Analysis and modeling of client order flow in limit order markets

Rama Cont[1], Mihai Cucuringu[1,2,3], Vacslav Glukhov[4], and Felix Prenzel*[1,4]

[1] *Mathematical Institute, University of Oxford*
[2] *Department of Statistics, University of Oxford*
[3] *The Alan Turing Institute*
[4] *JP Morgan*†

December 30, 2021

Abstract

Orders in major electronic stock markets are executed through centralised limit order books (LOBs). The availability of historical data have led to extensive research modelling LOBs. Better understanding the dynamics of LOBs and building simulators as a framework for controlled experiments, when testing trading algorithms or execution strategies are among the aims in this area. Most work in the literature models the aggregate view of the limit order book, which focuses on the volume of orders at a given price level using a point process. In addition to this aggregate view, brokers and exchanges also have information on the identity of the agents submitting the order to them. This leads to a more complicated representation of limit order book dynamics, which we attempt to model using a heterogeneous model of order flow.

We present a granular representation of the limit order book, that allows to account for the origins of different orders. Using client order flow from a large broker, we analyze the properties of variables in this representation. The heterogeneity of the order flow is modeled by segmenting clients into different *clusters*, for which we identify representative prototypes. This segmentation appears to be stable both over time, as well as over different stocks. Our findings can be leveraged to build more realistic order flow models that account for the diversity of market participants.

1 Introduction

Limit order books (LOBs) are the data structures that record the outstanding limit orders on an exchange, where different agents interact buying and selling a certain asset. This continuous buy and sell procedure leads to the asset's price formation. The different agents participating in the market follow different investment objectives and strategies when making trading decisions, which eventually end up as orders of different types sent to a LOB. Orders are sent to the exchange either through execution services or directly by the agents themselves. In particular, high-frequency traders (HFTs) and market makers (MMs) tend to have proprietary access to exchanges. Intuitively, this brings up the question about the heterogeneity of the order flow in LOBs. Yet, most LOB models based on stochastic processes like [23, 10, 2] assume homogeneous order flow and do not account for the potential differences between different agents with regard to their behaviour in LOBs.

The reason for this widely employed homogeneous modelling lies in the limited access to private data. In most settings, public data is being used, which, in the best case scenarios, shows orders on a tick level (i.e. order by order), such as the LOBSTER database[1]. This does not allow to account for the origin of a particular order, e.g. which agent (or what type of agent) submitted the order, and whether a limit order is part of a larger parent order which is being executed throughout a specified time horizon.

Figure 1 presents a view of this heterogeneous market ecology, which cannot be observed in public data. In first instance, one might separate traders into two groups – those trading with a proprietary access to exchanges, and those trading through execution services/brokers. The first group primarily contains high frequency traders and

*Corresponding author: prenzel@maths.ox.ac.uk
†Opinions expressed in this paper are those of the authors, and do not necessarily reflect the view of JP Morgan.
[1] https://lobsterdata.com/

Electronic copy available at: https://ssrn.com/abstract=3997109

market makers, generally trading with proprietary access to the exchange. The remainder are traders which generally trade through execution services and are not in full control of the actual placement of limit orders. In contrast, they send *parent orders* (also called *meta orders*) to a broker. The broker then uses different algorithms to slice the parent order into different child orders, which are then sent as *limit* or *market* orders to the exchange. Studies using public LOB data only see what is being sent to exchanges at the end of the order process (right hand side of Figure 1).

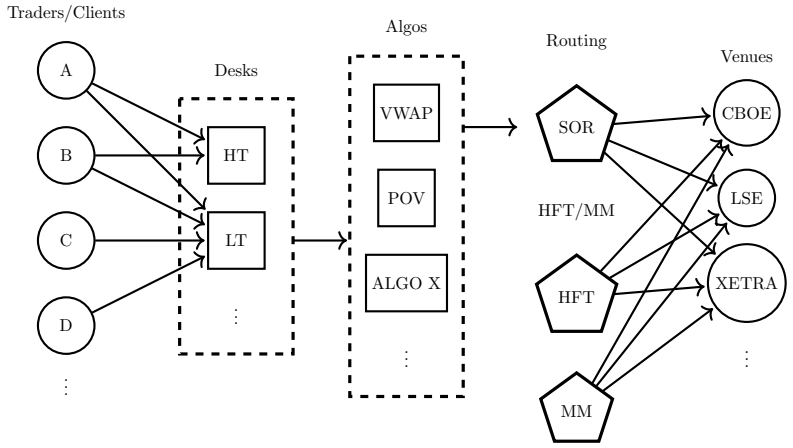


Figure 1: Flow chart depicting the process of an order from a client sent to a broker. Clients submit "parent orders" to execution services and different trading desks. For example, these could be high tough (HT) and low touch (LT) desks. At steps "Desks" and "Algos", the parent orders are sliced into child orders, and sent to different venues for cost-efficient execution. Usually, a routing system (SOR) determines to which venue a child order is sent.

This work contributes along two dimensions to the literature. Firstly, it introduces a new notation for LOBs which accounts for different agents submitting orders to LOBs. This notation allows for more detailed views on LOBs; in addition, common views such as the *public view* (queue size) can be derived from it. Second, one particular view on the LOB, the *broker view*, is analyzed in detail. We use trade execution data to segment traders into representative groups with similar attributes. These are then analyzed for stability in both the cross-sectional (i.e, cross-asset) and temporal dimensions. Following, the heterogeneity of the aggregated order flow which is induced by different trader types is investigated.

We find traders can be segmented into four different components, namely the quant, VWAP, signal and residual order flows. The different groups are stable both over time as well as over different instruments. This also holds for the traders themselves. Heterogeneity in both the agents' structure and the order flow they generate in LOBs allow the development of heterogeneous order flow models. In particular, we suggest a model for each agent types, based on the insights derived from the segmentation analysis which capture the most important characteristics of each component despite their simplicity.

This section concludes with a review of related work in the area. In Section 2, an alternative, more detailed notation for LOBs accounting for different origins of orders is introduced. The notation allows for different views on the LOB. Section 3 presents the data and pre-processing. Afterwards, the set of traders is segmented the representative agent types explained in detail and the segmentations' stability is shown. Section 4 presents results on the properties of the accumulated order flows for each agent type. A simple model for parent order flow capturing the main heterogeneity between the different agent types is presented in Section 5. Section 6 summarizes the results and present future research directions.

1.1 Related work

The existing literature covering LOBs is vast. This is particularly true for publicly available data. Comprehensive introductions to LOBs are given in [1, 11]. and among others [9, 5] analyse properties of the LOBs in general data.

Modelling, simulation and prediction have been the subject of a broad variety of studies in the literature. Most commonly known in literature, [23, 10] model the LOB via homogeneous Poisson processes. Hawkes processes – another type of point process – have been studied, for example, in [2], analyzing long-term properties of LOBs

Electronic copy available at: https://ssrn.com/abstract=3997109

driven by such self-exciting point processes. [22, 26] use deep learning to predict the mid price moves over some time interval. Several studies exist that aim to model LOBs with agent based models, e.g. [8, 25], but usually these models struggle with calibration since it is not clear how to calibrate the single agents.

The literature on market ecology is more scarce. The reason is partially data privacy and general access to non-public data. Few studies are available where researchers had access to non-public order book data, which were mainly used to analyse the effect of the newly arisen HFTs and MMs.

Two of the studies with non-public data have been presented by Brogaard et al. in [6, 7]. They study LOB data in which orders from HFTs and MMs are flagged, in order to analyse the impact of HFT and MM accounts on market quality and price discovery. The studies find evidence that HFTs increase market quality by potentially dampening intraday volatility among other properties. Furthermore, they argue that the trade directions of HFTs are based on public information such as "macro news announcements, market-wide price movements, and limit order book imbalances" [7]. [12] analyse LOB data with information about orders from HFTs and MMs, specifically outlining the differences between these two types of market participants in the way they tend to trade. They find MMs to take the majority of limit order traffic and to hold lower inventories compared to HFTs. [13] attempt to extract HFT trades by identifying the so-called "strategy runs", i.e., periods with similar inter-arrival times and order sizes, which is unique for HFTs, as the authors argue. Their suggested measure of low latency indicates that with increasing low latency trading, spreads decrease and the depth at the first level increases.

The richest analysis in the literature is [15]. The authors analyse audit trail data on transaction level from the S&P 500 eMini Future during 4 days, including the flash crash from May 2010. In particular, they separate the market into high frequency traders, market makers, fundamental buyers, fundamental sellers and opportunistic traders, as previously derived in [17]. The classification is done both based on transaction volume and scaled net positions. Despite having access to a very granular data set, the analysis is focused on high-frequency traders and market makers during the flash crash, and less on the properties of the remaining market participants.

Our present work aims to fill in the above gaps, by focusing on the remainder of the market participants, namely traders relying on execution services for their execution. In particular, we seek to analyse such traders which do not have direct market access (DMA), but rather trade through brokers/execution services.

2 Limit Order Book as a queuing system with different agents

To account for the origin of any order, we consider a finite set of agents A, representing different traders or agent types acting together in a limit order book. An agent is denoted as α. Each agent $\alpha \in A$ generates a flow of orders which affect the state of the LOB.

Definition 2.1 (Limit Order (LO))**.** A limit order $x = (t, p, q, \alpha)$ is characterized by

- an arrival time $t \in \mathbb{R}^+$,
- a price $p \in \delta\mathbb{N}$ which is a multiple of the price tick $\delta > 0$,
- a quantity $q \in \mathbb{Z} \setminus \{0\}$ with $q > 0$ denoting buy orders and $q < 0$ denoting sell orders,
- the identity $\alpha \in A$ of the agent submitting the order.

A limit order is considered "outstanding" as long as it has neither been cancelled nor fully executed.

Furthermore, we denote the unit point mass at x by ϵ_x and by

$$\mathcal{M}_+(\mathbb{R}_+ \times \delta\mathbb{N} \times A),$$

the space of positive measures on $\mathbb{R}_+ \times \delta\mathbb{N} \times A$, and by

$$\mathcal{M}(\mathbb{R}_+ \times \delta\mathbb{N} \times A),$$

the (vector) space of signed measures on $\mathbb{R}_+ \times \delta\mathbb{N} \times A$. With this in mind, we may represent the collection of outstanding orders as a signed measure

$$\begin{aligned} &\mu : \mathbb{R}_+ \times \delta\mathbb{N} \times A \to \mathbb{Z}, \\ &([0,T], \{p\}, \{\alpha\}) \mapsto \mu([0,T], \{p\}, \{\alpha\}), \end{aligned} \tag{1}$$

Electronic copy available at: https://ssrn.com/abstract=3997109

where $\mu([0,T] \times \{p\} \times \{\alpha\})$ represents the (net) volume of orders submitted by agent α at price p, between time 0 and T. The measure μ is an element of $\mathcal{M}(\mathbb{R}_+ \times \delta\mathbb{N} \times A)$, the (vector) space of signed measures on $\mathbb{R}_+ \times \delta\mathbb{N} \times A$, and its Jordan decomposition

$$\mu = \mu^+ - \mu^-,$$

corresponds to the distribution of outstanding sell and buy orders. The signed measure defined in Equation (1) allows to describe different views on the limit order book. The *omniscient view* and *public view* form extreme cases, while the third one, the *broker view*, builds a mixture between the first two.

2.1 Public View

The *public view* (i.e. anonymized view) on the limit order book corresponds to the information available to market participants who observe the volume of orders at each price. However, it does not contain any information regarding the origin nor the submission times of the single orders. Hence, most market participants only observe the *queue size* Q_p at each price level p

Definition 2.2 (Queue Size)**.** The queue size for any price $p \in \delta\mathbb{N}$ is denoted by

$$Q_p = \sum_{\alpha \in A} \mu([0,t], \{p\}, \alpha). \tag{2}$$

We call $Q : \delta\mathbb{N} \to \mathbb{Z}$ the anonymized limit order book, $Q \in \mathcal{M}(\delta\mathbb{N})$ and belongs to the state space $\mathbb{Z}^{\delta\mathbb{N}}$.

The *public view* is visualised in Figure 2a and corresponds to the sum of all orders' quantity at a particular level. The agent neither observes the number of orders nor the color (i.e. the agent id).

2.2 Omniscient View

The *omniscient view* of the limit order book is the collection of all outstanding limit orders, including the information about their time of submission and the identity of the submitter. This information is represented by the measure μ. E.g. in Figure 2c, the *omniscient view* not only contains the single orders with prices and quantities, but also the color (i.e. the agent id).

For example, the measure of the outstanding orders from agent $\alpha \in A$ is given by

$$\mu(\cdot, \cdot, \{\alpha\}) \in \mathcal{M}_+(\mathbb{R}_+ \times \delta\mathbb{N}). \tag{3}$$

While the *omniscient view* exists usually no one has access to it. The closest to the *omniscient view* is the view, which the corresponding exchange has on the order book because it has some origin about all the flow in the LOB. However, in many cases the exchange is not able to distinguish between the flow of different clients from a broker. Thus, the exchange sees all orders of a particular broker as one aggregated flow.

