Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors

Longbing Cao
Faculty of Engineering and IT
University of Technology
Sydney
lbcao@it.uts.edu.au

Yuming Ou
Faculty of Engineering and IT
University of Technology
Sydney.
yuming@it.uts.edu.au

Philip S Yu
Department of Computer
Science
University of Illinois at Chicago
psyu@cs.uic.edu

Gang Wei
Department of Surveillance
Shanghai Stock Exchange

ABSTRACT

In capital market surveillance, an emerging trend is that a group of hidden manipulators collaborate with each other to manipulate three trading sequences: buy-orders, sell-orders and trades, through carefully arranging their prices, volumes and time, in order to mislead other investors, affect the instrument movement, and thus maximize personal benefits. If the focus is on only one of the above three sequences in attempting to analyze such hidden group based behavior, or if they are merged into one sequence as per an investor, the coupling relationships among them indicated through trading actions and their prices/volumes/times would be missing, and the resulting findings would have a high probability of mismatching the genuine fact in business. Therefore, typical sequence analysis approaches, which mainly identify patterns on a single sequence, cannot be used here. This paper addresses a novel topic, namely *coupled behavior analysis* in hidden groups. In particular, we propose a coupled Hidden Markov Models (HMM)-based approach to detect abnormal group-based trading behaviors. The resulting models cater for (1) multiple sequences from a group of people, (2) interactions among them, (3) sequence item properties, and (4) significant change among coupled sequences. We demonstrate our approach in detecting abnormal manipulative trading behaviors on orderbook-level stock data. The results are evaluated against alerts generated by the exchange's surveillance system from both technical and computational perspectives. It shows that the proposed coupled and adaptive HMMs outperform a standard HMM only modeling any single sequence, or the HMM combining multiple single sequences, without considering the coupling relationship. Further work on coupled behavior analysis, including coupled sequence/event analysis, hidden group analysis and behavior dynamics are very critical.

KDD'10, July 25–28, 2010, Washington, DC, USA.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database applications—*Data Mining*

General Terms

Algorithms, Economics, Security

Keywords

Coupled behavior analysis, coupled sequence analysis, sequence item property, sequence change, hidden group discovery, coupled hidden Markov model, abnormal behavior detection, market manipulation

1. INTRODUCTION

Abnormal behavior detection plays an important role in capital market surveillance [5] and risk management. The ongoing global financial crisis and recession urge regulation bodies to undertake a deep investigation of trading behaviors in capital markets. An emerging abnormal trading situation is that a group of experienced market manipulators collaborate with each other to manipulate an instrument by fine-tuning its prices/volumes and trading time, in order to misguide other investors. Once the instrument's market price reaches a comfortable level, these manipulators immediately take advantage of the market movement. It is very challenging to detect such hidden group-based manipulative behaviors. In fact, similar coupled behaviors (as well as sequences and events) can be found in many domains, including intrusion detection, crime and national security.

In stock markets, trading transactions consist of multiple streams, in which three typical trading behavioral sequences – buy orders, sell orders and trades from the manipulators - are coupled with each other in terms of timing, prices and volumes etc., according to a market's trading model and investor intention [5]. Often only an individual sequence in such multiple coupled sequences (e.g., trades) is focused on for pattern analysis, while the quotes-related actions and the action price and volume information associated with trades are missing. As a result, we cannot detect those group-based manipulative trading behaviors. Alternatively, if buys and sells are also combined with trades, we may identify more informative patterns disclosing the full trading process and relevant investors' intentions from buys/sells to trades [20]. In addition, the engagement of sequence item properties such as prices/volumes into sequence analysis makes the findings much more meaningful for business analysts.

Typical sequence analysis algorithms such as GSP [18], PrefixSpan [12], Spam [11] and Spade [21] target single sequences. One may argue that the correlated sequences could be merged into one sequence and then analyzed by these methods; it would overlook the relationships that associate relevant sequences and the properties of sequence items. Another issue is that the current sequence-based behavior/event analysis methods ignore the associated behavioral properties. Therefore, the existing techniques cannot be directly used for coupled behavior analysis such as detecting group-based manipulative trading behaviors. It is essential to develop effective techniques, in particular, behavior informatics [6] and activity mining [14, 15], to analyze coupled behaviors considering item properties.

In addition, any significant change taking place in any sequence, the coupling relationships or behavioral properties could seriously affect the model performance. For instance, a sophisticated market manipulator may adjust trading strategies to manipulate an instrument, resulting in significant changes in the coupled sequences from time to time. It is important to understand such change and difference taking place in the time line, to check it against previous patterns, and to automatically tune the model as needed.

While group-based coupled behaviors, associated with correlated relationships and behavioral properties, are commonly and increasingly seen in complex business applications and social networks, it is very challenging to identify such coupled behavior patterns. To the best of our knowledge, we haven't found the related work that directly handles such an issue.

1.1 Contributions

In this paper, we first propose an effective Coupled Hidden Markov Model (CHMM) to identify abnormal patterns from multiple coupled sequences, and then an Adaptive CHMM (ACHMM) that can automatically detect significant behavior changes in the sequences.

First, we build a CHMM to model multiple sequences which are associated with each other in terms of certain relationships and involve item properties in constructing the CHMM. To the best of our knowledge, this is novel in the sequence analysis area.

Second, we further enhance the adaptability of CHMM to sequence dynamics by involving the automatic checking of difference from the benchmarks. Even though a single HMM's adaptability has been studied in areas such as video surveillance, we have not found any work on enhancing the CHMM's adaptability. Our model can adaptively retrain itself to fit in significant changes in coupled sequences.

Third, in evaluating the business impacts of identified trading behavior patterns, we estimate the business performance of trading those identified trading behavior patterns. This simulates traders' preferences and verifies the business impact of resulting patterns.

