Cryptocurrencies & Limit Order Books: Data, Stylised Facts and Model Building

Candidate Number: 1027984

April 4, 2021

Abstract

We investigate the limit order book of the OMG token market on the cryptocurrency exchange Coinbase Pro. We perform a statistical analysis and find consistency of the data with several well known stylised facts. The data allows us to identify the majority of traders as algorithmic traders. We can further classify them into two major classes trading either with constant volume or with constant funds.

The limit order book also exhibits a dynamical queue structure similar to the queue structure found for large tick stocks. These queues exhibit distinct statistical behaviour. In particular we find that the different classes of trading algorithms each place their orders predominantly in specific queues.

Finally we present a simple queuing model based on a Markov process. The invariant distribution of the model allows us to recover the volume distribution in the queues with reasonable accuracy.

Contents

1 Introduction

Limit Order Books are one of the most common ways to organise trading of financial assets. Many of today's electronic markets use limit order books to facilitate trade. Traders can place buy orders and sell orders at a price and volume of their choice on the electronic market. The order is then (partially) executed at the best available price. If the order is not (completely) executed it is kept on record in the limit order book as a so called *limit order* (hence the name limit order book) until it is executed or cancelled. Limit order book thus allow for a microscopic view into the basic mechanism of price finding.

In the present thesis we investigate the limit order book of the OMG token market on the cryptocurrency exchange Coinbase Pro. Cryptocurrencies and tokens are decentrelised digital assets which are based on a cryptographic tools such as distributed ledgers, block chains and digital signatures. The most prominent example of a cryptocurrency is probably Bitcoin, see [BTC08] for the original white paper written in 2008 by a still unknown author under the pseudonym Satoshi Nakamoto. We focus on the OMG token since the majority of recent studies focuse either on cryptocurrency indices or on the major cryptocurrencies such as Bitcoin, Ether or Litecoin, see literature review below.

Most cryptocurrency exchanges use limit order books to facilitate trade and many of them allow real time access to their trading data. The trading data offered by Coinbase Pro is particularly detailed and offers therefore an excellent opportunity to investigate the microstructure of the limit order book.

In section 2 we give a short overview of limit order books and their mathematical modelling. An explicit example how orders are placed into the limit order book is given in appendix A.

We review the characteristics of the OMG token and the cryptocurrency exchange Coinbase Pro in sections 3 and 4. We put articular emphasis on the fee structure of the exchange which appears to be a major incentive for algorithmic trading. The real time Level 3 data from Coinbase Pro which we collected from September to November 2021 allows us to investigate the life cycle of an order. In Level 3 data a unique order identification number identifies each order from its arrival at the exchange to its termination by cancellation or execution. This depth of information gives us the opportunity to investigate the fine structure of algorithmic traders and to identify certain types of trading bots.

We collect limit order book data directly from Coinbase Pro and perform in 5 a statistical analysis. The stylised facts that we observe in sections 5.2 and 5.3 are consistent with observations reported in other empirical studies for other assets and markets. For an extensive overview on limit order books, their statistical properties and mathematical modeling we refer to [GPWMFH13] and to [C01] for a review covering the central aspects of stylised facts.

In section 5.5 we observe that the depth profile of the limit order book has a queue-like structure close to the bid price and the ask price. We can identify four queues that are almost identical for the bid-side and the ask-side. From the change of spread distribution, section 5.4, we conclude that

most orders do not change the spread and therefore keep this queue structure stable.
The data quality and in particular the information provided by the Level 3 data allows us to devide traders into two main classes in section 5.6. We identify traders that trade with constant volume and traders that trade with constant funds, both following in general the best price. In section 5.7 we find evidence that the traders are probably automated, algorithmic traders and that their trading behaviour is intimately linked to the queue structure found in section 5.5.
Motivated by the preceding observations we employ a simple queuing model to the data in section 6. We follow closely the authors of [HLR15] who propose a queuing model based on a Markov process. The invariant distribution of the Markov process models the volume distribution at fixed price levels close to the best bid and best ask price. The model is related to the stochastic models presented in [CST10]. For large tick stocks the authors of [HLR15] find their model in good agreement with the empirical data. We apply the model to the queues we identified in section 5.5 and find good agreement with the data for the second, third and fourth queue but only relatively poor agreement for the first queue. We attribute this poor agreement to the sensitivity of the first queue to price change.
We conclude this thesis in section 7 with some concluding remarks and a list of open questions as well as some ideas how to proceed further based on our results.

Literature Review: The free availability of data from cryptocurrency exchanges allowed to analyse a new market with these new types of assets.
The authors of [HHR19] review the general mechanics of cryptocurrencies, presenting a high level exposition of the blockchain and cryptocurrency exchanges. They also provide summary statistics of cryptocurrency markets in general and some potential research directions. For a general overview of cryptocurrency trading with emphasis on technical aspects of trading and related research see for example [FVBKKMW20].
In [BBHN17] the authors study long range effects using daily to 5-hour Bitcoin data from 2011 to 2017. These results have been confirmed by [PCP18] where daily return data from 244 cryptocurrency indices have been investigated. Similar results have been found by [ZWLS18] and more recently by [HPR19].
A comprehensive study of stylised facts for the Bitcoin market on the exchanges BitFinex, Bitstamp and Coinbase can be found in [SRK19]. The authors retrieve real time data from December 2017 until October 2018 directly from the exchanges. They recover the majority of well known stylised facts.
In [PRH20] the authors investigate the question whether cryptocurrency markets as a whole are dominated by human traders or by autonomous algorithmic traders. They conclude that this is at least on average the case when considering intraday patterns in the CRIX (CRyptocurrency IndeX) as an indicator for human traders. This is in contrast to the authors of [SRK19] who find no such patterns for the Bitcoin market.

2 Limit Order Books

A limit order book (LOB) is a widely used protocol to organise trading of an asset for multiple traders. Traders can place different types of orders into the LOB. These orders are subject to a set of rules that govern the way how they are placed into the LOB. We will focus on a set of rules for LOBs which match the rules of the crypto exchange Coinbase Pro.

2.1 Limit Order Books

LOBs split into two separate sectors. The bid side contains all sell offers and the ask side contains all buy offers. A sell (buy) offer consists of a price and a volume of the asset which is to be sold (bought).

Definition 2.1 The set of all offers on an electronic exchange for a given asset at time $t \in \mathbb{R}$ is called *limit order book* or *LOB* and is denoted as $\mathcal{L}(t)$. The set of all sell offers is denoted as $\mathcal{B}(t)$ and the buy offers are denoted as $\mathcal{A}(t)$ such that $\mathcal{L}(t) = \mathcal{B}(t) \cup \mathcal{A}(t)$.

There are two types of orders that can be placed into a LOB. *Limit orders* specify a volume of the asset which shall be bought (sold) for at most (at least) a predefined price. *Market orders* specify a volume of the asset (an amount of funds) which shall be sold (for which the asset shall be bought). Trading on LOBs occurs in price and size increments which are specified by the exchange.

Definition 2.2 The *(incremental) lot size* $v_0 > 0$ is the smallest possible increment of traded volume. There may also be a minimal volume $v_{min} \geq 0$ which an order must at least contain. The *tick size* $p_0 > 0$ is the smallest possible price increment.

Limit and market orders are either buy or sell orders but they are not placed directly into the order book. They are preprocessed by the exchange and result in *matched orders* if they can be (partially) executed or in *open orders* if they cannot be completely executed. Open orders constitute the open buy or sell offers and are placed on the sell or bid side of the LOB accordingly. Trading takes place 24/7 without interruptions and open orders are kept indefinitely in the LOB, i.e. there is no predefined maximal life time of an open order.
Market orders are always executed and may result in several matched orders. Limit orders may be completely, partially or not executed, depending of the state of the LOB. They can therefore result in several matched orders and up to one open order.
The life cycle of an order ends if the order is completely executed. For market orders this is always the case and for limit orders this is the case if the order is either completely executed upon arrival or subsequently executed as an open order by incoming orders. The second way for an order to end its life cycle is being cancelled by the traders that placed the order. We call an order which has not been completely executed or cancelled an *active order*. All active orders at a given time constitute the limit order book, i.e. we have the following definition.

Definition 2.3 An *active order* is an n-tuple $x(t) := (\epsilon_x, p_x, v_x, t)$ with $t \in \mathbb{R}$. The *sign* of the active order $\epsilon_x = \pm 1$ indicates whether it is on the ask side ($\epsilon = +1$) or on the bid side ($\epsilon = -1$). The *price* $p_x \in \{kp_0 > 0 \,|\, k \in \mathbb{N}\}$ is a multiple of the tick size and the *volume* $v_x \in \{kv_0 > 0 \,|\, k \in \mathbb{N}\}$ is a strictly positive multiple of the lot size.
The set of all active orders constitutes the limit order book (LOB) $\mathcal{L}(t) := \{x(t) \,|\, x(t) \text{ is an active order}\}$. We have for the bid side of the LOB $\mathcal{B}(t) := \{x \in \mathcal{L}(t) | \epsilon_x = -1, \, x \text{ is active}\}$ and for the ask side $\mathcal{A}(t) := \{x \in \mathcal{L}(t) | \epsilon_x = +1, \, x \text{ is active}\}$ such that $\mathcal{L}(t) = \mathcal{B}(t) \cup \mathcal{A}(t)$.

Incoming orders change the LOB. But only open orders, matched orders and cancel orders have a direct effect on the LOB. Open orders add volume at a given price level of the LOB, while matched orders and cancel orders subtract volume. In this sense matched orders and cancel orders behave identically. We will call those orders which increase volume at a given price level (open orders) and those which decrease volume at a given price level (matched orders, cancel orders) *events* and model them as follows.

Definition 2.4 An *event* is an n-tuple $x := (\epsilon_x, p_x, v_x, t_x)$ where $\epsilon_x = \pm 1$ is the *sign* of the event indicating whether it is on the ask side ($\epsilon = +1$) or on the bid side ($\epsilon = -1$). The *price* $p_x \in \{kp_0 > 0 \,|\, k \in \mathbb{N}\}$ is a multiple of the tick size and the *volume* $v_x \in \{kv_0 \neq 0 \,|\, k \in \mathbb{Z}\}$ is a non-zero multiple of the lot size. The *submission time* $t_x \in \mathbb{R}$ is assumed to be continuous.

Events $x := (\epsilon_x, p_x, v_x, t_x)$ act on the $\mathcal{L}(t)$ by adding (or removing) volume v_x at price p_x on the LOB side given by the sign ϵ_x at time t_x. If there is an active order $y(t_x) \in \mathcal{L}(t_x)$ with $\epsilon_y = \epsilon_x$ and $p_y = p_x$ then the active order is updated to $y'(t \geq t_x) := (\epsilon_y, p_y, v_y + v_x, t)$ and we have $y(t) \in \mathcal{L}(t)$ for $t < t_x$ and $y'(t) \in \mathcal{L}(t)$ for $t \geq t_x$. The active order $y(t)$ gets removed from $\mathcal{L}(t \geq t_x)$ if $v_y + v_x = 0$. If there is no order $y(t_x) \in \mathcal{L}(t_x)$ with $\epsilon_y = \epsilon_x$ and $p_y = p_x$ then a new active order $x(t \geq t_x) := (\epsilon_x, p_x, v_x, t)$ is added to $\mathcal{L}(t \geq t_x)$.
Here we follow the convention that a change in $\mathcal{L}(t)$ becomes active at the arrival time t_x of the event x. So the resulting process $\mathcal{L}(t)$ is a *càdlág process* in t, i.e. continuous on the right and with limit existing on the left, see [BBDG18].

