Coupled Behavior Analysis with Applications

Longbing Cao, *Senior Member, IEEE*, Yuming Ou, and Philip S Yu, *Fellow, IEEE*,

Abstract—Coupled behaviors refer to the activities of one to many actors who are associated with each other in terms of certain relationships. With increasing network and community-based events and applications, such as group-based crime and social network interactions, behavior coupling contributes to the causes of eventual business problems. Effective approaches for analyzing coupled behaviors are not available, since existing methods mainly focus on individual behavior analysis. This paper discusses the problem of coupled behavior analysis and its challenges. A Coupled Hidden Markov Model (CHMM)-based approach is illustrated to model and detect abnormal group-based trading behaviors. The CHMM models cater for: (1) multiple behaviors from a group of people, (2) behavioral properties, (3) interactions among behaviors, customers and behavioral properties, and (4) significant changes between coupled behaviors. We demonstrate and evaluate the models on orderbook-level stock tick data from a major Asian exchange and demonstrate that the proposed CHMMs outperforms HMM-only for modeling a single sequence or combining multiple single sequences, without considering coupling relationships to detect anomalies. Finally, we discuss interaction relationships and modes between coupled behaviors, which are worthy of substantial study.

Index Terms—Coupled behavior analysis, coupled sequence analysis, hidden group discovery, coupled Hidden Markov Model, abnormal behavior detection.

1 Introduction

Behavior analysis is an essential activity in many fields, from social and behavioral sciences to computer science [32], [33], [34], [35], [36], [37]. Although there is an emerging focus on deep behavior studies such as periodic behavior analysis [31] and social network analysis [30], previous research has mainly focused on individual behaviors. In practice, behaviors from either the same, or different actors are often coupled with each other. Coupled behaviors play a much more fundamental role than individuals in the cause, dynamics and effect of business problems [28], [7], [29], [30], [37].

1.1 Coupled Behavior Applications

While very limited research outcomes can be identified in the literature, coupled behavior is widely researched. As well as the example in Section 3.1, the following are typical *coupled behavior applications*.

- Group-based criminal behaviors: a group of criminals conduct a series of activities in order to achieve their goal. The activities are associated with each other and aim for the same objective.
- Group-based insurance claims: a family or group of insurants lodge similar claims at the same time, or soon after. Another example is where a health care provider may collaborate with multiple customers to over-claim health benefits by approving frequent visits by the customers for a variety of services. Such group claims may lead to over-claims or over-use of services.
- Cross-reference citation analysis: from the references cross-cited by relevant groups, we find either genuine collaboration or manipulation of citations.
- Cross-market manipulation: investors in an underlying market manipulate a security so that an accomplice can take arbitrage on the corresponding instrument listed in a derivative market.
- Car transport system: at a busy intersection, many cars from different localities compete/cooperate with each other to move in their respective directions.
- Social network interactions: a group of users interact with each other in a social network.
- Intrusion detection: a large number of hackers collaborate to interfere with a website by applying multiple intrusion techniques.

With the deepening and widening of networking, these coupled behaviors are increasing in a wide range of circumstances, in particular, complex networks, communities, organizations and enterprise applications.

1.2 Challenges in Analyzing Coupled Behaviors

In the above applications, multiple traces of behaviors are often coupled in intrinsic and contextual relationships. The focus on any single trace of behaviors would not contribute to a full understanding of the underlying problem and its comprehensive solutions. It is very difficult to analyze such coupled behaviors.

- Behaviors refer not only to actions such as a buy quote, but also behavioral properties, for instance, the timing, price and volume associated with a buy. The engagement of behavioral properties in behavior analysis may make the findings much more workable for problem-solving.

Longbing Cao and Yuming Ou are with the University of Technology, Sydney, Australia. E-mails: {lbcao, yuming}@it.uts.edu.au. Philip S Yu is with the University of Illinois at Chicago, E-mail: psyu@cs.uic.edu.

- Behaviors are correlated in terms of certain coupling relationships. The difficulty is that the coupling relationships are often not obvious. A deep exploration of the relationships is necessary for us to understand how behaviors are correlated.
- It is important to model and analyze coupled behaviors as a whole. For this, both coupling relationships and behavioral properties need to be modeled.
- An additional challenge in coupled behavior analysis arises from behavior dynamics. Any significant change taking place in any behavior sequence, coupling relationships or behavioral properties could seriously affect a model's performance.

The above characteristics challenge the existing behavior-related analysis approaches including sequence analysis, multi-variate time series, and interactive process modeling. As a typical approach for understanding behavior, commonly-used sequence analysis algorithms such as GSP [16], PrefixSpan [11], Spam [10] and Spade [18] target only single sequences. Arguably, the correlated sequences could be merged into one sequence and then analyzed by these methods; however, this overlooks the relationships that associate the relevant sequences and the properties of sequence items. Another issue is that the classic sequence-based behavior/event analysis methods often ignore the associated behavioral properties. They cannot be directly used for coupled behavior analysis. Similarly, multi-variate time series analysis mainly correlates multiple numerical variables. It is not aimed at behavior analysis, and overlooks aspects such as coupling relationships and behavior properties.

1.3 Contributions

While coupled behaviors are increasingly seen in complex business applications and social networks, to the best of our knowledge, as outlined in [2], there is no related work that directly models and analyzes such coupled behaviors. Considering the complexities and difficulties in handling the problem, as a preliminary attempt, we formalize the problem, present a case study approach to detect anomalies from multiple coupled behavior sequences, and discuss interactions between behaviors. The main areas of our work are as follows.

First, the problem of coupled behavior analysis (CBA) is discussed, together with its applications and challenges. This provides a clear problem definition and explains the significance of exploring coupled behavior interactions.

Second, we illustrate the CBA problem by proposing a Coupled Hidden Markov Model (CHMM) based approach to model and analyze abnormal coupled trading behaviors. The CHMM model converts and extracts multiple sequences and item properties from order-book-level trading transactions, which are associated with each other in terms of certain relationships. To the best of our knowledge, this is a novel approach in sequence analysis.

Third, we enhance the adaptability of the CHMM by proposing an Adaptive CHMM (ACHMM) to monitor sequence dynamics and automatically check the difference against the benchmarks. It can adaptively retrain itself to accommodate significant changes in coupled behaviors. 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.

Substantial experiments have been conducted on real-life data from a major Asian stock market. The findings are evaluated in terms of not only technical performance (such as *recall*), but also the business performance (such as *abnormal return*) of trading on those identified trading behaviors. This verifies the workability of the results. From the business perspective, our approach leads to novel contributions towards *pattern-based* and *adaptive* market surveillance, which is currently not available but urgently needed to improve the regulation of globalized, volatile and cross-market operations.

2 Related Work

To the best of our knowledge, only limited efforts appear to have been made in deep analysis of coupled behaviors. A slightly relevant area is multivariate time series based analysis, but other than this, we have not found any related work that directly analyzes coupled behaviors as discussed in this paper. Below, we briefly discuss multivariate time series, sequence analysis and coupled Hidden Markov Model as it relates to this paper.

Multivariate time series Multivariate time series analysis is an emerging area for complex data analysis. [20] proposes methods for feature sub-set selection from multivariate time series based on common principal component analysis. [21] develops a temporal abstraction framework for generating multivariate time series features suitable for classification tasks. Grouping of variables in multivariate time series is discussed in [19]. [23] proposes a density based clustering method in the kernel feature space for clustering multivariate time series data of varying lengths. Clustering [25], [24], frequent pattern mining [26] and classification [21] have also been investigated in multivariate time series data. The main difference between the current multivariate time series analysis and coupled behavior analysis is in the consideration of coupling relationships and the mixed attributes involved in the coupled behaviors, which are beyond time series.

Sequence analysis In sequence analysis, typical algorithms including GSP [16], PrefixSpan [11], Spam [10] and Spade [18] mainly deal with single sequences. Although other 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 [9], it is difficult to mine coupled multiple sequences embedded with item properties.

Coupled Hidden Markov Model HMM cannot describe systems with multiple interacting processes such as the above three coupled trading sequences. CHMM [14] is proposed 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, not only the state of its own chain of HMM, but also on 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], [17] has become 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.

3 Problem Statement

In this section, the problem is first illustrated by a real-life example, and then formalized.

3.1 An Example: Coupled Trading Behaviors

In stock markets, a trading transaction consists of an investor's trading action on his/her 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). Professional investors and sophisticated manipulators often collaborate with each other to manipulate a stock by carefully placing quotes, prices, volumes and times to take advantage of associated, or opposite actions to maximise personal benefits. As a result, such carefully manipulated trading behaviors contribute to abnormal market dynamics. For instance, Table 1 illustrates several order transactions related to a group manipulation situation identified in an Asian stock market. Table 2 shows the corresponding trades, in which buys from investors (4) and (5) are executed against sell (2) for a total amount of 450 shares.

