Härdle, Wolfgang Karl; Chen, Shi; Liang, Chong; Schienle, Melanie

Working Paper

Time-varying Limit Order Book Networks

IRTG 1792 Discussion Paper, No. 2018-016

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Härdle, Wolfgang Karl; Chen, Shi; Liang, Chong; Schienle, Melanie (2018) : Time-varying Limit Order Book Networks, IRTG 1792 Discussion Paper, No. 2018-016, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230727

IRTG 1792 Discussion Paper 2018-016

Time-varying Limit Order Book Networks

Wolfgang Karl Härdle *
Shi Chen *[2]
Chong Liang *[2]
Melanie Schienle *[2]

* Humboldt-Universität zu Berlin, Germany
*[2] Karlsruher Institut für Technologie, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Time-varying Limit Order Book Networks

Abstract

This paper analyzes the market impact of limit order books (LOB) taking cross-stock effects into account. Based on penalized vector autoregressive approach, we aim to identify significance and magnitude of the directed network channels within and between LOBs by bootstrapped impulse response functions. Moreover, information on asymmetries and imbalances within the LOB over time would be derived. For the sample of a NASDAQ blue-chip portfolio during 06-07/2016 we find that LOB network effects crucially determine prices and bid-ask asymmetries are prevalent.

JEL classification: C02, C13, C22, C45, G12

Keywords: limit order book, high dimension, generalized impulse response, high frequency, market risk, market impact, network, bootstrap

1 Introduction

Advancements in trading technologies allow an extremely rapid placement of buy and sell orders. These rapid-fire trading algorithms can make decisions in milliseconds. The dynamic changes of the high frequency (HF) limit order book (LOB) gives us vital insights into the market behavior. In an LOB shown in Figure 1, the order book contains a quantity of limit orders and the corresponding price at which you would issue a "buy" or "sell" limit order. When an investor places an order to purchase or sell a stock, there are two fundamental execution options: place the order "at market" or "at limit." The market ones are orders of purchase or sale at the best available quote. On the other hand the limit orders are not immediately executed since they are placed at a quote which is less favorable than the best quote, e.g. the second level bid/ask order. The schematic representation of an LOB reflects the local decisions and interactions between thousands of investors, and thus generates a high dimensional dynamic and complex system. Insights into this highly dynamic LOB is therefore vital for pricing of assets, but requires skillful dimension reduction techniques in combination with generalized impulse response analysis.

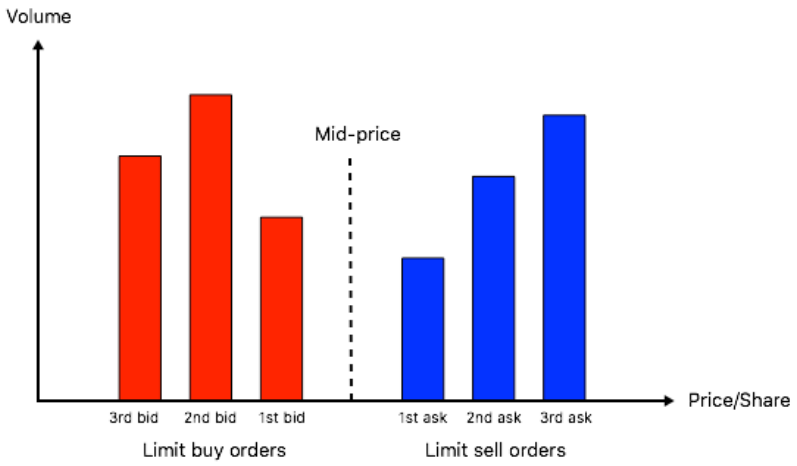

Figure 1: A simplified example of the three level LOB, with market order and first two levels of limit orders.

The limit order book has been analyzed in a variety of ways, theoretical analysis of limit

orders include Parlour and Seppi (2003), Foucault et al. (2005), Roşu (2009) etc. Empirical examples are Handa et al. (2003) and Bloomfield et al. (2005). These pieces of literature provide useful characterizations of limit orders, and discuss in detail the evolution of liquidity in an LOB market. Kavajecz and Odders-White (2004) suggests that limit orders may, in part, be informative about pockets of future liquidity. However empirical evidence on the actual market impact of limit order placements across stocks is virtually not existent, many questions of interest to regulators and traders are unsolved: i) How does the order flows interact with price dynamics, and further affect the market behavior? ii) Are the impacts on price responding to incoming ask and bid market/limit orders widely symmetric? iii) If not symmetric, how does the heterogeneous market impact caused by bid and ask order for various stocks affect the whole market? iv) How to measure the impact of market/limit order quantitatively? To address the arising questions, in this paper we conduct a comprehensive study on the interaction among price, bid and ask order sizes. LOB provides a more complicated scenario that inspires us to construct a high-dimensional object using both price and several levels of depth of order sizes with historical order flow. Of particular interest is vast directed network analysis based on the constructed high-dimensional object. The underlying assumption is that there exists a sparse representation of the data. This may help us to understand how information is impounded into price. For instance, the orders posted on the selected order levels that induce significant price impact would be treated as price drivers. In this way, investors' decision-making can be addressed by making trading price driven by order flows. The motivation to construct a network of LOB stems from the lack of both theoretical setup and empirical support.

To do so the vector autoregressive (VAR) model is without doubt one of the most useful tools that allows us to capture in a simple fashion their dynamic evolution. However it imposes challenges of high dimensionality when we incorporating a variety of time series, particularly where the vector observed at each time is high dimensional relative to the time period. Researchers have developed various penalized estimators to filter out less rel-

evant variables, key papers are on the Lasso Tibshirani (1996), SCAD Fan and Li (2001), adaptive Lasso Zou (2006), elastic net Zou and Hastie (2005), Dantzig selector Candes and Tao (2007). This paper is different from this structure of thoughts since it focuses on network connectivity, which is derived from generalized impulse response function. There has been a large literature discussing sparse VAR estimation through different penalty terms. For instance, Negahban and Wainwright (2011) imposed sparse dependence assumption on the transition matrix of VAR model and studied the theoretical properties. Kock and Callot (2015) discussed theoretical properties of Lasso and adaptive Lasso in VAR model that may reveal the correct sparsity pattern asymptotically. Basu et al. (2015) investigated theoretical properties of Lasso-type estimators for high-dimensional Gaussian processes. Wu and Wu (2016) studied the systematic theory for high-dimensional linear models with correlated errors. The Lasso-type estimators penalize the regression coefficients with the model size via a shrinkage procedure. Belloni et al. (2012) and Belloni et al. (2013) studied the post-model selection estimator that apply OLS to the first-step penalized estimators to alleviate shrinkage bias.

Diebold and Yılmaz (2014) proposed connectedness measures built from generalized forecast error variance decomposition (GFEVD) based on VAR systems, where the GFEVD is developed by Pesaran and Shin (1998) and Koop et al. (1996) with an intrinsic appeal to order-invariance. However the contributions of shares of forecast error variation in various locations do not add to unity, and it is restricted to Gaussian innovations. To solve this, we use the LN-GFEVD that has been recently proposed by Lanne and Nyberg (2016). The LN-GFEVD is economically interpretable, and can be implemented to both linear (Gaussian and non-Gaussian) or nonlinear models. To keep the sparsity structure of high-dimensional VAR estimator, we apply a bootstrap-based method rather than a moving-average (MA) transformation which is often done in fixed dimensional cases. In summary a new connectedness table is obtained, where the directed connectedness "from" and "to" are associated with the new forecast error variation for specific order book across various stocks when the arising shocks transmit from one stock to the others. This pa-

per contributes to network construction through high-dimensional VAR estimation, the resulting connectedness table facilitates convenient interpretation. At the same time, a parsimonious algorithm without MA transformation can help to improve the accuracy of final connectedness estimator.

We progress by focusing on the dynamics of LOB networks and their evolution. First, we find that the network involving the trading volumes is a better measure of the stock connectedness with the finance sector dominating the market in the sense of having a stronger influence on the others. Second, financial stocks are size-dominated, their price patterns are highly related to the market trading activity. The impact caused by ask and bid orders are statistically significant, substantial in size and significantly asymmetric. In particular, the NASDAQ market is more sensitive to the market sell pressure. Third, we investigate the LOB trading activity and find significant own-price and cross-price market impact. Moreover, we are able to identify the significant market impact caused by the arrival of a large market/limit order, and several robust risk transmission channels. Overall, our findings on the time-varying LOB networks yield a better understanding of market behavior.

The rest of the paper is organized as follows. Section 2 introduces NASDAQ LOB market and the non-synchronous LOB data, we then elaborate the data preparation based on volume-synchronization algorithm. In Section 3 we present the theoretical framework for high-dimensional VAR estimation, and construct the connectedness estimator based on our setting. The empirical results of time-varying network are illustrated in Section 4. Section 5 measures price dynamics under uncertainty shocks. Section 6 concludes, while more technical details are relegated to the Appendix.

2 Description of the Market and Data Preparation

2.1 NASDAQ Limit Order Book Market

Our sample consists of intraday trading data for selected NASDAQ stocks for the sample period spanning 1st, June 2016 to 30th, July of 2016. These data come from the LOBSTER academic data, which is powered by NASDAQ's historical TotalView using very detailed event information.

The basic structure of LOB is shown in Figure 2. The sample file has one time-stamped record for every order entered for each stock throughout the trading day. Trades are time stamped up to the nanosecond and signed to indicate whether they were initiated by a buyer or seller by the "Direction" ticker, i.e. sell trade direction are set to '-1' and buy trade direction are set to '1'. The ticker of "Event Type" indicates the trading type, for example, 1: Submission of a new order, 2: Cancellation (partial deletion of a order order), 3: Deletion (total deletion of a market/limit order), 4: Execution of a visible limit order, 5: Execution of a hidden limit order etc. Another important feature of this dataset is that each quote has been associated with trading information and limit order book. To be more specific, the k-th row in the "message" file (upper panel of Figure 2) describes the limit order event causing the change in the limit order book from line $k-1$ to line k in the "orderbook" file (lower panel).

The sample is stratified by market capitalization and industry sector. The industry breakdown of NASDAQ market is technology of 45.38%, Health Care of 11.43% and Financials of 8.42% (as of 23.02.2018). We consider a sample portfolio with 9 assets listed in Table 1, together with their market order and first two levels of limit orders, which attract the majority of trading activity, therefore becoming our research interest.

We present the summary statistics of sample dataset in Table 2. The data is collected for

Industry	Stock	Company	MktCap (billion $)
Technology	IBM	International Business Machines Corp.	171.72
	MSFT	Microsoft Corporation	499.35
	T	AT&T Inc.	257.53
Healthcare	JNJ	Johnson & Johnson	328.91
	PFE	Pfizer Inc.	206.69
	MRK	Merck & Co. Inc.	181.56
Finance	JPM	JP Morgan Chase & Co.	326.04
	WFC	Wells Fargo & Company	293.39
	C	Citigroup Inc.	168.06

Table 1: Sample data. MktCap is the market capitalization by Feb 25th, 2017.

	NumObs ($*10^3$)	AvgTrd ($*10^3$)	AvgAP1 (in $)	AvgBP1 (in $)	AvgAS1 (100 shrs)
IBM	118.25	5.82	153.07	153.04	1.92
MSFT	584.55	25.91	52.28	52.26	24.19
T	223.45	6.67	38.75	38.74	36.36
JNJ	172.77	8.17	113.99	113.98	4.11
PFE	427.51	12.49	34.83	34.82	41.96
MRK	188.84	5.82	56.70	56.68	7.43
JPM	414.35	11.49	65.48	65.46	9.47
WFC	275.29	10.91	50.90	50.89	18.02
C	472.90	12.19	46.82	46.81	14.19
	AvgBS1 (100 shrs)	AvgAS2 (100 shrs)	AvgBS2 (100 shrs)	AvgAS3 (100 shrs)	AvgBS3 (100 shrs)
IBM	2.17	1.95	2.26	2.09	2.26
MSFT	24.53	28.12	31.06	33.90	35.37
T	33.76	43.63	41.96	55.53	63.67
JNJ	3.62	5.86	4.44	7.74	4.90
PFE	42.29	48.07	48.09	50.94	55.68
MRK	7.36	14.34	11.30	24.20	13.87
JPM	9.45	13.10	11.82	17.41	15.09
WFC	17.01	20.68	17.72	23.58	19.05
C	12.97	18.58	16.48	22.23	19.60

Table 2: Summary statistics of selected stocks. *NumObs* denotes the average number of observation. *AvgTrd* is the average number of execution trades of a market/limit order. $AvgAP1$ gives the average ask price for the first order, and $AvgAS1$ represents the corresponding ask size.