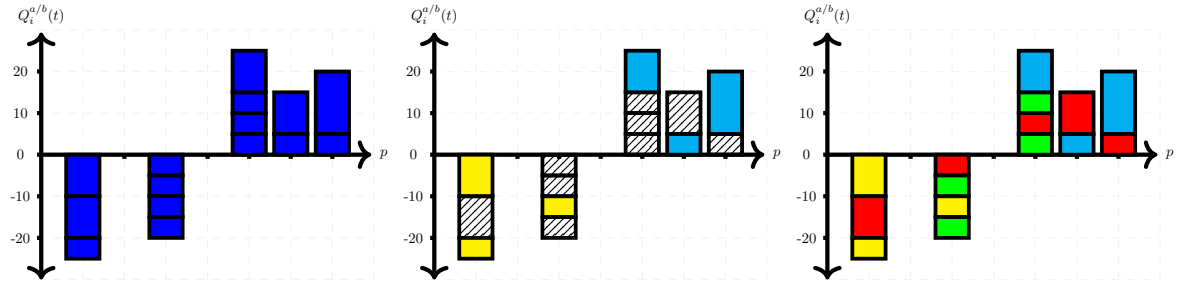
2.3 Broker View

The *omniscient view* and the anonymized limit order book represent two extreme cases of information on the limit order book. However, there are intermediate situations corresponding to partial information on the limit order book. An important case is the case of a broker who can observe the order submissions times and identities for a subset $B \subset A$ of agents. This broker will have a less detailed view of the limit order book, which corresponds to aggregating over all agents not in B. Denote by $B^* = B \cup \{\Delta\}$ the set obtained by adding one element to B; this element will represent all other agents not included in the broker's set of clients. The *broker view* of the limit order book may then be described by a measure $\mu_B \in \mathcal{M}(\mathbb{R}_+ \times \delta\mathbb{N} \times B^*)$ defined by

$$\mu_B(\cdot, \cdot, \{\alpha\}) = \mu(\cdot, \cdot, \{\alpha\}) \qquad \alpha \in B, \tag{4}$$

$$\mu_B(\cdot, \cdot, \{\Delta\}) = \sum_{\alpha \notin B} \mu(\cdot, \cdot, \{\alpha\}). \tag{5}$$

Electronic copy available at: https://ssrn.com/abstract=3997109



(a) Public view (b) Broker view (c) Omniscient view

Figure 2: Different views of the same LOB snapshot. Note, only the *omniscient view* has complete information about the queue priority of each order.

An exemplary *broker view* snapshot is visualised in Figure 2b. Yellow and light blue correspond to agents $\alpha \in B$, thus are from the broker's clients. These orders are observed with their particular ID and viewed via $\mu_B(.,.,\{\alpha\})$, while green and red orders would correspond to "all other agents" (Δ) which are not in B, the broker's set of clients. These orders are seen as $\mu_B(.,.,\{\Delta\})$. Furthermore, the broker does not observe the specific time as (5) suggests, but rather the relative time since the broker knows its queue position. Hence, the broker can (only) know that an order has been submitted between two orders of its own clients.

Most literature aims to model the *public view* of the LOB. These models have been shown to be useful for describing certain dynamics of limit order books. These dynamics include general price moves or distribution of queue sizes. The detailed queue and a specific order's queue position, however, is generally not available in both public data and the majority of the models in the literature. This has led to research aiming to estimate the queue position of an order such as [19]. It turns out, "queueing effects can be very significant"[19] and accounting for the queue position should not be ignored when designing trading algorithms. In accordance to this, keeping track of the queue size rather than the entire collection of orders does not allow cancellations to refer to particular orders. This makes keeping track of how orders flow through the queue impossible. Furthermore, models in the literature do not distinguish between different agents submitting these orders. However, different agents generally trade with different expectations and intentions. This is likely to be reflected in the way they place orders (close to the best prices or deeper in the book), as well as how they tend to cancel their orders. Instead, a LOB model which accounts for this heterogeneity allows to do this. We remark that the topic of identifying determinants of limit orders cancellation has been extensively studied recently, and would be an interesting research direction to explore, in itself.

As Figure 1 visualizes, limit orders sent through a broker part of a larger parent order. Since these parent orders are central in the sequel of the paper, we formally introduce the structure of a parent order.

Definition 2.3 (Parent order)**.** A parent order P is defined by a number of variables some of which are determined by the submitting trader, others are only known at the end of the execution.

- a submission or arrival time $t \in \mathbb{R}^+$,
- a target quantity $q^{target} \in \mathbb{Z} \setminus \{0\}$, the total quantity which shall be executed in the market,
- an executed quantity $q^{exec} \in \mathbb{Z}$ which corresponds to the total sum of executed child orders.

The parent order's target quantity q^{target} can be decomposed in its

1. absolute quantity $\tilde{q} = |q^{target}| \in \mathbb{N}$,
2. sign (buy or sell order)

$$\text{sign}(q^{target}) = \begin{cases} -1, & \text{if } q^{target} < 0 \\ 1, & \text{if } q^{target} > 0. \end{cases} \tag{6}$$

Each order, depending on its size and preferences, has an order placement schedule, which we denote by

$$\mathcal{X} = \{x_1, \dots, x_N\}, \quad x_i = (t_i, p_i, q_i, \alpha_i) \; \forall \; x_i \in \mathcal{X}, \tag{7}$$

Electronic copy available at: https://ssrn.com/abstract=3997109

where each element has the form of a LO as defined in Definition 2.1. The schedule contains all LOs placed in the LOB during to the parent order execution. Hence, $\alpha_i = \alpha$ since all limit orders are posted by the same agent. Note, $\mathcal{X}$ may contain effective market orders as executable limit orders. In other words, a buy limit order ($q_i < 0$ with price $p_i = \infty$) would correspond to a buy market order ($q_i > 0$ and $p_i = 0$ to a sell market order respectively).

Additionally, there is an execution schedule

$$\mathcal{X}^{exec} = \{x_1, \dots, x_N\}, \quad x_i = (t_i, p_i, q_i, \alpha_i) \; \forall \; x_i \in \mathcal{X}^{exec}, \tag{8}$$

which is comprised of the record of all executed limit orders for parent order P. This may be

1. a market order x with $p \in \{0, \infty\}$ sent within the order schedule eq. (7); in this case, $x \in \mathcal{X}$.
2. a market order x sent by any other agent which executes an outstanding order from $\mathcal{X}$; in this case, $x \in \mathcal{X}^{exec}$ but $x \notin \mathcal{X}$.

Any $x \in X^{exec}$ may not be the complete market order but only the fraction which affected a limit order in $\mathcal{X}$. For instance, if an incoming market order executes two limit orders from $\mathcal{X}$ currently in the limit order book, there are two entries in the data set. In contrast to $\mathcal{X}$, the execution schedule $\mathcal{X}^{exec}$ may contain orders from different agents. Note, (7) and (8) are generally determined during the execution of the parent order.

A full description of the market, the *omniscient view*, is not always available. The *broker view*, however, offers a partially more detailed view on the market, in particular of the broker's set of agents B. Thus, we want to shed some light into what B consists of to better understand the limit order market ecosystem. In case, the ecosystem can be separated into several groups of order flows, modelling the order flow of each group is a much more feasible and tractable task in comparison to modelling every single agent $\alpha \in B$. Thus, the remainder of this paper analyses the order flow of a broker, and separates a typical set of agents using execution services into different representative groups. These show homogeneous behaviour within their group but differ substantially across different groups.

3 Segmentation of agent types

This section considers the task of segmenting the population of traders that send orders to the brokers. Traders sending orders builds the very left hand side of Figure 1. Our analysis of this process stands in contrast to known works in the literature, since most studies only use anonymized order flow data, as outlined in Section 1.1. Few studies use non-public LOB data, and mainly [15] has access to the trading accounts. To the best of our knowledge, our study is the first one to provide an analysis of parent orders arriving at the broker. Consequently, the available data structure not only allows us to observe what arrives in the LOB and who sends it, but also whether several limit orders come from the same parent order. Uncovering latent similarities across traders and identifying different trader typologies can potentially facilitate better modelling of the parent order flow in LOBs. The main implication of our findings is that one can move beyond modeling each agent or trader independently, and pave the way for modelling each homogeneous group or cluster of agents.

3.1 Data set and Features

For the analysis, we use anonymized trade execution data from a large broker. This data builds a detailed view on a subset of the entire LOB as explained in Section 2. The universe is set to stocks from *STOXX 600* and buckets of 1-month duration will be employed in the segmentation. In other words, for each asset i and time period t, the set of traders that execute their trades via the broker is denoted as $B_{i,t}$ where each agent $\alpha \in B_{i,t}$ has sent at least one order to the broker which led to an execution. Clearly, $B_{i,t} \subseteq A_{i,t}$, the whole set of agents active in a LOB of the given ticker. Instruments are primarily analysed separately; joint analysis is used primarily for matters of comparison and stability, in particular in Section 3.4.

Generally speaking, the broker manages the execution and the child orders which are sent to the LOB. For the analysis, we thus rely on the parent order structure and the corresponding statistics for each agent. Features regarding the single child orders are not included. This means: when do agents send orders, how large are

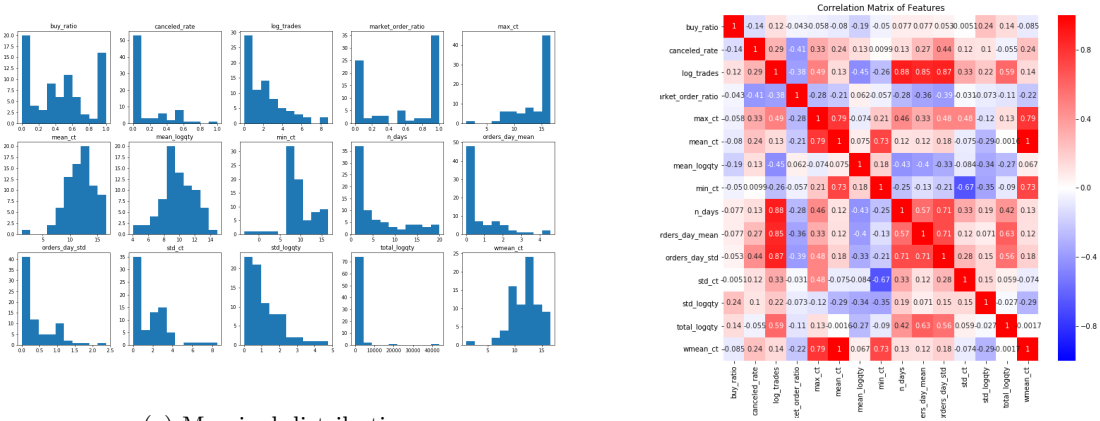
Electronic copy available at: https://ssrn.com/abstract=3997109

they, and how aggressive shall they be executed? The corresponding data fields are primarily *side* (buy/sell), *number of orders, time of submission* and *size*. The resulting features/statistics of this information include, for example, the average direction of an agent's trades, the number of orders an agent has submitted, distributional properties of the day time an agent submits orders, or an agent's order sizes' standard deviation. Information regarding external information or market conditions such as momentum, volatility, etc. are also used. A detailed description of each feature can be found in Table 14.

Altogether, this amounts to a data set of $n = |B_{i,t}|$ agents with p features

$$X \in \mathbb{R}^{n \times p} \text{ with } x_\alpha = (x_{\alpha,1}, \ldots, x_{\alpha,p}) \in \mathbb{R}^p \; \forall \; \alpha \in B_{i,t}$$

describing $B_{i,t}$. Each vector x_α describes the parent order structure of agent α in instrument i during period t.



(a) Marginal distribution

(b) Pearson correlation matrix

Figure 3: Features of data.

Figure 3 shows the distribution and correlation of computed features for one exemplary $B_{i,t}$ from the *STOXX 600*. Many features, such as the ratio to which a particular agent specifies a maximum price for execution (minimum price for sell orders respectively) – *market order ratio* – tend to be rather 0 or 1. I.e., most agents either always or never indicate a limit for executions of child orders. Other features are very skewed, for example *n_days*, the number of days an agent sends an order. This is because most agents only send very few orders during one month for a given ticker.

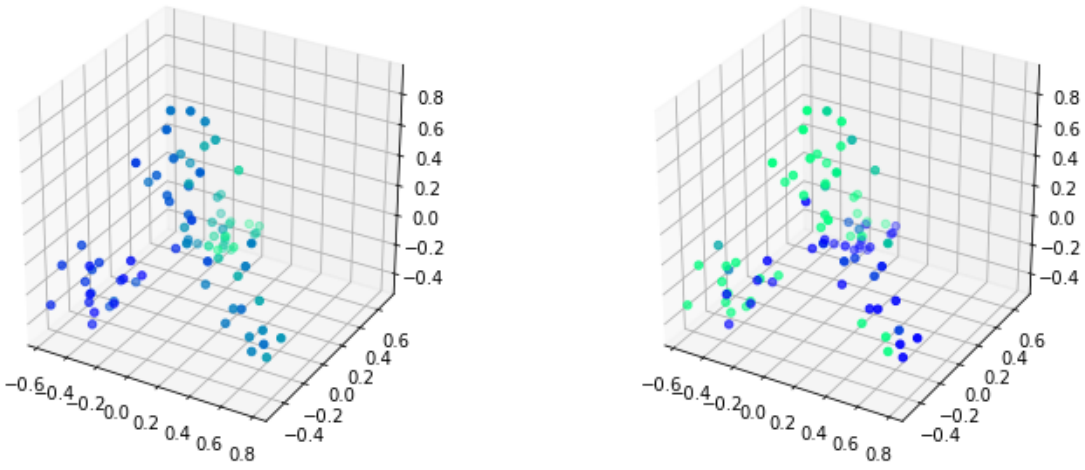
To visualize agents in a lower-dimensional space, methods such as principal component analysis (PCA) [14] or spectral embedding [4] may be applied to obtain further insights into the structure of the agents interacting with financial brokers. To this end, Figure 4 shows two exemplary features in the embedding space. The mean creation time of a client's orders in Figure 4a and the ratio indicating the fraction of orders of a particular trader which contains a maximum price for the execution (minimum price for sell orders respectively) in Figure 4b. For both features in Figure 4 there are areas of the embedding where traders with similar values of the corresponding features are clustered. For example, there exists an area of traders which have much higher mean creation time compared to that of other clusters. In Figure 4b, a group of traders can be encountered which tends to specify a limit price when sending orders for executions. Additionally, the fact that the point cloud in the embedding is not just a sphere indicates some structure in the underlying data which can be further exploited.

