

EMPIRICAL FINDINGS IN ASSET PRICE DYNAMICS REVEALED BY QUANTITATIVE MODELLING

A Thesis
Presented to
The Academic Faculty

by

Min Kyu Sim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2014

Copyright © 2014 by Min Kyu Sim

EMPIRICAL FINDINGS IN ASSET PRICE DYNAMICS REVEALED BY QUANTITATIVE MODELLING

Approved by:

Professor Shijie Deng, Advisor
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Xiaoming Huo
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Sebastian Pokutta
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Soohun Kim
College of Business
Georgia Institute of Technology

Dr. Kautilya Raval
Bank of America

Date Approved: September 30, 2014

TABLE OF CONTENTS

LIST OF TABLES	v
LIST OF FIGURES	vii
SUMMARY	ix
I INTRODUCTION	1
II A MEASURE OF CLUSTER STRUCTURE IN STOCK MARKET	7
2.1 Introduction	8
2.2 Construction of connectedness measures	11
2.3 Marketwide empirical evidence	14
2.3.1 Entire period of 2002-2011	14
2.3.2 Subperiods of bull/bear markets	17
2.3.3 The modularity measure and the cycle of the economy	22
2.4 Is <i>MOD</i> factor priced?	23
2.4.1 Classical Model	24
2.4.2 Model	26
2.4.3 Procedure to construct decile portfolios	26
2.4.4 Tests on decile portfolios — return difference	29
2.4.5 Estimation of premium on modularity risk	31
2.4.6 Investment perspectives (1) - with <i>MOD</i> basis portfolio of the difference in both decile ends	34
2.4.7 Investment perspectives (2) - with <i>MOD</i> basis portfolio constructed by minimum idiosyncratic risk procedure by [43]	37
2.5 Concluding remark	41
III ESTIMATION OF HIDDEN LIQUIDITY AND ITS EMPIRICAL ADVANTAGES IN PRICE IMPACT FUNCTION, ORDERBOOK PRESSURE MODEL, AND OPTIMAL ORDER EXECUTION STRATEGIES	42
3.1 Introduction	42

3.2	Background and literature review	44
3.2.1	Price impact function	45
3.2.2	Orderbook pressure: microstructure model of order imbalance and midprice movement	46
3.3	Model and estimation strategy	48
3.3.1	Review on the price impact function of Cont et al. (2013) . .	49
3.3.2	Model development	51
3.3.3	Two step estimation process	54
3.4	Empirical evidence	54
3.4.1	Enhancing the price impact function	55
3.4.2	The prediction power in market microstructure model	56
3.4.3	Intraday patterns	61
3.5	Optimal execution strategy	61
3.5.1	Setting and goal	62
3.5.2	Assumptions	62
3.6	The strategy tested on the market	65
3.7	Concluding remark	68
APPENDIX A — APPENDIX FOR CHAPTER 2		72
REFERENCES		81

LIST OF TABLES

1	Tabular view of Figure 1 with descriptions. The stocks are chosen based on 2012 Fortune 500 ranking	16
2	Returns and alphas of the decile portfolio (January 1992 - December 2011)	29
3	Betas of decile portfolio sorted by predicted modularity betas. (January 1992 - December 2011)	30
4	Composition of portfolio returns by alpha and exposures to four factors. (January 1992 - December 2011)	30
5	Estimated premium of modularity factor by GMM (with using three/four factors)	34
6	Weights in the ex-post tangency portfolio and monthly Sharpe ratios (January 1992 - December 2011)	35
7	Cross-correlation matrix of the five factors including MOD_S (January 1992 - December 2011)	36
8	Performance of enhanced market index portfolio (January 1992 - December 2011)	36
9	Weights in the ex-post tangency portfolio and monthly Sharpe ratios (January 1992 - December 2011)	39
10	Cross-correlation matrix of the five factors including MOD_P (January 1992 - December 2011)	40
11	Performance of enhanced market index portfolio (January 1992 - December 2011)	40
12	The probability of midprice going up given orderbook states (2012 January, BAC).	47
13	Average results of statistical properties in (35) (The intercept is added for testing)	57
14	Average results of statistical properties in price impact functions (from January 2012 to June 2012)	58
15	Average errors in orderbook pressure model by ours and by [7] (from January 2012 to June 2012)	60
16	Implementation Shortfall (in \$ per share)	67

17	(Equal-Weighted, corresponding to Table 2) Returns and alphas of the decile portfolio (January 1992 - December 2011)	77
18	(Equal-Weighted, corresponding to Table 3) Betas of decile portfolio sorted by predicted modularity betas. (January 1992 - December 2011)	77
19	(Equal-Weighted, corresponding to Table 4) Composition of portfolio returns by alpha and exposures to four factors. (January 1992 - December 2011)	77
20	Stocks with frequent appearances in low decile portfolios. (Stocks with low modularity beta)	79
21	Stocks with frequent appearances in high decile portfolios. (Stocks with high modularity beta)	80

LIST OF FIGURES

1	Cluster analysis with 10 years returns of 60 major stocks by MMC algorithm	15
2	Partial zoom-in for three cells in Figure 1. The blue solid lines indicate correlations within cells and the red dotted lines indicate correlations across the cells.	17
3	Heat map matrix representation of correlation structure in Figure 1 for 2002-2011. The difference of inner cells correlations and inter cells correlations are noticeable.	18
4	Heat map representation for a few bull/bear periods. Upper Left: 03/19/2009-06/11/2009, S&P 500 39.7% increase, The difference in average of diagonal and non-diagonal elements is 0.179. Upper Right: 06/11/2009-04/15/2010, S&P 500 28.3% increase, The difference in average of diagonal and non-diagonal elements is 0.171. Lower Left: 04/15/2010-07/02/2010, S&P 500 15.7% decrease, The difference in average of diagonal and non-diagonal elements is 0.090. Lower Right: 07/01/2011-08/10/2011, S&P 500 16.3% decrease, The difference in average of diagonal and non-diagonal elements is 0.085.	19
5	Partial zoom-in for three cells in Figure 4. The upper diagrams have higher modularity than the lower diagrams.	20
6	Top: The six months moving average series for monthly <i>MOD</i> measured based on cluster structure in Figure 1. Bottom: The six months moving average return for S&P 500. The two plots behave similarly for many subperiods, indicating that <i>MOD</i> can be a strong indicator for the cycle of economy.	22
7	Cumulative wealth growth on enhancement scenarios by MOD_S . . .	37
8	Cumulative return on enhancement scenarios with MOD_P	40
9	Orderbook with uniformly sized waiting orders and the effect of imbalance on it	49
10	Two equivalent scenarios of measuring e_n of [22].	51
11	Different perspectives of the two scenarios with the presence of hidden liquidity	52
12	Intraday pattern in median of visible and hidden liquidity in 2012H1 (125 trading days) (1/2)	69
13	Intraday pattern in median of visible and hidden liquidity in 2012H1 (125 trading days) (2/2)	70

14	Example of trading program comparison - BAC, April 2012	71
----	---	----

SUMMARY

This dissertation addresses the fundamental question of what factors drive equity prices and investigates the mechanisms through which the drivers influence the price dynamics. The studies are based on the two different frequency levels of financial data. The first part aims to identify what systematic risk factors affect the expected return of stocks based on historical data with frequency being daily or monthly. The second part aims to explain how the hidden supply-demand of a stock affects the stock price dynamics based on market data observed at frequency levels generally between a millisecond and a second. With more and more financial market data becoming available, it greatly facilitates quantitative approaches for analyzing asset price dynamics and market microstructure problems.

In the first part, we propose an econometric measure, terms as modularity, for characterizing the cluster structure in a universe of stocks. A high level of modularity implies that the cluster structure of the universe of stocks is highly evident, and low modularity implies a blurred cluster structure. The modularity measure is shown to be related to the cycle of the economy. In addition, individual stock's sensitivity to the modularity measure is shown to be related to its expected return. From 1992 to 2011, the average annual return of stocks with the lowest sensitivity exceeds that of the stocks with highest sensitivities by approximately 7.6%. Considerations of modularity as an asset pricing factor expand the investment opportunity set to passive investors.

In the second part, we analyze the effect of hidden demands/supplies in equity trading market on the stock price dynamics. We propose a statistical estimation

model for average hidden liquidity based on the limit orderbook data. Not only the estimated hidden liquidity explains the probabilistic property in market microstructure better, it also refines the existing price impact model and achieves higher explanation powers. Our enhanced price impact model offers a base for devising optimal order execution strategies. After we develop an optimal execution strategy based on the price impact function, the advantage of this strategy over benchmark strategies is tested on a simulated stock trading model calibrated by historical data. Simulation tests indicate that our strategy yields significant savings in transaction cost over the benchmark strategies.

CHAPTER I

INTRODUCTION

Over the past couple of decades, the breadth and the depth of the field of quantitative finance have been greatly extended by the wide availability of financial data, the growing numerical processing power, and the theoretical researches in relevant financial modeling. It has been recognized that there may exist a set of common risk factors which drive the return dynamics of the financial assets and in particular, the co-movements of the returns. Understanding the return drivers and how returns of various financial assets correlate with each other is one of the keys in addressing the fundamental questions of asset pricing and portfolio management. The first part of this dissertation tackles the inherit relationship of financial asset returns by quantitatively analyzing the clustering property of financial asset returns and consequently formulates a new asset pricing factor based on the proposed clustering measure. The second part of the dissertation investigates one particular aspect of the microstructure of equity trading markets which is the liquidity provided by bid and ask orders. It proposes an estimation strategy of invisible liquidity caused by the practice of hidden order placements in electronic based trading venues.

In part one, we explore the clustering property of equity returns and demonstrate that it can be interpreted as one common risk factor in driving stock returns. Factor based asset pricing model, which is the cornerstone of modern asset pricing theory, is to identify and to formulate marketwide or a common firm specific characteristic that contributes to the explanation of expected asset return. Since one factor model by [57][44][46] identified market portfolio as a common driver of the return of individual stock and decomposed the financial risk into systematic and unsystematic risk,

many studies on factor model have explored the idea of identifying more alternative systematic risk components. The three factor model by [26][27] established the firm characteristics, size and growth, into new factors and [14][41] found that historical returns add more explanation power. These studies established what is the so-called four factor model. Studies of factor model are often conducted either by finding better measurements of factors [56], by replacing established factors with new factors with higher explanation power [17], or by finding additional factors [28]. A new factor is often added by focusing on the contemporary interests in the financial market as well. For example, [51] shows that individual stock's sensitivity to market level liquidity innovation can lead to significant return difference. In similar vein, we show that individual stock's sensitivity to our measure of the market structure can be identified as an additional driver of stock return.

In order to identify an alternative asset pricing factor, we draw our attention on the interactions among the returns of financial assets and the changes of the interaction network. We quantify the level of interaction as the notion of distance. We say that two assets are close or have a small distance if they have relatively high co-movement tendency. On the other hand, we say that two assets are distant or have a large distance if they have relatively low co-movement tendency. This subject has been of great interests among researchers, to name a few, [10][12][24][55], after the financial crisis of 2007-2008. These studies show that the distances between financial assets are not static, and less distance is a sign of stressed markets in general. Their unique perspectives of distance structure provide pictures of dynamic evolution of financial market. Studies such as [40][45] have focus on visualization of network of financial assets and the relationship of changes in the network structure and the whole markets. In addition to such perspective in market level, we further investigate how individual assets react differently to the changes of market structure and how passive investors may utilize this observations.

In identifying our proposed risk factor, we utilize intuitions gained from multivariate statistical analysis [45][16], whose purposes are to identify and visualize fundamental structure of the network of returns of financial assets. In order to identify the structural backbone of the network of financial assets, we apply a numerical method of the combinatorial partitioning problem [59][47]. Identifying correlation structure alone provides view points at the market level, but further quantification is also important to draw out observations at individual asset level. Classical studies on correlation elements [30][31][58][62], which establish the robust ways to deal with Pearson’s correlation element, which is geometrically a cosine value of the angle that the two random vectors under the consideration create, are reviewed to find out how correlation elements must be manipulated. After defining our main measure of market structure, *modularity*, we test the possibility of this measure to be a new asset pricing factor. We support the factor model by presenting empirical evidence under standard methods used by [26][27][14][51].

The proposed method of visualization via our measure of market structure can be applied to many other types of financial data and provide dynamic marketwide perspectives in reduced dimension. As a good factor model should do, the consideration of our proposed factor can provide an additional instrument for portfolio diversification and risk management.

The second part of this dissertation studies the liquidity aspect of limit orderbook based trading markets. It focuses on the estimation of hidden limit orders through event-level limit orderbook data. Limit orderbook database, which is the real time database of quote and trade, provides snapshots of arrival processes of the supplies and the demands for financial securities and the transactions occurred in double-auction based electronic trading venues. Developments of data processing techniques have made these huge sets of records accessible to researchers for carrying out statistical data analysis and probability modelling.

The immediate transparency of the market snapshot brought by the virtue of information technology, however, does not always result in the enhancement of market efficiency. Market participants are often afraid of revealing their intended trade direction and quantities because of the possibility of losing their informational edges. The feature of hidden order placements is therefore enabled and have become a substantial portion of market liquidity. Therefore, this feature calls for reassessments and possible modifications on the observations and models regarding the orderbook mechanism.

Among the existing models that address various aspects of the limit orderbook mechanism, we focus on the *orderbook pressure* model and the *price impact* function. Orderbook pressure is the effect of the relative magnitude difference of ask and bid queue waiting orders to the future price change direction. The reduced form of orderbook pressure model by [23] assumes *no knowledge model* that only level-I waiting order sizes are relevant and that other variables are ignored. It also assumes the *Markovian property* that future transitions of orderbook states are independent to the historical transitions. This assumption is often challenged by empirical observation such as [6][37]. [23] assumes that orderbook transitions are independent Poisson Processes and derive the probability of midprice increase given the state variables. [21] extends the work with diffusion approximation. In light of work of [21], [7] incorporate the presence of hidden liquidity and propose estimation method. [7] therefore serves a benchmark of our study.

On the other hand, price impact function represents how prices change in response to the supply-demand imbalance caused by traders. Since price impacts account for a substantial portion in transaction costs, estimating this function alone is a critical task in presiding the design of order placement strategies. The approach of studying price impact function can be theoretical [22][38][42] or empirical [4][54]. [22] build a price impact model based on the theoretical mechanism of orderbook transition

which is well suited to [38] that proves that permanent linear price impact is only form under arbitrage-free condition. Based on the fact that hidden liquidity has less priority to visible counterpart, we modify mechanism based approach of [22] to propose an estimation method of hidden liquidity.

Bringing the attention back to the hidden order placement, there have been a few studies regarding the effect and the property of hidden liquidity. [32][13] provide empirical evidence such that more hidden liquidity results in greater liquidity and trading volume, and less impact toward market price changes are observed. Other than knowing its impact on the market, estimating its size in the market have been often studied with subscription based database [15][61]. In our search, [7] is so far the only study on the estimation method without relying on the exclusive historical database of hidden liquidity. They conceive that hidden liquidity must function similar to visible liquidity as far as *orderbook pressure* is concerned. In a similar vein, we develop our estimation method based on the fact that hidden liquidity must function similar to visible counterpart when *price impact* is concerned.

Our estimation of hidden liquidity can lead empirical advantages in many aspects including (but not limited to) a few that is presented in this study. First, our estimation based on the price impact function enhance the explanation power of price impact function, so that the possible impact of supply-demand imbalance to the market price can be better predicted. Second, our estimation can be used as an input value of the orderbook pressure model [21] so that the probability model of microstructure can be improved. Last, the enhanced price impact function can be a cornerstone of devising a large order trade execution strategy. Our simulated testing on the historical market data shows the significant savings in transaction cost than relevant studies [5][8][49].

Overall, this dissertation takes data-driven approach to analyze several important questions arising from the research fields of asset pricing and market microstructure modeling. We provide empirical evidence that cluster structure in stock market

changes over time and quantify the changes. The measure has a meaning of the quality of investment opportunity set and it is related to the cycle of economy in financial market. We provide empirical evidence that it can serve as a state variable for the return of individual stocks and that the consideration of this measure as an asset pricing factor enhances investment opportunity set to passive investors. The estimation model of hidden liquidity in second part is useful to estimate average hidden liquidity that has several empirical advantages. Incorporating the estimated hidden liquidity enhances orderbook pressure model that predicts the direction of future price movement based on orderbook states. Estimation hidden liquidity can be a component of price impact function that results in higher explanation power. Based on our observation that available liquidity in the market has certain intraday patterns, we devise analytic solution to make use of the observation. This optimal execution strategy saves a significant portion of transaction costs in our test under simulated market.

CHAPTER II

A MEASURE OF CLUSTER STRUCTURE IN STOCK MARKET

We propose an econometric measure, terms as *modularity*, to measure time-varying cluster tendencies of the returns in stock markets. Using cluster analysis on the U.S. major stocks, we find that stressed markets have a less evident clustering tendency in the returns of stocks (low modularity) and bullish markets have a clear clustering tendency (high modularity). Not only is the modularity measure related to the cycle of the economy, but the measure can be also used as a priced state variable for individual stocks or portfolios. During the years from 1992 to 2011, the average return of stocks with the lowest sensitivity to the modularity exceeds that of the stocks with highest sensitivities by approximately 7.6% annually. The consideration of modularity as an asset pricing factor can expand the investment opportunity set to passive investors.

2.1 Introduction

Do expanded financial markets expand the investment opportunity set? Granted, the scope of financial markets has been expanding during recent decades through introductions of vast amounts of stocks and diverse derivatives products, which should yield a more expanded investment opportunity set in general. However, because the levels of interactions within and among financial markets have also increased during the last decade [10], it is questionable whether the growing scope of financial markets has led to a more expanded investment opportunity set. That is, sizable but concentrated market structures do not necessarily provide diverse investment opportunities to market participants. Thus, market participants must comprehend the inter-related structures in markets in order to truly assess the investment opportunity set so that they can practice portfolio diversification and risk managements effectively.

There are growing interests in quantifying the level of association in financial assets in order to assess overall market structures in the perspective of a graph or a network. We call this line of research *the study of market connectedness*¹. For example, [10] and [24] construct their connectedness measures in financial institutions to measure the level of systemic risk during the global recession period in 2007-2008 and provided empirical evidence that their measures are related to the cycle of economy. Our study constructs connectedness measures to adopt network perspectives in stock markets, and we further explore the relationships between our proposed measure and the movements of individual stocks that the existing studies of connectedness have not provided.

We focus on the time-varying correlation structure of stock markets in order to assess the investment opportunity set as the stock markets change over time. There

¹The development of this line of study is largely grounded in the development of graph theory or network theory that are actively studied in the various disciplines such as combinatorics, computer science, physics, and (bio)-statistics.

have been many studies regarding the time variability of correlations and its relationship to the whole stock markets. Studies such as [12][55] report the phenomenon of soaring level of correlations in stressed stock markets in 2007-2008 period which are influenced by common drivers for stock markets, such as the stressed domestic/global economy, consumers' negative sentiments, major market participants' solvency issues, and so on. Furthermore, the relationship between the level of correlation and the cycle of economy is not limited to the periods of recession. [53] presents empirical evidence that the average correlation in the stock market has explanation power in return of the S&P 500 index. When considering correlation structures, we sense that these studies are missing an important property in market structure, which is *the clustering tendency* in stock markets. With cluster analysis, we classify the correlation elements into two kinds: ones that tend to be always naturally higher and the others that fluctuate a lot with the cycle of economy. Our measure built on the classifications of correlation elements enables to observe the time variations of clustering tendency and to develop an asset pricing model for individual stocks and portfolios which aforementioned studies did not achieve.

According to recent studies, the cluster property, where entities with similar characteristics tend to form a subgroup or a cell, is one of the most evident and important structural properties in financial markets. [45] provides empirical evidence that the major stocks in the U.S. can be drawn in tree structure, a special case of cluster structure, where branches connect the highly correlated stocks together. [40] utilizes tree structural view to assess the structural changes in foreign exchange markets as well. [52] classifies foreign exchange investing funds into two groups and proposed a few crowdedness measures for co-movement tendency of market participants. [16] conducts an empirical analysis in the U.S. stock market by hidden Gaussian graphical model which showed that the clustering tendency of across the universe of stocks is

observable even after eliminating a few common drivers of the stock market.² Since we believe that the findings of clustering tendency by aforementioned studies [45][16] can be refined in more quantifiable ways, our goal is to propose a quantitative measure based on the cluster structure.

Our study proposes a new form of measure for connectedness in a stock market. Conducting cluster analysis, we visualize and analyze the high dimensional space of stock markets with emphasis on the different levels of 1) correlations among the stocks' returns in the same cells and 2) correlations of stocks' returns across the different cells. In normal market conditions, the difference of the two sets of correlations is noticeable and this noticeable difference visualizes the evident cluster structure. When the market is stressed, the correlations across the cells increases much more than the correlations within the cells and the cluster structure becomes indistinct. Our proposed measure, modularity, quantifies the degree of clustering tendency by taking the difference of correlations within the cells and the correlations across the cells, and is shown to be closely related to fluctuations of market condition. We show that modularity can serve as a priced state variable, i.e., we consider the possibility of modularity as an extra factor to the Fama-French three factor model. We show that the decile portfolios, constructed by ranking individual stock's sensitivity to market modularity, makes 7.59% of annual return difference between its top and bottom decile portfolio. The result is statistically significant for the recent 20 years period and the risk premium is estimated as 6.89%-7.92%. We then provide ex-post empirical evidence that the consideration of modularity as an asset pricing factor can expand the investment opportunity set.

²Typical dimension reduction techniques such as principal component analysis and graphical model of [16] tend to lose its advantage of succinct modeling via identifying a few common drivers in case that several groups of entities are highly correlated within each group. That is, it generally costs an additional factor (or a principal component) to capture co-movement tendency in each group of highly correlated stocks. In our construction of connectedness measure, we aim to capture the strength of cluster as a whole without dealing with each cell by cell.

Our study is comparable to the studies of connectedness in financial markets such as [10][24][53][40]. We distinguish our study from them by providing a quantitative asset pricing model for individual stocks as well as providing qualitative implications in market level. Our study extends the areas of applications of the studies [45][16][59] on the analysis of cluster network. We believe our study's time varying cluster perspective can be also adopted to applications of cluster analysis such as [1]. Our motivation is similar to [14][51] in a sense that we attempt to expand a factor asset pricing model by introducing modularity as another factor.

This study proceeds as follows. Section 2 defines the measure of modularity. Section 3 empirically observes the modularity measure during the several subperiods as well as the recent decade as a whole. Section 4 examines whether the modularity can serve as an additional factor in modern factor-based asset pricing models by Fama-French and momentum factor. Section 5 concludes.

2.2 Construction of connectedness measures

In this section, we construct correlation-based connectedness measures that capture the structural changes of co-movement tendency in stock markets. Our bottom-to-top approach of building connectedness measures provide integrated perspectives on stock markets.