Remark 2.5 The trading rules of the exchange ensure that the volume $v_x > 0$ for each active order $x(t) \in \mathcal{L}(t)$ and that $\mathcal{B}(t) \cap \mathcal{A}(t) = \emptyset$ with $\max_{x \in \mathcal{B}(t)} p_x < \min_{x \in \mathcal{A}(t)} p_x$ for all $t \in \mathbb{R}$. See appendix A for an example how orders are placed into a LOB.

Definition 2.6 The best price at which the asset can be bought is given by the *bid-price* $b(t)$,

$$b(t) := \max_{x \in \mathcal{B}(t)} p_x.$$

The *ask-price* $a(t)$ is defined as the best price at which the asset can be sold,

$$a(t) := \min_{x \in \mathcal{A}(t)} p_x.$$

The *mid-price* $m(t)$ is the mean of the the bid-price and the ask-price

$$m(t) := \tfrac{1}{2}(a(t) + b(t))$$

and the *spread* $s(t)$ is the (strictly positive) difference between ask-price and bid-price,

$$s(t) := a(t) - b(t).$$

Note that $b(t) < a(t)$ for all t and that $a(t), b(t), m(t)$ and $s(t)$ are càdlág processes just as $\mathcal{L}(t)$.

3 Cryptocurrencies and Tokens

In this thesis the underlying cryptocurrency is the Ether coin, see [ETH18]. It is the proprietary cryptocurrency which is the monetary base of the Ethereum network. Creation and transfer of Ether is based on distributed ledgers living on the Ethereum block chain. The validation of transactions currently requires a proof-of-work similar to the Bitcoin but will change to a proof-of-stake verification in the near future. The proof-of-stake verification requires that the verifier of a transaction owns a certain amount of Ether, thus replacing trust built on computational work by trust built on wealth. As of 1. January 2021 Ether's transaction fee is ~ 3 \$ and its market capitalisation is ~ 40 billion \$, second only to Bitcoin.
A central feature of the Ethereum block chain is its ability to allow the creation of called tokens based on the ERC20 standard for smart contracts. Smart contracts are self executing programs living on the Ethereum network. Due to the flexibility of the ERC20 standard they can have a wide range of possible applications ranging from value storage over stock-like ownership tokens to voucher-like utility tokens. See section 2.5 in [HHR19] for an overview. We will focus on OMG utility token, see section 3.1.
Cryptocurrencies and tokens can be traded on cryptocurrency exchanges such as Kraken, Coinbase Pro, Bitfinex, etc. The present thesis we investigate trading of OMG tokens on the cryptocurrency exchange Coinbase Pro. Trading fees on cryptocurrency exchanges for coins or tokens held on accounts at the exchange are in general much lower than the above mentioned transaction fees, see section 4.1 since they are stored by the exchange. Only if the owner transfers the crypto asset to another location, the transaction has to be added to the distributed ledger on the corresponding block chain, thus resulting in the higher fees.

3.1 OMG Token

The OMG token is a utility token based on the ERC20 standard and lives on the Ethereum block chain. It serves as the basis for transaction verification by proof-of-stake on the OmiseGO network, see [OMG17]. The OmiseGO network aims to use cryptocurrencies to facilitate cross currency and cross country currency transaction with special focus on Southeast Asia. Money transfers between

countries and currencies on the OmiseGO network will have relatively low fees compared to ordinary bank transfers and do not require bank accounts. The security of the transfer is guaranteed by proof-of-stake of the verifiers on the OmiseGO network, i.e. by trustworthy participants holding sufficient amounts of OMG tokens. As a compensation for the verification of transactions the verifiers receive a fee.
The OMG tokens are fungible and are traded for example on Coinbase Pro. As of 1. January 2021 the market capitalisation of the OMG token is ~ 470 million \$.

4 Coinbase Pro

Coinbase Pro is one of the largest cryptocurrency exchanges. It is located in the United States of America in San Francisco CA. We chose Coinbase Pro for two main reasons. First, the exchange relatively well regulated. It is registered as a Money Services Business with FinCEN and is authorised by the Financial Conduct Authority under the Electronic Money Regulations 2011 (FRN: 900635) for the issuing of electronic money. Furthermore Coinbase Pro has filed a draft Form S-1 with the U.S. Securities and Exchange Commission (SEC). This formal draft is considered as a first step in becoming a registered stock exchange under US law.
Second, Coinbase Pro offers a public application programming interface (API) that allows to obtain real time data of LOBs in unprecedented detail of so called Level 3 order book updates. These Level 3 LOB updates allow the investigation of the full life cycle of an order from order submission to the exchange to cancellation or execution of the order.
Many cryptocurrency exchanges supply free historical price and trade data sets. Historical order book data sets can also be purchased with a resolution down to one minute. But there has been some criticism concerning the data quality of different sources [AD20]. We therefore opted to collect the data directly from the cryptocurrency exchange Coinbase Pro.
As of January 2021 36 cryptocurrencies and tokens can be traded on Coinbase Pro with an accumulated daily volume of ~ 5 billion \$. The OMG token can be traded since May 2020 in USD, GPB, EUR and Bitcoin and has on Coinbase Pro a traded volume of ~ 2.5 million \$ per day. We will use the data of the OMG-EUR pair in our investigation. For all tradable currency pairs Coinbase Pro offers a public trading dashboard, see for example [CBPomg] for the dashboard of the OMG Token traded in EUR.

4.1 Trading Rules and Fee Structure

Trading rules and in particular the fee structure are central to understand the functioning of the market on an exchange. The Coinbase Pro market works on the principles of a first-in-first-out (FIFO) LOB, see section 2. Order placement, execution and cancelation does non-discriminatory w.r.t. size of the order or trader placing the order. The official list of trading rules can be found in [CBPrules].

Pricing Tier	Taker Fee	Maker Fee
up to $ 10k	0.50%	0.50%
$ 10k - $ 50k	0.35%	0.35%
$ 50k - $ 100k	0.25%	0.15%
$ 100k - $ 1m	0.20%	0.10%
$ 1m - $ 10m	0.18%	0.08%
$ 10m - $ 50m	0.15%	0.05%
$ 50m - $ 100m	0.10%	**0.00%**
$ 100m - $ 300m	0.07%	**0.00%**
$ 300m - $ 500m	0.06%	**0.00%**
$ 500m - $ 1b	0.05%	**0.00%**
$ 1b +	0.04%	**0.00%**

Figure 1: Fee Structure valid September-November 2020, see [CBPfees] for the current fees.

The fee structure of Coinbase Pro is designed to encourage market makers to supply the market with liquidity by placing limit orders into the LOB that are not executed. The table for the fees structure valid from September 2020 to November 2020 in figure 1 shows the fees for executed orders (taker fees) and for limit orders (maker fees). The pricing tier represents the value of the accumulated volume over a trailing 30 day period placed on the exchange by a trader. Cancelling an existing order is free of charge.

Placing limit orders with a large volumes or at a high frequency thus results in drastically reduced (taker and maker) fees. This appears to be one plausible reason why the OMG market seems to be dominated by algorithmic traders or trading bots, as we will see below. Due to the open API of Coinbase Pro it is relatively easy for anyone to build a trading bot and act as a market maker thus reducing also the maker fees for selling and buying cryptocurrencies. Many tutorials how to build such trading bots are widely available on the internet as well as commercial offers.

4.2 OMG Token Data and Data Quality

We analyse the LOB of the OMG token traded in Euro, see [CBPomg] for the real time dashboard. The tick size of the OMG token is $p_0 = 0.0001$ € and the minimal traded volume is $v_{min} = 1$ OMG token with an (incremental) lot size size of $v_0 = 0.1$ OMG tokens. We measure the volume of orders, depth of the LOB, etc. in units of OMG tokens and write for example $v_{min} = 1$ OMG.

To ensure good data quality it is advantageous to stop the websocket and restart the data collection in one hour intervals. At each reconnection the websocket sends a snapshot of the full LOB as an initial condition followed by LOB updates as the occur on the exchange. But due to instabilities of the internet connection, slow reaction times of the technical setup or failures of the exchange,

LOB updates as well as LOB snapshots may be lost. See appendix B for a brief exposition of the technical setup we use to collect the data.
Data has been collected from September 9^{th} 2020 until November 1^{st} 2021. Full data consisting of LOB snapshots and LOB updates is collected in three one hour intervals each day between 4pm - 7pm CET to minimise these failures. This covers the busiest trading time on Coinbase Pro. During one hour of LOB updates $\sim$35 MB of data is collected. Therefore we decide to limit ourselves to collecting three hours of full data per day, i.e. $\sim$115 MB per day, to keep the amount of data manageable.
Each one hour interval contains at least one LOB snapshot which serves as an initial condition. It may happen that the exchange restarts the broadcasting during a one hour interval. This splits the interval into shorter intervals of less than one hour. For our statistical analysis this occasional splitting into shorter intervals has no consequences. We collected 162 intervals of full LOB data of one hour or less.
Out of the 162 intervals we find 16 intervals which exhibit negative volume or spread smaller than one tick at some instance of time. These 16 intervals are discarded from our analysis leaving us with 146 valid one our intervals of full LOB data.
These failures probably occur if orders are lost in the time between broadcasting of the LOB snapshot and the first LOB update. Usually the lists of sequence IDs of all orders are complete, i.e. no orders are missing after the first update has been received. Unfortunately Coinbase Pro does not broadcast the order ID of the first order arriving after the LOB snapshot. So we have to compromise and use the non-negativity of volume and the positivity of the spread as further measure for the validity of a LOB interval.
The 146 valid LOB intervals comprise $\sim 13 \cdot 10^6$ orders of all types, see figure 2, which gives an average of 25 orders arriving per second with up to 10 orders per millisecond at peak times.

Remark 4.1 Let T_i, $i = 1, ..., 146$ be the time intervals for which a valid LOB can be constructed. We will write these valid LOBs as $\mathcal{L}(T_i)$. Note that $\max(T_i) < \min(T_{i+1})$ and that $\max(T_i) - \min(T_i) \leq 60$ minutes.

To get further independent LOB information we collect snapshots of the full LOB in intervals of one minute during the time from 7pm - 4pm CET. The amount of this data is $\sim$45 MB per day.

5 LOB Statistics and Stylised Facts

In the following we analyse some statistical properties of the LOBs and the incoming orders from the OMG data. Some of these statistical properties confirm well known *stylised facts*, i.e. persistent qualitative and quantitative statistical patterns which emerge when observing time series of different asset classes in financial markets. For a comprehensive overview of stylised facts we refer to [C01] and [BBDG18].

	limit	market	open	matched	cancelled	change
# bid side	2241808	1045	2201103	21129	2196581	14
# ask side	2059134	1672	2040379	41194	2016981	14
total #	4300942	2717	4241482	62323	4213562	28

Figure 2: Absolute numbers per order type collected in 146 one hour LOB intervals from September 9^{th} until November 1^{st} 2020.