Finally, from the business perspective, our approach leads to novel contributions to adaptive market surveillance, which is currently not available from the community. The proposed CHMM and ACHMM have the potential for a deep understanding of *pattern-based surveillance* for the comprehensive analysis of multiple trading sequences with interactions, and the adaptation to stock data dynamics.

In fact, the underlying problem – coupled behavior analysis in hidden groups is interesting to be further explored for business applications and social networks associated with coupled sequences, coupled events, or correlated groups.

1.2 Related Work

We haven't found the related work that directly analyzes coupled behaviors as discussed in this paper. Here we discuss the relevant background, namely sequence analysis and hidden Markov Model.

Sequence analysis In sequence analysis, typical algorithms including GSP [18], PrefixSpan [12], Spam [11] and Spade [21] mainly deal with single sequences. Even though additional approaches such as sequence classification and sequence alignment [1] have been investigated for more informative sequence analysis, the underlying interactions and item properties associated with multiple sequences have not been considered. While mining multiple sequences is new [10], it is even difficult to mine coupled multiple sequences embedded with item property information. Also, the existing algorithms are not aimed at adaptively handling significant sequence changes. For this, it is essential to engage the theories of behavior informatics [6] and activity mining [14, 15] in coupled behavior analysis.

Hidden Markov Model The HMM is a statistical model suitable for describing behavior processes. It is widely used in many areas such as speech recognition, motion pattern analysis, and fraud and anomaly detection [2, 7, 13]. The HMM provides internal functions for modeling behavior processes, including clear Bayesian semantics, simple and efficient parameter estimation algorithms, computing the probability of an observation sequence given a model such as by *forward-backward procedure*, and *model training* by an iterative procedure such as the Baum-Welch method. However, the standard HMM only consists of a single variable to represent hidden states, which cannot be used to represent complex processes [9]. HMM cannot describe systems with multiple interacting processes such as the above three coupled trading sequences. CHMM [3, 16] is proposed on top of HMM to model multiple processes with coupling relationships. CHMM consists of more than one chain of HMMs representing different processes, in which the state of any chain of HMM at time t depends on not only the state of its own chain of HMM but also the states of other chains of HMMs at time $t-1$, namely the interaction between two modeled processes. In addition, to suit data and pattern changes, change detection [8, 19] is a recent focus. To the best of our knowledge, there is no existing work on utilizing the HMM for detecting abnormal coupled sequences and automatically handling sequence changes associated with item properties in data mining.

During the development of the Group-based Manipulative Behavior Analysis System (GMBAS) for an exchange, substantial experiments have been conducted on analyzing abnormal trading behavior patterns in one and a half years of orderbook-level trading activities from a stock market. The experimental results have been evaluated in terms of *accuracy*, *precision*, *recall* and *specificity* by taking the surveillance alerts as the benchmark, as well as *return* and *abnormal return* to indicate the business impact of those identified abnormal behaviors. It shows that the CHMM outperforms the standard HMM on any single sequence, and the HMM simply integrating multiple single HMMs without considering the coupling relationships between sequences, while the ACHMM can adapt to significant sequence changes in coupled sequences.

The rest of this paper is organized as follows. In Section 2, abnormal trading activity sequences are introduced as an example, followed by the definition of coupled sequences. Section 3 discusses the CHMM modeling coupled sequences, ACHMM for detecting sequence change, and algorithms for CHMM and ACHMM. Experiments on real-life stock exchange orderbook data are introduced and findings evaluated in Section 4. Finally, Section 5 concludes this paper.

2. PROBLEM DEFINITION

In this section, the problem is illustrated by a real-life example; We then discuss a coupled Hidden Markov Model for modeling three coupled sequences.

2.1 An Example: Coupled Trading Sequences

In stock markets, a trading transaction consists of an investor's trading action on its desired instrument at a particular trading price, volume and time point. Typical trading actions include 'place a buy order' ('buy' for short), 'place a sell order' ('sell' for short) and 'generate a trade' ('trade' for short, as an effect of matching a buy against a sell). For instance, Table 1 illustrates several order transactions related to a group manipulation situation identified in a stock market. Table 2 shows the corresponding trades, in which buys from investors (4) and (5) are traded against sell (2) at 10:02:02 at a total amount of 450 shares. Professional investors and sophisticated manipulators often collaborate with each other to manipulate a stock by carefully placing quotes and their prices, volumes and times to take advantage of or mislead other associated or opposite actions for maximizing personal benefits. As a result, such carefully manipulated trading behaviors contribute to abnormal market dynamics. For instance, as shown in Figure 1 which replays the trading behaviors in Table 1, investor (2) first placed a large buy at 10:00:35 to mislead other buyers after his/her partner (1)'s sell. To confuse other investors, (2) further placed a sell at 10:01:23 while (1) placed a buy at 10:01:38. After that more investors such as (4) and (5) followed up by submitting buy quotes at the same price as (2)'s sell. The above real-life story tells us that (a) buys, sells and trades are coupled with each other, reflecting investors' belief and intention. They need to be treated as a whole for anomaly analysis; if they are separated for scrutinization, or only trades are observed, it is challenging to capture the above manipulative cooperation between (1) and (2); (b) models for detecting abnormal trading behaviors should cater for the relationships amongst buys, sells and trades from both action and time perspectives, since the manipulation goal is achieved through a series of deliberate trading behaviors.