We build up connectedness measures based on the Pearson's pairwise correlation. We first define *connectedness of two stocks* as

$$C(i, j) = \rho_{i,j} \quad (1)$$

where $\rho_{i,j} = \frac{Cov(r_i, r_j)}{s.d.(r_i)s.d.(r_j)}$, r_i and r_j denote the returns of stock i and j , respectively. Using the pairwise correlation as a building block, we define the *connectedness between two groups of stocks*

$$C(A, B) := avg(\{C(i, j) | (i, j) \in (A, B), i \neq j\}) \quad (2)$$

where A and B are groups of stocks, and $avg(\cdot)$ is an averaging operator³. Note that the groups A and B are allowed to overlap (or, even identical) to each other by the condition $i \neq j$ which excludes trivial self-correlations for overlapping stocks.

We remark that connectedness for two groups can be used to define *total system connectedness* (*TSC*) for the entire stock market. Let V be a set of all stocks in the market, then $C(V, V)$ is the average of all possible pairwise correlations among all stocks in V , i.e. the average value of all off-diagonal elements in the $|V| \times |V|$ sized correlation matrix of all stocks return. [53] defined this quantity as ‘average correlation’ with all composite stocks of S&P 500 index, and provided the empirical evidence that it can predict the quarterly return of S&P 500 index.

Since our ultimate goal of construction of measures is to analyze multi-group structure in a stock market, we provide the notion of *partition* and *cluster*. For the set of entire stocks in the market, V , we consider disjoint subgroups of stocks in V . A *partition* P for the entire set V is defined as $P = \{V_1, V_2, \dots, V_k\}$, where $V = \bigcup_{c=1}^k V_c$, $V_i \cap V_j = \emptyset$, and V_i is a set of stocks in the i th cell. *Clustering* or *cluster analysis* on correlation matrix is a task of finding an adequate partition P for V such that stocks in same cells are highly correlated and stocks in different cells are less correlated.

As functions of a partition $P = \{V_1, V_2, \dots, V_k\}$, we define the marketwide connectedness measures⁴. We define the *inner-sector connectedness* (*INSC*) as an average of

³Taking an average of correlation elements is indeed not a simple task. One may simply think of taking an arithmetic average of all correlations under the consideration, but it is shown by [62] that this may result in a biased estimator. Pearson’s correlation is a cosine value of the angle that the two random vectors form in geometric sense, and thus it is not additive. Therefore, variance stabilization methods such as Fisher’s transformation [30][31] should be used as a means for estimating the average value. That is, with sample correlation elements of r_1, r_2, \dots, r_n under consideration, we define averaging operator $avg(r_1, r_2, \dots, r_n) := (exp(2\bar{z}) - 1) / (exp(2\bar{z}) + 1)$, where $\bar{z} = \frac{1}{n} \sum_{i=1}^n 0.5 \log \left(\frac{1+r_i}{1-r_i} \right)$. See appendix for more discussion on Fisher’s z -transformation.

⁴We interchangeably use the terms *cell* and *sector* throughout this paper. Although *cell* is the more technical and rigorous term, it bears the similar meaning to *sector* for stock markets and can be intuitively understood.

all pairwise correlations within the cells in the partition P .

$$INSC(P) := avg \left(\bigcup_{c=1}^k \{C(i, j) | (i, j) \in (V_c, V_c), i \neq j\} \right) \quad (3)$$

On the other hand, we define the *inter-sector connectedness* ($ITSC$) as an average of all correlations across the cells in the partition P .

$$ITSC(P) := avg \left(\bigcup_{c_1=1}^{k-1} \bigcup_{c_2=c_1+1}^k \{C(i, j) | (i, j) \in (V_{c_1}, V_{c_2})\} \right) \quad (4)$$

We note that $INSC(P)$ and $ITSC(P)$ can be combined to explain TSC , i.e. TSC is an weighted average of $INSC$ and $ITSC$.⁵ If a partition P exhibits a well-clustered structure in a correlation matrix, then $INSC(P)$ should be much higher than $ITSC(P)$, implying that the returns of stocks in a cell are much more independent to the stocks in other cells than to the stocks in the same cell. Thus, higher $INSC(P)$ accompanied with lower $ITSC(P)$ means highly evident cluster structure, and lower $INSC(P)$ with higher $ITSC(P)$ means less evident cluster structure. We define *modularity* as the difference of $INSC(P)$ and $ITSC(P)$ in order to capture the strength of partition in terms of how evident the cluster property is.

$$MOD(P) := INSC(P) - ITSC(P) \quad (5)$$

Clearly, identifying the most appropriate cluster structure is the foremost task in order to adopt cluster view on the financial market. With the universe of stock market, classification into cells by the Standard Industrial Classification (SIC) codes is a widespread and plausible practices. However, more technical approach based on correlation matrices is more suitable if one is interested in assessing the investment opportunity set solely based on the correlation structure.

⁵Suppose $P = \{V_1, \dots, V_k\}$ and $|V_i|$ denotes the number of element in set V_i . Then, $INSC(P)$ is an average of $a = \sum_{i=1}^k |V_i|(|V_i| - 1)/2$ correlations and $ITSC(P)$ is an average of $b = \sum_{c_1=1}^{k-1} \sum_{c_2=c_1+1}^k |V_{c_1}||V_{c_2}|$ correlations. Thus, $TSC(P) = \frac{a \cdot INSC(P) + b \cdot ITSC(P)}{a+b}$ where a, b are defined above.

Most clustering algorithms call a decision of the number of cells beforehand by researcher's, which can be seen as arbitrary. [59] proposes an objective method for identifying cluster structure without necessarily fixing the number of cells in prior, while still controlling the relative fineness and coarseness of cluster solutions. This allows cluster solutions to attain different number of cells depending on the variations of clustering tendency in different data. Thus, we adopt the *Modulated modularity clustering (MMC)* method⁶ proposed by [59]. The flexible number of cells that the solutions to MMC algorithm generates along with our modularity measure allow time-varying perspectives on the stock market. We appreciate the efficient nature of this method since the embedded eigenvalue decompositions in the algorithm is capable of processing high dimensional data.

2.3 *Marketwide empirical evidence*

In this section, we apply our marketwide connectedness measures to the space of 60 major U.S. stocks in order to investigate the characteristics of investment opportunity set. We also draw the relationship of the measures to the cycle of economy. This section uses the 60 major stocks⁷ for 10 years (1/1/2002-12/31/2011) picked from the top of Fortune 500 list published in 2012, which is ranked by the operating revenue in the year of 2011. The full list can be found at Table 1.

2.3.1 Entire period of 2002-2011

We first investigate the correlation structure of the 60 stocks daily return on the entire period from 1/1/2002 to 12/31/2011. Conducting cluster analysis with the MMC

⁶This clustering algorithm is discussed more in the appendix.

⁷We use CRSP (The Center for Research in Security Prices) database provided by WRDS (Wharton Research Data Services). We excluded companies without the common shares (WRDS share code 10 or 11) and whose stocks have missing daily return record during the 10 years period.

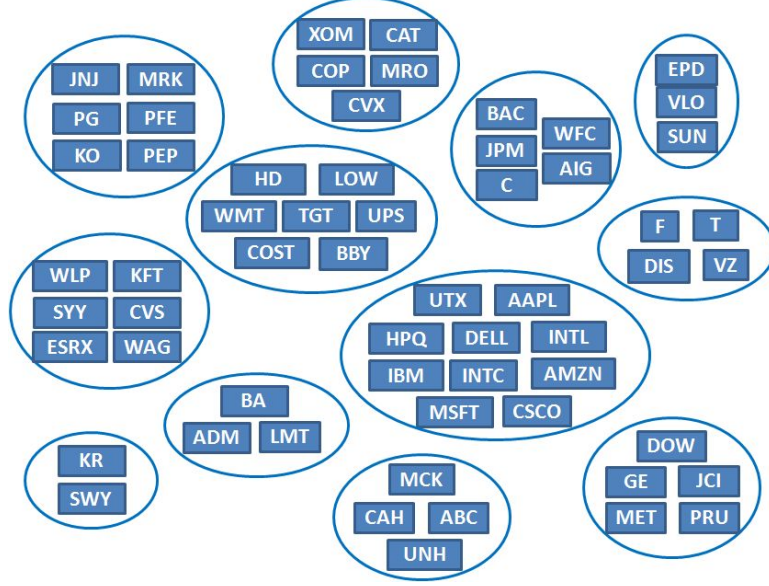


Figure 1: Cluster analysis with 10 years returns of 60 major stocks by MMC algorithm

algorithm on the sample correlation matrix of the 60 stocks' daily return during the ten years generates the partition of 12 cells in Figure 1. In Figure 1, the stocks with mutual high correlations are grouped as same cells. The partition is not exactly same as SIC codes presented in Table 1. However, this partition should be more meaningful observation if a market participant is solely interested in correlation structures to make investment decisions.

Figure 2 provides zoom-in view of the three cells (2^{nd} , 9^{th} , and 11^{th} cells in Table 1) with each numeric value⁸ representing the *connectedness between two groups* of (2). Connectednesses within the same cells (drawn by the blue solid lines) tend to be high and are components of $INSC(P)$. On the other hand, connectednesses across the different cells (drawn by the red dotted lines) tend to be low and are components of $ITSC(P)$.

Figure 3 provides a 12 by 12 dimensional connectedness *heat-map* matrix, $[C(V_i, V_j)]_{1 \leq i, j \leq 12}$,

⁸Specifically, $C(V_2, V_2) = 0.50$, $C(V_9, V_9) = 0.55$, $C(V_{11}, V_{11}) = 0.50$, $C(V_2, V_9) = 0.39$, $C(V_2, V_{11}) = 0.41$, and $C(V_9, V_{11}) = 0.44$

Table 1: Tabular view of Figure 1 with descriptions. The stocks are chosen based on 2012 Fortune 500 ranking

Cell	Ticker	Company Name	Revenues (\$ millions) (in 2011)	SIC code	SIC description
1	CVS	CVS Caremark	107,750	5912	Retail-Drug Stores and Proprietary Stores
1	WAG	Walgreen	72,184	5912	Retail-Drug Stores and Proprietary Stores
1	SYX	Sysco	39,324	5141	Wholesale-Groceries, General Line (merchandise)
1	ESRX	Express Scripts Holding	46,128	8093	Services-Specialty Outpatient Facilities, NEC
1	KFT	Kraft Foods	54,365	2052	Cookies & Crackers
1	WLP	WellPoint	60,711	6324	Hospital & Medical Service Plans
2	KO	Coca-Cola	46,542	2086	Bottled & Canned Soft Drinks & Carbonated Waters
2	PEP	PepsiCo	66,504	2086	Bottled & Canned Soft Drinks & Carbonated Waters
2	PG	Procter & Gamble	82,559	2841	Soap and Other Detergents
2	PFE	Pfizer	67,932	2834	Pharmaceutical Preparations
2	JNJ	Johnson & Johnson	65,030	2834	Pharmaceutical Preparations
2	MRK	Merck	48,047	2834	Pharmaceutical Preparations
3	GE	General Electric	147,616	3511	Turbines and Turbine Generator Sets
3	DOW	Dow Chemical	59,985	2821	Plastic Materials, Synth Resins & Nonvulcan Elastomers
3	JCI	Johnson Controls	40,833	2531	Public Bldg & Related Furniture
3	MET	MetLife	70,641	6311	Life Insurance
3	PRU	Prudential Financial	49,045	6311	Life Insurance
4	XOM	Exxon Mobil	452,926	2911	Petroleum Refining
4	COP	ConocoPhillips	237,272	2911	Petroleum Refining
4	CVX	Chevron	245,621	2911	Petroleum Refining
4	MRO	Marathon Petroleum	73,645	2911	Petroleum Refining
4	CAT	Caterpillar	60,138	3531	Construction Machinery & Equip
5	MSFT	Microsoft	69,943	7370	Services-Computer Programming, Data Processing, Etc.
5	DELL	Dell	62,071	3570	Computer & office Equipment
5	IBM	International Business Machines	106,916	3571	Electronic Computers
5	AAPL	Apple	108,249	3571	Electronic Computers
5	UTX	United Technologies	58,190	3724	Aircraft Engines & Engine Parts
5	HPQ	Hewlett-Packard	127,245	3571	Electronic Computers
5	INTC	Intel	53,999	3679	Electronic Components, NEC
5	SCO	Cisco Systems	43,218	3674	Semiconductors & Related Devices
5	INTL	INTL FCStone	75,498	6211	Security Brokers, Dealers & Flotation Companies
5	AMZN	Amazon.com	48,077	7370	Services-Computer Programming, Data Processing, Etc.
6	CAH	Cardinal Health	102,644	5122	Wholesale-Drugs, Proprietaries & Druggists' Sundries
6	MCK	McKesson	112,084	5122	Wholesale-Drugs, Proprietaries & Druggists' Sundries
6	ABC	AmerisourceBergen	80,218	5122	Wholesale-Drugs, Proprietaries & Druggists' Sundries
6	UNH	UnitedHealth Group	101,862	6324	Hospital & Medical Service Plans
7	WFC	Wells Fargo	87,597	6021	National Commercial Banks
7	JPM	J.P. Morgan Chase & Co.	110,838	6712	Bank Holding Companies
7	BAC	Bank of America Corp.	115,074	6021	National Commercial Banks
7	AIG	American International Group	71,730	6331	Fire, Marine & Casualty Insurance
7	C	Citigroup	102,939	6021	National Commercial Banks
8	SUN	Sunoco	45,765	2911	Petroleum Refining
8	VLO	Valero Energy	125,095	2911	Petroleum Refining
8	EPD	Enterprise Products Partners	44,313	4922	Natural Gas Transmission
9	TGT	Target	69,865	5331	Retail-Department Stores
9	WMT	Wal-Mart Stores	446,950	5331	Retail-Department Stores
9	LOW	Lowe's	50,208	5211	Retail-Lumber & Other Building Materials Dealers
9	HD	Home Depot	70,395	5211	Retail-Lumber & Other Building Materials Dealers
9	BBY	Best Buy	50,272	5731	Retail-Radio, TV & Consumer Electronics Stores
9	COST	Costco Wholesale	88,915	5330	Food & Drug Retailers
9	UPS	United Parcel Service	53,105	4215	Courier Services, Except By Air industry
10	ADM	Archer Daniels Midland	80,676	2075	Soybean Oil Mills
10	BA	Boeing	68,735	3721	Aircraft
10	LMT	Lockheed Martin	46,692	3764	Space Propulsion Units and Parts
11	F	Ford Motor	136,264	3711	Motor Vehicles & Passenger Car Bodies
11	DIS	Walt Disney	40,893	7996	Services-Amusement Parks
11	VZ	Verizon Communications	110,875	4813	Telephone Communications (No Radiotelephone)
11	T	AT&T	126,723	4813	Telephone Communications (No Radiotelephone)
12	KR	Kroger	90,374	5411	Retail-Grocery Stores
12	SWY	Safeway	43,630	5411	Retail-Grocery Stores

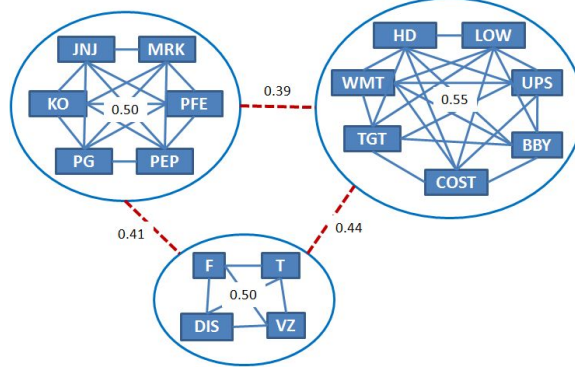


Figure 2: Partial zoom-in for three cells in Figure 1. The blue solid lines indicate correlations within cells and the red dotted lines indicate correlations across the cells.

where each element corresponds to the connectedness between the groups. The reduced dimensional heat-map matrix has its own advantage alone in viewing the entire market. It allows to view the correlation matrix in reduced dimension of 12, which is much more manageable than the bigger size original correlation matrix. Figure 3 shows the clear distinction between the level of connectedness within the same cells (diagonal elements in the matrix) and the level of connectedness across the different cells (off-diagonal elements in the matrix). The diagonal elements in the heat map matrix are components of $INSC(P)$ and the off-diagonal elements are the components of $ITSC(P)$. The modularity, $MOD(P)$, measures the difference of the two, and high value of $MOD(P)$ corresponds to highly evident cluster structure that can be seen as the noticeable contrast of darkness in on- and off-diagonal elements in Figure 3.

2.3.2 Subperiods of bull/bear markets

Although the correlation structure during the entire period showed an evident cluster structure, fluctuating market conditions during subperiods may exhibit different tendencies in the cluster structure. We intentionally choose four subperiods of which the two of them correspond to bull markets and the other two of them correspond to bear markets, and observe how the market structure changes based on the fixed

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.37	0.39	0.39	0.38	0.32	0.38	0.34	0.30	0.39	0.35	0.36	0.33
2	0.39	0.50	0.41	0.44	0.35	0.38	0.35	0.31	0.39	0.38	0.41	0.34
3	0.39	0.41	0.58	0.52	0.43	0.37	0.56	0.41	0.46	0.43	0.47	0.33
4	0.38	0.44	0.52	0.72	0.43	0.38	0.43	0.57	0.42	0.47	0.46	0.35
5	0.32	0.35	0.43	0.43	0.45	0.30	0.38	0.31	0.41	0.36	0.42	0.29
6	0.38	0.38	0.37	0.38	0.30	0.51	0.31	0.32	0.34	0.33	0.33	0.32
7	0.34	0.35	0.56	0.43	0.38	0.31	0.66	0.36	0.41	0.36	0.42	0.31
8	0.30	0.31	0.41	0.57	0.31	0.32	0.36	0.56	0.32	0.37	0.35	
9	0.39	0.39	0.46	0.42	0.41	0.34	0.41	0.32	0.55	0.37	0.44	0.36
10	0.35	0.38	0.43	0.47	0.36	0.33	0.36	0.37	0.37	0.42	0.39	0.28
11	0.36	0.41	0.47	0.46	0.42	0.33	0.42	0.35	0.44	0.39	0.50	0.33
12	0.33	0.34	0.33	0.35	0.29	0.32	0.31		0.36	0.28	0.33	0.66

Figure 3: Heat map matrix representation of correlation structure in Figure 1 for 2002-2011. The difference of inner cells correlations and inter cells correlations are noticeable.

partition of Figure 1.

Figure 4 presents the heat-map matrices for the four subperiods, and Figure 5 draws for the particular three cells (2^{nd} , 9^{th} , and 11^{th} cells in Table 1 again) for the same period. In both figures, the two top diagrams correspond to the bullish subperiods⁹, and the two bottom diagrams correspond to bearish subperiods¹⁰.

We remark two important observations on the subperiods by comparing the top diagrams and the bottom diagrams in Figure 4 and 5. First, the overall correlation levels hike in transitions from bull markets to bear markets. This phenomena were noted by many practitioners and studies such as [12][55]. The high correlation levels in bear market is a big warning signal to investors who may merely use the historical correlation matrices for the purpose of managing portfolio risks and devising diversification strategies based on the Markowitz' Mean-Variance framework. In a nutshell, the level of correlation changes over time and the bad news is that it tends to change

⁹Upper Left: 03/19/2009-06/11/2009 with S&P 500 39.7% increase; Upper Right: 06/11/2009-04/15/2010 with S&P 500 28.3% increase

¹⁰Lower Left: 04/15/2010-07/02/2010 with S&P 500 15.7% decrease; Lower Right: 07/01/2011-08/10/2011 with S&P 500 16.3% decrease

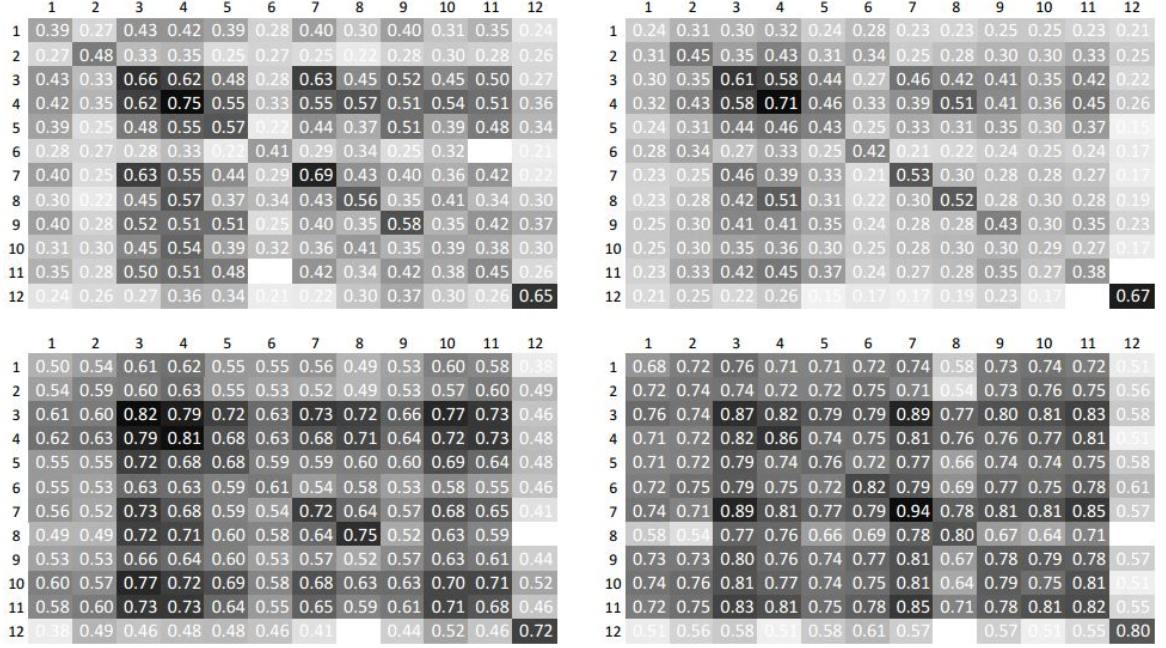


Figure 4: Heat map representation for a few bull/bear periods. Upper Left: 03/19/2009-06/11/2009, S&P 500 39.7% increase, The difference in average of diagonal and non-diagonal elements is 0.179. Upper Right: 06/11/2009-04/15/2010, S&P 500 28.3% increase, The difference in average of diagonal and non-diagonal elements is 0.171. Lower Left: 04/15/2010-07/02/2010, S&P 500 15.7% decrease, The difference in average of diagonal and non-diagonal elements is 0.090. Lower Right: 07/01/2011-08/10/2011, S&P 500 16.3% decrease, The difference in average of diagonal and non-diagonal elements is 0.085.

