Trade Co-occurrence, Trade Flow Decomposition, and Conditional Order Imbalance in Equity Markets

Yutong Lu[1], Gesine Reinert[1,3], and Mihai Cucuringu[1,2,3]

[1]Department of Statistics, University of Oxford, Oxford, UK

[2]Mathematical Institute, University of Oxford, Oxford, UK

[3]The Alan Turing Institute, London, UK

December 27, 2022

Abstract

The time proximity of high-frequency trades can contain a salient signal. In this paper, we propose a method to classify every trade, based on its proximity with other trades in the market within a short period of time, into five types. By means of a suitably defined normalized order imbalance associated to each type of trade, which we denote as *conditional order imbalance* (COI), we investigate the price impact of the decomposed trade flows. Our empirical findings indicate strong positive correlations between contemporaneous returns and COIs. In terms of predictability, we document that associations with future returns are positive for COIs of trades which are isolated from trades of stocks other than themselves, and negative otherwise. Furthermore, trading strategies which we develop using COIs achieve conspicuous returns and Sharpe ratios, in an extensive experimental setup on a universe of 457 stocks using daily data for a period of three years.

Keywords: Market microstructure; Co-occurrence analysis; Order imbalances; Trading strategies; Quantitative finance

1 Introduction

The transformation of major equity exchanges to electronic trading significantly reshapes the market microstructure landscape, by reducing latency up to nanoseconds (O'Hara (2015), Hirschey (2021)), and thus leading to market participants achieving unprecedented levels of profitability in their trading strategies. Every agent in the market can directly submit and cancel limit orders. Trades are settled when existing limit orders are executed by market orders/marketable limit orders. Trades, carrying distinct information and having their own impact on the price changes of the underlying stocks, have been classified into different types and studied separately by academics and practitioners. For example, grouping by directions of trading, Chordia, Goyal, and Jegadeesh (2016) study flows of buyer- and seller-initiated trades, thus decomposing into aggressive buys and aggressive sells. Kraus and Stoll (1972) and Lee et al. (2004) separate institutional trades from trades placed by individual investors. Different from these classifications, which are exclusively based on the characteristics of the individual trades, we are interested in classifying trades according to their time of placement relative to the arrival time of other trades across the market, both within the same asset and also cross-sectionally across the available universe.

Our motivation arises from the fact that market participants can make trading decisions by observing the trade flows in the market. Previous works (Kyle (1985), Kyle, Ou-Yang, and Wei (2011)) model the price formation at high frequency, and suggests that informed traders split large orders into many smaller orders in order to conceal their true purpose, while other market participants monitor order flows in the market in order to reach trading decisions. The development of high-performance trading systems has led to an astounding growth of high-frequency trading (HFT) and diversity of strategies (Hagströmer and Nordén (2013)). In this world, the reaction time plays an important role because opportunities can be transient if not acted upon within microseconds, and even nano-seconds. High-frequency trading strategies include anticipating trade flow (Hirschey (2021)) and preying on other market participants (Van Kervel and Menkveld (2019)). The questions we are interested in exploring concern whether certain trades, interacting with other trades in various different ways, contain useful information, and how they contribute to stock price movements, helping us shed light on the price formation mechanism at both short-term and long-term horizons.

We start with proposing the concept of *co-occurrence of trades*, defined in Section 3.1, which offers a tool to identify and group trades based on their interactions with other trades. For each given trade, we consider it to co-occur and interact with another trade if both trades are taking place close in time to each other. To define and quantify "closeness", we pre-define a neighbourhood size δ. If the time difference between two trades is lower than δ, they are close to each other and they co-occur. Notice that the threshold δ is an important parameter, determining the set of trades that co-occur. However, there is no strict rule to set its value. Intuitively, considering a scenario where an HFT preys on an institutional trader and trades in response to institutional marketable orders, we aim to capture these interactions and classify such trades into a category of, for example, actively interactive trades. With this in mind, an appropriate choice should be greater than the round-trip latency plus the time for the HFT to detect and make trading decisions, which is usually undisclosed. Therefore, we experiment with multiple values of δ, and compare and contrast the corresponding results. Note that δ should not be too large either, since a large neighbourhood is likely to incorporate irrelevant trades from the market. In this paper, we treat δ as a hyper-parameter; for simplicity, we only report the empirical results for $\delta = 500\mu s$ and make a comparison across different choices of δ values in Appendix A.

Using trade co-occurrence, we decompose daily trade flows by classifying all the trades of all stocks into subgroups. Given a trade, we determine to which group it belongs by asking the following two questions: Does it interact with other trades? If yes, does it interact with only trades of the same stock as itself, only with stocks different from itself, or with both kinds? Depending on the answer, a trade will be placed into one or two classes, for which detailed rules are explained in Section 3.2. After labeling all trades, we study the relations between returns and subgroups of trades.

We use order imbalance as a bridge connecting trade flows and stock returns, which has been thoroughly studied in the finance literature. An inventory paradigm (Stoll (1978), Spiegel and Subrahmanyam (1995), Chordia, Roll, and Subrahmanyam (2002)) suggests that, in intermediated markets, a difference, or so-called *imbalance*, between buyer-initiated and seller-initiated trades puts pressure on a market maker's inventory. In response, the market makers adjust inventories to maintain their market exposures, which drives the price to one direction.

Next, at a daily level, we investigate the properties of aggregated order imbalance of each category of trades and their relation with individual stock returns during normal trading hours. Data exploration indicates that all categories of conditional, as well as the unconditional, order imbalance are positively auto-correlated. The conditional order imbalances (COIs) all have strong positive correlations with the original order imbalance. However, they are not necessarily highly correlated with each other.

Our empirical results concentrate on the imbalance-return relations. By means of regression analysis, we discover positive and significant correlations between order imbalances and price changes within the same day. Furthermore, in comparison to a standard regression analysis, decomposing order flows leads to significantly higher adjusted R^2 in our multiple regression settings, which can be interpreted as better explanatory power in contemporaneous intraday open-to-close stock returns. To exploit predictability, we use the same regression analysis to fit order imbalances against future one-day ahead returns. In contrast to contemporaneous results, statistically significant relations only appear in a small proportion of stocks. Despite the low percentage of significant regression coefficients, we observe that order imbalances, for those trades that have a higher interaction with the rest of the market, appear to have negative relations with future returns. On the contrary, imbalances of trades isolated from other stocks in the market show weakly positive correlations.

These associations are amplified in our subsequent portfolio analysis, as follows. We leverage these imbalances to build trading strategies. In order to assess the economic value of the trade flow decomposition method, we construct signal-sorted portfolios using COIs as signals. In particular, if we make long/short decisions in alignment with the observed patterns in the predictive regressions, we attain profits in all of our portfolios, with the highest annualized Sharpe ratio reaching 2.38. As a benchmark, we build portfolio investing in order imbalances without decomposition, for which the Sharpe ratio is negative.

The remainder of this paper is organized as follows. Section 2 outlines our contributions to the finance literature. In Section 3, we introduce the definitions of trade co-occurrence, trade flow decomposition and COIs. We start our empirical studies with describing data sources and conducting exploratory analysis in Section 4. Subsequently, we uncover the relations between COIs and contemporaneous returns in Section 5 and

investigate the predictive power of COIs in Section 6, and economic value of COIs in Section 7. Section 8 provides robustness analysis and additional empirical findings. Finally, in Section 9, we summarizes the results and discuss our limitations and future research directions.

2 Related Literature

This paper contributes to three strands of literature. First, our study exploits a new financial application of co-occurrence analysis, which is a statistical method proven to be powerful in spatial pattern analysis and widely used in the fields of biology (Gotelli (2000), MacKenzie, Bailey, and Nichols (2004), Araújo et al. (2011)), natural language processing (NLP) (Dagan, Lee, and Pereira (1999), Kolesnikova (2016)), computer vision (Galleguillos, Rabinovich, and Belongie (2008), Aaron, Taylor, and Chew (2018)), and others (Appel and Holden (1998), Ye et al. (2017)). So far, the applications of co-occurrence analysis in finance literature concentrate on studying stocks co-occurring in news articles. Ma, Pant, and Sheng (2011) construct networks from company co-occurrence in online news and use machine learning models to identify competitor relationships between companies. Recent studies, including Guo et al. (2018); Tang, Zhou, and Hong (2019); Wu et al. (2019), build networks using stocks co-occurrence in news and employ them for tasks such as return predictions and portfolio allocation. We contribute by originating the idea of trade co-occurrence. By directly applying the co-occurrence of stock trades, we establish that this technique is beneficial for exploring and gaining insights from the financial market microstructure.

Second, our research adds to the studies of interactions among trading activities in the market. In Kyle (1985)'s model, market makers observe the aggregated order flows of informed and liquidity traders in the market to adjust their trading strategies. More aggressively, HFT traders can detect informed traders, such as institutions (Van Kervel and Menkveld (2019)) and predict trade flows of others (Hirschey (2021)). Various theoretical models (Grossman and Miller (1988), Brunnermeier and Pedersen (2005), Yang and Zhu (2020)) are proposed for the interplay between high-frequency and institutional traders. Van Kervel and Menkveld (2019) conduct an empirical study on the Swedish stock market and discover that HFT participants intend to trade against wind when

5

the institutional traders begin splitting large orders, and eventually trading in the same direction as the institutions.

We contribute to this topic by proposing the idea of trade co-occurrence and provide empirical evidence that the co-occurrence of stock trades is not coincident. Rather than studying interaction among traders, we innovate trade co-occurrence as a tool to analyze interactivity at the individual trade level. Our study of COIs conditional on co-occurrence shows that the interactions of trades at a granular level convey useful information on price formation.

Finally, this paper contribute to the literature of order imbalance and price formation. According to pioneering researches, persistence in order imbalance can arise in two ways. Firstly, as the model by Kyle (1985) states, traders intend to split large orders over time to minimize their market impacts, which leads to auto-correlated imbalances. Another source for order imbalance, as Scharfstein and Stein (1990) state, is the herd effect. To explore how order imbalance affects price changes, Chordia and Subrahmanyam (2004) propose a theoretical model to explain the positive relation between order imbalance and contemporaneous stock returns, arising from the market makers dynamically accommodating order imbalance. In addition, discretionary traders optimally splitting orders across days enables order imbalance to have strong positive auto-correlation and predictive power on future returns. Their empirical study, using daily data of stocks listed on New York Stock Exchange (NYSE) for a 10-year period from 1988 to 1998, confirms their theoretical results and shows profitability of using order imbalance as trading signals. However, there is controversy on the predictability. For example, Shenoy and Zhang (2007) and Lee et al. (2004) find order imbalances having no significant predictive power on future returns.