Table 1: An example of buy and sell orders

Investor	Time	Direction	Price	Volume
(1)	09:59:52	Sell	12.0	155
(2)	10:00:35	Buy	11.8	2000
(3)	10:00:56	Buy	11.8	150
(2)	10:01:23	Sell	11.9	200
(1)	10:01:38	Buy	11.8	200
(4)	10:01:47	Buy	11.9	200
(5)	10:02:02	Buy	11.9	250
(2)	10:02:04	Sell	11.9	500

Table 2: An example of trades

Investor	Time	Direction	Price	Volume
(4)	10:02:04	Buy	11.9	200
(5)	10:02:04	Buy	11.9	250
(2)	10:02:04	Sell	11.9	450

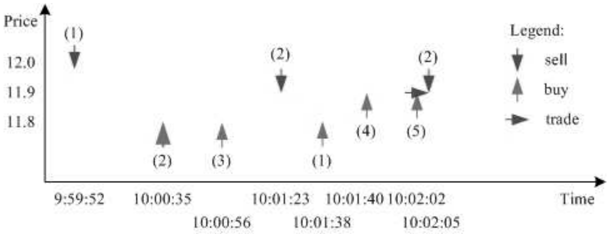

Figure 1: Coupled Trading Sequences

For the above cooperative manipulation, the usual sequence analysis tends to construct sequences by putting all actions from an investor together. While some interesting patterns for individual traders could be identified, it is not possible to detect abnormal behavioral patterns from related investors. It is also hard to model the coupling relationships among buys, sells and trades. Action properties such as prices and volumes cannot be fed into sequential analysis models. In order to detect such abnormal coupled trading behaviors, we convert the trading transactions from relevant investors into buy, sell and trade sequences, which are associated with corresponding prices/volumes. For instance, the transactions in Tables 1 and 2 are converted into the following buy, sell and trade sequences:

$$\{(buy, 11.8, 2000), (buy, 11.8, 150), (buy, 11.8, 200), (buy, 11.9, 200), (buy, 11.9, 250)\} \quad (1)$$

$$\{(sell, 12.0, 155), (sell, 11.9, 200), (sell, 11.9, 500)\} \quad (2)$$

$$\{(trade, 11.9, 200), (trade, 11.9, 250)\} \quad (3)$$

These three sequences enclose fruitful information about a group of related investors' beliefs and intentions (reflected through actions and related prices and volumes), as well as their implicit collaborations embodied through the price, volume and timing coupling. Also, they reflect the dynamic processes in which some traders intentionally adjust their behaviors and coupling relationships in the manipulation period. Abnormal behavior analysis needs to cater for such coupled sequences, their coupling relationships and dynamic adjustment of behaviors and relationships, otherwise it is almost impossible to identify such hidden group-based manipulation.

2.2 Coupled Hidden Markov Models Based Coupled Sequence Modeling

First, let us define the coupled sequences. Suppose we have C sequences: $\Phi_1 = \{\phi_{11}, \ldots, \phi_{1T}\}$, $\Phi_2 = \{\phi_{21}, \ldots, \phi_{2F}\}$ and $\Phi_C = \{\phi_{C1}, \ldots, \phi_{CG}\}$, T, F and G are the numbers of sequence items, the coupling relationship between two sequences Φ_i and Φ_j is $R_{ij}(\Phi_i, \Phi_j)$ which is a set (where $i, j \leq C, i \neq j$). $R_{ij} \subset R$, R is the set of coupling relationships for all C sequences.

Couple sequence modeling is carried out to understand and represent the coupling relationships R existing in C sequences. If $R_{ij}(\Phi_i, \Phi_j) = \varnothing$, we assume there is no coupling relationship in sequences Φ_i and Φ_j; otherwise, Φ_i and Φ_j are *coupled sequences.*

Further, a sequence item in a sequence, for instance ϕ_{11} in Φ_1, is often associated with the corresponding item properties, which we call *sequence item property.* For example, a buy action is associated with buy price, buy volume and buy time in a market. Let P_i represent the sequence item property set of the sequence Φ_i, its no$-k$ item ϕ_{ik} is further embodied in terms of its L item properties $\phi_{ik}(p_{ik,1}, \ldots, p_{ik,L})$.

Together with the sequence item properties, the coupled sequences form into *coupled behaviors.* It is certainly very complicated to model such coupled behaviors. In the following, we illustrate the coupled sequence modeling on the basis of Coupled Hidden Markov Models for market trading behaviors.

The three sequences shown in Section 2.1 cannot be separately observed. However, traditional sequence analysis cannot model such complex sequences and fit item properties such as prices and volumes. To this end, we use Coupled Hidden Markov Models (CHMM) to model these coupled sequences and their item properties, in order to detect abnormal coupled sequences and sequence changes within them.

In order to build CHMM for three coupled trading sequences, we define HMMs and CHMM as follows. We first build three HMMs: namely *HMM-B* for buy sequence Φ_B, *HMM-S* for sell sequence Φ_S and *HMM-T* for trade sequence Φ_T respectively.

Suppose there are N hidden states in an HMM, which are denoted as $S = \{S_1, S_2, \cdots, S_i, \cdots, S_N\}$, where S_i is an individual state. The state at time t is denoted as s_t. There are M distinct observation symbols per state in an HMM, denoted as $O = \{O_1, O_2, \cdots, O_i, \cdots, O_M\}$, where O_i is an individual symbol, the observation symbol at time t is denoted as o_t. An observation symbol corresponds to the output of the sequence being modeled. The probability distribution for the transition from state i to j is $X = \{x_{ij}\}$, where $x_{ij} = Pr(s_{t+1} = S_j | s_t = S_i), 1 \leq i, j \leq N$. The probability distribution for state j's observation is $Y = \{y_j(k)\}$, $y_j(k) = Pr(O_k | s_t = S_j), 1 \leq j \leq N, 1 \leq k \leq M$. Suppose the initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = Pr(s_1 = S_i), 1 \leq i \leq N$. As X and Y implicitly indicate N and M respectively, an HMM can be denoted as follows: $\lambda^{HMM} = (X, Y, \pi)$. For a trading sequence with T activities, according to [17], a model λ^{HMM} can be trained by the following re-estimated formulas with a set of observation sequences:

$$\bar{x}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) x_{ij} y_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(j)}, 1 \leq i, j \leq N \qquad (4)$$

$$\bar{y}_j(k) = \frac{\sum_{t=1, o_t = O_k}^{T} \alpha_t(j) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j) \beta_t(j)}, 1 \leq j \leq N \qquad (5)$$