Time (sec)	Event Type	Order ID	Size	Price	Direction
⋮	⋮	⋮	⋮	⋮	⋮
34713.685155243	1	206833312	100	118600	-1
34714.133632201	3	206833312	100	118600	-1
⋮	⋮	⋮	⋮	⋮	⋮

Ask Price 1	Ask Size 1	Bid Price 1	Bid Size 1	Ask Price 2	Ask Size 2	Bid Price 2	Bid Size 2	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1186600	9484	118500	8800	118700	22700	118400	14930	...
1186600	9384	118500	8800	118700	22700	118400	14930	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 2: Structure of LOBSTER data

the normal trading day involving both visible and hidden orders, which run from 9:30 a.m to 4 p.m ET. To avoid erratic effects during the market opening and closure, our sample period covers only the continuous trading periods between 9:45 and 16:00.

The main challenge in dealing with HFT data is the presence of microstructure noise arising from market frictions, where the noise-induced bias at very high sampling frequencies contaminates the observed price. Whereas infrequent sampling frequency leads to imprecise estimates, optimal sampling frequency is needed to acquire bias-variance tradeoff, see Bandi and Russell (2006), Aït-Sahalia et al. (2005), Bandi and Russell (2008). Here we implement the pre-averaging approach to exclude the impact of microstructure noise, technical details can be found in Appendix A.

2.2 Volume synchronization Algorithm

This section is devoted to the data preparation procedure by involving the order flows. We propose an algorithm that achieves volume synchronization for high-dimensional statistical setting.

As we know, the market order gets transacted at whatever price in that market, while the

limit order specify the price at which to execute the order. For larger orders placed in the market, it takes longer to fill and can actually move the market on their own. In contrast to a moderate time interval for price to reduce the microstructure noise, the time interval for trading volumes should be small enough to capture the large orders submitted by the market trader. Considering the facts, we propose a trading volume measure, size intensity, $\tilde{S}_{t_j}$ that captures the trading crowd that provides a substantial amount of liquidity at the quotes,

$$\tilde{S}_{t_j} = S_{t_j}(t_{j+1} - t_j) \tag{1}$$

where t_j denotes the time stamp of jth LOB, S_{t_j} is the corresponding tick size at t_j. By size intensity can be summed up over a given moderate time interval and therefore, matched with returns.

In the following we shall illustrate how to explicitly prepare the raw HF data. For ease of illustration, the volume synchronization algorithm can be divided into four steps,

Step 1: Set equally-spaced k time intervals starting at time T_0

$$T_0 + k\Delta T, \quad k = 0, 1, 2, \ldots, K$$

Step 2: Define the price and size at time $T_0 + k\Delta T$ as

$$\begin{aligned} \tilde{P}_{T_0+k\Delta T} &= P_{t_m},\ t_m = \max\{t_j;\ t_j \leq T_0 + k\Delta T\} \\ \tilde{S}_{T_0+k\Delta T} &= \sum_{T_0+(k-1)\Delta T \leq t_j \leq T_0+k\Delta T} S_{t_j}(t_{j+1} - t_j) \end{aligned}$$

the size variables denoted as $\tilde{S}_{T_0+k\Delta T}$ are the size intensity measure (1).

Step 3: Compute the changes of the log values,

$$\begin{aligned}\Delta p_{T_0+k\Delta T} &= \log \tilde{P}_{T_0+k\Delta T} - \log \tilde{P}_{T_0+(k-1)\Delta T} \\ \Delta s_{T_0+k\Delta T} &= \log \tilde{S}_{T_0+k\Delta T} - \log \tilde{S}_{T_0+(k-1)\Delta T}\end{aligned}$$

Step 4: Pre-averaging both $\Delta p_{T_0+k\Delta T}$ and $\Delta s_{T_0+k\Delta T}$ to remove microstructure noise,

$$\begin{aligned}\Delta \tilde{p}_{T_0+k\Delta T} &= \sum_{j=0}^{J} g_j \Delta p_{T_0+j\Delta T} \\ \Delta \tilde{s}_{T_0+k\Delta T} &= \sum_{j=0}^{J} g_j \Delta s_{T_0+j\Delta T}\end{aligned}$$

where $g_j \geq 0$ and $\sum_{j=0}^{J} g_j = 1$, the details are in Appendix A.

Preparing data in this way alleviates microstructure noise, matches the price to the size in a moderate interval and solves the problem of non-synchronicity.

For each stock, we take the mid price on the first level and the corresponding bid and ask sizes on the first three levels, i.e. market order, best limit order and 2nd best limit order. Then we construct the variable,

$$y_t^{(n)\top} = [\Delta\tilde{p}_t^{(n)}, \Delta\tilde{s}_t^{a1(n)}, \Delta\tilde{s}_t^{a2(n)}, \Delta\tilde{s}_t^{a3(n)}, \Delta\tilde{s}_t^{b1(n)}, \Delta\tilde{s}_t^{b2(n)}, \Delta\tilde{s}_t^{b3(n)}] \tag{2}$$

where $\Delta\tilde{p}_t^{(n)}$ is the prepared price factor for stock n, $\Delta\tilde{s}_t^{ar(n)}$ stands for the corresponding rth level of ask size factor, whereas $\Delta\tilde{s}_t^{br(n)}$ stands for the rth level of bid size factor for stock n.

By stacking the vector $y_t^{(n)\top}$ for different N stocks together, we define the large vector $Y_t^\top$ to estimate as

$$Y_t^\top = [y_t^{(1)\top}, y_t^{(2)\top}, \ldots, y_t^{(N)\top}] \tag{3}$$

Note that a critical assumption imposed to ensure the consistency of estimator is the observations are weakly dependence. In our setting we divide the trading period into 1-minute intervals and pre-average both $\Delta\tilde{p}_t^{(n)}$, $\Delta\tilde{s}_t^{br(n)}$ and $\Delta\tilde{s}_t^{ar(n)}$ to reduce microstructure noise over 15-min, yielding 375 observations per day.

3 Methodology

3.1 High-dimensional VAR estimation

Statistically, a high-dimensional (HD) VAR model facilitates consistent estimation and better finite-sample performance. Economically speaking, estimation results derived from a sparsity assumption help to explain the economic intuition. By incorporating the lags terms in the penalized VAR model, we aim to show the "sluggished" price adjustments caused by market/limit orders.

The standard VAR(p) model, Lütkepohl (2005) is,

$$\begin{aligned} Y_t &= A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + u_t \\ &= (A_1, A_2, \ldots, A_p) \left(Y_{t-1}^\top, Y_{t-2}^\top, \ldots, Y_{t-p}^\top\right)^\top + u_t \end{aligned} \tag{4}$$

where $Y_t = (y_{1t}, y_{2t}, \ldots, y_{Kt})^\top \in \mathbb{R}^K$ is a random vector, $t = 1, \ldots, T$. A_i are fixed $(K \times K)$ coefficient matrices. p is the lag and $u_t = (u_{1t}, u_{2t}, \ldots, u_{Kt})^\top \in \mathbb{R}^K$ is the i.i.d innovation process. In our LOB setting, dimension of $K = 7N$ with N is the number of stocks in the portfolio.

Assumption 1. *Assume* (4) *satisfies that,*

1. *The roots of* $|I_K - \sum_{j=1}^{p} A_j z^j| = 0$ *lie outside unit circle.*

2. u_t *are i.i.d innovations;*

   each element has bounded $(4+\delta)$*th moment,* $\delta > 0$.

3. $\|\Sigma_u\|_2 < \infty$ *and* $\|(A_1, A_2, \ldots, A_p)\|_2 < \infty$.

In practice, the coefficients $A_1, \ldots, A_p$ are unknown and have to be estimated from $\{Y_t\}_{t=1}^{T}$. Define,

$$\begin{aligned} Y &= (Y_1, Y_2, \ldots, Y_T) & A &= (A_1, A_2, \ldots, A_p) \\ Z_t &= (y_t, y_{t+1}, \ldots, y_{t-p+1})^\top & Z &= (Z_0, Z_1, \ldots, Z_{T-1}) \end{aligned} \tag{5}$$

Then equation (4) reads,

$$Y = AZ + U \tag{6}$$

with $U = (u_1, u_2, \ldots, u_T)$. The compact form (6) is equivalent to

$$\mathbf{y} = (Z^\top \otimes I_K)\beta + \mathbf{u} = \mathbf{x}\beta + \mathbf{u} \tag{7}$$

where the length of the parameter vector β is pK^2, the number of observations is KT.

In practice, the ration $\frac{Kp}{T}$ could be large due to high dimensionality, which deteriorates the accuracy of final estimate. Worse still, if $Kp > T$, the number of coefficients to be estimated increases quadratically in terms of the number of lags p, therefore the model cannot be identified with traditional methods such as OLS. Therefore variable selection techniques like Lasso is introduced to concentrate on a subset of non-zero parameters. For multiple time series data, especially high dimensional time series, it is preferred to use elastic net approach rather than pure Lasso to remedy potentially strong correlation among regressors. Besides, under normal assumption of error term, the upper bound of

estimated error is positively correlated in $\frac{\log(K^2p)}{T}$, part of oracle inequality. The methodologies introduced in the proceeding paragraph are of great importance in the sense that the true underlying model has a sparse representation.

The HD VAR estimates β by minimizing the objective function,

$$\arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{x}\beta\|_2^2 + \alpha_{1,T}\|\beta\|_1 + \alpha_{2,T}\|\beta\|_2^2 \right) \tag{8}$$

which is equivalent to,

$$\arg\min_{A_1, A_2, \ldots, A_P} \sum_{t=1}^{T} \|Y_t - \sum_{j=1}^{P} A_j Y_{t-j}\|_2^2 + \alpha_{1,T} \sum_{j=1}^{P} \|vec(A_j)\|_1 + \alpha_{2,T} \sum_{j=1}^{P} \|vec(A_j)\|_2^2 \tag{9}$$

where A_j is the $(K \times K)$ coefficient matrices of interest. $\alpha_{1,T}$ and $\alpha_{2,T}$ are the penalty parameters. Note that the notation $\|M\|_p$ depends on whether M is a vector or a matrix. To avoid confusion, we use $vec(M)$ here to tranform the object within $\|\|_p$ into a vector.

We choose a sequence of decreasing positive numbers $\alpha_{1,T}$ and $\alpha_{2,T}$ to control the regularization. In the case of regularization parameter is large, setting it too high will throw away useful information, whereas the estimated graph is not sparse when the α_T is small. To balance the sparsity and estimation accuracy, we choose a moderately small tuning parameter using the Bayesian information criterion (BIC). In addition, we apply OLS post-model selection estimator to the first-step penalized estimator (8) or (9) to reduce shrinkage bias and ensure better model model performance.

3.2 Structural Analysis of High-dimensional LOB Portfolio

This section discusses the effects of uncertainty shocks in the LOB. In general, uncertainty responds to all shocks through its relation to the lags of the LOB variables as specified in

the HD VAR model (8). Let us first consider the generalized impulse response function (GI) for the case of an arbitrary current shock.

For the multivariate case, following Koop et al. (1996) and Pesaran and Shin (1998), we assume shocks hitting only one equation at a time rather than all the shocks at time t. The effect on j-th equation of y_t of a one-standard deviation shock to perceived uncertainty are given by GI,

$$\begin{aligned} \delta_{jt} &: (\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^\top \sim \hat{u}^\star_{jt} e_j \\ GI(l, \delta_{jt}, \mathcal{F}_{t-1}) &= \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \end{aligned} \tag{10}$$

where $\hat{u}^\star_{jt}$ are independent draws with replacement from the set of residuals $\{\hat{u}_{jt}\}_{t=1}^T$ over the sample period, with $\{\hat{u}_{jt}\}$ is the model-implied residual of jth equation at time t. $\mathsf{E}(y_{t+1} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$ represents the expectation conditional on the history $\mathcal{F}_{t-1}$ and a fixed value of j-th shock δ_{jt} on time t at horizon l. $\mathcal{F}_{t-1}$ consists of the information used to compute the conditional expectations based on (4).