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

Fig. 1 further illustrates the above group manipulation process. Investor (2) first placed a large buy at 10:00:35 to mis-lead other buyers after the sell by his/her partner (1). To confuse other investors, (2) placed a sell at 10:01:23 while (1) placed a buy at 10:01:38. Subsequently, more investors, such as (4) and (5) followed up by submitting buy quotes at the same price as (2)'s sell.

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

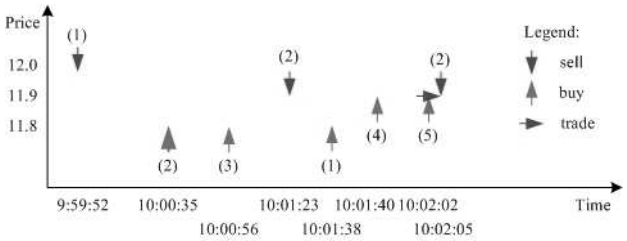

Fig. 1. Coupled Trading Behaviors

This real-life story tells us that: (a) Buys, sells and trades are coupled with each other, and need to be treated as a whole for anomaly analysis. If they are separated for scrutinization, or only trades are observed, it is difficult to capture the 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.

If the usual sequence analysis methods are used for the above group manipulation, sequences are constructed by putting all actions from an investor together. While some interesting patterns of individual traders could be identified, it is not possible to detect abnormal group behavioral patterns from the related investors. It is also hard to model the coupling relationships among buys, sells and trades. Behavior properties such as prices and volumes cannot be fed into sequential analysis models.

3.2 Coupled Behavior Analysis Problem

Behaviors refer to actions, operations, events and activity sequences conducted within certain contexts and environments in either a virtual or physical organization. We first define an abstract behavior model.

Definition 1. *A **behavior** ($\mathbb{B}$) is described as a four-ingredient tuple $\mathbb{B} = (\mathscr{E}, \mathscr{O}, \mathscr{C}, \mathscr{R})$,*

- *Actor $\mathscr{E} = \langle \mathcal{SE}, \mathcal{OE} \rangle$ is the entity that issues a behavior (subject, $\mathcal{SE}$) or on which a behavior is imposed (object, $\mathcal{OE}$).*
- *Operation $\mathscr{O} = \langle \mathcal{OA}, \mathcal{SA} \rangle$ is what an actor conducts in order to achieve certain goals; both objective ($\mathcal{OA}$) and subjective ($\mathcal{SA}$) attributes are associated with an operation. Objective attributes may include time, place, status and restraint; while subjective aspects may refer to action and its actor's belief and goal, etc. of the behavior and the behavior impact on business.*

- *Context $\mathscr{C}$ is the environment in which a behavior takes place.*
- *Relationship $\mathscr{R} = \langle \theta(\cdot), \eta(\cdot) \rangle$ is a tuple which reveals complex interactions within an actor's behaviors (named intra-coupled behaviors, represented by function $\theta(\cdot)$) and that between multiple behaviors of different actors (inter-coupled behaviors by relationship function $\eta(\cdot)$).*

For simplicity, operation and behavior is interchangeable in this paper. Accordingly, operation attributes indicate behavior properties.

Suppose there are I actors (customers) $\{\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_I\}$, an actor $\mathscr{E}_i$ undertakes J_i behaviors $\{\mathbb{B}_{i1}, \mathbb{B}_{i2}, \ldots, \mathbb{B}_{iJ_i}\}$, actor $\mathscr{E}_i$'s j^{th} behavior $\mathbb{B}_{ij}$ is a K-variable vector, its variable $[p_{ij}]_k$ reflects the k^{th} behavior property. From this perspective, we get a Behavior Feature Matrix $FM(\mathbb{B})$ as follows:

$$FM(\mathbb{B}) = \begin{pmatrix} \mathbb{B}_{11} & \mathbb{B}_{12} & \cdots & \mathbb{B}_{1J_{max}} \\ \mathbb{B}_{21} & \mathbb{B}_{22} & \cdots & \mathbb{B}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{B}_{I1} & \mathbb{B}_{I2} & \cdots & \mathbb{B}_{IJ_{max}} \end{pmatrix}$$

Where $J_{max} = max\{J_1, J_2, \cdots, J_I\}$, for every behavior set $\{\mathbb{B}_{ij} | J_i < J_{max}\}$, the corresponding element $\mathbb{B}_{ij}$ is recognized as $\emptyset$ when $J_i < j \leq J_{max}$. Further, each (i, j) element of this matrix $FM(\mathbb{B})$ is actually a row vector, expressed as $\overrightarrow{\mathbb{B}}_{ij} = ([p_{ij}]_1, [p_{ij}]_2, \cdots, [p_{ij}]_K)$, where $[p_{ij}]_k$ $(1 \leq k \leq K)$ is the k^{th} property of the behavior $\mathbb{B}_{ij}$. The intra-coupling is the relationship within one row of the above matrix, while how the behaviors interact is embodied among the columns of $FM(\mathbb{B})$, indicated as inter-coupling.

In the following, for an actor $\mathscr{E}_i$, we use the corresponding behavior amount J_{max} instead of J_i, then the relevant coupling functions constantly equal to 0 when $\mathbb{B}_{ij} = \emptyset$, i.e., $J_i < j \leq J_{max}$.

Definition 2. *(Intra-Coupled Behaviors) Actor $\mathscr{E}_i$'s behaviors $\mathbb{B}_{ij}$ $(1 \leq j \leq J_{max})$ are intra-coupled in terms of coupling function $\theta_j(\cdot)$,*

$$\mathbb{B}_{i\cdot}^{\theta} ::= \mathbb{B}_{i\cdot}(\mathscr{E}, \mathscr{O}, \mathscr{C}, \theta) | \sum_{j=1}^{J_{max}} \theta_j(\cdot) \odot \mathbb{B}_{ij} \quad (1)$$

$$|\theta_j(\cdot)| \geq \theta_0 \quad (2)$$

where θ_0 is the intra-coupling threshold, $\sum_{j=1}^{J_{max}} \odot$ means the subsequent behavior of $\mathbb{B}_i$ is $\mathbb{B}_{ij}$ intra-coupled with $\theta_j(\cdot)$, and so on, with nondeterminism.

Corollary 1. *If $\theta_j(\cdot) < 0$, $\mathbb{B}_i^{\theta}$ has negative intra-coupling; if $\theta_j(\cdot) > 0$ then there is a positive intra-coupling relationship; $\theta_j(\cdot) = 0$ indicates none intra-coupling.*

Definition 3. *(Inter-Coupled Behaviors) Actor $\mathscr{E}_i$'s behaviors $\mathbb{B}_{ij}$ $(1 \leq i \leq I)$ are inter-coupled with each other in terms of coupling function $\eta_i(\cdot)$,*

$$\mathbb{B}_{\cdot j}^{\eta} ::= \mathbb{B}_{\cdot j}(\mathscr{E}, \mathscr{O}, \mathscr{C}, \eta) | \sum_{i=1}^{I} \eta_i(\cdot) \odot \mathbb{B}_{ij} \quad (3)$$

$$|\eta_i(\cdot)| \geq \eta_0 \quad (4)$$

where η_0 is the inter-coupling threshold, $\sum_i^I \odot$ means the subsequent behavior of $\mathbb{B}_i$ is $\mathbb{B}_{ij}$ inter-coupled with $\eta_i(\cdot)$, and so on, with nondeterminism.

Corollary 2. *If $\eta_i(\cdot) < 0$, $\mathbb{B}_{j_1}$ and $\mathbb{B}_{j_2}$ have negative inter-coupling; if $\eta_i(\cdot) > 0$, $\mathbb{B}_{j_1}$ and $\mathbb{B}_{j_2}$ are positively inter-coupled; $\eta_i(\cdot) = 0$ indicates none inter-coupling.*

Definition 4. *(Coupled Behaviors) Coupled behaviors $\mathbb{B}_c$ refer to behaviors $\mathbb{B}_{i_1 j_1}$ and $\mathbb{B}_{i_2 j_2}$ that are coupled in terms of relationships $f(\theta(\cdot), \eta(\cdot))$, where $(i_1 \neq i_2) \vee (j_1 \neq j_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max})$*

$$\mathbb{B}_c = (\mathbb{B}_{i_1 j_1}^{\theta})^{\eta} * (\mathbb{B}_{i_2 j_2}^{\theta})^{\eta} ::= \mathbb{B}_{ij}(\mathscr{E}, \mathscr{O}, \mathscr{C}, \mathscr{R}) |$$
$$\sum_{i_1, i_2=1}^{I} \sum_{j_1, j_2=1}^{J_{max}} f(\theta_{j_1 j_2}(\cdot), \eta_{i_1 i_2}(\cdot)) \odot (\mathbb{B}_{i_1 j_1}, \mathbb{B}_{i_2 j_2}) \quad (5)$$

where $f(\theta_{j_1, j_2}(\cdot), \eta_{i_1 i_2}(\cdot))$ is the coupling function denoting the corresponding relationships between $\mathbb{B}_{i_1 j_1}$ and $\mathbb{B}_{i_2 j_2}$, $\sum_{i_1, i_2=1}^{I} \sum_{j_1, j_2=1}^{J_{max}} \odot$ means the subsequent behaviors of $\mathbb{B}$ are $\mathbb{B}_{i_1 j_1}$ coupled with $f(\theta_{j_1}(\cdot), \eta_{i_1}(\cdot))$, $\mathbb{B}_{i_2 j_2}$ with $f(\theta_{j_2}(\cdot), \eta_{i_2}(\cdot))$, and so on, with nondeterminism.