α_t and β_t are the $forward$ and $backward$ variables at time t, $\bar{\pi}_i$, $\bar{x}_{ij}$ and $\bar{y}_j(k)$ are the expected parameters of model.

$$\bar{\pi}_i = \frac{\alpha_1(i) \beta_1(i)}{\sum_{j=1}^{N} \alpha_1(j) \beta_1(j)}, 1 \leq i \leq N \qquad (6)$$

After the model $\lambda^{HMM} = (X, Y, \pi)$ is trained, the probability that an observation sequence $Q = q_1, q_2, \cdots, q_T$ of a trading sequence is generated by λ^{HMM} can be computed by the formula:

$$Pr(Q | \lambda^{HMM}) = \sum_{i=1}^{N} \alpha_T(i) \qquad (7)$$

The coupling matrix between two coupled trading sequences is represented by $Z = \{z_{ij'}\}$, where $z_{ij'}$ represents the effect of S_i on $S_{j'}$. $z_{ij'} = Pr(s_{t+1}^{'} = S_{j'} | s_t = S_i)$, where S_i and $S_{j'}$ denote the hidden states of two interacting sequences Φ_i and Φ_j respectively. Correspondingly, a CHMM modeling three trading sequences can be expressed as $\lambda^{CHMM} = (X, Y, Z, \pi)$.

To evaluate the observation and train CHMM models, we use the $forward$-$backward\ analysis$. However, in a CHMM with 3 chains the joint state trellis is N^3 states wide, which leads to a computational complexity $O(TN^6)$ for computing $forward$ and $backward$ variables. To reduce the computational complexity, we use the approximate inference algorithms - N-heads dynamic programming [3], which relaxes the assumption that every transition must be visited, and thereby achieves $O(T(3N)^2)$.

3. CHMM MODELS AND ALGORITHMS

This section first introduces the details of CHMM model structure, hidden states and observation sequences. An adaptive CHMM (ACHMM) is introduced to capture sequence changes. Finally, algorithms for implementing the CHMM and ACHMM are discussed.

3.1 CHMM Model Structure

The CHMM model structure is shown in Figure 2, and consists of three chains of HMM modeling buy-orders Φ_B, sell-orders Φ_S and trades Φ_T respectively, which are fully coupled with each other via interactions. The circles denote the hidden states of the three trading sequences, for instance, S_{t-1}^{buy} denotes the hidden state for buy sequence at time $t-1$; the squares stand for the observation sequences of an HMM chain, for example, IA_{t-1}^{sell} indicates the observation of the sell sequence. The definition of hidden states and the construction of observation sequences are presented in Sections 3.2 and 3.3.

3.2 Hidden States

Based on domain knowledge, we define the hidden states in the CHMM in terms of an investor's belief, desire and intention (BDI), which are embodied through trading actions and their corresponding behavior characteristics.

- For the buy process, its hidden state S^{buy} is defined on the buy side, in which *Positive Buy*, *Neutral Buy* and *Negative Buy* are categorized in terms of profitable potential at the buy end according to domain knowledge.

  $S^{buy} = \{Positive\ Buy, Neutral\ Buy, Negative\ Buy\}$

- For the sell process, its hidden state S^{sell} denotes the investors' BDI on the sell side, which are embodied in terms of *Positive Sell*, *Neutral Sell* or *Negative Sell.*

  $S^{sell} = \{Positive\ Sell, Neutral\ Sell, Negative\ Sell\}$

- For the trade process, its hidden states S^{trade} stands for the trends of the market, labelled by *Market Up* or

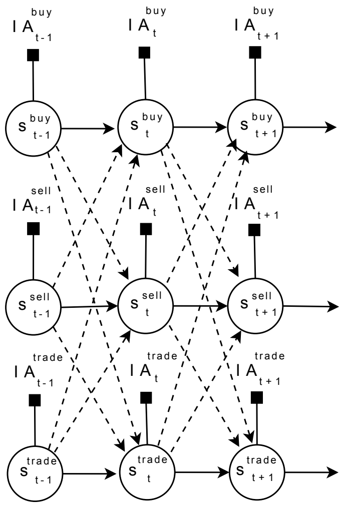

Figure 2: Architecture of CHMM

Market Down.

$$S^{trade} = \{Market\ Up, Market\ Down\}$$

The above hidden states reflect investors' BDI in manipulating a stock, which may shift from one to another, captured by the CHMM with particular probabilities.

3.3 Observation Sequences

The construction of observation sequences in the CHMM is based on two concepts: *activity* (A) and *interval activity* (IA) as follows, which involve human intention information embodied through sequence item property sets P_B for the Buy sequence Φ_B, P_S for the Sell sequence Φ_S, and P_T for the Trade sequence Φ_T, respectively. The item property P is actually embodied through factors such as trading prices, volumes and times in stock markets.

Definition 1. *Activity (A) represents a subject's individual behavior. In capital markets, A is a trading behavior, which consists of an atomic trading action ($a = \{buy \mid sell \mid trade\}$) taken by an investor, associated with the investor's BDI information embodied through the sequence item properties p and v at time t. The three variables a, p and v reflect the cause and effect of an investor's trading behavior in a market. $A = \{a_1, a_2, \ldots, \}$), where $a_i = (a(t_i), p(t_i), v(t_i))$. $a(t_i) = \{buy \mid sell \mid trade\}$, which represents one of the three trading actions in capital markets: buy-order, sell-order or trade at time t_i, its associated BDI variables $p(t_i)$ and $v(t_i)$ are defined as follows: $p(t_i) = \{buy\ price|sell\ price|$ $trade\ price\}$ is the price of the corresponding trading action $a(t_i)$ at t_i. Similarly, $v(t_i) = \{buy\ volume|sell\ volume|$ $trade\ volume\}$ is the trade volume of $a(t_i)$ at t_i.*

Definition 2. *Interval Activity (IA) represents a collective behavior embodied through the collective properties of a behavior sequence. $IA = (\mathcal{A}, \bar{p}, \bar{v}, f)$, $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, $A_i(a) = A_j(a)$ consists of a set of trading activities taking place in window l with size winsize. The variables $\bar{p}$, $\bar{v}$ and f represent the accumulative activities and their average prices and volumes of an investor or a group of investors:*

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i}{f} \tag{8}$$

$$\bar{v} = \frac{\sum_{i=1}^{n} v_i}{f} \tag{9}$$

$$f = |\mathcal{A}| = n \tag{10}$$

n is the number of activities in the window l.