To measure the persistent effect of a shock on the behaviour of a series, the basic object in 10 is the conditional expectation. However the sparse estimation of HD VAR is non-linear, the GI functions cannot be expressed in closed form. Therefore we use bootstrap methods to produce simulated realizations that can be used to form draws from the joint distribution of shocks. The steps for computing the conditional expectations in GI are described in Appendix B.

3.3 Network Construction

The LN-GFEVD denoted as $\lambda_{ij,\mathcal{F}_{t-1}}(h)$ is defined by j-th shock hitting i-th variable at time t,

$$\lambda_{ij,\mathcal{F}_{t-1}}(h) = \frac{\sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}{\sum_{j=1}^{K} \sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}, \quad i, j = 1, \ldots, K \tag{11}$$

where h is the horizon, $\mathcal{F}_{t-1}$ refers to the history. Therefore $\lambda_{ij,\mathcal{F}_{t-1}}(h) \in [0, 1]$, measuring the relative contribution of a shock δ_{jt} to the j-th equation in relation to the total impact of all K shocks on the i-th variable in y_t after h periods. Compared to traditional GFEVD, LN-GFEVD has the attractive property that the proportions of the impact accounted for by innovations in each variable sum to unity. The LN-GFEVD is thus economic interpretable.

Many literature characterizes connectedness of the variables in the VAR systerm, for instance, Diebold and Yılmaz (2014) and Demirer et al. (2017) proposed connectedness measures built from GFEVD for both univariate and multivariate cases. However, to our knowledge, the combination of bootstrap-based GI analysis and network construction seems to be new to the literature: Upon the HD VAR estimation of (8) and (9), we use the sparsity concept that filters out less relevant variables. Instead of transforming into a MA process, which is often done in fixed dimensional cases, we apply a bootstrap-based method to compute $\lambda_{ij,\mathcal{F}_{t-1}}(h)$, then naturally produce the population connectedness, see Table 3. By this way, the connectedness table can be constructed for both linear and nonlinear models. Besides, the bootstrapped LN-GFEVD relies neither on the ordering of the variables nor on the distribution of the innovations. At the same time, a parsimonious algorithm without MA transformation can help to improve the accuracy of final connectedness estimator.

The details for computation steps can be found in Appendix B. In particular, the numer-

	x_1	x_2	...	x_K	From others
x_1	$\lambda^b_{11}(h)$	$\lambda^b_{12}(h)$	...	$\lambda^b_{1K}(h)$	$\sum_{j=1}^K \lambda^b_{1j}(h), j \neq 1$
x_2	$\lambda^b_{21}(h)$	$\lambda^b_{22}(h)$	...	$\lambda^b_{2K}(h)$	$\sum_{j=1}^K \lambda^b_{2j}(h), j \neq 2$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_K	$\lambda^b_{K1}(h)$	$\lambda^b_{K2}(h)$	...	$\lambda^b_{KK}(h)$	$\sum_{j=1}^K \lambda^b_{Kj}(h), j \neq K$
To	$\sum_{i=1}^K \lambda^b_{i1}(h)$	$\sum_{i=1}^K \lambda^b_{i2}(h)$	...	$\sum_{i=1}^K \lambda^b_{iK}(h)$	$\frac{1}{K}\sum_{i=1,j=1}^K \lambda^b_{ij}(h)$
others	$i \neq 1$	$i \neq 2$		$i \neq K$	$i \neq j$

Table 3: Connectedness table of interest, estimated by bootstrap-based methods.

ical techniques for conditional mean forecast from nonlinear models for more than one period ahead are implemented in this paper, we use bootstrap to calculate $GI(l, \delta_{jt}, \mathcal{F}_{t-1})$, see more details in Teräsvirta et al. (2010).

We then have the directional connectedness "from" and "to" associated with the forecast error variation $\lambda^b_{ij}(h)$ for a specific order book across various stock when the arising shocks transmit from one stock to the others. These two connectedness estimators can be obtained by adding up the row or column elements. Hence the pairwise directional connectedness from j to i can be written as,

$$C_{i \leftarrow j} = \lambda^b_{ij}(h) \tag{12}$$

Furthermore, the total directional connectedness "from" $C_{i \leftarrow \cdot}$ (others to i) given by

$$C_{i \leftarrow \bullet} = \sum_{j=1}^{K} \lambda^b_{ij}(h), i \neq j \tag{13}$$

equals to unity based on (11), and the total directional connectedness "to" $C_{\cdot \leftarrow j}$ (j to others) is defined as

$$C_{\bullet \leftarrow j} = \sum_{i=1}^{K} \lambda^b_{ij}(h), i \neq j \tag{14}$$

The corresponding net total directional connectedness

$$C_i = C_{to,i} - C_{from,i} = C_{\bullet \leftarrow i} - C_{i \leftarrow \bullet} \tag{15}$$

measures the direction and magnitude of the net spillover impacts.

4 Network Analysis

Upon the estimates of the sparse HD VAR model, we calculate the bootstrapped LN-GFEVD and corresponding connectedness at horizon $h = 30$ for every trading day.

4.1 Individual Stock Network

4.1.1 Pairwise Connectedness of Stocks

Let us first focus on the individual stock network to understand how the impact of a shock originating in one stock can be transmitted and amplified to the other stocks.

Basically a network can be considered as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of two core items: nodes (or vertexs) $\mathcal{V}$ and edges $\mathcal{E}$. Nodes are the entities we are evaluating and edges are the connections between them. Here we first consider the cross-stock network $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$ with only price factors $p^{(n)}$,

$$\mathcal{V}_p = p^{(n)}, \quad n = 1, \ldots, N \quad \text{and} \quad \mathcal{E}_p = C_{i \leftarrow j}, \quad i, j \in \mathcal{V}_p \tag{16}$$

We model each trading day as a separate network and extract the pairwise connectedness estimate for each stock. To understand the behavior of networks, there are various approaches for evaluating the node importance. We employ the centrality measures proposed by Freeman (1978) to evaluate the relative importance of nine stocks,

- degree centrality $deg(\mathcal{V})$: refers to the number of edges attached to one node. This is simplest measure of node connectivity, but it is can be interpreted as a form of popularity. We use "out-degree" centrality $outdeg(\mathcal{V})$, i.e. the number of ties that the node directs to others to measure the impact of "to"-connectedness, and "in-degree" centrality $indeg(\mathcal{V})$ (number of inbound links) to measure the impact of "from"-connectedness.

- closeness centrality $Clos(\mathcal{V})$: is defined as the inverse of the sum of its distances to all other nodes, it scores each node based on their closeness to all other nodes within the network. Thus we are able to identify the nodes who are best placed to influence the entire network most quickly. The more central a node is, the closer it is to all other nodes. This centrality measure will be useful to distinguish influencers in the network.

- betweenness centrality $Bet(\mathcal{V})$: quantifies the number of times a node lies on the shortest path between other nodes. Nodes that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness. This centrality measure is helpful to decide which nodes act as "bridges" between nodes in a network, and can potentially influence the spread of information through the network.

To better grasp the results, given the large amount of estimation results, we will use the summary results in tables throughout the paper. Table 4 provides the summary of the corresponding centrality measures. Citigroup, AT&T and Johnson&Johnson are central in the network, in the sense that nodes with higher "out-degree" play the role of choice maker. Meanwhile JNJ is a choice receiver with high "in-degree" value of 3.57, slightly smaller than 3.88 of Microsoft. JP Morgan and IBM are the nodes who are best placed to influence the entire network most quickly, with IBM acts as "bridge" between nodes at the same time. The above conventional centrality measures are helpful to understand the evolution of the pairwise network, but we cannot accurately classify the most important nodes demonstrating the high centrality values with above results. Even though each

measure works well for probing certain phenomena, it fails to capture the node's spreading potential, e.g. Johnson&Johnson.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{outdeg(\mathcal{V}_p)}(0.25)$	2.00	3.00	1.25	2.00	2.00	1.00	1.00	2.00	2.00
$Q_{outdeg(\mathcal{V}_p)}(0.75)$	4.00	4.00	4.00	5.00	4.00	4.00	4.00	4.00	5.00
$\mu_{outdeg(\mathcal{V}_p)}$	3.26	3.33	2.52	3.33	3.07	2.81	2.83	2.50	3.40
$Q_{indeg(\mathcal{V}_p)}(0.25)$	2.25	1.00	2.00	2.00	1.00	0.00	1.00	0.00	2.00
$Q_{indeg(\mathcal{V}_p)}(0.75)$	6.00	4.75	5.75	5.00	3.00	3.75	5.00	4.75	5.00
$\mu_{indeg(\mathcal{V}_p)}$	3.88	2.86	3.43	3.57	2.38	1.69	3.21	2.60	3.45
$Q_{Clos(\mathcal{V}_p)}(0.25)$	12.56	20.08	19.72	15.26	18.94	14.27	16.93	20.08	15.33
$Q_{Clos(\mathcal{V}_p)}(0.75)$	254.51	228.13	265.74	257.15	242.92	211.66	283.35	237.15	237.09
$\mu_{Clos(\mathcal{V}_p)}$	167.70	163.99	173.45	159.09	159.43	154.56	175.89	171.95	157.81
$Q_{Bet(\mathcal{V}_p)}(0.25)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Q_{Bet(\mathcal{V}_p)}(0.75)$	10.50	14.75	14.75	6.50	5.00	4.75	10.00	9.00	10.00
$\mu_{Bet(\mathcal{V}_p)}$	6.00	7.55	7.98	4.33	4.43	3.02	5.98	5.24	6.07

Table 4: Summary of different centrality measures for $\mathcal{G}_p$ from 06.2016 to 07.2016. $Q_{\cdot}(\alpha)$ is the quantile function, $\mu_{\cdot}$ is the mean.

4.1.2 Including Order Flows

We now investigate how the network is affected by the presence of liquidity effects, i.e. by including the order volumes in the book.

We take the first trading day after Brexit as an example. In accordance with the discussion in section 3.3, we depict the estimated full sample directional connectedness Table 3 in left panel of Figure 3. Directed connectedness are drawn as directed lines connecting two nodes. The price factor and size factors that belong to the same company appear in the same colour, the width of edges between two nodes represents the connectedness. The full sample network plot reveals that the stocks with LOB factors are massively connected, it is quite informative about the total directional connectedness of each factor. However, it is not easy to decipher all pairwise connectedness. On the right panel, each stock is a node in the network, links between nodes represent the overall "from"' and "to" impacts on the system, i.e., aggregating the connectedness measure of both price and size factors for each stock. The respective links of Citigroup and JP Morgan and Wells Fargo reveal that they are the stocks that generated highest "to"-connectedness, whereas the other six stocks are mainly risk receiver.