Corollary 3. *Further, coupled behaviors can be represented by behavior attributes $\{[p_{ij}]_k \mid 1 \leq k \leq K\}$, then we have the corresponding **behavior adjoint matrix**:*

$$AM(\mathbb{B}_c) ::= AM(\mathbb{B}) \mid \sum_{i_1, i_2=1}^{I} \sum_{j_1, j_2=1}^{J_{max}} f(\theta_{j_1 j_2}(\cdot), \eta_{i_1 i_2}(\cdot)) \odot (\overrightarrow{\mathbb{B}}_{i_1 j_1}^{T} \overrightarrow{\mathbb{B}}_{i_2 j_2})$$
$$= AM(\mathbb{B}) \mid \sum_{i_1, i_2=1}^{I} \sum_{j_1, j_2=1}^{J_{max}} f(\theta_{j_1 j_2}(\cdot), \eta_{i_1 i_2}(\cdot)) \odot ([mp_{i_1 i_2 j_1 j_2}]_{k_1 k_2})_{K \times K} \quad (6)$$

where $\mathscr{E}_{i_1}$ and $\mathscr{E}_{i_2}$ refer to two distinct actors, $\overrightarrow{\mathbb{B}}_{i_1 j_1}^{T} = ([p_{i_1 j_1}]_{k_1 s})_{K \times 1}$ and $\overrightarrow{\mathbb{B}}_{i_2 j_2} = ([p_{i_2 j_2}]_{s k_2})_{1 \times K}$ refer to two distinct behavior vectors with corresponding behavior attributes; $[mp_{i_1 i_2 j_1 j_2}]_{k_1 k_2} = [p_{i_1 j_1}]_{k_1 1} \cdot [p_{i_2 j_2}]_{1 k_2}$ is the (k_1, k_2) element of the matrix multiplication $\overrightarrow{\mathbb{B}}_{i_1 j_1}^{T} \overrightarrow{\mathbb{B}}_{i_2 j_2}$; $\sum_{i_1, i_2=1}^{I} \sum_{j_1, j_2=1}^{J_{max}} \odot$ means the subsequent behavior adjoint matrix of $AM(\mathbb{B})$ is $\overrightarrow{\mathbb{B}}_{i_1 j_1}^{T} \overrightarrow{\mathbb{B}}_{i_2 j_2}$ coupled with $f(\theta(\cdot)_{j_1, j_2}, \eta(\cdot)_{i_1 i_2})$, and so on, with nondeterminism; and the following constraints hold: $(i_1 \neq i_2) \vee (j_1 \neq j_2) \vee (k_1 \neq k_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max}) \wedge (1 \leq k_1, k_2 \leq K)$.

In practice, coupled behaviors may be grouped in terms of different coupling relationships. In particular, coupled behaviors can be segmented into behavior sequences which are coupled with certain relationships.

Definition 5. *(Coupled Behavior Sequences) Suppose $\mathbb{B}_c$ is partitioned into M coupled behavior sequences:*

$$\Phi(\mathbb{B}_c) ::= \Phi(\mathbb{B}) | \sum_{t_1=1}^{T_1} \sum_{t_2=1}^{T_2} \cdots \sum_{t_M=1}^{T_M} f(\theta(\cdot), \eta(\cdot)) \odot \Phi_{12 \cdots M} \quad (7)$$

where $\{\Phi_m = \{\phi_{11}, \ldots, \phi_{mT_m}\} \mid 1 \leq m \leq M\}$, and T_m is the number of sequence items (behavior instances) for the m^{th}

behavior sequence. $f(\cdot)$ function indicates that the coupling relationship between two sequences Φ_i and Φ_j is $R_{ij}(\Phi_i, \Phi_j)$ which is a set (where $1 \leq i, j \leq M$). $R_{ij} \subseteq R$, R is the set of coupling relationships for all M 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*.

Similarly, coupled behavior sequences can be further represented by behavior properties associated with each sequence item.

Corollary 4. *Let SP_i represent the sequence item property set of the sequence Φ_i, its j^{th} item ϕ_{ij} is further embodied in terms of its K item properties $\phi_{ik}([p_{ij}]_1, \ldots, [p_{ij}]_K)$.*

$$AM(\Phi(\mathbb{B}_c)) ::= AM(\Phi(\mathbb{B}))| \sum_{t_{i_1}=1}^{T_{i_1}} \sum_{t_{i_2}=1}^{T_{i_2}} \sum_{t_{j_1}=1}^{T_{j_1}} \sum_{t_{j_2}=1}^{T_{j_2}} f(\theta(\cdot), \eta(\cdot)) \odot ([mp_{i_1 i_2 j_1 j_2}]_{k_1 k_2})_{K \times K} \quad (8)$$

Note that the denotations here are the same as in Corollary 3.

Theorem 1. *(Coupled Behavior Analysis (CBA)) The analysis of coupled behaviors (*CBA Problem *for short) is to build the objective function $g(\cdot)$ under the condition that behaviors are coupled with each other by coupling function $f(\cdot)$, and satisfy the following conditions.*

$$f(\cdot) ::= f(\theta(\cdot), \eta(\cdot)), \quad (9)$$

$$g(\cdot)|(f(\cdot) \geq f_0) \geq g_0 \quad (10)$$

The above behavior vector-based representation and behavior property-based quantification indicate the roadmap from the understanding to modeling of CBA problems. In the following section, we discuss the challenge of the CBA problem and the reason that existing approaches are not suitable.

3.3 Research Issues and Challenges

Solving the CBA problem is challenging because of its special data characteristics and corresponding analytical tasks. First, coupled behaviors have the following particular data characteristics:

- a data record consists of variables from multiple actors (customers), which is different to the common data management structures in which data is organized in terms of distinct customers;
- a behavior is embodied in multiple dimensional behavior properties; it likely involves heterogeneous data structures;
- typically, behaviors take place in a temporal order;
- behaviors are associated with each other by certain relationships; such couplings need to be considered in behavior data construction and analysis.

Second, the analysis of the above characterized behavior data requires specific analytical tasks, e.g.:

- behavior coupling: need to deeply understand both intra and inter-coupling relationships;
- data preparation: the existing transactional data needs to be represented and converted into coupled behavior-centered data in order to explicate behavior relationships and properties;
- analytical goals: determine coupled behavior-centered analytical objectives and evaluation performance mechanisms;
- behavior modeling: develop proper objective functions to model/learn coupled behaviors and coupling relationships.

Third, handling of the above data characteristics and tasks challenges existing behavior analysis approaches, for instance:

- behavior interaction modeling: representation and learning of complex behavior interactions is a new and challenging topic;
- frequent pattern mining: item-sets come from multiple actors, frequency-based test needs to be built on top of coupling relationships, and Apriori properties may be challenged;
- sequence analysis: because a sequence may consist of items from multiple actors, new methods are necessary for constructing and modeling sequences mixing items from actors;
- clustering: how the clustering idea can be applied to coupled behavior data; new data structures, similarity measures and methods are essential;
- classification: a label is linked to coupled behaviors; how to prepare the data and build a classifier to classify the coupled behaviors;
- post-analysis: different post-processing methods may be needed to analyze relationships between behaviors, and to generate coupled behavior patterns.

In practice, however, it is often very hard to identify both functions $\theta(\cdot)$ and $\eta(\cdot)$ and the coupling function $f(\cdot)$. For simplicity, one usually considers only one of them in the modeling. In fact, intra-coupled behaviors and inter-coupled behaviors are two special cases of coupled behaviors. Existing approaches such as sequence analysis focus mainly on intra-coupling behavior analysis ($J = 1$) and ignore coupling relationships, which makes them unsuitable here. In Section 4, we will discuss a case study of modeling coupled trading behaviors by CHMM, in which all traders' trading behaviors are reconstructed into buy, sell and trade sequences, and observation functions (ie $f(\cdot)$ function) are built for each other by including the couplings with each other. A group of trading behaviors are abnormal if they satisfy objective functions (41) and (42) (ie $g(\cdot)$ function).

