Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulations

Yin Song*
Faculty of Engineering and IT, University of Technology, Sydney

Longbing Cao
Faculty of Engineering and IT, University of Technology, Sydney
longbing.cao@uts.edu.au

Xindong Wu†
Hefei University of Technology, China;
University of Vermont, Burlington

Gang Wei
Department of Research Centre, Shanghai Stock Exchange
weigang@sse.com.cn

Wu Ye
Department of Research Centre, Shanghai Stock Exchange
yewu@sse.com.cn

Wei Ding
Department of Computer Science, University of Massachusetts Boston
ding@cs.umb.edu

ABSTRACT

In stock markets, an emerging challenge for surveillance is that a group of hidden manipulators collaborate with each other to manipulate the price movement of securities. Recently, the coupled hidden Markov model (CHMM)-based *coupled behavior analysis* (CBA) has been proposed to consider the coupling relationships in the above group-based behaviors for manipulation detection. From the modeling perspective, however, this requires overall aggregation of the behavioral data to cater for the CHMM modeling, which does not differentiate the coupling relationships presented in different forms within the aggregated behaviors and degrade the capability for further anomaly detection. Thus, this paper suggests a general CBA framework for detecting *group-based market manipulation* by capturing more comprehensive couplings and proposes two variant implementations, which are *hybrid coupling* (HC)-based and *hierarchical grouping* (HG)-based respectively. The proposed framework consists of three stages. The first stage, *qualitative* analysis, generates possible qualitative coupling relationships between behaviors with or without domain knowledge. In the second stage, *quantitative* representation of coupled behaviors is learned via proper methods. For the third stage, anomaly detection algorithms are proposed to cater for different application scenarios. Experimental results on data from a major Asian stock market show that the proposed framework outperforms the CHMM-based analysis in terms of detecting abnormal collaborative market manipulations. Additionally, the two different implementations are compared with their effectiveness for different application scenarios.

*Email: yin.song@student.uts.edu.au.
†Corresponding author (xwu@uvm.edu).

KDD'12, August 12–16, 2012, Beijing, China.

Categories and Subject Descriptors

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

General Terms

Algorithms, Economics, Security

Keywords

Coupled behavior analysis, anomaly detection, market manipulation, coupled hidden Markov model, hierarchical clustering, relational learning

1. INTRODUCTION

For many decades, human behavior analysis has been extensively investigated in many fields, such as social and behavioral sciences [5, 13] and computer science [10]. For simplicity, 'behavior' in this paper is used as a synonym of 'human behavior', which refers to an action from a human and usually interacts with behaviors of his/her own and of other actors [3]. 'Coupled' in this paper refers to actors and their behaviors having certain relationships which are not independent. The interactions within an actor are referred to as intra-coupled relationships (interchangeable with 'intra-couplings') while inter-coupled relationships (interchangeable with 'inter-couplings') are between behaviors of different actors [3]. Taking the couplings into account is critical for a deep understanding of group behaviors in many real-life scenarios. For instance, in stock markets, investors' trading behaviors are not isolated but affected by each other, which is caused by the supply-and-demand nature of the markets. Most of existing research efforts on behavior studies, however, focus mainly on individual behaviors [4]. A comprehensive analysis of intra and inter-coupled relationships is beyond current individual behavior analysis techniques, to the best of our knowledge. In addition, the intra- and inter- couplings between behaviors are

usually implicit or hidden, making it even more challenging to define, model and analyze the coupling relationships among behaviors in a group.

One useful application of considering these intra and inter-couplings is to detect abnormal coupled behaviors in stock markets [4]. As mentioned before, the behaviors of investors are inherently coupled with each other. Meanwhile, on some trading days, some investors (termed as 'manipulators') may intentionally arrange trading behaviors for the purpose of exceptionally high profit, which is not allowed by the corresponding regulations and could bring great losses to other investors. Any financial market regulators are keen to effectively detect these coupled trading behaviors and discover such manipulations. As initial attempts for this purpose, the CHMM-based CBA [4] suggests a method to implicitly represent the couplings in statistical models, such as CHMM [2]. It regards all the actors as a whole group, aggregates their behaviors for each time interval and analyzes the aggregated behaviors based on a CHMM. This is feasible to some extent but has some limitations: e.g., segmentation and aggregation of the behaviors may lose important coupling information within these aggregated behaviors, which may further deteriorate the performance of anomaly detection. In other words, capturing richer couplings for coupled behavior analysis may enhance the capability of detecting manipulations. Motivated by this, this paper proposes a general CBA framework to capture richer coupling relationships between behaviors, which is further used to detect anomalies in group-based behaviors. By capturing more comprehensive couplings between the behaviors, better anomaly detection performance is expected.

The main contributions of this paper are summarized as follows.

- We extend the group-based CBA to a general framework to cater for more flexible and comprehensive analysis of coupled behaviors, which is expected to have better capability to detect abnormal behaviors. To achieve this, two stages: *qualitative* analysis (for reducing the possible coupled relationships space) and *quantitative* analysis (for proper numerical modeling of the couplings) are proposed to efficiently model the underlying rich coupling relationships in the behaviors.

- Two variant implementation approaches are proposed to model more comprehensive couplings. One is data-driven, which learns the coupled relationships directly from the data and the other is domain-driven, which integrates some domain knowledge for learning the couplings. These two approaches are useful for different application scenarios when domain knowledge is unavailable or available.

- The proposed framework has been tested on a real-world data set from a major Asian stock market to detect collaborative manipulations in stock markets, covering around 550,000 tick-based transactions on 388 valid trading days. We test different technical significance of the identified suspicious manipulations against the benchmark of miscellaneous alerts fired on the real-time transactions. The results show the advantages of our proposed framework compared to the previous CHMM-based framework. More specifically, the two proposed approaches are compared to reach the conclusion that they are superior to the CHMM-based CBA framework in different application settings.

The remainder of this paper is organized as follows. Section 2 describes the proposed general CBA framework. Then the hybrid coupling-based implementation of the proposed framework is described in Section 3, when assuming there is no domain knowledge of the coupling relationships. Section 4 describes another variant of the proposed framework based on a hierarchical grouping representation, when some domain knowledge is available. Section 5 reports the experimental results while conclusions are drawn and future work is discussed in Section 6.