3.2 Spectral Clustering

As outlined in Section 2, a broker has a detailed view on a subset of the market, i.e. knows the trader identity and the specific time of an order, for all agents $\alpha \in B_{i,t}$. For ease of notation, we will refer to $B_{i,t}$ as B in the sequel. To better understand the structure of a typical set of traders which use a broker, this section segments B into a partition $\mathcal{C} = \{C_1, \ldots, C_K\}$, in order to obtain representative agent types that best describe the structure of a broker's clients.

Clustering algorithms are designed to do exactly what we want to do. They separate the observations into different, typically unknown, classes or clusters such that observations within the same cluster are more similar

Electronic copy available at: https://ssrn.com/abstract=3997109



(a) Mean creation time

(b) Ratio of specified limit price

Figure 4: Exemplary embedding using first three eigenvectors of the normalized Laplacian matrix using the spectral embedding method from [4]. The color of the nodes indicates the values of the associated feature. The brighter the points, the higher the value of the corresponding feature.

to each other, while observations that end up in different clusters are rather different or dissimilar to each other. A partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ is sought after, such that

1. $C_k \subseteq B$,
2. $C_k \neq \emptyset \; \forall \; k \in \{1, \ldots, K\}$,
3. $C_k \cap C'_k = \emptyset \; \forall \; k \neq k'$,

and furthermore is optimal with respect to some objective. In our case, this means every trader would belong to a certain type. These types should be heterogeneous, but traders of the same type should rather form a homogeneous population with similar characteristics. Referring to Figure 1, we would like to shrink down the number of nodes of many individual traders submitting parent orders on the left hand side to just a few nodes representing each trader type. K indicates the number of types the agents shall be separated into. $\bar{x}_1, \ldots, \bar{x}_K \in \mathbb{R}^p$ are the corresponding cluster centers indicating the average features (i.e. coordinates) of the of the data points (i.e. agents) affiliated with the corresponding cluster, so $\bar{x}_{k,j} = \frac{1}{|C_k|} \sum_{\alpha \in C_k} x_{\alpha,j} \; \forall \; j \in \{1, \ldots, p\}$. Essential to each clustering algorithm is a distance matrix $W \in \mathbb{R}^{n \times n}$ which contains the distance $W_{\alpha,\alpha'}$ between observation x_α and $x_{\alpha'}$, for some distance measure d. This distance is often the Euclidean distance.

The spectral clustering technique used in this study combines two methods, spectral embedding and the K-means clustering algorithm [21, 20, 18]. In other words, it creates the partition via an iterative ascent algorithm on a p'-dimensional, non-linear embedding of the data set, describing the agents' trading behaviour.

1. **Construction of the adjacency graph**

   The adjacency graph indicates for each pair of observations whether these are connected or not. In particular,

   $$A \in \mathbb{R}^{n \times n} \text{ where } A_{i,j} = \begin{cases} 1 \text{ if } \|x_i - x_j\|^2 < \epsilon \\ 0 \text{ else,} \end{cases} \qquad \forall \; i, j \in \{1, \ldots, n\}. \tag{9}$$

   Two vertices are connected by an edge if their Euclidean distance in the actual space $\mathbb{R}^p$ is below a certain threshold ϵ. Alternatively, one may connect a vertex i to its l nearest neighbors where $l \leq n, l \in \mathbb{N}$.

2. **Weighting the similarities**

Electronic copy available at: https://ssrn.com/abstract=3997109

This step weighs the connections between observation i (row) and j (column). The common choice is the heat/rbf kernel. In this case, the matrix $W \in \mathbb{R}^{n \times n}$ is computed with

$$W_{i,j} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\sigma}} \text{ if } A_{i,j} = 1 \\ 0 \text{ else} \end{cases} \quad \forall\, i, j \in \{1, \dots, n\}, \tag{10}$$

for some user-tuned bandwidth parameter $\sigma \in \mathbb{R}_+$. The closer (i.e. more similar) the observations x_i and x_j are in Euclidean distance, the higher the value W_{ij}. Alternatively, whenever the input matrix is binary, with $A_{i,j} = 1$ for connected vertices, we set $W_{i,j} = 1$ if $A_{i,j} = 1$.

3. **Compute eigenmaps via first p' eigenvectors**

   Next, the generalised eigenvector problem

   $$Lf = \lambda D f, \tag{11}$$

   is solved. $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the row-sums of W, i.e. $D_{i,i} = \sum_j W_{i,j}$. We define $L = D - W$ to be the unnormalized Laplacian matrix (also referred to as the Combinatorial Laplacian).

   The solution of eq. (11) is a set of (sorted) eigenvalues $\lambda = (\lambda_0, \dots, \lambda_{n-1})$ for which $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$ holds. For the corresponding eigenvectors $f = (f_0, \dots, f_{n-1})$ from eq. (11) $\lambda_0 = 0$ and $f_0 = (1, \dots, 1) \in \mathbb{R}^n$ holds [24]. The multiplicity of the $\lambda_0 = 0$ indicates the number of connected components in the graph. Finally, the eigenvectors $f_1, \dots f_{p'}$ are then used for the p'-dimensional embedding and $y_i = (f_1(i), \dots, f_{p'}(i)) \in \mathbb{R}^{p'}$.

4. **Clustering the low-dimensional embedding**

   Finally, the agents are clustered in this low-dimensional embedding via the K-means algorithm [16]. The objective function optimized by the algorithm is given by

   $$\min_{\mathcal{C}, \{\bar{y}_k\}_k} \sum_{k=1}^{K} |C_k| \sum_{\alpha \in C_k} \|y_\alpha - \bar{y}_k\|^2, \tag{12}$$

   where $y_\alpha, \bar{y}_k \in \mathbb{R}^{p'}$, the space of the embedding. The objective function (12) corresponds to the minimization of the variance within the clusters with respect to partition $\mathcal{C}$. Optimization is done via iterative descent, alternating between recomputing cluster centers $\bar{y}_k$ and reassigning the observations y_α to the nearest cluster center[2].

In contrast to K-Means, spectral clustering is a non-linear clustering method. This enables the algorithm to potentially uncover clusters which are not convex [24] since the clustering is not performed in the original feature space $\mathbb{R}^p$, but rather in the embedding space $\mathbb{R}^{p'}$. This capability of uncovering non-linear relationships makes spectral clustering more flexible in comparison to K-Means.

The features prepared in Section 3.1 are now used to cluster the agents. Several clustering algorithms, such as K-means and spectral clustering, are used to gain insights into a potential segmentation of the agents' structure. For this, the representative agent types, or centroids, are analyzed.

Extracting centroids when using spectral clustering is not as straight forward as for K-Means. This is because the embedding procedure described above is not invertible for any arbitrary point. Hence, a point $y \in \mathbb{R}^k$ cannot be mapped into the actual feature space $\mathbb{R}^p$. To overcome this, one can use the center of a cluster's observations in the actual space from the data $X \in \mathbb{R}^{n \times p}$, i.e. $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \;\forall\; k \in \{1, \dots, K\}$. Alternatively, the prototype for cluster k can be defined as $x_i \in \mathbb{R}^p$ with $i = \arg\min_{i \in C_k} \|y_i - \bar{y}_k\|^2$. In other words, the observation x_i is the one whose embedding y_i is the closest to the cluster center $\bar{y}_k$ in the embedding space. In this study, the first method is used to obtain cluster centers, as it is less prone to outliers for some feature of the prototype x_i. The cluster centers are then used to give details about properties and provide a comparison of the different agent types.

Several algorithms were used to cluster $B_{i,t}$ and find an optimal partition $\mathcal{C}$. In the following, we particularly outline observations and statistics when comparing the partitions between K-means and spectral clustering in more detail for one exemplary $B_{i,t}$. The number of clusters K ranges from 2 to 5. The number of agents for the ticker is ~ 80. Some observations may be made:

[2] See [14] for the detailed algorithm.

Electronic copy available at: https://ssrn.com/abstract=3997109

1. **Cluster sizes:** Generally, the number of observations in the clusters is neither very large nor very small. There is no cluster with only very few observations.

2. **Consistency:** The partitions using K-means and spectral clustering are very similar. This can quantified by the adjusted rand index (ARI) which indicates the consistency across two partitions. Let $\mathcal{C}^{(1)} = \{C_1^{(1)}, \dots C_K^{(1)}\}$ and $\mathcal{C}^{(2)} = \{C_1^{(2)}, \dots C_K^{(2)}\}$ be two partitions. The ARI reads

$$ARI = \frac{\sum_{k,k'} \binom{n_{k,k'}}{2} - \left[\sum_k \binom{a_k}{2} \sum_{k'} \binom{b_{k'}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_k \binom{a_k}{2} + \sum_{k'} \binom{b_{k'}}{2}\right] - \left[\sum_k \binom{a_k}{2} \sum_{k'} \binom{b_{k'}}{2}\right] / \binom{n}{2}}, \tag{13}$$

   where $n_{k,k'} = |C_k^{(1)} \cap C_{k'}^{(2)}|$ denotes the number of observations inside $C_k^{(1)}$ and $C_{k'}^{(2)}$ and $a_k = \sum_j n_{k,k'} = |C_k^{(1)}|$ ($b_{k'} = \sum_k n_{k,k'} = |C_{k'}^{(2)}|$). The index takes a maximum value of 1 if $\mathcal{C}^{(1)} = \mathcal{C}^{(2)}$.
   Table 1 shows the ARI for different values of K. A particularly high consistency is indicated for 2 and 4 clusters. Noteworthy is the high consistency taking into account that K-means is performed on the actual feature space ($\mathbb{R}^p$), while the spectral clustering is based on the non-linear embedding in $\mathbb{R}^{p'}, p' \ll p$. In other words, regardless of the feature space employed, the recovered partitions are very similar.

# Clusters	2	3	4	5
ARI	**0.7979**	0.5107	**0.7968**	0.6153

Table 1: Table indicating the ARI between K-means and spectral clustering. The higher the number, the higher the consistency between to partitions; the maximum value ARI can attain is 1, indicating a perfect matching of the clusters.

3. **Stability of clusters:** Increasing K leads to a sequential splitting of the data cloud into clusters. E.g. one cluster gets further split when increasing K by one unit. The other clusters and their affiliated traders remain stable.

4. **Number of clusters:** Setting $K = 2$ or $K = 4$ leads to the best scores. First of all, the consistency is better than for 3 and 5 clusters indicated by higher ARIs in Table 1. Lastly, Table 2 shows the variance between agents and their corresponding cluster center. In particular, the more the variance decreases, the more justified is the addition of another cluster. The second differences of the variance within the clusters can indicate how the variance reduction of an additional new cluster changes. This helps to identify a K, for which an increase leads to a much lower variance reduction. In particular, values for K would be either 2 or 4, which goes in line with the other metrics.

# Clusters	1	2	3	4	5	6	7	8	9
Variances	3.914	3.507	3.210	2.981	2.871	2.738	2.640	2.596	2.536
2nd Diffs.	-	**0.108**	0.069	**0.118**	-0.021	0.034	0.053	-0.016	-

Table 2: The variance between an observation and its corresponding cluster. The stronger the decrease when K is increased by one cluster, the larger is the marginal effect of the new cluster to separate the data.

3.3 Representative agent types

Despite the final goal being the heterogeneity of the aggregated order flow caused by different clusters, the cluster centers $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$ are computed to interpret the agent types. Such resulting cluster centers for a subset of the features used are shown in Table 3, for one exemplary $B_{i,t}$.

In fact, as shown in Table 4, the four agent types may be summarised as follows:

1. **Quantitative agents** (C-Quant): The most distinguishable agent type is cluster 1, which contains those agents that submit many trades within the period, with a smaller trade size. Despite trading rather small amounts, the total volume traded is by far the highest in this cluster. Also, the number of days during which the cluster's agents trade is 12 days, which is several times higher than the second highest

Electronic copy available at: https://ssrn.com/abstract=3997109

	1: C-Quant	2: C-Day VWAP	3: C-Signal	4: C-Res
Buy ratio	0.63	**0.55**	0.63	**0.65**
Cancellation ratio	**0.26**	**0**	0.15	0.07
# Trades per month	**219.7**	**1.72**	4.08	5.03
Maximum order creation time	**16:10:47**	**08:32:04**	15:02:03	14:34:05
Mean order creation time	12:03:50	**08:25:34**	**14:20:00**	11:31:56
Mean order size (in ADV)	**0.01**	0.03	**0.06**	**0.06**
Mean momentum (bps)	2.285	**-17.87**	29.94	**38.26**
Mean volatility	22.2	23.03	22.42	22
Minimum order creation time	**07:58:08**	08:19:10	**13:34:31**	08:40:12
# Active days per month	**13.47**	**2.1**	3.71	4.95
St. dev. of order creation time	02:24:49	**00:07:12**	00:42:16	**02:40:46**

Table 3: Exemplary centroids setting $K = 4$ for one of the 25 most liquid STOXX 600 instruments during December 2019. Liquidity in this case refers to the total number of trades in the data base. The highest (lowest) value of the corresponding features are highlighted in blue (red).

Cluster Name	Cluster Description
C-Quant	Mainly quantitative traders, many trades, small volumes, orders sent throughout the day, execution with few child orders leading to a large POV, net inventory closer to 0
C-Day VWAP	Mostly VWAP as execution, few orders, only sent in the morning, almost no cancellation
C-Signal	Typically trading in the afternoon (especially US market opening), large trades, large amount traded in dark venues, large POV in general
C-Res	Large orders, medium frequency, sent throughout the day

Table 4: Summary of representative agents.

value. Cluster 1 exhibits the highest average cancellation rate, indicating that this agent type perhaps tracks the execution of its trades more actively than other types. Also, the day time of the trades is uniformly distributed across the whole day, with a mean creation time around noon. This type of agent may be summarised as a "*quantitative agent*", called C-Quant in the following sections.