To map IAs to the observation symbols of CHMM, we quantize $\bar{p}, \bar{v}$ and f based on the k-means clustering algorithm, and generate the observation variable $IA(p^{'}, v^{'}, f^{'})$.

$$IA(\mathcal{A}, \bar{p}, \bar{v}, f) \xrightarrow{\text{quantization}} IA^{'}(p^{'}, v^{'}, f^{'}) \tag{11}$$

$IA^{'}(p^{'}, v^{'}, f^{'})$ is calculated as follows. Taking the dimension $\bar{p}$ of IA as an example, the values of $\bar{p}$ are first grouped into several clusters through the k-mean clustering algorithm. Let θ_i^p be the centroid of the i^{th} generated cluster. Then the discrete values of $p^{'}$ are given by the formula:

$$p^{'} = argmin_i|\bar{p} - \theta_i^p| \tag{12}$$

Similarly, we quantize the dimensions V and W as follows:

$$v^{'} = argmin_i|\bar{v} - \theta_i^v| \tag{13}$$

$$f^{'} = argmin_i|f - \theta_i^f| \tag{14}$$

where θ_i^v and θ_i^f are the centroids of clusters for $\bar{v}$ and f respectively.

3.4 Adaptive CHMM for Detecting Sequence Changes

As the trading activities in stock markets change frequently due to the dynamics in investor sentiment and the external market environment, it is important to make the CHMM adaptive to significant changes of trading sequences. For this purpose, we improve the CHMM by adding an automatically adaptive mechanism to form an Adaptive CHMM (ACHMM). The idea is as follows. During the course of detecting abnormal trading behaviors, the CHMM model is updated if a significant change in the outputs of CHMM is detected by the Output Analyzer. The automatic and adaptive detection and adjustment in ACHMM rely on the problem-solving of two issues: how to detect the significant change, and how to update the model instantly.

To solve the first problem, we use t test to check if there is a significant difference between the current outputs and its benchmark. The current benchmark consists of the outputs generated right after the last update of the CHMM model. A significant difference indicates the CHMM cannot capture the corresponding changes in trading activities properly, and needs to be updated. As shown in Figure 3, dataset DS_1 is drawn from the trading *window 1* $[t_1, t_2]$. The outputs generated by *model 1* on DS_1 is taken as the benchmark for detecting the abnormal sequences in *window 2* $[t_3, t_4]$ with the same size as *window 1*. If there is a significant difference in the t-test result between the outputs generated on DS_2 and the benchmark, then *model 1* should be updated and the outputs generated by *model 2* updated on DS_2 are treated as the benchmark *window 3*. The model update is based on a sliding strategy. We first use the parameters of model 1 on window 1 as the initial settings to train model 2 on window 2 only, and then update model 1 based on the parameters gained for model 2. In other words, we retrain the model on the most recent dataset rather than the whole

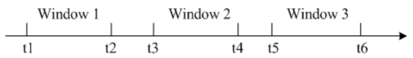

Figure 3: Update Point of ACHMM

training dataset. This strategy enables us to avoid the great expense of model retraining. The retraining of the CHMM parameters x, y, z and π is based on the following formulas to update the model:

$$x_{ij}^{update} = (1-w)x_{ij}^{old} + w * x_{ij}^{new} \tag{15}$$

$$y_{ij}^{update} = (1-w)y_{ij}^{old} + w * y_{ij}^{new} \tag{16}$$

$$z_{ij'}^{update} = (1-w)z_{ij'}^{old} + w * z_{ij'}^{new} \tag{17}$$

$$\pi_{i}^{update} = (1-w)\pi_{i}^{old} + w * \pi_{i}^{new} \tag{18}$$

where w is a weight that reflects the bias towards the most current dataset, x_{ij}^{old}, y_{ij}^{old}, $z_{ij'}^{old}$ and π_{ij}^{old} are the parameters of the previous model, x_{ij}^{new}, y_{ij}^{new}, $z_{ij'}^{new}$ and π_{ij}^{new} are parameters of the new model.

3.5 The Algorithms

The key algorithms for ACHMM based abnormal trading behavior analysis consist of two aspects: (a) extracting and spliting trading sequences from the trading transactional data to construct the observation sequences for the CHMM, and (b) detecting abnormal trading activity sequences by feeding the three trading sequences into the CHMM model.

In stock markets, orders are placed by investors after the market opens and expire after the market closes if they are not traded. Trades are based on the orders placed on the same day only. This market mechanism indicates that all orders and trades for a stock on the same day are closely related. We segment a trading day into multiple windows. Within each trading window, its trading transactions are first grouped into buy, sell and trade *interval activities* and then put together to generate buy, sell and trade sequences respectively. Algorithm 1 extracts trading activity sequences, which form the observation sequences of the CHMM.