2. CBA FOR GROUP-BASED MANIPULATION DETECTION

The proposed framework consists of three key stages: *qualitative* analysis, which converts the transactional data to proper representations and provides a flexible coupling structure for the next-stage quantitative coupling relationships modeling; *quantitative* analysis, which characterizes the coupled behaviors by learning the corresponding model and is helpful for further analysis; and anomaly detection, which is the final stage of checking whether or not the new coupled behaviors are abnormal. The following sections describe how the proposed framework works, which is depicted in Figure 1.

2.1 Qualitative Analysis

Here we first briefly review the concept of *coupled behaviors*. Suppose there are I actors, and an actor i undertakes m_i behaviors $\mathbf{b_{i1}}, \mathbf{b_{i2}}, \cdots, \mathbf{b_{im_i}}$. Each actor i's j^{th} behavior $\mathbf{b_{ij}}$ is associated with a behavioral type $T(\mathbf{b_{ij}}) = t_{\mathbf{b_{ij}}}$ (e.g., buy, sell and trade.). Each behavioral type $t \in T$ has a number of associated properties $\mathbf{P^t} = (P_1^t, P_2^t, \cdots, P_n^t)(n$ may vary for different t value). Thus, each behavior $\mathbf{b_{ij}}$ is associated with a set of behavioral property value (a vector) $(p_1^{t_{\mathbf{b_{ij}}}}, \cdots, p_n^{t_{\mathbf{b_{ij}}}})$ (e.g., price and volume) determined by its behavioral type $t_{\mathbf{b_{ij}}}$. A *behavior feature matrix* $FM(\mathbf{b})$ for all actors for a specific period of time can be then represented as follows [3]:

$$FM(\mathbf{b}) = \begin{pmatrix} \mathbf{b_{11}} & \mathbf{b_{11}} & \cdots & \mathbf{b_{1m_{max}}} \\ \mathbf{b_{21}} & \mathbf{b_{21}} & \cdots & \mathbf{b_{2m_{max}}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{b_{I1}} & \mathbf{b_{I1}} & \cdots & \mathbf{b_{Im_{max}}} \end{pmatrix}. \quad (1)$$

where $m_{\max} = \max\{m_1, m_2, \ldots, m_{\mathrm{I}}\}$, and for each actor i, if $m_i < m_{\max}$ the corresponding element $b_{ij}(m_i < j \leq m_{\max})$ is defined as $\emptyset$, which means no action taken. Thus, the *intra-couplings* are reflected by the relationship between elements within one row of the above matrix, whereas the relationships between elements of different rows indicate the *inter-couplings*. Actor i's behaviors $\mathbf{b_{ij}}$ are intra-coupled with other behaviors of the same actor in terms of the corresponding function $\theta_k^i(\cdot)(1 < j \leq m_i, k \neq j)$ and inter-coupled with other actors' behaviors in terms of the corresponding function $\eta_k^i(\cdot)(1 < k \leq \mathrm{I}, k \neq i)$, with non-determinism.

The behavior feature matrix defined in Equation 1 represents a group of behaviors that are coupled for analysis. To consider the coupled relationships, the space for analyzing the couplings of these behaviors is almost infinite. For a

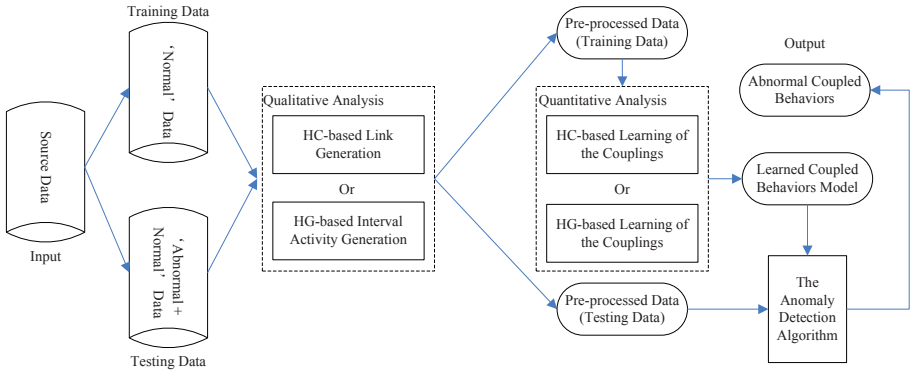


Figure 1: The Workflow of the Proposed Framework.

behavior $\mathbf{b_j}$ among n coupled behaviors, it can be coupled to any one of the remaining $n-1$ behaviors and the corresponding search space is $O(C_{n-1}^1) = O(n-1)$. Generally, if it is considered to be coupled to any k, $1 \leq k \leq n-1$, the possible search space can be $O(C_{n-1}^k) = O(\frac{(n-1)!}{k!\,(n-1-k)!})$. Thus, the sum of the above search space could be $O(2^{n-1})$, which means the computational complexity is exponential to the increase of the number of coupled behaviors. This is intractable when the number of coupled behaviors is large. To avoid this, the CHMM-based framework aggregates all the behaviors within time intervals and considers to model the couplings between these interval aggregated activities. The above approach may lose important coupling information within these aggregated behaviors, which may be useful for further anomaly detection. In this paper, to enrich the modeling capability of the CBA framework, we propose to integrate *qualitative* analysis into the CBA framework. To achieve this, two different strategies are designed, which are further discussed in Section 3 and Section 4. The former implementation is designed for the settings with no prior domain knowledge about the coupling relationships. Then we assume hybrid couplings exist between the behaviors, which means behaviors are associated with each other in a complicated structure of multiple different coupling relationships [3]. By contrast, when some domain knowledge is known about the couplings, the latter implementation is adopted. In this paper, we consider the hierarchical coupling structure [3], which means behaviors are coupled with each other in a hierarchical structure (determined by the corresponding grouping structure of the actors). All the above considerations are advantageous compared to the CHMM-framework because the possible coupled relationships we consider here are more comprehensive.