2. **Day VWAP agents** (C-Day VWAP): The most distinct cluster from C-Quant is cluster 2 as its features exhibit the largest anti-correlation to features of C-Quant. The number of trades per agent is the lowest, and all trades are submitted very early in the morning. The average trade size is larger than for C-Quant, yet not the largest across the agents for this partition. Most noticeable apart from the few number of trades with larger size is the creation time of the order, which typically only occurs before market open. This indicates that these orders are large orders typically sent before market opening, and with an execution target over the entire day. Looking at the execution algorithm, one can primarily find rather passive algorithms such as VWAP, which further supports the previous statement. This trading behaviour is further reflected in the cancellation rate, which is the lowest across all clusters, while having most child orders during the execution. This type of agent is summarised as a "*Day VWAP agent*", called C-Day VWAP in the following sections.

3. **Signal agents** (C-Signal): Cluster number 3 is most distinguishable due to its creation time, which shows a minimum of $\sim$ 13:30 and a maximum of $\sim$ 15:00 for the corresponding data slice. Hence, this is an agent type which is typically active in the afternoon (which corresponds to the opening of the US market since it is a STOXX 600 instrument). Similar to C-Day VWAP, the cancellation rate is quite small, only $\sim 5\%$, when compared to C-Quant which is around $\sim 14\%$. This agent type tends to execute large order sizes (5% of the average daily volume of the last 20 days). Additionally, the cluster in this example has a high percentage of volume, and significant fractions are executed via dark venues which leads us to assume the agent type is less concerned about execution costs, but has a high urgency – hence also executes a lot in dark venues to mitigate traces in the market. The reason may be that this agent type wants to act on trading signals and is thus referred to as a "*Signal agent*", called C-Signal in the following sections.

4. **Residual agents** (C-Res): Cluster 4 indicates agents with the largest average trade size, and only about

Electronic copy available at: https://ssrn.com/abstract=3997109

five trades per month. The trades are submitted during the entire day, which is also indicated by the high standard deviation of the creation time. Furthermore, the standard deviation of the sizes, as measured in the logarithm of the average daily volume percentage, is the highest. This is the "least distinguishable" agent type, and seems to be in-between the other three clusters. One possible reason may be that some agents trade different strategies, and thus do not act very homogeneously. Hence, these agents may be referred to as "*residual agent*", denoted as C-Res in the following sections.

3.4 Stability of clusters

Section 3.3 shows that the set of agents $B_{i,t}$, trading asset i during period t through a broker, can be segmented into different clusters. Setting $K = 4$ gives an appropriate number of agent types which can be interpreted, and are also very distinct in their representative features. Moreover, the partitions do not change significantly when changing the clustering algorithm or the feature standardisation.

What is unknown is how the cluster affiliations and the representative agent types change over time or different instruments. Are the results in Section 3.3 random or is the partition rather stable? This section addresses the stability of the agents and clusters, both across different instruments as well as different time slices. One problem arising when comparing two partitions in the present case is that most often $B_{i,t} \neq B_{j,t'}$. In other words, traders which are active in asset i during period t do not necessarily trade asset j during period t'. This is expected to hold in particular for traders which tend to trade at lower frequency or rather long-term strategies. Nonetheless, the following two questions are of particular interest:

1. How stable is the affiliation of an agent to a particular cluster? Do agents change their behaviour and act differently across time or different tickers?
2. How stable are the representative agent types and their features presented in Section 3.3? How do they vary across time and different tickers?

Two approaches are pursued to answer these questions:

1. **Joint clustering of several data slices**:

   As before $B_{i,t}$ is the set of agents α with at least one trade in asset i during period t. For example, $\bar{x}_{i,\alpha} = \frac{1}{|T_{i,\alpha}|} \sum_{t \in T_{i,\alpha}} x_{i,t,\alpha}$, where $T_{i,\alpha} = \{t | \alpha \in B_{i,t}\}$ denotes the average of a trader's features across all time periods in which they have traded at least once.

   If T periods of one asset i, i.e. $B_{i,t}\ \forall\ i \in \{1, \dots, T\}$ are clustered together, a high concentration of observations from one agent, e.g. $x_{i,t,\alpha}, T_{i,\alpha}$ in one cluster C_k indicates consistency of the agent across different time spans, and similar for different assets in the same period. Denoting $P_\alpha(\alpha \in C_k)$ as the empirical probability for an observation of agent α to be in cluster C_k, the entropy

   $$H_\alpha = - \sum_{k \in \{1,\dots,K\}} P_\alpha(\alpha \in C_k) \log(P_\alpha(\alpha \in C_k)) \tag{14}$$

   is used to measure a client's concentration in one cluster. Additionally, the maximum affiliation probability, defined as

   $$\max_k P_\alpha(\alpha \in C_k), \tag{15}$$

   indicates a more interpretable degree of concentration. For evaluation, we compute the log-weighted average of both (14) and (15) across different traders. This assigns a higher weight to agents active in many tickers, and none to agents which are only active in one ticker.

2. **Meta clustering**:

   For stability of the representative agent types, their representative features should coincide over different tickers (or time spans). Matching the clusters of two tickers/time frames is difficult for two reasons. Firstly, an agent does not necessarily behave equally in different tickers (or time) which makes a mapping based on trader affiliation not necessarily correct. Secondly and more importantly, many agents are potentially active only every couple of months and not in every ticker $B_{i,t} \neq B_{j,t'}$. Hence, the intersection of $B_{i,t}$ and $B_{i,t'}$ (or $B_{j,t}$) may not contain enough samples to correctly match the clusters from two different partitions.

Electronic copy available at: https://ssrn.com/abstract=3997109

To circumvent this problem, meta clustering can be applied. In particular, every $B_{i,t}$ is clustered as before. The result are several different cluster partitions, T different cluster partitions for each month clustered of one particular instrument.

$$\mathcal{C}^{i,t} \text{ with } \mathcal{C}^{i,t} = \{C_0^{i,t}, \dots, C_K^{i,t}\} \; \forall \, i \in \{1, \dots, N\}, t \in \{1, \dots, T\}$$

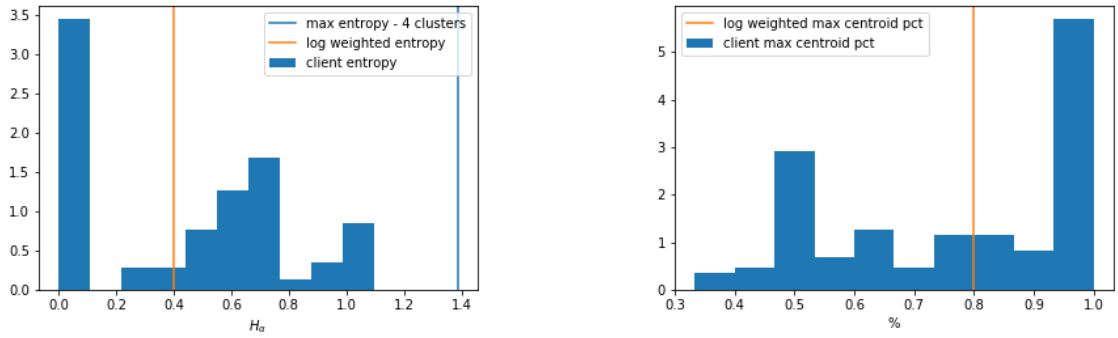
where each partition $C_k^{i,t}$ has its cluster center $\bar{x}_k^{i,t} = \frac{1}{|C_k^{i,t}|} \sum_{\alpha \in C_k^{i,t}} x_{i,t,\alpha}$. The $K \cdot |T|$ or $(K \cdot |N|$ for instrument dimension respectively) cluster centers obtained in the first step are then clustered in the second step. A high concentration of the first stage cluster centers indicates a rather high stability of the agent types and their representative features. In other words, a very clear partition with low variance within the cluster, as well as a high variance between values from different cluster centers, are expected.

In our computations, the number of clusters is set to $K = 4$ as outlined in Section 3.3. The analysis is performed for the 25 most liquid instruments in the *STOXX 600* index, in terms of the numbers of parent orders submitted in the period. Two years of data corresponds to 24 partitions per ticker. Figure 5 shows the distribution of both entropy and the maximum affiliation probability across the agents. Table 5 shows log-weighted values for entropy and maximum affiliation, in both ticker and time dimension. Log-weights with respect to the number of tickers (respectively, time periods) a trader has been active in are used to give higher weights to such traders that span across several distinct sets.

The log-weighted average of the entropy of ~ 0.4, which is relatively low compared to the maximum value of entropy for 4 clusters of 1.4. The log weighted maximum affiliation probability is $\sim 0.8 = 80\%$. In other words, for the cluster C_k which contains most of some agent's observations, the average probability of that agent's observation to be in this cluster is 80%. This indicates a fairly strong stability of the agent's affiliation to a cluster, across different tickers. This indicates, the agents have similar behaviour throughout time and instruments.

Dimension	Entropy	Maximum Affiliation
Ticker	0.4	0.8
Time	0.5	0.76

Table 5: Log-weighted average of agents' entropy and maximum affiliation probability. The first row indicates both values for the consistency of the 10 different tickers, for an exemplary time slice. The second row indicates the values for 24 time slices, of one exemplary ticker.



(a) Entropy.

(b) Maximum affiliation percentage.

Figure 5: Entropy and maximum affiliation probability for different agents. The orange vertical lines indicate the log-weighted averages. Log-weights are based on how many tickers an agent has been active in. Hence, agents with only one observation are given a weight of equal to zero.

Figure 6 shows traders for several different tickers in their embedding space. In particular, the observations of two traders, one with low (orange) and and one with high (green) entropy are accentuated. The observations of the low entropy trader can be observed to be very concentrated in the embedding space as expected. The observations of the high entropy trader are more spread out than for the low entropy trader. However, the observations of they do not spread across the entire embedding space but are rather clustered as well. This indicates that even those agents with a higher entropy do have some consistency in their trading.

Electronic copy available at: https://ssrn.com/abstract=3997109

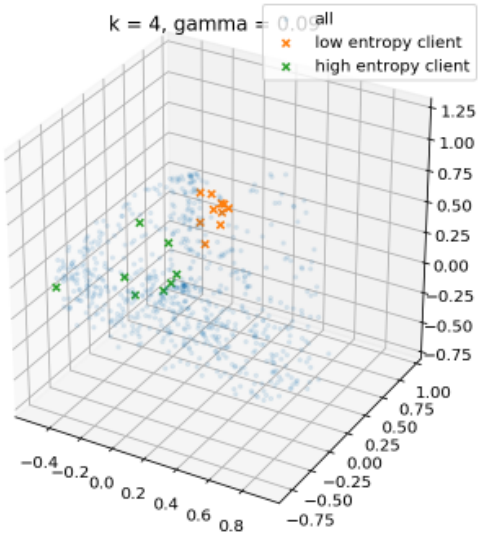


Figure 6: Observations of two agents visualised in embedding space. Despite one agent exhibiting one of the largest entropy values, the points are still concentrated in one area of the embedding supporting the evidence that agents act similarly across different tickers.

For the temporal comparison, 24 months of one ticker are clustered together. In particular, for the project's aim to use the clustered trader types in order to simulate markets, a certain degree of stability of the clusters would be beneficial to design a model which represents and captures the heterogeneity which can consistently be found in LOBs. While an individual agent may change its behaviour, it would be desirable if the cluster centers and the clusters' aggregated properties are rather stable.

As outlined above, an agent may have several observations (one for each month, which is used for the analysis). These can be clustered together, and the entropy along with the maximum affiliation probability can be computed. The second row in table Table 5 shows a slightly higher entropy and a slightly lower maximum affiliation probability. This indicates that agents tend to have a slightly higher consistency across different tickers at the same time, than for the same ticker but across different time slices. Potentially, many agents may not act every month, for instance, due to the typical holding period of the purchased instrument (e.g. trading frequency in general). Thus, despite similar trading behaviour, they do not occur quite as often in a data.

Concerning the stability of the representative features of agent types, meta clustering is applied. In this setting, cluster centers obtained (1) for different tickers at the same time slice, and (2) the same ticker and different time slices, are now considered as single observations. If the clusters are consistent, the meta-clusters (the clusters of the representative agent types) should be very clear and distinguishable, and each cluster centre should belong to the corresponding meta cluster. Figure 7 shows the cluster centers in their embedding in $\mathbb{R}^2$, with the colour indicating the cluster from the first stage. Cluster centres of the same type are very much concentrated – e.g. all C-Quant cluster centers are located closely together in the embedding.Just few points are close to the cluster centers of other types. For some partitions some clusters are closer to the C-Res point cloud.

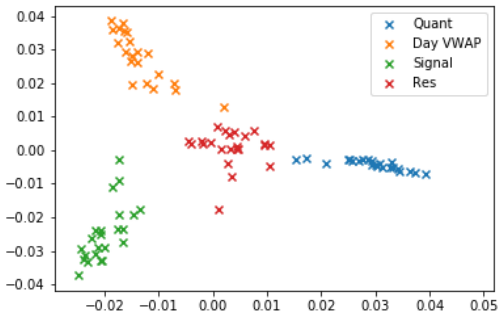
This is also confirmed by the meta clustering using $K = 4$ meta clusters. Both Table 6 and Table 7 show very small confusion where cluster centres are not correctly grouped in their meta clusters. The most prominent confusion is occurs in table 6. In particular, for three instruments, the C-Quant cluster is allocated to the Meta Residual cluster.

