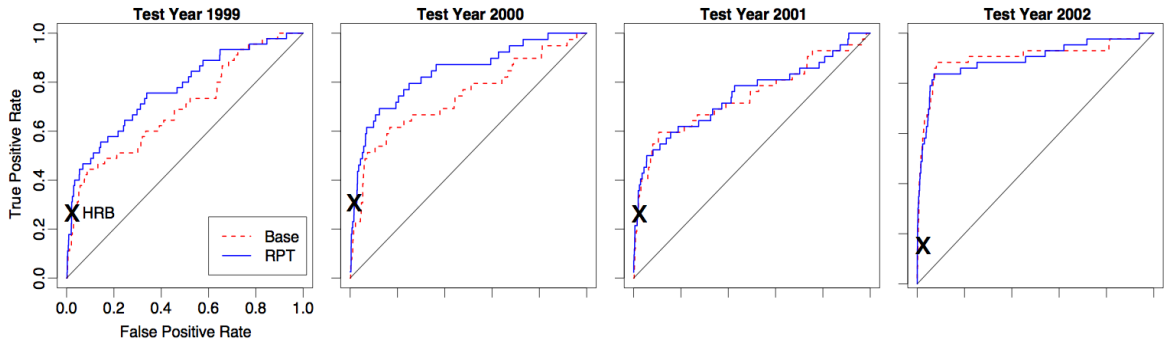

Figure 4: ROC curves for baseline and full RPT models on 1999-2002 test samples.

subgraphs that reflect their relational neighborhood at the end of each year.

From these merged samples, we produced 10 *pseudosamples* for bagging. Each pseudosample was produced using stratified bootstrap resampling (i.e., the positive and the negative examples were sampled separately). For both positive and negative samples, we used sampling with replacement from the original sample. The number of positive examples in the pseudosamples was chosen to be the same as the original sample, but we limited the negative examples to 1500 to increase the overall proportion of positive examples. This also limits the number of times each broker is added to the sample (since brokers are unlikely to have a positive class label for three consecutive years).

Table 1: Temporal sample information.

Year	Positive	Negative	Total	HRB
1996	33	4062	4095	65
1997	63	4110	4173	106
1998	56	4059	4115	109
1999	45	4195	4240	118
2000	39	4257	4296	97
2001	42	4092	4134	119
2002	43	4227	4270	129

The RPT models had 55 attributes available for classification, including information on the broker (e.g., has other business), past disclosures (e.g., event date), current employment (e.g., branch location), past employment (e.g., termination reason), and coworkers (e.g., time in industry).

6.2 Predicting Serious Disclosures

In this section, we evaluate the performance of RPT models in predicting whether a broker will have a serious disclosure in the following calendar year.

We present results on four test samples from 1999-2002[2]. The RPTs were trained on samples that combined the samples for the three previous years, using the procedure outlined above. For example, the test year 1999 means that the model was trained on samples dated 1996, 1997, and 1998, and the test sample was the sample dated 1999. Recall that each training or test sample builds subgraphs using the data available at the end of the sample year, and assigns the class label using the disclosures filed in the following calendar year.

We examine four measures of performance. We use Receiver Operating Characteristic (ROC) curves and area under the ROC curve (AUC) to evaluate the ranking of the different models. In addition, because the HRB does not provide a ranking but only a binary classification of the brokers, we present precision and recall results.

Figure 4 shows the ROC curves on the four test samples. ROC curves show the quality of the ranking provided by the classifier [13]. The curve shows how the false positive rate and the false negative rate vary as the probability threshold between classes is varied between zero and one. If a model dominates the ROC space it can be regarded as the model that provides the best ranking of the brokers. A random ranking is expected to produce a diagonal line with equal true positive and false positive rates.

The figure shows ROC curves for Base and RPT, but only a single point for HRB. This is because both Base and RPT provide a ranking of the brokers, therefore multiple ways of setting the threshold between classes, while HRB provides only a binary class label and therefore a single threshold.