2.2 Quantitative Analysis

After the *qualitative* analysis of the coupled behaviors, the possible coupling relationships between behaviors are expanded and efficiently constrained, compared to the CHMM-based framework. Then how to quantitatively model the couplings becomes the key point and we solve it by modeling the autocorrelations that exist in coupled behaviors. More formally, for a set of coupled behaviors $\mathbf{b}$, there could be possible coupled relationships $(\theta(\cdot), \eta(\cdot))$; Then the autocorrelation for coupled behaviors with respect to one behavioral attribute P can be defined as follows:

Definition 1 (Coupled Autocorrelation). *It measures the dependence among the values of a behavioral variable $P \in \mathbf{P^t}$ defined on the coupled behavior pairs $(\theta(\cdot), \eta(\cdot))$. Given a set of coupled behavior pairs $(\theta(\cdot), \eta(\cdot))$, the autocorrelation of a continuous variable can be calculated as:*

$$ca = \frac{\sum_{i_1,j_1,i_2,j_2 s.t.(b_{i_1j_1}, b_{i_2j_2}) \in (\theta(\cdot),\eta(\cdot))} (p_{i_1j_1} - \bar{P})(p_{i_2j_2} - \bar{P})}{\sum_{is.t.i,js.t.b_{ij} \in (\theta(\cdot),\eta(\cdot))} (p_{ij} - \bar{P})^2}$$

Motivated by the considerations of the coupled autocorrelations for *quantitative* analysis, different strategies are proposed for different variant CBA frameworks to efficiently consider these coupled autocorrelations for modeling the coupled behaviors.

2.3 Anomaly Detection Techniques

To determine whether the new coupled behaviors $\mathbf{b^k}$ are normal or abnormal, we choose to calculate the likelihood given the observations of the coupled behaviors based on the established normal model M. The higher the likelihood of the coupled behaviors $\mathbf{b^k}$, the more likely $\mathbf{b^k}$ conforms to be normal. The following two sections will describe two variant implementations for the general framework proposed in this section.

3. THE HC-BASED FRAMEWORK

When there is no prior knowledge of how the behaviors are coupled, to comprehensively capture the couplings with reasonable search space is challenging. In addition, to avoid the loss of coupling information of aggregating all the behaviors within time intervals, we alternatively consider the possible hybrid couplings within the behaviors. To achieve this, we use links to indicate possible coupled relationships that are suggested by some of the qualitative behavioral properties of the behaviors, which can be seen as *qualitative* analysis. This is advantageous compared to the CHMM-framework because we do not forcedly aggregate the behaviors within one time intervals and makes it possible to consider the couplings between them. Then the remaining behavioral properties can be defined as quantitative properties and used for learning the coupling relationships between the behaviors from the perspective of *quantitative* analysis. The formal definition of two such properties is as follows.

Definition 2 (Qualitative Property). *A qualitative property $R \in \mathbf{P}$ refers to the behavioral property which is*

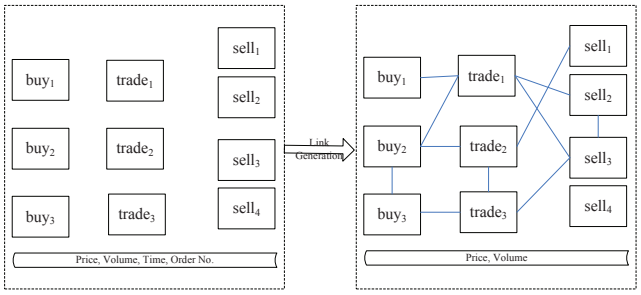


Figure 2: An Example of Qualitative Analysis.

used to generate the possible underlying coupling relationships between behaviors, usually user-defined.

DEFINITION 3 (QUANTITATIVE PROPERTY). *An quantitative property $A \in \mathbf{P}$ refers to the behavioral property which is used to learn the coupling relationships between behaviors.*

3.1 Qualitative Analysis: Link Generation

Based on the qualitative properties, Algorithm 1 describes the procedure to generate links according to the qualitative properties. In Algorithm 1, steps 2 to 8 form an inner loop process to generate links between behaviors which have the same values as the qualitative property and steps 1 to 9 form an outer loop process to generate the corresponding links for the behaviors according to every reference property. A toy example of the link generation process is shown in Figure 2, which transforms the raw behaviors[1] in the left to the linked behaviors in the right. To be more specific, in Figure 2, the behavioral properties of 'buy', 'sell' and 'trade' behaviors are: 'Price', 'Volume', 'Time' and 'Order No.'. The qualitative properties used are 'Time' and 'Order No.' while the quantitative properties used are 'Price' and 'Volume' after the link generation. Then a group of coupled behaviors can be represented by a typed attributed graph $G_B = (V_B, E_B)$. The nodes V_B represent behaviors (e.g., $buy_i (1 \leq i \leq 3)$, $sell_i (1 \leq i \leq 4)$ and $trade_i (1 \leq i \leq 3)$) and the edges E_B represent potential coupled relations among the behaviors (e.g., the edges in Figure 2).

Algorithm 1 The Link Generation Algorithm

Input: A Group of Behaviors $\{\mathbf{b_i}\}$,
A Set of Qualitative Properties $\{R_j\}$,
Output: A Set of Generated Links $\{l\}$.
1: $\{l\} \rightarrow \emptyset$
2: **for all** R_j in the set of qualitative properties **do**
3: **for all** values r_{jk} of the qualitative property R_j **do**
4: **if** exists $R_j^{b_m} == r_{jk}$ **then**
5: Add links $\{l_n\}$ between the behaviors $\{b_m\}$
6: $\{l\} \rightarrow \{l_n\} \cup \{l\}$
7: **end if**
8: **end for**
9: **end for**

3.2 Quantitative Analysis: Modeling Coupled Behaviors via Relational Learning

[1]Each behavior here refers to a transaction record.