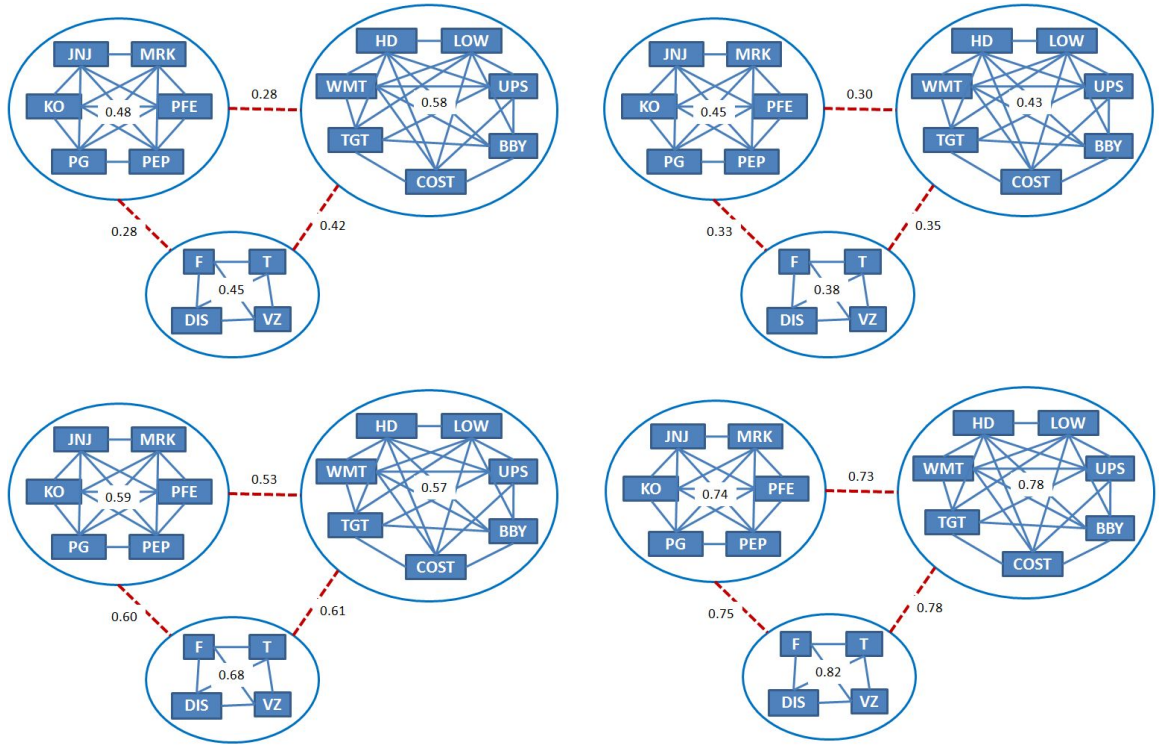


Figure 5: Partial zoom-in for three cells in Figure 4. The upper diagrams have higher modularity than the lower diagrams.

into investors' unfavorable direction when investor most desperately wish not to.

Second and more importantly to the point of our study, the cluster structure in the stressed subperiods is much less evident. Although both $INSC(P)$ and $ITSC(P)$ experience significant increases in the transitions from bull markets to bear markets, the level of increase in $ITSC(P)$ (represented in off-diagonal elements in the matrices of Figure 4 and illustrated by red dotted lines in the diagrams of Figure 5) is much higher than the level of increase in $INSC(P)$ (represented in on-diagonal elements in the matrices of Figure 4 and illustrated by blue solid lines in the diagrams of Figure 5). This phenomena are also noticeable from Figure 4 in that the visual contrast of on-diagonal elements over off-diagonal elements in top diagrams are no longer observable in the bottom diagrams. In such structural changes of the market from bull market to bear market, the widespread sector-based diversification practices of allocating wealth into the different sectors no longer work because of the destruction of the cluster structure.

Should the stressed markets lead to destructions of existing cluster structure, one needs a new cluster structure to understand market better. However, redoing clustering analysis to find a new partition makes it harder, though maybe possible, to compare the new structure with the old structure in a quantitative way. Thus, we propose the marketwide connectedness measure, modularity, to quantify the deviation level of current structure from the fixed partition structure. The remarkable structural changes in the clustering tendency at the different market conditions are captured and quantified by the proposed measures of inner-sector connectedness ($INSC$), inter-sector connectedness ($ITSC$), and modularity (MOD), of which modularity is the ultimate measure that quantifies the intensity of the clustering tendency.

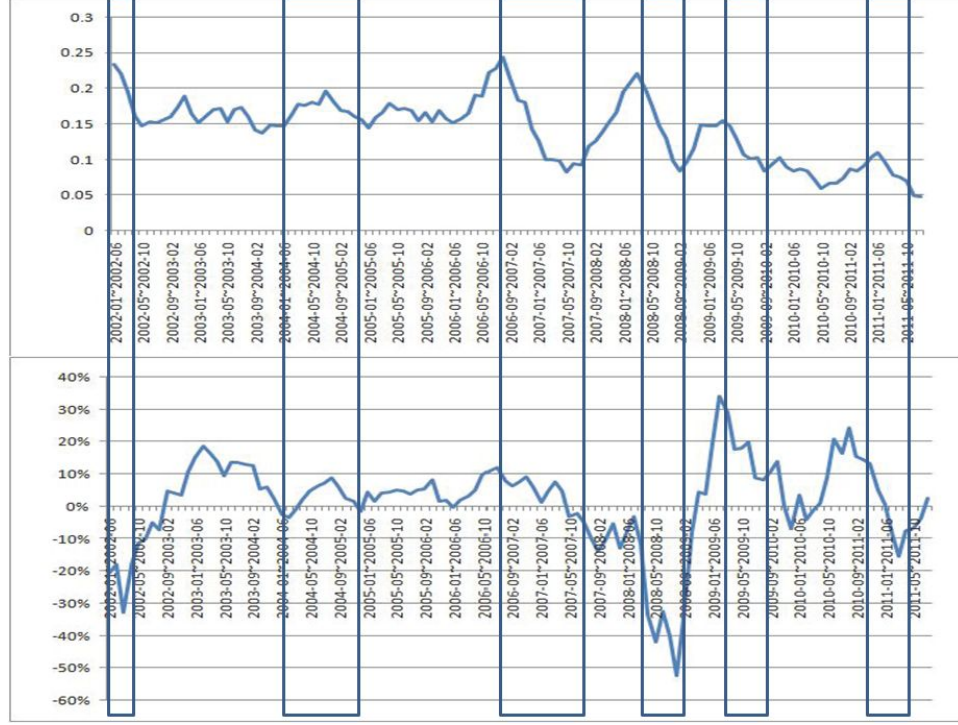


Figure 6: Top: The six months moving average series for monthly MOD measured based on cluster structure in Figure 1. Bottom: The six months moving average return for S&P 500. The two plots behave similarly for many subperiods, indicating that MOD can be a strong indicator for the cycle of economy.

2.3.3 The modularity measure and the cycle of the economy

Motivated by the observations on the four subperiods, we further investigate the time varying market structure of the entire period in light of the proposed measure modularity. We present the marketwide empirical evidence revealed from proposed measure by providing historical innovations of modularity with the return of S&P 500 index.

The top plot in Figure 6 presents six months moving average time series of $MOD(P)$ where the partition P is fixed as shown in Figure 1. The bottom plot presents six months moving average returns of S&P 500 for the corresponding period. We note that the periods of 6 months negative return in bottom plot tends to coincide with the periods that MOD drops in the top plot. The highlighted rectangular boxes

are where the two plots particularly behaved in the similar way. Thus, modularity that measures the clustering tendency of the market structure can serve as a indicator for the cycle of economy as well as an indicator for the quality of investment opportunity set.

Given that the modularity is revealed as a market indicator and is able to provide qualitative understanding of structural changes in markets, we further investigate whether it is related to returns of individual stocks or portfolios in the following section.

2.4 Is MOD factor priced?

This section studies whether a stock's or a portfolio's sensitivity to the market modularity can explain its expected return. The sensitivity β_i^{MOD} is defined as the coefficient of market modularity in multiple linear regression where the other explanatory variables are established popular factors. We construct annually updating sensitivity-sorted decile portfolios based on the ranking of β_i^{MOD} of individual stocks. The excess returns of each decile portfolio are regressed on 1-,3-,4-factor models, and the existence of nonzero intercept varying across the decile portfolios substantiates the existence of modularity factor in the market. We ensure our findings by estimating the premium of modularity factor by generalized method of moments (GMM). We then present empirical result of modularity mimicking portfolio.

This section proceeds as follows. Section 2.4.1 briefly reviews the essence of 1-,3-,4-factor models. Section 2.4.2 proposes our asset pricing factor model. Section 2.4.3 outlines the construction procedure of sensitivity rank sorted decile portfolio. Using the portfolio constructed in section 2.4.3, section 2.4.4 shows that the return difference in decile portfolio is evident. That is, modularity factor generates market anomaly that are not explained by the established popular factors models. Finally,

section 2.4.5 shows that the return premium of top decile portfolio over the bottom decile portfolio is indeed explained by the introduction of modularity factor. The premium is estimated by Generalized Method of Momentum proposed by [34]. Section 2.4.6 shows that introducing MOD spread portfolio, defined as the long-short portfolio of top and bottom decile portfolio, expands the investment opportunity. Section 2.4.7 shows that *MOD* mimicking portfolio constructed by *minimum idiosyncratic procedure* can achieve similar enhancement in investment opportunity.

2.4.1 Classical Model

The traditional one factor asset pricing model, also known as the Capital Asset Pricing Model (CAPM), claimed that returns of assets or portfolios can be solely described by the return of the whole market. [26][27] observed that the stocks of small caps and stocks with high book-to-market ratio tend to yield higher returns. Adding this observation to linear factor equation gives (6) for the return of asset i

$$r_i - r_f = \alpha_i + \beta_i^M MKT + \beta_i^S SMB + \beta_i^H HML + \epsilon_i \quad (6)$$

where r_i is return of asset i , r_f is risk free rate, MKT is return of market minus risk free rate, and SMB and HML are factors that represent returns of the corresponding replicating portfolios. α_i is abnormal return of asset i , and β_i^M, β_i^S , and β_i^H are the sensitivities of asset i to the corresponding factors. SMB stands for small minus big in market capitalization and measures the excess return of small cap stocks over large cap stocks. HML stands for high minus low in book-to-market ratio and measures the excess return of growth stocks.

Given the three factors in multiple regression (6), an asset or portfolio's expected return r_i is determined by its coefficient to the three factors where each of coefficients β_i^M, β_i^S , and β_i^H quantifies the exposure or the sensitivity to each factor. Specifically, the pricing model assumes expected return should be expressed in terms of linear

combination of betas. Then it becomes,

$$E(r_i - r_f) = \beta_i^M \lambda_M + \beta_i^S \lambda_S + \beta_i^H \lambda_H \quad (7)$$

where λ_M , λ_S , and λ_H are market premium for each additional unit of exposure to corresponding factors. By taking expectation on both hand sides of (6) and comparing to (7), it is clear that $\lambda_M = E(MKT)$, $\lambda_S = E(SMB)$ and $\lambda_H = E(HML)$ where the true value of abnormal return α_i is assumed to be zero. Thus, price of stock i is determined by its coefficient to factors and market expectation of the level of factors.

Since wide acceptances of the three factors model, many attempts have been made to expand the model. [14] observed that “past loser” portfolio tend to underperform than “past winner”, and proposed the momentum factor. The three factors with momentum factor, MOM , is now widely accepted as the four factor model. With MOM factor, factor model equation (6) would include $+\beta_i^{MOM} MOM$ term. [51] applies the observation that “less liquid stocks should yield higher returns.” After constructing so-called liquidity measure in market level, they showed that the stocks with higher sensitivity to market liquidity measure, which are effectively more illiquid stocks, should yield higher expected return.

The flow and method of our analysis is very similar¹¹ to that of liquidity factor evidence in [51] in that our modularity is not measured in return, but constructed as a non-traded factor. [51] first estimated the sensitivity of each stock to the liquidity measure in multiple linear regression. Based on the ranking of the sensitivities of stocks, they constructed the standard decile portfolios where the portfolios are updated every year. The top decile and the bottom decile portfolio exhibit significant difference in their returns which supports the existence of the additional factor of market liquidity.

Factor model validation via constructions of rank sorted portfolios is rather norm

¹¹Specifically, followings are similar: the construction process of decile portfolio, testing against four factor model, formulation detail for the GMM test, and the analysis of mean variance framework.

than exception, and is used by the most of factor model studies [27][14][51]. One should not run separate time series regressions for each stock because individual stocks may not have stable sensitivities on the factor over time. This is why the factor model studies allow stocks to make transitions from one decile portfolio to another, typically in annual basis.

2.4.2 Model

We define β_i^{MOD} as the coefficient of modularity factor MOD_t in a multiple linear regression (8).

$$r_{i,t} - r_{f,t} = \beta_i^0 + \beta_i^{MOD} MOD_t + \beta_i^M MKT_t + \beta_i^S SMB_t + \beta_i^H HML_t + \epsilon_{i,t} \quad (8)$$

where $r_{i,t}$ is an asset i 's return, $r_{f,t}$ is an risk free return, MKT_t , SMB_t , HML_t are the Fama-French three factors, and β_i^M , β_i^S , β_i^H are the corresponding coefficients. Since the modularity measure MOD_t is not expressed in return, the intercept term β_i^0 has a different meaning than the alpha as abnormal return in 1-,3-,4-factor models. We use historical estimate of β_i^{MOD} to sort stocks and construct decile portfolios.

2.4.3 Procedure to construct decile portfolios

Step 1. At 1/1/1992, we will construct decile portfolios for next 12 months.

Step 2. Collect historical data of all stocks.

We collect historical stock prices from the Center for Research in Security Prices (CRSP). The historical data of the Fama French three factors and momentum factor is obtained from Kenneth R. French - Data Library.¹² We use all ordinary common stocks (CRSP share code of 10 and 11) on the NYSE, AMEX, and NASDAQ. Stocks

¹²http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

with price below \$5 or above \$1000 are excluded.¹³

Step 3. Collect residuals of recent history

First, regress all stocks by three factor model equation using past monthly return of 60 months¹⁴.

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^M MKT_t + \beta_i^S SMB_t + \beta_i^H HML_t + \eta_{i,t} \quad (9)$$

Using estimated betas $\hat{\beta}_i^M$, $\hat{\beta}_i^S$, and $\hat{\beta}_i^H$ from (9), collect residual of each stock as follows

$$e_{i,t} = r_{i,t} - r_{f,t} - \hat{\beta}_i^M MKT_t - \hat{\beta}_i^S SMB_t - \hat{\beta}_i^H HML_t \quad (10)$$

Therefore, $e_{i,t}$ contains both alpha and noise that are not captured by the Fama-French three factors¹⁵. In this step, the stocks with missing record on any month in the 60 months are excluded.

Step 4. Construct historical MOD series

Based on the top 60 available stocks in Fortune 500 list published at the beginning of year 1992, which is ranked by annual operating revenue of the previous year¹⁶, conduct the cluster analysis by MMC algorithm with correlation matrix of daily return in recent 60 months¹⁷. Using the cluster solution of the algorithm, construct a monthly historical modularity of the same period. Specifically, we obtain the sample correlation matrix for each month based on daily returns of 60 stocks and apply the

¹³This data filtering criteria is same as [51].

¹⁴Indeed, we use 36 months of history for the first year, 48 months for the second year, and 60 months for the all other years. This criteria is consistent to [?]

¹⁵We only use three factors for estimating sensitivity and this is same process of [51]. We tested for four factors as well, but the results are very similar.

¹⁶We believe that this top portion of Fortune 500 is a good representative set for the domestic economy. The number of stocks are set to 60 for the computational efficiency of clustering algorithm. Also, considering too many stocks for constructing correlation matrix is not desirable because the number of variable should be moderated by the number of sample size for the stability of estimated values.

¹⁷Again, we use 36 months of historical data for the year of 1992, 48 months for the year of 1993, and 60 months for the all other years.

fixed partition structure to compute the modularity¹⁸.

Step 5. *Regress residual by modularity to estimate historical sensitivity*

Using the same 60 months period of data, we run following simple regression for each stock i :

$$e_{i,t} = \psi_i^0 + \psi_i^1 MOD_t + \pi_{i,t} \quad (11)$$

Then, $\hat{\psi}_i^1$ is the historical sensitivity of stock i to MOD factor.

Step 6. *Construct a decile portfolio*

Based on the ranking of $\hat{\psi}_i^1$, we construct the value-weighted decile portfolios (or, equal-weighted, whose results are presented in the appendix), i.e. $\hat{\psi}_i^1$ serves as ‘predicted modularity beta’ of stock i for the next twelve months.¹⁹ The top decile portfolio has stocks with lowest sensitivities.

Step 7. *Collect return for 12 months*

Using historical data, collect return of each decile portfolio for next 12 months. In case of missing record, assume that the return of the month was zero.

Step 8. *Continue to next year*

After 12 months, go to Step 1 and set the date to the next year and repeat Step 2-Step 7. Repeat this until it reaches 1/1/2011 so that we collect monthly return of decile portfolio for the 20 years from 1992 to 2011.

¹⁸For each month, there are only 18-23 trading days, and this number of observations is statistically insufficient to guarantee the stability of estimate based on correlation matrix. In order to boost the stability, we use 3 months moving average value of MOD instead of original monthly time series.

¹⁹[51] not only used historical betas to predict future beta, but also used forecast betas with other market variables. The result with forecast betas was slightly better. In our study, we simply set historical beta as predicted beta because we have no plausible candidate or conjecture on what characteristics should predict future modularity beta yet.

Table 2: Returns and alphas of the decile portfolio (January 1992 - December 2011)

	1	2	3	4	5	6	7	8	9	10	'1-10'
Return (p.a.)	13.99	10.23	10.29	11.41	10.06	8.21	8.48	11.11	7.72	6.4	7.59
s.d. (p.a.)	20.94	17.1	14.29	14.25	14.45	14.78	14.99	15.67	18.5	20.9	14.8
CAPM alpha (t-statistics)	4.25 (1.66)	1.43 (0.79)	2.51 (1.58)	3.54 (2.39)	2 (1.46)	0.25 (0.15)	0.35 (0.23)	2.6 (1.84)	-1.69 (-0.94)	-3.66 (-1.66)	7.92 (2.38)
3-factor alpha (t-statistics)	4.82 (1.99)	1.76 (0.97)	2.12 (1.41)	2.62 (1.95)	1.93 (1.48)	-0.41 (-0.32)	-0.45 (-0.33)	1.71 (1.33)	-1.94 (-1.07)	-2.35 (-1.15)	7.17 (2.13)
4-factor alpha (t-statistics)	5.76 (2.38)	1.94 (1.06)	2.21 (1.46)	3.33 (2.51)	1.86 (1.4)	-0.6 (-0.46)	0.6 (0.46)	1.98 (1.52)	-0.41 (-0.24)	-1.13 (-0.56)	6.89 (2.02)

2.4.4 Tests on decile portfolios — return difference

Our conjecture is that the decile portfolio sorted by the individual stocks's sensitivities to *MOD* shows significant differences in return. In this section, we investigate whether the difference of returns in the both ends of decile portfolio is significant and not explained by the other factor models.

Table 2 shows that sensitivity sorted decile portfolio creates systematic return difference that are not explained by existing factor models. The first row presents the annualized return and standard deviation of the each decile portfolio. The last column corresponds to the difference of '1' portfolio and '10' portfolio that is equivalent to the net zero investment portfolio where investors buy the first decile and sell the last decile by same amount. In the rows from the second to the fourth, we present the level of alphas with respect to the established factor models. Specifically, let R^j be the excess return of decile portfolio $j = 1, 2, \dots, 10$, and '1-10', then α^j from the following regression (12) is presented.

$$R_t^j = \alpha_j + B_j F_t + \eta_{j,t} \quad (12)$$

In (12), F_t is either 1-, 3-, or 4-factors at time t , corresponding to the second, the third, and the fourth row of Table 2, respectively. B_j is the vector for the sensitivities to the factors. From the last column of Table 2, the '1-10' portfolio exhibits 7.59% of annualized return and $\alpha_{1-10'}$ are all significantly positive ranging from 6.89%–7.92% of annual return depending on the 1-, 3-, 4-factor models.

We also test the hypothesis whether all alphas are jointly equal to zero where this

Table 3: Betas of decile portfolio sorted by predicted modularity betas. (January 1992 - December 2011)

	1	2	3	4	5	6	7	8	9	10	'1-10'
β_{MKT}	1.02	0.99	0.85	0.84	0.89	0.93	0.89	0.97	1.03	1.07	-0.05
(t-statistics)	(20.99)	(27.06)	(28.02)	(31.72)	(33.64)	(35.76)	(34.31)	(37.41)	(29.98)	(26.59)	(-0.71)
β^{SMB}	0.27	-0.14	-0.16	-0.1	-0.15	-0.27	-0.16	-0.1	0	0.16	0.11
(t-statistics)	(4.52)	(-3.01)	(-4.18)	(-2.94)	(-4.6)	(-8.43)	(-4.88)	(-3.02)	(-0.07)	(3.28)	(1.28)
β^{HML}	-0.21	-0.02	0.12	0.19	0.06	0.21	0.17	0.19	0.01	-0.32	0.11
(t-statistics)	(-3.33)	(-0.44)	(3)	(5.27)	(1.79)	(6.18)	(5.03)	(5.5)	(0.25)	(-6.11)	(1.22)
β^{MOM}	-0.1	-0.02	-0.01	-0.08	0.01	0.02	-0.11	-0.03	-0.16	-0.13	0.03
(t-statistics)	(-2.53)	(-0.67)	(-0.39)	(-3.5)	(0.37)	(0.98)	(-5.26)	(-1.34)	(-5.77)	(-3.98)	(0.54)

Table 4: Composition of portfolio returns by alpha and exposures to four factors. (January 1992 - December 2011)

	1	2	3	4	5	6	7	8	9	10	'1-10'
Excess return (p.a.)	10.86	7.1	7.16	8.28	6.93	5.08	5.35	7.98	4.59	3.27	7.59
α	5.76	1.94	2.21	3.33	1.86	-0.6	0.6	1.98	-0.41	-1.13	6.89
$\beta^{MKT} * \overline{MKT}$	5.9	5.77	4.94	4.9	5.18	5.4	5.16	5.64	6.01	6.18	-0.28
$\beta^{SMB} * \overline{SMB}$	0.77	-0.39	-0.45	-0.28	-0.43	-0.77	-0.45	-0.28	-0.01	0.46	0.31
$\beta^{HML} * \overline{HML}$	-0.93	-0.09	0.52	0.81	0.27	0.93	0.75	0.82	0.05	-1.41	0.48
$\beta^{MOM} * \overline{MOM}$	-0.65	-0.13	-0.06	-0.49	0.05	0.13	-0.72	-0.18	-1.05	-0.84	0.19

test is proposed by [33]. For the null hypothesis of $\alpha_1 = \alpha_2 = \dots = \alpha_{10} = 0$, we found that this hypothesis is rejected at 5% significance level in CAPM and 3-factor alphas, and 1% significance level in 4-factor alphas.²⁰

In order to reassure that the differences in returns are not captured by the established factors, Table 3 presents B_j in (12) where 4×1 vector F_t corresponds to the four factors at time t . The last column in Table 3 shows that the sensitivities of the '1-10' portfolio is statistically close to zeros (The t-statistics are all in statistically non-significant range), meaning that '1-10' portfolio is neutral to the four factors.