The figure shows that all three models performed better than random. The steep slope of the curves in the true positive range [0.0, 0.4] indicates that the models accurately rank brokers at the top of the list. In addition, the figure reveals that the three models are roughly comparable at the single threshold produced by the HRB list. The relational information produces the largest improvement when ranking the 1999 and 2000 samples. The improvement is most pronounced for the true positive range [0.4,0.9].

On average the RPT model produces an equivalent, or better, ranking when compared to the baseline model. The ROC information is summarized in Figure 5, which plots the AUC for both the baseline model and the RPT. The benefit of the relational information differs significantly between 1999-2000 and 2001-2002. We are still investigating the reason for this change in performance. Our initial hypothesis, first suggested by NASD staff familiar with the disclosures during this period, is that it is due to the bursting of the "tech

[2] We obtained class label information for the 2002 sample after the second iteration of model development and evaluation. We include post-hoc evaluation on this sample to improve understanding of the evaluations reported in section 6.3.

bubble" in mid-2000, which may have changed the nature and pattern of disclosures and caused concept drift.

Figure 6 shows precision and recall results. To obtain these results, the rankings provided by RPT and Base models were used to generate broker lists of the same size as the HRB list. For example, in the 1999 sample, which included 118 brokers on the HRB list, the RPT list included the 118 brokers ranked most highly by the RPT model. Table 1 lists the size of the HRB list for each sample.

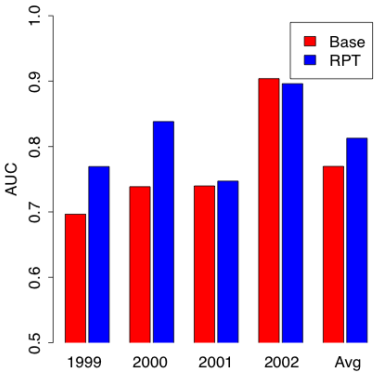

Figure 5: AUC performance comparison of the baseline and full RPT models.

Precision refers to the proportion of brokers on the list who have a positive class label. Recall is the proportion of brokers with a positive class label who appear on the list. Due to the small number of positives and the size of the HRB list, 0.40 is the maximum precision any model can hope to achieve. Also given the low proportion of positives, random performance would result in approximately 0.01 precision and 0.03 recall. Clearly all the models are performing above random. In all test samples, RPT precision and recall performance were equivalent to or higher than HRB. The relative performance of RPT and Base paralleled the ROC-AUC performance reported in the previous section.

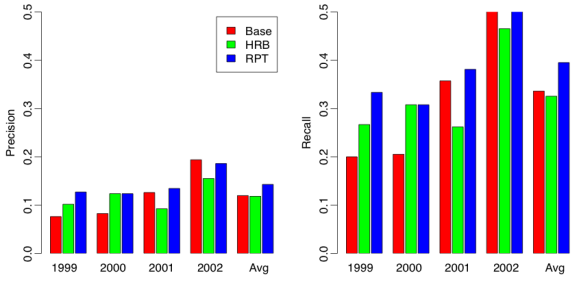

Figure 6: Precision-recall performance comparison of the RPT models and the high-risk broker list.

To quantify the amount of relational information included in the RPT models, we computed the proportion of tree nodes that use relational features weighted by the proportion of training instances that traveled through the node. In each RPT we learned, the weighted proportion of nodes that used relational information was more than 50%. This indicates that the RPTs made substantial use of the relational information. Recall that the baseline models, which ignored the relational information, performed substantially worse in two of the four years evaluated.

6.3 Correlation with Examiner Ratings

We next evaluate the RPT models with subjective ratings of brokers from NASD examiners. These ratings were not part of the CRD data, but were produced on a small set of brokers in February 2005 with the sole purpose of evaluating our models.