After obtaining the graph structure of the coupled behaviors, we explore the learning of couplings between the behaviors in a numerical form. For a *quantitative* model to describe the coupling relationships, we choose to learn the joint probability distribution of these behaviors' attributes considering their coupled autocorrelations. Exact learning of this joint probability is very computationally intensive and we adopt an approximate approach by learning a set of conditional prolixity distributions (CPDs) [12]. For each quantitative behavioral property $A \in \mathbf{A}$ we try to learn the probability distribution of its values conditioned on other possible coupled behaviors' behavioral attribute values (i.e., the CPD of A). Consequently, we approximate the joint probability distribution of the coupled behaviors with a set of CPDs. For this purpose, we introduce relational dependency network (RDN) [12], to model the joint probability distribution of the coupled behaviors since its key idea is based on a set of CPDs.

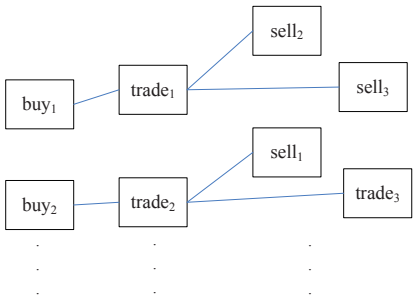
3.2.1 Approximating the Joint Probability Distribution

The RDN learning algorithm uses pseudo-likelihood techniques [12] to efficiently approximate the joint probability distribution. Unlike relational Bayesian network (RBN) [8] and relational Markov network (RMN) [14], it learns conditional distributions independently, rather than jointly, using local conditional probability models, such as Relational Bayesian Classifiers (RBCs). For example, considering the coupled behaviors in Figure 2, for each behavioral property of these coupled behaviors, the RDN learns a CPD model and obtain a set of CPD models. The joint probability distribution model is formed by the integration of these models, which becomes the quantitative model of coupled behaviors. The next section describes how to learn these CPDs.

3.2.2 Modeling of the CPDs

In consequence, in order to model the coupled behaviors, we could estimate a set of the CPDs of the quantitative behavioral properties conditioned on other possible coupled behaviors' behavioral attribute values, and underline which coupled autocorrelations between behaviors are considered. To learn the CPD is challenging because of two issues: each quantitative behavioral attribute instance could be conditioned on different linked behaviors' behavioral property values (e.g., heterogeneous structure of links and various behavioral property types) and the linked behaviors to consider could be limitless. To cater for the second issue, we must determine how much should be modeled for the CPD of each quantitative behavioral attribute instance. For computational simplicity, we may only consider two related behavior links from the target behaviors for modeling the CPD. The whole graph can then be decomposed into subgraphs according to each quantitative behavior property and this could be done by the visual query language QGraph [1]. A toy example can be seen in Figure 3(a) and the quantitative behavioral property A is one of the property of 'trade' behaviors ($trade_1$ and $trade_2$ in Figure 2). Then $trade_1$ and $trade_2$ are transformed to two subgraphs with consideration of coupled behaviors 1-link away (we could consider n-link ($n \geq 1$), but we only depict the 1-link situation for simplicity).

After that, the aforementioned first issue can be solved by "flatten" the subgraphs into propositional instances consist-



(a) An Example of the Subgraphs for Coupled Behaviors

	A	RF_1	RF_2	$\cdots$	RF_n
$trade_1$	x_1	rf_{11}	rf_{21}	$\cdots$	rf_{n1}
$trade_2$	x_2	rf_{12}	rf_{22}	$\cdots$	rf_{n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(b) An Example of the Relational Features for Coupled Behaviors

Figure 3: An Example of the "Flattened" Propositional Coupled Behavioral Data

ing of relational features $RF_1, RF_2, \cdots, RF_n$ and the quantitative behavioral property A. Here we choose RBCs to generate the relational features for simplicity and efficiency. The relational features for one quantitative behavioral attribute are the linked behaviors' attributes and each attribute contains multiple sets of values within the subgraph; for example, considering the quantitative property A of the 'trade' behaviors as in Figure 3. The RBC considers all the attributes associated with the linked behaviors 'buy', 'sell' and 'trade' and treats them as independent relational features $RF_1, \cdots, RF_n$. For any specific instance of RF_i $(1 \leq i \leq n)$, for example rf_{i1} is a multi-set of values $rf_{i11}, \cdots, rf_{i1m}$ and these values are assumed to be independently drawn from $p(RF_1|A)$. Then $p(RF_1|A)$ can be estimated by counting the frequency of the values in all the multi-set instances of RF_1 and similar estimation can be done on $p(RF_2|A), \cdots, p(RF_n|A)$. The CPD $p(A|RF_1, \cdots, RF_n)$ can be estimated as

$$\alpha p(A)p(RF_1|A)p(RF_2|A)\cdots p(RF_n|A) \quad (2)$$

where α is the normalized constant [12]. For example, with respect to the quantitative property A, the coupled behaviors in Figure 3(a) can be transformed into Table 3(b) by generating the relational features $RF_1, \cdots, RF_n$. CPD_A for A then becomes $p(A|RF_1, RF_2, \cdots, RF_n)$ and the joint probability distribution of coupled behaviors becomes $p(\mathbf{b}) = \prod_{A \in \mathbf{A}} CPD_A$, which model the coupled behaviors.

3.3 Efficient Abnormal Coupled Behavior Detection

Training on a set of coupled behaviors ($\{\mathbf{b^i}\}$) is computationally intensive and the coupling relationships in some $\{\mathbf{b^i}\}$ may be similar. Thus, in order to relief the computational complexity and modeling efficiency, we propose a match pursuit [11] like algorithm to only pick up a subset of coupled behaviors to model.

Matching pursuit [11] is a type of numerical technique widely used in signal processing. Informally speaking, the basic idea is to find the best representation of a signal using a subset of elements provided by a dictionary D. Similarly, our aim is to find the best representation of a subset of coupled behaviors among all groups of coupled behaviors. As described in Algorithm 2, step 1 initializes the dictionary D to all the set of training coupled behaviors. Then steps 2 to 11 choose the most representative groups of coupled behaviors and train a set of models $M_i (1 \leq i \leq n)$. Finally, steps 12 to 20 detect the anomaly based on the set of models. The training computational complexity is very intensive when the number of coupled behaviors increase and our proposed algorithm efficiently choose a small set of representative ones, which improves the efficiency of the learning process.