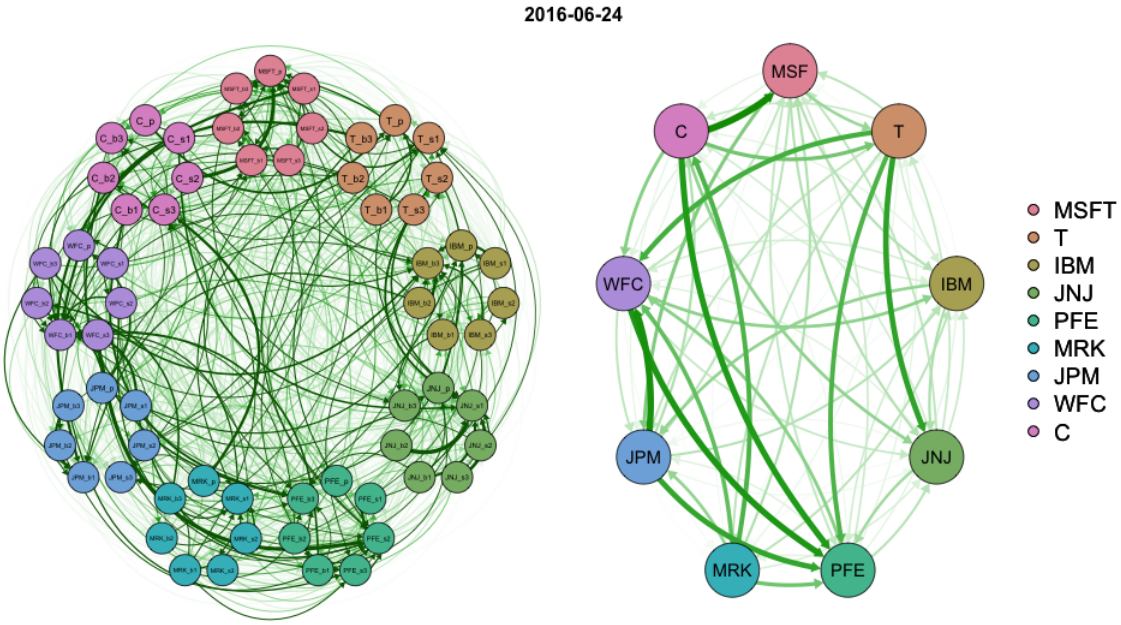

Figure 3: Left panel: the full sample network plot. Right panel: the aggregated network plot of nine stocks, on 24.06.2016

To formalize the analysis we construct the network based on (16), the aggregated individual stock network is given by $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$ consisting of,

$$\mathcal{V}_g = v_g^{(n)} \tag{17}$$

$$v_g^{(n)} = p^{(n)} + \sum_r bs_r^{(n)} + \sum_r as_r^{(n)}, \quad n = 1, \ldots, N \tag{18}$$

$$\mathcal{E}_g = C_{i \leftarrow j}, \quad i, j \in \mathcal{V}_g \tag{19}$$

where $as_r^{(n)}$ and $bs_r^{(n)}$ are the r-th level ask/bid size factors for stock n. By including the size factors from LOB, we are able to investigate how the network is affected by the presence of liquidity effects. For a network with smaller number of nodes, it is easy and appealing to identify the characteristics and patterns between individual stock.

Table 5 tabulates the centrality measures based on aggregated nine stock network, which produces different results comparing to those obtained for pairwise stock network in 4.1.1. The primary reason is that these conventional centrality measures are rarely accurate

when the majority of nodes are not highly influential in the network. Each centrality measure assess the node's importance based mostly on the path lengths and distances. The impacts caused by the less important nodes may be neglected, this will potentially cause inaccuracy and thus result in the poor performance. Therefore we use net total directional connectedness proposed in (15) as a refined centrality measure to capture the most influential spread in the following full sample network analysis.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{outdeg(\mathcal{V}_g)}(0.25)$	1.00	1.00	1.00	1.00	0.00	0.25	0.00	0.00	0.00
$Q_{outdeg(\mathcal{V}_g)}(0.75)$	114.50	131.00	115.75	111.00	110.00	103.75	105.75	98.00	105.25
$\mu_{outdeg(\mathcal{V}_g)}$	128.83	147.02	132.71	129.76	127.95	125.48	120.31	113.50	123.00
$Q_{indeg(\mathcal{V}_g)}(0.25)$	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Q_{indeg(\mathcal{V}_g)}(0.75)$	100.50	89.75	97.00	100.75	96.50	90.50	111.25	108.00	98.00
$\mu_{indeg(\mathcal{V}_g)}$	136.29	122.50	121.24	118.31	121.88	121.00	136.79	133.98	136.60
$Q_{Clos(\mathcal{V}_g)}(0.25)$	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02
$Q_{Clos(\mathcal{V}_g)}(0.75)$	0.08	0.08	0.07	0.08	0.09	0.08	0.08	0.08	0.09
$\mu_{Clos(\mathcal{V}_g)}$	0.13	0.13	0.12	0.13	0.13	0.13	0.13	0.13	0.13
$Q_{Bet(\mathcal{V}_g)}(0.25)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Q_{Bet(\mathcal{V}_g)}(0.75)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\mu_{Bet(\mathcal{V}_g)}$	4.17	3.26	2.36	2.07	2.79	2.19	3.12	3.57	3.69

Table 5: Summary of different centrality measures for $\mathcal{G}_g$ from 06.2016 to 07.2016. $Q_{\cdot}(\alpha)$ is the quantile function, $\mu_{\cdot}$ is the mean.

Specifically, the element in the connectedness table measures the total impact of all K shocks on the i-th variable, and these contributions sum to unity, which suggests the row sum of the pairwise connectedness produces one unit of "from"-connectedness for each factor, therefore the "net"-connectedness C_i is associated with "to"-connectedness and measures the share of volatility shocks to other. To understand the dynamic behavior of the risk transmission in the system, Table 6 reports the net spillover effects for each stock using the quantile functions,

$$\begin{aligned} C_i &= C_{\bullet \leftarrow i} - 2r - 1 = \sum_j C_{j \leftarrow i} - 2r - 1, \quad i, j \in \mathcal{V}_g \\ Q_{C_i}(\alpha) &= F^{-1}(\alpha) = \inf\{C_i : F(C_i) \geq \alpha\} \end{aligned} \tag{20}$$

In the table, JP Morgan is the stock with the highest "net" connectedness to others, with mean value of 0.97 over the sample period, followed by Citigroup 0.35, Wells Fargo 0.34,

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{C_i}(0.05)$	-3.19	-3.51	-3.72	-3.79	-2.84	-3.38	-2.57	-2.80	-3.35
$Q_{C_i}(0.15)$	-2.15	-3.01	-3.24	-2.99	-2.16	-2.62	-1.97	-2.30	-2.61
$Q_{C_i}(0.50)$	0.17	-0.70	-0.91	-0.50	0.14	-0.71	0.97	0.34	0.35
$Q_{C_i}(0.85)$	2.01	1.47	1.31	2.27	1.78	2.47	3.68	3.21	3.04
$Q_{C_i}(0.95)$	3.84	2.54	2.99	2.70	3.81	3.28	5.11	4.63	4.74

Table 6: The net spillover of nine-stock aggregation from 06.2016-07.2016

Microsoft 0.17, Pfizer 0.14. The "net" total connectedness of the left four stocks are all negative. As an evident result one see that that the JP Morgan is most influential in the network, while the technology companies like IBM and AT&T are main risk receivers in the aggregated system. Even though the magnitude of financial stock estimates differs to some extent, their "net"-connectedness are larger than the other in most cases. This suggests that financial companies are dominant stocks driving the networks over time. We conclude that the sign and magnitude of "net"-connectedness provide different information regarding the role for each stock in the network. The aggregated individual stock network is a better measure of how central a stock is within the network since it takes into consideration the trading volumes.

4.1.3 Total Connectedness and Volatility

We now turn our focus on time-varying pattern of the aggregated individual stock network in comparison with daily volatility estimates using the full sample high-frequency observations. Estimating volatility in this context is important as they are commonly known as proxies of market fear, a high degree of volatility is likely to correspond to increasing market risk and represent the market consensus on the expected future uncertainty.

Inspired by a voluminous literature such as Andersen et al. (2000), Andersen and Bollerslev (1998), Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2001), the realized volatility (RV) is illustrated as measure of daily volatility in high-frequency setting.

In literature, several main approaches to improve the realized volatility (RV) estimator include the preaveraging estimator of Jacod et al. (2009), the realized kernel estimator of Barndorff-Nielsen et al. (2008), the two scales estimator of Zhang et al. (2005) and multiscale estimator of Zhang et al. (2006) and Zhang (2011). Here we compute the two-scale realized variance (TS-RV) proposed by Zhang et al. (2005) as a robust estimator of the RV. The TS-RV estimator computes a subsampled RV on one slower time scale and then combine with another subsample RV calculated on a faster time scale to correct for microstructure noise.

Figure 4 compares the total net connectedness with estimated daily volatility, where the dotted lines illustrate the total connectedness estimates of (20), and the barplots indicate the TS-RV estimates. Visual inspection of the time series plots suggests, for all stocks, a rising volatility phase since the beginning of June, with the peak volatility observed around 24th of June, after that volatility decreases given the selloff in stocks following the Brexit vote followed by a rebound to record highs. The findings are consistent with the results of net connectedness measures, where a very small value of C_i is usually observed near Brexit: in other words the stocks are less connected when high market volatility occurs. In addition, we observe another peaks in volatility appear around 18th of July 2016 for three technology stocks, when Turkish shares closed down by 7.1% following the attempted coup in Turkey on 17th of July 2016. Then the volatility level come back in as the market fear caused by after coup attempt in Turkey is resolved. While important events play out, investors are likely to join a selloff as geopolitical risk is always important for decision-making in financial market. Since the peaks of volatility are generally correspond to a very low net connectedness value, this can be a signal for market investors because a peak in volatility is followed by a market rally in most cases.

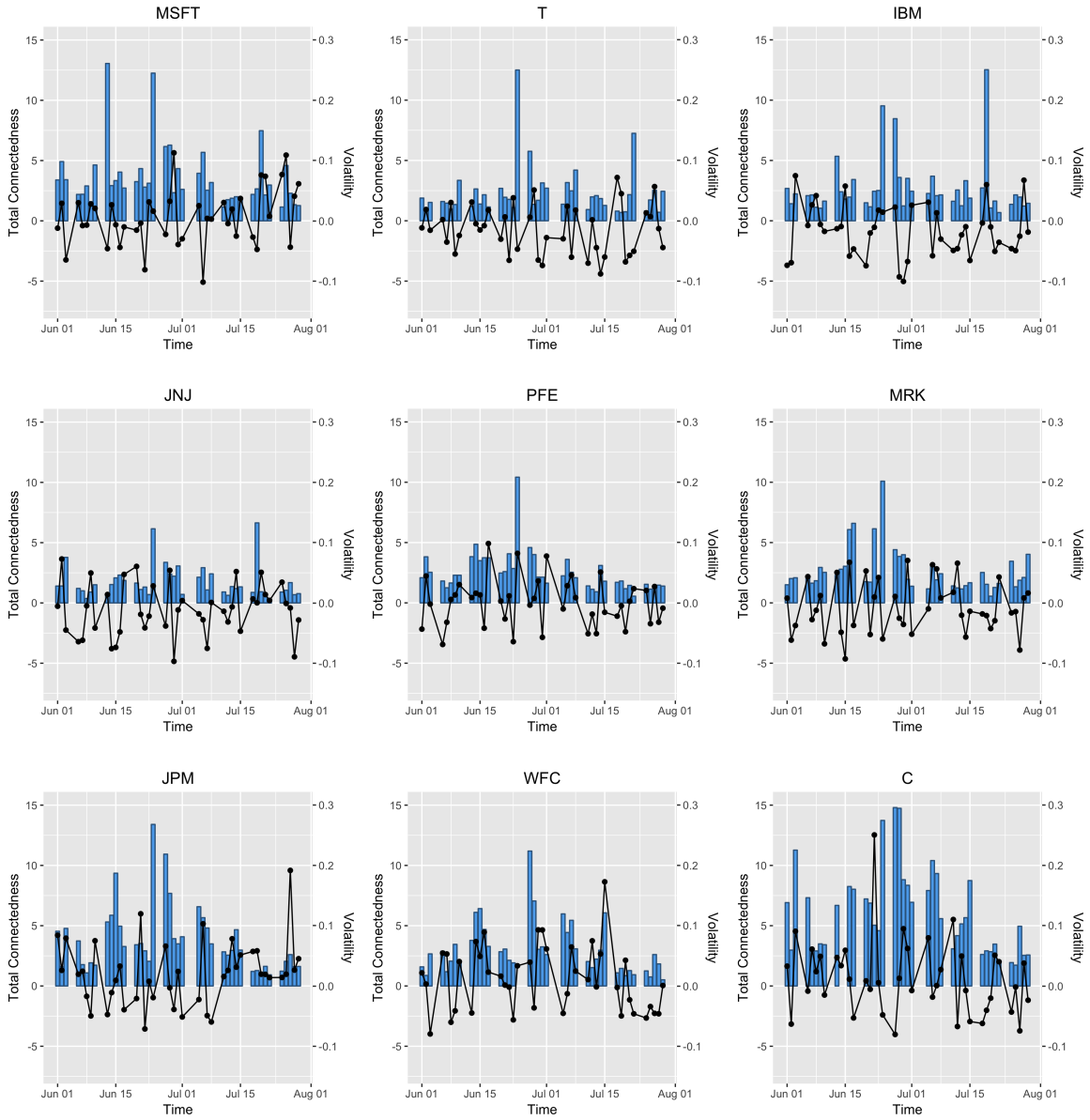

Figure 4: The time varying total net connectedness and volatility measure, 06-07.2017

4.2 Limit Order Book Network

4.2.1 Asymmetric Market Sell/Buy Pressure

Besides the purpose of studying the impacts between individual stocks, the information contained in the LOB is very valuable. Limit orders are stored in the LOB and are executed in sequence according to price priority, large trading quantities may cause a price drop or rise. The intuition behind a typical mechanism resulting in mid-price movement can be illustrated in combination with Figure 1. If there is an arrival of a market order that is sufficiently large to match all of the best bids, then the limit order will be updated with a lower best bid price.