3.4 Coupled Behavior Representation and Preparation

The above analysis shows the importance of converting the normal transactions into behavior-oriented data. This may be done in various ways depending on the analytical goals and methods, by considering the fact that

coupled behaviors consist of multiple traces of associated behaviors; each behavior instance corresponds to a sequence item; a behavior presents properties embodied through sequence item properties. it is certainly very complicated to convert transactional data to behavior-oriented data. Taking sequence construction as an example, we here discuss two data structures to reconstruct the trading behaviors in Section 3.1.

Data Structure 1. *Data Structure 1 is*

$$\frac{Actor - Operation}{Attributes} \tag{11}$$

Sequences in a time window $winsize$ are constructed as per

$$\{\frac{Actor_i - Operation_i}{Attributes_i} \xrightarrow{\eta} \frac{Actor_j - Operation_j}{Attributes_j}\}_{i,j=1;winsize}^{I,J} \tag{12}$$

Following this structure, the coupled trading behaviors in Tables 1 and 2 are converted into the sequences presented in Fig. 2.

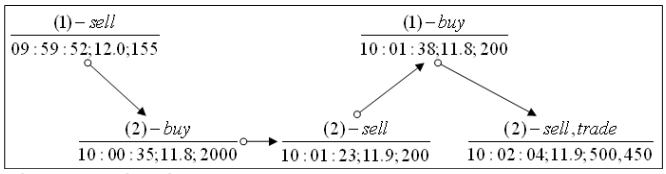

Fig. 2. Behavior sequences - Data Structure 1

Data Structure 2. *Data Structure 2 is as follows, here 'category' refers to the type of operations, which is extracted in transforming transactions into behavior sequences.*

$$Category : \frac{Actor - Operation}{Attributes} \tag{13}$$

Sequences are constructed in terms of

$$Category : \{\frac{Actor_i - Operation_i}{Attributes_i} \xrightarrow{\eta} \frac{Actor_j - Operation_j}{Attributes_j}\}_{i,j=1;winsize}^{I,J} \tag{14}$$

Fig. 3 illustrates the coupled trading behaviors in terms of Data Structure 2. This structure is used in Section 4.3, in which CHMM is built to model coupled trading behaviors.

4 CHMM-based Abnormal Coupled Behavior Analysis

This section presents a case study: a CHMM-based approach for abnormal coupled behavior analysis. Based on the discussion about anomalies in trading, the system framework and CHMM model structure for abnormal coupled behavior analysis are then introduced, which is followed by details of the CHMM-based coupled trading behavior modeling. An adaptive CHMM (ACHMM) is introduced to capture sequence changes.

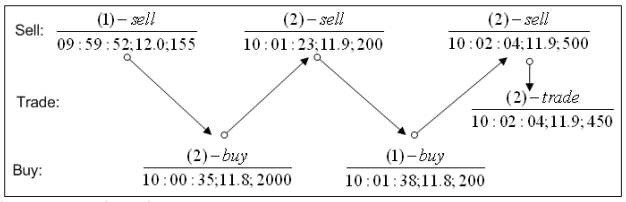

Fig. 3. Behavior sequences - Data Structure 2

4.1 Abnormal Trading Behaviors

In stock markets, trading transactions showing anomalies are expected to be detected by market surveillance rules built into surveillance systems. Consequently, alerts are generated for those transactions showing 'abnormal' trading behaviors. In practice, it is often very costly, and sometimes not even realistic, to identify genuine 'abnormal' behaviors in stock markets. This is due to the complexity whereby abnormal tradings are isolated and mixed in normal transactions. Another fact is that transactions that do not fire any alerts are likely to be 'normal' based on the available surveillance capability. With the alerts as indicators, it is easy to extract all 'normal' transactions. We are readily provided with so-called 'normal' data verified by domain analysts, which hopefully reflects the characteristics and dynamics of 'normal' business and transactions.

Given the above domain knowledge, in particular, the theory of semi-supervised learning [5] and learning from positive and unlabeled data [6], [4], it is assumed that abnormal transactions would demonstrate different characteristics and dynamics from such labeled 'normal' transactions. A model learned only from the normal data cannot fit the abnormal data very well. The output of the model will inform us whether the input data is normal. To evaluate the model, it is acceptable to business that all alerts aggregated on a target can be used as a reference to evaluate new models, and to determine whether a model can identify possible anomalies against indicators used in the surveillance system.

Below, we build and evaluate CHMM-based models based on the above assumption to detect anomalies in group-based market manipulation.

4.2 The System Framework

A CHMM based system has been built for detecting abnormal group-based manipulative trading behaviors. The CHMM captures and models a group of investors' buy, sell and trade sequences and their relationships in trading a stock, by converting order-book-level transactions into behavior sequences following Data Structure 2. The CHMM model structure is shown in Fig. 4, consisting of three chains of HMMs modeling buy-orders Φ_B, sell-orders Φ_S and trades Φ_T respectively, which are coupled with each other via interactions. The circles denote the hidden states of the three trading sequences; for instance, S_{t-1}^B 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}^S indicates the observation of the sell sequence.

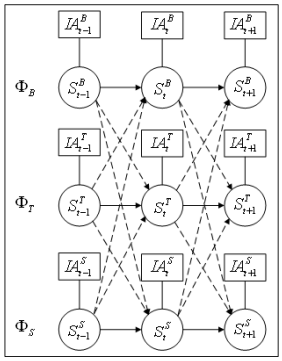

Fig. 4. Architecture of CHMM

Experienced and sophisticated market manipulators may sometimes make significant changes in manipulating a series of orders, to make their trading behaviors so volatile that they cannot be captured by predefined models. It is important to capture such significant changes by making models adaptive. For this purpose, we build an Adaptive CHMM (ACHMM) on top of the CHMM, which is tuned to adjust to significant sequence changes (see Section 4.6) amongst the three trading sequences and their coupling relationships.

The system for CHMM and ACHMM based abnormal coupled behavior detection is illustrated in Fig. 5. It consists of the following key components: Sequence Extractor, CHMM, Output Analyzer, Model Adjustor and Change Detector. Its working process is as follows. During the training period, only 'normal' trading sequences are fed into the Sequence Extractor and converted into three sequences by following Data Structure 2. Such sequences are fed into the CHMM, which learns the relationship and dynamics of the sequential data. Since the CHMM is trained on 'normal' data, it can fit the 'normal' trading data very well but not the abnormal transactions. The new trading sequences of a group are judged to be abnormal or not according to the probability of fitting in the learned CHMM models. If the fit probability is low, it means that the group's behaviors do not fit the model well, and they are treated as anomalies. This makes sense to business people because the model is specifically learned from normal cases. Further, the CHMM output is explored by the Output Analyzer, to analyze the outputs of the CHMM. If the Change Detector identifies a significant change from the benchmarks (using t-test), it notifies the Model Adjustor to immediately update (retrain) the CHMM in line with the change.

The system has the following key features:

- First, it models three trading activity sequences by catering for their interactions with each other, rather than a single sequence. In this way, it closely captures the nature of market trading, in which buys, sells and trades are coupled with each other and affect each other. However, many existing sequence analysis methods either split multiple sequences into separate ones to be modeled individually, or combine them into a single sequence, which destroys the intrinsic interactions between sequences.
- Second, CHMM-based sequence modeling considers the sequence item properties, namely price and volume of a trading action, as well as the effect of interactions between three trading sequences reflected through price and volume changes. In existing sequence analysis methods, item properties are usually ignored.
- Third, we only use 'normal' data to train the CHMM, because it is too costly to obtain negative data reflecting anomalies in business, while it is relatively easier for domain experts to justify whether a transaction looks normal. Modeling normal data and then checking the difference between the new data and the normal data can effectively avoid the need for, and cost of, acquiring negative data.
- Finally, this system can adapt automatically to the significant changes in coupled sequences. This feature is extremely valuable since real-life data is usually dynamic and can challenge models in general.

4.3 CHMM-based Coupled Behavior Modeling

The representation of coupled behavior sequences in Equation (7) and its quantification in (8) show the possibility of using CHMM to model CBA problems. A CHMM model $\lambda^{CHMM} = (X, Y, Z, \pi)$ can be built based on the following mapping relationships:

$$CBA\ problem \rightarrow CHMM\ model \tag{15}$$

$$\Phi(\mathbb{B}_c)|category \rightarrow X \tag{16}$$

$$M(\Phi(\mathbb{B}_c))|\phi_{ik}([p_{ij}]_1, \ldots, [p_{ij}]_K) \rightarrow Y \tag{17}$$

$$f(\theta(\cdot), \eta(\cdot)) \rightarrow Z \tag{18}$$

$$Initial\ distribution\ of\ \Phi(\mathbb{B}_c)|category \rightarrow \pi \tag{19}$$

On the other hand, the conversion of transactional data into behavior sequences following Data Structure 2 makes the data suitable for CHMM-based modeling. In order to build CHMM for three coupled trading sequences as described in Data Structure 2, we define HMMs and CHMM as follows.