Algorithm 2 Matching Pursuit-like Anomaly Detection

Input: A Training set $\{\mathbf{b^i}\}$ $(1 \leq i \leq N)$,
A Testing set $\{\mathbf{b^k}\}$ $(1 \leq k \leq M)$,
Two Threshold Th_0, Th_1.
Output: An anomaly set $\mathcal{A}$.

1: $n \leftarrow 1$,
$D \leftarrow \{\mathbf{b^1}, \mathbf{b^2}, \cdots, \mathbf{b^N}\}$
2: **repeat**
3: Train one M_n model on the first training sample (sorted by temporal attributes, such as date.) chosen from D.
4: **for all** $\mathbf{b^d}$ $(1 \leq d \leq N)$ in D **do**
5: Compute the likelihood of $\mathbf{b^d}$ given the model M_n: $p(\mathbf{b^d}|M_n)$.
6: **if** $p(\mathbf{b^d}|M_n) > Th_1$ **then**
7: Delete $\mathbf{b^d}$ from D
8: **end if**
9: **end for**
10: $n \leftarrow n + 1$
11: **until** stop condition (e.g., $D = \emptyset$)
12: **for all** $\mathbf{b^k}$ in the Testing set **do**
13: **for** $i = 1 \rightarrow n$ **do**
14: Compute the likelihood of $\mathbf{b^k}$ given the model M_i: $p(\mathbf{b^k}|M_i)$
15: **end for**
16: $D_k = \max\{p(\mathbf{b^k}|M_i)\}$ $(1 \leq i \leq n)$
17: **if** $D_k < Th_0$ **then**
18: $\mathbf{b^k} \rightarrow \mathcal{A}$
19: **end if**
20: **end for**

4. THE HIERARCHICAL GROUPING-BASED FRAMEWORK

Sometimes, some prior domain knowledge is known about the possible coupling relationships among behaviors of some investors. This provides a possible method to explore couplings between behaviors of actors within certain subgroups (at a local level), rather than taking the whole population as a group (at a global level). Such analysis of behaviors is crucial for capturing more comprehensive couplings existing in behaviors, which may be helpful to further detecting abnormal coupled behaviors. As mentioned before, the CHMM-based CBA framework may fail to explore some detailed information of the couplings between behaviors, due to the aggregation of all the actors' behaviors, which makes the modeling relatively coarse and in turn may render it less

capable to detect abnormal coupled behaviors. For this purpose, with the presence of some useful domain knowledge, we propose the corresponding hierarchical grouping-based CBA framework below.

The *qualitative* analysis stage is to partition the whole investor population according to prior domain knowledge so that the couplings between the behaviors in the resultant sub-groups will be modeled more precisely. We define these groups as 'Particle Groups'. Performing CBA on behaviors of these particle groups can lead to more precise modeling of local interaction characteristics within local groups than directly on the global population. However, dividing actors into particle groups only cater for the couplings between behaviors within each particle group, it does not necessarily consider the couplings between behaviors from different particle groups. To cater for between-group couplings, we further propose a hierarchical clustering-based algorithm to merge those relevant particle groups into larger groups until finally into one super-group. Through this manner, the couplings between behaviors from different particle groups are expected to be captured. Then in the *quantitative* analysis stage, the coupled relationships in different group levels are also modeled. This hierarchical grouping strategy captures behavior interactions at different group levels, therefore is more effective to capture the more comprehensive couplings within the coupled behaviors, which makes it possible for detecting abnormal couplings more accurately. In the next sections, we illustrate the details of this variant framework for the general CBA.

4.1 Qualitative Analysis: Domain Knowledge-driven Initial Grouping

In stock markets, investors may be intentionally or unintentionally grouped. Based on domain knowledge, stock market surveillance experts often create rules according to their judgement and group investors who are likely to cooperate to have manipulative behaviors. As a result, an empirical blacklist is often generated for further monitoring. Such domain knowledge is very helpful for our initial understanding of investors grouping. We model the normal couplings in each group of investor behaviors, and build initial corresponding particle groups. This initial grouping of the investors aims to divide them into small particle groups so that the most likely coupled investors form into particle groups with similar coupled relationships. We define the result of these groups as particle groups.

Definition 4 (Particle Groups). *The particle groups, which are represented by* $\{PG_j\}$ $(1 \leq j \leq N)$ *are the partitioning result of actors* $\{A_i\}$ $(1 \leq i \leq I)$ *by the rule* $R(\cdot)$ *made by domain experts:*

$$R(\cdot)|\{A_1, A_2, \cdots, A_I\} \rightarrow \{PG_1, PG_2, \cdots, PG_N\}. \quad (3)$$

This *qualitative* analysis stage is domain knowledge driven and any useful information related to investors can be utilized to group the investors. In this paper, except for trading record data, there is no additional information available about the investors. Domain knowledge experts advise us to consider the average ordering/trading volume of a trading day for each investor as an initial grouping rule. Based on such information and related domain knowledge, investors are grouped as if they have similar ordering/trading volume of a security. Because the abnormal collaborative behaviors are more likely to happen in these predefined groups, it is reasonable to model the normal couplings within the groups and use the model to check the anomaly. By dividing the actors into particle groups and performing CBA on them separately, we can avoid the influence of roughly analyzing all the investors as a whole group, which may weaken the performance of the modeling. According to [4], we can convert the behavioral data for each particle group into three behavioral sequences, the 'buy', 'sell' and 'trade' sequence, in terms of the trading behavior types. For a particle group, we can transfer the trading records related to them into three coupled sequences. Further, in order to fit the behavior sequences to CHMM observation sequences for modeling, we also convert them to interval activities to reflect the characteristic of behaviors within the particle group during a period, similar to [4]. This makes modeling the couplings within these particle groups possible. However, this mechanism of dividing actors into particle groups can only describe the couplings between behaviors within each particle group, while it may omit the couplings between behaviors from different particle groups.