	Meta Quant	Meta Day VWAP	Meta Signal	Meta Res
C-Quant	22	0	0	**3**
C-Day VWAP	0	25	0	0
C-Signal	0	0	24	**1**
C-Res	**1**	0	0	24

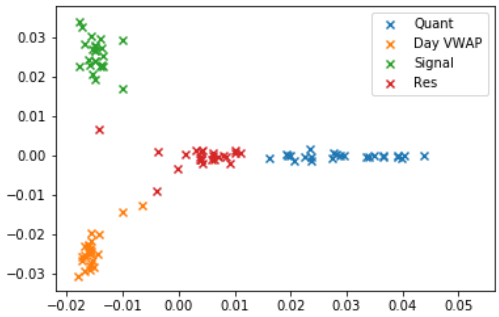
Table 6: Affiliation of clusters to meta clusters for different tickers.

To further support the evidence for the agent types' stability, one may look at certain representative features

Electronic copy available at: https://ssrn.com/abstract=3997109



(a) Different tickers, same time period.



(b) Same ticker, different time periods.

Figure 7: Embedding of first stage cluster centers from the single time slice clustering process. Colours indicate the first stage clustering result. As discovered in Table 3, C-Res is positioned in-between the other three clusters.

	Meta Quant	Meta Day VWAP	Meta Signal	Meta Res
C-Quants	23	0	0	**1**
C-Day VWAP	0	23	0	**1**
C-Signal	0	0	24	0
C-Res	0	0	0	24

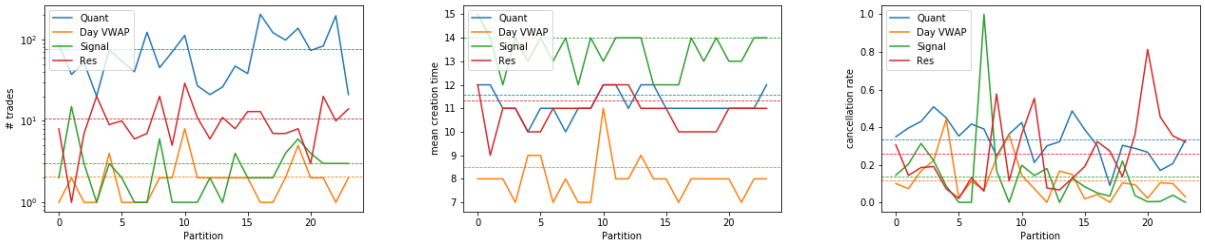
Table 7: Affiliation of clusters to meta clusters different time periods.

from the clusters across different months or tickers. Figure 8 visualises three of the clusters' representative features for 24 different months:

- The average number of orders (Figure 8a) was one of the most relevant features in the clustering. Also the centroid values over time show strong discriminative behaviour here. C-Quant is trading by far the most in all months, but one where C-Res trades similarly often. C-Day VWAP and C-Signal show a similar number of trades on the lowest level. This further supports our interpretation above and the embedding in Figure 7 that C-Res lies in-between all clusters, and sometimes much closer to C-Quant in terms of the number of trades.
- The mean creation time depicted in Figure 8b shows a both a discriminative and stable behaviour, similar to the number of trades in Figure 8a. C-Quant and C-Res have a mean creation creation time around noon both types typically trade throughout the entire day. In contrast, C-Day VWAP trades very early in the morning and thus shows a mean submission time much earlier than the remaining agent types. For C-Signal, the mean submission time for different months fluctuates to a certain degree around the US market open. In general, the feature values from C-Day VWAP and C-Signal are well separated from C-Quant and C-Res.
- For mean percentage of volume, the situation differs slightly. In general, the centroids' mean value fluctuates less and the time series of the values throughout different months are more overlapping. The dashed lines indicating the mean values of the different partitions still indicate the same order, where in particular C-Quant tends to have a high POV since the parent orders often lead to just one child order execution leading to a large POV. Second highest is C-Res, which in several periods contains traders which tend to behave like C-Quant agents increasing the average percentage of volume. C-Day VWAP, apart from one large outlier, has the lowest average POV of all clusters. In general, mean POV illustrates that not all of the features used in the clustering process are stable or have the same ranking throughout different partitions.

This section's results indicate a high stability of the agent types presented in Section 3.3. This leads us to assume that the observations are not random, but rather follow a consistent pattern that exhibits different types of traders acting in the limit order market ecosystem.

Electronic copy available at: https://ssrn.com/abstract=3997109



(a) Number of orders (log scale) (b) Mean submission time (c) Cancellation rate

Figure 8: Exemplary centroid features of one ticker for several months. The solid lines indicate the actual values over the months, the dashed horizontal lines indicate the average over all months.

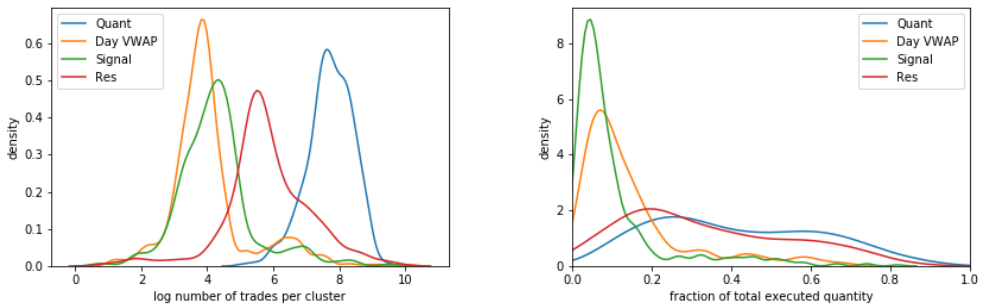
4 Decomposition of order flow components

This section reviews properties of the order flow components which were segmented by clustering the agents in Section 3, both on parent and child order level. This shall answer the question whether the components which are found to have strong heterogeneity in their features, also lead to different dynamics in the LOB. This section builds a change of perspective as we do not look at different traders anymore but aggregate over a particular cluster as one component. First, we look at the sizes and activities of the components' flow over different data slices. Following this, child order properties, profitability and the correlation of inventory with the market's price move are analyzed to outline notable differences between the components.

Figure 9a illustrates the distribution of the number of orders for each of the order flow components, throughout 25 different instruments of the *STOXX 600* in two years. The distributions differ substantially. While C-Day VWAP and C-Signal show a similar distribution in terms of numbers of trades per month, C-Quant exhibits the largest numbers of orders. C-Res is once more located in-between the C-Quant and C-Day VWAP/C-Signal. This matches and further intensifies our previous interpretation that C-Res is some mixture of the first three agent types. Table 8 also indicates that the majority of the trades are done by C-Quant, and only a small fraction by C-Day VWAP and C-Signal.

The fraction of the total executed quantity from each of the single data slices is indicated in Figure 9b. The distributions of the fractions from C-Quant and C-Res show a similar shape. The same holds for C-Day VWAP and C-Signal. In fact, the sum of the executed quantity from C-Quant and C-Res does not vary much due to their strong negative correlation. The fraction of C-Res tends to increase when a trader from C-Res behaves rather like a C-Quant trader or potentially as a mixture of C-Quant, C-Day VWAP and C-Signal.

In comparison to the very small number of orders submitted by C-Day VWAP and C-Signal, the actual executed quantity is much higher. In other words, while C-Day VWAP and C-Signal tend to account for only a small fraction of the number of orders, their contribution to the overall executed quantity is much larger since the orders are generally larger and/or lead to higher execution size (for example, due to fewer cancellations and longer execution). The mean fractions are displayed in the "Executed qty" row of Table 8.



(a) Number of submitted orders (b) Fraction of executed quantity

Figure 9: Distribution of the number of trades and the executed quantity.

Electronic copy available at: https://ssrn.com/abstract=3997109

	C-Quant	C-Day VWAP	C-Signal	C-Res
Number of orders	0.69	0.06	0.06	0.20
Executed qty	0.41	0.15	0.11	0.34

Table 8: Average values of the distributions of clients, trades and executed quantities across different clusters

4.1 Heterogeneity of child orders

So far, we primarily focused on the structure of the parent orders which were also used for the segmentation of the traders. This section reviews the differences of the four different components regarding the child orders which are sent to different venues. This may give further insights to the degree to which the order flow components differ not only at parent order, but also at the child order level. Table 9 shows a summary of the orders submitted to exchanges for one exemplary $B_{i,t}$. These sample results are for BATS.L for the month of December 2019. The numbers are normalized either by the mean of the corresponding statistic or by their sum.

In line with the observations regarding the number of orders and executed quantity from Figure 9a and Figure 9b, the number of child orders submitted by C-Quant is by far the largest. Furthermore, the quantitative agents have the highest direct market access (DMA) ratio. This means these child orders come from parent orders that skip the processing of the broker's execution algorithms, hence primarily using the broker as a platform to send their orders to a venue. The DMA ratio is also high for the C-Signal component. In this particular segmentation, C-Signal contains one agent which heavily and exclusively trades with DMA orders leading to both an unusually high DMA and cancellation rate. This may be an agent which rather belongs to the C-Quant cluster and impacts the statistics of the C-Signal order flow component. For the C-Res and C-Day VWAP clusters, the DMA ratio is almost zero, and only around 75% of the child orders get cancelled.

Regarding the order and execution sizes of child orders, the orders of C-Quant have the lowest mean and standard deviation, as indicated in Table 9. C-Signal has by far the highest mean order and execution quantity. A significantly higher mean compared to median indicates the presence of outliers for all clusters, which is potentially due to larger orders in dark venues. In particular, component C-Signal exhibits many executions in dark venues for this month as mentioned in Section 3.3, hence rendering a large mean order quantity more plausible. As for the median, the execution quantities are significantly lower than the order sizes. One reason for this may be partially filled orders – in particular, in dark venues. Note that we exclude child orders without any partial fills for the present statistics.

Lastly, *Median Distance to Best Price* indicates the relative price level at which the child orders of the particular component tend to be placed (i.e., this can be construed as a proxy for aggressiveness or urgency). Again, the median is displayed due to its robustness. C-Quant child orders are placed closest to the best price. The median distance from the best price for C-Res and C-Signal is roughly similar, while the orders from C-Day VWAP are placed deeper in the book, when compared to child orders from the other clusters.

	C-Quant	C-Day VWAP	C-Signal	C-Res
# Child Orders	0.88	0.02	0.03	0.06
Mean Order Qty.	0.23	0.42	2.74	0.61
Median Order Qty.	0.29	0.28	3.11	0.33
Std Order Qty.	0.52	0.61	2.02	0.85
Mean Exec Qty.	0.12	0.18	3.53	0.17
Median Exec Qty.	0.70	0.75	1.76	0.79
Std Exec Qty.	0.33	0.40	2.95	0.31
DMA Ratio	2.27	0.03	1.63	0.07
Cancellation Ratio	1.07	0.90	1.09	0.94
Median Distance to Best Price	0.52	1.81	0.90	0.77

Table 9: Aggregated statistics of child orders by clusters. This table shows the total number of submitted child orders, the ratio of orders which where sent due to direct market access, the ratio of child orders which were cancelled, the mean, median and standard deviation for both order quantity and the executed quantity of orders. The last row shows the median distance of a limit order from the best opposite price. The number of child orders is normalized by the sum of all clusters; the remainder of the statistics are normalized by the mean of the four clusters.

Electronic copy available at: https://ssrn.com/abstract=3997109

Lastly, we look at daily net inventory of the different components' child orders, the sum of the signed traded sizes (positive for buy, negative for sell) submitted by a cluster during one day. Table 10 indicates the mean cumulative inventory for one exemplary month. In particular, C-Quant appears to have been a net seller, while the remaining clusters were net buyers in that particular month during which the corresponding stock showed a positive return.

	C-Quant	C-Day VWAP	C-Signal	C-Res
Average inventory	-0.62	0.12	0.11	0.16

Table 10: Table indicating the mean cumulative inventory of the clusters. Values are normalized with the absolute inventory of the clusters.

Table 11b indicates the correlation between the components over the whole duration and all instruments for both the net inventory as well as the order flow imbalance[3]. Clusters C-Signal and C-Res show the strongest dependency with a negative correlation of -0.09. However, the correlations obtained are mostly statistically insignificant. Regressing the net inventory or the order flow imbalance of one cluster on any of the three other components, the coefficients rarely show any significant deviations from zero on instrument basis. For only very few instruments, weak significant correlations can be found, but the average p-value over all instruments considered here is around 0.25. We furthermore fit regressions over several instruments but a smaller time horizon, under the assumption that the correlations may change over time. Again, few variables show a persistent significance throughout time and the resulting R^2 are very small. This indicates that, despite significance for few variable combinations, the explanatory power is small.

	C-Quant	C-Day VWAP	C-Signal	C-Res
C-Quant	1.00	-0.05	-0.02	-0.04
C-Day VWAP	-0.05	1.00	0.00	-0.03
C-Signal	-0.02	0.00	1.00	-0.09
C-Res	-0.04	-0.03	-0.09	1.00

(a) Net inventory.

	C-Quant	C-Day VWAP	C-Signal	C-Res
C-Quant	1.00	-0.07	0.00	-0.11
C-Day VWAP	-0.07	1.00	-0.02	-0.04
C-Signal	0.00	-0.02	1.00	-0.03
C-Res	-0.11	-0.04	-0.03	1.00

(b) Order flow imbalance.

Table 11: Tables indicating the correlation of the order flow between different components.

The resulting both inconclusive as well as insignificant correlation between the net inventories (OFIs respectively) indicates that not only the behaviour between the cluster differs quite substantially but they are also independent from each other in the way the accumulate inventory. One possible reason for this may be that different trader types trade different strategies which are either not correlated at all (leading to insignificant correlations). Another explanation would be that changes depend on the market environment as some strategies may only correlate during certain market conditions. In particular the first makes sense since the trading types seem to act on different time scales in the market. The most persistent observation is a slight negative correlation between C-Quant and the remainder of the order flow components which, however, is relatively weak.