As an extension to Table 3, Table 4 presents how the returns of decile portfolios can be decomposed by the exposures to the four factors. By multiplying each estimated betas in Table 3 by the historical average value of each corresponding factor, Table 4 is obtained. Specifically, by taking average of both sides in (12) over the 240 months,

²⁰For equal-weighted decile portfolios, CAPM and 4-factor alphas are rejected at 1% level and 3-factor alphas are rejected at 5% level.

we have

$$\overline{R^j} = \alpha_j + \beta_j^{MKT} \overline{MKT} + \beta_j^{SMB} \overline{SMB} + \beta_j^{HML} \overline{HML} + \beta_j^{MOM} \overline{MOM} \quad (13)$$

where \overline{factor} implies the historical average value of the factor. As Table 3 indicates that ‘1-10’ portfolio has only insignificant exposure to the four factors, Table 4 confirms that those exposures, if any, do very little contribution to the return difference of 7.59%. Thus, we conclude that the ‘1-10’ portfolio is fairly factor neutral for our purpose, and the existing factor models fail to capture the systematic difference in the sensitivity sorted decile portfolio.

2.4.5 Estimation of premium on modularity risk

Now that the modularity decile portfolios exhibit the systematic difference in returns that are not explained by the popular four factors, it still remains to show that the variations are indeed captured by the suggested new factor, MOD . Therefore, we estimate the modularity risk premium with all 10 decile portfolios in this section.

We rewrite the multivariate regression as follows (only notationally modified from (8)),

$$R_t = \beta_0 + \mathbf{B}\mathbf{F}_t + \beta^{MOD} MOD_t + e_t \quad (14)$$

where R_t is 10×1 vector containing the excess return on the decile portfolios; \mathbf{F}_t is 4×1 vector of realized factor, MKT, SMB, HML, and MOM; \mathbf{B} is 10×4 matrix of loadings; and β_0 and β^{MOD} are 10×1 vectors. We also consider using only Fama-French three factors excluding the momentum factor. Assuming that each decile portfolio is linearly priced by the returns’ sensitivity to the tradable factors \mathbf{F} and non-tradable modularity factor MOD , the expected excess return of each decile portfolio can be expressed as a linear combination of loading \mathbf{B} and β^{MOD} as in (15), where λ_F and λ_{MOD} are the price of the exposure to corresponding factor.

$$ER_t = \mathbf{B}\lambda_F + \beta^{MOD}\lambda_{MOD} \quad (15)$$

Note that $\lambda_F = E\mathbf{F}_t$ because \mathbf{F}_t represents the return of tradable portfolios. However, $\lambda_{MOD} \neq E[MOD]$ because MOD is a non-tradable factor.²¹

Based on (15), λ_{MOD} indicates the unit price of exposure to modularity factor, and the premium²² of ‘1’ portfolio over ‘10’ portfolio due to the sensitivity difference is equal to $(\beta_1^{MOD} - \beta_{10}^{MOD}) \lambda_{MOD}$. Taking expectation on the both hand sides of (14), then equating its right hand side to the right hand side of (15) gives following equation.

$$\beta_0 = \beta^{MOD}(\lambda_{MOD} - E\mathcal{MOD}_t) \quad (16)$$

Our hypothesis is that λ_{MOD} should be negative. Recall that in (15), each element of λ_F is known to be positive. They are positive because the high sensitivities to the four factor implies higher risk, where each of the four factors being high means more risky market condition. On the other hand, high modularity factor generally corresponds to less volatile market condition as shown in Figure 4. Thus, low sensitivity to the modularity factor should imply higher risk and higher expected return. Hence, we expect λ_{MOD} to be negative and $(\beta_1^{MOD} - \beta_{10}^{MOD})\lambda_{MOD}$ to be positive as the first decile portfolio contains the stocks with the smallest modularity beta.

In order to test our hypothesis, we estimate λ_{MOD} and β^{MOD} using the generalized method of moments (GMM) of [34]. Our specific procedure is same as the test on the liquidity factor by [51]. Let θ denote the set of unknown parameters: $\lambda_{MOD}, \beta^{MOD}, \mathbf{B}$, and $E(\mathcal{MOD}_t)$. We use the moment conditions, given as $Eg(\theta) = 0$, where $g(\theta) :=$

²¹In classical four factor model, the pricing equation should be a linear combination of loadings such as $Er_t = \beta^{MKT}\lambda_{MKT} + \beta^{SMB}\lambda_{SMB} + \beta^{HML}\lambda_{HML} + \beta^{MOM}\lambda_{MOM}$. In this case, $\lambda_F = E\mathbf{F}_t$ holds for all factors. However, this is not the case for our model because of β_0 term in factor model equation (14). This is essentially related to MOD being a non-tradable factor.

²²This term is derived from (15). Since $ER_t^1 = \mathbf{B}_1\lambda_F + \beta_1^{MOD}\lambda_{MOD}$ and $ER_t^{10} = \mathbf{B}_{10}\lambda_F + \beta_{10}^{MOD}\lambda_{MOD}$ where \mathbf{B}_1 and \mathbf{B}_{10} are the first and last rows of 10×4 matrix \mathbf{B} , it reduces to $E(R_t^1 - R_t^{10}) = (\mathbf{B}_1 - \mathbf{B}_{10})\lambda_F + (\beta_1^{MOD} - \beta_{10}^{MOD})\lambda_{MOD}$. From here, $\mathbf{B}_1 - \mathbf{B}_{10} = 0$ is assumed because factor model assumes that factors are independent to each other. This assumption is backed by the last columns of Table 3 and Table 4 where ‘1-10’ portfolio has negligible loadings on the four factors.

$$(1/T) \sum_{t=1}^T f_t(\theta),$$

$$f_t(\theta) = \begin{pmatrix} h_t \otimes e_t \\ MOD_t - EMOD_t \end{pmatrix}, \quad (17)$$

$$h_t = \begin{pmatrix} 1 \\ \mathbf{F}_t \\ MOD_t \end{pmatrix}, \quad (18)$$

$$\begin{aligned} e_t &= r_t - \beta_0 - \mathbf{B}\mathbf{F}_t - \beta^{MOD} MOD_t \\ &= r_t - \beta^{MOD} [\lambda_{MOD} - EMOD_t] - \mathbf{B}\mathbf{F}_t - \beta^{MOD} MOD_t \text{ by (16)} \end{aligned} \quad (19)$$

For the estimation process, we use the standard two steps approaches suggested by [18].²³ First, we find an estimator of θ that minimizes $g(\theta)'g(\theta)$, and let W as the inverse matrix of $(1/T) \sum_{t=1}^T f_t(\theta)f_t'(\theta)$. For the second stage, we find an estimator of θ again that minimizes $g(\theta)'Wg(\theta)$. The estimated premium λ_{MOD} and the difference in return $(\beta_1^{MOD} - \beta_{10}^{MOD})\lambda_{MOD}$ are reported in Table 5. $(\beta_1^{MOD} - \beta_{10}^{MOD})\lambda_{MOD}$ is reported as annual percentage return. For the asymptotic t-statistics, we use bootstrapping method for given data set of decile portfolio's return. Bootstrapping repetitions were conducted for 1,000 times.

In the previous section, we constructed the time series of MOD in 60-month rolling basis. However, we need a consistent time series measure for MOD in ex-post test such as GMM. Therefore, we therefore concatenate the MOD series into 240 months vector with scaling method same as the one for the consumer price index. That is, whenever there is an update of the basket, we scale new MOD series in the way that the new MOD at the beginning month is same as the old MOD at the ending month²⁴.

²³The first step solves (I) $\min g(\theta)'g(\theta)$ subject to $Eg(\theta) = 0$, then it proceeds to solve (II) $\min g(\theta)'Wg(\theta)$ subject to $Eg(\theta) = 0$ where W is an inverse matrix of $(1/T) \sum_{t=1}^T f_t(\theta)f_t'(\theta)$ using solution θ from (I). It is known that $Tg(\theta)'Wg(\theta) \sim \chi^2(\# \text{ of moments} - \# \text{ of parameters})$ where θ

Table 5: Estimated premium of modularity factor by GMM (with using three/four factors)

	λ_{MOD} ($t - statistics$)	$(\beta_1^{MOD} - \beta_{10}^{MOD})\lambda_{MOD}$ ($t - statistics$)	chi-square statistics
Using the four factors	-0.0314 (-113.58)	1.96% (2.20)	110.9
Using the three factors	-0.0316 (-398.0)	2.30% (2.16)	94.0

In Table 5, the unit price of modularity factor, λ_{MOD} is estimated as a significantly negative value, and the corresponding return premium of the ‘1’ portfolio over ‘10’ portfolio, $(\beta_1^{MOD} - \beta_{10}^{MOD})\lambda_{MOD}$, is estimated as 1.96%-2.30% depending on the basis of three or four factor models. Thus, even when the premium is estimated by all decile portfolios, the contribution of sensitivity sorting still remains valid.²⁵ The chi-square statistics for overall regression model is highly significant ($> 99\%$ level), where $g(\theta^*)'Wg(\theta^*)$ serves as a chi-square statistic and θ^* is the minimizer of the second stage optimization.²⁶

2.4.6 Investment perspectives (1) - with MOD basis portfolio of the difference in both decile ends

Under the assumptions of linear factor based pricing model, investors only need to construct basis factor portfolios and consider investing on the basis portfolios. This section tests whether the ‘1-10’ portfolio constructed in section 2.4.3 expands the

is the solution to (II).

²⁴We understand the possibility of not smooth transitions of basket that may lose some implications of modularity factor. However, this is common challenge to cope with measures whose base is updated periodically, such as many production and consumer indices. We conducted the GMM test with fixed basket from a few fixed years and the results were similar.

²⁵The return difference is not as high as 6.89% level in Table 2. We believe that this is due to having concatenation of modularity through 20 years that may not perfectly capture changes in the state of economy. When constructing the decile portfolio in previous section, we annually updated the basket of 60 stocks to create the time series of MOD .

²⁶The degree of freedom of the chi-square statistics is 1, and the 99% threshold is 6.635.

Table 6: Weights in the ex-post tangency portfolio and monthly Sharpe ratios (January 1992 - December 2011)

Number of instruments	<i>MKT</i>	<i>SMB</i>	<i>HML</i>	<i>MOM</i>	<i>MOD_S</i>	Sharpe Ratio
1	100					0.109
2	64.3	35.7				0.116
2	43.1		56.9			0.176
2	55.2			44.8		0.172
2	42.1				57.9	0.189
3	26.6	25.5	47.8			0.197
4	26.3	14.3	38.7	20.7		0.252
4	22.6	17.4	35.4		24.7	0.242
5	23.5	10.3	31.3	17.2	17.7	0.286

investment opportunity set. We use the standard Mean-Variance tangent portfolio analysis, similar to the approach employed by [51].

Unlike the other four established factors, the time series of *MOD* is not the return of portfolio. Thus, we use the ‘1-10’ portfolio as alternative to the original *MOD* factor to the investment universe of the four factors, and test whether doing so would provide better investment opportunity set. We notate this annually updated tradable mimicking portfolio ‘1-10’ as *MOD_S*, where the subscript *S* implies spread of the top and the bottom decile portfolio.

Table 6 presents the maximum ex-post Mean-Variance efficient portfolio using several different combinations of the five factor portfolios under consideration. We use historical returns of 4 factors and ‘1-10’ portfolio *MOD_S* for the 20 years from January 1992 to December 2011, and obtained the weight of the ex-post tangent portfolio by the Mean-Variance optimization formula $w = \frac{\Sigma^{-1}(r-r_f)}{1^t \Sigma^{-1}(r-r_f)}$ where r and Σ are the historical average return vector and the historical covariance matrix of the factor portfolios. In Table 6, investing only on *MKT* factor based on one factor model (CAPM) yields monthly Sharpe ratio of 0.109. If an investor were to pick another factor in addition to *MKT*, adding *MOD_S* will result in the monthly Sharpe ratio of 0.189, higher than that of adding any other factor. Overall, Table 6 suggests that

Table 7: Cross-correlation matrix of the five factors including MOD_S (January 1992 - December 2011)

	MKT	SML	HML	MOM	MOD_s
MKT	1	0.24	-0.24	-0.26	-0.06
SML		1	-0.35	0.09	0.05
HML			1	-0.14	0.06
MOM				1	0.05
MOD_s					1

Table 8: Performance of enhanced market index portfolio (January 1992 - December 2011)

	MKT	$MKT + 10\% MOD_s$	$MKT + 20\% MOD_s$
Return (p.a.)	8.94%	9.69%	10.45%
Std (p.a.)	15.46%	15.44%	15.57%
SR (monthly)	0.109	0.123	0.136

the MOD_S should be an attractive investment instrument regardless of the current factor usages.

Table 7 presents the correlation matrix of the four factors and MOD_S . The MOD_S tends to be highly independent of existing factors, which is already noted from Table 2 - Table 4. MOD_S has a good average annual return of 7.59% for being net zero investment portfolio, and it has very low correlation of -0.07 with respect to the MKT . This makes MOD_S as a strong candidate to add to portfolios that have higher exposures to MKT factor, such as market index funds.

Table 8 exhibits the enhancing effect of market index portfolio by adding 10% or 20% portion of MOD_S . From MKT , adding 10% portion of MOD_S would result in increase of annual return while still sustaining the similar level of standard deviation, thereby increasing monthly Sharpe ratio from 0.109 to 0.123. Adding 20% of MOD_S shows a similar enhancing effect of the market portfolio. Figure 7 confirms our finding that adding some portions of MOD_S to MKT result in stable enhancements. The

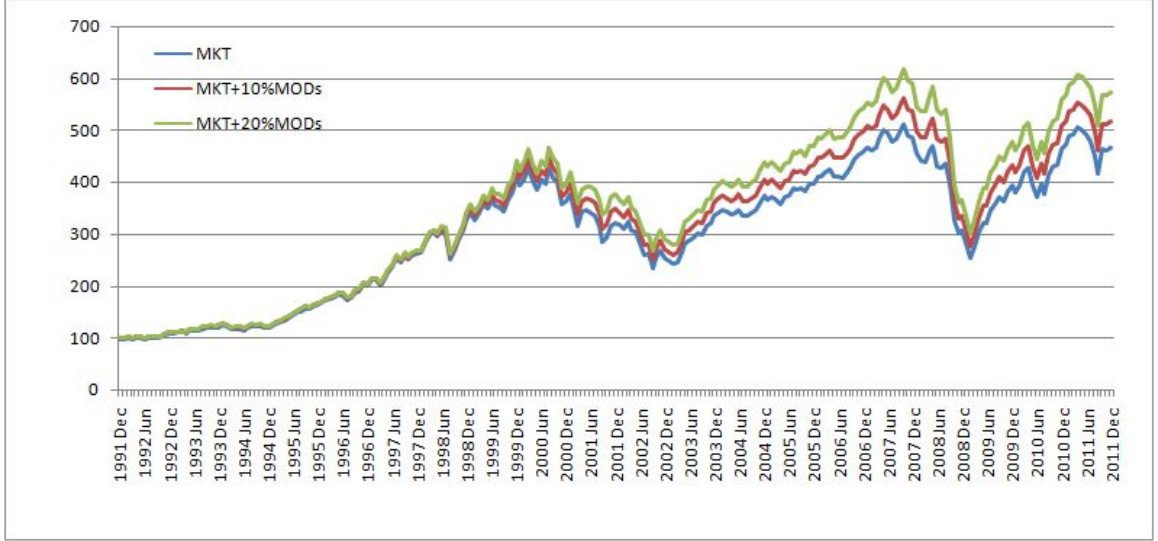


Figure 7: Cumulative wealth growth on enhancement scenarios by MOD_S

face value of the portfolio are assumed to be 100 at the beginning of the 20 years of investment horizon.

2.4.7 Investment perspectives (2) - with MOD basis portfolio constructed by minimum idiosyncratic risk procedure by [43]

There are a few widely accepted construction methods for the factor mimicking portfolios. The return difference in both decile ends, exhibited in the previous section 2.4.6, is one of the popularly accepted methods. However, using only top 10% and bottom 10% of sensitivity-sorted stocks for the basis portfolio can be seen as arbitrary. Another popular method is based on Fama-MacBeth mimicking portfolio construction [29].

This section briefly discusses the theory and the caveat of Fama-MacBeth portfolio formation and advocates the minimum idiosyncratic portfolio construction method by [43]. We construct a MOD factor mimicking portfolio by the idiosyncratic method by [43] in order to exhibit that the consideration of modularity factor enhances the

investment perspective even with the different mimicking method than previous section.

Fama-MacBeth [29] noted that the estimated coefficients from cross-sectional regressions on returns can be interpreted as portfolio returns, since they are linear functions of the dependent variable. Consider a linear regression of the k -factor model.

$$r = BF \tag{20}$$

$$B'r = B'BF \tag{21}$$

$$\hat{F} = (B'B)^{-1}B'r = [B(B'B)^{-1}]'r =: W'r \tag{22}$$

where r is a vector of excess returns, F is a vector of factor return, and B is *exposure matrix* whose i -th row implies i -th asset's (stock's) exposure to the k -factors. Since $\hat{F}_t = W'r_t$, [19] interprets the matrix B as a set of k portfolio vectors with the j -th estimated factor equal to the excess return on the j -th portfolio.

However, [43] points out that there is no guarantee that the Fama-Macbeth factor mimicking portfolio will be sufficiently highly correlated to the original factor in finite samples. They show that the presence of sampling error in estimation of regression coefficient leads Fama-MacBeth mimicking portfolio to contain a significant amount of idiosyncratic risks. As a result, the factor mimicking portfolio is less attractive as a basis portfolio.

[43] proposes the *minimum idiosyncratic risk procedure* that the proposed basis portfolio is constructed by the difference of marketwide equally weighted portfolio and the portfolio that is orthogonal to one factor under consideration. The orthogonal portfolio to the one factor solves the following problem.

$$\min w'_{orth} w_{orth} \tag{23}$$

$$s.t. w'_{orth} \mathbf{1} = 1 \tag{24}$$

$$w'_{orth} \beta = 0 \tag{25}$$

Table 9: Weights in the ex-post tangency portfolio and monthly Sharpe ratios (January 1992 - December 2011)

Number of instruments	<i>MKT</i>	<i>SMB</i>	<i>HML</i>	<i>MOM</i>	<i>MOD_P</i>	Sharpe Ratio
1	100					0.109
2	64.3	35.7				0.116
2	43.1		56.9			0.176
2	55.2			44.8		0.172
2	55.5				44.5	0.177
3	26.6	25.5	47.8			0.197
4	26.3	14.3	38.7	20.7		0.252
4	28.1	20.0	34.2		17.7	0.226
5	27.1	13.3	33.2	17.2	9.2	0.262

where $\mathbf{1}$ is a vector of ones, and β is obtained by the factor linear regression. The solution to this problem is given as $w_{orth} = \frac{1}{N} \left[\frac{\sigma_{\beta}^2 + \bar{\beta}^2}{\sigma_{\beta}^2} \mathbf{1} - \frac{\bar{\beta}^2}{\sigma_{\beta}^2} \beta \right]$ where N is the number of stocks used for construction of portfolio, σ_{β} is the standard deviation of β , and $\bar{\beta}$ is the mean of β . The factor mimicking portfolio is therefore $w_{mimic} = \frac{1}{N} \mathbf{1} - w_{orth} = \frac{\bar{\beta}}{N\sigma_{\beta}^2} (\beta - \mathbf{1}\bar{\beta})$.

We estimate a $N \times 1$ vector β^{MOD} from (8) and use the above method to construct the modularity factor mimicking portfolio. We call this portfolio as MOD_P where the subscript ‘ P ’ implies *portfolio*. Following Table 9 - Table 11 and Figure 8 repeat the same analysis²⁷ in section 2.4.6. The correlation between MOD_S and MOD_P is about 0.725 and the analysis results are very similar. The MOD_P is also an attractive instrument in Mean-Variance sense (Table 9). The low correlation of MOD_P with respect to the market portfolio promotes the usage of MOD_P as overlaying portfolio on the market index portfolios (Table 10, Table 11, and Figure 8). We conclude that the both mimicking methods lead similar enhancements in investment perspectives by providing the basis portfolio that mimics modularity factor.

²⁷In order to present sensible numbers, MOD_P is scaled to have same standard deviation to MOM .

Table 10: Cross-correlation matrix of the five factors including MOD_P (January 1992 - December 2011)

	MKT	SML	HML	MOM	MOD_P
MKT	1	0.24	-0.24	-0.26	-0.31
SML		1	-0.35	0.09	-0.09
HML			1	-0.14	0.26
MOM				1	0.28
MOD_P					1

Table 11: Performance of enhanced market index portfolio (January 1992 - December 2011)

	MKT	$MKT + 10\% MOD_P$	$MKT + 20\% MOD_P$
Return (p.a.)	8.94%	9.56%	10.18%
Std (p.a.)	15.46%	14.98%	14.72%
SR (monthly)	0.109	0.124	0.139

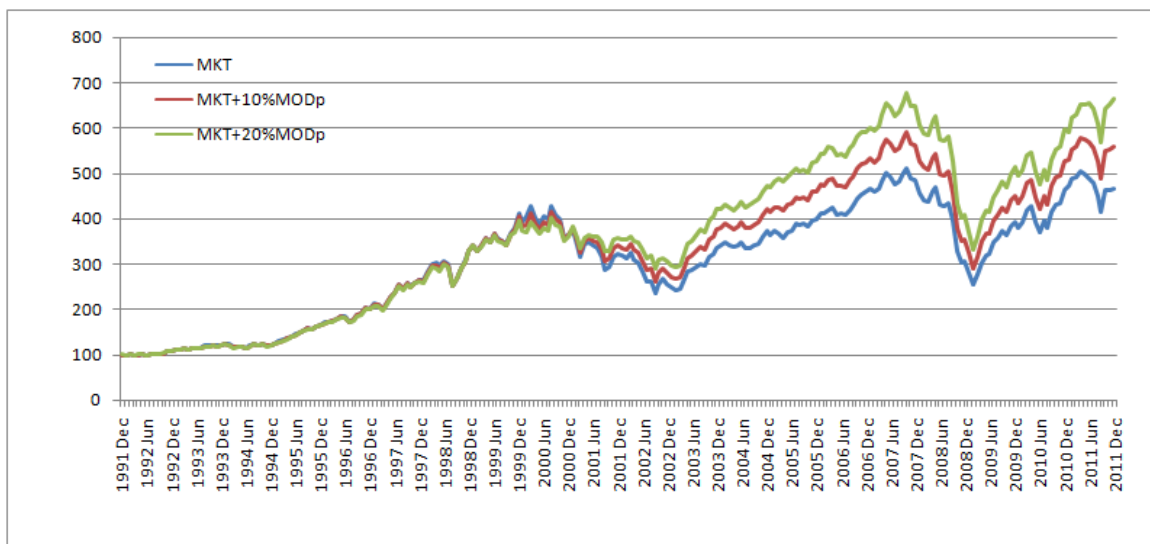


Figure 8: Cumulative return on enhancement scenarios with MOD_P

2.5 *Concluding remark*

We propose the measure of modularity that quantifies the strength of cluster structure in stock markets. This measure is related to the cycle of economy. Yet, it represents a rather independent dimension to the other connectedness measures such as the average of marketwide correlations. Thus, we test our modularity measure as an alternative of systematic factor which influence the expected returns of common stocks.