We 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. 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

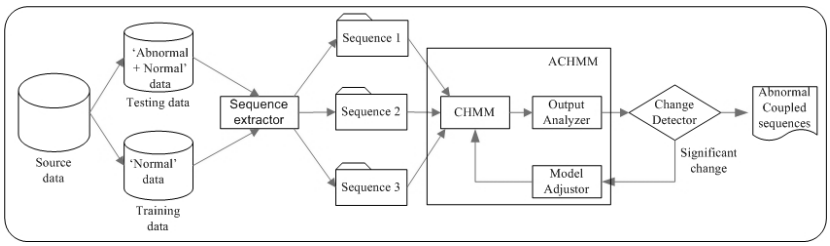

Fig. 5. Framework of abnormal coupled behavior detection

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 [15], a model λ^{HMM} is trained by the following re-estimated formulas with a set of observation sequences:

$$\bar{x}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) x_{ij} y_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(j)}, 1 \leq i, j \leq N \quad (20)$$

$$\bar{y}_j(k) = \frac{\sum_{t=1, o_t=O_k}^{T} \alpha_t(j) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j) \beta_t(j)}, 1 \leq j \leq N \quad (21)$$

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.

$$\bar{\pi}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{N} \alpha_1(j)\beta_1(j)}, 1 \leq i \leq N \quad (22)$$

After the model $\lambda^{HMM} = (X, Y, \pi)$ is trained, the probability of an observation sequence $Q = \{q_1, q_2, \cdots, q_T\}$ is computed as follows based on caching calculation [15]:

$$Pr(Q|\lambda^{HMM}) = \sum_{i=1}^{N} \alpha_T(i) \quad (23)$$

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 $\lambda^{CHMM} = (X, Y, Z, \pi)$.

The $forward\text{-}backward\ analysis$ is used to evaluate the observation and train three chains of CHMM. To reduce the computational complexity ($O(TN^6)$), we use the approximate inference algorithms - N-heads dynamic programming, which relaxes the assumption that every transition must be visited, and achieves $O(T(3N)^2)$.

4.4 Hidden States

Following the mapping relationship (16), the categories of behaviors correspond to the states in CHMM. Based on domain knowledge, we define the hidden states in terms of an investor's belief, desire and intention (BDI), which are embodied through trading actions and their corresponding behavior characteristics.

- For *HMM-B*, its hidden state S^B 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.

$$S^B = \{Positive\ Buy, Neutral\ Buy, Negative\ Buy\} \quad (24)$$

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

$$S^S = \{Positive\ Sell, Neutral\ Sell, Negative\ Sell\} \quad (25)$$

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

$$S^T = \{Market\ Up, Market\ Down\} \quad (26)$$

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.

4.5 Observation Sequences

In the mapping relationship (17), behavior properties are the base for building observation sequences. As we can see, there may be quite a few items in a behavior sequence. The behaviors in a trading day are many. The construction of observation sequences in CHMM is as follows. We partition the data in terms of time windows. All behavior instances in each time window are aggregated into one representative element. All such elements in a day form the CHMM observation sequences. Each element is a vector, with its properties to be used for calculating the observation probability. In order to construct observation sequences for trading sequences, we define two concepts: $activity$ (A) and $interval\ activity$ (IA). They 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 embodied through factors such as trading prices, volumes and times in stock markets.

Definition 6. *Activity (A) represents an actor's individual behavior.*

In capital markets, A is a trading behavior, which consists of an atomic trading operation ($a = \{buy \mid sell \mid trade\}$) taken by an investor, associated with behavior properties price (p) and volume (v) at time t. These variables a, p, v and t 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 operations in capital markets: buy-order, sell-order or trade at time t_i, its associated properties $p(t_i)$ 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; $v(t_i) = \{buy\ volume|sell\ volume|trade\ volume\}$ is the trade volume of $a(t_i)$ at t_i. The CHMM observations reflect the cumulative effect over all states at the previous step, which corresponds to all relevant actors. For this, we define *Interval Activity*.

Definition 7. *Interval Activity (IA) represents a collective behavior embodied through the collective properties of a behavior sequence.*

For trading behaviors, $IA = (\mathcal{A}, \bar{p}, \bar{v}, f)$, $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, $A_i(a)$ consists of a set of trading behaviors taking place in window l with size $winsize$. The variables $\bar{p}$, $\bar{v}$ and f capture the characteristics, distributions and accumulative activities of an investor, or a group of investors:

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i}{f} \tag{27}$$

$$\bar{v} = \frac{\sum_{i=1}^{n} v_i}{f} \tag{28}$$

$$f = |\mathcal{A}| = n \tag{29}$$

n is the number of activities in the window l.

To map IAs to the observation symbols of CHMM to obtain the probability of observations, we quantize $\bar{p}, \bar{v}$ and f based on the k-means clustering algorithm. This identifies the most representative activity, and generates the observation variable $IA(p^{'}, v^{'}, f^{'})$.

$$IA(\mathcal{A}, \bar{p}, \bar{v}, f) \xrightarrow{\text{quantization}} IA^{'}(p^{'}, v^{'}, f^{'}) \tag{30}$$

$IA^{'}(p^{'}, v^{'}, f^{'})$ 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, the discrete values of $p^{'}$ are given by:

$$p^{'} = argmin_i|\bar{p} - \theta_i^p| \tag{31}$$

Similarly, we quantize v and f as follows:

$$v^{'} = argmin_i|\bar{v} - \theta_i^v| \tag{32}$$

$$f^{'} = argmin_i|f - \theta_i^f| \tag{33}$$

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

4.6 Adaptive CHMM for Detecting Behavior Changes

To make the CHMM adaptive to significant changes in trading sequences, we improve the CHMM by adding an automatically adaptive mechanism to form an Adaptive CHMM (ACHMM). 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 their 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 properly capture the corresponding changes in trading activities and needs to be updated. As shown in Fig. 6, 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

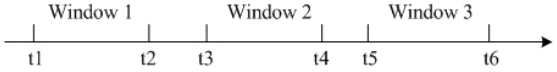

Fig. 6. Update Point of ACHMM

on a sliding window 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 CHMM parameters x, y, z and π on the most recent dataset rather than the whole training dataset. This strategy is consistent with the Markov assumption, i.e. the current state is dependent only on the previous state, and enables us to avoid the great expense of model retraining.

$$x_{ij}^{update} = (1-w)x_{ij}^{old} + w * x_{ij}^{new} \tag{34}$$

$$y_{ij}^{update} = (1-w)y_{ij}^{old} + w * y_{ij}^{new} \tag{35}$$

$$z_{ij^{'}}^{update} = (1-w)z_{ij^{'}}^{old} + w * z_{ij^{'}}^{new} \tag{36}$$

$$\pi_{i}^{update} = (1-w)\pi_{i}^{old} + w * \pi_{i}^{new} \tag{37}$$

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.

4.7 The Algorithms

The key algorithms for CHMM/ACHMM based abnormal trading behavior analysis consist of two features: (a) extracting and splitting trading sequences from the trading transactions to construct the observation sequences by following Data Structure 2, and (b) detecting abnormal trading activity sequences by feeding the three trading sequences into the CHMM/ACHMM models. *Algorithm 1* extracts trading activity sequences, which form the observation sequences of a CHMM.

Algorithm 1 Constructing observation sequences

Step 1: Segment a 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 sequence 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

$$IA^{type} = IA_1^{'type}, IA_2^{'type}, \cdots, IA_L^{'type} \tag{38}$$

where $type \in \{buy, sell, trade\}$. IA^{buy}, IA^{sell} and IA^{trade} are the observation sequences in the day.
Step 5: Repeat Steps 1-4 for each trading day

Algorithm 2 shows the procedure for constructing an abnormal trading activity sequence. It calculates 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.

Algorithm 2 CHMM/ACHMM for detecting abnormal coupled 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 CHMM/ACHMM models on the training sequences;
Step 3: Compute the mean (μ) and standard deviation (σ) of probability of training sequences according to the following formulas:

$$\mu = \frac{\sum_{i=1}^{K} Pr(Seq_i|CHMM(ACHMM))}{K} \tag{39}$$

$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K} Pr(Seq_i|CHMM(ACHMM))) - \mu} \tag{40}$$

where K is the total number of training sequences, mean μ represents the centroid of model CHMM/ACHMM, and the standard deviation σ represents the radius of models CHMM/ACHMM.
Step 4: For each test sequence $Seq_i^{'}$, calculate its distance D_i to the centroid of model by

$$D_i = \frac{\mu - Pr(Seq_i^{'}|\mathcal{M})}{\sigma} \tag{41}$$

Consequently, $Seq_i^{'}$ is an exceptional pattern, if it satisfies:

$$D_i > \psi_0 \tag{42}$$

where ψ_0 is a given threshold.