4.2 Profitability

In Section 3, we showed that traders using execution services may be summarised into different clusters. These trader types have different properties when it comes to the type of orders they send. This leads to the assumption that the objectives of the segmented agent types may differ as well, for example, with respect to their horizon of investment. To this end, we analyze the hypothetical profit and loss (PnL) for each order flow component, in order to investigate structural differences in the returns of the components' trades. We remark that this PnL is hypothetical as it does not refer to the actual inventory of the trader. For example, it may be that a trader is not holding a position for the respective future horizon of time, or that it is actually unwinding a short position instead of building a long position, or the holding period is different. It much rather represents the average PnL of the respective component, at a certain fixed time horizon.

To investigate the hypothetical profitability of each component, we compute the PnL of each trade via

$$PnL_t^l = -\text{sign}(q^{target}) \log\left(\frac{p_{t+l}}{p^{exec}}\right), \tag{16}$$

[3]The order flow imbalance is computed with the net inventory divided by the total trade volume of the component. A more detailed explanation can be found in Section 4.3.

Electronic copy available at: https://ssrn.com/abstract=3997109

where p^{exec} is the volume-weighted execution price of the corresponding parent order. q^{target} is the target quantity of the parent order, as specified in Definition 2.3, and $-\text{sign}(q^{target})$ the negative sign of the return. If, for example, $q^{target} > 0$ the order is a sell order, thus we multiply the log-return with -1. For some timestep $t + l$, p_{t+l} denotes the closing price at $t + l$, where we consider trading days as increments. For $l = 0$, p_t corresponds to the close price of the day when the corresponding trade happens.

The volume weighted execution price p^{exec} of one parent order is computed via

$$p^{exec} = \frac{1}{q^{exec}} \sum_{x \in \mathcal{X}^{exec}} q_x \cdot p_x, \tag{17}$$

where q_x and p_x indicate the quantity and the price of each execution in $\mathcal{X}^{exec}$ of the corresponding parent order. Finally, we compute the expected PnL for each component $k \in \{1, \dots, K\}$ of trades at day t, as $\mathbb{E}(PnL^l_{t,k})$ by averaging over all returns for a given l, t, k. The result is a daily time series for different lags $l \in \{0, 1, 10, 20\}$ and different components $k \in \{1, \dots, K\}$, where each element is the average PnL of the trades occurred on that particular day. The same computation is done using market excess return, where we subtract the future market return from the instrument's future return, for the PnL computation in Equation (16).

Table 12 indicates the average expected PnL in basis points for 2018 and 2019 for 25 instruments. From trade to close, C-Quant appears to be the only order flow component generating a slight profit over the period covered here. This indicates, C-Quant is potentially pursuing more intraday like strategies. For C-Signal, the 20-day return (both raw return and market excess return) are the largest. Even though this is not a realized return, it gives reason to assume this agent type has some information medium-frequency comparison to the other components. This is supported by the observation that C-Signal shows the smallest PnL on the same trading day ($l = 0$), which further supports our interpretation of C-Signal given in Section 3 that this component is not aiming for a profit realizing the same day. It may well be C-Signal trades mean reversion signals which realize only after around a month. Apart from that, C-Day VWAP shows a slight outperformance on the $l = 1$ horizon.

	$l = 0$	$l = 1$	$l = 10$	$l = 20$	$l = 1$, excess	$l = 10$, excess	$l = 20$, excess
C-Quant	**0.38**	0.72	3.39	3.76	0.78	3.78	2.83
C-Day VWAP	-0.25	**1.54**	-2.04	-3.95	**1.42**	0.15	1.44
C-Signal	-0.47	-0.43	3.09	**13.17**	0.32	2.65	**10.63**
C-Res	0.08	0.37	-1.36	-1.51	0.52	-2.24	-2.19

Table 12: Table indicating the average $\mathbb{E}(PnL^l_k)$ for agent types in basis points (bps). The suffix *_excess* indicates market excess returns over the *STOXX 600* baseline.

Additionally, Figure 10 shows the cumulative hypothetical PnL for each component. In line with the observations from Table 12, C-Signal is clearly outperforming the other agent types on a 20-day horizon (lower right plot) while having the worst performance from trade to close and to $t + 1$ (upper plots).

4.3 Order flow imbalances during volatile periods

While some components show differences in their expected returns with respect to the trading horizon, it stands to question to which degree the components' order flow on a particular day is correlated with the return of the day. In particular, how does the order flow of different agent types behaves when markets move substantially. To this end, we compute the order flow imbalance (OFI) for each instrument via the following measures

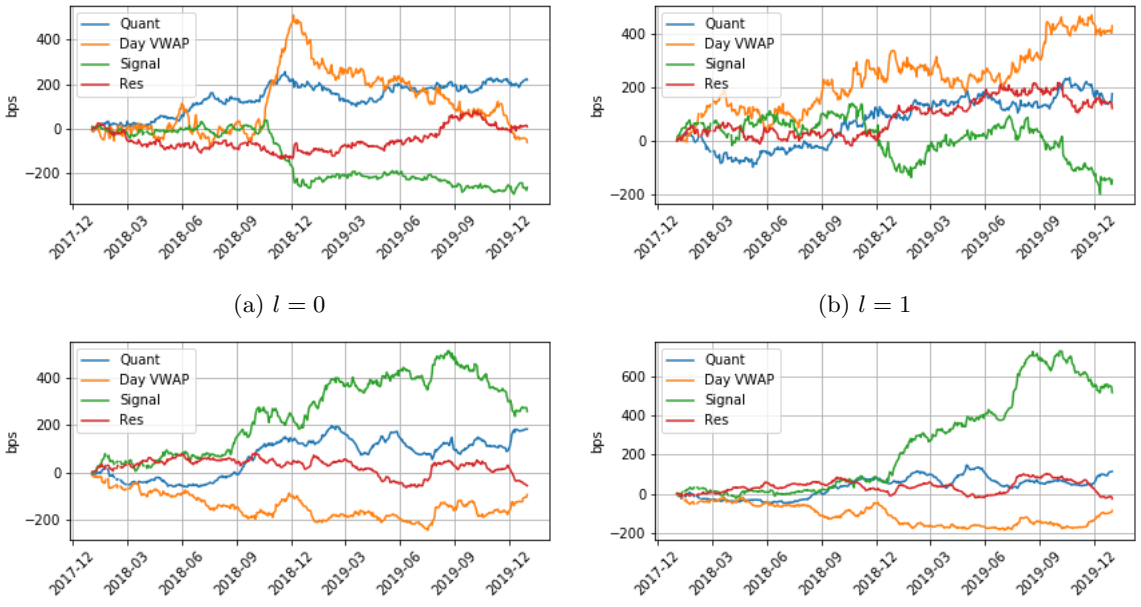
$$OFI^{C_k} = \frac{\sum_{p \in \mathcal{X}^k} q^{exec}}{\sum_{p \in \mathcal{X}^k} |q^{exec}|}, \tag{18}$$

and

$$OFI^{ADV} = \frac{\sum_{p \in \mathcal{X}^k} q^{exec}}{\text{ADV}}, \tag{19}$$

where $\mathcal{X}^k$ is the set of parent orders from cluster k on a given day. We compute the imbalance of the order flow in two ways. The first imbalance shown in Equation (18) is normalized by the total executed quantity of the

Electronic copy available at: https://ssrn.com/abstract=3997109



(a) $l = 0$ (b) $l = 1$

(c) $l = 10$ (d) $l = 20$

Figure 10: Cumulative PnL $\mathbb{E}(PnL^l_{t,k})$ of agent types over different time horizons, $l = \{0, 1, 10, 20\}$.

cluster, hence called OFI^{C_k}. The second imbalance shown in Equation (19) is normalized by the average daily volume (ADV) from the last 20 days and denoted as OFI^{ADV}. This is to take into account that a large OFI, as defined in Equation (18), does not necessarily mean a large impact to the market during the particular day. That is because the total traded volume of the cluster might only be a small fraction of the daily volume. In contrast, Equation (19) makes different values of the same day more comparable across different clusters, and is large only if a component's net inventory is large in relation to the historical ADV. E.g. one cluster might have a large OFI in terms of own orders but still a very small OFI measured on the ADV due to small traded volume.

Figure 11 shows a histogram of the OFI as in Equation (18) to simplify comparison. C-Quant stands in contrast to the other three components. The net order flow peaks around zero, while the other order flow components have two peaks at zero and one. In particular, the aggregated order flow from C-Quant tends to be rather neutral in terms of order flow imbalance, which can be due to two reasons. First, C-Quant trades much more often (albeit smaller sizes) and thus facilitates order flow imbalances closer to zero. Second, C-Quant pursues more intraday/medium frequency strategies without accumulating larger positions.

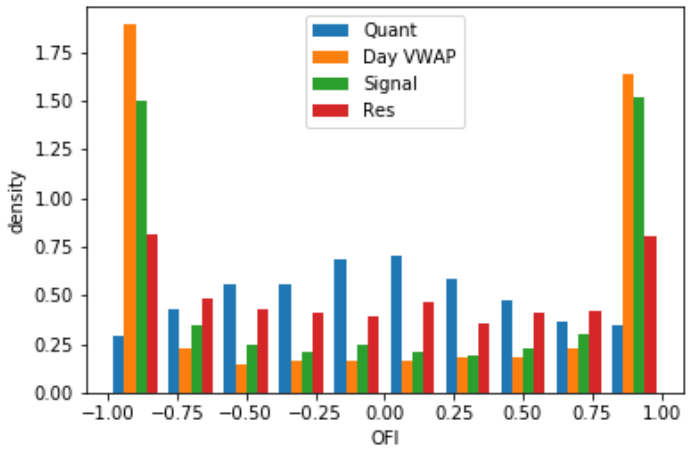


Figure 11: Order flow imbalance by cluster, during days of returns with large magnitude.

Table 13 shows the correlation of the OFI computed via Equation (18) in Table 13a and Equation (19) in Table 13b with the log return from open to close (logOC). C-Day VWAP and C-Res show the clearest picture.

Electronic copy available at: https://ssrn.com/abstract=3997109

For both the case of days of large returns of the particular instrument (left columns), as well as days of large returns of the index (right columns), C-Day VWAP and C-Res exhibit a fairly strong positive correlation. This holds for the net inventory scaled by the traded volume of the particular component (Table 13a), as well as the net inventory scaled by the market volume (Table 13b). In fact, the correlation of OFI^{ADV} with both the instrument and the index return are larger than the correlation of OFI^{C_k} with the return. In other words, when these clusters accumulate net inventory which is also large in terms of the average daily volume, the instrument is likely to also exhibit a larger return. A possible reason for this may be that the order flow becomes a driving factor in the market.

cluster	logOC	logOC_index
C-Quant	-0.0301	0.0113
C-Day VWAP	**0.1590**	0.0229
C-Signal	**-0.1506**	0.0499
C-Res	**0.1102**	**0.1616**

(a) OFI^{C_k} as in Equation (18).

cluster	logOC	logOC_index
C-Quant	0.1368	0.0833
C-Day VWAP	**0.2044**	0.1416
C-Signal	**-0.0441**	**-0.0794**
C-Res	**0.2189**	0.2390

(b) OFI^{ADV} as in Equation (19).

Table 13: Correlation between OFI and returns of stocks (column logOC) and return of the index (column logOC_index), for two different types of normalized OFI.

C-Signal is the only cluster exhibiting a fairly large negative correlation with negative returns of the instruments. This correlation seems to be less strong when the net inventory is large as defined in Equation (19). A reason for this may be that due to the high participation rate, C-Signal becomes a driving factor of the market, similar to C-Day VWAP and C-Res above. The correlation with the index is less clear and varies between OFI^{C_k} and OFI^{ADV}. However, it seems that the index is more likely to decrease if the net inventory is large, also based on the average daily volume of the corresponding stock.

C-Quant shows the lowest correlation of net inventory scaled with total trade size, with both the daily return as well as the index (-0.03 and 0.01, respectively). This is to be expected, taking into account Figure 11 that shows the tendency of C-Quant to keep a net inventory closer to zero. For net inventories which are large measured on the total market volume following Equation (19), however, the correlation also seems to turn stronger. Similar as for C-Day VWAP and C-Res, if the net inventory is large measured on the ADV, the cluster becomes more of a market driver and a higher correlation with the daily return can be observed.

5 Heterogeneous parent order model

As illustrated in the order process in Figure 1, limit orders sent through execution services are usually part of a larger parent order. This section proposes a simple model which captures the most important heterogeneous properties for each of the components extracted in Section 3. Once these parent orders can be modeled, their scheduling and execution in the LOB may be simulated. This can be done by replicating the flow of the parent orders orders into the LOB as depicted in Figure 1 via known execution and order routing algorithms, some of which are well studied in the literature. Modelling the flow of these components itself without the direct scheduling and execution can additionally be of high interest to brokers, as this can possibly improve the broker's knowledge and service to clients. Rather than aiming to replicate and fit the data to the full extent, the following model is designed to outline a starting point which has good fits from a marginal perspective for several stocks, despite being very simple. It also further underpins the structural differences between the different clusters which have been outlined in the previous sections. Furthermore, we assume independence between the flows presented in the following, due to the absence of a significant correlation structure between the components, as detailed in Table 11.

5.1 C-Quant

As outlined in Section 3, the C-Quant order flow component is submitting orders throughout the day. Figure 12c suggests the U-shaped intra-day pattern commonly observed in trading behaviour of limit order books [9]. In the morning and towards closing time of the market, the intensity of the orders increases. C-Quant order sizes are of mostly small size and typically executed in just a few child orders. This, however, does not imply that only one order is sent to the exchange.