In order to guarantee a sound statistical interpretation of the statistical properties of a given time series $X(t)$, we need in general that $X(t)$ is stationary, see [BBDG18]. By stationarity we mean that for any set of times t_1,...,t_n the joint distribution of $X(t_1)$,...,$X(t_n)$ coincides for any $\tau > 0$ with the joint distribution of $X(t_1 + \tau), ..., X(t_n + \tau)$. Showing that a given time series has this property is often not easy due to periodic effects and it may require a proper definition of the time scale [CS01]. We therefore assume that any time series is indeed stationary.
To be able to calculate expectations we assume furthermore ergodicity for any time series $X(t)$, i.e. that the ensemble average $\mathbb{E}[X(t)] = \mu$ is independent of t and coincides with the time average $\langle X(t) \rangle$,

$$\langle X(t) \rangle := \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} X(t) \mathrm{dt}.$$

Ergodicity is typically satisfied by iid random variables, see [CS01] but may fail in general.
For the statistical analysis in this thesis we use Python and here in particular the standard packages numpy.py and statistics.py as well as Pandas data frames.
As an illustration of the time series we obtain from the OMG data, we display the mid-price and the spread for a time period of one hour on September 29^{th} in figure 3.
We compute all statistical properties that involve time averages separately for the time intervals T_i, $i = 1, ..., 146$ which allow the construction of valid LOBs $\mathcal{L}(T_i)$, see remark 4.1. Then we take the arithmetic average over the resulting 146 time averages. This applies in particular to the computation of the various autocorrelation functions in the following subsections.

5.1 Market Orders and Matched Orders

Let us first investigate the distribution of market orders and matched orders compared to limit and open orders. In figure 4 we show the ratio of the number of market orders to limit orders and of matched orders to open orders for each day of the observation period.
Apart from October 20^{th} market play an extremely insignificant role w.r.t. the total number of orders amounting to less than 0.1 % of all the incoming orders. In comparison the ratio of matched

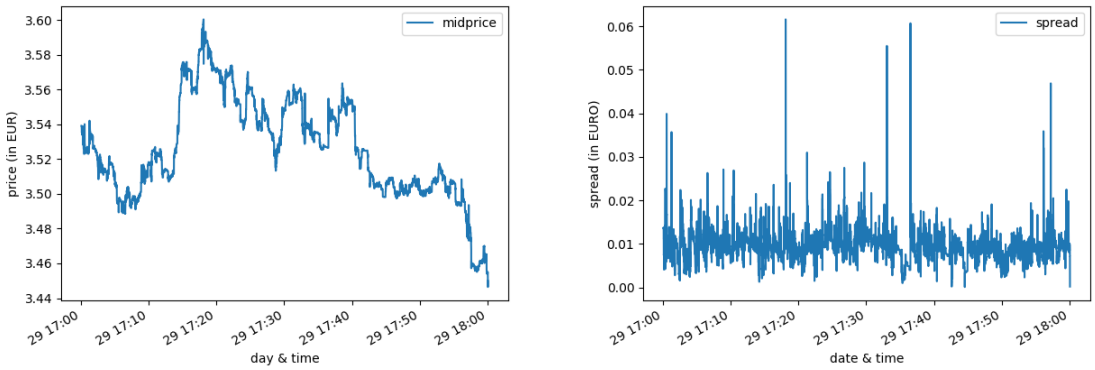


Figure 3: Mid-price (left) and spread (right) of OMG token in EUR

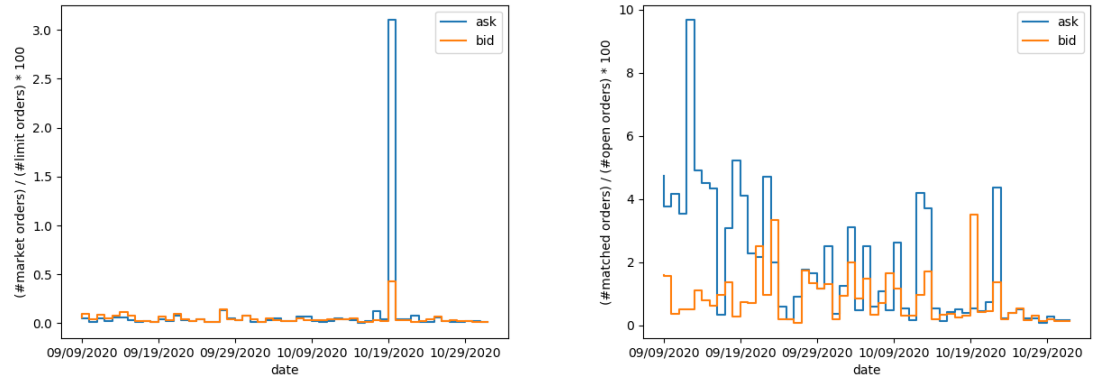


Figure 4: Daily ratio market to limit orders (left) and matched to open orders (right)

to open orders is roughly one order of magnitude larger. On average each market order generates 1.75 matched orders.

Level 3 updates allows us to determine via the order ID which matched orders are linked to market orders. We find that matched orders originating from market orders make up $\sim 8\%$ of all matched orders, i.e. $\sim 92\%$ of matched orders derive from limit orders placed on the opposite side of the order book. Yet, market orders account for $\sim 21\%$ of the matched order volume. This is reflected in the mean volume of 171.0 OMG for the generated matched orders while matched orders originating from limit orders have a mean volume of 52.8 OMG.

The anomaly on October 20^{th} originates from an extremely large number of ~ 857 market orders of volume 1 OMG out of a total of 909 market orders this day. The relatively small mean volume of 22.8 OMG per market order this day explains on the other hand the absence of the anomaly on the ratio of matched to open orders.

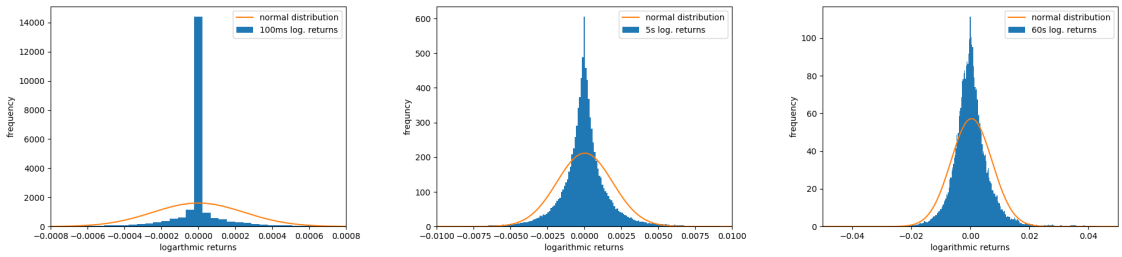


Figure 5: Distribution of mid-price logarithmic returns $r^m(t, \Delta)$ with $\Delta = 100$ms (left), $\Delta = 5$s (middle) and $\Delta = 60$s (right).

5.2 Returns and Volatility

Next we investigate some statistical properties of the logarithmic returns. We concentrate on the returns of the mid-price.

Definition 5.1 Let $\Delta > 0$ be a time interval. The *logarithmic return* of $m(t)$ at time scale Δ is defined as

$$r^m(t, \Delta) := \ln\left(\frac{m(t+\Delta)}{m(t)}\right) = \ln m(t+\Delta) - \ln m(t). \tag{1}$$

The logarithmic returns for the ask-price and bid-price can be defined in the same manner.

To investigate the return time series we first resample the mid-price time series $m(t)$ obtained from OMG data with a sampling rate of $\delta s = 10$ milliseconds. We thus obtain an evenly spaced time series.
The distribution of the logarithmic returns for time scales $\Delta = 10\delta s = 100$ms, $\Delta = 500\delta s = 5$s and $\Delta = 6000\delta s = 60$s are shown in figure 5. We superimpose onto each distribution a best fit normal distribution from which suggests that the distribution of the logarithmic returns has a power-law tail:

Definition 5.2 Let X be a random variable and $F_X(u) := \mathbb{P}[X \leq u]$ its (unconditional) distribution function. If there exists an $\alpha > 0$ such that

$$F_X(u) \sim \mathcal{O}(z^{-\alpha}) \quad \text{as} \quad u \to \infty$$

then F_X is said to have a *power-law tail* with *tail index* α.

To see the power-law tails more clearly we plot the absolute value of the logarithmic returns on a logarithmic scale, see figure 6, and fit them with power law functions with tail index $\alpha = 4$ for $\Delta = 100$ms, $\alpha = 3.7$ for $\Delta = 5$s and $\alpha = 3$, for $\Delta = 60$s. The fit is certainly not conclusive

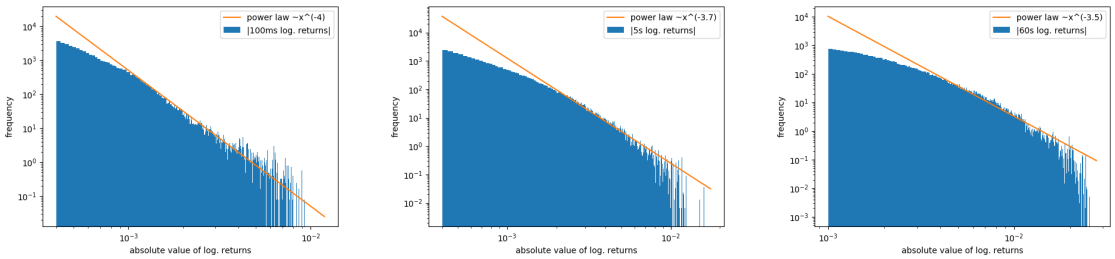


Figure 6: Distribution of mid-price log-returns $r^m(t, \Delta)$ with $\Delta = 100$ms (left), $\Delta = 5$s (middle) and $\Delta = 60$s (right) in EUR.

and gives only a strong suggestion in favour of a power-law behaviour and seems to agree with the generally observed tail index $2 \leq \alpha \leq 5$, see [C01] and [BBDG18].
This observation suggests that the short time logarithmic returns of the OMG token have a distribution with heavy tails.

Definition 5.3 Let X be a random variable and $F_X(u) := \mathbb{P}[X \leq u]$ its (unconditional) distribution function. The distribution is called heavy tailed iff

$$\int_{-\infty}^{\infty} e^{tu} dF_X(u) = \infty \quad \text{for all } t > 0.$$

In particular random variables with power-law behaviour fall into the class of heavy tailed distributions. Return series with heavy tailed distributions are very common and has been observed in a multitude of assets, see [C01].

From the above observations it seem reasonable to suggest that the short time returns of display the following stylised fact:

Stylised Fact 5.1 The distribution function of logarithmic returns $F_{r^m(t,\Delta)}(u)$ for short time scales of less than a minute is heavy tailed and can be described by a distribution with power-law tail. For the empirically observed tail index we find $2 \leq \alpha \leq 5$, in agreement with observations for longer time scales, see [C01].

It is also apparent that the distributions in figure 5 tend to resemble the normal distribution more closely for increasing Δ. To get a simple quantitative measure we calculate the normalised kurtosis

$$\kappa(\Delta) := \frac{\langle (r^m(t,\Delta) - \langle r^m(t,\Delta) \rangle)^4 \rangle}{\mathrm{var}(\mathrm{r^m(t,\Delta)})^2} - 3.$$

of the logarithmic returns. We find $\kappa(\Delta = 100\mathrm{ms}) = 106.5$, $\kappa(\Delta = 5\mathrm{s}) = 8.8$ and $\kappa(\Delta = 60\mathrm{s}) = 8.8$. This certainly does not allow to conclude that the kurtosis goes to zero for large Δ. Therefore we can only conjecture that aggregational Gaussianity is also displayed by the logarithmic returns.