5 Experiments and Evaluation

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

5.1 Experimental Data

The experimental data is from an Asian stock market. Tables 3 and 4 illustrate an excerpt of stock order-book data used in this paper. It covers 388 valid trading days from 1 June 2004 to 31 December 2005. Following Data Structure 2 and Algorithm 1, 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:

$$\{Buy : ((2), 11.8, 2000), ((3), 11.8, 150), ((1), 11.8, 200), ((4), 11.9, 200), ((5), 11.9, 250)\}$$
$$\{Sell : ((1), 12.0, 155), ((2), 11.9, 200), ((2), 11.9, 500)\}$$
$$\{Trade : ((2), 11.9, 450), ((4), 11.9, 200), ((5), 11.9, 250)\}$$

The data is partitioned into two sets by referring to the domain expert opinion. The training data set consists of transactions collected from 1 June 2004 to 31 December 2004, by filtering those transactions associated with the identified alerts. We treat it as 'normal' data. HMM-based models are trained on such labeled normal data to capture the characteristics and dynamics in so-called 'normal' trading. The transactions acquired from 1 January 2005 to 31 December 2005 are entered into the test set. Those transactions that triggered alerts are retained in the test set, making the test data a mixture of both normal and abnormal behaviors. The trained models are deployed on the mixed test data to detect abnormal behaviors. Alerts fired on those possibly problematic transactions are treated as a rough benchmark for us to evaluate the CHMM and ACHMM against the detection performance of existing market surveillance rules. This is practical and acceptable to business people, especially when it is very costly to obtain labeled data.

5.2 Benchmark Models

In order to evaluate the performance of CHMM and ACHMM, we build another four HMM models as the 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 between 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.

TABLE 3
An excerpt of stock orderbook data

$Account\_Id$	$Security\_Code$	$Order\_No$	$Order\_Date$	$Order\_Time$	$Trade\_Direction$	$Order\_Price$	$Order\_Volume$
B12894940	61234	201293233	2004/6/24	100435	Buy	10.35	10000
A93940201	72392	328193944	2004/6/29	144523	Sell	28.50	300000

TABLE 4
An excerpt of stock orderbook data (continued)

$Order\_Balance$	$Trade\_Balance$	$Start\_Trade\_Time$	$Withdraw\_Time$	End_Trade_Time	$Withdraw\_Volume$	Alert
35000	20000	100512	0	113648	0	0
550000	600000	144529	151741	145838	150000	1

- *ACHMM*: a CHMM model on trade, buy-order and sell-order sequences considering the coupled relationships and with an 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.

5.3 Technical Performance

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

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (43)$$

$$Precision = \frac{TP}{TP+FP} \quad (44)$$

$$Recall = \frac{TP}{TP+FN} \quad (45)$$

$$Specificity = \frac{TN}{FP+TN} \quad (46)$$

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.

We test the six models on the data by setting various window sizes ($winsize$). Figs 7, 8, 9 and 10 show their technical performance, where the horizontal axis ($P\text{-}Num$) 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 benchmarks at any window size ($winsize$), while ACHMM performs the best.

For instance, in Fig 8, 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 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 is generally significantly better than CHMM. When the number of detected sequences increases, more false positive (FP, abnormal alerts) may be captured, which correspondingly

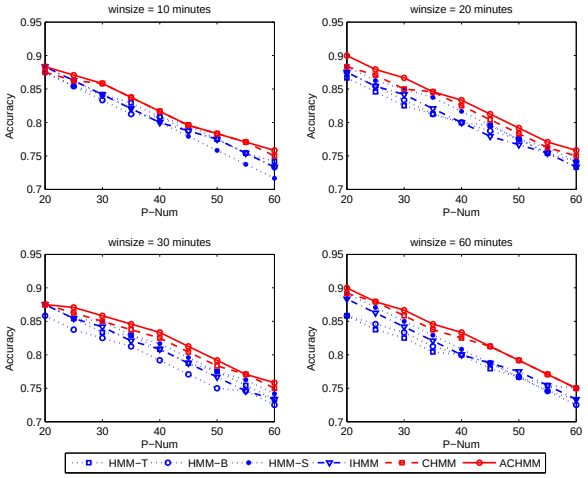

Fig. 7. Accuracy of Six Models

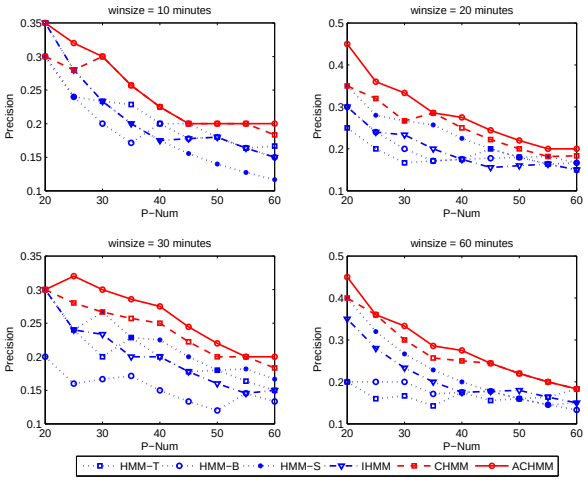

Fig. 8. Precision of Six Models

reduces aspects of the model's performance, such as Precision. However, ACHMM retains its advantage over all others (see Figs 5 and 6, $P-Num = 60$). In particular, the recall increases with P-Num rising, which shows the HMMs trained on 'normal' data can contribute to a lower false negative (FN, i.e. false 'normal'). 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.

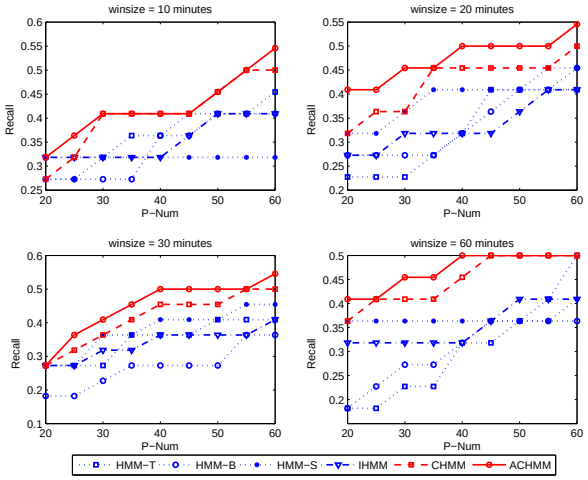

Fig. 9. Recall of Six Models

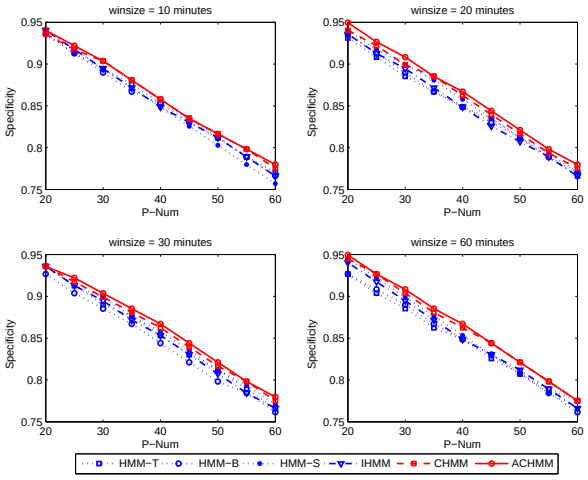

Fig. 10. Specificity of Six Models

5.4 Business Performance

We further evaluate CHMM and ACHMM against the benchmark models in terms of the business performance of trading on the detected abnormal coupled trading behaviors as shown below.

$$\begin{cases} 84993, 9:25:05, 36.00 \\ A32920236, 10:23:20, buy, 36.50, 990,000, withdraw \\ A32920236, 10:23:22, buy, 36.55, 80,000 \\ A32920974, 10:23:57, buy, 37.00, 500,000 \\ A67923702, 10:25:39, buy, 36.55, 100,000 \\ A09934523, 10:27:15, buy, 37.05, 200,000 \\ A18949234, 10:30:28, buy, 37.05, 200,000, tradepartially \\ A49920093, 10:31:34, buy, 37.05, 500,000, tradepartially \\ A29984884, 10:32:52, buy, 37.05, 800,000, tradepartially \\ \ldots \\ A40299387, 10:39:42, sell, 37.05, 1,000,000, tradepartially \\ \ldots \\ 84993, 10:54:47, 35.40 \end{cases} \quad (47)$$

Two business metrics widely used in capital markets are introduced for evaluating the business performance. They are *return* and *abnormal return* [3]. *Return* refers to the gain or loss for a single security or portfolio over a specific period, which is calculated by

$$Return = ln\frac{p_t}{p_{t-1}} \quad (48)$$

where p_t and p_{t-1} are the trade prices at time t and $t-1$, respectively. *Abnormal return* is defined as the difference between the *actual return* of a single security or portfolio and the *expected return* over a given time period. The expected return is the estimated return based on an asset pricing model, using a long-term historical average, or multiple valuation. The formula to compute *abnormal return* is as follows:

$$Abnormal\ Return = Return - (\gamma + \xi Return^{market}) \quad (49)$$

where $Return^{market}$ is the observed return for the market index, γ and ξ are the estimated parameters using previous return observations. Empirically, the trading days with exceptional patterns are more likely to incur a higher return and abnormal return than those without exceptional trading. This is consistent with the findings in Figs 11 and 12, which show the *return* and *abnormal return* in trading those abnormal sequences detected by CHMM and ACHMM outperform the other four models with different $winsize$. As shown in Fig 12 ($winsize = 20, P-Num = 20$), trading abnormal sequences detected by ACHMM can lead to over 50% additional abnormal return over trading from HMM-T. Therefore, Figs 11 and 12 show that CHMM, and ACHMM in particular, can better detect abnormal trading behaviors with a higher business impact than other models.