Electronic copy available at: https://ssrn.com/abstract=3997109

As per modelling the C-QUANT order flow component, we suggest a non-homogeneous Poisson process. In particular, $\{N^{Quant}(t), t > 0\}$ denotes the counting process of the parent order submissions indicating the number of submitted parent orders up to time t. The arrival time of the n-th parent order is denoted as t_n^{Quant}. New orders arrive proportionally to the conditional intensity rate $\lambda^{Quant}(t)$, $t \in [0, T]$ similar to the shape in Figure 12c. This is to properly reflect the intraday pattern of the C-QUANT order arrivals. It must hold that

$$\int_0^T \lambda^{Quant}(s)ds = 1,$$

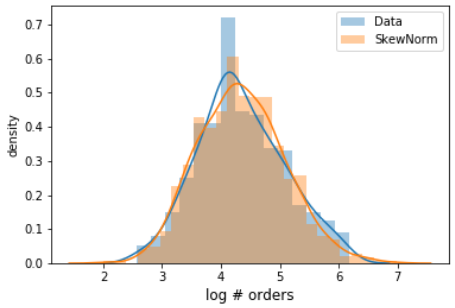
so that the expected number of orders within $[0, T]$ equals one. The conditional intensity function $\lambda^{Quant}(t)$ is then scaled by the number of expected order submissions for the corresponding day

$$\lambda^{Quant}(t) \cdot n^{Quant}.$$

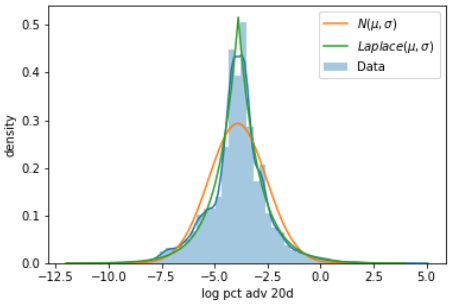
Each of the arriving parent orders additional "marks" or properties as specified in Definition 2.3, in particular a sign and a target quantity. For the C-QUANT parent orders, the following considerations are in place

- the logarithm of the expected number of parent orders, parameter N follows a skewed normal distribution, $\log(n^{Quant}) \sim sN(a, \mu, \sigma^2)$ with skewness a, mean μ and variance σ^2 (Figure 12a),
- the target quantities are fitted with a Laplacian distribution, i.e. $q^{target} \sim \text{Laplace}(\mu, b)$, as illustrated in Figure 12b,
- the signs of the order is Bernoulli distributed $sign_i \sim B(1, p)$,
- the probability of an order being a buy order, p, differs across days and is modelled by a Beta(a, b) distribution; Figure 12d indicates the fit.

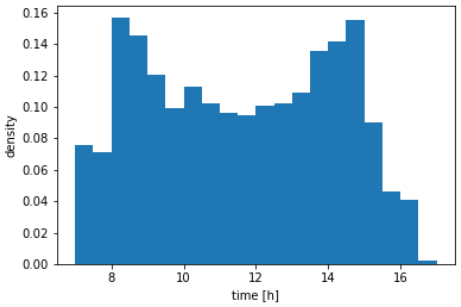
The execution of the C-QUANT orders is done with the Almgren-Chriss optimal execution framework, first presented in [3]. The execution generally consists of only very few if not just one child order and the execution time is very short.



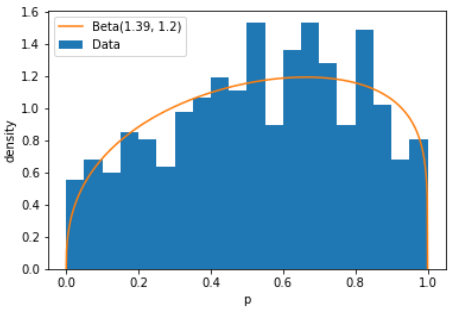
(a) Order counts: QUANT order flow.



(b) Order size distribution: QUANT order flow.



(c) Submission times



(d) Fraction of buy orders

Figure 12: Aggregated parent order flow distributions for one stock, over two years, for C-QUANT.

Electronic copy available at: https://ssrn.com/abstract=3997109

5.2 C-Day VWAP

As outlined in Table 3, C-Day VWAP mainly sends parent orders in the early morning around the market opening. The total number of parent orders is quite small, while execution generally takes places throughout the entire day. A model for these orders is thus quite simple and can be done without any time dependence, since we assume all orders to be submitted at market open (i.e. $t = 0$). The orders are then executed throughout the day proportional to some estimated volume profile.

For C-Day VWAP, it suffices to know how the sum of all buy (respectively, sell) orders is distributed, which is then executed as one large buy order (respectively, sell order). Denoting $X^{DayVWAP}$ as the set of all parent orders from the C-Day VWAP component for a given day, the quantities of interest are

$$\sum_{p \in X^{DayVWAP}} (q^{target})_- \quad \text{and} \quad \sum_{p \in X^{DayVWAP}} (q^{target})_+,$$

where the first term builds the cumulative size of all C-Day VWAP buy orders, and the second term all C-Day VWAP sell orders. The C-Day VWAP component hence consists of two (aggregated) parent orders with

- submission time $t = 0$,
- target quantities for both orders, where the logarithm of the absolute value, $\log(q^{target}) \sim sN(a, \mu, \sigma^2)$ where $sN(a, \mu, \sigma)$ follow a skewed normal distribution with skew a, mean μ and variance σ^2 (as shown in Figure 13),
- the sign of the order, which is -1 for the aggregated buy order and 1 for the aggregated sell order.

The aggregated buy and sell C-Day VWAP orders are executed following to the volume profile of the previous day.

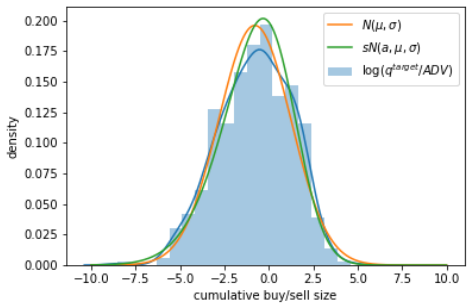


Figure 13: The aggregated parent order sizes (on a log scale) for one stock over two years for C-Day VWAP.

5.3 C-Signal

The C-Signal cluster as outlined in Table 3 sends quite large orders and executes them in a relatively short horizon. This lets assume C-Signal is a more aggressive player in the LOB, moving the LOB in a certain time horizon.

As per modelling the C-Signal component, we suggest a non-homogeneous Poisson process similar to C-Quant. As before, $\{N^{Signal}(t), t > 0\}$ denotes the counting process of the parent order submissions indicating the number of submitted parent orders up to time t. New orders arrive proportionally to the conditional intensity rate $\lambda^{Signal}(t)$, $t \in [0, T]$, similar to the shape in Figure 14c. The conditional intensity function $\lambda^{Signal}(t)$ is then scaled by the number of expected order submissions for the corresponding day

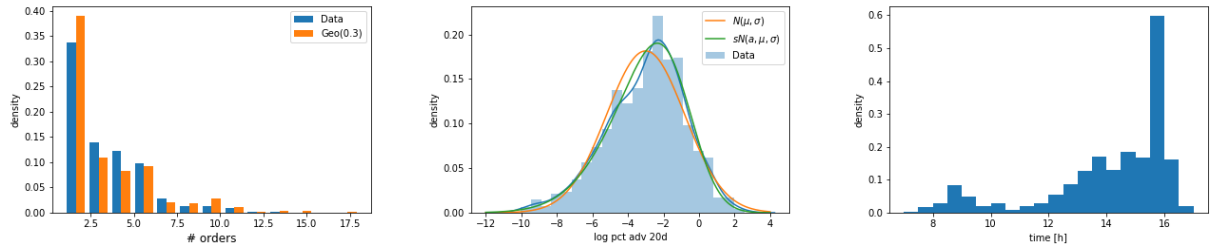
$$\lambda^{Signal}(t) \cdot n^{Signal}.$$

The C-Signal parent orders come with the following "marks" or properties

Electronic copy available at: https://ssrn.com/abstract=3997109

- the expected number of arriving parent orders at a day, n^{Signal}, follows a geometric distribution, i.e. $n^{Signal} \sim \text{Geo}(p)$ (Figure 14a),
- the logarithm of the target quantities C-Signal parent orders $\log(q^{target})$ is skewed normal distributed, $\log(q^{target}) \sim sN(a, \mu, \sigma^2)$ with skew a, mean μ and variance σ^2 similarly to C-Day VWAP (Figure 14b).
- the sign of the orders $sign_i \sim B(1, p)$, where p denoted the probability of a buy orders.

A large proportion of the C-Signal component consists of close related orders. These orders are removed because they do not form part of the continuous trading session. We suggest the execution of the C-Signal orders is done similarly to the C-Quant orders using the Almgren-Chriss framework [3]. The detailed execution, however, is not part of this work. Note, orders from C-Signal are substantially larger than those from C-Quant, which – together with the frequency of orders and the intraday pattern – constitutes the main difference between the two clusters.



(a) Order counts (b) Order sizes (c) Submission times

Figure 14: Aggregated parent order flow distributions for one stock over two years for C-Signal.

5.4 C-Res

Both the representative features in Table 3 as well as the embedding of the first stage clustering in Figure 7 indicate that C-Res lies in between the remainder of the other clusters. In addition, the confusion matrices in Table 6 and Table 7 confirm that whenever there exists instability in the market segmentation, one of the remaining clusters is mistaken with C-Res. To this end, we model C-Res as a random mixture of C-Quant, C-Day VWAP and C-Res. We suggest random attribution of the mixture following a dirichlet distribution

$$f(x) = \frac{1}{B(\alpha)} \Pi_{i=1}^{3} x_i^{\alpha_i - 1} \quad \text{where} \quad B(\alpha) = \frac{\Pi_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K})} \tag{20}$$

and $x = (x_1, x_2, x_3), \alpha = (\alpha_1, \alpha_2, \alpha_3)$ are the weights between the three clusters. The random allocation is then proportionally added to the first three components.

6 Conclusion

Many different agents act together in limit order books with different intentions, trading horizons and information sets. This gives reason to assume that order flow in limit order books is not homogeneous but rather of different types. To account for different agents submitting orders an alternative notation for limit order books was given in Section 2. This notation allows to derive different views on the order book with different granularity ranging from an anonymised *public view* to the fully detailed *omniscient view*.

To investigate the heterogeneity of those agents which make use of brokers, trade execution data was analysed. Results show these agents may be summarised in four representative clusters which differ substantially in both their trading behaviour as well as the order flow induced in the limit order book by the parent orders. In particular, trading frequency, trade size but also order submission time and execution strategies show notable differences between these agent types which gives evidence that some heterogeneity may be assumed. The insights were used to propose a simple model for parent order flow of each cluster in order to capture some of the heterogeneous dynamics of the different trader types in Section 5.

Electronic copy available at: https://ssrn.com/abstract=3997109

In relation to [15], the results may be seen as additive rather than comparative. In particular, the results presented in this work refers more to the "fundamental buyers/sellers" and "opportunistic traders" classes from [15] as the data in this study excludes HFT and market maker agents which have their own execution platforms. Kirilenko et al. [15] mainly focus on HFTs and MMs in their study. Looking at the entire market one thus would have to aggregate both studies to get the full picture of the heterogeneity in limit order markets.

In contrast to [15], this study is able to distinguish agents on their actual parent orders for several tickers and a longer time horizon. This enables to show that the agent types presented in Section 3.3 are consistent over longer time periods and also exists in different stocks. The only noteworthy confusion factor is between C-Res and some of the other clusters which seem to move towards C-Res under certain conditions.

The results of this study and [15] provide evidence that order flow in limit order books show strong heterogeneity indicating that modelling LOBs under the assumption of homogeneity is not very valid. In contrast to most order flow models in literature, such models should hence include some degree of heterogeneity.

This study builds a first step towards better understanding limit order markets and their heterogeneity. We focus on the parent order process, the very left hand side of the order process depicted in Figure 1. Our results provide a foundation for many future research directions. The interplay between traders with proprietary access to the exchange (HFTs and MMs) and traders acting through brokers is yet to be analyzed in detail. The authors of [15] remark, for instance, that HFTs keep their trading patterns stable also in times of increased market volatility. More detailed studies are yet missing. It is also to be investigated how exactly orders are processed and arrive in the LOB. This would correspond to the center part of Figure 1. While this study provides indications in which fashion each agent type tends to execute orders, it is not exactly clear how much liquidity from each venue tends to go to the lit or dark venues etc. Lastly, the combination of parent order flow and HFTs and MMs may be used to create heterogeneous models for entire LOBs. Under the assumption of similarity of the trader structure between different brokers, these order flows may be scaled up to the entire volume of all brokers. It remains to be investigated whether realistic LOB models can be developed, which intend to incorporate the entire process shown in Figure 1.

Electronic copy available at: https://ssrn.com/abstract=3997109

References

[1] F. Abergel, M. Anane, A. Chakraborti, A. Jedidi, and I. M. Toke. *Limit order books*. Cambridge University Press, 2016.

[2] F. Abergel and A. Jedidi. Long-time behavior of a hawkes process–based limit order book. *SIAM Journal on Financial Mathematics*, 6(1):1026–1043, 2015.

[3] R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[5] J.-P. Bouchaud, M. Mézard, M. Potters, et al. Statistical properties of stock order books: empirical results and models. *Quantitative finance*, 2(4):251–256, 2002.

[6] J. Brogaard et al. High frequency trading and its impact on market quality. *Northwestern University Kellogg School of Management Working Paper*, 66, 2010.

[7] J. Brogaard, T. Hendershott, and R. Riordan. High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306, 2014.

[8] D. Byrd, M. Hybinette, and T. H. Balch. Abides: Towards high-fidelity market simulation for ai research. *arXiv preprint arXiv:1904.12066*, 2019.

[9] R. Cont. Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazine*, 28(5):16–25, 2011.

[10] R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations research*, 58(3):549–563, 2010.