Although Chordia and Subrahmanyam (2004) do not differentiate trade flows, subsequent studies have shown that marketable orders, placed at different time, by different agents, with distinct properties can have different impacts on price changes. Most evidence stems from the Chinese market (Lee et al. (2004), Bailey et al. (2009), Zhang, Gu, and Zhou (2019)), where private data of identification of trader types are available, and they find indications that order imbalances of institutional trade flows have higher pressure on prices than imbalances of individual traders. Same results are found in the US market by Cox (2021)'s recent study of S&P 500 stocks during 2015 to 2016, which split

6

trades into binary classes depending on whether or not they are inter-market sweeping orders, which are mainly adopted by institutions (Chakravarty et al. (2012)).

Our research complements these works by supplementing the study of order imbalances in the US market using data of the most recent period before Covid-19 and proposing a novel method to decompose the unconditional trade flows without requiring an additional private data set. We show that order imbalances, without differentiating trades, no longer have forecasting power on future returns, which is evidence for an evolution of the market microstructure over the past decades (Chordia, Roll, and Subrahmanyam (2002); Chordia and Subrahmanyam (2004)). However, trade flows decomposed with our proposed method carry different information content, and their COIs do possess forecasting power.

3 Co-occurrence of Trade and Flows Decomposition

3.1 Trade Co-occurrence

We first introduce the definition of trade co-occurrence. For each trade x_i occurring at time t_i, with a pre-specified δ, every trade, other than x_i itself, that arrives within time period $(t_i - \delta, t_i + \delta)$ is defined as having co-occurred with trade x_i. We denote the threshold δ as the *neighbourhood size*, and the set of all trades co-occurred with x_i as δ-neighbourhood of trade x_i. Figure 1 sketches an example, where trade x_i co-occurs with trades x_j and x_l, while it does not co-occur with trade x_k. We note that co-occurrence is not an equivalence relation. It is perfectly possible for x_i and x_j to co-occur, and for x_i and x_l to co-occur, without x_j and x_l co-occurring.
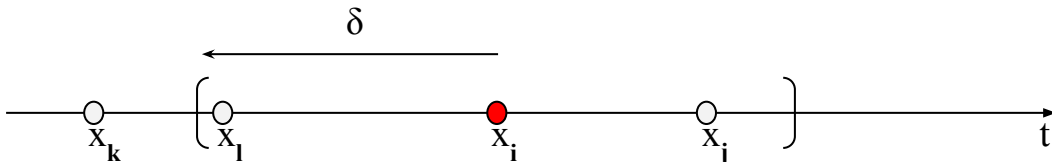
Figure 1: Illustration of trade co-occurrence. This figure visualizes the idea of trade co-occurrence; given a user-defined neighbourhood size δ, trade x_j arrives within the δ-neighbourhood of trade x_i, and thus they co-occur. In contrast, trade x_k locates outside x_i's neighbourhood, and thus the two trades do not co-occur. Both trades x_j and x_l co-occur with trade x_i, but they do not co-occur with each other

3.2 Trade Flows Decomposition

Based on co-occurrence, we next split trades into different classes characterized by their δ-neighbourhood, with the protocol illustrated in Figure 2. Initially, we partition all trades into two groups, isolated (iso) and non-isolated (nis) trades. Trades are labelled as *isolated* if they do not co-occur with any other trades. Otherwise, trades are labelled as *non-isolated*.

We further decompose the non-isolated trades according to properties of the trades within their δ-neighbourhood.

Each non-isolated trade x_i can be classified into one of the following three categories

1. *non-self-isolated* (nis-s): the δ-neighbourhood of trade x_i contains **only** trades (at least one) of the <u>same</u> stock as the one from trade x_i;

2. *non-cross-isolated* (nis-c): the δ-neighbourhood of trade x_i contains **only** trades of stocks which are <u>different</u> than the stock corresponding to trade x_i;

3. *non-both-isolated* (nis-b): the δ-neighbourhood of trade x_i contains **both** at least one trade of the <u>same</u> stock, and at least one other trade of a <u>different</u> stock.

Theses three classes form a partition of the set of non-isolated trades, as illustrated in the last line in of Figure 2. We refer to this process of separating trades into categories as **trade flow decomposition**.

3.3 Conditional Order Imbalance

With the decomposition of trade flows, we proceed to study the price impact of trades with different characteristics. A bridge connecting trading activities and price changes is given by the order imbalance quantity, defined as the normalized difference between the volume of buyer- and seller-initiated trades (Chordia and Subrahmanyam (2004)). For a given stock i, we derive conditional daily order imbalances, as follows

$$COI_{i,t}^{type} = \frac{N_{i,t}^{type,buy} - N_{i,t}^{type,sell}}{N_{i,t}^{type,buy} + N_{i,t}^{type,sell}}, \tag{1}$$

where $N_i^{buy,type}$ and $N_i^{sell,type}$ denote the total number of market buy orders and market sell orders of stock i in day t respectively. If the denominator is 0, which happens when
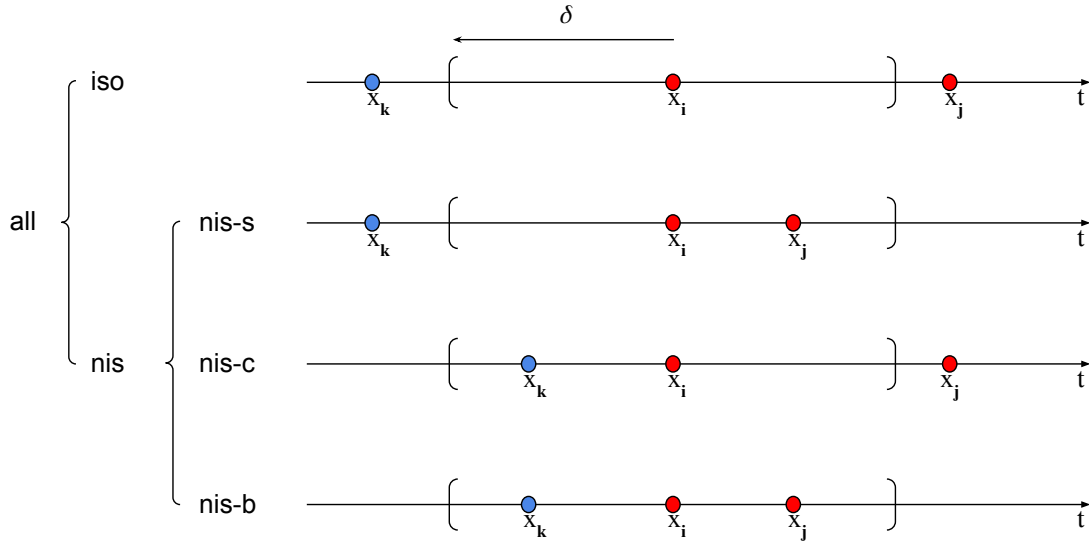
Figure 2: Illustration of trade types, conditioning on co-occurrence. We showcase the distinct categorical labels of trade x_i. Color indicates the stock corresponding to a trade. Thus, x_j is for the same stock as x_i, while x_k is for a different stock. First line: x_i is an isolated (iso) trade with empty δ-neighbourhood; second line: x_i is a non-isolated (nis) trade with nonempty δ-neighbourhood; third line: x_i is a non-self-isolated ('nis-s') trade with only other trades for the same stock in its δ-neighbourhood; x_i is an non-cross-isolated ('nis-c') trade with only other trades for the different stocks in its δ-neighbourhood; last line: x_i is a non-both-isolated ('nis-b') trade with both other trades for the same and different stocks in its δ-neighbourhood.

there are no trades of a certain type, we define the COI in this case to be 0. We consider six types of COIs and the superscript *type*, which takes a value in {all, iso, nis, nis-s, nis-c, nis-b}, indicates the group of trades used to calculate the imbalance. Note that the 'all' label corresponds to using the entire universe of trades without decomposing based on trade co-occurrence. Thus, the 'all' COI is the same as order imbalance in the number of transactions, scaled by total transactions, studied by Chordia and Subrahmanyam (2004).

4   Exploratory Data Analysis

In this section, we provide a brief description of the data employed in our empirical study. Through exploratory analysis, we uncover salient patterns of trade co-occurrence. Furthermore, we show that the order imbalances of decomposed trade flows are weakly correlated with each other, which indicates the trade decomposition we propose is meaningful.

4.1 Data Source and Preprocessing

Our study is based on 457 US stocks during the period from 2017-01-03 to 2019-12-10. The selected stocks are those companies included in Standard & Poor's (S&P) 500 index for which both order book data and price data is available over the entire sample period.

4.1.1 Limit Order Books

We obtain limit order book data from the LOBSTER database (Huang and Polak (2011)), which provides detailed records of limit orders for all stocks traded in the NASDAQ exchange. The records include limit order submissions, cancellations and executed trades, indexed by time with precision up to nanoseconds. For each stock on each trading day, a record contains the time stamp, event type (submissions/cancellations/executions), direction (buy/sell), size and price for a limit order event. By filtering for limit order executions and reversing their directions, we infer the buyer- and seller-initiated trades, e.g. execution of a limit buy order implies placement of a market sell order/marketable limit sell order. Noticing that a large market order simultaneously consumes multiple existing limit orders, we merge inferred trades with identical timestamps. Given LOBSTER's high time resolution, we assume different trades cannot have exactly the same timestamps.

4.1.2 Prices and Returns

We acquire daily price data for our stock universe under consideration, from the Center for Research in Security Prices (CRSP) database, and calculate daily open-to-close logarithmic returns as

$$R_{i,t} = \log \frac{P_{i,t}^{Close}}{P_{i,t}^{Open}}, \tag{2}$$

where $P_{i,t}^{Open}$ and $P_{i,t}^{Close}$ are daily open and close prices of stock i on day t. To alleviate the effect of the market component, we also consider *market excess returns* in this study, denoted as $r_{i,t}$, calculated as follows

$$r_{i,t} = R_{i,t} - R_{SPY,t}, \tag{3}$$

where $R_{SPY,t}$ is the daily return of SPY ETF, which tracks the S&P 500 index. For simplicity, here we assume all stocks have the same market *beta equal to 1.*

4.2 Summary Statistics of Trades

After building our data set of trades, we label every trade with its corresponding type. A summary of the data for the different types of trades is presented in Table 1; the chosen neighbourhood size for co-occurrence is 500 microseconds ($\delta = 500\mu s$). Panel A shows descriptive statistics of the raw data, where each number is calculated by averaging daily time series and then considering the cross-sectional mean, median or standard deviation over all stocks. On average, isolated trades account for 38.29% of the total number of transactions, while the majority of trades are non-isolated in one of the three defined types (nis-s, nis-c, or nis-b). Approximately half of the non-isolated trades, 29.83% of all trades, are non-self-isolated. The proportions of non-cross-isolated and non-both-isolated trades are similar, with the mean of 15.58% and 16.30%, respectively. The large standard deviation for the number of trades could be seen as an indication that the population is heterogeneous. Panel B shows the percentages of different groups of trades in terms of volumes, which are very similar to those reported in Panel A. With this in mind, it is reasonable to concentrate on the count of trades as a liquidity measure.