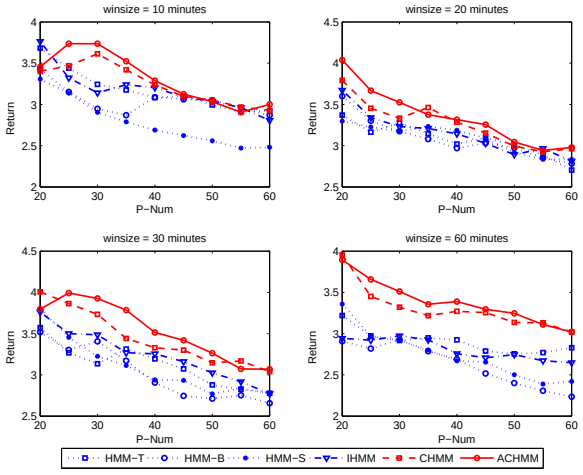

Fig. 11. Return of Six Models

5.5 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 5, 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.

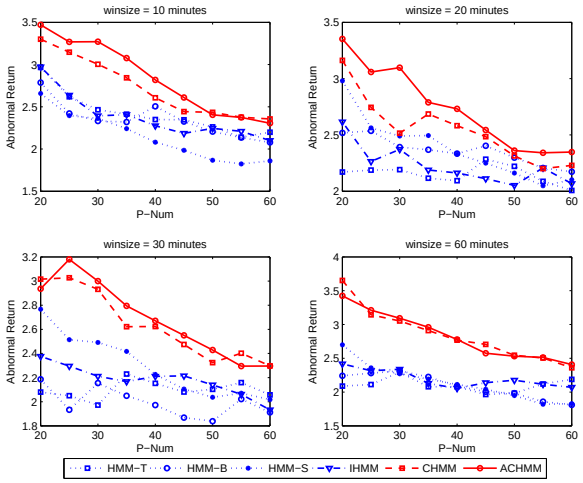

Fig. 12. Abnormal Return of Six Models

TABLE 5
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

6 Discussions about Behavior Interactions

6.1 Behavior Coupling Relationships

Coupled behaviors may be caused by different factors. Some are unconditional, while others are conditional. *Unconditional coupling* often appears in loosely coupled groups, and no strict engagement exists. Parties often take an ad hoc, on-demand attitude to collaborating with others. A typical example of unconditional coupling is frequent behavioral itemsets identified by association rule mining. They are associated in a light coupling relationship. On the other hand, *conditional coupling* exists in those groups with clear and strong engagement, in particular, in terms of certain engagement rules, through which parties in each group know and fulfill pre-arranged roles. For example, two behaviors take place in a parallel relationship. For both unconditional and conditional coupled groups, behaviors are coupled because of either a cooperative or competitive nature. *Cooperative behaviors* are formed towards the same objectives and benefits, while *competitive behaviors* are associated with opposite objectives or a conflict of interest. An example of cooperative behaviors would be a group of investors collaborating in trading towards a mutual profit-making goal, whereas competitive behaviors occur as the result of a manipulator's trading behaviors and a regulator's anti-manipulation interventions in the market.

The *coupling relationships* in group-based behavior present in different forms. We list a few typical ones that exist in various correlations and collaborations from the perspective of logic, interaction and combination.

- Inter-leaving coupling: e.g., $\{a_1, a_2\}$, parties in a group conduct respective behaviors without any constraints; e.g., two traders take different positions in trading a security without knowing each other's intention.
- Parallel coupling: e.g., $\{a_1 \parallel a_2\}$, parties in the group conduct their behaviors a_1 and a_2 in parallel; e.g., two manipulators placing large buy quotes on two target securities respectively at the same time.
- Serial coupling: e.g., $\{a_1; a_2\}$, parties conduct their behaviors in a serial order, a_2 follows a_1; e.g., two traders inter-weave their buys and sells on a security by following a pre-arranged manipulation strategy.
- Causal coupling: e.g., $\{a_1 \Rightarrow a_2\}$, parties conduct their behaviors in a causal order, a_1 causes a_2; consequent coupling is a strong serial coupling with inter-weaving relationships among the parties' behaviors.
- Exclusive coupling: e.g., $\{a_1 \nparallel a_2\}$, parties are associated with each other with different behaviors occurring mutually and exclusively; e.g., one manipulator keeps buying security A, while the other keeps selling security B.
- Negative coupling: e.g., $\{a_1 = \bar{a_2}\}$, if a party's behavior a_1 appears, the same behavior by another should not take place; e.g., a manipulator puts a large buy order, so no other group member will place a big buy at the same time or later.
- Hierarchical coupling: e.g., $\{a_1; (a_2 \parallel a_3)\}$, parties are associated with each other in a hierarchical structure, two parallel behaviors a_2 and a_3 follow a_1; e.g., a manipulator initiates a mark-the-close strategy, and several others follow it up by iteratively placing and withdrawing buy orders.
- Hybrid coupling: parties are associated with each other in a complicated structure which consists of multiple different coupling relationships.

From a reaction perspective, behaviors may be coupled in an asynchronous or synchronous manner.

- Asynchronous coupling: behaviors from two relevant parties take place asynchronously; e.g., manipulators in two markets with a time difference take respective actions according to a pre-agreed strategy.
- Synchronous coupling: behaviors from two parties take place synchronously; e.g., one manipulator waits for the indicator triggered by another before any action is taken.

6.2 Behavior Interaction Modes

Behaviors coupled in terms of the above coupling relationships may interact in different modes, e.g.:

- Peer-to-peer mode: two behavior instances a_1 and a_2 are associated with each other in an equal posi-

tion; e.g., two manipulators from different brokerage firms undertake respective manipulative strategies.
- Master-slave mode: a behavior instance a_2 is triggered or coordinated by another a_1; e.g., a manipulator initiates a manipulation and drives a group of traders to take follow-up actions.
- Underlying-derivative mode: e.g., $\{a_1\} - \{a_1; a_2\}$ a behavior instance follows another in either a prefix or postfix manner in terms of a coupling relationship; e.g., activities of a manipulation initiator and his/her followers' activities.
- Contrast mode: $\{a_1\} - \{a_2 = \bar{a_1}\}$, two behavior instances take place in opposite positions; e.g., a manipulator places buys while another places sells.

These interaction modes are helpful for understanding the behavior interaction between behavior sequences once coupling relationships are determined. For instance, through replaying the trading actions of identified abnormal tradings as illustrated in (47), the group-based manipulation is as follows: (1) Stage 1: In a certain market movement situation, a trader (initiator) lodged a large market buy order; several group members (followers) then started to lodge incremental buy quotes until the security price showed clear movement. (2) Stage 2: After the markable price movement took place, all group members stop trading. (3) Stage 3: When the price movement stopped after several rising limits, the initiator placed a large sell, followed by sells from the rest of the group members. The behaviors of the initiator and the followers in Stage 1 interacted in the master-slave and underlying-derivative modes. The occurring buys at Stage 1 and the non-occurring buy behaviors (negative sequences [27]) at Stage 2 form an underlying-derivative relationship. The large buy in Stage 1 and the following large sell in Stage 3 are also in a strong underlying-derivative coupling.

6.3 Prospects

With the above discussions about coupling relationships and behavior interaction modes, it is very likely that many novel types of coupled patterns will be discovered by developing corresponding approaches and tools. More research is encouraged to mine for interesting behavior patterns coupled in various forms.