[11] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[12] B. Hagströmer and L. Nordén. The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4):741–770, 2013.

[13] J. Hasbrouck and G. Saar. Low-latency trading. *Journal of Financial Markets*, 16(4):646–679, 2013.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[15] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

[16] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[17] S. Mankad, G. Michailidis, and A. Kirilenko. Discovering the ecosystem of an electronic financial market with a dynamic machine-learning method. *Algorithmic Finance*, 2(2):151–165, 2013.

[18] M. Meila and J. Shi. A random walks view of spectral segmentation. 2001.

[19] C. C. Moallemi and K. Yuan. A model for queue position valuation in a limit order book. *Columbia Business School Research Paper No. 17-70*, 2016.

[20] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[22] J. Sirignano and R. Cont. Universal features of price formation in financial markets: perspectives from deep learning. *Quantitative Finance*, 19(9):1449–1459, 2019.

[23] E. Smith, J. D. Farmer, L. s. Gillemot, S. Krishnamurthy, et al. Statistical theory of the continuous double auction. *Quantitative finance*, 3(6):481–514, 2003.

Electronic copy available at: https://ssrn.com/abstract=3997109

[24] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[25] S. Vyetrenko, D. Byrd, N. Petosa, M. Mahfouz, D. Dervovic, M. Veloso, and T. H. Balch. Get real: Realism metrics for robust limit order book market simulations. *arXiv preprint arXiv:1912.04941*, 2019.

[26] Z. Zhang, S. Zohren, and S. Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019.

Electronic copy available at: https://ssrn.com/abstract=3997109

A Feature table

Feature Name	Explanation
Buy ratio	Pct. of client orders which is a buy order
Cancellation ratio	Pct. of client orders which are cancelled before full execution
# Trades per month	Number of trades per month.
Inventory	Mean inventory accumulation of a client on a given day
NO Limit price ratio	Percentage of the client's orders for which no limit price has been specified (maximum/minimum price for execution of child orders)
Maximum order creation time	Latest time at which a client submits an order
Mean order creation time	Average time at which a client submits an order
Mean order size (ratio of ADV)	Mean order size measured on the average daily volume of the last 20 days, execution may be less.
Mean momentum (bps)	Mean momentum of entire trading day measured in basis points (bps)
Mean percentage of volume	Mean percentage of traded volume during trade horizon (visible fills + dark fills) / (visible market volume), exceeding 100 is indicator for larger placements in dark venues
Mean volatility	Mean volatility during which a client trades measured on the last 20 days
Minimum order creation time	Earliest time a client creates an order
# Active days per month	Number of days a client trades per month
Mean # orders per active day	Number of orders a client trades if it trades during a day
Standard deviation # orders per active day	Standard deviation of the number of orders of days during which a client places at least one order
Standard deviation of order creation time	Standard deviation of a client's creation time
Standard deviation order size	Standard deviation of a clients order size
Total order size	Cumulative trade size measured on the average daily volume of the last 20 days. Indication of how much a client in average trades at all

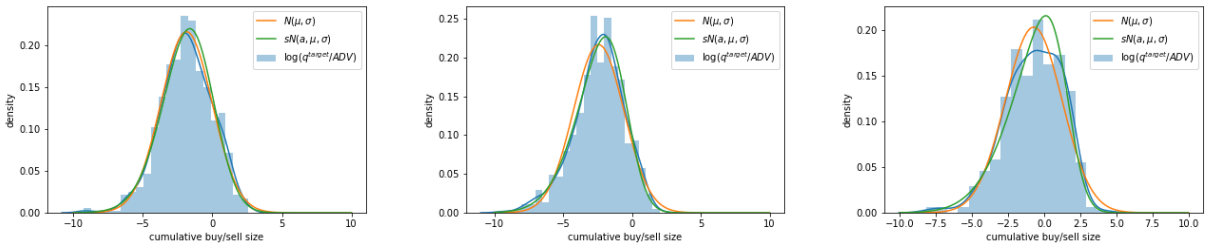
Table 14: Feature list used for the clustering. All features are computed on the base of a one month data set. For instance, #Trades indicates the number of trades for a particular client per month. The creation time, measured from the time passed since midnight is set to a minimum of 7 as some clients, sending their orders on the evening before disrupt the feature distribution. 7am in this case involves all orders sent before market opening.

Electronic copy available at: https://ssrn.com/abstract=3997109

B Plots – Parent order model

Additional fits for suggested distributions for a high market cap stock (left), high volume stock (middle) and low volume stock (right).

B.1 C-Day VWAP

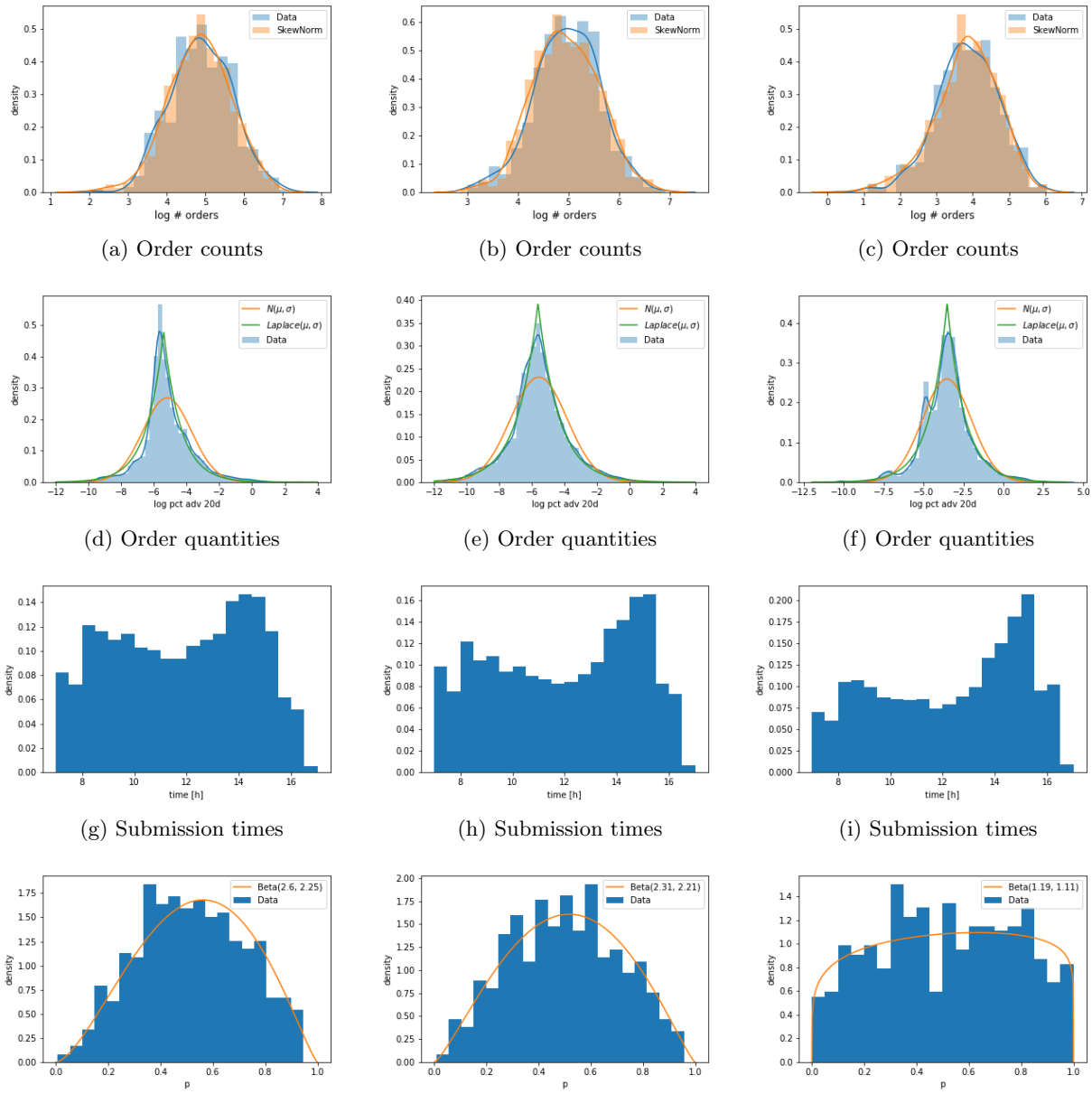


(a) Order quantities (b) Order quantities (c) Order quantities

Figure 15: Order quantities of the C-Day VWAP trader type for a high market cap stock (left), high volume stock (middle) and low volume stock (right).

Electronic copy available at: https://ssrn.com/abstract=3997109

B.2 C-Quant



(a) Order counts (b) Order counts (c) Order counts

(d) Order quantities (e) Order quantities (f) Order quantities

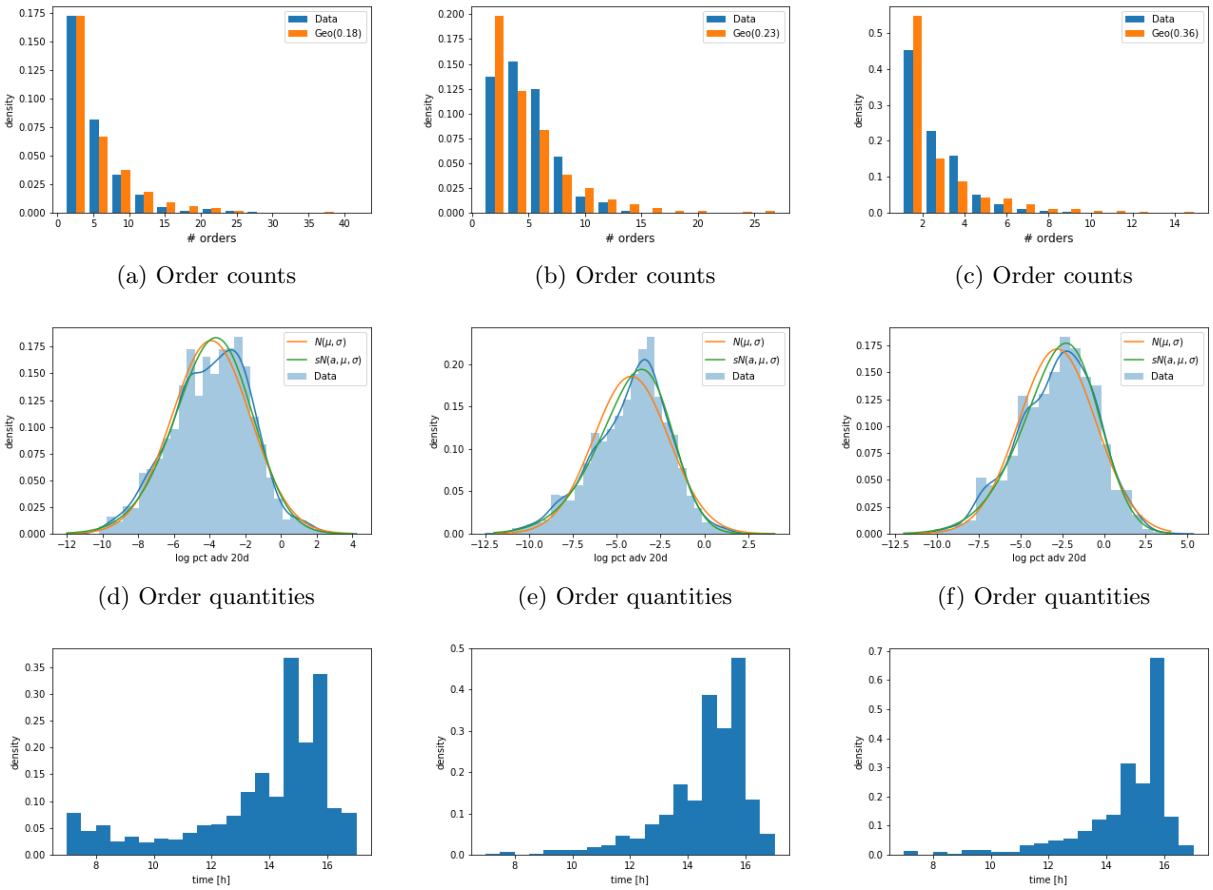
(g) Submission times (h) Submission times (i) Submission times

(j) Fraction of buy orders (k) Fraction of buy orders (l) Fraction of buy orders

Figure 16: Order counts, order quantities, submission time and fraction buy orders distribution of the C-Quant trader type for a high market cap stock (left), high volume stock (middle) and low volume stock (right).

Electronic copy available at: https://ssrn.com/abstract=3997109

B.3 C-Signal



(a) Order counts (b) Order counts (c) Order counts

(d) Order quantities (e) Order quantities (f) Order quantities

(g) Submission times (h) Submission times (i) Submission times

Figure 17: Order counts, order quantities, submission time and buy ratio distribution of the C-Signal trader type for a high market cap stock (left), high volume stock (middle) and low volume stock (right).

Electronic copy available at: https://ssrn.com/abstract=3997109

C Disclaimer

Opinions and estimates constitute our judgement as of the date of this Material, are for informational purposes only and are subject to change without notice. This Material is not the product of J.P. Morgan's Research Department and therefore, has not been prepared in accordance with legal requirements to promote the independence of research, including but not limited to, the prohibition on the dealing ahead of the dissemination of investment research. This Material is not intended as research, a recommendation, advice, offer or solicitation for the purchase or sale of any financial product or service, or to be used in any way for evaluating the merits of participating in any transaction. It is not a research report and is not intended as such. Past performance is not indicative of future results. Please consult your own advisors regarding legal, tax, accounting or any other aspects including suitability implications for your particular circumstances. J.P. Morgan disclaims any responsibility or liability whatsoever for the quality, accuracy or completeness of the information herein, and for any reliance on, or use of this material in any way.

Electronic copy available at: https://ssrn.com/abstract=3997109