Highlighting the empirical fact that the trading activity is higher at the start and end of a trading day, Figure 3 plots the intraday distributions of trades, revealing slightly different temporal behaviours of different trade types. The plot exhibits the number of each type of trades over every half hour, with the y-axis indicating percentages of the total number of trades. We observe that all types of trades increase drastically in the last half an hour. It is noteworthy that, after the decomposition, the flow of isolated trades is smoother than the flow of non-isolated-trades, with a lower slope for the last-half-hour climb. By further separating the sub-types of non-isolated trades, we find that non-self-isolated trades contribute more at the start of a day, while the line of other two types are flat except at the end of days.

4.3 Co-occurrence Probabilities: Null v.s. Empirical

With order book data, we first answer the following fundamental questions. Do trades really co-occur or are their arrivals simply random and independent of each other? Does

Table 1: Summary statistics for all groups of trades.
This table documents the time-series average of the daily cross-sectional statistics of each type of trades. Our data include records of trades within normal trading hours of 457 stocks from 2017-01-03 to 2019-12-10.

Panel A: Average daily statistics of trade counts

	Mean	Median	Std. dev.
Number of trades	3378.06	3022.07	1651.50
Percentage of iso trades	38.29	38.14	7.27
Percentage of nis trades	61.71	61.86	7.27
Percentage of nis-s trades	29.83	29.49	7.02
Percentage of nis-c trades	15.58	14.70	5.66
Percentage of nis-b trades	16.30	15.85	5.38

Panel B: Average daily statistics of trade volumes

	Mean	Median	Std. dev.
Volume of trades	522301.02	458738.54	275668.69
Percentage of iso volumes	38.81	38.52	7.29
Percentage of nis volumes	61.19	61.48	7.29
Percentage of nis-s volumes	29.75	29.50	6.81
Percentage of nis-c volumes	16.00	15.01	5.87
Percentage of nis-b volumes	15.44	15.09	5.11

our trade flows decomposition capture a signal? In this part, we develop a null model under the assumption of completely random order arrival. By comparing the theoretical co-occurrence probabilities (Donges et al. (2016)) under the null model and the empirical values derived from data, we confirm the existence of co-occurrence among stock trades at the level of 0.5 milliseconds, supporting the idea that the overall trading volume has a strong cross-asset interaction component. From an economic perspective, this is perhaps to be expected, given the large presence in nowadays markets of index-arbitrage traders who **simultaneously** trade an index ETF against a basket of constituents.

We develop the null model under the assumption that, for stock i, the arrivals of trades within a time interval of length T, follow independent Poisson processes with the same intensity λ_T. Let N_i denote the number of trades of stock i in $[0, T]$. Conditional on $N_i = n_i$, the arrival time of the n_i trades are independent and follow a uniform distribution on $[0, T]$. Hence, for each trade, the probability that another trade falls in its δ-neighbourhood during the time period T is
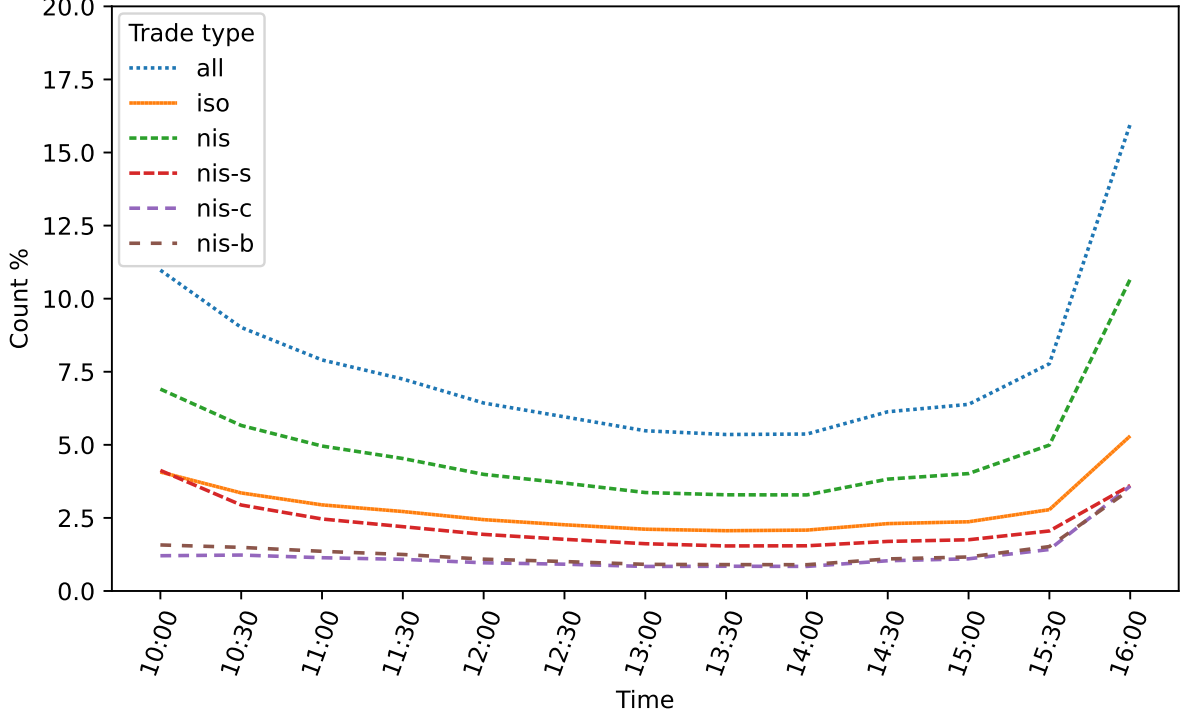
$$p = \frac{2\delta}{T}. \tag{4}$$

Figure 3: Intraday distributions of the number of each type of trades.
We calculate COIs for non-overlapping 30 minute intervals during normal trading hours from 9:30 to 16:00 for all stock over the period from 2017-01-03 to 2019-12-10. This figure plots intraday 30-minute COIs, averaged over both time series and cross-section.

Next, we derive the probabilities of different types of trade flows, as follows

$$
\begin{aligned}
\mathbb{P}_i(iso) &= (1-p)^{(N_i+N_{-i}-1)}, \\
\mathbb{P}_i(nis) &= 1 - (1-p)^{(N_i+N_{-i}-1)}, \\
\mathbb{P}_i(nis\text{-}s) &= [1-(1-p)^{N_i-1}](1-p)^{N_{-i}}, \\
\mathbb{P}_i(nis\text{-}c) &= (1-p)^{N_i-1}[1-(1-p)^{N_{-i}}], \\
\mathbb{P}_i(nis\text{-}b) &= [1-(1-p)^{N_i-1}][1-(1-p)^{N_{-i}}],
\end{aligned}
\tag{5}
$$

where N_{-i} denotes the number of trades for all stocks in the market other than stock i. The next step is to derive the probabilities for each stock. As illustrated in Figure 3, the intraday intensities are not constant, and hence we calculate the probabilities for every half-hour ($T = 30$ min) interval, and consider their averages (weighted by the intensities), as the final daily probabilities.

Table 2 shows the null and empirical daily probabilities averaged over time and stocks. Given the tiny neighbourhood size ($\delta = 500\mu s$), 92.67% of trades should be isolated if there is no co-occurrence. However, there are only 38.29% isolated trades in the market.

13

Table 2: Null and empirical probability of each type of trade flows.
This table shows the percentage, averaged over both days and stocks, of each type of trades under the null model and from the real data respectively.

	Null	Empirical
iso	92.67	38.29
nis	7.33	61.71
nis-s	0.02	29.83
nis-c	7.31	15.58
nis-b	0.00	16.30

In conclusion, there is empirical evidence that the notion of trade co-occurrence captures a signal. This serves as motivation to decompose trade flows and study them separately.

4.4 Descriptive Statistics of Order Imbalances

With trades labeled according to their co-occurrence types, we compute daily order imbalances and report descriptive statistics in Table 3. Panel A documents summary statistics of each category of order imbalance, averaged over time and stocks. Overall, the average unconditional order imbalances are positive. After the decomposition, the isolated and non-self-isolated order imbalances tend to be positive, with both higher means and variances compared to their unconditional counterparts. In contrast, the means of non-cross-isolated and non-both-isolated imbalances are negative, but with even higher variance. Hence, our study essentially constructs features with different behaviours by conditioning on the co-occurrence of trades. However, the standard deviations are much larger than the means, so statistically, the means are not significantly different from zero. Hence the means can only be taken as a very weak indication of a potential signal.

Panel B presents average partial autocorrelations of each type of order imbalance. It can be seen that all the order imbalances are positively auto-correlated. The lag 1 auto-correlations for COIs are substantial. Among the conditional imbalances, the non-cross-isolated order imbalance, corresponding to trades that closely co-occur with trades of other stocks in the market, has relatively higher auto-correlation. In contrast, the auto-correlation for the order imbalance from non-self-isolated trades is comparatively lower. These partial auto-correlations decay drastically with increasing lags.

Figure 4 shows the Pearson correlations, averaged over all stocks, of COIs, with $\delta = 500$ microseconds. All types of order imbalances are positively correlated with each other

Table 3: Summary statistics for all groups of trades and order imbalances.
This table shows the summary statistics of COIs from 2017-01-03 to 2019-12-10 for the selected 457 stocks. Panel A documents the mean, median and standard deviations of COIs. Panel B presents the partial autocorrelations of COIs, averaged over all stocks.

Panel A: Statistics of daily order imbalances			
	Mean	Median	Std. dev.
all	0.0024	0.0026	0.1163
iso	0.0108	0.0105	0.1493
nis	-0.0037	-0.0028	0.1256
nis-s	0.0090	0.0087	0.1471
nis-c	-0.0265	-0.0225	0.1847
nis-b	-0.0163	0.0128	0.1985

Panel B: Average partial autocorrelation of order imbalances			
lag	1	2	3
all	0.270	0.089	0.040
iso	0.259	0.094	0.045
nis	0.261	0.092	0.044
nis-s	0.213	0.084	0.045
nis-c	0.316	0.124	0.068
nis-b	0.243	0.096	0.059

while the strengths are different and can be fairly low. An exception is the unconditional order imbalance, which is strongly associated with every other type. The correlations between isolated imbalance and non-isolated imbalance, as well as its sub-types, are low.