Algorithm 1 Constructing observation sequences

```
Step 1: Segment the whole trading day into $L$ intervals by a time window with the length $winsize$.
Step 2: Calculate $IA$ for buy-order, sell-order and trade activities respectively in each window. They are denoted as $IA_l^{buy}$, $IA_l^{sell}$ and $IA_l^{trade}$, respectively.
Step 3: Obtain $IA_l^{'buy}$, $IA_l^{'sell}$ and $IA_l^{'trade}$ by quantizing $IA_l^{buy}$, $IA_l^{sell}$ and $IA_l^{trade}$.
Step 4: Obtain the trading activity sequnce $IA^{buy}$ for buy-order by putting all $IA_l^{'buy}$ in a trading day together. Obtain $IA^{sell}$ and $IA^{trade}$ in the same way. We obtain
    $IA^{type} = IA_1^{'type}, IA_2^{'type}, \cdots, IA_L^{'type}$    (19)
  where $type \in \{buy, sell, trade\}$. $IA^{buy}$, $IA^{sell}$ and $IA^{trade}$ are the observation sequences of CHMM in the day.
Step 5: Repeat Step 1-4 for each trading day
```

With the ACHMM trained, an algorithm is developed to detect abnormal trading behaviors. Its purpose is to calculate the distance from a sequence to the centroid of a model. If the distance is larger than a user-specified threshold ψ_0, then the sequence is considered to be abnormal. The procedure of constructing an abnormal trading activity sequence is shown in Algorithm 2.

Algorithm 2 Detecting abnormal trading sequences

```
Step 1: Construct trading sequences including training sequences $Seq_1, Seq_2, \cdots, Seq_K$ and test sequences $Seq_1^{'}, Seq_2^{'}, \cdots, Seq_{K'}^{'}$.
Step 2: Train the ACHMM model on the training sequences;
Step 3: Compute the mean ($\mu$) and standard deviation ($\sigma$) of probability of training sequences according to the following formulas:
    $\mu = \frac{\sum_{i=1}^{K} Pr(Seq_i|ACHMM)}{K}$    (20)
    $\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K} Pr(Seq_i|ACHMM)) - \mu}$    (21)
  where $K$ is the total number of training sequences, mean $\mu$ represents the centroid of model ACHMM, and the standard deviation $\sigma$ represents the radius of model ACHMM.
Step 4: For each test sequence $Seq_i^{'}$, calculate its distance $D_i$ to the centroid of model by
    $D_i = \frac{\mu - Pr(Seq_i^{'}|\mathcal{M})}{\sigma}$    (22)
  Consequently, $Seq_i^{'}$ is an exceptional pattern, if it satisfies:
    $D_i > \psi_0$    (23)
  where $\psi_0$ is a given threshold.
```

4. EXPERIMENTS AND DISCUSSIONS

This section discusses the experimental data and benchmark models, followed by performance evaluation.

4.1 Experimental Data

The trading transactions for the experiments are from a stock market and cover 388 trading days from June 2004 to December 2005. We use the data from June 2004 to December 2004 as training data, and the remaining data as test data. In the experimental data, there are some trading days associated with alerts generated by the surveillance system used in the stock exchange. These alerts are not absolutely true positive but they can be used as a reference to evaluate our models, to see whether our model can identify anomalies which have been alarmed by other indicators used in the surveillance system. We remove the data with alerts from the training data to enable our model to be trained on 'normal' data alone, and leave the data with alerts to the test data, in which the alerts are used for evaluating the model detection performance (we understand some alerts may be false positive, but the alerts provide a rough benchmark for us to evaluate our models against the detection performance of market surveillance rules, especially when it is very costly to obtain labeled data).

4.2 Benchmark Models

In order to evaluate the performance of the CHMM and ACHMM, we build another four HMM models as the experimental benchmarks, namely HMM-B, HMM-S, HMM-T and IHMM. They are explained as follows.

- *HMM-B*: an HMM on buy sequence including buys from all investors, without adaptive mechanism.
- *HMM-S*: an HMM on sell sequence including sells from all investors, without adaptive mechanism.
- *HMM-T*: an HMM on trade sequence including all trades, without adaptive mechanism.
- *IHMM*: an integrated HMM combining *HMM-B*, *HMM S* and *HMM-T*. The probability of *IHMM* is the sum of the probability values of the three models. It does not consider the coupled relationships amongst the three processes and also has no adaptive mechanism.
- *CHMM*: a CHMM model on trade, buy-order and sell-order sequences. It considers the coupled relationships, but has no adaptive mechanism.
- *ACHMM*: a CHMM model on trade, buy-order and sell-order sequences considering the coupled relationships and adaptive mechanism.

As we can see, HMM-B, HMM-S or HMM-T only capture one sequence, while IHMM somehow represents a traditional way of modeling multiple sequences through a simple combination. CHMM represents a new mechanism for modeling multiple sequences with coupling relationships; ACHMM is an adaptive model which can cater for changes in multiple sequences. The comparison between HMM-B, HMM-S, HMM-T/IHMM and CHMM/ACHMM indicates not only the importance of the new approach to modeling coupled sequences, but also the limitation of either modeling a single sequence or merging multiple sequences into one sequence in a coupled sequence.

4.3 Technical Performance

The technical performance evaluation of a model is based on *accuracy*, *precision*, *recall*, and *specificity*.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (24)$$

$$Precision = \frac{TP}{TP+FP} \quad (25)$$

$$Recall = \frac{TP}{TP+FN} \quad (26)$$

$$Specificity = \frac{TN}{FP+TN} \quad (27)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative. TP, TN, FP and FN are counted in terms of the abnormal cases identified in the data and verified by domain experts. These metrics measure the detection quality of a model.

We test the six models on the experimental data by setting various window sizes ($winsize$). Figures 4, 5, 6 and 7 show their technical performance, where the horizontal axis (*P-Num*) stands for the number of detected abnormal activity sequences, and the vertical axis represents the values of technical measures. CHMM and ACHMM outperform the other four benchmark models with any window sizes ($winsize$), while ACHMM performs the best.