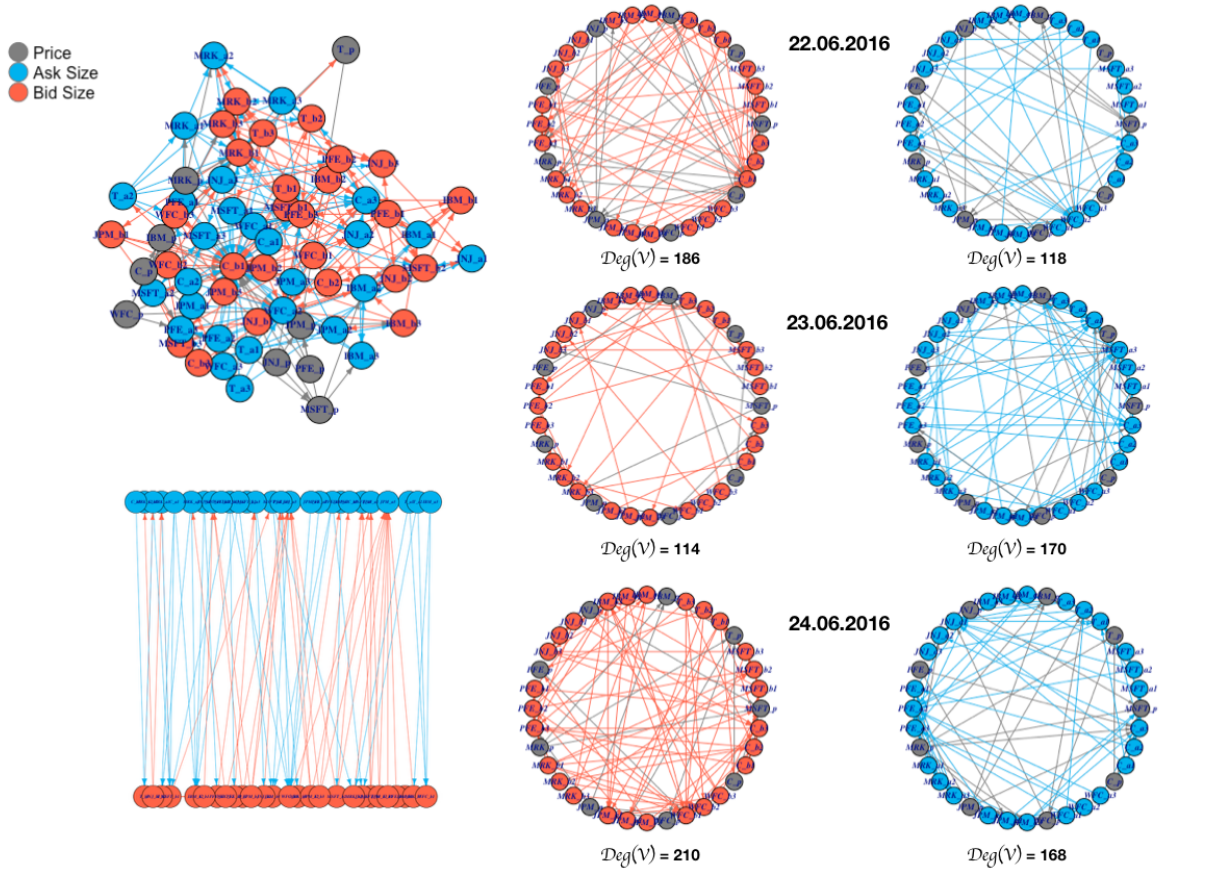

Figure 5: Plots of LOB networks from 22.06.2016-24.06.2016

Figure 5 shows the graphical display of the networks consisting of price factors, ask size factors and bid size factors, with the connectedness $C_{i \leftarrow j}$ color-coded by the type of fac-

tors that is causing the relationship, i.e., the factor j which has an impact on the others. Blue indicates the ask size factors, red indicates the bid size factors, and grey indicates the price factors. The upper left panel of Figure 5 depicts the full-sample connectedness on 22.06.2016, which is hard to decipher important pairwise connectedness. Therefore we decompose the full-sample connectedness into two parts, the price&ask size connectedness graph and price&bid size connectedness graph as shown in colored circles on the right panel. It shows how the LOB network changed during Brexit announcement, we typically observe changes in the behavior of bid size factors. The price&bid size factor network is less connected on 23.06.2016, while the price&ask size factor network is slightly tightly connected on the same day. This result could indicate that, when there is a risk caused by political uncertainty, the buying pressure is much weaker and selling pressure slightly stronger.

The impacts on returns respond to ask and bid limit orders are not symmetric. Recent studies have showed that limit orders and cancelations, not just trades, have a tangible effect on prices, see Hautsch and Huang (2012) and Eisler et al. (2012). Building on these ideas, we construct a graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$ to study the asymmetric impact from aggregated size factors to price factors,

$$\mathcal{V}_s = \left(p^{(n)}, \sum_n bs_r^{(n)}, \sum_n as_r^{(n)} \right) \quad n = 1, \ldots, N \tag{21}$$

$$\mathcal{E}_s = C_{i \leftarrow j} \quad i \in \{p^{(n)}\}, \quad j \in \left\{ \sum_n bs_r^{(n)}, \sum_n as_r^{(n)} \right\} \tag{22}$$

In Table 7 we provide the summary of $\mathcal{E}_s$ in (22), i.e. impacts from aggregated size factors to the stock price factor. The higher are the values in this table, the stronger are the stocks affect by trading activities over time. We notice that JP Morgan on average is more likely to be affected by ask side trading activity, while Wells Fargo is most sensitive to the bid side trading activity. In addition, both best bid and ask limit orders (i.e. 2nd level of ask/bid size) exhibit opposite results, with JP Morgan and Wells Fargo are less

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{C_{p^{(n)}\leftarrow\sum as1}}(0.25)$	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.00
$Q_{C_{p^{(n)}\leftarrow\sum as1}}(0.75)$	0.31	0.41	0.39	0.26	0.42	0.34	0.72	0.45	0.43
$\mu_{C_{p^{(n)}\leftarrow\sum as1}}$	0.44	0.43	0.45	0.34	0.45	0.68	0.86	0.53	0.83
$Q_{C_{p^{(n)}\leftarrow\sum as2}}(0.25)$	0.02	0.02	0.00	0.01	0.01	0.02	0.01	0.01	0.01
$Q_{C_{p^{(n)}\leftarrow\sum as2}}(0.75)$	0.61	0.50	0.58	0.43	0.26	1.00	0.30	0.24	0.40
$\mu_{C_{p^{(n)}\leftarrow\sum as2}}$	0.53	0.41	0.59	0.48	0.60	0.65	0.33	0.50	0.64
$Q_{C_{p^{(n)}\leftarrow\sum as3}}(0.25)$	0.02	0.02	0.00	0.01	0.01	0.02	0.03	0.01	0.02
$Q_{C_{p^{(n)}\leftarrow\sum as3}}(0.75)$	0.71	0.52	0.54	0.34	0.37	0.50	0.88	0.45	0.46
$\mu_{C_{p^{(n)}\leftarrow\sum as3}}$	0.90	0.69	0.54	0.33	0.58	0.89	1.03	0.54	0.39
$Q_{C_{p^{(n)}\leftarrow\sum bs1}}(0.25)$	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01
$Q_{C_{p^{(n)}\leftarrow\sum bs1}}(0.75)$	0.11	0.34	0.21	0.20	0.23	0.28	0.26	0.66	0.42
$\mu_{C_{p^{(n)}\leftarrow\sum bs1}}$	0.12	0.44	0.47	0.40	0.32	0.41	0.46	0.61	0.40
$Q_{C_{p^{(n)}\leftarrow\sum bs2}}(0.25)$	0.00	0.02	0.00	0.00	0.01	0.01	0.01	0.02	0.01
$Q_{C_{p^{(n)}\leftarrow\sum bs2}}(0.75)$	0.26	0.29	0.31	0.11	0.46	0.27	0.33	0.16	0.26
$\mu_{C_{p^{(n)}\leftarrow\sum bs2}}$	0.50	0.35	0.38	0.20	0.37	0.22	0.48	0.20	0.28
$Q_{C_{p^{(n)}\leftarrow\sum bs3}}(0.25)$	0.01	0.02	0.00	0.01	0.00	0.01	0.01	0.03	0.01
$Q_{C_{p^{(n)}\leftarrow\sum bs3}}(0.75)$	0.43	0.40	0.15	0.20	0.32	0.21	0.49	0.71	0.45
$\mu_{C_{p^{(n)}\leftarrow\sum bs3}}$	0.63	0.37	0.24	0.39	0.51	0.38	0.43	0.73	0.47

Table 7: Summary of the aggregated impacts from size factors to the stock price factor from 06.2016-07.2017. $Q_{\cdot}(\alpha)$ is the quantile function, $\mu_{\cdot}$ is the mean.

likely to be affected by best ask and bid limit order respectively. We find this result very interesting, because it brings into question how best limit orders are correlated with the order flow preceding their arrival and therefore have very little impacts on the price. This may be explained by the assumption that both market and limit orders tend to drive prices, while prices tend to impact best limit orders and their cancellations in the book. We conclude that the financial stocks are size-dominated stocks, their price patterns are highly related to the market trading activity. When selling pressure increases, the Health-care stocks are more stable. While the technology stocks appear to be more stable for buying pressure.

It follows that the depth of the book at which limit orders are submitted is driving the price. Accordingly, we calculate the impacts from the aggregated ask/bid size factors to

the aggregated price factors given by,

$$C_{\sum_N p \leftarrow \sum_N s_r^{(n)}} = \sum_{i=1}^{N} C_{i \leftarrow j} \tag{23}$$

$$i \in \{p^{(n)}\}, \quad j \in \left\{ \sum_n bs_r^{(n)}, \sum_n as_r^{(n)} \right\} \tag{24}$$

Table 8 compares the aggregated impacts for six types of size factors in our study. The impacts on return (aggregated price factors) respond to incoming ask and bid market/limit orders are not symmetric. In general, the impacts from ask orders are larger than the bid orders, ranging from the lowest value of 0.30 for aggregated impacts of bs_2 to the highest value of 0.59 for as_3 on average. Please note that this results are consistent with the results of Table 7, indicating that we can observe stronger impacts on prices caused by market sell pressure.

	$Q_C(0.25)$	$Q_C(0.50)$	$Q_C(0.75)$	μ_C
$C_{\sum_N p \leftarrow \sum_N as_1^{(n)}}$	0.12	0.27	0.67	0.50
$C_{\sum_N p \leftarrow \sum_N as_2^{(n)}}$	0.19	0.30	0.55	0.47
$C_{\sum_N p \leftarrow \sum_N as_3^{(n)}}$	0.17	0.35	0.83	0.59
$C_{\sum_N p \leftarrow \sum_N bs_1^{(n)}}$	0.09	0.18	0.39	0.36
$C_{\sum_N p \leftarrow \sum_N bs_2^{(n)}}$	0.08	0.16	0.41	0.30
$C_{\sum_N p \leftarrow \sum_N bs_3^{(n)}}$	0.13	0.29	0.63	0.42

Table 8: Summary of the impacts from aggregated size factors to the aggregated price factor from 06.2016-07.2017. $Q_\cdot(\alpha)$ is the quantile function, $\mu_\cdot$ is the mean.