As expected, conditioning on isolation and non-isolation produces distinct features. Furthermore, the three order imbalances obtained by decomposing non-isolated trades are also strongly correlated with the aggregated non-isolated order imbalance, but weakly correlated with each other. Upon exploring their relations in more detail, we find that the non-self-isolated order imbalances derived from orders which are not co-traded with other stocks in the market, are relatively more correlated with isolated order imbalances. In contrast, the order imbalances of non-cross-isolated and non-both-isolated trades, which are more connected with the market, are less correlated with the isolated and non-self-isolated order imbalances. Therefore, we are confident that the decomposed order imbalances are distinguishable features, with all pairwise correlations smaller than 0.6, that they can reveal insights about structural properties of the equity market which cannot otherwise be inferred by looking at the aggregated order flow.
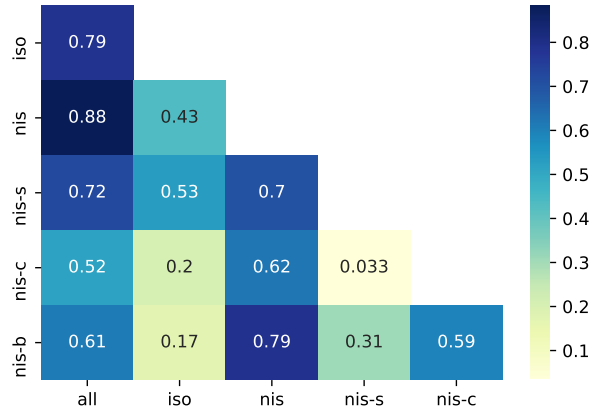
Figure 4: Pearson correlation of order imbalances. For each type of order imbalance, we first consider the vector of daily values during 2017-01-03 to 2019-12-10, then compute the correlation matrix, and finally average across all stocks.

5 Contemporaneous Price Impacts of Conditional Order Imbalances

To assess the contemporaneous effects of each type of order imbalance on future returns, we employ the following time-series regression for each individual stock i

$$r_{i,t} = \alpha_i + \beta_i COI_{i,t}^* + \epsilon_{i,t}, \tag{6}$$

where $r_{i,t}$ is the return of stock i at time t; $\epsilon_{i,t}$ is assumed to be mean zero normally distributed and the random variables $\{\epsilon_{i,t}\}$ are assumed to be independent. For inference, we apply the Newey–West estimator (Newey and West (1994)) to correct for heteroscedasticity and auto-correlation in the residual terms $\epsilon_{i,t}$. We apply two-tailed t-tests and use a 5% significance level.

In Table 4, we report results of the contemporaneous regressions. Consistent with previous research, the unconditional order imbalances are positively and significantly related to returns, for almost all stocks. Furthermore, our conditional order imbalances (COI) also express significantly positive influence on the same-day *contemporaneous* returns, especially isolated COI. It is noteworthy that impacts of the three types (nis-s, nnis-c and nis-b) of order imbalances derived from decomposing non-isolated trades have comparatively weaker influences with respect to mean value of coefficients and proportion of significant coefficients.

Focusing on the percentage of variance explained by the regression on contemporane-

Table 4: Contemporaneous Regression.
This table summarizes the results of 457 regressions using Equation (6). 'Average coefficient' denotes the mean of all regressions coefficients. 'Percentage' denotes proportion of selected stocks. 'Significant' denotes statistically significant at 5% significance level using two-tailed t test. 'Average R^2' denotes the regression R^2 averaged across all stocks.

	Average coefficient	Percentage positive	Percentage positive and significant	Percentage negative and significant	Average $R^2(\%)$
all	2.56	100.00	97.16	0.00	7.16
iso	2.00	99.34	96.06	0.00	7.15
nis	1.80	97.81	91.03	0.22	4.55
nis-s	1.11	88.62	69.80	2.64	2.83
nis-c	0.82	93.87	76.37	0.87	2.48
nis-b	0.84	95.62	81.62	0.66	2.61

ous returns, we find that the 'all' COI, calculated with all trades, generates the highest R^2 of 7.16%. After decomposing trade flows, the 'iso' COI, although calculated with only 38.29% of trades, explains a comparable amount of variance as 'all' COI. Regressing returns against 'nis' COIs achieves a lower R^2 than regressing against 'all' or 'iso' COIs. Hence, the price impact is not proportional to the quantity but appears to be driven by the types of trades. It indicates that price pressures generated by trades with distinct co-occurrence relations with other trades in the market are inhomogeneous and warrant studying separately.

In addition to the significant effect of individual conditional order imbalances on returns, we are also interested in the extra information gained from decomposing aggregated order imbalances. To this end, we fit the following three regressions

$$
\begin{aligned}
r_{i,t} &= \alpha_i + \beta_i^{all} COI_{i,t}^{all} + \epsilon_{i,t} \\
r_{i,t} &= \alpha_i + \beta_i^{iso} COI_{i,t}^{iso} + \beta_i^{nis} COI_{i,t}^{nis} + \epsilon_{i,t} \\
r_{i,t} &= \alpha_i + \beta_i^{iso} COI_{i,t}^{iso} + \beta_i^{nis-s} COI_{i,t}^{nis-s} + \beta_i^{nis-c} COI_{i,t}^{nis-c} + \beta_i^{nis-b} COI_{i,t}^{nis-b} + \epsilon_{i,t},
\end{aligned}
\tag{7}
$$

where β_i^* are regression coefficients, and $\epsilon_{i,t}$ are residual terms assumed to be Gaussian noise. Each regression takes as input a group of COIs derived from a partition of the entire trade flow. We use R^2 to evaluate the performances. Taking the influence of feature numbers into account, we also use the adjusted R^2 as an evaluation metric.

From Table 5 we can observe evident improvements in both R^2 and adjusted R^2 when

Table 5: Contemporaneous Regression R^2.

We run the three regressions in Equation (7) for each of the selected 457 stocks and report the R^2 and adjusted R^2, averaged over stocks. We document the increases in adjusted R^2 of each regression compared with regression of its previous row in the column '$\Delta adj.R^2$'. The superscript *** indicates significant at 1% using one-tailed paired t-test and corresponding t-values are reported in the last column.

	R^2	adj. R^2	Δ adj. R^2	t value
all	7.15	7.03	—	—
iso + nis	8.20	7.95	0.92***	19.45
iso + nis-s + nis-c + nis-b	9.50	9.00	1.05***	19.02

taking trade type into account. Using the unconditional model as benchmark, splitting market orders into *isolated* and *non-isolated* explains 0.92% more of total variance. By further decomposing *non-isolated* trades, an extra 1.05% growth in R^2 is achieved. Both improvements are statistically significant. In conclusion, we successfully separate trades with different contemporaneous price impact from the entire trade flow, and the decomposition helps explain daily price changes.

6 Predictive Power of Imbalances on Future Returns

In conjunction with contemporaneous effects of order imbalances, it is also important to study their forecasting power. In this section, we show that iso and nis-s order imbalances are positively related to future returns, while nis, nis-c and nis-b COIs are negatively correlated with future returns. Moreover, we discover that decomposing trade flows and simultaneously using multiple COIs contain signals for forecasting next-day returns. We provide evidence, using both regression and portfolio sorting approaches.

6.1 Time Series Regression

We repeat the regression analysis procedures in the previous section to explore the connection between COIs and one-day ahead market-excess returns. Firstly, we regress future returns against current COIs individually, under the model

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t}COI^*_{i,t} + \epsilon_{i,t+1}, \tag{8}$$

Table 6: Predictive Regression.

This table summarizes the results of the linear regressions, using Equation (8), one for each of the selected 457 stocks. Average coefficient denotes the cross-sectional mean of all regressions coefficients. 'Percentage' denotes proportion of selected stocks. 'Significant' denotes statistically significant at 5% significance level using two-tailed t test. 'Average R^2' denotes the regression R^2 averaged over all stocks.

	Average coefficient	Percentage positive	Percentage positive and significant	Percentage negative and significant	Average $R^2(\%)$
all	-0.04	47.48	2.84	5.25	0.15
iso	0.05	56.89	5.47	2.84	0.17
nis	-0.09	42.89	1.97	6.13	0.15
nis-s	0.03	54.92	4.81	3.28	0.16
nis-c	-0.10	33.26	1.31	10.50	0.20
nis-b	-0.08	34.79	1.31	6.13	0.15

where $\beta_{i,t}$ is the regression coefficient, and $\epsilon_{i,t}$ are the residual terms assumed to be independent and identically distributed with mean zero normal distributions.

Table 6 documents the regression results. As expected, unlike contemporaneous impact, both the magnitudes and percentages of significant coefficients are low, with average coefficients for unconditional order imbalances being approximately equal to zero. Over our study period, we do not find evidence to support the theoretical model put forth by Chordia and Subrahmanyam (2004), which would yield a positive relationship between imbalances and one-day ahead returns, in the absence of future order imbalance. However, with our decomposition of trades into categories, we can slightly strengthen the above signals. Our findings suggest that the price pressures which arose from *isolated* and *non-self-isolated* order executions show moderate predictive power. Additionally, *non-isolated (nis)*, *non-cross-isolated* and *non-both-isolated* trade imbalances are negatively associated with future price changes. In particular, *nis* COI has the largest mean coefficients, with 10.5% of estimated coefficients for individual stocks being negative and significant while only 1.31% of the coefficients are positive and significant. In term of R^2, all COIs of the decomposed trade flows outperform the COI of the undecomposed (i.e., aggregated) trade flow.

In the next step, we regress future 1-day stock returns against each of the three groups

Table 7: Predictive Regressions R^2.

We run the three regressions in Equation (9) for each of the selected 457 stocks and report the R^2 and adjusted R^2, averaged over stocks. We document the increases in adjusted R^2 of each regression compared with regression of its previous row in the column '$\Delta adj.R^2$'. The superscript $***$ indicates significant at 1% using one-tailed paired t test and corresponding t-values are reported in the last column.

	R^2	adj. R^2	Δ adj. R^2	t value
all	0.15	0.016	—	—
iso + nis	0.34	0.061	0.045***	15.92
iso + nis-s + nis-c + nis-b	0.64	0.094	0.033***	19.82

of COIs, along the following models

$$r_{i,t+1} = \alpha_i + \beta_i^{all} COI_{i,t}^{all} + \epsilon_{i,t}$$

$$r_{i,t+1} = \alpha_i + \beta_i^{iso} COI_{i,t}^{iso} + \beta_{i,t}^{nis} COI_{i,t}^{nis} + \epsilon_{i,t}$$

$$r_{i,t+1} = \alpha_i + \beta_i^{iso} COI_{i,t}^{iso} + \beta_i^{nis-s} COI_{i,t}^{nis-s} + \beta_i^{nis-c} COI_{i,t}^{nis-c} + \beta_i^{nis-b} COI_{i,t}^{*,nis-b} + \epsilon_{i,t},$$

(9)

where the β_i^* are regression coefficients, and $\epsilon_{i,t}$ denote residual terms assumed to be independent and identically distributed mean zero Gaussian noise.