Our empirical study results demonstrate that the modularity measure is indeed a valid risk factor driving the asset returns. The stocks with negative sensitivity to modularity factor have considerably higher expected return even after exposure to Fama-French three factors and momentum factor are accounted for. The difference in decile portfolio or mimicking method of [43] creates modularity factor portfolios that can be used to widen investment opportunity to passive investors.

Future extension of our study can be done by agglomerating current construction of modularity to statistical techniques such as principal component analysis or hidden graphical models. The latent structure of stock market can be further refined and so can be the modularity factor.

Although we use the daily frequency data of stock markets in our analysis, the framework for analyzing the time-varying cluster property of the measure of modularity can be applied to other types of financial market data, including but not limited to the high-frequency data taken from other financial markets such as fixed income market and foreign exchange market.

CHAPTER III

ESTIMATION OF HIDDEN LIQUIDITY AND ITS EMPIRICAL ADVANTAGES IN PRICE IMPACT FUNCTION, ORDERBOOK PRESSURE MODEL, AND OPTIMAL ORDER EXECUTION STRATEGIES

We propose a statistical model for quantifying hidden liquidity at the best bid and ask prices in electronic trading venues based on Level-I limit orderbook data. Not only the estimated hidden liquidity can explain the probabilistic property such as orderbook pressure better, it also refines the existing price impact model by [22] thereby achieving higher explanation powers. Our enhanced price impact function can serve as a foundation for devising optimal intraday order execution strategies. Simulation tests of our model using historical data show an average saving of 29.7% in transaction cost compared to the execution strategies of equally-splitting as proposed by [8][5] and an average savings of 36.5% over the trading strategy which accounts for the resiliency of liquidity in limit orderbook by [49].

3.1 Introduction

As Electronic Communications Networks (ECNs) have become the popular trading venues in various financial markets, market participants experience clearer market information available in lower latency than before. However, the transparency and the immediacy of available market information often makes market participants to be afraid of exposing their trading willingness due to being possibly “picked off” or “arbitraged” [61]. Therefore, some exchange platforms allow market participants to

hide a partial or an entire portion of their limit orders. Such practice is termed as placements of *hidden limit orders*, and the liquidity provided by the orders in the market is referred to as *hidden liquidity*. The hidden orders have become a popular practice, and currently accounts for a substantial portion of total liquidity available in financial markets. According to [35], 15% of market orders were executed against hidden orders of stocks traded on Inet in 2004. The portion increased to 17.3% in 2007 and 19.0% in 2008.

The presence of hidden liquidity influences the market dynamics similarly as does the visible liquidity. Empirical evidence provided by [32] shows that more hidden liquidity results in greater total liquidity, greater trading volume, and smaller price impact of individual orders. [7] shows the same effect by hidden liquidity on price dynamics in ultra-high-frequency level (tick level). [13] demonstrates how hidden liquidity affects the trading costs.

Since market participants face incomplete information setting when available liquidity is only partially observed, the presence of hidden liquidity creates new challenges for making trading decisions. Under the incomplete information settings, market participants need to estimate the level of available hidden liquidity in order to accurately estimate the potential market impact caused by their own trading. One approach is to use the “ping” detection method (placing a small unit of market order to see if there is any counterpart hidden liquidity), but this method is *ad hoc* and cannot reveal the total size of hidden liquidity. Other than such detection methods, gauging the quantity of hidden liquidity is an important task especially prior to placing a substantial size of orders. [15][61] present evidence that the size of hidden liquidity can be explained by stock characteristics such as size of trade, spread, volatility using the database NASDAQ ModelView. Compared to this study, we develop estimation process based solely on physical mechanism of orderbook without relying on the informational contents of market characteristics. By incorporating hidden liquidity, we

aim to extend the price impact model as proposed by [22].

We propose a statistical model for estimating the size of hidden liquidity (section 3.3). By incorporating estimates of hidden liquidity, we propose a refined price impact function that connects the amount of orders with the resulting quote price changes (section 3.4.1). The estimates also lead to an accurate modelling of the probabilistic property of orderbook (section 3.4.2). Finally, we illustrate that the hidden liquidity help market participants make more informed trading decisions under the incomplete information setting (section 3.6) and save transaction costs.

3.2 Background and literature review

High-frequency financial data consists of trades and quotes data in electronic order-driven financial markets. It has become widely available through the development of information technology, and the increased availability of high-frequency data raised the importance of data processing, statistical modeling, and stochastic modeling.

[20] classifies high-frequency data further by its frequency level. Ultrahigh-frequency data is tick level frequency of limit orderbook data driven by the physical mechanisms of individual orders' arrivals and cancellations. Modelling the limit orderbook characteristics and their most respective effects on market prices are the main problems of interests in analyzing ultrahigh-frequency data. On the other hand, (non-ultra) high-frequency data is the data of frequency in 1-100 seconds, and its main interests include trade executions and trades' impacts on quote prices, price jumps, and the information contents of price fluctuations. Our estimation process should be classified as a study of ultrahigh-frequency.

The two most practically important high-frequency trading applications are statistical arbitrages and order placement strategies. Statistical arbitrageurs aim to identify anomalies in the market by quantitative analysis on ultrahigh-frequency or

(non-ultra) high-frequency data and to exploit them by making trades in low latency. Typically for brokerage firms, the large-sized (parent) orders need to be strategically split into a series of small-sized (child) orders by assessing the potential impact of each order to market price changes.

3.2.1 Price impact function

In general, high buying demands of a security typically result in increases of subsequent quote/trade prices. In this sense, market tends to move against a trader's intended trading direction, i.e. consecutive buying actions of the trader consume the supply of stocks from the lowest available prices and it will result in the higher quote price. Therefore, large size of orders are not likely to be completely executed at the best bid (or, ask) price due to this adverse price movements.

The price impact function measures the relationship between aggregated actions by traders and the change in market price. The price impact alone accounts for a large portion of the total transaction cost associated with trading. Therefore, estimating an accurate price impact function is a critical foundation prior to devising optimal order placement strategies. Relevant studies on optimal order placements based on price impact functions include [4][5][8][49][3][39]. Price impact functions in practice typically suffer very low explanation power (The R-squares are typically below 30%) due to a lot of noises in supply-demand imbalance measurements such as [42]. [22] takes a rather unconventional approach¹ of constructing a high-frequency price impact function based on the orderbook mechanism in ultrahigh-frequency level, and build a price impact function reaching approximately 65% of R-square on average. Their study is characterized by the usage of quote data rather than trade data, which allows them to utilize ample data set and to gather more accurate supply-demand

¹Will be discussed later in detail in section 3.3.1.

imbalance than the traditional tick-test of [42].

3.2.2 Orderbook pressure: microstructure model of order imbalance and midprice movement

The Level-I orderbook contains (best) ask and bid prices and waiting limit orders at the two prices. The Level-I orderbook is updated whenever there is a change to the orderbook state caused by either a market order arrival, a limit order arrival, or a limit order cancellation. We notate the state of Level-I orderbook at time n by

$$OB_n := (P^B(n), P^A(n), q^B(n), q^A(n))$$

, where $P^B(\cdot)$ and $P^A(\cdot)$ imply (best) bid and ask prices, $q^B(\cdot)$ and $q^A(\cdot)$ denote the size of waiting orders at the prices, and $P(n) := \frac{P^B(n) + P^A(n)}{2}$ is the midprice.

Imbalanced respective total amount of orders waiting at the best bid and the best ask price levels likely pushes the price to move in the direction of less amount of orders, with all other things being equal. This phenomena is called as the *orderbook pressure*. For example, $q^B(n) \gg q^A(n)$ implies a significantly high chance of midprice moving up in the near future. The effect of orderbook pressure can be modelled by considering the probability of midprice moving up as a function of the sizes of the waiting ask and bid orders.

$$p_{up}(OB_n) := \mathbb{P}[P(\inf_{t>n}\{P(t) \neq P(n)\}) > P(n) | OB_n] = f(q^B(n), q^A(n))$$

This function is of high interest in microstructural modeling with ultrahigh-frequency data. It characterizes a *no-knowledge model* in which only level-I waiting order sizes are relevant to the future direction of price change, while all other variables are ignored. It assumes the *Markovian property* in limit orderbook without needing to incorporate the history of orderbook transitions is ignored. Although this Markovian assumption is often challenged by empirical studies such as [6][37], it still serves as the

Table 12: The probability of midprice going up given orderbook states (2012 January, BAC).

Size of bid orders	Size of ask orders				
	(83900,107450]	(107450,127350]	(127350,148700]	(148700,174400]	(174400,213650]
(83900,107450]	0.522	0.462	0.401	0.343	0.302
(107450,127350]	0.563	0.537	0.481	0.405	0.344
(127350,148700]	0.62	0.604	0.498	0.422	0.411
(148700,174400]	0.644	0.629	0.558	0.516	0.476
(174400,213650]	0.674	0.733	0.664	0.596	0.523

foundation for characterizing price fluctuations in reduced forms. The functional form of orderbook pressure aims to describe empirical results such as Table 12, which shows the effect of orderbook imbalance to the directional tendency in midprice movement. In Table 12, each column corresponds to the range of size of ask orders and each row corresponds to the range of size of bid orders. In upper right corner of the table, midprice is likely to go down because of the small size in existing bid orders and the large size in existing ask orders.

There have been a few studies on the orderbook pressure resorting to stochastic models. [23] models the innovations of $q^A(\cdot)$ and $q^B(\cdot)$ as queuing processes where both order arrivals and cancellations are assumed to be independent Poisson processes.² They derive the probabilities of increase in midprice such as Table 12 by Monte Carlo simulations and Laplace transform. [21] proposes diffusion approximations to the queuing models used in [23], then presents limit theorems and the reduced form of orderbook pressure model as follows.

$$\begin{aligned}
p_{up}(x, y) &= \mathbb{P}[P(\inf_{t \geq n} \{P(t) \neq P(n)\}) > P(n) | (q^B(n), q^A(n)) = (x, y)] \\
&= \frac{1}{2} \left(1 - \frac{\arctan\left(\sqrt{\frac{1+\rho}{1-\rho}} \frac{y-x}{y+x}\right)}{\arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right)} \right), \tag{26}
\end{aligned}$$

where $x = \frac{X}{\langle X \rangle}$, $y = \frac{Y}{\langle Y \rangle}$, X, Y are bid and ask queue sizes, and $\langle X \rangle$ and

²A (simple) Poisson process is a counting process that has identically and independently distributed exponential inter-arrival times between unit-sized arrivals.

$< Y >$ are medians of X and Y . Assuming $dx_t = \sigma dW_t^{(1)}$, $dy_t = \sigma dW_t^{(2)}$, and $\mathbb{E}(dW^{(1)}dW^{(2)}) = \rho dt$ where $W^{(1)}$ and $W^{(2)}$ are independent Brownian motions, the probability of midprice moving up given the orderbook states are derived.

[7] is the first study that considers the existence of hidden liquidity in a orderbook pressure model. They incorporate the presence of hidden liquidity by replacing x and y in (26) by $x + H$ and $y + H$, where H denotes the average size of hidden liquidity at the best ask and best bid prices. By maximum likelihood estimations, H is estimated and it fits empirical data better than the previous model [23], which do not consider the existence of hidden liquidity.

In the following sections, we use different approach to estimate the level of hidden liquidity and present the evidence that our estimates fit empirical data of orderbook pressure better than does [22].

3.3 Model and estimation strategy

In this section, we first review the construction process of price impact model proposed in [22], then propose an improved model by incorporating the presence of hidden liquidity, which enables us to estimate average hidden liquidity at the best bid and ask prices. In level-I orderbook, let P^B be the bid price, q^B be the size of waiting bid orders at P^B , P^A be the ask price, q^A be the size of waiting ask orders at P^A , and $P = (P^A + P^B)/2$ be the (bid-ask) midprice. Let $N(t)$ be the counting function for the number of orderbook events during the time $[0, t]$ where an orderbook event is defined as any event that changes the vector of orderbook state $OB := (P^A, P^B, q^A, q^B)$, i.e., limit order arrivals, limit order cancellations, and market order arrivals.

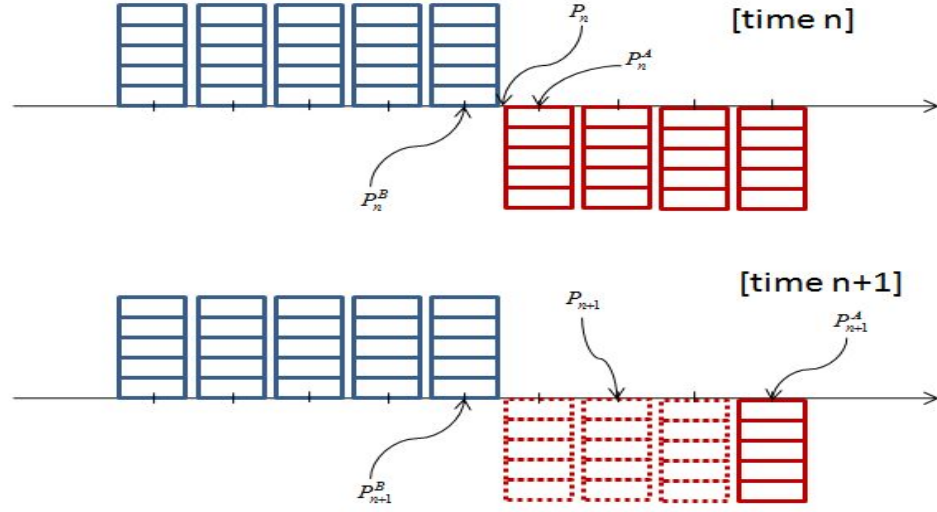


Figure 9: Orderbook with uniformly sized waiting orders and the effect of imbalance on it

3.3.1 Review on the price impact function of Cont et al. (2013)

This section reviews the assumptions and processes of price impact function construction by [22]. Consider an orderbook with assumptions that i) waiting orders at all prices are uniformly sized and ii) limit order arrivals and cancellations take places only at (best) bid and ask prices. Then, the orderbook (supply-demand) imbalance will change the midprice linearly as illustrated in Figure 9. In the upper diagram of Figure 9, waiting orders at each price level are five. In the lower diagram of Figure 9, a market buy order of fifteen (or, equivalently, a cancellation of ask limit order of fifteen) is placed, and it consumes the sell side of liquidity for the three adjacent price levels. As a result, ask price is shifted by three ticks and the midprice is moved by three half ticks. Likewise, each orderbook event changes the midprice in linear fashion.³

³The round-off discrepancies may come into play, but these are offset away when aggregated to a larger time interval.

The *contribution of n -th orderbook event to the orderbook imbalance* can be therefore measured by e_n ,

$$e_n = I_{(P_n^B \geq P_{n-1}^B)} q_n^B - I_{(P_n^B \leq P_{n-1}^B)} q_{n-1}^B - I_{(P_n^A \leq P_{n-1}^A)} q_n^A + I_{(P_n^A \geq P_{n-1}^A)} q_{n-1}^A \quad (27)$$

and spelling out the indicator functions for each case will specify (27) by the following five cases,

- i)* If $P_n^B = P_{n-1}^B$ and $P_n^A = P_{n-1}^A$ (Both ask and bid prices stay the same),
then, $e_n = (q_n^B - q_{n-1}^B) - (q_n^A - q_{n-1}^A)$
- ii)* If $P_n^B > P_{n-1}^B$ and $P_n^A = P_{n-1}^A$ (New bid arrival narrows the spread),
then, $e_n = q_n^B - (q_n^A - q_{n-1}^A)$
- iii)* If $P_n^B < P_{n-1}^B$ and $P_n^A = P_{n-1}^A$ (Depletion at bid widens the spread),
then, $e_n = -q_{n-1}^B - (q_n^A - q_{n-1}^A)$
- iv)* If $P_n^B = P_{n-1}^B$ and $P_n^A < P_{n-1}^A$ (New ask arrival narrows the spread),
then, $e_n = (q_n^B - q_{n-1}^B) - q_n^A$
- v)* If $P_n^B = P_{n-1}^B$ and $P_n^A > P_{n-1}^A$ (Depletion at ask widens the spread),
then, $e_n = (q_n^B - q_{n-1}^B) - (-q_{n-1}^A)$

Let each k for $(t_{k-1}, t_k]$ be a small time interval such as 10 seconds and $N(\cdot)$ be a counting function for the number of orderbook events, the summing e_n over the time interval $(t_{k-1}, t_k]$ defines the *order flow imbalance (OFI)* as follows.

$$OFI_k = \sum_{N(t_{k-1})+1}^{N(t_k)} e_n \quad (28)$$

Let $\Delta P_k := P_k - P_{k-1}$, then the relationship of *order flow imbalances (OFI)* and the resulting price changes is given as the following linear price impact function.

$$\Delta P_k = \beta_i OFI_k + \epsilon_k \quad (29)$$

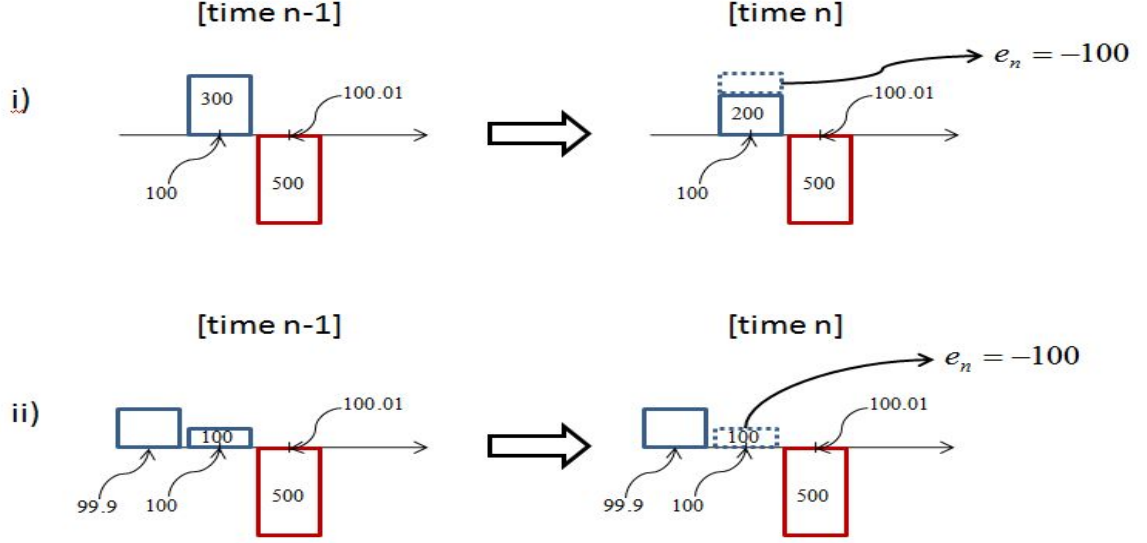


Figure 10: Two equivalent scenarios of measuring e_n of [22].

where each i represents a longer time interval such as 30 minutes interval during which the linear relationship is assumed to be sustained. [22] presents the empirical evidence that the price impact function of (29) has a high explanation power around 65% of R-square on average.

3.3.2 Model development

We modify the price impact function (29) of [22] by incorporating the presence of hidden liquidity. We first question a suitable imbalance measure for the following two scenarios of decrease in waiting bid order. This is illustrated in upper and lower diagram of Figure 10.

- i) From $q_{n-1}^B = 300$, it becomes $q_n^B = 200$ while bid price stays the same.
($P_n^B = P_{n-1}^B$)
- ii) From $q_{n-1}^B = 100$, all of the bid orders at the bid price are depleted and bid price decreases. ($P_n^B < P_{n-1}^B$).

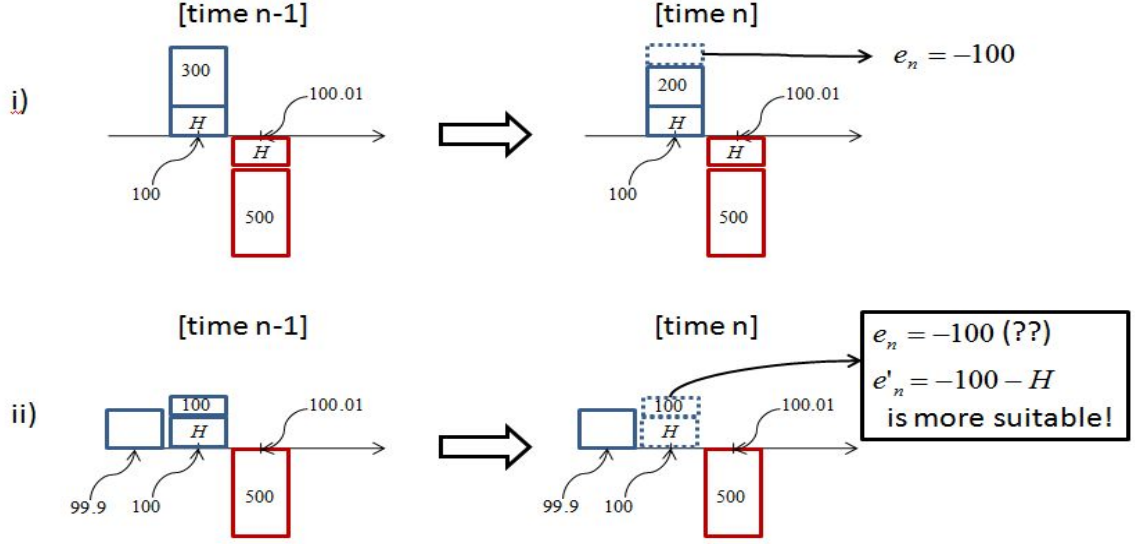


Figure 11: Different perspectives of the two scenarios with the presence of hidden liquidity

In both cases, the bid side of liquidity are consumed. The measure e_n in [22] treats the above two events in the exactly same way, i.e. $e_n = -100$ in (27). However, the contribution to imbalance of the second case should be measured as a larger amount than that of the first case if one considers that the hidden liquidity is also depleted in addition to the visible 100 unit of bid order. This is illustrated in the upper and lower diagram of Figure 11. In the upper diagram, 100 unit of bid order is removed and we agree with [22] that the imbalance should be -100 . In the lower diagram, however, the contribution to orderbook imbalance from n -th event should be $-100 - H$ rather than -100 because visible liquidity has a higher priority of execution than does hidden liquidity. By ignoring the presence of hidden liquidity, e_n in [22] tends to underestimate the contribution of event to orderbook in the case of iii) and v) in section 3.3.1 (the spell-out for (27)). In a nutshell, price changing events should be measured as higher contributions to orderbook imbalance, because those events may include the consumption of hidden liquidity. We therefore newly define e'_n as *the contribution of n -th event to orderbook imbalance in the presence of*

hidden liquidity for the following cases.