More precisely, let μ_1 be the mean of the overall impacts from selling orders over the sample period ($T = 42$), and μ_2 the corresponding mean of the overall impacts from buying orders, i.e.,

$$\mu_1 = \frac{1}{3T}\left(C_{t,\sum_N p \leftarrow \sum_N as_1^{(n)}} + C_{t,\sum_N p \leftarrow \sum_N as_2^{(n)}} + C_{t,\sum_N p \leftarrow \sum_N as_3^{(n)}} \right) \tag{25}$$

$$\mu_2 = \frac{1}{3T}\left(C_{t,\sum_N p \leftarrow \sum_N bs_1^{(n)}} + C_{t,\sum_N p \leftarrow \sum_N bs_2^{(n)}} + C_{t,\sum_N p \leftarrow \sum_N bs_3^{(n)}} \right) \tag{26}$$

therefore the hypothesis of interest can be expressed as,

$$H_0 \ : \ \mu_1 - \mu_2 = 0$$
$$H_a \ : \ \mu_1 - \mu_2 > 0$$

Table 9 suggests that both the pooled t-test and the Welsh t-test give roughly the same results. Since the p-value is very low, we reject the null hypothesis, indicating that there is strong evidence of a significant larger impact from selling orders in the market.

	t-statistics	p-value
Pooled t-test	2.7557	0.003144
Welsh t-test	2.7557	0.003168

Table 9: Comparison of two hypothesis tests, when assuming/not assuming equal standard deviation.

4.2.2 Own-price and Cross-price Market Impact

The discussion in section 4.2.1 concludes that the impacts on return respond to different level of depth of the book are widely asymmetric. In this subsection we provide further empirical evidence of own-price and cross-price market impact at the level on the individual stock. First, we analyze the market impacts of their own trades for each stock, and then we undertake a detailed analysis of the impact of trades in one stock on the prices of other stocks.

At first, we consider own-price market impact for different levels of depth of the book for the selected stocks, i.e. the own-price market impacts are caused by their own order flows. The results are presented in Table 10. Based on the averaged connectedness over two months, JP Morgan receives highest market impact from its own ask orders, especially when the orders are placed in the market order or the 2nd best limit order. Even though Wells Fargo and Microsoft are the two stocks receiving highest market impacts from their own bid trades, the market impacts from their ask trades are high as well. In addition,

the Johnson&Johnson responds weakly to both ask orders and bid orders.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$\mu_{C_{p^{(n)}\leftarrow as1^{(n)}}}$	0.47	0.90	1.28	0.06	0.31	2.95	3.58	2.47	0.26
$\mu_{C_{p^{(n)}\leftarrow as2^{(n)}}}$	0.34	0.31	0.47	0.13	0.07	1.42	0.38	2.17	1.43
$\mu_{C_{p^{(n)}\leftarrow as3^{(n)}}}$	2.57	0.26	1.30	0.25	1.69	0.90	5.26	0.74	0.32
$\sum \mu_{C_{p^{(n)}\leftarrow as^{(n)}}}$	3.38	1.47	3.05	0.44	2.07	5.27	9.22	5.38	2.01
$\mu_{C_{p^{(n)}\leftarrow bs1^{(n)}}}$	0.09	0.59	0.14	0.05	0.60	1.14	1.12	2.47	0.70
$\mu_{C_{p^{(n)}\leftarrow bs2^{(n)}}}$	1.35	0.18	0.29	0.05	0.43	0.08	0.83	0.89	0.42
$\mu_{C_{p^{(n)}\leftarrow bs3^{(n)}}}$	2.58	0.07	0.23	1.20	0.11	2.16	1.37	2.14	1.42
$\sum \mu_{C_{p^{(n)}\leftarrow bs^{(n)}}}$	4.02	0.84	0.66	1.30	1.14	3.38	3.32	5.50	2.54

Table 10: The mean of own-price market impacts caused by market orders $\{as1^{(n)}, bs1^{(n)}\}$, best limit orders $\{as2^{(n)}, bs2^{(n)}\}$ and 2nd best limit orders $\{as3^{(n)}, bs3^{(n)}\}$ for each stock n from 06.2016-07.2017. All numbers are multiplied by 100. $\mu_{\cdot}$ is the mean.

In contrast to (21), we measure the cross-price market impacts by adding up the impacts from all ask/bid orders for each stock. The graph we construct is denoted as $\mathcal{G}_{cross} = (\mathcal{V}_c, \mathcal{E}_c)$, with cross-stock market impacts from the aggregated size factors to the price factor given by,

$$\mathcal{V}_c = \left(p^{(m)}, \sum_r bs_r^{(n)}, \sum_r as_r^{(n)}\right) \tag{27}$$

$$\mathcal{E}_c = C_{i\leftarrow j} \quad i \in \{p^{(m)}\}, \quad j \in \left\{\sum_r bs_r^{(n)}, \sum_r as_r^{(n)}\right\} \tag{28}$$

$$m, n \in \{1, \ldots, N\} \quad r = 1, 2, 3 \quad m \neq n \tag{29}$$

When $j \in \left\{\sum_r as_r^{(n)}\right\}$ in (29), we compare the cross-price market impacts of ask trades for each stock in Table 11. Obviously, the diagonal elements measuring the market impacts of their own trades are the same as $\sum \mu_{C_{p^{(n)}\leftarrow as^{(n)}}}$ summerised in Table 10. We observe three large values on the diagonal, indicating that JP Morgan, Merck and Wells Fargo have higher own-price market impacts than cross-price market impacts. Furthermore, JP Morgan is the stock with the highest cross-price market impact to Microsoft and Citigroup. IBM receives stronger cross-price market impact from Wells Fargo and

Citigroup. The price of Pfizer is more sensitive to the ask order flows of Merck and Wells Fargo. Therefore we conclude that the stock price can be affected not only by their own ask order flows, but also by the ask order flows of financial stocks.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$\mu_{C_{p^{(MSFT)}\leftarrow\sum_r as_r^{(n)}}}$	3.38	0.68	1.82	0.65	1.76	0.73	5.46	0.54	3.66
$\mu_{C_{p^{(T)}\leftarrow\sum_r as_r^{(n)}}}$	3.08	1.47	1.00	0.62	1.86	2.58	1.08	1.29	2.38
$\mu_{C_{p^{(IBM)}\leftarrow\sum_r as_r^{(n)}}}$	1.52	0.38	3.06	1.33	0.91	1.57	1.41	3.58	2.05
$\mu_{C_{p^{(JNJ)}\leftarrow\sum_r as_r^{(n)}}}$	1.69	0.62	1.04	0.45	1.47	1.05	1.37	0.31	3.49
$\mu_{C_{p^{(PFE)}\leftarrow\sum_r as_r^{(n)}}}$	1.07	0.96	0.44	0.13	2.06	4.83	1.86	2.89	2.12
$\mu_{C_{p^{(MRK)}\leftarrow\sum_r as_r^{(n)}}}$	3.18	1.17	0.43	0.83	2.44	5.27	4.15	2.25	2.57
$\mu_{C_{p^{(JPM)}\leftarrow\sum_r as_r^{(n)}}}$	2.09	1.10	1.81	0.72	2.68	1.13	9.22	1.34	2.16
$\mu_{C_{p^{(WFC)}\leftarrow\sum_r as_r^{(n)}}}$	1.22	2.38	1.70	0.55	1.93	1.22	0.79	5.37	0.60
$\mu_{C_{p^{(C)}\leftarrow\sum_r as_r^{(n)}}}$	2.55	1.11	2.37	0.84	2.57	1.33	4.51	1.23	2.01

Table 11: The mean of the market impacts caused by ask orders of stock m for each stock n. All numbers are multiplied by 100. $\mu.$ is the mean.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$\mu_{C_{p^{(MSFT)}\leftarrow\sum_r bs_r^{(n)}}}$	4.02	2.26	0.62	0.53	0.59	1.67	0.41	0.89	1.61
$\mu_{C_{p^{(T)}\leftarrow\sum_r bs_r^{(n)}}}$	1.36	0.84	0.22	1.04	1.41	3.67	0.92	1.10	1.03
$\mu_{C_{p^{(IBM)}\leftarrow\sum_r bs_r^{(n)}}}$	0.79	1.29	0.66	0.13	0.58	0.97	3.47	1.85	1.15
$\mu_{C_{p^{(JNJ)}\leftarrow\sum_r bs_r^{(n)}}}$	0.63	0.85	0.30	1.30	0.86	0.99	0.50	1.90	2.59
$\mu_{C_{p^{(PFE)}\leftarrow\sum_r bs_r^{(n)}}}$	2.12	0.36	1.10	0.19	1.13	0.37	1.43	4.08	1.23
$\mu_{C_{p^{(MRK)}\leftarrow\sum_r bs_r^{(n)}}}$	0.72	0.49	0.25	0.25	1.35	3.37	1.59	0.84	1.27
$\mu_{C_{p^{(JPM)}\leftarrow\sum_r bs_r^{(n)}}}$	1.66	0.47	1.25	0.97	1.39	0.59	3.32	1.87	2.16
$\mu_{C_{p^{(WFC)}\leftarrow\sum_r bs_r^{(n)}}}$	1.99	1.29	0.30	0.73	1.37	0.83	1.75	5.50	1.67
$\mu_{C_{p^{(C)}\leftarrow\sum_r bs_r^{(n)}}}$	1.02	1.80	0.42	1.24	0.84	1.08	1.41	1.12	2.54

Table 12: The mean of the market impacts caused by bid orders of stock m for each stock n. All numbers are multiplied by 100. $\mu.$ is the mean.

We proceed with the summary of the market impacts of bid trades for each stock when $j \in \left\{\sum_r bs_r^{(n)}\right\}$. Table 12 reports the results. The table reveals that financial stocks have stronger cross-price market impacts compared with healthcare and technology stocks. For

example, the bid trades of Citigroup and Well Fargo have strong cross-price market impact on Johnson & Johnson, IBM receives stronger cross-price market impact from the bid order flows of JP Morgan and Wells Fargo.

So far we analyze the individual stock network with and without the order flows in the book, the network study enables us to investigate the interaction between order flows and price dynamics. Furthermore, we discover both bid and ask trading volumes of the limit order book affect the price. Hence we are able to answer the first three questions proposed in the very beginning, i) How does the order flows interact with price dynamics? ii) Are the impacts on return responding to incoming ask and bid limit orders widely symmetric? iii) If not symmetric, how does the heterogeneous market impact caused by bid and ask order for various stocks affect the whole market? Our model has implied that in an LOB market, the huge sell/buy volume queued on the ask/bid side could induce strong sell/buy pressure on the market and therefore changing the price. In the following, we will focus on the last question, iv) How to measure the impact of market/limit order quantitatively?

5 Measuring Price Direction under Uncertainty Shock

When a large market order to buy or sell a stock arrives, the market order will automatically execute, this causes a temporary market impact. Even though sufficiently large market order immediately affects the price direction, the bid/ask sizes alone do not give enough information on price direction. To solve this, we use structural analysis proposed in section 3.2 to measure the persistent effect of shock in the LOB. In this section, our aim is to gain some insights into the details of the price formation and explore the existence of arbitrage opportunities.

To measure the impacts of market/limit order and whether the impacts identified by our model are temporary or robust over time, we resort to generalized impulse response

analysis similar to the GI defined in (10). However we assume a unit shock hitting only one equation at a time, its impact on jth equation of y_t is the following,

$$\begin{aligned} \delta_{jt} &: (\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^\top \sim e_j \\ GI(l, \delta_{jt}, \mathcal{F}_{t-1}) &= \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \end{aligned} \tag{30}$$

where $\mathsf{E}(y_{t+1} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$ represents the expectation conditional on the history $\mathcal{F}_{t-1}$ and a fixed value of j-th shock δ_{jt} on time t at horizon l. $\mathcal{F}_{t-1}$ consists of the information used to compute the conditional expectations based on bootstrap method.

Our starting point is based on market impacts regarding their own trading activities. To measure the market impacts of the order flows on price factor at a given horizon l for a stock n, the response of price factor $\Delta\tilde{p}_t^{(m)}$ are quantified by equation (30) when the shock δ_{jt} is treated as one of the size factors ($\Delta\tilde{s}_t^{a1(n)}$, $\Delta\tilde{s}_t^{a2(n)}$, $\Delta\tilde{s}_t^{a3(n)}$, $\Delta\tilde{s}_t^{b1(n)}$, $\Delta\tilde{s}_t^{b2(n)}$, $\Delta\tilde{s}_t^{b3(n)}$) hitting the system. With a moderate sparse structure selected by BIC after post-LASSO, we are able to identify not only the existence of significant market impact, but also the pattern of own-price market impact when $m = n$ and cross-price market impact when $m \neq n$. Here we use $l = 30min$ and calculate the corresponding bootstrapped GI estimation for every trading day.