To overcome this, we not only model the behavioral couplings within particle groups but also merge the particle groups into larger groups and consider the coupled relationships in different group levels. To do this, we further define the distance/similarity measure based on the coupling patterns of the two particle groups. Because of the variety and dynamics of coupled behaviors, we merge those groups having most similar coupling patterns first and then join the remaining groups progressively into a super hierarchical group, within which there are many different levels of groups. Through this way, the coupled relationships between the behaviors are hopefully captured in a finer granularity. The detailed definition for the similarity measure for coupling patterns is discussed in the following. Each CHMM represents the corresponding coupling pattern for behaviors in each particle group, our proposed similarity measure is based on the distance/similarity between two CHMMs. Inspired by [7], we put forward a novel similarity measure based on the Kullback-Leibler (KL) divergence [9, 6], which is a standard measure for the similarity between probability density functions. To be more specific, for two particle groups i and j, represented by λ_i and λ_j respectively, there are corresponding interval activities sets $\{\mathbf{b}_{\mathbf{IA}}^{\mathbf{n,i}}\}$ $(1 \leq n \leq N)$ and $\{\mathbf{b}_{\mathbf{IA}}^{\mathbf{n,j}}\}$ $(1 \leq n \leq N)$, where N is the number of the trading days. The likelihood of the behaviors $\mathbf{b}_{\mathbf{IA}}^{\mathbf{n,i}}$ under λ_j is denoted as ξ_{ij}^n. For all the coupled behaviors in the set $\{\mathbf{b}_{\mathbf{IA}}^{\mathbf{n,i}}\}$, we obtain a likelihood subspace $\xi_{ij} = \{\xi_{ij}^1, \xi_{ij}^2, \ldots, \xi_{ij}^N\}$, which are "intelligently" sampled points from the model space representing the fitness of the coupled behaviors set to the CHMM λ_j. Similarly, we obtain $\xi_{ii} = \{\xi_{ii}^1, \xi_{ii}^2, \ldots, \xi_{ii}^D\}$ to denote the fitness of the coupled behaviors set $\{\mathbf{b}_{\mathbf{IA}}^{\mathbf{n,i}}\}$ to the CHMM λ_i. If we normalize ξ_{ij} and ξ_{ii}, the corresponding probability density functions $f_{\xi_{ij}}$ and $f_{\xi_{ii}}$ can be obtained. Then, the distance/similarity between two set of coupled behaviors is converted to the similarity measure between probability density functions, for which the KL divergence is a suitable choice. Its formulation for the discrete case is as follows:

$$D_{KL}(f_{\xi_{ii}}|f_{\xi_{ij}}) = \sum_m f_{\xi_{ii}}(m) \log \frac{f_{\xi_{ii}}(m)}{f_{\xi_{ij}}(m)} \quad (4)$$

In the same way, $D_{KL}(f_{\xi_{jj}}|f_{\xi_{ji}})$ reflects the similarity be-

tween the CHMMs λ_i and λ_j from the angle of the coupled behaviors set $\{\mathbf{b_{IA}^{n,j}}\}$. Finally, the symmetric distance between the coupled behaviors of two particle groups of λ_i and λ_j is defined as:

$$D_{KL}(\lambda_i|\lambda_j) = \frac{1}{2}[D_{KL}(f_{\xi_{ii}}|f_{\xi_{ij}}) + D_{KL}(f_{\xi_{jj}}|f_{\xi_{ji}})] \quad (5)$$

By calculating the similarities between particle groups, hierarchical clustering [15] is used on the basis of the similarity matrix to reveal the coupling structure in a hierarchical way, which is expected to make a full-scale modeling and can reflect different levels of couplings between behaviors.

4.2 Quantitative Analysis: Hierarchical Modeling of Coupled Behaviors

After the qualitative analysis of hierarchically grouping the investors and their corresponding behaviors. For the coupled behaviors of each group in different levels, we can learn the corresponding CHMM (for further details, please refer to [4, 3]). Then these CHMMs are the quantitative models that capture the couplings between the behaviors and provide helpful information for further anomaly detection.

4.3 The Anomaly Detection Algorithm

After hierarchically grouping (HG) of all the investors and quantitatively modeling their coupled behaviors in multi-level groups, we are also able to detect the abnormal coupled behaviors. Algorithm 3 illustrates the process.

Algorithm 3 The Anomaly Detection Algorithm of HG-based CBA

Input: A Training set $\{\mathbf{b^i}\}$ $(1 \le i \le N)$,
A Testing set $\{\mathbf{b^k}\}$ $(1 \le k \le M)$,
An Initial Grouping $\{PG_l\}$,
A Threshold Th_0.
Output: An anomaly set $\mathcal{A}$.

1: **for all** Particle group PG_l in the Training set **do**
2: Construct its interval activity behavioral sequences $\{\mathbf{b_{IA}^{1,PG_1}}, \mathbf{b_{IA}^{2,PG_1}}, \cdots, \mathbf{b_{IA}^{N,PG_1}}\}$.
3: **end for**
4: Hierarchically cluster the Particle group $\{PG_l\}$ and generate a hierarchical grouping $\{G_{l'}\}$;
5: **for all** Groups $\{G_{l'}\}$ in the Training set **do**
6: Train the corresponding CHMM $\lambda_{G_{l'}}$ for each group's behaviors $\{\mathbf{b_{IA}^{1,G_{1'}}}, \mathbf{b_{IA}^{2,G_{1'}}}, \cdots, \mathbf{b_{IA}^{N,G_{1'}}}\}$.
7: **end for**
8: **for all** Groups $\{G_{l'}\}$ in the Testing set **do**
9: Construct its interval activity behavioral sequences $\{\mathbf{b_{IA}^{1,G_{1'}}}, \mathbf{b_{IA}^{2,G_{1'}}}, \cdots, \mathbf{b_{IA}^{M,G_{1'}}}\}$
10: Calculate their likelihood $p(\mathbf{b_{IA}^{k,G_{1'}}}|\lambda_{G_{l'}})$ given the corresponding model.
11: **end for**
12: **for all** Trading day k in the testing set, **do**
13: we choose $L_k = \arg\min_l\{p(\mathbf{b_{IA}^{k,G_{1'}}}|\lambda_{G_{l'}})\}$
14: **if** $L_k < Th_0$ **then**
15: $\mathbf{b^k} \rightarrow \mathcal{A}$
16: **end if**
17: **end for**