From Table 7, we observe stable improvements in both R^2 and adjusted R^2 values. Cross-sectionally, the growth in adjusted R^2 is statistically significant at the 1% level. Therefore, we conclude that the order imbalances conditioning on co-occurrence can contribute to short-term return forecasting. We conjecture that decomposing trade flows according to such COIs improves predicting future returns.

6.2   Imbalance-Based Portfolio Sorting

To bolster our findings on the positive and negative relations between future returns and different types of COI, we apply the portfolio sorting methods (Cattaneo et al. (2020), Fama and French (1993)) to translate order imbalances into portfolios. For each type of COI, we sort stocks according to their imbalance values, from low to high, into 5 quintile portfolios. Taking multiple features into account, we further create 5 × 5 double-sort portfolios, for every pair of COIs. The imbalance-sorted portfolios are equally weighted and have only long positions on stocks, with daily portfolio returns calculated as the average returns of all stocks in them. Backtests of imbalance-sorted portfolios, over the

entire sample period from 2017-01-03 to 2019-12-10, reinforce the finding that *iso* and *nis-s* imbalances are momentum signals, while the *nis, nis-c and nis-b* imbalances are reversal signals, and that they have different influence on future returns.

6.2.1 Single-Sort Portfolios

Panel A of Table 8 documents the annualized returns of single-sort portfolios. We note, in the first row, that the returns of the unconditional-imbalance-sorted portfolios are negative and fluctuate along quintiles, which confirms the absence of clear linear relations between unconditional order imbalance and future return. However, after performing the decomposition, we find that the growth in returns with increasing *iso* order imbalance is monotonic, which reinforces its positive correlation with future returns. Similarly, there is also an increasing trend for *nis-s*, which is a sign of positive correlation. In contrast, we observe declines in average returns along other types of COIs, which echos our time series regression results and confirms negative correlations.

Panel B shows daily COIs averaged over stocks in each portfolio. The COIs are signed, denoting that 'Low' and 'High' portfolios correspond to strong signals with opposite signs. We observe that the distributions of all signal strengths are roughly symmetric and centered around 0. In each row, there are no quintile portfolios consisting of stocks with indistinguishable average COI values. However, the portfolio returns are neither symmetric nor monotonic along quintiles (except *iso*). By comparing returns in each row of Panel A, we observe that the magnitudes of the most positive returns are always smaller than the absolute values of the most negative returns. Therefore, we conjecture that the positive and negative impacts of COIs on future returns are asymmetric, with negative values being more influential.

Furthermore, for negative impacts on future returns, the highest magnitudes in COIs do not lead to the largest next day decreases. For example, the 'Low' and 2nd quintile portfolios of *iso* COI have similar returns and, for portfolios of *nis COI*, the 4th quintile reaches the lowest average return of -3.85%, while the return of the highest quintile rise to -2.63%. As interpretation of this phenomenon we propose that extreme imbalances can lead to strong reversal on the following day, because some investors aim to maintain stable levels of risk exposures.

Table 8: Summary of single-sort portfolios.

This table shows the statistics of COI-sorted quintile portfolios. Each row contains five portfolios constructed by sorting all stocks every day by the type, indicated by its row index, of COI on the previous day from low to high and allocating each stock to the corresponding quintile portfolio indicated by the column names. The breakpoints are $20\%, 40\%, 60\%$ and 80% of each type of COI calculated daily. Panel A presents the annualized return of each portfolio calculated by averaging its daily returns, from 2017-01-03 to 2019-09-10, and multiplying by 252. Panel B, reports the average daily COIs of stocks included in portfolio over the sample period.

Pannel A: Annualized returns					
	Low	2	3	4	High
all	-1.57	-1.77	-1.95	-3.61	-0.49
iso	-4.25	-4.11	-2.13	-0.91	2.00
nis	0.54	-1.76	-1.69	-3.85	-2.63
nis-s	-4.09	-2.56	-2.28	-2.70	2.25
nis-c	4.67	-0.56	-2.75	-5.72	-5.03
nis-b	3.60	-1.91	-1.87	-4.55	-4.66

Pannel B: Average daily COIs					
	Low	2	3	4	High
all	-0.16	-0.05	0.00	0.06	0.17
iso	-0.20	-0.06	0.01	0.08	0.22
nis	-0.18	-0.06	0.00	0.06	0.17
nis-s	-0.20	-0.07	0.01	0.08	0.21
nis-c	-0.24	-0.10	-0.03	0.05	0.19
nis-b	-0.28	-0.11	-0.01	0.08	0.24

6.2.2 Double-Sort Portfolios

To future investigate the interplay between COIs, we build portfolios by independently double-sorting on every pair of imbalances of decomposed trade flows. Table 9 presents the annualized returns of all portfolios, where each block contains 25 portfolios by sorting on a pair of signals indicated by row and column names.

In each column of the iso−nis block, the average returns rise from low to high COIs of isolated trades. In contrast, controlled with *iso* COI, the returns typically fall from low to high *non-isolated* COI. Double-sorting on the strongest signals generates the highest and lowest returns, on the upper-right and bottom-left corner of the block. The magnitudes of the strongest returns, 11.37% and -14.53%, are also amplified compared with sorting on one single signal. The same patterns and improvements appear when double-sorting on every pair of momentum and reversal COI features, except *nis-s−nis-b*. Nonetheless, the patterns in the pairs of COIs with the same signs of relation with future returns

are obscure. For example, looking at the blocks of iso−nis-s sorts, we cannot observe monotonic patterns along rows and columns. Therefore, we conjecture that momentum and reversal signals carry distinct information and incorporating them simultaneously boosts predictive powers.

7 Economic Value of Conditional Order Imbalances

As discussed in previous sections, there is evidence that conditional order imbalances contain signals for explaining and forecasting individual stock returns. In this section, we exploit their economic values by forming long-short portfolios using sorts. Our imbalance-based trading strategies generate conspicuous profits and significant abnormal returns. High trading profits also provide important evidence of the predictive power which the COIs of the decomposed trade flows possess.

7.1 Long-Short Portfolio Construction and Evaluation

We design practical trading strategies based upon imbalance-sorted quintile portfolios. At 9:30am of each trading day, we buy the first (resp., last) and short sell the last (resp., first) quintile portfolios for momentum (resp., reversal) signals with the same amount such that they are self-rebalancing. Every day, we close all position at 16:00pm to avoid overnight effects. Overall, the daily returns are the differences between the returns of the long and short imbalance-sorted portfolios.

To evaluate profitability, we compare the annualized returns of the portfolios, as well as the annualized Sharpe ratio (Sharpe (1994)), defined as

$$SR_p := \frac{mean(R_{p,t}) - R_f}{std(R_{p,t})} \times \sqrt{252}, \tag{10}$$

where $R_{p,t}$ are daily returns of the portfolios and R_f is the average daily risk-free rate, which equals 0.00625% during the period of interest.

7.2 Profitability

We construct long-short portfolios and report their profitability measures in Table 10. Panel A displays the annualized returns, with on- and off-diagonal values for single-

Table 9: Annualized returns of double-sort portfolios.

This table presents annualized returns of double-sort portfolios based on every pair of COIs. Each panel contains 5 × 5 double-sort portfolios. To construct the portfolios, we sort all stocks every day by the two types, indicated by its row index and column name, of COIs on the previous day, from low to high, to five quintiles independently. Then intersections of the two sorts create 25 double-sort portfolios. The annualized return of each portfolio is calculated by averaging its daily returns, from 2017-01-03 to 2019-09-10, and multiplying by 252.

Rows: iso, Columns: nis

	Low	2	3	4	High
Low	-0.99	-3.96	-7.91	-10.97	-14.53
2	-1.85	-6.37	-1.72	-4.35	-9.36
3	1.53	1.40	-0.36	-3.42	-9.84
4	6.67	-1.32	-0.78	-4.34	-2.51
High	11.37	5.12	2.63	-0.14	2.13

Rows: iso, Columns: nis-c

	Low	2	3	4	High
Low	0.26	-3.51	-7.36	-6.75	-11.63
2	7.20	-1.49	-7.10	-8.76	-9.40
3	5.94	-0.05	-1.30	-5.75	-6.93
4	3.54	0.82	3.63	-7.87	-4.87
High	17.65	8.04	-0.13	3.56	-1.16

Rows: nis, Columns: nis-s

	Low	2	3	4	High
Low	-1.50	2.03	9.51	3.92	12.41
2	-8.91	-1.40	-0.68	5.32	10.06
3	-6.94	-5.46	-2.40	0.48	9.35
4	0.15	-9.50	-5.94	-3.67	3.16
High	-19.3	1.73	-11.61	-9.62	0.44

Rows: nis, Columns: nis-b

	Low	2	3	4	High
Low	2.88	-4.62	-6.16	0.16	0.95
2	5.12	-3.16	-1.69	-12.87	-1.21
3	10.16	-0.26	-1.99	-7.03	2.52
4	6.24	8.17	-1.41	-6.59	-7.73
High	-7.31	13.17	-3.48	2.61	-4.79

Rows: nis-s, Columns: nis-b

	Low	2	3	4	High
Low	1.21	-10.21	-8.62	-9.36	-4.83
2	1.10	-0.92	-1.23	-12.11	-3.38
3	7.68	-4.22	-0.69	-5.57	-11.59
4	5.73	2.84	-2.01	-2.46	-10.15
High	6.14	11.71	1.79	-0.02	-0.16

Rows: iso, Columns: nis-s

	Low	2	3	4	High
Low	-4.04	-3.37	-3.54	-11.71	-4.70
2	-8.91	-4.41	-3.37	0.18	0.84
3	-2.00	-2.34	-1.10	-4.17	1.85
4	5.85	-1.85	-2.93	-4.71	2.26
High	-9.24	3.73	3.19	1.83	2.54

Rows: iso, Columns: nis-b

	Low	2	3	4	High
Low	0.58	-4.41	-6.89	-8.08	-9.22
2	1.06	-3.93	-3.07	-9.25	-6.86
3	5.64	-1.27	0.59	-7.17	-6.85
4	6.90	1.07	-0.69	-5.00	-5.66
High	9.39	2.96	0.78	5.28	0.27

Rows: nis, Columns: nis-c

	Low	2	3	4	High
Low	3.90	-1.35	-4.49	-4.44	-8.27
2	2.94	-2.89	0.09	-7.59	-0.93
3	7.58	3.77	-8.48	-4.87	-7.06
4	3.42	-4.54	-0.17	-7.20	-5.95
High	5.49	-3.71	3.11	-2.51	-4.50

Rows: nis-s, Columns: nis-c

	Low	2	3	4	High
Low	2.49	-4.20	-6.03	-8.86	-8.27
2	1.15	-1.41	-1.86	-6.83	-3.61
3	6.24	-2.66	-4.14	-4.35	-6.79
4	4.95	0.67	-5.2	-5.53	-6.93
High	8.40	5.82	6.08	-1.66	-2.84

Rows: nis-c, Columns: nis-b

	Low	2	3	4	High
Low	6.42	1.68	2.42	5.58	6.17
2	2.27	-1.12	1.63	-5.22	-4.60
3	1.50	-2.99	-1.39	-7.6	-1.43
4	4.19	-5.67	-8.07	-6.04	-5.22
High	-9.01	-4.11	-1.08	-3.45	-6.83

Table 10: Profitability of long-short portfolios.