The ratings required an examiner to spend a considerable amount of time (approximately 30 minutes per broker). As a consequence, we could obtain ratings for only a small set of brokers. Because of this limitation, we chose to evaluate a single RPT model. We trained this model[3] using the 1999-2001 samples, which constituted the most recent three-year span in the CRD data available to us.

Using this model, we obtained predictions for the 2002 sample. We selected 80 brokers from this sample and asked four NASD examiners to rate these brokers on a five-point scale, indicating the degree to which each broker warranted additional attention from an NASD examiner in 2003. A score of 1 indicated that the broker deserved no additional attention; a score of 5 indicated they deserved the highest attention. We asked examiners to use any information to which they have access, including the data accumulated since 2003 and any useful sources outside of CRD.

We believe that these ratings provide a better measure of the utility of the RPT rankings than the class label we used to train our models (i.e., whether a broker will have a serious disclosure in the following year): They reflect the judgment of experienced examiners in the light of extensive information (not limited to the CRD data) and the hindsight provided by the data accumulated from 2003 until February 2005.

We selected the brokers using the HRB and RPT lists and partitioned the 2002 sample into the following categories:

- *Both*—Brokers who appeared on both HRB and RPT lists.
- *RPT only*—Brokers who appeared on only the RPT list.
- *HRB only*—Brokers who appeared on only the HRB list.
- *Neither*—Brokers who did not appear on either list.

Table 2 contains the number of brokers in each category. We selected 20 brokers from each category as follows. We ranked the brokers in each category with respect to the probability estimates produced by the RPT model. Within each category, we created 20 bins by frequency (i.e., we placed an equal number of brokers in each bin) and selected the broker with the median probability value in each bin as representative of that bin.

Each of the four NASD examiners independently rated the 80 selected brokers. The examiners were not aware of the procedure used to select the set of brokers; and the UMass author communicating the results to examiners was not aware

[3] This RPT model was learned at an earlier stage of our analysis than those reported in Section 6.2, and used a less conservative p-value cutoff (α=0.05) than those reported in previous sections.

of which category each broker belonged to. Furthermore, to avoid systematic biases caused by evaluation order, each examiner received the list of brokers in a different random order.

Table 2: Overlap between HRB and RPT lists.

	On HRB list	Off HRB list
On RPT list	62	67
Off RPT list	67	4074

We present the agreement between the evaluations of the examiners in Table 3. The pairwise correlations among the examiner scores indicate a relatively consistent ranking of brokers, with the exception of examiner 3. This examiner rated 70% of the brokers with a score of 1, while the other examiners rated brokers more uniformly in the range [1, 5].

Table 3: Pairwise correlations of examiner ratings.

	Exam. 1	Exam. 2	Exam. 3	Exam. 4
Exam. 1	1	0.738	0.473	0.675
Exam. 2	0.738	1	0.398	0.609
Exam. 3	0.473	0.398	1	0.456
Exam. 4	0.675	0.609	0.456	1

Figure 7 shows the distribution of mean ratings in each of the four categories. Table 4 shows the mean of the distribution for each category and the results of two-tailed t-tests comparing the *RPT-only* distribution to each of the other three distributions. The distributions for the *RPT-only* and *HRB-only* categories are nearly identical, but significantly higher than the distribution for the *Neither* category and significantly lower than the distribution or the *Both* category ($p < 0.01$).

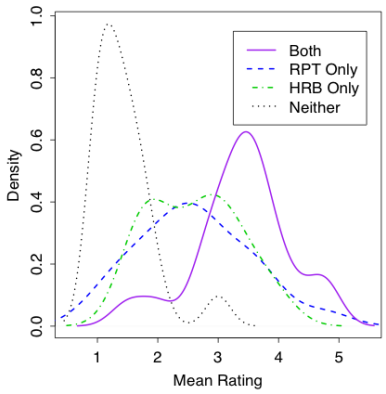

Figure 7: Distribution of mean ratings in each category.