We identify in total 10 days where there are significant own-price market impacts for Wells Fargo. As an example, Figure 6 depicts the result on 25th of July. We observe a negative correlation between the magnitude of its ask market order and price factor. It is normal for financial market in the sense that the investors will start marking down their bid price when there is a wave of sell orders coming into the order book. As expected, the price (average of bid and ask quotes) factor tends to decrease significantly after the arrival of a large ask market order. This argument holds for the case of bid market order as well. In Figure 7, we observe a positive market impact from bid market order on 19th of July, 2016. Both impacts can last for almost 10 minutes before the price

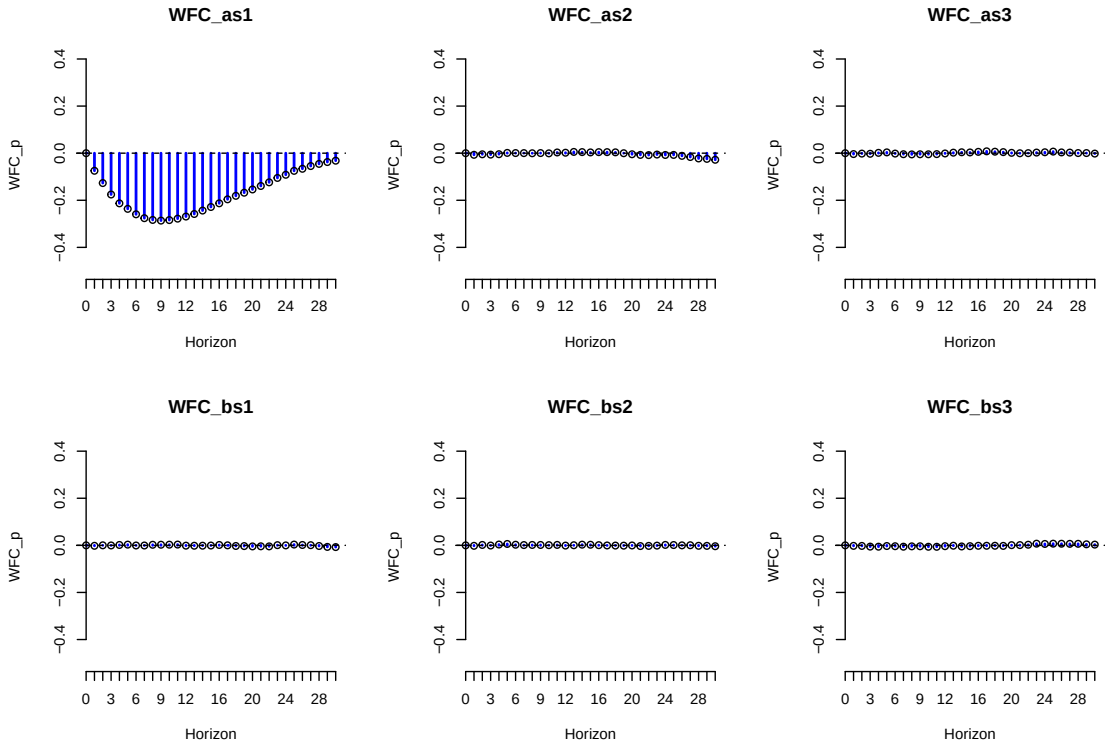

Figure 6: Own-price market impact of WFC (Wells Fargo) on 25th of July, 2016.

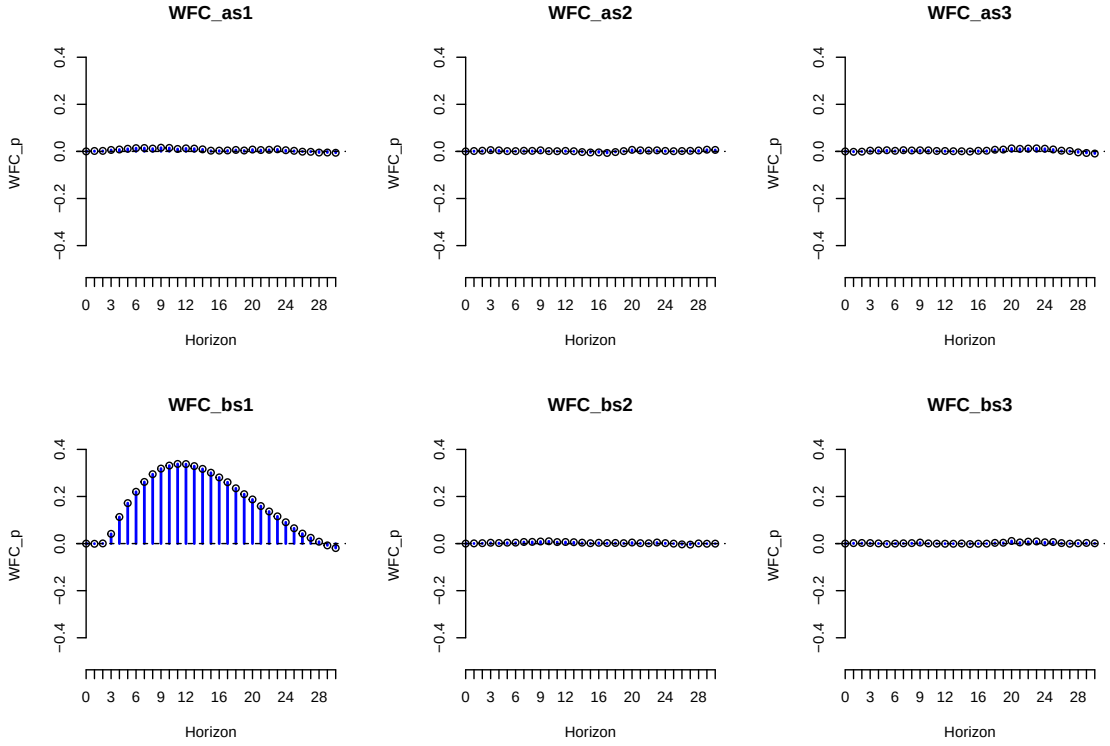

Figure 7: Own-price market impact of WFC (Wells Fargo) on 19th of July, 2016.

shifts back, this gives the HF investors enough time of reaction to arbitrage opportunities.

w[h!]

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$as1$		$\ominus\ominus\ominus\ominus\ominus$		$\oplus\oplus\oplus$		$\oplus\ominus$	$\ominus\ominus\ominus$	$\ominus\ominus\ominus$	$\ominus$
$as2$	$\ominus\ominus\ominus$	$\oplus$	$\oplus\oplus$		$\oplus$			$\ominus\ominus$	$\ominus\ominus\ominus\ominus\ominus\ominus$
$as3$	$\oplus\oplus$	$\oplus$	$\oplus$	$\oplus$	$\oplus\oplus\oplus$		$\ominus$	$\oplus\oplus$	$\ominus$
$bs1$	$\oplus\ominus\ominus$	$\oplus\ominus\ominus\ominus$		$\oplus$	$\oplus$	$\oplus\oplus$	$\oplus\oplus\oplus\ominus$	$\oplus\oplus$	$\oplus$
$bs2$	$\ominus$	$\ominus$			$\oplus\ominus$	$\oplus$	$\oplus$		$\oplus$
$bs3$		$\ominus$	$\ominus\ominus$	$\ominus\ominus$	$\ominus$	$\ominus\ominus$	$\ominus$		
r_{size}	44%	46%	0%	14%	25%	57%	80%	78%	100%

Table 13: The summary of own-price market impacts.

Figure 8 shows the market impacts of orders posted deeper in the book for Citigroup. This implies the positive pile-on effect where larger ask order may further perpetuating a price decrease, the orders may not necessarily set at the current market price of the stock (i.e. they are not market orders, they are limit orders). The estimated market impact lasts for almost 20 minutes, the price goes up after 10 minutes because the market investors may buy trades picking up the posted volume or by cancellations on the ask side.

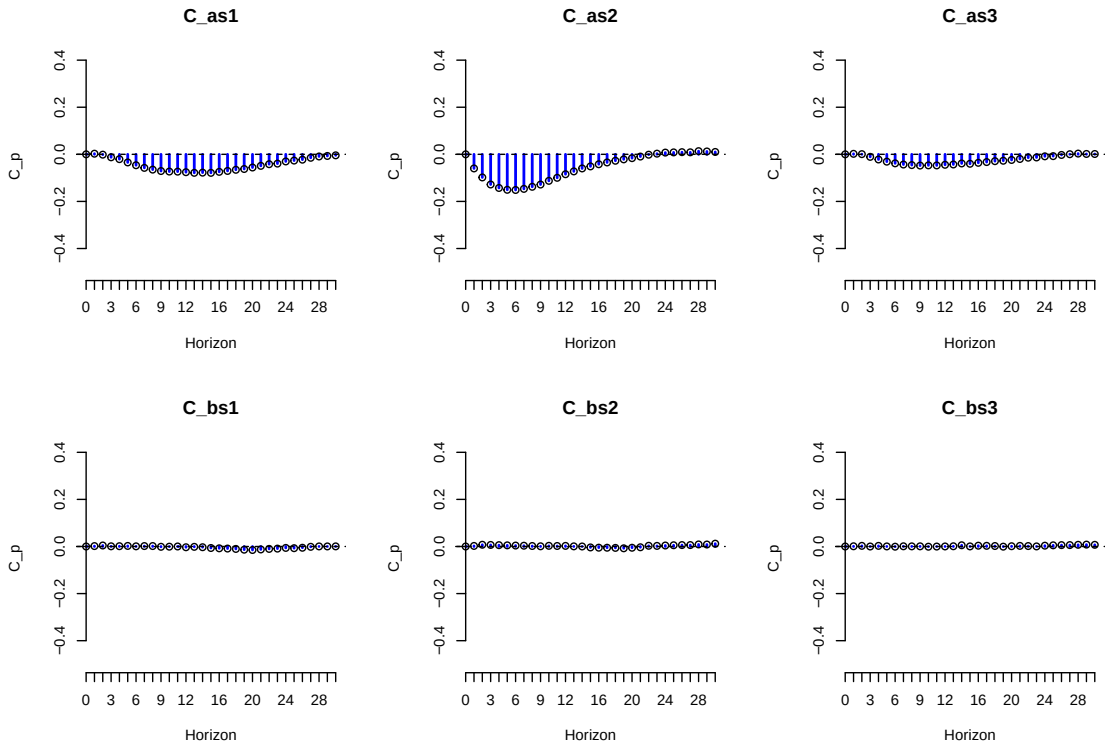

Figure 8: The bootstrapped market impact of Citigroup on 1st of June, 2016.

Table 13 reports the summary of significant market impacts identified by our model. For each trading day, we use $\ominus$ and $\oplus$ to represent the significant negative and positive response of price after the arrival of a market/limit order. Specifically, we define a ratio denoted as r_{size} to measure the price direction of market impacts,

$$\begin{aligned} r_{size} &= \frac{|\operatorname{sgn}(GI_t)|}{42} \\ \operatorname{sgn}(GI_t) &= \begin{cases} -1 & -GI_t(h) > Q_{0.05}(GI_t(h)) \\ 0 & |GI_t(h)| \leq Q_{0.05}(GI_t(h)) \\ 1 & GI_t(h) > Q_{0.05}(GI_t(h)) \end{cases} \\ t &= 1 \ldots T, \quad h = 1, \ldots, 30 \end{aligned} \tag{31}$$

The results suggest that the group of financial stocks is of higher r_{size} values, this may be explained by the fact that finance sector is leading the market, the history information indicates that their response of price to trading volumes is stable and thus robust for statistical arbitrage, see Hautsch and Huang (2012). The Citigroup performs well among them. Interestingly, the healthcare and technology stocks sometimes show opposite results, we notice that their prices are positively linked to ask order flows in some cases. This is because they are price-dominated stocks, i.e., they have multiple risk sources except for their own trading activity. This result is consistent with our main findings in subsection 4.1.2 and 4.2.1 where we conclude that financial stocks are size-dominated stocks and they are influencers in the system. Alternatively, the price of healthcare and technology stocks are risk receivers. Based on our methodology, it would be more profitable to invest in financial stocks for algorithm traders.