- iii) If $P_n^B < P_{n-1}^B$ and $P_n^A = P_{n-1}^A$ (Depletion at bid widens the spread),
then, $e'_n = -(q_{n-1}^B + H) - (q_n^A - q_{n-1}^A)$
- v) If $P_n^B = P_{n-1}^B$ and $P_n^A > P_{n-1}^A$ (Depletion at ask widens the spread),
then, $e'_n = (q_n^B - q_{n-1}^B) - (-q_{n-1}^A - H)$

These changes formulate the definition of e'_n :

$$\begin{aligned} e'_n &= I_{(P_n^B \geq P_{n-1}^B)} q_n^B - I_{(P_n^B \leq P_{n-1}^B)} q_{n-1}^B - I_{(P_n^A \leq P_{n-1}^A)} q_n^A + I_{(P_n^A \geq P_{n-1}^A)} q_{n-1}^A \\ &\quad - H I_{(P_n^B < P_{n-1}^B)} + H I_{(P_n^A > P_{n-1}^A)} \end{aligned} \quad (30)$$

$$= e_n - H I_{(P_n^B < P_{n-1}^B)} + H I_{(P_n^A > P_{n-1}^A)} \quad (31)$$

As a result, the new definition of *order flow imbalance (under the presence of hidden liquidity)*, OFI' , is derived as:

$$\begin{aligned} OFI'_k &= OFI_k - H \sum_{n=N(t_{k-1})+1}^{N(t_k)} \left(I_{(P_n^B < P_{n-1}^B)} - I_{(P_n^A > P_{n-1}^A)} \right) \\ &= OFI_k - H \cdot ODI_k, \end{aligned} \quad (32)$$

where the *order depletion imbalance*, ODI , measures the difference between the number of depletions in bid and ask queues. The OFI' leads to a price impact function that takes the presence of hidden liquidity into account, where k is the frequency of imbalance aggregation and i is the interval of which the price impact relationship is assumed to be sustained. (Again, we set k to be 10 seconds and i to be 30 minutes in our tests just as [22].)

$$\begin{aligned} \Delta P_k &= \beta_i OFI'_k + \eta_k \\ &= \beta_i (OFI_k - H \cdot ODI_k) + \eta_k \end{aligned} \quad (33)$$

3.3.3 Two step estimation process

Based on our price impact function (33), one can estimate the average level of hidden liquidity using historical orderbook data. We use two step estimation processes for hidden liquidity level, H , at bid and ask prices. One step approach would be a multiple linear regression of (33) such as $\Delta P_k = a + bOFI_k + cODI_k + \eta_k$ where b should serve as an estimate of β_i and c should serve as an estimate of $\beta_i H$. However, it would be less straightforward to observe the statistical property of H alone in this approach. Thus, we propose the following two step approach. First, we conduct the linear regression of original price impact function, $\Delta P_k = \beta_i OFI_k + \epsilon_k$, of [22] and collect its residuals.

$$\epsilon_k = \Delta P_k - \hat{\beta}_i OFI_k \quad (34)$$

Second, based on (33), we conduct a simple linear regression on the unexplained part of [22].

$$\epsilon_k = H_i(-\hat{\beta}_i ODI_k) + \eta_k, \quad (35)$$

where the H_i is the average size of *hidden liquidity at bid/ask prices* over the 30 minutes interval i . With (32), our main price impact function is estimated as follows.

$$\Delta P_k = \beta_i OFI'_k + \eta_k \quad (36)$$

3.4 Empirical evidence

We use trades and quotes (TAQ) database from January 2012 to June 2012. Following twenty actively traded stocks in the U.S. exchanges are selected: AAPL (Apple Inc.), BA (The Boeing Company), BAC (Bank of America Corporation), BP (BP plc), COP (ConocoPhillips), CSCO (Cisco Systems, Inc.), CVS (CVS Caremark Corporation), DELL (Dell Inc.), DIS (The Walt Disney Company), DOW (The Dow

Chemical Company), JNJ (Johnson and Johnson), JPM (JPMorgan Chase and Co.), KO (The Coca-Cola Company), KR (The Kroger Co.), MCK (McKesson Corporation), MSFT (Microsoft Corp.), PG (The Procter and Gamble Company), TGT (Target Corp.), WFC (Wells Fargo and Company), and XOM (Exxon Mobil Corp.). The data was obtained from TAQ consolidation quotes and trades databases provided by the Wharton Research Data Services (WRDS). We aggregate quotes updates from all the U.S. exchanges to estimate the National Best Bid and Offer (NBBO) data set by the suggested method of WRDS⁴.

Our consideration of hidden liquidity should achieve a higher explanation power than the models that ignore the presence of hidden liquidity. In section 3.4.1, we present the evidence that our price impact function with consideration of hidden liquidity has improved statistical properties than the one by [22]. In section 3.4.2, we present the evidence that our estimation method is better than that of [7] in explaining the orderbook pressure. In section 3.4.3, we present intraday pattern of estimated average hidden liquidity.

3.4.1 Enhancing the price impact function

In this section, we compare the statistical properties of the following two price impact functions. The price impact function (I) is proposed by [22], and we incorporate the hidden liquidity into the function (II). The intercepts are added for the statistical tests.

$$(I) \Delta P_k = \alpha_i + \beta_i OFI_k + \epsilon_k \quad (\text{intercept added to (29)})$$

$$\begin{aligned} (II) \Delta P_k &= \alpha_i + \beta_i OFI'_k + \eta_k \quad (\text{intercept added to (33)}) \\ &= \alpha_i + \beta_i (OFI_k - H \cdot ODI_k) + \eta_k \end{aligned}$$

⁴The detail can be found at http://wrds-web.wharton.upenn.edu/wrds/research/applications/microstructure/NBBO_derivation/

We first test whether the unexplained part of (I) is explained by introducing a variable of hidden liquidity, H . After ϵ_k is obtained in (34), the result of linear regression $\epsilon_k = \alpha_i + H_i(-\hat{\beta}_i ODI_k) + \eta_k$,⁵ is presented in Table 13. The results shown in the columns of t -value and P -value support our conjecture that $\alpha_i = 0$ and $H_i > 0$. Incorporating the presence of hidden liquidity can explain about 11%-34% of the unexplained portion of price impact function (I) as shown in the column of R^2 . Having confirmed the significance of hidden liquidity to the price impact function, we compare the overall statistical property of (I) and (II) in Table 14. Since (II) has an additional variable, we fix H in (II) as same as the estimated value of previous 30 minute interval. By doing so, Table 14 presents the statistical properties of the both simple (one variable) linear regressions. The price impact functions (II) across different stocks and days are shown to have higher explanation powers than (I) based on the R^2 column. Also, the variable OFI' in (II) is more significant than OFI in (I) based on the t -statistics of coefficients (columns of $t(\hat{\beta}_i)$).

3.4.2 The prediction power in market microstructure model

[7] estimates the size of hidden liquidity which improves the orderbook pressure model of [21]. We test whether our hidden liquidity estimations can improve the orderbook pressure model of [21] even further.

Assuming $\rho \rightarrow -1$ for simplicity, (26) is reduced to

$$p_{up}(x, y, H) = \frac{x + H}{x + y + 2H} \quad (37)$$

[7] estimates H by maximum likelihood method that minimizes the least square error of $\min_H \sum_{i,j} \left[\left(u_{ij} - \frac{i+H}{i+j+2H} \right)^2 d_{ij} \right]$, where i and j represent the i -th decile of bid queue(a row in Table 12) and j -th decile of ask queue(a column in Table 12) respectively, u_{ij} is empirical probability, and d_{ij} is the number of corresponding observations.

⁵An intercept is added to (35).

Table 13: Average results of statistical properties in (35) (The intercept is added for testing)

Ticker	Average results					P-value	
	$\hat{\alpha}_i$	$t(\hat{\alpha}_i)$	\hat{H}_i	$t(\hat{H}_i)$	R^2	$\alpha_i \neq 0$	$\beta_i \neq 0$
AAPL	0.0251	0.0328	5.2	5.39	0.146	0.659	0.0008
BA	-0.0032	-0.0624	12.1	6.04	0.173	0.644	0.0001
BAC	0.0002	0.0157	6703.5	7.63	0.240	0.734	0.0088
BP	-0.0004	-0.027	67.7	7.78	0.253	0.659	0.0004
COP	-0.0014	-0.0424	20.9	6.37	0.188	0.689	0.0000
CSCO	-0.0002	-0.0244	2434.5	9.14	0.299	0.664	0.0005
CVS	-0.0015	-0.0481	65.2	7.73	0.250	0.672	0.0000
DELL	-0.0005	-0.0389	1404.1	9.55	0.324	0.628	0.0001
DIS	-0.0015	-0.0548	89.6	7.35	0.231	0.693	0.0001
DOW	0.0002	-0.0016	81.8	8.55	0.288	0.656	0.0000
JNJ	0.0000	0.004	54.4	6.75	0.206	0.690	0.0020
JPM	-0.0014	-0.0303	98.2	5.80	0.162	0.760	0.0004
KO	-0.0009	-0.0328	45.7	7.42	0.236	0.671	0.0032
KR	0.0002	0.0311	208.7	9.89	0.336	0.611	0.0000
MCK	-0.0020	-0.0268	5.1	4.92	0.127	0.691	0.0020
MSFT	-0.0003	-0.0132	446.0	8.08	0.261	0.684	0.0003
PG	-0.0004	-0.0157	39.0	6.29	0.187	0.721	0.0090
TGT	-0.0025	-0.0626	19.6	6.27	0.183	0.694	0.0002
WFC	-0.0008	-0.0257	199.8	6.64	0.198	0.714	0.0008
XOM	-0.0028	-0.062	14.1	4.39	0.108	0.769	0.0308
Average	0.0003	-0.0243	600.8	7.0997	0.2198	0.6851	0.0030

Table 14: Average results of statistical properties in price impact functions (from January 2012 to June 2012)

	Price impact function (I) by [22]					Price impact function (II)				
Ticker	$\hat{\alpha}_i$	$t(\hat{\alpha}_i)$	$\hat{\beta}_i$	$t(\hat{\beta}_i)$	R^2	$\hat{\alpha}_i$	$t(\hat{\alpha}_i)$	$\hat{\beta}_i$	$t(\hat{\beta}_i)$	R^2
AAPL	0.016	0.000	0.358	11.55	0.422	0.034	0.052	0.289	13.91	0.513
BA	-0.002	-0.035	0.035	13.11	0.482	-0.004	-0.075	0.030	15.97	0.579
BAC	-0.001	-0.031	0.000	21.19	0.669	0.000	-0.009	0.000	26.29	0.763
BP	-0.001	-0.026	0.004	21.40	0.695	-0.001	-0.026	0.004	27.45	0.789
COP	-0.006	-0.099	0.020	14.36	0.527	-0.006	-0.117	0.016	17.81	0.630
CSCO	0.000	0.015	0.000	20.56	0.655	0.000	0.019	0.000	28.71	0.774
CVS	-0.005	-0.121	0.007	14.54	0.528	-0.005	-0.142	0.006	19.50	0.667
DELL	0.001	0.043	0.001	18.65	0.623	0.000	0.027	0.000	26.96	0.754
DIS	-0.006	-0.163	0.006	15.18	0.551	-0.006	-0.186	0.005	19.83	0.672
DOW	0.003	0.079	0.006	13.62	0.499	0.003	0.089	0.005	19.32	0.666
JNJ	0.001	0.027	0.004	24.38	0.745	0.001	0.038	0.003	29.43	0.810
JPM	-0.004	-0.084	0.003	17.66	0.622	-0.004	-0.091	0.003	21.10	0.701
KO	-0.003	-0.097	0.007	21.62	0.696	-0.003	-0.096	0.006	26.95	0.782
KR	0.000	-0.004	0.003	12.71	0.458	0.000	0.011	0.003	19.74	0.652
MCK	-0.002	-0.032	0.085	13.46	0.498	-0.003	-0.056	0.072	15.49	0.567
MSFT	-0.001	-0.037	0.001	22.87	0.698	-0.001	-0.030	0.000	29.51	0.794
PG	0.002	0.076	0.005	25.07	0.757	0.002	0.062	0.005	29.32	0.805
TGT	-0.007	-0.148	0.020	14.42	0.528	-0.008	-0.180	0.017	17.77	0.628
WFC	-0.001	-0.060	0.002	26.17	0.763	-0.001	-0.055	0.001	31.34	0.821
XOM	-0.007	-0.139	0.010	26.33	0.759	-0.008	-0.175	0.009	28.22	0.787
Average	-0.001	-0.042	0.029	18.44	0.609	0.000	-0.047	0.024	23.23	0.707

After estimating H , substituting the estimate back to (37) produces the estimated orderbook pressure table such as Table 12. We compare the orderbook pressure table of the method of [7] against ours by following steps.

- *Step 1.* With the orderbook data of January 2012, estimate hidden liquidity by our method and the method by [7] for each 30 minutes interval of all 20 trading days.
- *Step 2.* Using the medians of the estimated hidden liquidity values in each method from January 2012, create 5 by 5 probability tables for orderbook pressure based on (37). This table is now a probability table forecast for February 2012.
- *Step 3.* Do *Step 1* with February 2012 data, then do *Step 2* with March 2012 data. Continue until *Step 1* with May 2012 and *Step 2* with June 2012.
- *Step 4.* Compute the distances i) between our tables and actual empirical tables and ii) between tables by method of [7] and empirical tables from February 2012 to June 2012.

We report two different types of average distance between the two 5 by 5 probability matrices. The first distance is the average standard errors of the 25 elements, and the second distance is the average value in Frobenius metric. Specifically, Let $P^{(1)}$ be our predicted probability matrix, $P^{(2)}$ be a predicted probability matrix by the method of [7], and $P^{(emp)}$ be an actual empirical matrix. The first distance is $\frac{1}{25} \sum_{1 \leq i \leq 5} \sum_{1 \leq j \leq 5} \frac{|P_{ij}^{(k)} - P_{ij}^{(emp)}|}{\sqrt{P_{ij}^{(emp)}(1 - P_{ij}^{(emp)})/n_{ij}}}$ where $P_{ij}^{(\cdot)}$ is the i -th row and j -th column element of the matrix $P^{(\cdot)}$ and n_{ij} is the corresponding number of observations. The second distance is $\|P^{(k)} - P^{(emp)}\|_F$ for $k = 1, 2$, where $\|A\|_F = \sqrt{\sum_i \sum_j |a_{ij}^2|}$ is the Frobenius norm for a square matrix A . The distances how far the estimated orderbook pressure matrices are from the actual ones are presented in Table 15. Our

Table 15: Average errors in orderbook pressure model by ours and by [7] (from January 2012 to June 2012)

Ticker	Average standard error		Frobenius metric	
	Ours	By [7]	Ours	By [7]
AAPL	18.46	11.88	0.404	0.198
BA	10.17	16.60	0.165	0.236
BAC	26.95	39.07	0.301	0.430
BP	13.84	21.98	0.150	0.246
COP	14.21	14.23	0.203	0.179
CSCO	12.83	26.98	0.148	0.317
CVS	9.71	18.63	0.144	0.252
DELL	13.00	24.77	0.184	0.347
DIS	13.76	21.39	0.161	0.238
DOW	14.45	26.28	0.153	0.276
JNJ	13.17	26.43	0.149	0.289
JPM	24.20	46.94	0.187	0.336
KO	11.41	12.74	0.181	0.187
KR	16.38	22.32	0.272	0.369
MCK	7.79	10.95	0.239	0.275
MSFT	15.77	31.52	0.161	0.312
PG	8.88	19.49	0.120	0.266
TGT	8.84	12.58	0.146	0.202
WFC	21.97	40.98	0.209	0.369
XOM	17.73	20.44	0.178	0.205
Average	14.68	23.31	0.193	0.276

hidden liquidity estimations produce orderbook pressure estimates closer to actual data than the estimations by [7], in terms of both distance measures. The magnitude of error is reduced to about the two third. We remark that all stocks except AAPL show that our estimation method produces closer result to the actual values. We believe that the anomaly in AAPL is due to its high price resulting in less effective number of observations.⁶

⁶This poor statistical performance on AAPL is consistent through our observations in following discussions.

3.4.3 Intraday patterns

The levels of visible and hidden liquidity change drastically during a trading day. Since these variations of liquidity will affect price impact functions, it is crucial to observe this intraday liquidity variations for the practical use of price impact function. Figure 12 and Figure 13 present the average 30 minute liquidity level for the first half year of 2012. The median levels of both visible and hidden liquidity tend to increase during a trading day, and this tendency is highly noticeable when comparing the first few hours and the last few hours.⁷

3.5 *Optimal execution strategy*

The drastic intraday variations of the level of liquidity shown in Figure 12 and Figure 13 imply the drastic changes in the price impact function of (33) during a day. In this section, we develop an optimal order execution strategy based on the intraday variations in liquidity.

There have been many efforts to devise optimal programs of splitting a large order into smaller pieces of ‘child’ orders. Too mention a few, [8] assumes the linear price impact function with no-knowledge price dynamics of simple random walk, and derived the naive evenly-splitting trading strategy that minimizes the expected total cost of purchase. [5] incorporates temporary price impact component where initial price changes by a trade die down as time progresses. The solution is still an evenly splitting strategy if traders are risk-neutral, while risk-averse traders would naturally choose more front-loaded solutions in order to stay away from market fluctuations. [39] also shows that risk averse traders should post more orders in the beginning stages. [49] discusses the liquidity components observed in the limit orderbook, where

⁷We have not found any plausible explanation on the characteristics of variations in hidden liquidity, because there are few studies regarding the intraday variation of hidden liquidity yet.

three basic characteristics of liquidity are concerned: bid-ask spread, market depth, and resilience. Focusing particularly on the resilience of limit orderbooks, the expected transaction cost minimization takes place by a solution where a trader should 1) place a substantial order at first, then 2) capture the continuously replenishing liquidity by placing consecutive small-sized child orders, which is followed by 3) a last substantial child order to complete the program at the very last trading period. Under the assumption of risk neutrality, our time-varying price impact function will lead to an optimal solution that considers the intraday variations of available liquidity, which results in less transaction cost than the evenly splitting strategy by [8] and [5]. Our test using historical data will provide a comparison against [49] as well, which tells which aspect of liquidity, whether it is the sheer depth of market or the resilience of liquidity, is practically more important.

3.5.1 Setting and goal

Without loss of generality⁸, suppose that a trader wants to buy X shares of a particular stock during a trading day. Let X_n , $1 \leq n \leq N$ be the size of the child orders at each time n . The trader is allowed to place child market orders at each discrete time n such that $X_1 + X_2 + \dots + X_N = X$. The trader aims to minimize the expected total transaction cost.

3.5.2 Assumptions

We assume the price dynamics of simple random walk and that the deterministic price impact from each trade is inversely proportional to the depth of orderbook V_t

⁸Selling program would be simply a mirror image.

as follows.

$$P_{t+1} - P_t = \frac{1}{2} \frac{X_t}{V_t} + \epsilon_{t+1}, \quad (38)$$

where P_t is the bid-ask midprice in unit of tick⁹ at time t , ϵ_{t+1} is white noise governing the fluctuation of the price during time interval $(t, t+1]$, and the coefficient $\frac{1}{2}$ of the term $\frac{X_t}{V_t}$ is derived by the assumption of uniform sized orderbook and this assumption is empirically verified by [22]. Based on the assumption of uniformly sized orderbook, we assume that a buy order of size X_t at time t will occur at the average price of $P_t + \frac{1}{2}s_t + \frac{1}{2V_t}X_t$, where s_t is the size of bid-ask spread at time t ¹⁰.

Suppose that a trader wants to execute X units of buy order throughout the discrete time horizon, n , $1 \leq n \leq N$. The expected total cost minimization problem of finding (X_1, \dots, X_N) subject to $X_n \geq 0$ for all n and $X_1 + \dots + X_N = X$ is defined as

$$(P) \quad \min_{(X_1, \dots, X_N)} \mathbb{E}_1 \left[\sum_{n=1}^N \left(P_n + \frac{1}{2}s_n + \frac{1}{2V_n}X_n \right) X_n \right] \quad (39)$$

$$= \min_{(X_1, \dots, X_N)} \mathbb{E}_1 \left[\sum_{n=1}^N \frac{1}{2}s_n X_n \right] + \min_{(X_1, \dots, X_N)} \mathbb{E}_1 \left[\sum_{n=1}^N \left(P_n + \frac{1}{2V_n}X_n \right) X_n \right] \quad (40)$$

$$= \min_{(X_1, \dots, X_N)} \frac{1}{2}\bar{s}X + \min_{(X_1, \dots, X_N)} \mathbb{E}_1 \left[\sum_{n=1}^N \left(P_n + \frac{1}{2V_n}X_n \right) X_n \right], \quad (41)$$

where $P_n + \frac{1}{2}s_n + \frac{1}{2V_n}X_n$ in (39) is the average execution price per unit at n -th execution, $\bar{s} := \frac{\sum_{n=1}^N s_n X_n}{X}$ is the weighted average of bid-ask spreads. We assume that we have no prediction power to the size of bid-ask spread over time and that the midprice dynamics are deterministic as the series of zero-mean white noises do not affect the expected total costs. Under these assumptions, the problem is reduced to

⁹For example, midprice of \$5.87 is denoted as $P_t=587$.

¹⁰ $P_t + \frac{1}{2}s_t$ is the best ask price where the first execution will begin. For the market order of $X_t \in (nV_t, (n+1)V_t]$ for $n \in \mathbb{N}$, the X_t consumes n consecutive ask prices completely and remaining $X_t - nV_t$ unit is executed at the price of $P_t + \frac{1}{2}s_t + nV_t$. Ex-post average execution price of X_t will be then somewhere close to the middle of beginning ask price of $P_t + \frac{1}{2}s_t$ and the ending ask price of $P_t + \frac{1}{2}s_t + \frac{X_t}{V_t}$. Error up to half tick, we approximate the average execution price by $P_t + \frac{1}{2}s_t + \frac{1}{2V_t}X_t$.

the following recursive form that considers only the second term in (41).

$$C_1(W_1) = \min_{(X_1, \dots, X_N)} \mathbb{E}_1 \left[\sum_{n=1}^N \left(P_n + \frac{1}{2V_n} X_n \right) X_n \right] \quad (42)$$

$$= \min_{X_1} \mathbb{E}_1 \left[\left(P_1 + \frac{1}{2V_1} X_1 \right) X_1 + C_2(W_2) \right] \quad (43)$$

$$C_2(W_2) = \min_{X_2} \mathbb{E}_2 \left[\left(P_2 + \frac{1}{2V_2} X_2 \right) X_2 + C_3(W_3) \right] \quad (44)$$

...

...

$$C_n(W_n) = \min_{X_n} \mathbb{E}_n \left[\left(P_n + \frac{1}{2V_n} X_n \right) X_n + C_{n+1}(W_{n+1}) \right] \quad (45)$$

...

...

$$C_N(W_N = W_{N-1} - X_{N-1}) = \left(P_N + \frac{1}{2V_N} W_N \right) W_N, \quad (46)$$

where $W_1 = X$, $W_n = W_{n-1} - X_{n-1}$ for $2 \leq n \leq N$ that counts the number of remaining shares to be traded at time n and $C_n(W_n)$ is the expected acquiring costs for remaining quantity W_n since the time n , excluding the costs associated with bid-ask spread.