These results indicate that the RPT model is competitive with the HRB list, because it identifies brokers with similar ratings. We note here that examiners' subjective judgments and the HRB criteria are likely to be highly correlated—the HRB list was created to correlate well with examiners' judgments, and their judgments may have been influenced by the existing criteria for the HRB list—so it not surprising that brokers on the HRB were scored highly by examiners. What is more surprising is that brokers identified by the RPT (which was tuned to a surrogate class label) did as well as HRB in terms of examiner ratings.

Table 4: Mean rating in each category.

Category	Mean Rating	t-test (vs RPT)
B	3.387	p=0.007
RPT	2.600	--
HRB	2.600	p=1.000
Neither	1.425	p=3.019e-05

Furthermore, the results indicate that the RPT model identifies novel cases, previously unidentified by the HRB list but with equivalent ratings. This suggests that the RPT model can be used successfully to extend the set of brokers currently assessed by examiners. And finally, the results show that brokers identified by the combination of models (*Both*) have significantly higher scores than those identified by either model in isolation. This indicates that an ensemble of models may be useful to prioritize examiner attention.

These results prompted us to add membership on the HRB list as a feature to the RPT learning algorithm. The evaluation of this modified RPT model revealed that its performance was slightly better than the RPT model presented here, in particular on the 2001 test set. Future work will include additional investigation in this direction.

Figure 8a shows a scatterplot of the mean rating assigned to each broker and the probability of a positive class label assigned to the same broker by the RPT model. Note that we have not sampled uniformly from this space. The bottom left corner of the plot is a very dense region that contains a large number of brokers that are not on the HRB list and who also have a very low probability of positive class label. These brokers are in the *Neither* category in Table 2. We selected only 20 brokers from this category of over 4000 individuals.

Figure 8b shows a variation of the same plot in which brokers are placed into ten bins with respect to their RPT probabilities (bin width = 0.1). The figure shows a scatterplot of the mean rating and the mean RPT probability in each bin. Both figures reveal that RPT probabilities correlate well with the ratings of NASD examiners. This indicates that the RPT model can be used to rank brokers in a manner that would be consistent with examiners ratings if the examiners were to rate each case individually. In other words, the RPT model can be used to prioritize examiners attention on brokers more likely to be involved in misconduct.

Examiners also provided some anecdotal evidence that the model produced ranking that corresponded to their expert judgments. Without prompting, one NASD examiner made the following comment when returning his ratings:

> *One broker I was highly confident in ranking as 5 – because not only did I have the pleasure of meeting him at a shady warehouse location, I also negotiated his bar from the industry. If the model predicted this person, it would be right on target. This person actually used investors' funds to pay for personal expenses including his trip to attend a NASD compliance conference!*

We note that this was the only comment made on an individual broker. Further examination revealed that this broker was identified only by the RPT model and had an average score of 4.75 from the four examiners.

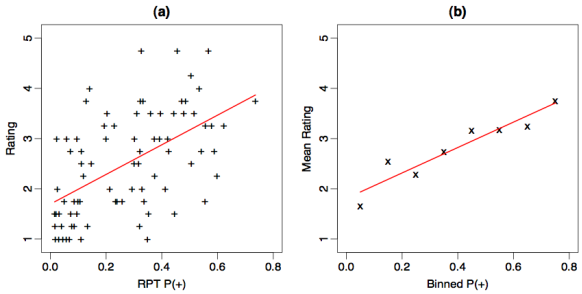

Figure 8: Correlation between RPT predictions and examiner ratings. (a) corr=0.549, R^2=0.293, p-value=1.30e-07, (b) corr=0.945, R^2=0.876, p-value=3.94e-04.

Finally, to evaluate the utility of the surrogate class label we examined the distribution of examiner ratings in light of additional class label information provided for the 2002 sample. Figure 9 shows the distribution of mean scores for the brokers with positive and negative class labels. There are only six brokers in our evaluation set with positive class labels but the scores for these brokers are significantly higher than for the rest of the brokers. However, the average score for negative examples is still much higher than we expected. This indicates that while there is useful information in our surrogate class label, there is a significant amount of untapped information in the negative examples that could be used to improve the models.