6 Conclusion

This paper build upon and extend current literature where the connectedness measures are often estimated by MA transformation of VAR systems and restricted to Gaussian innovations. We combine bootstrap-based generalized impulse response analysis with network construction. In this way, the network we construct relies neither on the ordering of the variables nor on the distribution of the innovations, the resulting connectedness measures is economic interpretable. Furthermore, given the HF LOB NASDAQ data, network analysis of LOB across stocks becomes interesting. Throughout the paper, we first show how network for LOB can be constructed in the presence of microstructure noise and non-synchronous trading, then we progress by focusing on the models that capture the dynamics of LOB and their influence over time. Our primary finding is that the network that involving the trading volumes is a better measure of the stock connectedness. With our methodology, we identify the significant market impact caused by the arrival of a large limit order, and order imbalance generally exists across stocks, bootstrapped market impacts can be quantified. The financial institutions are connected more closely compared with the firms come from other industry.

A Pre-averaging estimation

Suppose that we observe non-synchronous noisy data Y_t following,

$$Y_t = X_t + \varepsilon_t, \quad t \geq 0 \tag{32}$$

with efficient log price X_t is latent. The error term ε_t represents microstructure noise and is assumed to be independent and identically distributed with

$$\mathsf{E}(\varepsilon_t) = 0, \quad \mathsf{E}\left(\varepsilon_t^2\right) = \psi \tag{33}$$

The price process X_t follows a semi-martingale form, Delbaen and Schachermayer (1994),

$$X_t = X_0 + \int_0^t a_s ds + \int_0^t \sigma_s dW_s \tag{34}$$

where $(a_s)_{s\geq 0}$ is a càdlàg drift process, $(\sigma_s)_{s\geq 0}$ is an adapted càdlàg volatility process, $(W_s)_{s\geq 0}$ is a Brownian motion. In addition, we assume X_t and ε_t are independent, i.e.

$$\mathsf{E}(\varepsilon_t \mid X) = 0 \tag{35}$$

If one can only observe Y_i^n at discrete times t, i indexes the time points with interval length Δ_n, the returns $\Delta_i^n Y$ is thus defined as,

$$Y_i^n = Y_{i\Delta_n}, \quad \Delta_i^n Y = Y_i^n - Y_{i-1}^n, \quad i = 1, \ldots, n \tag{36}$$

A pre-averaging is conducted to alleviate microstructure noise and solve non-synchronicity, we follow the notations originally used by Jacod et al. (2009). The basic idea is to construct smoothing functions to diminish the impact of the noise induced by ε_t. Specifically, there is a sequence of integers denoted as k_n which satisfies,

$$\exists \theta > 0, \quad k_n \sqrt{\Delta_n} = \theta + \mathcal{o}\left(\Delta_n^{\frac{1}{4}}\right) \tag{37}$$

and a continuous weight function $g : [0,1] \mapsto \mathbb{R}$. g is piecewise C^1 with a piecewise derivative $g^{'}$, $g(0) = g(1) = 0$, and $\int_0^1 g^2(s)ds > 0$. Furthermore, the following real-valued numbers and functions are associated with function g on $\mathbb{R}_+$,

$$\begin{aligned}
\psi_1 &= \int_0^1 \{g^{'}(u)\}^2 du, \quad \psi_2 = \int_0^1 \{g(u)\}^2 du \\
\Phi_1(s) &= \int_s^1 g^{'}(u) g^{'}(u-s) du, \quad \Phi_2(s) = \int_s^1 g(u) g(u-s) du \\
\Phi_{ij} &= \int_0^1 \Phi_i(s) \Phi_j(s) du, \quad i,j = 1,2, \quad u \in [0,1]
\end{aligned} \tag{38}$$

Here we choose $g(x) = x \wedge (1-x)$, as in Podolskij et al. (2009), Christensen et al. (2010) and Hautsch and Podolskij (2013). Therefore we have

$$\begin{aligned}
\psi_1 &= 1, \quad \psi_2 = \frac{1}{12}, \quad \Phi_{11} = \frac{1}{6} \\
\Phi_{12} &= \frac{1}{96}, \quad \Phi_{22} = \frac{151}{80640}
\end{aligned} \tag{39}$$

The pre-averaged returns $\overline{Y}_i^n$ associated with the weight function g are given as,

$$\begin{aligned}
\overline{Y}_i^n &= \sum_{j=1}^{k_n - 1} g\left(\frac{j}{k_n}\right) \Delta_{i+j}^n Y \\
&= -\sum_{j=0}^{k_n - 1} \left\{ g\left(\frac{j+1}{k_n}\right) - g\left(\frac{j}{k_n}\right) \right\} Y_{i+j}^n, \quad i = 0, \ldots, n - k_n + 1
\end{aligned} \tag{40}$$

The window size k_n defined in equation (37) is chosen of $\mathcal{O}\left(\sqrt{\frac{1}{\Delta_n}}\right)$, balance the noise $\overline{\varepsilon}_i^n = \mathcal{O}_p\left(\sqrt{\frac{1}{k_n}}\right)$ and the efficient price $\overline{X}_i^n = \mathcal{O}_p\left(\sqrt{k_n \Delta_n}\right)$.

B Bootstrap-based multistep forecast methods

Here we describe the computational steps to obtain the $\mathsf{E}(y_{t+1}|u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$, GI, GFEVD via Bootstrap method, more details can be found in Koop et al. (1996), Lanne and Nyberg (2016), Teräsvirta et al. (2010).

1. Denote $\mathcal{F}_{t-1}$ as all the information prior to Y_t, and select a forecast horizon h.

2. Randomly sample N_B vectors of shocks $(\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^\top$ from the residuals of estimated model,

$$\delta_{jt} : (\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^\top \sim \hat{u}^\star_{jt} e_j \tag{41}$$

$$\hat{u}^\star_{jt} = Y_t - \left(\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_p\right) \left(Y^\top_{t-1}, Y^\top_{t-2}, \ldots, Y^\top_{t-p}\right)^\top = Y_t - g(Y_{t-1}) \tag{42}$$

3. Compute conditional multistep forecast $\mathsf{E}(y_{t+l}|\mathcal{F}_{t-1})$,

$$\begin{aligned} f_{t,0} &= g(Y_{t-1}) \\ f_{t,1} &= \mathsf{E}[Y_{t+1} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,0} + \hat{u}^\star_t) \mid \mathcal{F}_{t-1}] \\ f_{t,2} &= \mathsf{E}[Y_{t+2} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,1} + \hat{u}^\star_{t+1}) \mid \mathcal{F}_{t-1}] \\ &\ldots \end{aligned} \tag{43}$$

   with $\hat{u}^\star_{t+l}, l = 1, \ldots, h$ are independent draws with replacement from the set of residuals $\{\hat{u}_{t+l}\}_{t=1}^T$ over the sample period.

4. Repeat steps 3 for all N_B vectors of estimated innovations with bootstrap methods, iterating on the estimated model,

$$\begin{aligned} fb_{t,1} &= \frac{1}{N_B} \sum_{i=1}^{N_B} g(f_{t,0} + \hat{u}^{\star(i)}_t) \\ fb_{t,2} &= \frac{1}{N_B} \sum_{i=1}^{N_B} g(g(f_{t,0} + \hat{u}^{\star(i)}_t) + \hat{u}^{\star(i)}_{t+1}) \\ &\ldots \end{aligned} \tag{44}$$

5. By the same logic, we compute $\mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$ when the shock is given as

$\delta_{jt} = \hat{u}^{\star}_{jt} e_j$,

$$\begin{aligned} f_{t,0} &= g(Y_{t-1}) \\ f_{t,1} &= \mathsf{E}[Y_{t+1} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,0} + \hat{u}^{\star}_{jt} e_j) \mid \mathcal{F}_{t-1}] \\ f_{t,2} &= \mathsf{E}[Y_{t+2} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,1} + \hat{u}^{\star}_{j,t+1} e_j) \mid \mathcal{F}_{t-1}] \\ &\dots \end{aligned} \tag{45}$$

with $\hat{u}^{\star}_{j,t+l}, l = 1, \dots, h$ are independent draws with replacement from the set of residuals $\{\hat{u}_{j,t+l}\}_{t=1}^{T}$ over the sample period.

6. Repeat steps 5 for all N_B vectors of estimated innovations with bootstrap methods, iterating on the estimated model,

$$\begin{aligned} fb_{t,1} &= \frac{1}{N_B} \sum_{i=1}^{N_B} g(f_{t,0} + \hat{u}^{\star(i)}_{jt} e_j) \\ fb_{t,2} &= \frac{1}{N_B} \sum_{i=1}^{N_B} g(g(f_{t,0} + \hat{u}^{\star(i)}_{jt} e_j) + \hat{u}^{\star(i)}_{j,t+1} e_j) \\ &\dots \end{aligned} \tag{46}$$

7. Plug in the GI function

$$GI(l, \delta_{jt}, \mathcal{F}_{t-1}) = \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \tag{47}$$

to obtain the relative contribution of a shock δ_{jt} to the i-th variable with horitzon h at time t,

$$\lambda_{ij,\mathcal{F}_{t-1}}(h) = \frac{\sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}{\sum_{j=1}^{K} \sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}, \quad i, j = 1, \dots, K \tag{48}$$

8. Repeat steps 2-6 for all histories.

9. Construct table 3 using averaged $\lambda_{ij,\mathcal{F}_{t-1}}(h)$ generated from step 7.

If there is a unit shock,

$$\delta_{jt} \quad : \quad (\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^{\top} \sim e_j \tag{49}$$

then we can simiply replace $\hat{u}^{\star}_{jt} e_j$ of (41) with e_j of (49), and repeat the steps from 1 to 6 stated above, the generalized impulse response can be calculated based on (47), i.e.

$$GI(l, \delta_{jt}, \mathcal{F}_{t-1}) = \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \tag{50}$$

We should note that if K is extremely large in empirical study, the denominator of equation (48) might be unnecessarily large due to accumulated noise caused by the large amount of irrelevant variables. Therefore one more step of prescreening is preferred to filter out less relevant variables.

References

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies*, 18(2):351–416.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, pages 885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2000). Exchange rate returns standardized by realized volatility are (nearly) gaussian. Technical report, National Bureau of Economic Research.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association*, 96(453):42–55.

Bandi, F. M. and Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, 79(3):655–692.

Bandi, F. M. and Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies*, 75(2):339–369.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241.

Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Bloomfield, R., OÕhara, M., and Saar, G. (2005). The Òmake or takeÓ decision in an electronic market: Evidence on the evolution of liquidity. *Journal of Financial Economics*, 75(1):165–199.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Christensen, K., Kinnebrock, S., and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116–133.

Delbaen, F. and Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische annalen*, 300(1):463–520.

Demirer, M., Diebold, F. X., Liu, L., and Yılmaz, K. (2017). Estimating global bank network connectedness. Technical report, National Bureau of Economic Research.

Diebold, F. X. and Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Eisler, Z., Bouchaud, J.-P., and Kockelkoren, J. (2012). The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance*, 12(9):1395–1419.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *The review of financial studies*, 18(4):1171–1217.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Handa, P., Schwartz, R., and Tiwari, A. (2003). Quote setting and price formation in an order driven market. *Journal of financial markets*, 6(4):461–489.

Hautsch, N. and Huang, R. (2012). The market impact of a limit order. *Journal of Economic Dynamics and Control*, 36(4):501–522.

Hautsch, N. and Podolskij, M. (2013). Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence. *Journal of Business & Economic Statistics*, 31(2):165–183.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications*, 119(7):2249–2276.

Kavajecz, K. A. and Odders-White, E. R. (2004). Technical analysis and liquidity provision. *Review of Financial Studies*, 17(4):1043–1071.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, 74(1):119–147.

Lanne, M. and Nyberg, H. (2016). Generalized forecast error variance decomposition for linear and nonlinear multivariate models. *Oxford Bulletin of Economics and Statistics*, 78(4):595–603.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097.

Parlour, C. A. and Seppi, D. J. (2003). Liquidity-based competition for order flow. *The Review of Financial Studies*, 16(2):301–343.

Pesaran, H. H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29.

Podolskij, M., Vetter, M., et al. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli*, 15(3):634–658.

Roşu, I. (2009). A dynamic model of the limit order book. *The Review of Financial Studies*, 22(11):4601–4641.

Teräsvirta, T., Tjostheim, D., Granger, C. W., et al. (2010). Modelling nonlinear economic time series. *OUP Catalogue*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wu, W. B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

Zhang, L. et al. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.