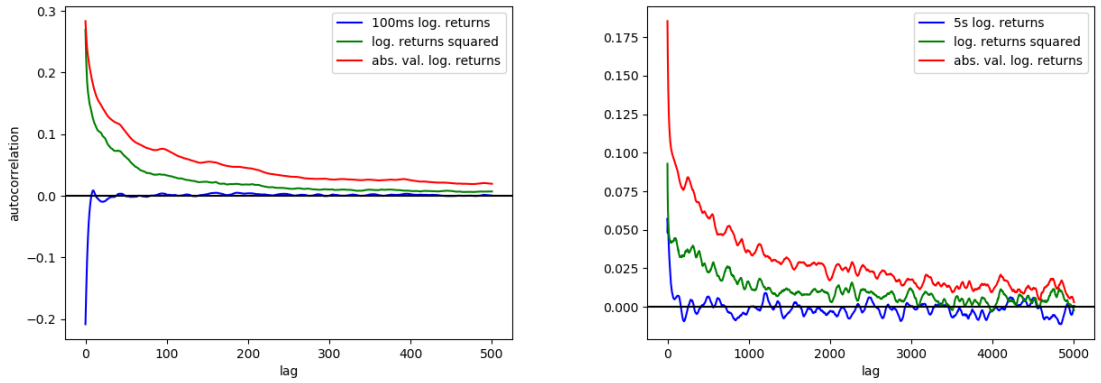


Figure 7: Autocorrelation of mid-price logarithmic returns $r^m(t, \Delta)$, its absolute value $|r^m(t, \Delta)|$ and ist square $|r^m(t, \Delta)|^2$ with $\Delta = 100$ms and lag increment $\delta\tau = 10$ms (left) as well as $\Delta = 5$s and lag increment $\delta\tau = 100$ms (right).

i.e. that for large time scales as $\Delta \to \infty$ logarithmic returns approach a normally distributed random variable as $r^m(t, \Delta) \sim \mathcal{N}(\mu, \sigma)$.
Finally we investigate the autocorrelation function of the logarithmic returns and their volatilities.

Definition 5.4 Let $\tau > 0$ be a time interval. For a time series $X(t)$ the *autocorrelation function* with lag τ is defined as

$$C(X(t), \tau) := \frac{\text{cov}[X(t), X(t+\tau)]}{\text{var}[X(t)]}$$

given that the covariance $\text{cov}[\cdot, \cdot]$ and and the variance $\text{var}[\cdot]$ are well defined for the time series. If a lag increment $\delta\tau > 0$ is given such that $\tau = n \cdot \delta\tau$, $n \in \mathbb{N}$, then we write $C(X(t), n)$ instead of $C(X(t), n \cdot \delta\tau)$.

As a measure for the (realised) volatility of the logarithmic returns we choose the square $r^m(t, \Delta)^2$ and the absolute value $|r^m(t, \Delta)|$ of the time series. We calculate the autocorrelation functions for $r^m(t, \Delta)$, $r^m(t, \Delta)$ and $|r^m(t, \Delta)|$ with $\Delta = 100$ms and $\Delta = 5$s for each of the 146 one hour LOB intervals and take the average. In figure 7 one can clearly see that the autocorrelation of the returns drops for $\Delta = 100$ms and $\Delta = 5$s quickly to zero while both volatility measures decay slowly.
This phenomenon is known as *volatility clustering* and is also a well known stylised fact, see [C01]:

Stylised Fact 5.2 The autocorrelation of volatility measures is positive over long periods of time. So high-volatility events tend to come in clusters.

Furthermore we see in figure 8 that the autocorrelation functions of the volatility measures of the logarithmic returns also show power-law and therefore heavy tails. Their tail index appears to be $0 < \alpha < 1$ with decreasing α for increasing Δ.

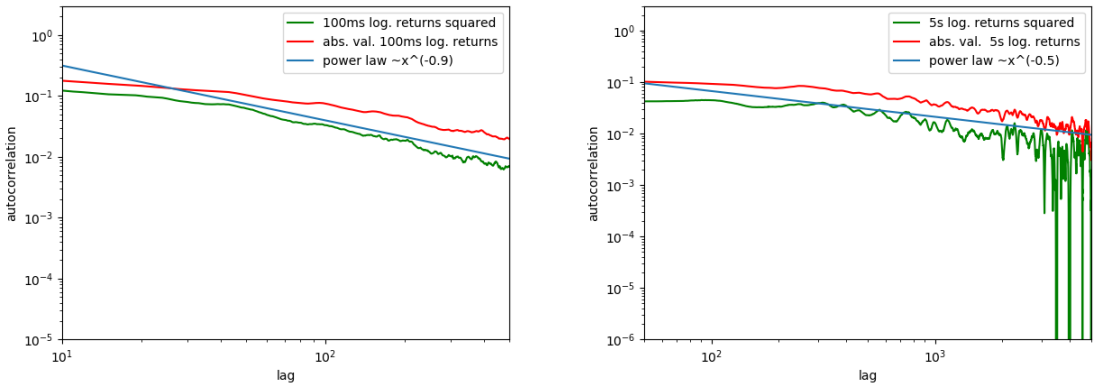


Figure 8: Autocorrelation of the absolute value of mid-price logarithmic returns $|r^m(t,\Delta)|$ and the square of mid-price logarithmic returns $|r^m(t,\Delta)|^2$ with $\Delta = 100$ms and lag increment $\delta\tau = 10$ms (left) as well as $\Delta = 5$s and lag increment $\delta\tau = 100$ms (right).

Stylised Fact 5.3 The autocorrelation functions of volatility measures for absolute returns display a heavy tails. We find that the tail index is $0 < \alpha < 1$, corresponding to long memory processes, [PB03] and [LF04].

5.3 Order Series

Now we define the order sign series, see for example [BBDG18].

Definition 5.5 The order sign series of a given order type is $\{S_1, S_2, ...\}$ where for $j = 1, 2, ...$

$$S_j := \begin{cases} -1 \text{ if the } j^{th} \text{ order of the given type is a buy order (bid side)} \\ +1 \text{ if the } j^{th} \text{ order of the given type is a sell order (ask side)} \end{cases}$$

The autocorrelation function of the order sign series for limit orders and cancel orders is shown in figure 9. We see again that the series display a power law behaviour with tail index $\alpha \approx 0.35$ and we can confirm the following stylised fact [PB03].

Stylised Fact 5.4 The autocorrelation function of limit and cancel order sign series display power-law tails with empirically observed tail index $0 < \alpha < 1$ corresponding to long memory processes, [PB03] and [LF04].

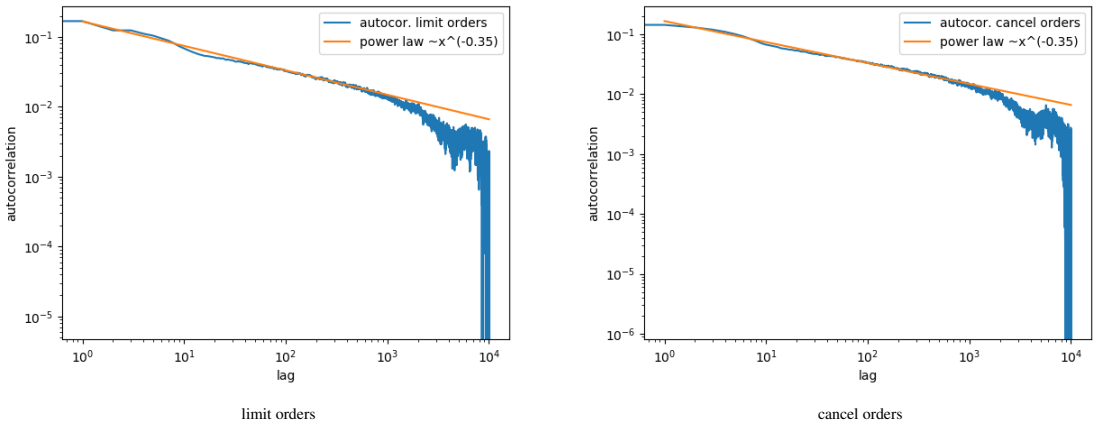


Figure 9: Autocorrelation order sign series for limit orders and cancel orders

5.4 Spread and Change of Spread

The distribution of the spread and the change of spread is a partial motivation for the choice of the queueing model which we will use the volume distribution of the OMG order books. We start by resampling the spread time series $s(t)$ to the moments in time $t_j^c \in \mathbb{R}$, $j = 0, 1, 2, ...$ when LOB changing events arrive, i.e. an open order, a matched order or a cancel order. At each instant t_k^c we have the spread $s(t_k^c)$ and calculate the *change of spread* $c^s(t_k^c) := s(t_k^c) - s(t_{k-1}^c)$ for $k \geq 1$. Both are measured in units of EUR.

The distribution of the spread has its maximum at 0.006 € and drops off exponentially to either side with a second local maximum at 0.0002 €, i.e. at a spread of two ticks $p_0 = 0.0001$, see left panel in figure 10. The arithmetic mean of the spread is 0.0063 €, i.e. $63p_0$ times the tick size. The arithmetic mean of the mid-price over the time period under consideration is $\langle m(t) \rangle = 2.9$ €, so the OMG token can be categorised as a *small tick asset*.

Of particular importance for the applicability of the queueing model is the relative scarce change of spread if a LOB changing event arrives. As can be seen in the right panel of figure 10, the change of spread is sharply peaked around $0 \pm 2p_0$ and drops off exponentially from there on. Indeed 86.9% of the incoming events do not change the spread at all and the events which change the spread be $\pm 2p_0$ account for 93.4% off all events. This relatively static behaviour of the spread w.r.t. incoming events could partially explain of the structures which appear in the volume profile of the order book close to zero relative price as described in the next section.

5.5 LOB Profiles and Queues

We use the LOB snapshots collected in one minute intervals from 19:00 to 16:00 CET to obtain the distribution of the average bid- and ask-side relative depth profiles with respect to the *relative price*,

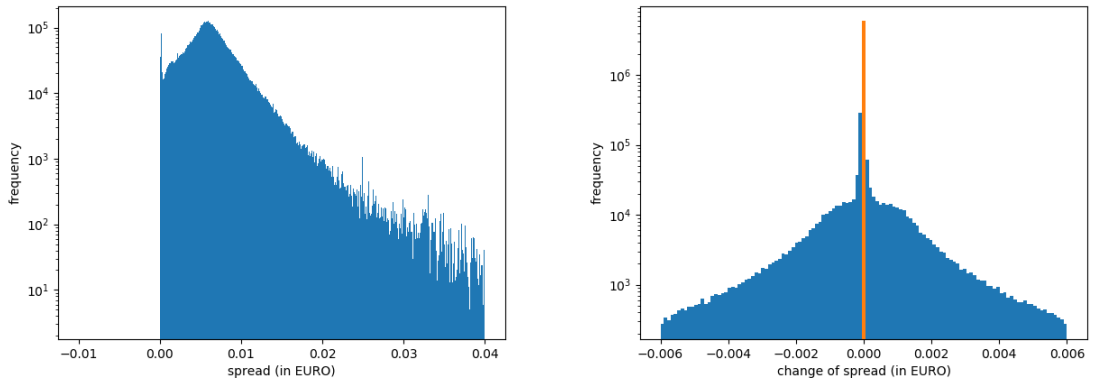


Figure 10: Non-normalised distribution of the spread $s(t^c)$ in EUR (left) and of the change of the spread $c^s(t^c)$ in EUR (right)

which is defined as follows:

Definition 5.6 The *ask-relative price* of an order $x \in \mathcal{L}(t)$ as $\delta^a(p_x, t) := p_x - a(t)$ and the *bid-relative price* as $\delta^b(p_x, t) := b(t) - p_x$. If we refer to general orders (bid- or ask-relative) we will suppress the super-scripts a and b and simply write $\delta(p_x, t)$.
The *bid-side depth available at relative price* p (at time t) is

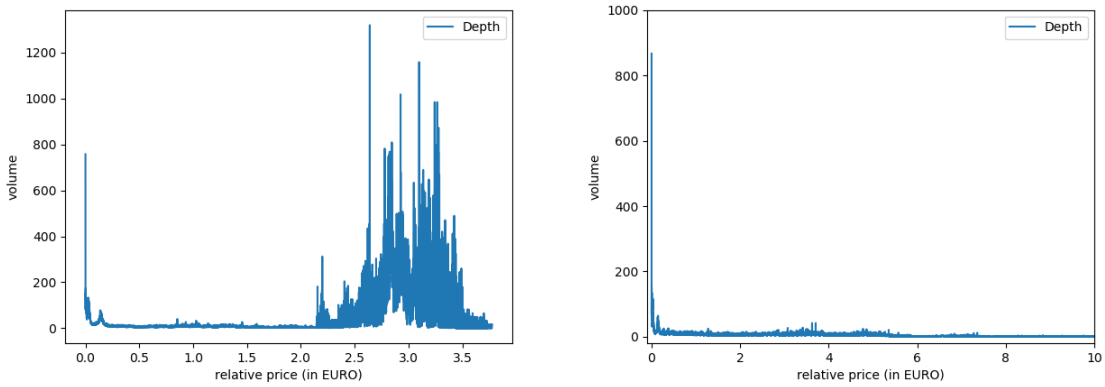
$$N^b(p,t) := \sum_{\{x \in \mathcal{B}(t) | \delta^b(p_x,t)=p\}} v_x$$

and $N^a(p,t)$ is defined similarly. Furthermore we define the *bid-side accumulated depth up to relative price* p (at time t) as

$$\Sigma N^b(p,t) := \sum_{\{x \in \mathcal{B}(t) | \delta^b(p_x,t)\leq p\}} v_x$$

and $\Sigma N^a(p,t)$ similarly.