5. EXPERIMENTS

5.1 Experimental Data

Our algorithms are tested on a real data set from a major Asian stock exchange. The tick data covers 388 valid trading days from 1 June 2004 to 31 December 2005. It consists of 58333 traders, and 174416 buy orders, 178464 sell orders, and 189148 trades. The data is partitioned into two sets suggested by domain experts. The training data set is extracted from the transactions from 1 June 2004 to 31 December 2004 and those transactions associated with the identified alerts is filtered. Models are trained on such labeled normal data to capture the characteristics in so-called 'normal' coupled trading behaviors. The test set consists of the remaining transactional data and is made up of both normal and abnormal coupled trading behaviors. For evaluating the performance of the proposed approaches, true positive TP, true negative TN, false positive FP and false negative FN are counted in terms of treating the abnormal cases as the positive class. Then four generally accepted measures, accuracy ($\frac{TP+TN}{TP+FN+FP+TN}$), precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), and specificity ($\frac{TN}{FP+TN}$) are adopted as the technical performance measures.

5.2 The Performance of the HC-based Framework

We tested the HC-based CBA framework using the RDN model (denoted as 'CBA-RDN') on the test data set and compared it to the CHMM-based CBA framework (denoted as 'CBA-CHMM')[2]. Figure 4 shows the technical performance. We vary the threshold Th_0 in Algorithm 2 for detecting different numbers of anomalies and compare the corresponding performance for both the algorithms. By doing this, we expect to provide a comprehensive comparison without considering the influence of the threshold. The horizontal axis (P-Num) stands for the number of detected group-based abnormal behaviors (i.e., the number of trading days with abnormal coupled behaviors), and the vertical axis represents the values of technical measures (accuracy, precision, recall or specificity). Figure 4 shows the four technical measures of the HC-based and CHMM-based CBA frameworks. Generally speaking, the former performs better than the latter in terms of all the metrics. For instance, the precision[3] of the CBA-RDN framework of $P - num = 45$ is 0.38 while that of the CBA-HMM framework is 0.28, which improves the precision of 35% higher than that of the CHMM-based framework. With respect to other measures, the CBA-RDN has better results as well, which proves that the HC-based CBA framework can characterize the coupled behaviors with more comprehensive information than the CHMM-based framework. This may lead to better anomaly detection performance to some extent.

5.3 The Performance of the HG-based Framework

For the HG-based framework, here we study seven variant strategies for detecting abnormal coupled behaviors.

[2]The performance results are the averaged values of different time sliding windows [4].
[3]The precision looks very low, since we use the alerts from the market surveillance system as the benchmark, which is known with high overall false positive rate.

Models \ P-Num	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
CBA-RDN	**0.88**	**0.88**	**0.88**	**0.87**	**0.87**	**0.87**	**0.86**	**0.86**	**0.86**	**0.85**	**0.85**	**0.85**	**0.84**	**0.84**	**0.84**	**0.83**
CBA-CHMM	0.85	0.85	0.85	0.85	0.84	0.84	0.84	0.84	0.83	0.83	0.82	0.82	0.81	0.81	0.81	0.80

(a) Accuracy

Models \ P-Num	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
CBA-RDN	**0.38**	**0.39**	**0.38**	**0.37**	**0.36**	**0.36**	**0.35**	**0.34**	**0.34**	**0.33**	**0.33**	**0.32**	**0.31**	**0.31**	**0.30**	**0.29**
CBA-CHMM	0.28	0.28	0.28	0.28	0.27	0.26	0.26	0.26	0.26	0.25	0.24	0.24	0.23	0.23	0.23	0.22

(b) Precision

Models \ P-Num	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
CBA-RDN	**0.92**	**0.91**	**0.91**	**0.90**	**0.90**	**0.90**	**0.89**	**0.89**	**0.88**	**0.88**	**0.88**	**0.87**	**0.87**	**0.86**	**0.86**	**0.85**
CBA-CHMM	0.90	0.90	0.89	0.89	0.89	0.88	0.88	0.87	0.87	0.87	0.86	0.86	0.85	0.85	0.84	0.84

(c) Specificity

Models \ P-Num	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
CBA-RDN	**0.52**	**0.55**	**0.56**	**0.56**	**0.56**	**0.57**	**0.58**	**0.58**	**0.58**	**0.58**	**0.59**	**0.59**	**0.59**	**0.60**	**0.60**	**0.60**
CBA-CHMM	0.39	0.40	0.41	0.42	0.42	0.42	0.43	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.45	0.45

(d) Recall

Figure 4: Technical Performance of the Two Frameworks

- CBA-CHMM: The behaviors of all the investors are collected together to construct corresponding coupled behavior sequences and a CHMM is trained to represent the characteristics of the group-based coupled behaviors.
- CBA-PG, CBA-PG3: The investors are divided into different particle groups based on different predefined domain knowledge and then their behaviors are analyzed separately for each particle group. A set of CHMMs are trained for each particle group.
- CBA-PG2: The investors are randomly divided into different particle groups. The other processing is the same as that of CBA-PG and CBA-PG3.
- CBA-HG, CBA-HG3: On top of the CBA-PG and CBA-PG3 respectively, the proposed hierarchical clustering method is adopted to provide a comprehensive profile of the investors' behaviors. Consequently, the behaviors are analyzed in different scales and the corresponding CHMMs for each group is set up, which is expected for more accurate analysis of anomalies.
- CBA-HG2: On the basis of CBA-PG2, particle groups are randomly merged into a hierarchical grouping structure. The other parts of processing are the same as CBA-HG and CBA-HG3.