This table shows the annualized returns and Sharpe ratios of the long-short portfolios sorted on COIs indicated by the corresponding row indices and column names. The on- and off- diagonal values are for single- and double-sort portfolios respectively. Panel A presents the annualized return of portfolios calculated by averaging their daily returns, from 2017-01-03 to 2019-09-10, and multiplying by 252. Panel B reports the annualized Sharpe ratios over the sample period calculated by Equation (10).

Panel A: Annualized returns					
	iso	nis	nis-s	nis-c	nis-b
iso	6.25	25.85	6.58	29.39	18.55
nis		3.18	31.71	8.41	7.65
nis-s			6.34	16.74	11.10
nis-c				9.70	13.25
nis-b					8.26

Panel B: Annualized Sharpe ratios					
	iso	nis	nis-s	nis-c	nis-b
iso	1.29	1.67	0.99	2.38	1.77
nis		0.41	2.09	1.19	1.26
nis-s			1.36	1.65	0.72
nis-c				1.62	1.78
nis-b					1.58

and double-sort portfolios based on COIs indicated by row and column names. We find that incorporating multiple COIs improves the profit of portfolios, which is supporting evidence that the trade flow decomposition technique creates profitable COI signals. For example, the return of the long-short strategy corresponding to $iso-nis$ double-sort is 25.85%, which is 19.60% and 22.67% higher than simply sorting on iso and nis COI separately. The highest annualized return hits 31.71% by double-sorting on $nis\text{-}s$ and nis COIs. Generally, when comparing off-diagonal values with the corresponding diagonal values on their rows and columns, we find that the returns of all 10 double-sort long-short portfolios are higher than at least one of their single-sort benchmarks, and 9 out of 10 are higher than both values.

The Sharpe ratios in Panel B strengthen our findings on the economic value of COI signals. Adjusted for risks, our trading strategies remain profitable, and double-sorting outperforms trading on signals individually. The portfolio sorted on iso and $nis\text{-}c$ achieves the highest Sharpe ratio of 2.38, followed by 2.09 of $nis\text{-}s-nis$ sorted portfolio. Moreover, 80% of off-diagonal values are higher than at least one of their corresponding diagonal

Table 11: Abnormal returns of long-short portfolios.
This table documents the abnormal returns, α, of long-short portfolios with respect to Fama-French 5 factor models (Fama and French (2015)). For each long-short portfolio, we run time series regressions on portfolio excess returns against 5 factors,

$$R_{p,t} - R_{f,t} = \alpha_p + b_p(R_{M,t} - R_{f,t}) + s_p SMB_t + h_p HML_t + r_p RMW_t + c_p CMA_t + e_{p,t},$$

where α_p is the abnormal return of the portfolio, the explanatory variables are the market, size, value, profitability and investment factors and $e_{p,t}$ is the idiosyncratic term. The on- and off-diagonal values are for single- and double-sort long-short portfolios respectively. The superscripts *, ** and *** indicate statistical significance at 10%, 5% and 1%, and the corresponding t-values are reported in the parentheses.

Annualized alpha					
	iso	nis	nis-s	nis-c	nis-b
iso	5.23**	5.23***	5.11*	29.11***	17.27***
	(2.46)	(2.79)	(1.70)	(4.21)	(3.06)
nis		1.54	31.02***	7.29**	6.27**
		(0.68)	(3.63)	(2.14)	(2.19)
nis-s			4.92**	15.32***	9.72
			(2.38)	(2.82)	(1.25)
nis-c				7.78***	12.05***
				(2.65)	(3.10)
nis-b					6.63***
					(2.65)

values, and 50% are higher than both values. Therefore, there is some indication that for investors it is economically beneficial to incorporate multiple types of COIs when making trading decisions.

From the perspective of asset pricing, COIs are unique and significant sources of abnormal returns. We regress the excess returns of the long-short portfolios against the Fama-French five factors[1] (Fama and French (2015)) and show their alphas in Table 11. All of the portfolios, except single-sort on *nis* COI, generate statistically significant abnormal returns, providing evidence that the profits cannot be explained by common risk factors.

We also compare our strategies with three benchmark portfolios. First, we construct a long-short portfolio of *'all'* COI to assess the economic value of trade flow decomposition. Second, we build another benchmark long-short portfolio based on returns of previous days. Because COIs and contemporaneous returns are significantly correlated,

---

[1]We obtain the data of factors from Kenneth R. French − Data Library.
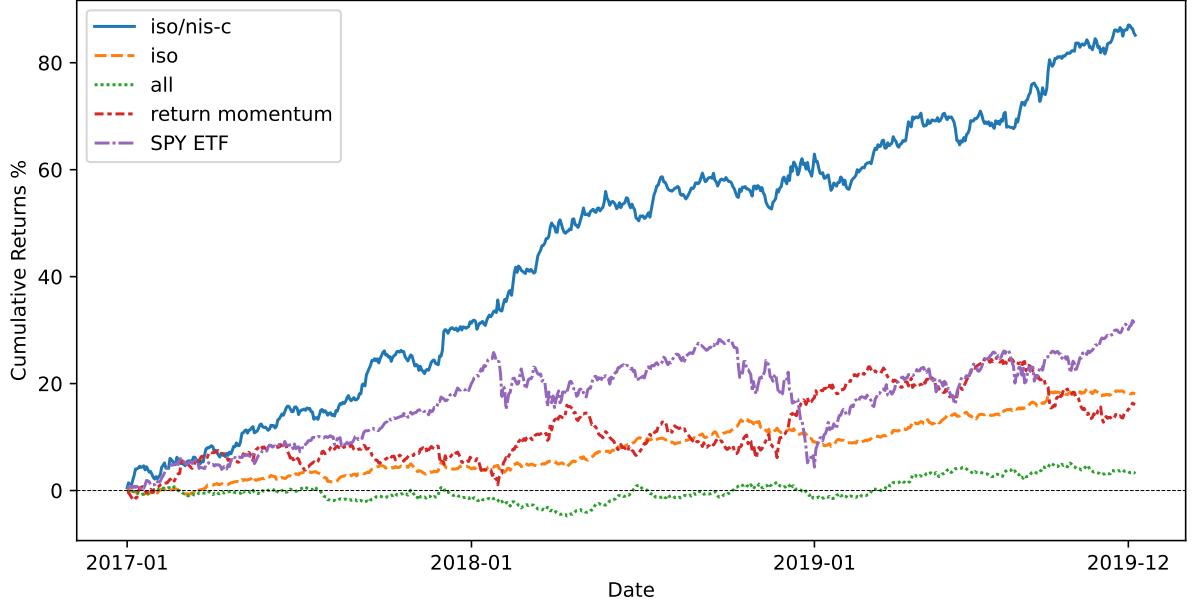
Figure 5: Cumulative returns of portfolios.
This figure plots cumulative returns of five portfolios from 2017-01-03 to 2019-12-10. The portfolios include (1) 'iso/nis-c': the long-short portfolio double-sorted on iso and nis-c COIs; (2)'iso': the long-short portfolio single-sorted on iso COI; (3)'all': the long-short portfolio single-sorted on COI of decomposed trade flows; (4)'return momentum': the long-short portfolio single-sorted on previous days' returns; (5)'SPY ETF': the SPDR S&P 500 ETF Trust which tracks the S&P 500 Index.

it is necessary to show that the profitability is not fully revealed by prices. Third, we choose SPY as a tradeable market portfolio to benchmark with overall market performance. From the COI-based strategies, we select the single-sort portfolio of *iso* COI and the double-sort portfolio of *iso* and *nis-c* COIs as representatives. Figure 5 visualizes cumulative returns of selected COI-based long-short portfolios and benchmarks. Over the test period, we observe that using COIs of the decomposed trade flow attains no profits. The return momentum long-short portfolio and SPY have Sharpe ratios of 0.45 and 0.73 respectively. The *iso* single-sort portfolio has similar annualized return as return momentum but much less volatility, and attains a Sharpe ratio of 1.29. Clearly, the double-sort portfolio surpasses all other portfolios with superior return and Sharpe ratio.

8 Robustness Analysis

In this section, we briefly comment on the robustness of the identification of trade co-occurrences, and the construction of conditional order imbalances. Further details are provided in the Appendix.

27

8.1 Neighbourhood Size Effect

The definition of trade co-occurrence and classification of individual trades depends on the choice of neighbourhood size δ. When considering the extreme case of $\delta = 0$, all trades are isolated. As we progressively increase δ, an isolated trade turns into one sub-type of non-isolated trades. Meanwhile, both non-self-isolated and non-cross-isolated trades can only become non-both-isolated. Eventually, when δ is large enough, all trades are non-isolated; to be specific, they all become non-both-isolated. Hence, with the value δ increasing, the number of isolated trades decreases and the numbers of non-isolated and non-both-isolated trades increase monotonically. Thus, the quantities of non-self-isolated and non-cross-isolated trades initially increase; after reaching their respective maximum, they begin to decrease.

We replicate our analysis for eight values of δ's in Appendix A. The patterns in contemporaneous impact and predictive power are robust for small neighbourhood sizes. Nevertheless, when δ reaches 50 milliseconds, the performance of trade co-occurrence as a filter drops. In addition, we achieve the best results of different types of COIs at different δ values, hinting at the potential benefit of the approach to combine signals derived from multiple values of δ.

8.2 Time-of-Day Effect

Trading activities during different intraday periods have different impact on prices. As we notice, trading activities are more intensive during the first and the last half hour of each trading day. Some recent works, such as Cont, Cucuringu, and Zhang (2021), exclude these volatile periods when they calculate imbalances for robustness, while others (Chu and Qiu (2021)) pay special attention to imbalances during these half-hour intervals. Taking this time-of-day effect into account, we study COIs within three time intervals, namely $9:30 - 10:00$, $10:00 - 15:30$, and $15.30 - 16:00$ separately, and document our findings in Appendix B.

Our findings on contemporaneous return-imbalance relations hold for every period. Additionally, we find that the predictive power of the decomposed trade flows originates from different time periods. The *iso* and *nis-s* COIs of the last hour contribute to forecasting future returns. On the other hand, the *nis-c* COI's forecasting power stems

from periods other than the last half an hour. Moreover, for the *nis-b* trades, only the COI pertaining to 10:00−15:30 help anticipate the next-day open-to-close market excess returns.