The plot of the average bid- and ask-side relative depth profiles $(p, N^{b/a}(p,t))$ in figure 11. This clearly shows that most of the volume of the bid side is placed more than 2 € away from the best bid-price $b(t)$. The same is true for the ask side, yet it is less visible since the volume can be distributed more evenly as the possible price range of the ask side is not bounded from above at any given instant in time.
But the central observation for the present investigation the structure of the average relative depth profiles close to the the bid- and ask-relative price. In figure 12 we see the bid side and the ask side for $0 \cdot p_0 \leq \delta^{a/b} \leq 650 \cdot p_0$ with tick size 0.0001 €.
The first striking observation is the similarity of bid side and ask side. Apart from small variations the average relative depth profiles close $\delta^{a/b} = 0$ to are almost identical. We will therefore combine



bid-side ask-side

Figure 11: Average relative depth profiles of bid-side LOB and ask-side LOB

the bid side and the ask side when investigating statistical features close to $\delta^{a/b} = 0$.

The second striking observation is the distribution of maxima and in particular of the minima at $\delta^{a/b} \approx 6 \cdot p_0$, $\delta^{a/b} \approx 50 \cdot p_0$ and $\delta^{a/b} \approx 120 \cdot p_0$, see first three red vertical lines in figure 12. This distribution is a stable feature of the LOBs over the whole period of observation of 54 days.

It appears therefore natural to assume that this particular structure is induced by the particular trading patterns of the traders. It is a first hint that trading on the OMG-EUR market is dominated by algorithmic traders. We will call such traders *trading Bots* or simply *Bots* and will analyse them in more detail in the following sections.

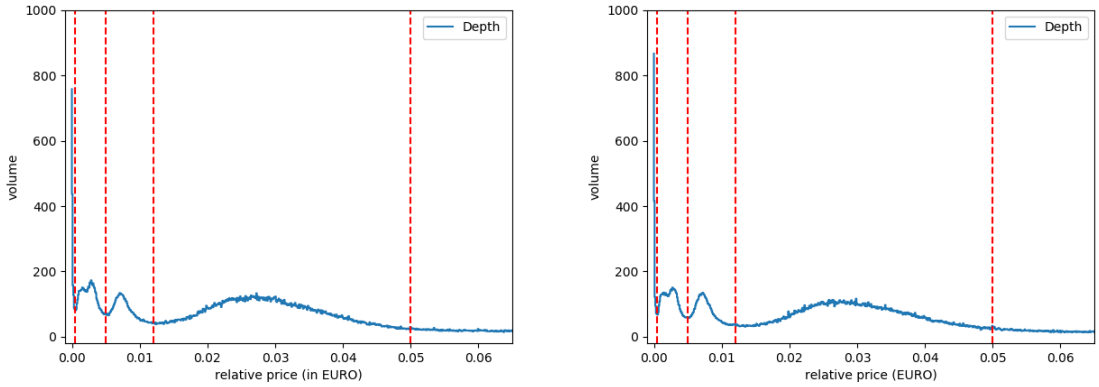
We will assume that the four intervals

$$\begin{aligned} Q_1 &= [0 \cdot p_0, 6 \cdot p_0] \\ Q_2 &= (6 \cdot p_0, 50 \cdot p_0] \\ Q_2 &= (50 \cdot p_0, 120 \cdot p_0] \\ Q_4 &= (120 \cdot p_0, 350 \cdot p_0] \end{aligned} \tag{2}$$

are essential for the dynamical properties of the OMG-EUR market on Coinbase Pro. Here we have chosen $\max(Q_4) = 350 \cdot p_0$ arbitrarily in such a way that the local maximum is comfortably included in the interval.

These intervals will serve as an equivalent of of one-tick sized queues for large tick stocks used in various queueing models as for example in [CST10] and [HLR15]. It is assumed that many important features of the dynamics of a LOB are mainly governed by the dynamics close to $\delta^{a/b} = 0$ and that we can employ models from queuing theory for their description.

For large tick stocks one usually assumes that the queues are defined simply by the tick size. We assume instead that the intervals Q_i, $i = 1, ..., 4$ defined in equations (2) serve as queues for the small tick OMG token. Consequently the queues are not fixed but are formed dynamically and may vary over time. But for the present investigation we assume that the four queues Q_i, $i = 1, ..., 4$



bid-side ask-side

Figure 12: Zoom into average relative depth profiles of bid-side LOB and ask-side LOB

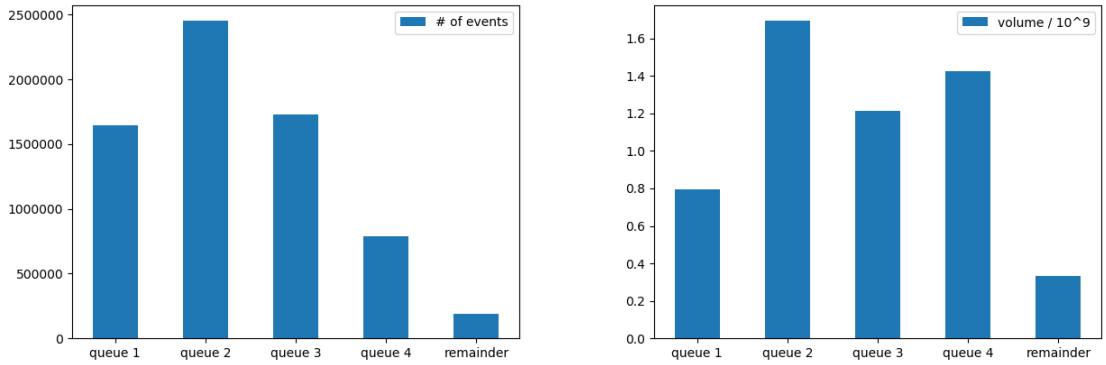


Figure 13: Number of events arriving in queues (left) and total volume arriving in queues (right) during the time of observation

remain fixed. The four queues show distinct behaviour regarding the events arriving in the LOB. In the left panel figure 13 the total number of events are plotted for each queue and for the remaining relative price range $\delta^{a/b} > 350 \cdot p_0$. We see that most events are placed into Q_2 and very few events are placed into the remaining relative price range $\delta^{a/b} > 350 \cdot p_0$. The events being placed in Q_1 have on average a smaller volume while in particular the event placed in Q_4 and in the relative price range $\delta^{a/b} > 350 \cdot p_0$ have on average a larger volume.

The distribution of the volume per queue w.r.t. LOB snapshots during observation period of 54 days is shown in figure 14. We see that the volume distribution in Q_1 has is peaked roughly around $v \approx 100$ OMG, $v \approx 1000$ OMG and $v \approx 2800$ OMG. Q_2 and Q_3 have similar distributions. Q_4 on the other hand has a very distinct distribution peaked roughly around $v \approx 29000$ OMG. In section 6 we model these distributions using a simple queue model from [HLR15].

Finally, note that most of the volume in the LOB is aggregated far away from $\delta^{a/b} = 0$ as we can