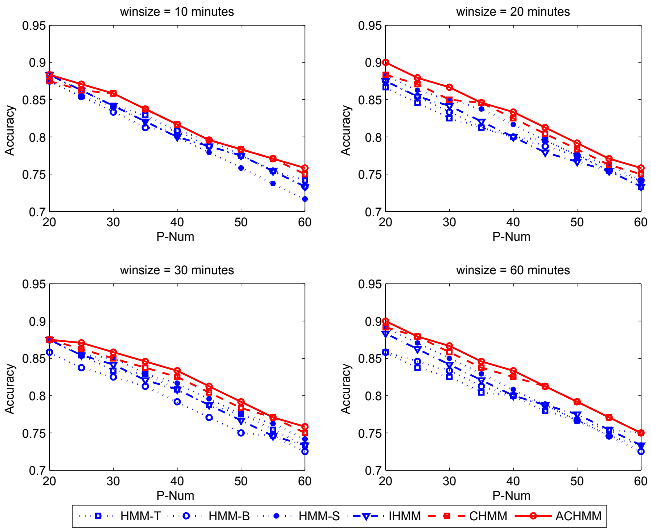

Figure 4: Accuracy of Six Models

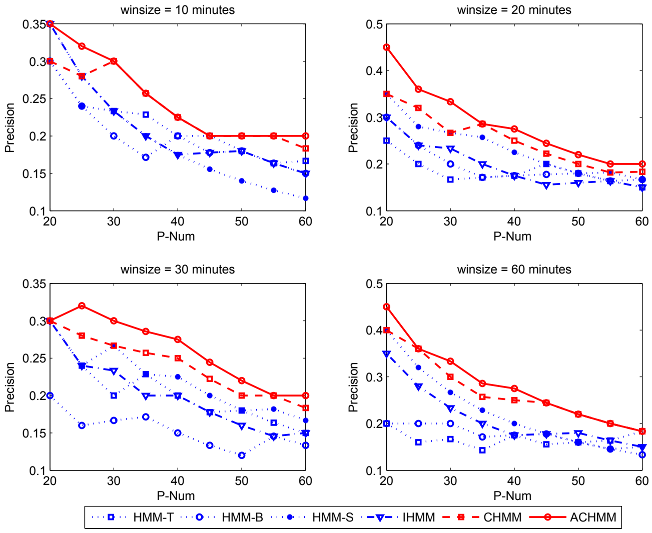

Figure 5: Precision of Six Models

For instance, in Figure 6, when $P-Num = 20$ and $winsize = 20$, the precision of CHMM is 0.35, ACHMM is 0.45, while HMM-T is only 0.25, so the precision of ACHMM can be as much as over 50% better than HMM-T. This shows that performance of the HMM only modeling trade sequence is much lower than the CHMM modeling three coupled sequences, as well as the ACHMM catering for sequence changes. Further, ACHMM generally beats CHMM. When the number of detected sequences increases, more false positive (FP, abnormal alerts) may be captured, which correspondingly reduces aspects of the model's performance, such as Precision. However, ACHMM retains its advantage over all others (see Figures 5 and 6, $P-Num = 60$). In particular, the recall increases with P-Num rising, which shows the

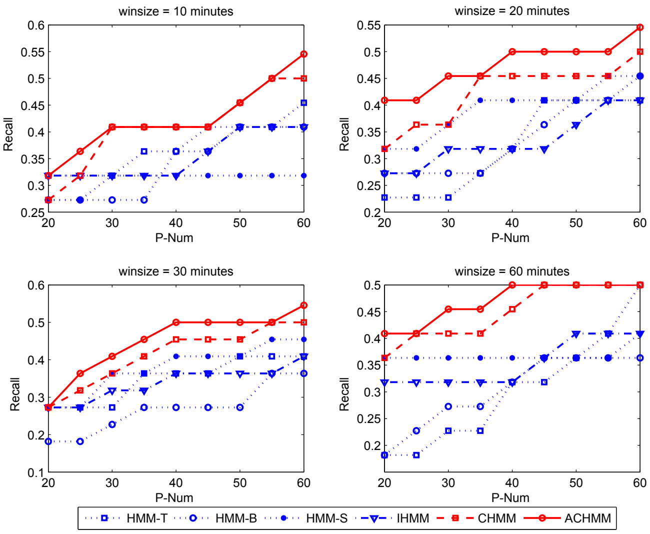

Figure 6: Recall of Six Models

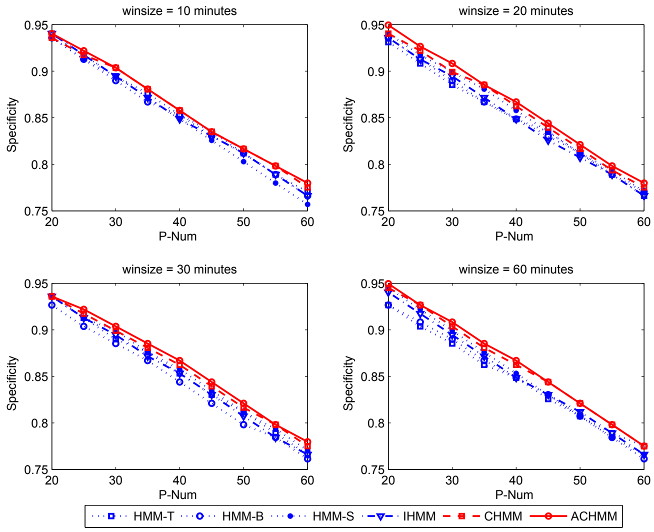

Figure 7: Specificity of Six Models

HMMs trained on 'normal' data can contribute to a lower false negative (FN, i.e. false 'normal'). Therefore, CHMM and ACHMM can detect abnormal trading sequences better than any HMM modeling a single sequence only.

4.4 Computational Performance

Finally, we evaluate the computational performance of IHMM, CHMM and ACHMM. As *HMM-B*, *HMM-S* and *HMM-T* only model one trading activity sequence, they are excluded from the computational performance evaluation. As shown in Table 3, CHMM and ACHMM cannot outperform IHMM, which is understandable. CHMM and ACHMM need much more time in general to calculate the coupling matrix and to adjust models.