8.3   COI Measured by Volumes

Apart from incorporating the number of transactions, it is also common to define order imbalance as the normalized difference between volumes of buyer- and seller-initiated trades. We study the relation between individual stock returns and volume order imbalances, and analyze the corresponding trading strategies. Further details are included in Appendix C.

Our findings are robust under the volume measure. We observe the same patterns as count COIs, but notice that the R^2's of contemporaneous regressions against volume imbalances are approximately 1% lower than those against count imbalances, for all types of trades. This finding is in line with previous research (Chan and Lakonishok (1995), Chordia and Subrahmanyam (2004)) which provided evidence that the number of transactions better capture the price pressure from institutions who intend to split their orders for optimal execution.

9   Conclusion and Future Directions

In this paper, we propose the idea of trade co-occurrence, which relates trades arriving close to each other in time, and enables the study of interactions among stock transactions at a granular level. Conditional on co-occurrence with other trades, we classify every single trade into five groups. We calculate order imbalances for each type of decomposed trade flow (COI), and investigate their contemporaneous impacts and forecasting power on individual stock returns, as well as their economic value.

Our empirical results show that the decomposed trade flows have different price impacts. The COI of *iso* trade flow alone can explain a comparable amount of variation in same-day returns as using COI of all trades without the decomposition, while incorporating COIs of other trade flows further improves the explainability. For predictability, we observe that future returns, on average, are positively related with *isolated* and *nis-s* COIs, while negatively related with *nis, nis-c* and *nis-b* COIs. Furthermore, the trade

flow decomposition has significant economic value, and constructing long-short portfolios based on the directions of previous days' COIs leads to conspicuous enhancements in the profitability of trading strategies.

Finally, we suggest two future directions, particularly motivated by our current limitations concerning data availability and computation power. First, we empirically show the significance of decomposing trades based on their co-occurrence with other trades, but we cannot identify who initiates certain types of trades. It would be an interesting research direction to distinguish different types of traders by leveraging private data sets (Tumminello et al. (2012), Cont, Cucuringu, Glukhov, et al. (2021)), and discover the mechanics behind the interaction of trades. For example, it would be of interest to detect whether informed traders, such as institutions, may successfully hide their trading purpose, leading to their transactions most likely to be isolated from those of others. If high-frequency traders can be identified, it is worth applying the co-occurrence analysis to understand how HFT react to trading activities of other market participants. Second, for data reduction purposes we only study the executions of limit orders, rather than all limit order book events. Past studies have found that submissions and cancellations of limit orders also lead to price impact. It may be interesting to extend our idea to the co-occurrence of limit orders and look deeper into order flow imbalances (Eisler, Bouchaud, and Kockelkoren (2012), Cont, Kukanov, and Stoikov (2014), Xu, Gould, and Howison (2018), Cont, Cucuringu, and Zhang (2021)), which are the analogues of our COIs.

References

[1]   Jesse S Aaron, Aaron B Taylor, and Teng-Leong Chew. "Image co-localization–co-occurrence versus correlation". In: *Journal of cell science* 131.3 (2018), jcs211847.

[2]   Anne E Appel and George W Holden. "The co-occurrence of spouse and physical child abuse: a review and appraisal." In: *Journal of family psychology* 12.4 (1998), p. 578.

[3]   Miguel B Araújo et al. "Using species co-occurrence networks to assess the impacts of climate change". In: *Ecography* 34.6 (2011), pp. 897–908.

[4] Warren Bailey et al. "Stock returns, order imbalances, and commonality: Evidence on individual, institutional, and proprietary investors in China". In: *Journal of Banking & Finance* 33.1 (2009), pp. 9–19.

[5] Markus K Brunnermeier and Lasse Heje Pedersen. "Predatory trading". In: *The Journal of Finance* 60.4 (2005), pp. 1825–1863.

[6] Matias D Cattaneo et al. "Characteristic-sorted portfolios: Estimation and inference". In: *Review of Economics and Statistics* 102.3 (2020), pp. 531–551.

[7] Sugato Chakravarty et al. "Clean sweep: Informed trading through intermarket sweep orders". In: *Journal of Financial and Quantitative Analysis* 47.2 (2012), pp. 415–435.

[8] Louis KC Chan and Josef Lakonishok. "The behavior of stock prices around institutional trades". In: *The Journal of Finance* 50.4 (1995), pp. 1147–1174.

[9] Tarun Chordia, Amit Goyal, and Narasimhan Jegadeesh. "Buyers versus sellers: who initiates trades, and when?" In: *Journal of Financial and Quantitative Analysis* 51.5 (2016), pp. 1467–1490.

[10] Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. "Order imbalance, liquidity, and market returns". In: *Journal of Financial Economics* 65.1 (2002), pp. 111–130.

[11] Tarun Chordia and Avanidhar Subrahmanyam. "Order imbalance and individual stock returns: Theory and evidence". In: *Journal of Financial Economics* 72.3 (2004), pp. 485–518.

[12] Xiaojun Chu and Jianying Qiu. "Forecasting stock returns using first half an hour order imbalance". In: *International Journal of Finance & Economics* 26.3 (2021), pp. 3236–3245.

[13] Rama Cont, Mihai Cucuringu, Vacslav Glukhov, et al. "Analysis and modeling of client order flow in limit order markets". In: *Available at SSRN* (2021).

[14] Rama Cont, Mihai Cucuringu, and Chao Zhang. "Price impact of order flow imbalance: multi-level, cross-sectional and forecasting". In: *arXiv e-prints* (2021), arXiv–2112.

[15]   Rama Cont, Arseniy Kukanov, and Sasha Stoikov. "The price impact of order book events". In: *Journal of financial econometrics* 12.1 (2014), pp. 47–88.

[16]   Justin Cox. "ISO order imbalances and individual stock returns". In: *Journal of Financial Research* 44.1 (2021), pp. 5–23.

[17]   Ido Dagan, Lillian Lee, and Fernando CN Pereira. "Similarity-based models of word cooccurrence probabilities". In: *Machine learning* 34.1 (1999), pp. 43–69.

[18]   Jonathan F Donges et al. "Event coincidence analysis for quantifying statistical interrelationships between event time series". In: *The European Physical Journal Special Topics* 225.3 (2016), pp. 471–487.

[19]   Zoltan Eisler, Jean-Philippe Bouchaud, and Julien Kockelkoren. "The price impact of order book events: market orders, limit orders and cancellations". In: *Quantitative Finance* 12.9 (2012), pp. 1395–1419.

[20]   Eugene F Fama and Kenneth R French. "A five-factor asset pricing model". In: *Journal of financial economics* 116.1 (2015), pp. 1–22.

[21]   Eugene F Fama and Kenneth R French. "Common risk factors in the returns on stocks and bonds". In: *Journal of Financial Economics* 33.1 (1993), pp. 3–56.

[22]   Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. "Object categorization using co-occurrence, location and appearance". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[23]   Nicholas J Gotelli. "Null model analysis of species co-occurrence patterns". In: *Ecology* 81.9 (2000), pp. 2606–2621.

[24]   Sanford J Grossman and Merton H Miller. "Liquidity and market structure". In: *the Journal of Finance* 43.3 (1988), pp. 617–633.

[25]   Li Guo et al. "News co-occurrence, attention spillover, and return predictability". In: *Attention Spillover, and Return Predictability (November 18, 2018)* (2018).

[26]   Björn Hagströmer and Lars Nordén. "The diversity of high-frequency traders". In: *Journal of Financial Markets* 16.4 (2013), pp. 741–770.

[27]   Nicholas Hirschey. "Do high-frequency traders anticipate buying and selling pressure?" In: *Management Science* 67.6 (2021), pp. 3321–3345.

[28]   Ruihong Huang and Tomas Polak. "Lobster: Limit order book reconstruction system". In: *Available at SSRN 1977207* (2011).

[29]   Olga Kolesnikova. "Survey of word co-occurrence measures for collocation detection". In: *Computación y Sistemas* 20.3 (2016), pp. 327–344.

[30]   Alan Kraus and Hans R Stoll. "Parallel trading by institutional investors". In: *Journal of Financial and Quantitative Analysis* 7.5 (1972), pp. 2107–2138.

[31]   Albert S Kyle. "Continuous auctions and insider trading". In: *Econometrica: Journal of the Econometric Society* (1985), pp. 1315–1335.

[32]   Albert S Kyle, Hui Ou-Yang, and Bin Wei. "A model of portfolio delegation and strategic trading". In: *The Review of Financial Studies* 24.11 (2011), pp. 3778–3812.

[33]   Yi-Tsung Lee et al. "Order imbalances and market efficiency: Evidence from the Taiwan Stock Exchange". In: *Journal of Financial and Quantitative Analysis* 39.2 (2004), pp. 327–341.

[34]   Zhongming Ma, Gautam Pant, and Olivia RL Sheng. "Mining competitor relationships from online news: A network-based approach". In: *Electronic Commerce Research and Applications* 10.4 (2011), pp. 418–427.

[35]   Darryl I MacKenzie, Larissa L Bailey, and James D Nichols. "Investigating species co-occurrence patterns when species are detected imperfectly". In: *Journal of Animal Ecology* 73.3 (2004), pp. 546–555.

[36]   Whitney K Newey and Kenneth D West. "Automatic lag selection in covariance matrix estimation". In: *The Review of Economic Studies* 61.4 (1994), pp. 631–653.

[37]   Maureen O'Hara. "High frequency market microstructure". In: *Journal of Financial Economics* 116.2 (2015), pp. 257–270.

[38]   David S Scharfstein and Jeremy C Stein. "Herd behavior and investment". In: *The American Economic Review* (1990), pp. 465–479.

[39]   William F Sharpe. "The sharpe ratio". In: *Journal of Portfolio Management* 21.1 (1994), pp. 49–58.

[40]   Catherine Shenoy and Ying Jenny Zhang. "Order imbalance and stock returns: Evidence from China". In: *The Quarterly Review of Economics and Finance* 47.5 (2007), pp. 637–650.

[41] Matthew Spiegel and Avanidhar Subrahmanyam. "On intraday risk premia". In: *The Journal of Finance* 50.1 (1995), pp. 319–339.

[42] Hans R Stoll. "The supply of dealer services in securities markets". In: *The Journal of Finance* 33.4 (1978), pp. 1133–1151.

[43] Yi Tang, Yilu Zhou, and Marshall Hong. "News co-occurrences, stock return correlations, and portfolio construction implications". In: *Journal of Risk and Financial Management* 12.1 (2019), p. 45.

[44] Michele Tumminello et al. "Identification of clusters of investors from their real trading activity in a financial market". In: *New Journal of Physics* 14.1 (2012), p. 013041.

[45] Vincent Van Kervel and Albert J Menkveld. "High-frequency trading around large institutional orders". In: *The Journal of Finance* 74.3 (2019), pp. 1091–1137.