One possible way is *combined mining* [12] to identify *combined patterns* with components from multiple aspects, for example, behaviors from multiple actors. In [12], general frameworks and algorithms are discussed for identifying combined patterns such as pair patterns and cluster patterns. For instance, we applied it to identify combined debt arrangement-repayment patterns with behavior components from different actors in the social security area:

$$\{(a_1 \parallel a_3) \Rightarrow (b_2; b_6)\}. \tag{50}$$

It reflects the following scenario: A debtor (government customer) received overpayments from the government; a government debt management officer made serial arrangements b_2 and b_6 (they form relationship $\{b_2; b_6\}$) for the customer to pay back the debt. However, the customer actually took two parallel repayment actions a_1 and a_3 (they form relationship $\{a_1 \parallel a_3\}$) to pay it back. In another example [13], we expand frequent pattern mining to identify *impact-targeted activity patterns*. For instance, *underlying-derivative activity sequences* are mined in terms of intra-coupling relationships:

$$\{(a_1; a_2 \Rightarrow I), (a_1; a_2; a_5 \Rightarrow \bar{I})\} \tag{51}$$

where actions $a_1; a_2$ lead to positive impact I, while the addition of a_5 converts the impact from positive to negative ($\bar{I}$). Similary, in *contrast activity patterns*

$$\{(a_1; a_7 \Rightarrow I), (a_1; \bar{a_7} \Rightarrow \bar{I})\} \tag{52}$$

actions $a_1; a_7$ are likely associated with positive impact, while the non-appearance of a_7 (ie. $\bar{a_7}$) following a_1 has a high probability of generating a negative impact. We will further explore combined mining of coupled behaviors.

In addition, besides the CHMM based approach for analyzing group-based behavior, it is worthwhile to try other techniques, for instance multi-variate time series based analysis [19], [20], [21], advanced Bayesian networks, and agent mining-based methods [22]. Techniques in multi-agent coalition formation, for example, may also be helpful for understanding group-based behavior formation and evolution.

7 Conclusion

This paper discusses a challenging issue - coupled behavior analysis. With the illustration of relevant applications, a formal definition and discussions about the emerging challenges are given. As a case study, we propose a coupled Hidden Markov Model-based approach to detect anomalies in group-based trading manipulation. The proposed CHMM and an adaptive CHMM model multiple 'normal' coupled trading sequences for detecting abnormal group based manipulative trading behaviors. They consider interactions between sequences, sequence item properties, and significant changes in the coupled sequences. A system has been developed and intensively tested on real-life order-book-level data. The results have shown that the CHMMs outperform a single HMM only modeling any single trading sequence or an integrative HMM combining multiple single-sequences, but ignoring interactions between them, in terms of both technical and business performance in detecting trading anomalies. Finally, we discuss the coupled behavior interactions, which are widely seen in behavior-oriented applications but haven't been addressed. Coupled behavior analysis brings about great challenges and opportunities in many areas such as representing, reasoning and learning behavior coupling and interactions, and mining behavior interaction patterns.

8 ACKNOWLEDGEMENTS

This work is sponsored in part by Australian Research Council Discovery Grants (DP1096218, DP0988016, DP0773412) and ARC Linkage Grant (LP100200774, LP0989721, LP0775041), as well as American NSF through grants IIS-0905215 and IIS-0914934.

REFERENCES

[1] A. Karwath and N. Landwehr. Boosting relational sequence alignments. ICDM08, pages 857-862, 2008.
[2] L. Cao, Y. Ou, P. Yu and G. Wei. Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors. KDD10, 85-93, 2010.
[3] 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.
[4] X. Li and B. Liu, Learning to classify texts using positive and unlabeled data, in IJCAI03, 587-592, 2003.
[5] O. Chapelle, B. Scholkopf, and A. Zien (Eds.). *Semi-Supervised Learning*, Cambridge, MA: MIT Press, 2006.
[6] H. Yu, J. Han, and K. C. Chang. Pebl: positive example based learning for web page classification using svm, in KDD02, 239248, 2002.
[7] L. Cao and P. Yu. Behavior informatics: An informatics perspective for behavior studies. *The Intelligent Informatics Bulletin*, 10(1):6-11, 2009.
[8] D. Kifer and J. Gehrke. Detecting change in data streams. VLDB04, 180-191, 2004.
[9] R. Gwadera and F. Crestani. Discovering significant patterns in multi-attribute sequences. ICDM08, 827-832, 2008.
[10] J. Ayres, J. Flannick and T. Yiu. Sequential pattern mining using a bitmap representation. KDD02, 429-435, 2002.
[11] J. Pei, J. Han and M. C. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. ICDE01, 215-226, 2001.
[12] L. Cao, H. Zhang, Y. Zhao, D. Luo and C. Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data. *IEEE Trans. SMC Part B*, http://ieeexplore.ieee.org /stamp/stamp.jsp?arnumber=05621927, 2011.
[13] 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.
[14] 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.
[15] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 275–286, 1989.
[16] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. EDT96, 3-17, 1996.
[17] X. Song, M. Wu and S. Ranka. Statistical change detection for multi-dimensional data. KDD07, 667-676, 2007.
[18] M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42:31–60, 2001.
[19] A. Tucker, S. Swift and X. Liu. Variable grouping in multivariate time series via correlation. *IEEE Trans. SMC B*, 31(2): 235245, 2001.
[20] H. Yoon, K. Yang and C. Shahabi. Feature subset selection and feature ranking for multivariate time series. *IEEE Trans. on Knowledge and Data Engineering*, 17(9): 11861198, 2005.
[21] I Batal, L Sacchi, R Bellazzi and M Hauskrecht. Multivariate time series classification with temporal abstractions. *Int J Artif Intell Tools*, 22:344-349, 2009.
[22] L. Cao, V Gorodetsky and P. A. Mitkas. Agent mining: The synergy of agents and data mining. *IEEE Intelligent Systems*, 24(3): 64-72, 2009.
[23] S. Chandrakala and C. Chandra Sekhar A density based method for multivariate time series clustering in kernel feature space, *IJCNN*, pages 1885-1889, 2008.
[24] Keogh, J. Lin and W. Truppel. Clustering of time series subsequences is meaningless: Implications for past and future research. ICDM03, 115-122, 2003.
[25] A. Singhal and D. Seborg. Clustering of multivariate time-series data. *Proceedings of the American Control Conference*, 3931-3936, 2002.
[26] G. Tatavarty, R. Bhatnagar and B. Young. Discovery of temporal dependencies between frequent patterns in multivariate time series. CIDM07, 688-696, 2007.
[27] Y. Zhao, H. Zhang, L. Cao, C. Zhang and Hans Bohlscheid. Efficient mining of event-oriented negative sequential rules. WI08, 336-342, 2008.
[28] Y.J. Park and K.N. Chang. Individual and group behavior-based customer profile model for personalized product recommendation, *Expert Systems with Applications*, 36(2): 1932-1939, 2009.
[29] L.B. Cao. In-depth behavior understanding and use: the behavior informatics approach, *Information Science*, 180(17); 3067-3085, 2010.
[30] T Hogg and G Szabo. Diversity of online community activities, HT08, 227-228, 2008.
[31] H. Cao, N. Mamoulis, and D. W. Cheung. Discovery of periodic patterns in spatiotemporal sequences. *IEEE Trans. Knowl. Data Eng.*, 19(4): 453 - 467, 2007.
[32] D.E. Hinkle, W. Wiersma and S.G. Jurs. *Applied Statistics for the Behavioral Sciences: Applying Statistical Concepts* (5th edn), Wadsworth Publishing, 2002.
[33] W.D. Pierce and C.D. Cheney *Behavior Analysis and Learning*, Psychology Press, 2008.
[34] G.L. Zacharias and J. MacMillan. (Eds.) *Behavioral Modeling and Simulation: From Individuals to Societies*, National Academies Press, 2008.
[35] D.R. Ilgen and C.L. Hulin. (Eds.) *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline*, American Psychological Association, 2000.
[36] Y.S. Xu and K.C. Lee. *Human Behavior Learning and Transfer*, CRC Press, 2005.
[37] H. Liu, J. Salerno and M.J. Young. (Eds.) *Social Computing, Behavioral Modeling, and Prediction*, Springer, 2008.
[38] L. Getoor and B. Taskar (Eds.) *Introduction to Statistical Relational Learning*, MIT Press, 2007.

Longbing Cao Dr. Longbing Cao is a Professor at the University of Technology Sydney, and the Data Mining Research Leader of the Australian Capital Markets Cooperative Research Centre. He got one PhD in Intelligent Sciences and another in Computing Sciences. His research interests include data mining and machine learning and their applications, behavior informatics, multi-agent technology, open complex intelligent systems, and agent mining.

Yuming Ou Dr. Yuming Ou is a research fellow in Data Sciences & Knowledge Discovery Research Lab, Faculty of Engineering & IT, University of Technology, Sydney, Australia. His research interests include sequential pattern mining, abnormal trading pattern recognition and pattern-based market surveillance system design. He has over 20 publications, including 2 book chapters and 3 journal articles.

Philip S. Yu Dr. Philip S. Yu is a Professor at the University of Illinois at Chicago and also holds the Wexler Chair in Information Technology. His research interests include data mining, privacy preserving data publishing, data stream, Internet applications and technologies, and database systems. Dr. Yu has published more than 590 papers in refereed journals and conferences. He holds or has applied for more than 300 US patents. Dr. Yu is a Fellow of the ACM and the IEEE.