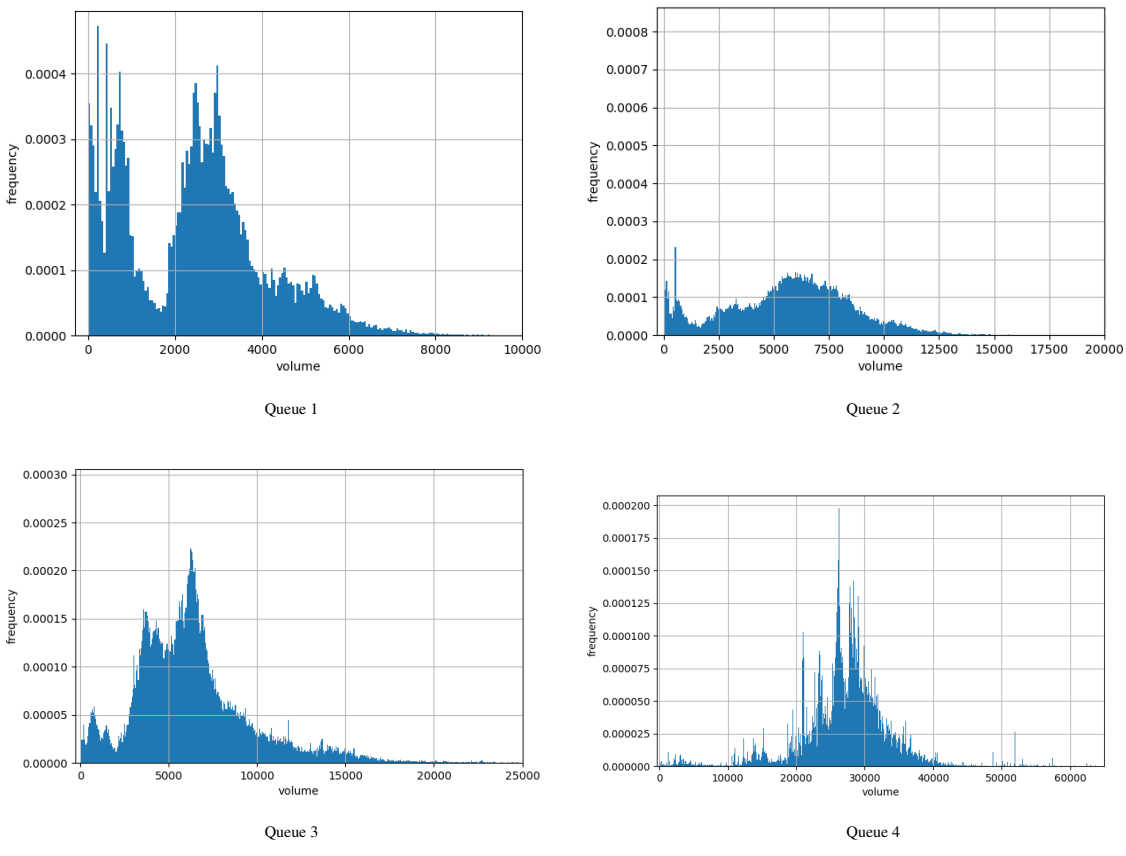
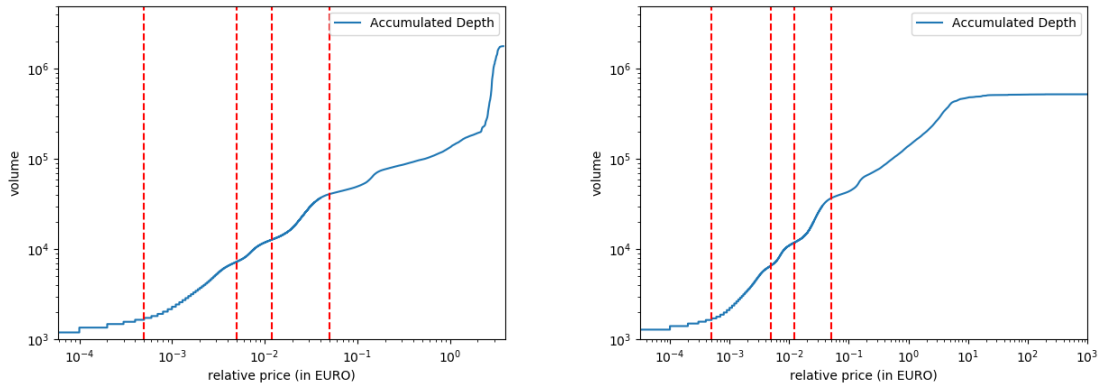
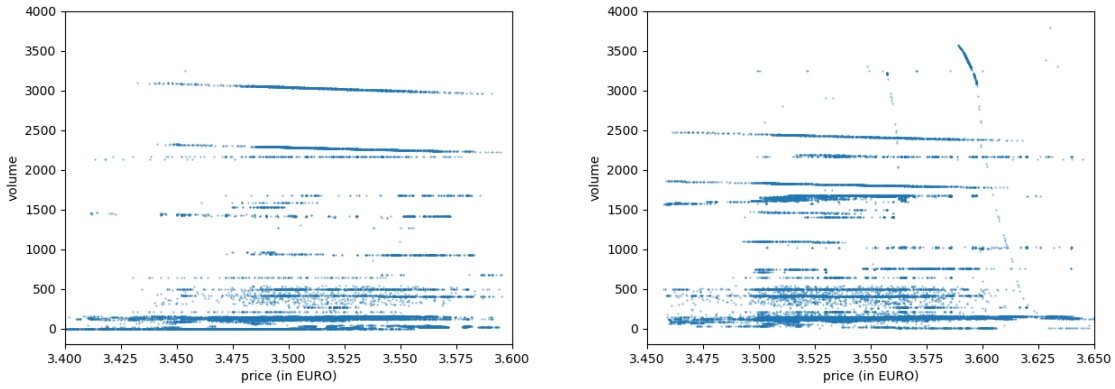


Figure 14: Distribution of the volume per queue w.r.t. LOB snapshots during observation period of 54 days, bin size = 50 OMG



bid-side ask-side

Figure 15: Average accumulated depth profiles of bid-side LOB and ask-side LOB



bid-side ask-side

Figure 16: Price-volume scatter plots for bid-side and ask-side. Data set from September 29^{th} 2021 between 17:00 and 18:00 CET.

easily see in the plot of the average bid-side relative depth profile in figure 11. On the ask side the depth profile is stretched out over a wide rage of the relative price. By plotting the accumulated depth profiles $(p, \Sigma N^{b/a}(p,t))$ in figure 15 the accumulation of volume far away from $\delta^{a/b} = 0$ becomes apparent.

Since relatively few events arrive in the price range $\delta^{a/b} \geq 350 p_0$, see figure 13, we conclude that orders placed far into the LOB are rarely cancelled. This leads to an accumulation of volume, i.e. volume is parked at a very good price if executed but with very little chance of execution.

5.6 Trading Bots

A price-volume scatter plot 16 of the incoming limit orders shows horizontal patterns which can be attributed to high frequency algorithmic traders. We identify these patterns as two basic types of traders. The first type consists of traders which place and cancel orders with fixed volume into the LOB. The second type consists of traders which place and cancel orders into the LOB for which the product of price and volume is within a small interval, i.e. the product is almost fixed. By high frequency trading we mean in this context that the time between placing an cancelling an order is in general less than a minute. As we show later, that the cancellation time of orders is on average indeed much shorter.

In the following we define these algorithmic traders which we will call *Bots* and present an algorithm to classify the incoming orders into the different Bot types. The definition of Bots uses in particular the order ID which is provided by the Level 3 updates from Coinbase Pro. In the context of Bots we focus on limit orders only and do not take market orders into account.

Let us write l for a limit order, o for an open order, m for a matched and c for a cancel order. For an order $x = l, o, m$ or c define $SID(x)$ to be the (ongoing) sequence ID of the order and $OID(x)$ be its order ID, see section C and appendix D for further details.

Let $\mathcal{O}$ be a set containing n-tuples $X = (x^1, x^2, ..., x^n)$ of order with $x^i = l$ for some i and $OID(x^1) = OID(x^2) = \cdots = OID(x^n)$. Note that n is not fixed so $\mathcal{O}$ contains in general tuples of different lengths. In particular $x^1 = l$. Furthermore we define the order ID of the n-tuple as $OID(X) := OID(x^1)$ and, since all orders with the same order ID share the same price, the price of the n-tuple as $p_X := p_{x^1}$ where p_{x^1} is the price of x^1. Note that due to the way the orders are generated on the exchange any n-tuple in $\mathcal{O}$ can be arranged such that $SID(x^1) < SID(x^2) < \cdots < SID(x^n)$. We assume that all n-tuples in $\mathcal{O}$ are ordered in such a way with respect to the sequence ID.

Let $\mathcal{T} \subset \mathcal{O}$ be a subset of 3-tuples fo the form $X = (l, o, c)$, i.e. a set containing limit orders which are not (partially) executed before being cancelled. For such 3-tuples we can define the volume of the 3-tuple $v_X := v_l$ with v_l being the volume of the limit order. Furthermore we define the *funds* of a 3-tuple $X \in \mathcal{T}$ as $f_X := p_X \cdot v_X$. Each $X \in \mathcal{T}$ contains exactly two events, an open order o and a cancel order c.

We can now define the two main Bot types that we could identify.

Definition 5.7 A subset $CVB \subset \mathcal{T}$ is called a *constant volume Bot* (CVB) if the following three conditions hold:

1. there is a $v > 0$ such that for all triples $X = (l, o, c) \in CVB$ we have $v_X = v$.
2. the cardinality $|CVB| \geq \kappa$, with $\kappa > 0$ fixed.
3. for all 3-tuples $X_i = (l_i, o_i, c_i) \in CVB$ the indexing $i = 1, 2, ...$ can be chosen such that $SID(c_i) < SID(l_{i+1})$.

The set of all constant volume Bots in $\mathcal{T}$ is denoted as $\mathcal{CVB}$.

Condition 1 in definition 5.7 ensures that all orders of a CVB have indeed the same volume. Condition 2 ensures that there are at least κ consecutive orders in a CVB and condition. Choosing κ correctly is basically a matter of experience. With condition 3 we aim to ensure that the CVB cannot double spend its trading volume by placing a new limit order before canceling the active one.

Definition 5.8 A subset $CFB \subset \mathcal{T}$ is called a *constant funds Bot* (CFB) if the following three conditions hold:

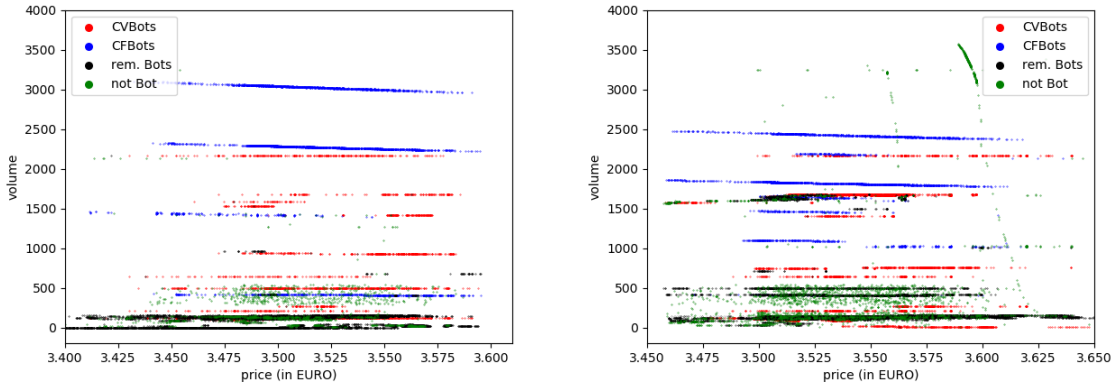
1. for all triples $X = (l, o, c) \in CVB$ we have $f_X \in I$ for an interval $I \subset \mathbb{R}_+$.
2. the cardinality $|CFB| \geq \kappa$, with $\kappa > 0$ fixed.
3. for all triples $X = (l, o, c) \in CFB$ we have $v_X \geq V$ with $V > 0$ fixed.

The set of all constant funds Bots in $\mathcal{T}$ is denoted as $\mathcal{CFB}$.

Constant funds Bots are defined w.r.t. an interval $I \subset \mathbb{R}_+$ since the trading Bot will adjust the price and the volume of each limit order according to the state of the LOB it observes. From the data we see that this adjustment is not so accurate that the funds stay fixed but they fluctuate slightly. Therefore we demand in condition 1 that the funds stay in a suitably chosen interval. As in definition 5.7, condition 2 ensures that there are at least κ orders in a CFB and condition. An equivalence of condition 3 from definition 5.7 turns out to be not very practical since in many cases funds are split into several equal sized CFBs trading simultaneously. This phenomenon does not seem to be common for CVBs. Condition 3 has been added since the density of low volume 3-tuples is generally much higher than the density of high volume 3-tuples. Thus we would easily miss-identify low volume 3-tuples as CFBs.

Definition 5.9 A subset $RBOT \subset \mathcal{T}$ is called a a *remaining Bot* (rem.Bot) if it fulfills the conditions of definition 5.7 or 5.8 but with the negation of the respective condition 3. The set of all remaining Bots in $\mathcal{T}$ is denoted as $\mathcal{RB}$.
Finally we define $\mathcal{NB} := \mathcal{O} \setminus (\mathcal{CVB} \cup \mathcal{CFB} \cup \mathcal{RB})$ as the set of orders which by our definition are *not Bots*.

The remaining Bots are those 3-tuples which we consider as being Bots but which cannot be identified clearly enough as CVBs or CFBs. The set $\mathcal{NB}$ may certainly still contain many algorithmically traders, but we do not classify them as Bots.
To extract the Bot types from the Level 3 LOB updates we employ a simple three step algorithm. As a preliminary choose the same minimal length $\kappa > 0$ for CVBs and CFBs. For CFBs choose possibly several intervals $I_f \subset \mathbb{R}_+$ and a minimal volume V. We adapt the intervals I_f to the range of the funds in which we are searching for CFBs. Then employ the following algorithm:



bid-side ask-side

Figure 17: Price-volume scatter plots for bid-side BOTs and ask-side BOTs. Data set from September 29^{th} 2021 between 17:00 and 18:00 CET. Both plots have been cropped to the displayed price and volume range.