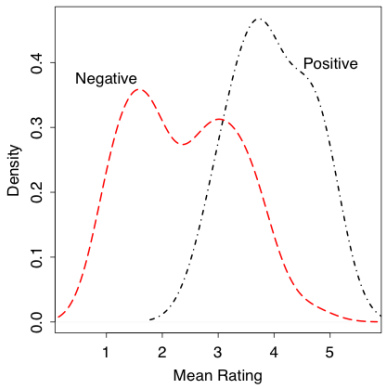

Figure 9: Mean examiner ratings by class label (Positive: mean=3.96, Negative: mean=2.39).

7. DISCUSSION

NASD staff began this project contending that information about the professional and organizational networks that connect brokers would provide useful information for determining their risk for serious violations of securities regulations. The results of this paper have borne out those beliefs.

Our relational models provide predictions that are competitive with, but significantly different from, the predictions provided by NASD's hand-tuned rules, which only examined brokers and their disclosures, ignoring additional relational information such as coworkers at present and past firms. These models show important potential for NASD's screening process. They identified higher-risk brokers not previously identified by the NASD rules, and thus provided additional targets for NASD examinations. Furthermore, being identified as higher-risk by both our models and the HRB model was found to be more predictive of future problems than being identified by either model alone, thus permitting NASD to focus examinations on those most likely to have a disclosure in the near future. And finally, the probability estimates assigned to brokers by our models in general agreed with the subjective ratings of NASD examiners, thus the ranking provided by our models can be used to prioritize examiners' attention.

Our models made substantial use of relational features. In addition, we showed how the statistical models that ignored this relational information performed substantially worse in two of the four years evaluated.

That said, the data provide only relatively weak abilities to exploit the relational component of the data. In CRD, individual brokers are directly related only through firms. Even branch relationships have to be inferred from address information, although this limitation will be obviated beginning this October, when each broker will be systematically linked to a branch. More importantly, we do not know which individual brokers work together directly, nor what other social or organizational relationships they may share. NASD is investigating other technologies to enhance their knowledge of potential links among individuals, most notably the NORA (Non-Obvious Relationship Awareness) system produced by Systems Research and Development (SRD), a Nevada-based company recently acquired by IBM. Such relationships could add substantially to the data analyzed in the work reported here, which could only use branch and firm relations present in CRD.

The work reported here also exemplifies a framework that may be useful to other projects that seek to develop screening tools to aid field examiners working in other domains such as heath care, insurance, banking, and environmental health and safety. In such cases, development of a labeled training set may be impractical in the initial stages of a project. While the most accurate class labels would be the judgments of examiners, examiners' time is typically limited and organizations may be understandably skeptical about devoting large amount of examiners' time to creating labeled data sets.

As we did here, however, initial classifiers can be developed with a surrogate for the ideal class label (here we used the occurrence of serious disclosures as a surrogate for examiners' judgments about the utility of an examination). Evaluations of models constructed with this surrogate class label can determine how well it matches examiners' judgments and can serve to guide and motivate additional work.

8. FUTURE WORK

Our research to date suggests a wide variety of directions for future work. First, the inferences described in this paper did

not exploit a key feature of relational data — the potential of inferences about one object to inform inferences about others. This approach, called *collective inference* [3][15][10] has been shown to improve the accuracy of inferences in relational data. We suspect that this approach could improve accuracy if inferences were made collectively about all brokers and firms.

Second, some of the knowledge conveyed by NASD examiners to the UMass researchers was too complex to be captured by the features currently available to RPTs. In particular, examiners described a set of temporal changes in employment that they believed were strongly associated with higher risk brokers. We suspect that representing and using this temporal information would significantly improve model accuracy. Temporal-relational models are a promising direction for future research that researchers have only recently started to explore [15]. In a similar way, we hope to use additional sorts of connections among brokers to enhance our knowledge of the social and professional networks that affect broker behavior.