Table 3: Computational performance

		IHMM	CHMM	ACHMM
winsize =10 (m)	Training time (s)	0.574	11.978	11.988
	Test time (s)	0.056	1.296	3.576
winsize =20 (m)	Training time (s)	0.256	4.929	4.933
	Test time (s)	0.047	0.655	3.486
winsize =30 (m)	Training time (s)	0.206	4.121	4.119
	Test time (s)	0.042	0.447	2.429
winsize =60 (m)	Training time (s)	0.109	2.003	2.004
	Test time (s)	0.036	0.221	1.206

5. CONCLUSION

The financial crisis has warned market regulators of the critical need to develop effective detection techniques for identifying hidden group based manipulative trading behaviors in capital markets. Although sequence analysis can play an important role in this regard, typical existing work mainly focuses on analyzing single sequences or combining sequences into one single sequence. This inevitably ignores coupling relationships in which multiple sequences are coupled with each other for various reasons and properties such as trade prices are associated with sequence items. Furthermore, such coupled sequences often involve changes that can significantly challenge the performance of a trained model. This raises a challenging and critical research issue, namely, analyzing dynamic and coupled behaviors in hidden groups.

This paper has presented an effective coupled Hidden Markov Model and an adaptive CHMM model trained on multiple 'normal' coupled sequences for detecting abnormal group based manipulative trading behaviors: (a) from multiple sequences interacting with each other, (b) involving the sequence item properties, and (c) adapting to the significant change of such sequences. A system GMBAS has been developed and intensively tested on real-life orderbook-level data. The results have shown that the CHMM can outperform a single HMM only modeling any single trading sequence such as trade activities, and can beat an HMM combining multiple single-sequence-based HMMs, ignoring interactions among them, in terms of both technical performance and business impact. The ACHMM built on top of the CHMM with the adaptive mechanism can adapt to changes in coupled sequences, and beat the CHMM in both technical and business performance aspects. Most importantly, our models can identify anomalies that cannot be detected by current market surveillance systems and sequence analysis methods.

In fact, the findings are helpful for dealing with coupled behavior analysis with multiple sequences and item properties, hidden group analysis, and behavior dynamics in other similar areas such as social networks and crime networks. We are currently investigating the detection of hidden groups and the possibility of updating existing sequence analysis methods for analyzing multiple coupled sequences, and improving the computational performance of CHMM and ACHMM.

6. ACKNOWLEDGMENTS

This work is sponsored in part by Australian Research Council Discovery Grants (DP1096218, DP0988016, DP0773412) and ARC Linkage Grant (LP0989721, LP0775041), as well as supported in part by American NSF through grants IIS-0905215 and IIS-0914934.

7. REFERENCES

[1] A. Karwath and N. Landwehr. Boosting relational sequence alignments. In *The Eighth IEEE International Conference on Data Mining Proceedings*, pages 857–862, 2008.

[2] Y. Bengio. Markovian models for sequential data. *Neural Computing Surveys*, pages 129–162, 1999.

[3] M. Brand. Coupled hidden markov models for modeling interacting processes. *Tech. Rep., MIT Media Lab*, 1997.

[4] S. J. Brown and J. B. Warner. Using daily stock returns: The case of event studies. *Journal of Financial Economics*, 14(1):3–31, 1985.

[5] L. Cao and Y. Ou. Market microstructure patterns powering trading and surveillance agents. *Journal of Universal Computer Sciences*, 14(14):2288–2308, 2008.

[6] L. Cao and P. Yu. Behavior informatics: An informatics perspective for behavior studies. *The Intelligent Informatics Bulletin*, 10(1):6–11, 2009.

[7] S. B. Cho and H. J. Park. Efficient anomaly detection by modeling privilege flows using hidden markov model. *Computer and Security*, 22(1):45–55, 2003.

[8] D. Kifer and J. Gehrke. Detecting change in data streams. In *The Thirtieth International Conference on Very Large Data Bases Proceedings*, pages 180–191, 2004.

[9] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, pages 245–275, 1997.

[10] R. Gwadera and F. Crestani. Discovering significant patterns in multi-attribute sequences. In *The Eighth IEEE International Conference on Data Mining Proceedings*, pages 827–832, 2008.

[11] J. Ayres, J. Flannick and T. Yiu. Sequential pattern mining using a bitmap representation. In *The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining Proceedings*, pages 429–435, 2002.

[12] J. Pei, J. Han and M. C. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *The 17th International Conference on Data Engineering Proceedings*, pages 215–226, 2001.

[13] S. Joshi and V. Phoha. Investigating hidden markov models capabilities in anomaly detection. In *The 43rd Annual Southeast Regional Conference Proceedings*, pages 98–103, 2005.

[14] L. Cao, Y. Zhao and H. Zhang. Activity mining: from activities to actions. *International Journal of Information Technology and Decision Making*, 7(2):259–273, 2008.

[15] L. Cao, Y. Zhao and C. Zhang. Mining impact-targeted activity patterns in imbalanced data. *IEEE Trans. on Knowledge and Data Engineering*, 20(8):1053–1066, 2008.

[16] M. Oliver and A. P. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:831–843, 2000.

[17] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 275–286, 1989.

[18] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *The 5th International Conference on Extending Database Technology Proceedings*, pages 3–17, 1996.

[19] X. Song, M. Wu and S. Ranka. Statistical change detection for multi-dimensional data. In *The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Proceedings*, pages 667–676, 2007.

[20] Y. M. Ou, L. B. Cao and C. Q. Zhang. Adaptive anomaly detection of coupled activity sequences. *The IEEE Intelligent Informatics Bulletin*, 10(1):12–16, 2009.

[21] M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42:31–60, 2001.