Proposition. The solution and the value function to the optimal execution problem (42)-(46) are

$$X_n = \beta_n W_n, \quad C_n(W_n) = P_n W_n + \gamma_n W_n^2, \quad \text{for } 1 \leq n < N \quad (47)$$

with $X_N = W_{N-1} - X_{N-1}$ and $C_N(W_N) = \left(P_N + \frac{1}{2V_N} X_N \right) X_N$, where the coefficients are recursively determined as follows with $\alpha_n = 1/(2V_n)$ for $1 \leq n \leq N$ and the terminal condition $\gamma_N = 1/(2V_N)$,

$$\beta_n = \max \left(1 - \frac{\alpha_n}{2\gamma_{n+1}}, 0 \right) \quad (48)$$

$$\gamma_n = \alpha_n \beta_n + \gamma_{n+1} (1 - \beta_n)^2 \quad (49)$$

Proof of the proposition. Let $\alpha_n = \frac{1}{2V_n}$ for all $1 \leq n \leq N$, and $\gamma_N = \alpha_N$. At time $N - 1$,

$$\begin{aligned}
& C_{N-1}(W_{N-1}) \\
&= \min_{X_{N-1}} (P_{N-1} + \alpha_{N-1}X_{N-1})X_{N-1} + C_{N-1}(W_{N-1} - X_{N-1}) \\
&= \min_{X_{N-1}} (P_{N-1} + \alpha_{N-1}X_{N-1})X_{N-1} + (P_{N-1} + \alpha_{N-1}X_{N-1} + \gamma_N(W_{N-1} - X_{N-1}))(W_{N-1} - X_{N-1}) \\
&= P_{N-1}W_{N-1} + \min_{X_{N-1}} \alpha_{N-1}X_{N-1}^2 + \alpha_{N-1}X_{N-1}(W_{N-1} - X_{N-1}) + \gamma_N(W_{N-1} - X_{N-1})^2
\end{aligned}$$

The optimum occurs at $X_{N-1} = \max\left(1 - \frac{\alpha_{N-1}}{2\gamma_N}, 0\right) W_{N-1} = \beta_{N-1}W_{N-1}$, and the resulting value function is $C_{N-1} = P_{N-1}W_{N-1} + (\alpha_{N-1}\beta_{N-1} + \gamma_N(1 - \beta_{N-1})^2) W_{N-1}^2$. Letting $\gamma_{N-1} = \alpha_{N-1}\beta_{N-1} + \gamma_N(1 - \beta_{N-1})^2$ reduce the problem at time $N - 2$,

$$\begin{aligned}
& C_{N-2}(W_{N-2}) \\
&= \min_{X_{N-2}} (P_{N-2} + \alpha_{N-2}X_{N-2})X_{N-2} + C_{N-1}(W_{N-2}) \\
&= \min_{X_{N-2}} (P_{N-2} + \alpha_{N-2}X_{N-2})X_{N-2} + (P_{N-2} + \alpha_{N-2}X_{N-2} + \gamma_{N-1}(W_{N-2} - X_{N-2}))(W_{N-2} - X_{N-2}) \\
&= P_{N-2}W_{N-2} + \min_{X_{N-2}} \alpha_{N-2}X_{N-2}^2 + \alpha_{N-2}X_{N-2}(W_{N-2} - X_{N-2}) + \gamma_{N-1}(W_{N-2} - X_{N-2})^2
\end{aligned}$$

The optimum occurs at $X_{N-2} = \max\left(1 - \frac{\alpha_{N-2}}{2\gamma_{N-1}}, 0\right) W_{N-2} = \beta_{N-2}W_{N-2}$, and the resulting value function is $C_{N-2}(W_{N-2}) = P_{N-2}W_{N-2} + (\alpha_{N-2}\beta_{N-2} + \gamma_{N-1}(1 - \beta_{N-2})^2) W_{N-2}^2 = P_{N-2}W_{N-2} + \gamma_{N-2}W_{N-2}^2$. Continuing recursively until time 1 proves the proposition. \square

3.6 The strategy tested on the market

We have devised the optimal execution strategy that can utilize the property of time-varying available liquidity in limit orderbook. We apply this strategy to historical market data in order to compare the total execution cost against aforementioned strategies of evenly-splitting by [8][5] and resiliency-concerned heavily loaded at the both ends strategy by [49].

Since the level of liquidity even for the nearest future is not observable, our strategy relies on the forecastability of total available liquidity. In order not to rely too much on forecasting techniques, we simply use the three months simple moving average method. Specifically, our forecast of total (both visible and hidden) liquidity at 9:30-10:00 in all trading days of April is the average of total liquidity at 9:30-10:00 during all trading days from January to March. The forecast of particular 30 minutes interval in the days of May is the average total liquidity during the same time period in days during February-April. And the same goes for the forecast of 30 minute intervals in June.

Since we have estimated hidden liquidity for each 30 minutes interval, we set the trades interval to be 30 minutes apart. This sets setting the optimal execution problem to $N = 13$. Each child order will be placed as a form of market order, and the size of child order will consecutively consume the available liquidity from the best offer price to the next price levels under the uniform orderbook assumption until all shares of the child order are executed. For example, if 1) 100 shares are to be bought at a moment where the best offer price is \$10, 2) the size of waiting orders at the price is 30 shares, and 3) the estimated hidden liquidity at the best prices of the moment is 5 shares, then 35 shares are executed at \$10, 35 shares are executed at \$10.01, and the remaining 30 shares are executed at \$10.02.

For the comparison against other execution strategies, we use the ex-post implementation shortfall per share. The implementation shortfall is the difference between the actual revenue of executions and the ideal revenue of execution when all orders are executed at the midprice at the beginning of program. Minimizing implementation shortfall is equivalent to minimizing total acquiring cost (or, maximizing revenue for selling program). For our set of twenty stocks, we test with some portion (ranging from 11%-24%) of daily total trading volume for each stock for every trading day from April 2012 to June 2012. Although we devise the optimal executing strategy

Table 16: Implementation Shortfall (in \$ per share)

	Total shares to trade (X)	% in total daily volume	Buy program			Sell program		
			Ours	Even ([8][5])	[49]	Ours	Even ([8][5])	[49]
AAPL	2500000	14%	10.812	10.83	10.609	13.089	12.959	12.662
BA	1000000	19%	0.938	2.32	2.524	1.047	2.42	2.625
BAC	50000000	15%	0.369	0.456	0.564	0.416	0.506	0.617
BP	1000000	18%	0.419	0.507	0.57	0.519	0.612	0.669
COP	2500000	17%	1.54	3.479	3.879	1.571	3.574	3.935
CSCO	5000000	18%	0.123	0.16	0.208	0.178	0.21	0.253
CVS	1000000	11%	0.28	0.557	0.763	0.135	0.423	0.615
DELL	2500000	20%	0.06	0.109	0.164	0.105	0.14	0.192
DIS	2500000	18%	0.989	1.855	2.398	0.889	1.772	2.307
DOW	2500000	24%	0.722	1.482	1.991	0.794	1.539	2.052
JNJ	1000000	15%	0.338	0.532	0.677	0.27	0.469	0.613
JPM	5000000	11%	0.952	1.464	1.699	1.111	1.585	1.823
KO	1000000	14%	0.754	1.249	1.532	0.607	1.147	1.454
KR	1000000	20%	-0.007	0.093	0.19	0.033	0.132	0.225
MCK	500000	21%	1.522	2.138	2.228	1.078	1.691	1.815
MSFT	10000000	19%	0.509	0.716	0.917	0.553	0.758	0.974
PG	1000000	12%	0.326	0.719	0.908	0.386	0.737	0.933
TGT	1000000	22%	1.019	1.747	2.042	0.997	1.764	2.061
WFC	5000000	19%	0.821	1.204	1.417	0.825	1.219	1.442
XOM	2500000	14%	3.602	6.042	6.547	3.636	5.948	6.44
Average		17.05%	1.304	1.883	2.091	1.412	1.98	2.185

for $N = 13$, it is undesirable to place such a large sized child order at once. Hence, we split each child order of 30 minutes further into 30 equal pieces of evenly split grand-child orders that we place at every 1 minute.

Table 16 presents the average implementation shortfall during the 62 trading days from April 2012 to June 2012. Our strategy based on the time variations of available static liquidity is shown to outperform the benchmark strategies. Average saving is 29.7% over the evenly splitting strategy [8][5] and 36.5% over the strategy [49]. Figure 14 presents an example of trading programs under consideration. Our strategy is roughly characterized as ‘more execution in deeper market’. This solution is plausible in market equilibrium sense that a rational trader should devise trading program somewhat similar to the variation of the intraday liquidity. That is, if each trader does not have incentive to deviate from the market prevalent liquidity pattern,

the combined outcome of traders' action will lead equilibrium state of the intraday variation. Comparison against [49] that focuses on the resiliency aspect of liquidity, our result shows that static aspect of liquidity is still important and likely needs to be primarily considered before concerning the dynamic aspects of market liquidity.

3.7 Concluding remark

The feature of hidden liquidity has growing importance in high-frequency trading. Not with exclusive subscription-based historical data, we propose an estimation method for the average size of hidden liquidity. The estimates can be a useful source whether market participants want to build orderbook pressure model or construct intraday trading program based on price impact functions. Future extension of our study can be made by identifying other market variables and conditions that affect the innovations of liquidity level.

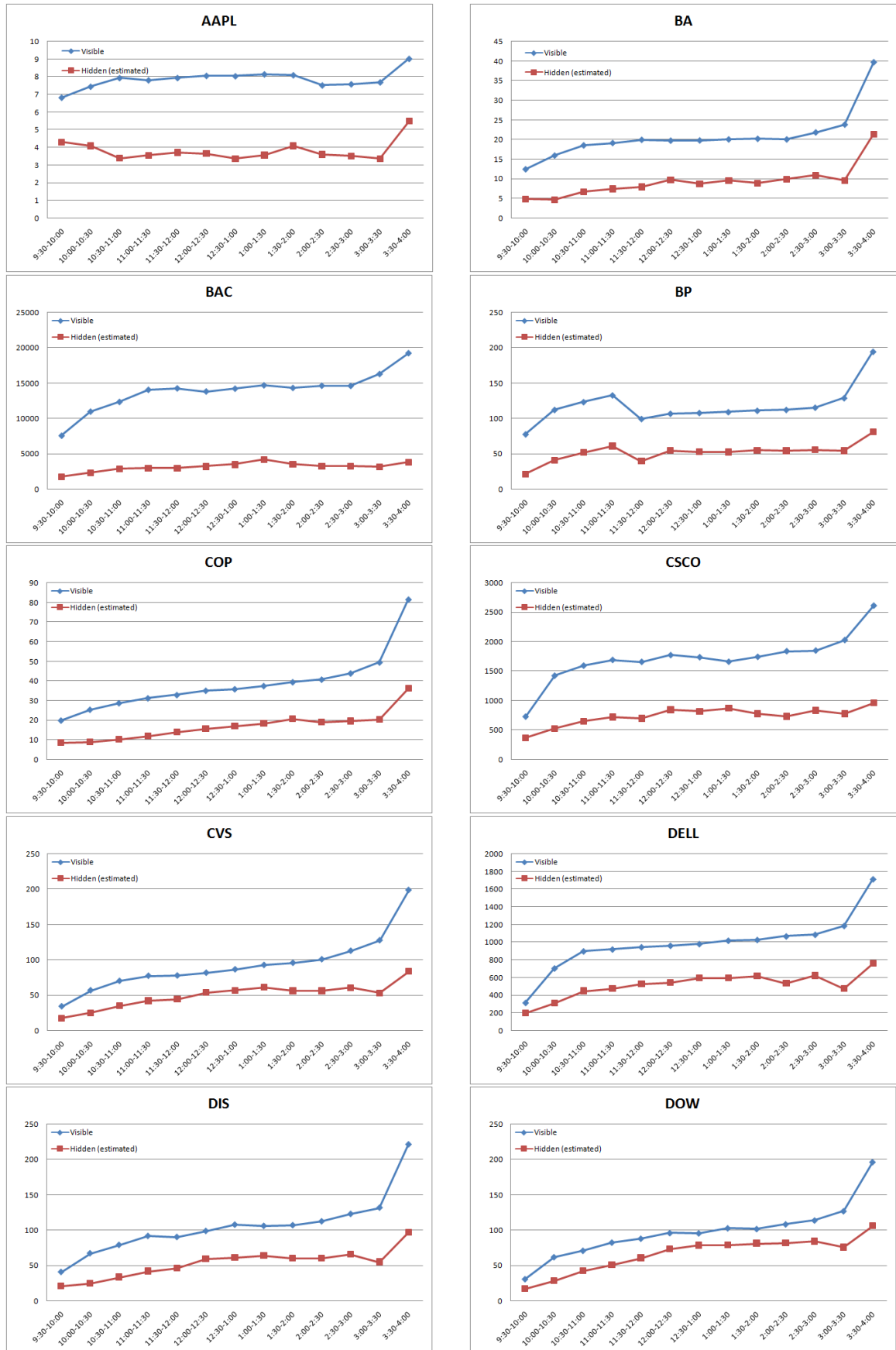


Figure 12: Intraday pattern in median of visible and hidden liquidity in 2012H1 (125 trading days) (1/2)

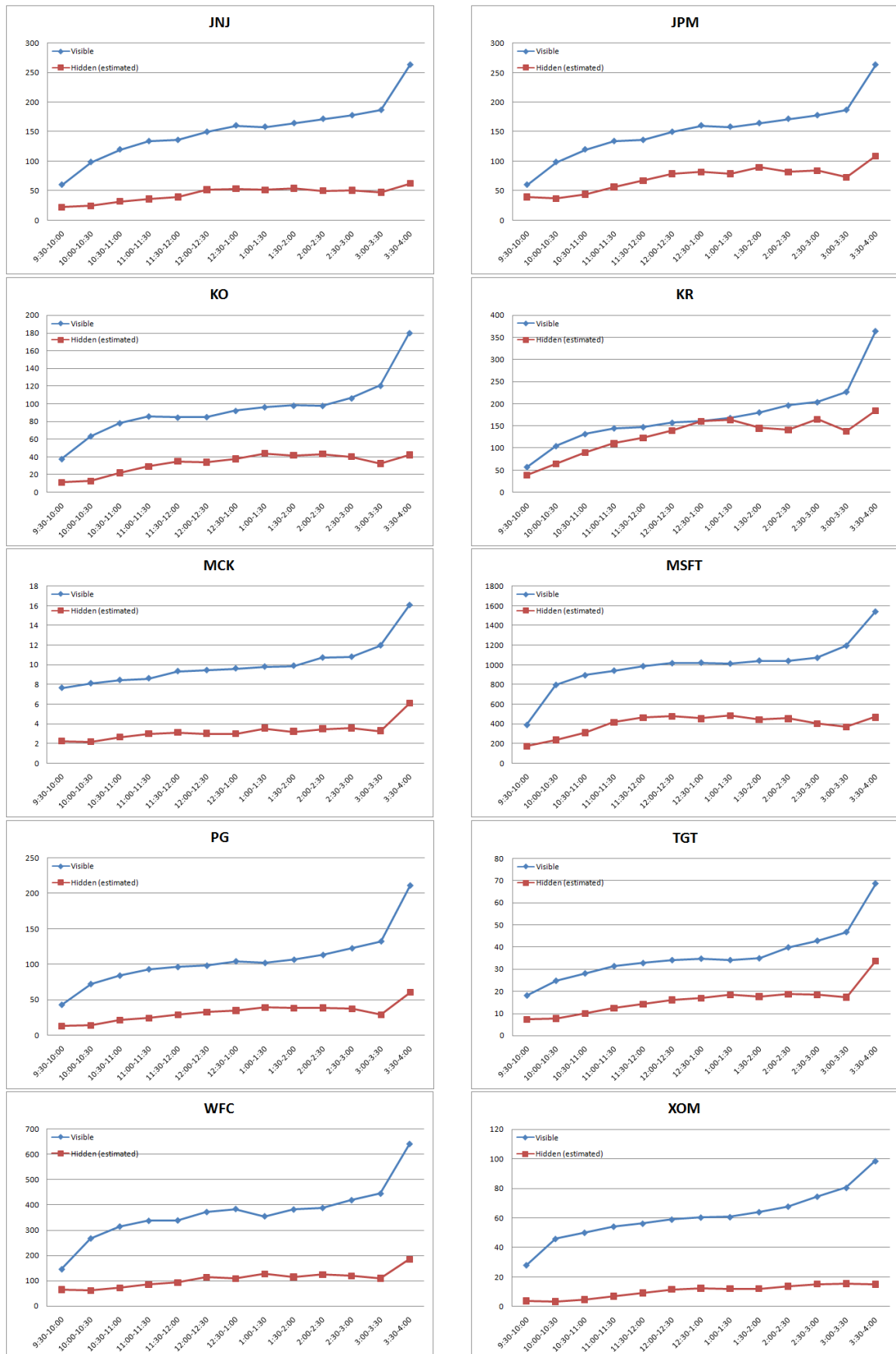


Figure 13: Intraday pattern in median of visible and hidden liquidity in 2012H1 (125 trading days) (2/2)

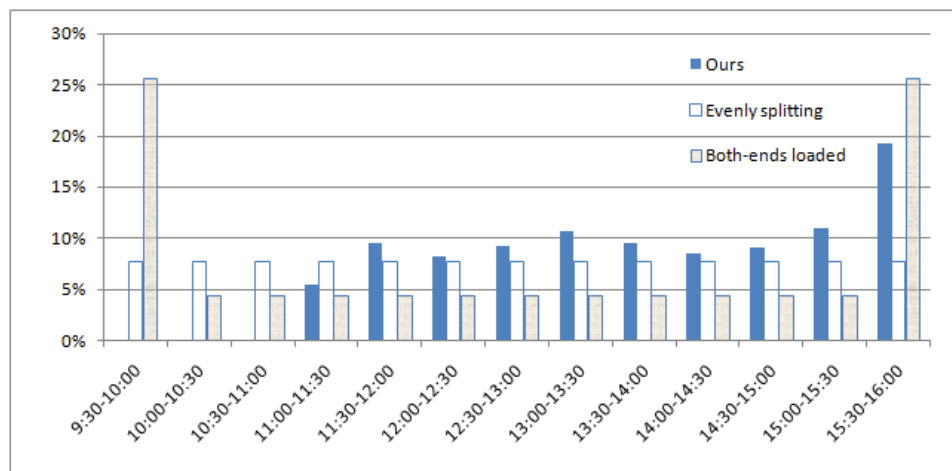


Figure 14: Example of trading program comparison - BAC, April 2012

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Fisher transformation

The Pearson's correlation coefficient $\rho_{X,Y}$ between two random variables X and Y is defined as $\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$ where μ_X and μ_Y are the means, and σ_X and σ_Y are the standard deviations. For the statistical hypothesis testing of $H_0 : \rho_{X,Y} = 0; H_1 : \rho_{X,Y} \neq 0$, test statistic $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$ is used where n is the number of samples and r is the sample correlation coefficient. However, this t-test cannot be used for null hypothesis that the correlation is equal to some non-zero value. This is because the sample correlation is not an unbiased estimator of ρ . Sampling from a normal distribution, [62] shows that the bias is almost eliminated by an estimator suggested by [30] as a remedy.

Fisher's transformation [62] of a sample correlation r to z is an inverse hyperbolic tangent function and can be written as $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$. Its standard error is given by $SE_z = \sqrt{\frac{1}{n-3}}$ and z is known to follow a normal distribution if r is from sample data following a bivariate normal distribution. Throughout this variance stabilizing transformation (i.e. SE_z is not dependent on the value of r), one can test hypothesis such as $H_0 : \rho = 0.3; H_1 : \rho \neq 0.3$ or $H_0 : \rho_1 = \rho_2; H_1 : \rho_1 \neq \rho_2$ where ρ_1 and ρ_2 are correlations from independent populations.¹ It is also possible to construct confidence interval using Fisher's transformation. This technique is built in software packages such as MATLAB and SAS.

¹For example, $H_0 : \rho_1 = \rho_2, H_1 : \rho_1 \neq \rho_2$ can be tested by test statistic $\frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$, where z_i is Fisher's transformation of sample correlation r_i and n_i is sample size for each population

Since the Fisher's Z-transformation of sample correlation coefficient value is additive, [58] recommends to obtain an average value of correlation elements through Z-transformation rather than through arithmetic average of sample correlations. The standard error of average $\frac{1}{k} \sum_{i=1}^k z(r_i)$ is given as $\frac{1}{k} \sqrt{\sum_{i=1}^k \frac{1}{n_i-3}}$ in Z-domain where each r_i is drawn from independent population of bivariate normal distribution. In our case, however, each r_i is not drawn from independent distribution and comes from each element in one correlation matrix. Only upper bound of the standard error of $\frac{1}{k} \sum_{i=1}^k z(r_i)$ can be said to be $\sqrt{\frac{1}{n-3}}$.

Based on the additive property of correlation coefficient in Z-domain, it is appropriate to consider the measures, TSC , $INSC$, and $ITSC$, in Z-domain. In this way, we have homogeneous standard error for each time series observation. Our pivotal measure for factor model, modularity also needs to be assessed in Z-domain. $MOD = z^{-1}(z(INSC) - z(ITSC))$, where $z(\cdot)$ is Fisher's Z-transformation, would be a more appropriate estimate because $INSC$ and $ITSC$ are not additive quantity, while are $z(INTC)$ and $z(ITSC)$.

A.2 MMC and modularity measure

In a graph, partitioning² is a task to identify disjoint subsets of nodes where each subset (called as *cell*) has nodes that are close to each other, and within different cells the vertices are not close to each other. This section briefly discusses the line of partitioning studies by [48][47][59] and their similarities to our modularity measure.

Most heuristic partitioning algorithms utilize top-down or bottom-up approaches with some stopping criteria. However, the aforementioned three studies contributed this line of research by setting a partitioning problem as a combinatorial problem with single objective function where solutions can be found in analytic procedure. [48]

²Interchangeable terms include *clustering* and *community structure identification* depending on the various fields of disciplines.

defines modularity for an unweighted 0-1 graph as follows.

The modularity is, up to multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random.

[47] proceeds to develop an algorithm of maximizing the modularity defined by [48] based on the spectral algorithm. [59] defines the modularity for weighted graphs by modifying the definition of [48]. The definition of modularity by [59] is, up to multiplicative constant, the sum of edge weights within groups minus the expected sum of edges weights in a equivalent network with edges placed at random. With some modulation on their modularity measure, [59] provides an algorithm for an analytic solution based on eigenvalue decomposition.

The modularities defined by [48][59] not only serve as an objective function but can measure the goodness to the specific fixed partitions. Under fixed partition structure, they measure the goodness of the partition. Therefore, they can measure the structural changes in time series with respect to cluster structure under a fixing partition. Under this motivation, our study measures the structural changes in U.S. stock market by measuring our modularity measure.

Our modularity definition is the average weight of edges for inner-groups minus average weight of edges for inter-groups. Ours and the modularity measure by [59] are motivationally similar, and relative innovations are close. The measure by [59] enjoys the analytic algorithm, but our measure is more intuitive and easier to present because we use the notion of spread between inner-group average (INSC) and inter-group average (ITSC). Also, our modularity definition extends connectivity study such as [10] and [53] by simply using correlation structure of market. By no means, we claim our modularity is always superior to the measure of [59]. We claim our measure as an intuitive way to quantify the general notion of modularity.