Third, the evaluation we conducted with the help of examiners indicates that it would be possible to obtain class labels directly from examiners. This would allow us to abandon our surrogate label (serious disclosures) and attempt to reproduce examiners' screening judgments directly. In the ideal case, we would faithfully reproduce a consensus judgment on the part of examiners, allowing them to focus on in-depth examinations, rather than initial screening of high-risk brokers.

Fourth, we hope to focus on a wider range of firms in future work. Here we examined only the brokers who work at small to medium-sized firms. The promising results we obtained in this task encourages us to continue to develop models for larger firms.

Finally, we hope to account for the apparent concept drift that caused the relational information to show greater improvement in 1999 and 2000. Preliminary investigations show that the 2001-2002 period has a different profile of disclosures, perhaps resulting from the precipitous decline in tech stocks in 2000 and the subsequent rash of complaints from customers in subsequent months and years. NASD staff suggested normalizing disclosure rates based on market performance, and this seems a promising approach.

9. ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of Ted Senator, additional NASD staff (John Cogswell, Steven Cohen, Don Lopezi, Pragnya Mahadwar, Victoria Pawelski, Carl Rubin, David Troutner, and George Walz), and several members of Knowledge Discovery Laboratory at the University of Massachusetts Amherst (Hannah Blau, Matthew Cornell, Amy McGovern, Matthew Rattigan, and Agustin Schapira).

This effort is supported by DARPA and NSF under contract numbers IIS0326249 and HR0011-04-1-0013. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, NSF, or the U.S. Government.

10. REFERENCES

[1] Blau, H., N. Immerman, and D. Jensen. (2001). A Visual Query Language for Relational Knowledge Discovery. University of Massachusetts Amherst Computer Science Technical Report 01-28.

[2] Brodley, C., and P. Smyth (1997). Applying classification algorithms in practice. *Statistics and Computing,*7:45-56.

[3] Chakrabarti, S., B. Dom and P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. *Proceedings of ACM SIGMOD98*, 307-318.

[4] Cortes, C., D. Pregibon, and C. Volinsky (2001). Communities of interest. *Proceedings of the 4th International Symposium of Intelligent Data Analysis*.

[5] Fawcett, T. and F. Provost (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291-316.

[6] Fayyad, U. and G. Piatetsky-Shapiro, and P. Smyth (1996). From data mining to knowledge discovery in databases. *AI Magazine*, Fall:37-54.

[7] Getoor, L., N. Friedman, D. Koller, and A. Pfeffer (2001). Learning probabilistic relational models. In *Relational Data Mining*, Dzeroski and Lavrac, Eds., Springer-Verlag.

[8] Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*, Springer, New York.

[9] Jensen, D. and J. Neville (2002). Linkage and autocorrelation cause bias in relational feature selection. *Proceedings of the 19th International Conference on Machine Learning*, 259-266.

[10] Jensen, D., J. Neville and B. Gallagher (2004). Why Collective Inference Works Well. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[11] Jensen, D., J. Neville and M. Hay (2003). Avoiding bias when aggregating relational data with degree disparity. *Proceedings of the 20th International Conference on Machine Learning*, 274-281.

[12] Neville, J., D. Jensen, L. Friedland and M. Hay (2003). Learning Relational Probability Trees. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 625-630.

[13] Provost, F. and P. Domingos (2003). Tree Induction for Probability-based Rankings. *Machine Learning*, 52:3.

[14] Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 43-48.

[15] Sanghai, S., P. Domingos and D. Weld (2003). Dynamic Probabilistic Relational Models. *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 992-1002.

[16] Taskar, B., P. Abbeel and D. Koller (2002). Discriminative probabilistic models for relational data. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 485-492.