[46] Qiong Wu et al. "A deep learning framework for pricing financial instruments". In: *arXiv.org* (2019).

[47] Ke Xu, Martin D Gould, and Sam D Howison. "Multi-level order-flow imbalance in a limit order book". In: *Market Microstructure and Liquidity* 4.03n04 (2018), p. 1950011.

[48] Liyan Yang and Haoxiang Zhu. "Back-running: Seeking and hiding fundamental information in order flows". In: *The Review of Financial Studies* 33.4 (2020), pp. 1484–1533.

[49] Shujing Ye et al. "Co-occurrence and interactions of pollutants, and their impacts on soil remediation—a review". In: *Critical reviews in environmental science and technology* 47.16 (2017), pp. 1528–1553.

[50] Ting Zhang, Gao-Feng Gu, and Wei-Xing Zhou. "Order imbalances and market efficiency: New evidence from the Chinese stock market". In: *Emerging Markets Review* 38 (2019), pp. 458–467.

A   Neighbourhood Size Effect

To study the effect of neighbourhood size on conditional order imbalances, we repeat the regression and portfolio analysis for each $\delta \in \{0.05$ ms, 0.075 ms, 0.125 ms, 0.25 ms, 0.5
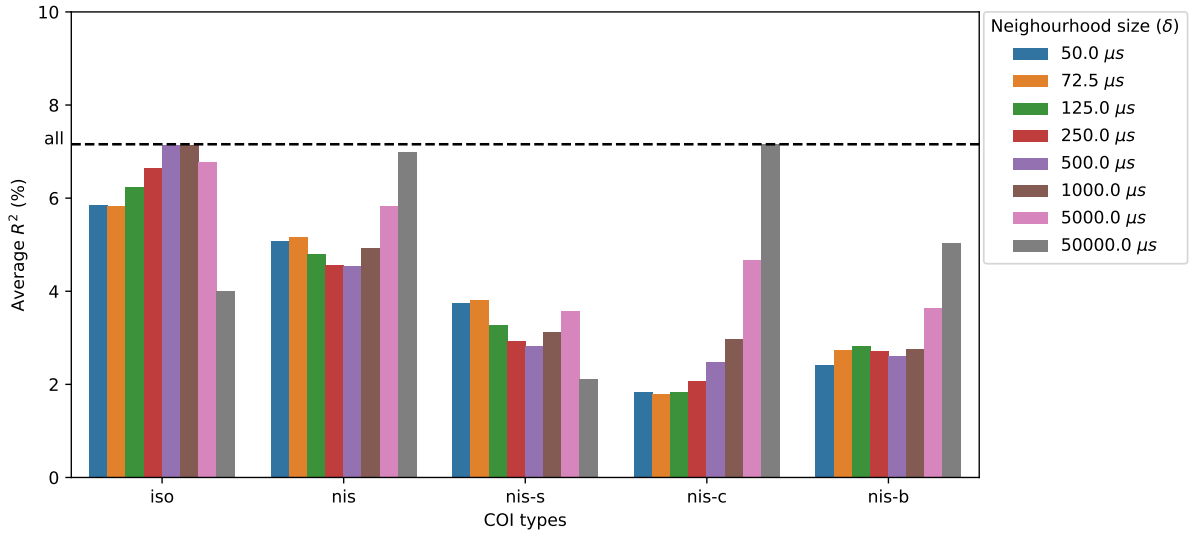
(a) Average R^2 of regressing contemporaneous returns against COIs for different δs.
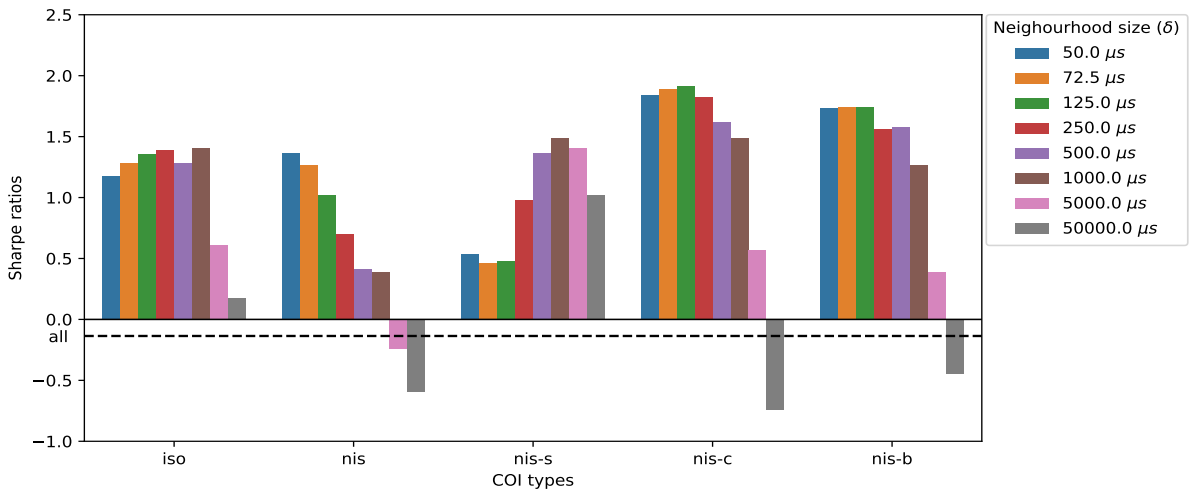
(b) Sharpe ratios of COI-based long-short portfolios for different δs.

Figure 6: Empirical results for different δs.

ms, 1 ms, 5 ms, 50 ms}, and display the results in Figure 6.

Figure 6a illustrates the average R^2 of contemporaneous regressions. Isolated order imbalances achieve the highest R^2 at $\delta = 1$ ms. In contrast, the histograms of R^2 of the non-isolated imbalances have a U-shape with minimum at $\delta = 0.5$ ms. For the three types of non-isolated order imbalances, the R^2s for non-self-isolated and non-both-isolated imbalances have downward trends with growth in values of δ. Non-cross-isolated imbalances explain more variance in returns as δ increases.

Figure 6b details the Sharpe Ratios of long-short portfolios of different COI types ordered by δ. We remark that the Sharpe Ratios of each type of order imbalance peak at different values of δ.

Table 12: COIs by time period.

We calculate COIs of $9:30 - 10:00$, $10:00 - 15:30$, and $15:30 - 16:00$ each day from 2017-01-03 to 2019-12-10 for the selected 457 stocks. Panel A summarizes the R^2, averaged over all stocks, of linear regressions on contemporaneous open-to-close market-excess? returns against each type of COIs, using Equation (6), one for each stock. Panel B presents the annualized Sharpe Ratios, given by Equation (10), of single-sort long-short portfolios based on COIs of different intraday time periods. The last column of both panels reports the daily COIs as a benchmark.

Panel A: Contemporaneous regression $R^2(\%)$				
	$9:30 - 10:00$	$10:00 - 15:30$	$15:30 - 16:00$	$9:00 - 16:00$
all	1.84	6.41	2.98	7.16
iso	3.23	6.21	2.06	7.15
nis	0.68	4.14	2.30	4.55
nis-s	0.61	2.48	1.15	2.83
nis-c	0.51	2.37	1.03	2.48
nis-b	0.21	2.38	1.39	2.61
Panel B: Annualized Sharpe ratios				
	$9:30 - 10:00$	$10:00 - 15:30$	$15:30 - 16:00$	$9:00 - 16:00$
all	-0.87	-0.65	0.79	-0.14
iso	-0.78	0.44	2.31	1.29
nis	0.73	0.67	-0.39	0.41
nis-s	-0.55	0.92	2.03	1.36
nis-c	1.21	1.39	0.29	1.62
nis-b	-0.18	1.57	0.33	1.57

B Time-of-Day Effect

We investigate the COIs of different intraday time intervals. Firstly, we evaluate their influences on same-day price change by regressing contemporaneous open-to-close market access returns against each COI individually. Panel A of Table 12 presents the R^2 of all such regressions. Excluding the first and last half hours of trades does not explicitly change the imbalance-return relations we discover. Regardless of periods, deriving COIs with only *iso* trades is enough to explain a comparable amount of variance as when using all trades. Note that, especially for the first hour, the price impact mainly stems from isolated trades.

Secondly, we trade on each COI by constructing single-sort long-short portfolios and present annualized Sharpe Ratios in Panel B. It is reasonable to expect that trading activities towards the end of the normal trading period contribute more to forecasting future returns. We observe that the signal corresponding to the *iso* and *nis-s* COIs of the last hour leads to a 1.03 and 0.67 increase in Sharpe Ratios, significantly enhancing

36

the portfolio profits. Conversely, the last half-hour of non-cross-isolated COI is not a good signal for predicting future returns. For the *nis-b* trades, the future returns are only predicted by the COI during less volatile trading hours.

C COI Measured by Volumes

Instead of considering the number of trades, in this section we analyze volume order imbalances defined as

$$COI_{i,t}^{type} = \frac{V_{i,t}^{type,buy} - V_{i,t}^{type,sell}}{V_{i,t}^{type,buy} + V_{i,t}^{type,sell}}, \tag{11}$$

where $V_{i,t}^{type,buy}$ and $V_{i,t}^{type,sell}$ denote the total volume of market buy orders and market sell orders of stock i on day t. We repeat the analysis on volume imbalances and present the results in Table 13.

Table 13: COIs measured by volume.

We calculate COIs measured by volumes, as Equation (11), from 2017-01-03 to 2019-12-10 for the selected 457 stocks. Panel A summarizes the results of 457 regressions using Equation (6). Average coefficient denotes the mean of all regressions coefficients. 'Percentage' denotes proportion of selected stocks. 'Significant' denotes statistically significant at 5% significance level using two-tailed t test. 'Average R^2' denotes the regression R^2 averaged over all stocks. Panel B shows the annualized Sharpe Ratios of the long-short portfolios sorted on COIs indicated by the corresponding row indices and column names. The on- and off- diagonal values are for single- and double-sort portfolios respectively. The annualized Sharpe Ratios over the sample period are given by Equation (10).

Panel A: Time series regression					
	Average coefficient	Percentage positive	Percentage positive and significant	Percentage negative and significant	Average $R^2(\%)$
all	2.14	100.00	97.59	0.00	6.27
iso	1.71	100.00	97.37	0.00	6.16
nis	1.52	98.25	91.68	0.00	3.97
nis-s	1.07	92.78	77.24	0.66	3.03
nis-c	0.58	93.43	65.65	0.66	1.22
nis-b	0.72	97.16	81.18	0.66	2.09

Panel B: Annualized Sharpe ratios					
	iso	nis	nis-s	nis-c	nis-b
iso	2.11	1.83	2.35	2.17	1.96
nis		-0.13	0.99	0.88	0.93
nis-s			1.39	1.79	1.22
nis-c				1.42	1.36
nis-b					1.18