We tested the above seven strategies on the test dataset by setting various window sizes (*winsize*) and the results reported here are the averaged ones over different window sizes. Figure 5 shows their technical performance. The horizontal axis and the vertical axis have the same meanings in Figure 4. The performance of CBA-PG, CBA-PG2 and CBA-PG3 are worse than CBA-CHMM. This is because dividing investors into small particle groups only models the couplings within these groups and ignores the coupled relationships between particle groups. By contrast, while CBA-HG and CBA-HG3 perform better than other schemes in most cases, CBA-HG2 does not. A possible explanation is that the former two schemes integrate a hierarchical group structure based on reasonable similarity measures rather than a random merger. CBA-HG generally performs best out of the three strategies at general, which indicates suitable domain knowledge and a reasonable hierarchical structure is valuable for a more accurate grouping and corresponding anomaly detection of the investors. Although the technical measures of CBA-HG is not much higher than those of other strategies when P-Num is small (smaller than 45), they are significantly higher as P-Num increases. For instance, as shown in Figure 5, when $P - Num = 40$, the precision of CBA-HG is 0.33, while that of CBA-PG is 0.24 and that of CBA-CHMM is 0.25. The precision of CBA-HG can be 32% higher than that of CBA-CHMM. It reveals that our proposed framework performs best and is most stable than other strategies.

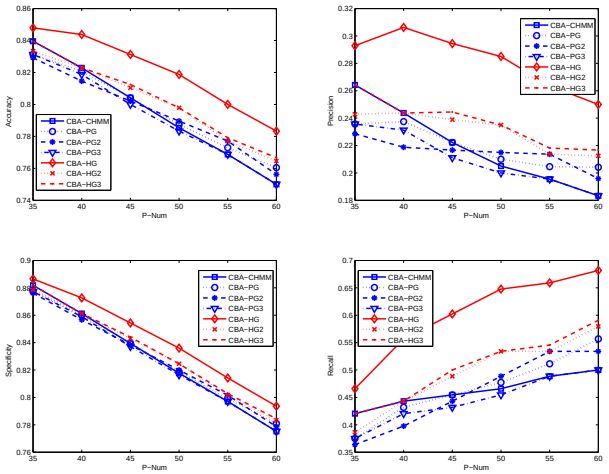


Figure 5: Comparison of the Seven Strategies

5.4 Comparisons of the Three Frameworks

The above experiments show the performance of the two variant CBA frameworks respectively. To further compare their performance, Figure 6 shows the performance comparison of the two variant frameworks and the CHMM-based framework. As can be seen from the figure, the CBA-HG framework obtains the best performance in terms of all the

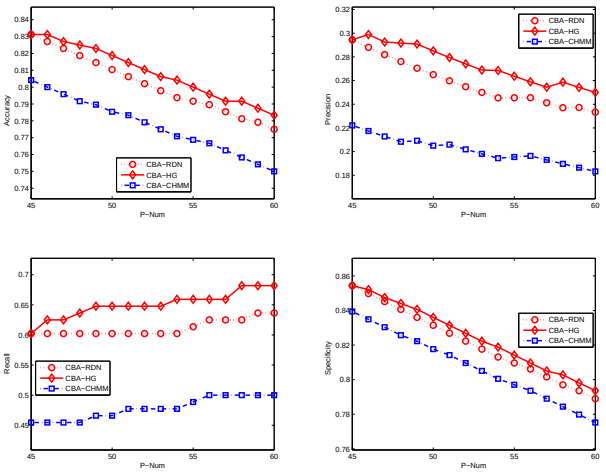


Figure 6: Comparison of the Three Frameworks

technical measures. This is because the proper domain knowledge directs us to model proper and more comprehensive couplings between behaviors for anomaly detection. In addition, this proves integrating the proper domain knowledge for CBA could enhance its performance to some extent. The CBA-RDN performs better than the CBA-CHMM framework, which may benefit from the modeling of more comprehensive coupling relationships. When there is no prior domain knowledge, it is reasonable to assume hybrid couplings between the behaviors and give a comprehensive modeling. This method is still advantageous compared to the CBA-CHMM framework, which aggregates all investors' behaviors and may omit important coupled relationships. To sum up, the HC-based CBA considers hybrid couplings between behaviors for modeling is helpful when there is little knowledge about the coupling structures while HG-based CBA could provide better analysis of the coupled behaviors when the proper domain knowledge is set up for the underlying couplings.

6. CONCLUSIONS

This paper examined a challenging issue of detecting group-based market manipulations from the perspective of coupled behavior analysis. In order to analyze the rich couplings among behaviors and detect anomalies more accurately, we proposed a three-stage general CBA framework for abnormal behavior detection, which consists of *qualitative* analysis, *quantitative* analysis and anomaly detection stages. To cater for different situations of prior domain knowledge about the couplings, two variant implementation approaches have been proposed. The experimental results on a real-world data set in a major Asian stock market exhibited that the proposed approaches generally outperform the previous CHMM-based one by taking the miscellaneous alerts fired in the market surveillance system as a benchmark. In addition, we also found that integrating domain knowledge into abnormal coupled behavior detection can significantly improve the performance while detecting anomalies without additional domain knowledge is still possible. Future research could be on exploration of how to integrate more sophisticated domain knowledge for CBA and considering other application domains.

7. ACKNOWLEDGEMENT

This work is supported by the Australian Research Council (ARC) Discovery Grant (DP1096218), ARC Linkage Grant (LP100200774), the National 863 Program of China (2012AA011005), and the US NASA Research Award (NNX09AK86G).

8. REFERENCES

[1] H. Blau. A visual query language for relational knowledge discovery. Technical report, University of Massachusetts Amherst, 2001.

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997.

[3] L. Cao, Y. Ou, and P. Yu. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering*, 2011.

[4] L. Cao, Y. Ou, P. Yu, and G. Wei. Detecting abnormal coupled sequences and sequence changes in group-based manipulative trading behaviors. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 85–94. ACM, 2010.

[5] D. Cohen. *JB Watson, the founder of behaviourism: a biography*. Routledge and Kegan Paul, 1979.

[6] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

[7] D. Garcia-Garcia, E. Hernandez, and F. Diaz de Maria. A new distance measure for model-based sequence clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1325–1331, 2009.

[8] L. Getoor and B. Taskar. *Introduction to statistical relational learning*. The MIT Press, 2007.

[9] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[10] H. Liu, J. Salerno, and M. Young. *Social computing, behavioral modeling, and prediction*. Springer-Verlag New York Inc, 2008.

[11] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[12] J. Neville. *Statistical models and analysis techniques for learning in relational data*. PhD thesis, 2006.

[13] B. Skinner. About behaviorism, alfred a, 1974.

[14] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02)*, pages 895–902, 2002.

[15] J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.