A.3 Model criteria in asset pricing factor models

All of Fama-French three factors and momentum factor are the returns of tradable portfolios. The *MKT* is the return of market portfolio, the *SMB* is the return of portfolio where one goes long on small stocks and goes short on big stocks, and the *HML* is the return of portfolio by going long on growth stocks and going short on non-growth stocks. The *MOM* is the return of portfolio by going long on past winner and short on past loser stocks.

Having tradable portfolio as a factor has big advantage because it would imply that excess return over risk free rate of each individual asset can be expressed by a linear combination of the tradable factors plus alpha as abnormal return. If factor model is good enough in terms of explaining individual asset or portfolio's return variation by linear combination of factors, then investors only need to consider investing on the factors without having to deal with individual stocks. Investors' only job would be to figure out the weights onto each factor according to investment objectives.

On the other hand, the factor *MOD* is the variable created by linear combination of elements of correlation matrix.³ Therefore, the factor *MOD* is not a tradable portfolio. By introducing non-tradable factor *MOD* into the four factor model, now we have (8) instead of (6). Note that (6) has α_i and (8) has β_i^0 . β_i^0 is not interpreted as 'abnormal return' because of the existence of the *MOD* term. If we were to make factor *MOD* as return form, then we must make the MOD factor something like high MOD stocks minus small MOD stocks or low MOD stocks minus high MOD stocks. However, this is impossible because MOD is not a variable for an individual asset but a market level variable, while constructions of HML and SMB are based on accounting variables of individual stocks. The exactly same issue was experienced by [51] because they defined market level liquidity as a factor. Following same approach

³Although z-transformation is applied, basically it is linear combination with intermediate usages of z-transformation based on cluster structure.

as [51], we define the decile portfolio sorted by stock’s sensitivity to the market level MOD.

Having non-tradable portfolio as a factor is often regarded as weak form in academia. This is where the most criticisms on factor model development speak such as “I have seen thousands of factor models like this,” “Learning linear regression is the license for developing factor model”. But the creation of MOD factor is based on theoretical conjecture that connectedness structure of the market should play a role in asset pricing and backed by the empirical evidence in our study. The creation is based on the statistical techniques of identifying cluster structure and processing correlation matrices accordingly.

In order to validate the linear factor model, people tend to put less focus on the R-square statistics. Instead, ones are more interested in the questions such as “Do small beta stocks yield higher expected return than high beta stocks?” so that asset pricing anomaly can be resolved. When answering such question, ones tend to group the stocks into five or ten groups so that non-systematic behavior of each stock can offset each other. Since assuming the betas of each stocks staying constant forever to the factors are not realistic assumptions, ones should allow each stock to update the sensitivities to factors and the decile portfolios to be updated periodically.

A.4 Tests with value-weighted portfolio

This section presents Table 17-19 of the test results with equal-weighted portfolio, instead of value-weighted portfolio in the main context.

A.5 Stocks with frequent appearances in both ends deciles

Table 20 and Table 21 list stocks whose number of appearances out of 20 years more than or equal to ten times and average decile less than or equal to 3 (Table 20) or

Table 17: (Equal-Weighted, corresponding to Table 2) Returns and alphas of the decile portfolio (January 1992 - December 2011)

	1	2	3	4	5	6	7	8	9	10	'1-10'
Return (p.a.)	16.34	13.81	13.55	13.08	12.97	12	11.25	11.52	11.76	11.44	4.9
s.d.	18.69	15.61	14.62	14.69	14.22	14.2	14.6	15.04	16.83	19.36	8.88
CAPM alpha	7.15	5.5	5.62	5.08	5.2	4.18	3.27	3.37	2.98	1.94	5.21
(t-statistics)	(3.36)	(3.32)	(3.52)	(3.25)	(3.27)	(2.74)	(2.11)	(2.16)	(1.73)	(0.92)	(2.61)
3-factor alpha	5.13	2.64	2.78	2.28	2.34	1.49	0.59	0.5	0.58	0.37	4.77
(t-statistics)	(3.41)	(2.26)	(2.44)	(2.06)	(2.07)	(1.35)	(0.51)	(0.47)	(0.43)	(0.23)	(2.36)
4-factor alpha	6.38	3.39	3.41	2.74	2.99	1.89	1.59	1.07	2.05	2.2	4.18
(t-statistics)	(4.46)	(2.98)	(3.03)	(2.48)	(2.69)	(1.7)	(1.47)	(1.01)	(1.71)	(1.53)	(2.06)

Table 18: (Equal-Weighted, corresponding to Table 3) Betas of decile portfolio sorted by predicted modularity betas. (January 1992 - December 2011)

	1	2	3	4	5	6	7	8	9	10	'1-10'
β_{MKT}	0.91	0.87	0.83	0.84	0.82	0.83	0.83	0.86	0.9	0.94	-0.03
(t-statistics)	(31.8)	(38.02)	(36.7)	(38.15)	(36.62)	(37.32)	(38.2)	(40.66)	(37.22)	(32.76)	(-0.71)
β^{SMB}	0.62	0.38	0.3	0.29	0.22	0.22	0.23	0.32	0.42	0.58	0.05
(t-statistics)	(17.39)	(13.33)	(10.69)	(10.51)	(7.81)	(7.98)	(8.63)	(11.91)	(13.88)	(16.12)	(0.88)
β^{HML}	0.15	0.39	0.41	0.42	0.44	0.42	0.4	0.42	0.28	0.06	0.08
(t-statistics)	(3.85)	(13)	(13.95)	(14.23)	(15.07)	(14.2)	(13.9)	(14.87)	(8.74)	(1.69)	(1.52)
β^{MOM}	-0.13	-0.08	-0.07	-0.05	-0.07	-0.04	-0.11	-0.06	-0.16	-0.19	0.06
(t-statistics)	(-5.67)	(-4.33)	(-3.66)	(-2.75)	(-3.8)	(-2.32)	(-6.01)	(-3.46)	(-7.98)	(-8.31)	(1.87)

Table 19: (Equal-Weighted, corresponding to Table 4) Composition of portfolio returns by alpha and exposures to four factors. (January 1992 - December 2011)

	1	2	3	4	5	6	7	8	9	10	'1-10'
Excess return (p.a.)	13.21	10.68	10.42	9.95	9.84	8.87	8.12	8.39	8.63	8.31	4.9
α	6.38	3.39	3.41	2.74	2.99	1.89	1.59	1.07	2.05	2.2	4.18
$\beta^{MKT} * \overline{MKT}$	5.29	5.02	4.79	4.89	4.74	4.81	4.81	5	5.2	5.46	-0.17
$\beta^{SMB} * \overline{SMB}$	1.76	1.07	0.85	0.82	0.61	0.63	0.66	0.89	1.18	1.63	0.13
$\beta^{HML} * \overline{HML}$	0.63	1.7	1.8	1.81	1.93	1.81	1.74	1.81	1.21	0.28	0.36
$\beta^{MOM} * \overline{MOM}$	-0.85	-0.52	-0.43	-0.32	-0.45	-0.27	-0.69	-0.39	-1.01	-1.25	0.4

more than or equal to 8 (Table 21).

It is still premature to draw a strong conclusion from this observations, but we feel that more economy sensitive stocks tend to be in high modularity beta category. From Table 20-Table 21, there is a little tendency that the list of high beta stocks contain stocks in IT sectors and consumer goods while the list of low beta stocks contain stocks in energy industry.

Table 20: Stocks with frequent appearances in low decile portfolios. (Stocks with low modularity beta)

Company Name	SIC code	Description	Mkt Cap. (in B\$)	Appearance	Avg. Decile
EXPRESS SCRIPTS INC	8093	Services-Specialty Outpatient Facilities, NEC	21685	14	2.29
ARCHER DANIELS MIDLAND CO	2075	Soybean Oil Mills	19104	20	2.75
SANDISK CORP	3570	Computer & office Equipment	11936	11	2.36
NEWFIELD EXPLORATION CO	1311	Crude Petroleum & Natural Gas	5078	13	2.92
SUNOCO INC	2911	Petroleum Refining	4379	20	2.85
FIDELITY NATIONAL FINANCIAL INC	6361	Title Insurance	3931	11	3.00
MICROS SYSTEMS INC	7373	Services-Computer Integrated Systems Design	3723	14	2.93
BERRY PETROLEUM CO	1311	Crude Petroleum & Natural Gas	2174	13	2.77
FILENET CORP	7372	Services-Prepackaged Software	1478	15	2.47
WHITNEY HOLDING CORP	6020	National Commercial Banks	1298	14	2.79
STONE ENERGY CORP	1311	Crude Petroleum & Natural Gas	1292	11	2.45
N L INDUSTRIES INC	2816	Inorganic Pigments	631	14	3.00
PEP BOYS MANNY MOE & JACK	5531	Retail-Auto & Home Supply Stores	580	11	2.91
CONSOLIDATED GRAPHICS INC	2752	Commercial Printing, Lithographic	501	12	3.00
WILMINGTON TRUST CORP	6712	Offices of Bank Holding Companies	412	19	2.89
LILLIAN VERNON CORP	5961	Retail-Catalog & Mail-Order Houses	60	11	2.27
MILACRON INC	3559	Special Industry Machinery	13	11	2.82
C P I CORP	7384	Services-Photofinishing Laboratories	13	17	2.71

Table 21: Stocks with frequent appearances in high decile portfolios. (Stocks with high modularity beta)

Company Name	SIC code	Description	Mkt Cap. (in B\$)	Appearance	Avg. Decile
T J X COMPANIES INC NEW	5651	Retail-Family Clothing Stores	24344	20	8.00
KOHL'S CORP	5311	Retail-Department Stores	12507	14	8.14
IMMUNEX CORP NEW	2830	bio-pharmaceutical products and services	12334	11	8.09
XEROX CORP	3577	Computer Peripheral Equipment, NEC	11041	11	8.00
ROSS STORES INC	5650	Retail-Family Clothing Stores	10867	16	8.19
WASHINGTON MUTUAL INC	6035	Savings Institution, Federally Chartered	6882	17	8.29
SYNOPSYS INC	7370	Services-Computer Programming, Data Processing, Etc.	3922	14	8.36
CONTINENTAL AIRLINES INC	4512	Air Transportation, Scheduled	3491	12	8.08
N V R INC	1531	Operative Builders	3414	13	8.15
IAC INTERACTIVECORP	4833	Television Broadcasting Stations	3286	13	8.23
ROBBINS & MYERS INC	3561	Pumps & Pumping Equipment	2223	11	8.09
BIO RAD LABORATORIES INC	3826	Laboratory Analytical Instruments	2206	14	8.29
21ST CENTURY INSURANCE GROUP	6331	Life Insurance	1940	16	8.44
APPLEBEES INTERNATIONAL INC	5810	Retail-Eating & Drinking Places	1905	13	8.00
F E I COMPANY	3826	Laboratory Analytical Instruments	1544	11	8.55
QLOGIC CORP	3670	Electronic Components & Accessories	1484	11	8.00
VISHAY INTERTECHNOLOGY INC	3674	Semiconductors & Related Devices	1292	17	8.59
FAIR ISAAC CORP	7389	Services-Business Services, NEC	1280	13	8.31
ADVENT SOFTWARE INC	7280	software products and related services	1242	11	8.18
KNIGHT TRANSPORTATION INC	4213	Trucking (No Local)	1241	11	8.00
INTERMAGNETICS	3490	Miscellaneous Fabricated Metal Products	1170	11	8.00
WALLACE COMPUTER SERVICES INC	2761	Manifold Business Forms	1097	12	8.42
PLEXUS CORP	3670	Electronic Components & Accessories	948	15	8.27
CALGON CARBON CORP	2813	Industrial Gases	890	11	8.00
JONES GROUP INC	2337	Women's, Misses', and Juniors' Suits, Skirts, and Coats	853	13	8.15
LIZ CLAIBORNE INC	2335	Women's, Misses', and Juniors' Dresses	816	17	8.88
ELECTRONICS FOR IMAGING INC	3570	Computer & office Equipment	650	14	8.36
C E C ENTERTAINMENT INC	5812	Retail-Eating Places	650	17	8.18
EMULEX CORP	3576	Computer Communications Equipment	586	12	8.75
M S C SOFTWARE CORP	7372	Services-Prepackaged Software	383	14	8.21
CALLAWAY GOLF CO	3949	Sporting & Athletic Goods, NEC	359	14	8.43
INTERVOICE INC	7370	Services-Computer Programming, Data Processing, Etc.	321	11	8.00
SUPERTEX INC	3679	Electronic Components, NEC	228	12	8.33
ADVANTA CORP	6140	Personal Credit Institutions	10	11	8.55

REFERENCES

- [1] AHN, D., CONRAD, J., and DITTMAR, R., “Basis assets,” *Review of Financial Studies*, vol. 22, no. 12, pp. 5133–5174, 2009.
- [2] AITKEN, M., BERKMAN, H., and MAK, D., “The use of undisclosed limit orders on the australian stock exchange,” *Journal of Banking and Finance*, vol. 25, pp. 1589–1603, 2001.
- [3] ALFONSI, A. and SCHIED, A., “Optimal trade execution and absence of price manipulations in limit order book models,” *Journal on Financial Mathematics*, vol. 1, pp. 490–522, 2010.
- [4] ALMGREN, R., “Optimal execution with nonlinear impact functions and trading-enhanced risk,” *Applied Mathematical Finance*, vol. 10, pp. 1–18, 2003.
- [5] ALMGREN, R. and CHRISS, N., “Optimal execution of portfolio transactions,” *Journal of Risk*, vol. 3, pp. 5–39, 2000.
- [6] ANDERSEN, A., CONT, R., and VINKOVSKAYA, E., “A point process model for the high-frequency dynamics of a limit order book,” *Working Paper*, 2010.
- [7] AVELLANEDA, M., REED, J., and STOIKOV, S., “Forecasting prices from level-i quotes in the presence of hidden liquidity,” *Algorithmic Finance*, vol. 1, pp. 35–43, 2011.
- [8] BERTSIMAS, D. and LO, A., “Optimal control of execution costs,” *Journal of Financial Markets*, vol. 1, p. 50, 1998.
- [9] BESSEMBINDER, H., PANAYIDES, M., and VENKATARAMAN, K., “Hidden liquidity: an analysis of order exposure strategies in electronic stock markets,” *Journal of Financial Economics*, vol. 94, pp. 361–383, 2009.
- [10] BILLIO, M., GETMANSKY, M., LO, A., and PELIZZON, L., “Econometric measures of connectedness and systemic risk in the finance and insurance sectors,” *Journal of Financial Economics*, 2012. doi:10.1016/j.fineco.2011.12.010.
- [11] BLOOMFIELD, R., O’HARA, M., and SAAR, G., “Hidden liquidity: some new light on dark trading,” *Working paper, Cornell University*, 2011.
- [12] BURASCHI, A., PORCHIA, P., and TROJANI, F., “Correlation risk and optimal portfolio choice,” *The Journal of Finance*, vol. LXV, no. 1, pp. 393–420, February 2010.
- [13] BURGHARDT, G., HANWECK, J., and LEI, L., “Measuring market impact and liquidity,” *The Journal of Trading*, vol. 1, no. 4, pp. 70–84, 2006.

- [14] CARHART, M., “On persistence in mutual fund performance,” *The Journal of Finance*, vol. 52, no. 1, pp. 57–82, 1997.
- [15] CEBIRROGLU, G. and HORST, U., “Determinanats and impact of hidden liquidity,” *Working paper*, 2013.
- [16] CHANDRASEKARAN, V., PARRILO, P., and WILLSKY, A., “Latent variable graphical model selection via convex optimization,” *The Annals of Statistics*, vol. 40, no. 4, pp. 1935–1967, 2012.
- [17] CHEN, L. and ZHANG, L., “A better three-factor model that explains more anomalies,” *Journal of Finance*, vol. LXV, no. 2, 2010.
- [18] COCHRANE, J., *Asset Pricing*. Princeton University Press, 2010.
- [19] CONNOR, G., GOLDBERG, L., and KORAJCZYK, R., *Portfolio Risk Analysis*. Princeton University Press, 2005.
- [20] CONT, R., “Statistical modeling of high-frequcney financial data,” *IEEE Signal Processing Magazine*, pp. 16–25, September 2011.
- [21] CONT, R. and DE LARRARD, A., “Orderbook dynamics in liquid market(limit therems and diffusion approximations,” *Working Paper*, 2012.
- [22] CONT, R., KUKANOV, A., and STOIKOV, S., “The price impact of order book events,” *Journal of Financial Econometrics*, vol. 12, no. 1, pp. 47–88, 2013.
- [23] CONT, R., STOIKOV, S., and TALREJA, R., “A stochastic model for order book dynamics,” *Operations research*, vol. 58, no. 3, pp. 549–563, 2010.
- [24] DIEBOLD, F. X. and YILMAZ, K., “On the network topology of variance decompositions: measuring the connectedness of financial firms,” *Working paper*, 2011.
- [25] ESSER, A. and MONCH, B., “The navigation of an icerberg: The optimal use of hidden orders,” *Finance research letters*, vol. 4, pp. 68–81, 2007.
- [26] FAMA, E. and FRENCH, K., “The cross-section of expected stock returns,” *Journal of Finance*, vol. 47, no. 2, pp. 427–465, 1992.
- [27] FAMA, E. and FRENCH, K., “Common risk factors in the returns on stocks and bonds,” *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.
- [28] FAMA, E. and FRENCH, K., “A five-factor asset pricing model,” *Working paper*, 2014.
- [29] FAMA, E. and MACBETH, J., “Risk, return, and equilibrium: empirical tests,” *Journal of Political Economy*, vol. 81, no. 3, pp. 607–636, 1973.

- [30] FISHER, R., “Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population,” *Biometrika (Biometrika Trust)*, vol. 10, no. 4, pp. 507–521, 1915.
- [31] FISHER, R., “The distribution of the partial correlation coefficient,” *Metron*, vol. 3, no. 3-4, pp. 329–332, 1924.
- [32] FREY, S. and SANDAS, P., “The impact of hidden liquidity in limit order books,” *CFS Working paper, Goethe University Frankfurt*, vol. 48, 2008.
- [33] GIBBONS, M., ROSS, S., and SHANKEN, J., “A test of the efficiency of a given portfolio,” *Econometrica*, vol. 57, pp. 1121–1152, 1989.
- [34] HANSEN, L., “Large sample properties of generalized method of moments estimators,” *Econometrica*, vol. 50, pp. 1029–1054, 1982.
- [35] HASBROUCK, J. and SAAR, G., “Technology and liquidity provision: the blurring of traditional definitions,” *Journal of Financial Markets*, vol. 12, pp. 143–172, 2009.
- [36] HAUTSCH, N. and HUANG, R., “On the dark side of the market: identifying and analyzing hidden order placements,” *Working Paper*, 2012.
- [37] HEWLETT, P., “Clustering of order arrivals, price impact and trade path optimization,” *Working Paper*, 2006.
- [38] HUBERMAN, G. and STANZL, W., “Price manipulation and quasi-arbitrage,” *Econometrica*, vol. 72, no. 4, pp. 1247–1275, 2004.
- [39] HUBERMAN, G. and STANZL, W., “Optimal liquidity trading,” *Review of Finance*, vol. 9, pp. 165–200, 2005.
- [40] JANG, W., LEE, J., and CHANG, W., “Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree,” *Physica A*, vol. 390, pp. 707–718, 2011.
- [41] JEGADEESH, N. and TITMAN, S., “Returns to buying winners and selling losers: Implications for stock market efficiency,” *The Journal of Finance*, vol. 48, no. 1, pp. 65–91, 1993.
- [42] LEE, C. and READY, M., “Inferring trade direction from intraday data,” *Journal of Finance*, vol. 46, pp. 733–746, 1991.
- [43] LEHMANN, B. and MODEST, D., “Diversification and the optimal construction of basis portfolios,” *Management Science*, vol. 51, no. 4, pp. 581–598, 2005.
- [44] LINTNER, J., “The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets,” *The Review of Economics and Statistics*, vol. 47, no. 1, pp. 13–37, 1965.

- [45] MATERASSI, D. and INNOCENTI, G., “Unveiling the connectivity structure of financial networks via high-frequency analysis,” *Physica A*, vol. 388, pp. 3866–3878, 2009.
- [46] MOSSIN, J., “Equilibrium in a capital asset market,” *Econometrica: Journal of the Econometric Society*, pp. 768–783, 1966.
- [47] NEWMAN, M., “Modularity and community structure in networks,” *PNAS (Proceedings of the national academy of sciences USA)*, vol. 103, no. 23, pp. 8573–8574, 2006.
- [48] NEWMAN, M. and GIRVAN, M., “Finding and evaluating community structure in networks,” *Physical Review*, vol. 69, no. 026113, 2004.
- [49] OBIZHAEVAA, A. and WANG, J., “Optimal trading strategy and supply/demand dynamics,” *Journal of Financial Markets*, vol. 16, 2013.
- [50] PARDO, A. and PASCUAL, R., “On the hidden side of liquidity,” *European journal of finance*, vol. 18, no. 10, pp. 949–967, 2012.
- [51] PASTER, L. and STAMBAUGH, R., “Liquidity risk and expected stock returns,” *The Journal of Political Economy*, vol. 111, no. 3, pp. 642–685, 2003.
- [52] POJARLIEV, M. and LEVICH, R., “Detecting crowded trades in currency funds,” *Working paper*, 2010.
- [53] POLLET, J. and WILSON, M., “Average correlation and stock market returns,” *Journal of Financial Economics*, vol. 96, pp. 364–380, 2010.
- [54] POTTERS, M. and BOUCHAUD, J., “More statistical properties of order books and price impact,” *Physica A: Statistical Mechanics and its Applications*, vol. 324, pp. 133–140, 2003.
- [55] SANDOVAL JUNIOR, L. and FRANCA, I., “Correlation of financial market in times of crisis,” *Physica A*, vol. 391, pp. 187–208, 2012.
- [56] SAVOV, A., “Asset pricing with garbage,” *Journal of Finance*, vol. LXVI, no. 1, 2011.
- [57] SHARPE, W., “Capital asset prices: A theory of market equilibrium under conditions of risk,” *The Journal of Finance*, vol. 19, no. 3, 1964.
- [58] SILVER, C. and DUNLAP, W., “Averaging correlation coefficients: Should fisher’s z transformation be used?,” *Journal of Applied Psychology*, vol. 72, no. 1, 1987.
- [59] STONE, E. and AYROLES, J., “Modulated modularity clustering as an exploratory tool for functional genomic inference,” *PLoS Genet*, vol. 5, no. 5, 2009. e1000479.doi:10.1371/journal.pgen.1000479.

- [60] WIESEL, A., ELDAR, Y., and HERO III, A., “Covariance estimation in decomposable gaussian graphical models,” *IEEE transaction on signal processing*, vol. 58, no. 3, 2010.
- [61] WINNE, R. and D’HONDT, C., “Hide-and-seek in the market: placing and detecting hidden orders,” *Review of Finance*, vol. 11, pp. 663–692, 2007.
- [62] ZIMMERMAN, D., ZUMBO, B., and WILLIAMS, R., “Bias in estimation and hypothesis testing of correlation,” *Psicologica*, vol. 24, pp. 133–158, 2003.