I. find all CVBs in $\mathcal{T}$

II. find all remaining Bots in $\mathcal{T} \setminus \mathcal{CVB}$

III. find all CFBs in $\mathcal{T} \setminus (\mathcal{CVB} \cup \mathcal{RB})$

We start with the CVBs and the remaining Bots because they are easy to identify and the subsequent identification of the CFBs becomes much cleaner.

For the OMG data we find by visual comparison that $\kappa = 50$ and $V = 200$ OMG optimise the identification of BOTs in the data sets . For the CFB intervals we choose I_1 with $\max(I_1) - \min(I_1) = 3$ for funds in the range from 0 €·OMG to 1000 €·OMG. For funds in the range from 1000 €·OMG to 2000 €·OMG we choose I_2 with $\max(I_2) - \min(I_2) = 12$. And for funds larger than 2000 €·OMG we choose I_3 with $\max(I_3) - \min(I_3) = 25$. The progressive increase in the interval length for larger funds comes from the observation that the CFBs have a wider price spread for larger funds.

We then apply the algorithm to those Level 3 LOB updates which constitute the 146 error free LOB intervals, see remark 4.1. For each time interval T_i, $i = 1, .., 146$ we obtain a set $\mathcal{CVB}_i$, $\mathcal{CFB}_i$, $\mathcal{RB}_i$ and $\mathcal{NB}_i$. Then we take the union over all respective sets for each Bot type to obtain $\mathcal{CVB}$, $\mathcal{CFB}$, $\mathcal{RB}$ and $\mathcal{NB}$.

In figure 17 we illustrate the result of the algorithm on the data for the data of second time interval collected on September 29^{th} 2021 between 17:00 and 18:00 CET.

As we can see, the algorithm performs well for larger volumes (funds) since the 3-tuples are less dens compared to smaller volumes (funds). The discrimination is particularly difficult for small volumes with integer values. Furthermore we can see that there are probably also algorithmic traders left in the set we identified as not Bots. For example a green vertical trace on the ask side in the price range between $3.600 - 3.625$ € which we cannot attribute to any type of Bot.

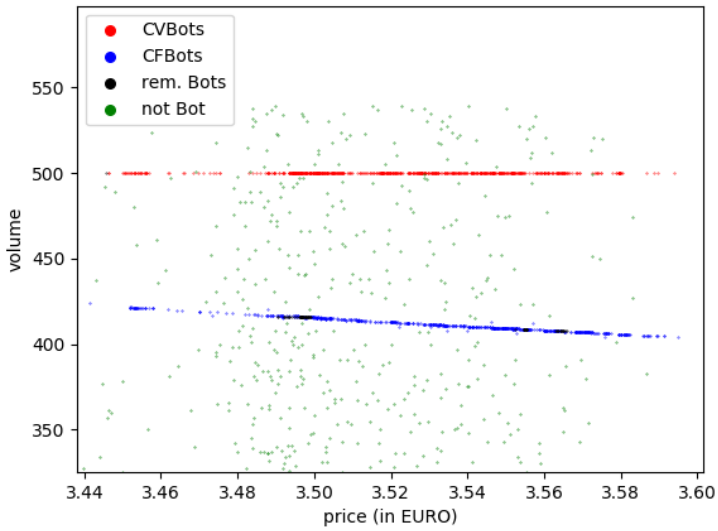


Figure 18: Price-volume plot bid-side BOTs

The algorithm is also not perfect in detecting all 3-tuples which belong to a Bot and may miss-attribute 3-tuples. See for example the CFB (blue dots) in figure 18. Some of the 3-tuples have been miss-attributed as rem.Bots. Changing simply the order of steps in the execution of algorithm unfortunately does not improve the situation. Due to long computation times we refrained for the time being to construct a more precise algorithm which would certainly have to be more complicated.

5.7 Bot Statistics

The distribution of cancellation times of 3-tuples which we identified as either CVBs, CFBs or rem.Bots is shown in figure 19. We find that 75% of the 3-tuples identified as Bots get cancelled within less than 4 seconds. The maximum of the distribution is ~ 50 ms which is a clear sign for algorithmic traders and well below the reaction time of any human. The volume distributions of the 3-tuples from the three Bot types and those not identified as Bots are shown in figure 20. We can make the following observations for the different Bot types. The volume distribution of 3-tuples from CVBs is widely spread up to ~ 11000 OMG. There are two main clusters, one around ~ 500 OMG and a second one around ~ 2000 OMG as well as several smaller ones.

The volume distribution of of 3-tuples from CFBs on the other hand shows two clear clusters. Most of the 3-tuples sit in the second cluster with its maximum around ~ 2500 OMG. The first, smaller cluster has its maximum also around ~ 500 OMG.

For the 3-tuples from rem.Bots and those not identified as Bots the volume distribution for volumes below 1000 OMG seems to be almost identical. This strengthens the assumption that most of the orders that we could not identify as Bots nevertheless have their origin in algorithmic traders. For

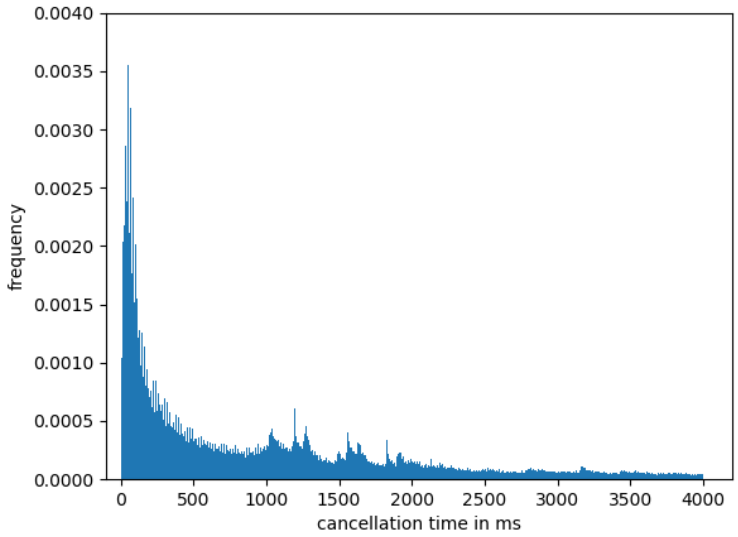


Figure 19: Distribution of cancellation times of less than 4000 ms for CVBs, CFBs and rem.Bots.

rem.Bots we see additionally a relatively broad cluster ranging $\sim$ 1800 OMG to $\sim$ 3500 OMG followed by smaller clusters.

Overall there seem to be close similarities in the shape of the the volume distributions for the three Bot types for volumes below $\sim$ 4500 OMG. Those orders which we cannot identify as Bots seem to have mainly small volumes of less than 1000 OMG.

In figure 21 we show the distribution of events (open, matched and cancel orders) and the absolute value of their volume arriving at the four queues defined in equation (2). We note that Bots generate for more than 86% of all events arriving in Q_1 to Q_4. Taking into account that the number of events arriving in the remainder of the LOB, see figure 13 is very small, we conclude that Bots play a dominant role in the dynamics of the LOB.

From the left panel we see that in Q_1, Q_2 and Q_3 the majority of the incoming events have their origin in rem.Bots but they account in general for less than 15% of the volume arriving at the queues. The number of events and the volume deposited in the queues is monotonically falling from Q_1 to Q_4.

The number of events coming from QVBs is rising monotonically from Q_1 to Q_4 and the same is true for the absolute value of their volume, see left panel. QVBs account for $\sim$ 80% of the volume in Q_4

The number of events coming from QFBs stays approximately constant from Q_1 to Q_3 and drops off in Q_4. Although QFBs account only for $\sim$ 20% of the events in Q_1 to Q_3 they dominate the volume which arrives in these queues with $\sim$ 80% in Q_2. In contrast QFBs contribute a negligible amount of volume to Q_4.

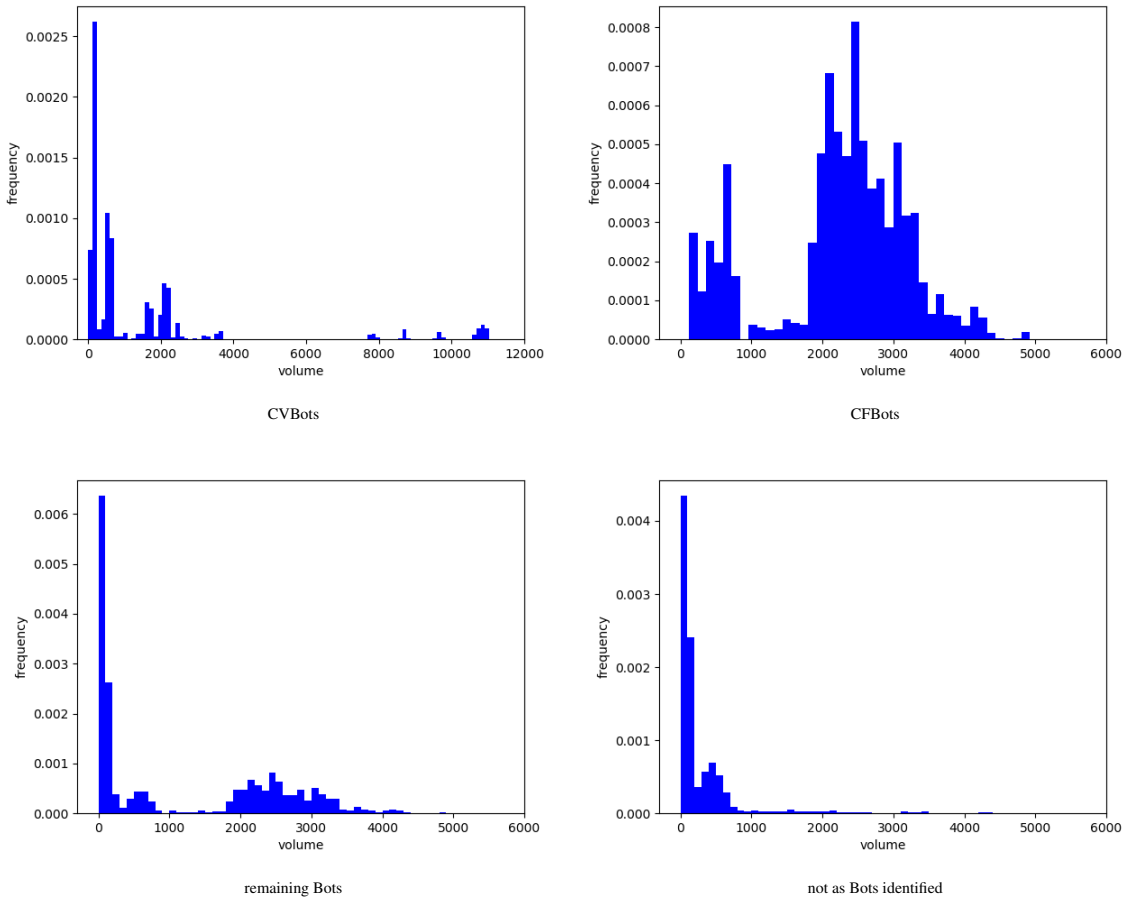


Figure 20: Distribution of absolute volume $|v_x|$ of all events arriving via different Bot types, bin size = 100

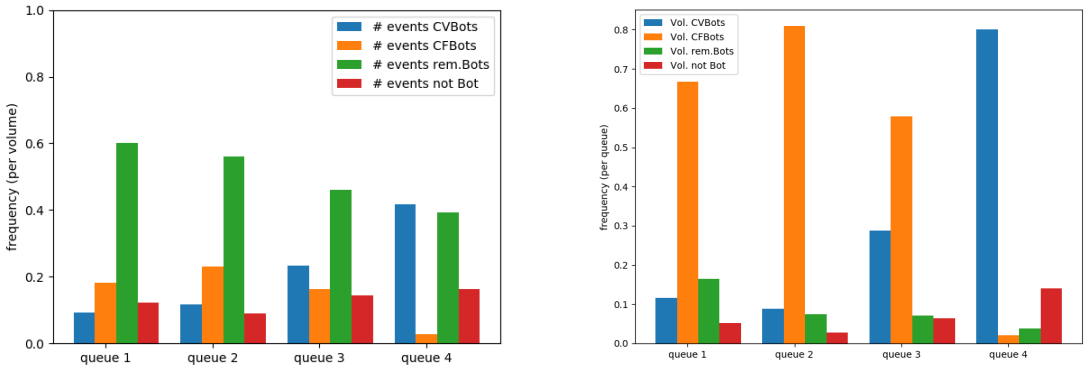


Figure 21: Ratio of number of events (left) and ratio of absolute value of volume (right) w.r.t. bot types per queue

Those events which we cannot identify with Bots stay almost constant from Q_1 to Q_4 w.r.t. to either the number of events or the volume arriving at the queues.
All these statistical features fit well with the assumption that one purpose of Bots is the exploitation of the Coinbase Pro fee structure as described in section 4.1. CFBs place and cancel a large number of orders with relatively large volume close to the bid-price $b(t)$ or ask-price $a(t)$. So if a match occurs the adjusted volume guarantees that the desired funds are spent or received. If no match occurs CFBs generate volume flow to reduce the trading fees. CVBs with large volumes on the other hand place their orders relatively far from the bid-price and ask-price. We suspect that CVBs serve a double purpose just as CFBs. On the one hand they park the tokens at a safe distance while waiting for a better price. On the other hand they also generate volume flow thus reducing the trading fees.

6 A simple Queuing Model

The authors of [HLR15] propose that the LOB close to the bid price and the ask price can be described by a Markov queuing system consisting of several (possibly interacting) queues. In particular the average a volume distribution for each queue can be seen as the invariant distribution of the Markov processes which drive the dynamics of the queues.
In [HLR15] the assets under consideration are large tick stocks such as France Telecom and Alcatel Lucent traded on Euronext Paris. The queue size is defined to be 1 tick and the authors consider the first four queues to be the first four price levels of one tick from relative price $\delta^{a/b} = 0$. Furthermore the authors assume that incoming and outgoing orders have the same size for each queue.
The authors of note that the average volume distribution of the queues depends very little on the side of LOB, i.e. they are almost identical for the bid-side and the ask-side. So they combine both sides into and model the volume change in the queues by a single 4-dimensional continuous time Markov

process. The authors show that under some general assumptions the Markov process is ergodic and has therefore an invariant distribution.
In the simple case of non-interacting queues each queue can then be modeled by an independent $G/G/1$ queuing model. So the distribution of the incoming and outgoing orders is taken to be the empirical distribution given by the order data. The invariant volume distribution can be derived explicitly from the balance equations of the queuing model, see [GSTH08].
Even in the simple case of non-interacting queues the authors find that the invariant volume distribution derived from the order flow is in good agreement with the average volume distribution obtained from LOB snapshots taken over a given time period.
Although the OMG token is certainly a small tick asset, see section 5.4, we nevertheless propose that the simple queuing model of [HLR15] can be applied. We suggest that instead of using queues of size 1 tick, we can use the queue structure that we found in the average relative depth profiles in section 5.5, in particular in figure 12. We use the similarity of the bid-side and the ask-side to justify that we can combine both sides just as the authors in [HLR15] did.
We then assume that the average volume distributions for each in figure 14 can be interpreted stable distributions of an ergodic Markov process driving the simple model treating each queue independently.
Next we define the Markov process of the incoming and outgoing order flow and show that the process is ergodic and thus has an invariant distribution. In section 6.2 we use the balance equations of the queuing model to derive explicit formulas that allow to calculate the invariant distribution from the order flow. Then we apply these results to our data in section 6.3.

6.1 The Ergodic Markov Process

In the following section we will follow closely [HLR15]. The volume change in the queues is modeled by a k-dimensional continuous time Markov process $Q(t) = (Q_1(t), Q_2(t), ..., Q_k(t))$. Each individual queue $Q_\alpha(t)$ will be treated as an independent 1-dimensional continuous time Markov jump process with values in $\mathbb{N}_0$, where the unit of measure is the average size of an incoming event denoted as AES_j (average event size). The AES depends on the queue and the source of the events. The state space of the process $Q(t)$ is $\Omega = \mathbb{N}_0^k$ and is in particular countable. Each incoming event will be modeled by a jump of one AES_α unit in queue Q_α, $\alpha = 1, ..., k$.
In the following we restrict ourselves for simplicity to one queue with average event size $AES = 1$. Given a particular state of the process $q = (q_1, .., q_j, ..., q_k) \in \Omega = \mathbb{N}_0^k$, an incoming limit order in Q_α, $\alpha = 1, ..., k$ is modeled by the jump $q \to q + e_\alpha$ where e_α is the α-th (row) unit vector. An incoming cancellation or market order in queue α is modeled by the jump $q \to q - e_\alpha$.
Since the state space Ω is countable we can define the infinitesimal generator as the transition rate

matrix $\mathcal{Q}$ as follows

$$\begin{aligned} \mathcal{Q}_{q,q+e_\alpha} &= \lambda_\alpha(q_\alpha) \\ \mathcal{Q}_{q,q-e_\alpha} &= \mu_\alpha(q_\alpha) \\ \mathcal{Q}_{q,q} &= -\sum_{p\in\Omega, p\neq q} \mathcal{Q}_{q,p} \\ \mathcal{Q}_{q,p} &= 0 \quad \text{otherwise} \end{aligned} \tag{3}$$

for each $q, p \in \Omega$.
The transition probabilities of the Markov jump process from state q to state p in time t form a semi-group with elements $P_{q,p}(t) = \exp(t\mathcal{Q}_{q,p})$. See figure 22 for an illustration of the process for two time steps of a single queue. In this figure the integers are to be understood as multiples of the average event size $AES = 1$ of the queue.
Our aim is now to show that the Markov jump process admits a stable limiting distribution π, i.e. a distribution satisfying $\pi P = \pi$ and

$$\lim_{t\to\infty} P_{p,q} = \pi_q.$$

Such a Markov process with countable state space is called ergodic.
We nee two further assumptions on the process which will guarantee the process is sufficiently well behaved. The first assumption ensures that queue volumes tend to decrease if they become too large.

Assumption 6.1 (Assumption 1 in [HLR15]) There exist an integer $C \in \mathbb{N}$ and a $\delta > 0$, such that for all $\alpha = 1, ..., k$ and all $q \in \Omega$, if $q_\alpha > C$ then

$$\lambda_\alpha(q_\alpha) - \mu_\alpha(q_\alpha) < -\delta.$$

The second assumption ensures the boundedness of incoming events and thus guarantees that the Markov process is non-explosive.

Assumption 6.2 (Assumption 2 in [HLR15]) There exists an $H > 0$ such that for any $q \in \mathcal{Q}$

$$\sum_{\alpha\in\{1,...,k\}} \lambda_\alpha(q_\alpha) \leq H$$

Under these two assumption we can prove that the following central theorem is true:

Theorem 6.3 (Thm. 2.1 in [HLR15]) *Under assumption 6.1 and assumption 6.1 the k-dimensional continuous time Markov jump process $Q(t) = (Q_1(t), Q_2(t), ..., Q_k(t))$ is ergodic.*

The first key ingredient in the proof of theorem 6.3 is theorem 4.2 from [MT93] from we can deduce that $Q(t)$ is an irreducible Markov process with invariant distribution. The second key ingredient is theorem 3.6.2 from [N98] from which the ergodicity of the Markov process follows.
Theorem 4.2 from [MT93] is formulated for Markov processes with arbitrary state spaces and requires the notion of petite sets. To simplify the presentation we restrict ourselves to the case of countable state spaces. This allows us to replace the notion of petite sets by the notion of small sets and eventually by any subset of the state space. Following [RR01], we first recall the standard definition of a small set.

Definition 6.4 Let $M(t)$ be a Markov process with state space Ω and with transition probability P_M. A set $S \in \Omega$ is *small* if there is an $n_0 \in \mathbb{N}$, an $\epsilon > 0$ and a probability distribution ν on Ω, such that

$$P_M^{n_0}(x, \cdot) \geq \epsilon\, \nu(\cdot), \quad \forall\, x \in C. \tag{4}$$

Every small set is a petite set, see for example [RR04]. By the following proposition we see that any subset of a countable state space is small.

Proposition 6.5 (Prop. 4 in [RR01]) *Let $M(t)$ be a Markov process with countable state space Ω and let $S \subset \Omega$ be any subset. Then S is small.*

Under the additional assumption (CD2) it is shown in theorem 4.2 in [MT93] that a non-explosive right process is positive Harris recurrent and admits a unique invariant probability density. For our purpose it enough to note that positive Harris recurrence implies irreducibility of the Markov process, see section 3.2 in [MT93]. So we formulate the following corollary of theorem 4.2 in [MT93]:

Corollary 6.6 *Let $M(t)$ be a non-explosive Markov jump process with countable state space Ω and infinitesimal generator $\mathcal{Q}$. Assume that for $c, d > 0$, some measurable set $S \subset \Omega$, some functions $f, V : \Omega \to \mathbb{R}$ with $f \geq 1$ and $V \geq 0$ bounded on S*

$$\mathcal{Q}V(x) \leq -cf(x) + d\mathbb{1}_S(x) \quad \forall x \in \Omega \tag{5}$$

holds. Then the process $M(t)$ is irreducible and admits a unique invariant probability density π_M.

If corollary 6.6 can be shown to hold, theorem 3.6.2 from [N98] can be used to deduce ergodicity.

Theorem 6.7 (Thm. 3.6.2 in [N98]) *Let $M(t)$ be an irreducible non-explosive Markov process with state space Ω, infinitesimal generator $\mathcal{Q}$ and transition probabilities $P_{p,q}(t)$, $p, q \in \Omega$. Assume that $M(t)$ has an invariant distribution π. Then we have for all states $p, q \in \Omega$*

$$\lim_{t\to\infty} P_{p,q}(t) = \pi_q.$$

We are now in the position to sketch the main steps of the proof of theorem 6.3. For computational details we refer to Appendix A in [HLR15].

Proof of theorem 6.3:
Let $z > 1$ and define a positive function $V : \Omega \to \mathbb{R}$ by

$$V(q) = \sum_{\alpha \in \{1,\dots,k\}} z^{|q_\alpha - C|}$$

where $C \in \mathbb{N}$ from assumption 6.1. Now apply the infinitesimal generator $\mathcal{Q}$ defined in 3 to V. For any $q \in \Omega$ one finds

$$\begin{aligned}
\mathcal{Q}V(q) &= \sum_{p \neq q} \mathcal{Q}_{p,q}[V(p) - V(q)] \\
&= \sum_{\alpha \in \{1,\dots,k\}} [\lambda_\alpha(q_\alpha)(z^{|q_\alpha+1-C|} - z^{|q_\alpha - C|}) + \mu_\alpha(q_\alpha)(z^{|q_\alpha - 1 - C|} - z^{|q_\alpha - C|})] \\
&= (z-1) \sum_{\substack{\alpha \in \{1,\dots,k\}, \\ q_\alpha = C}} \lambda_\alpha(q_\alpha) + (z-1) \sum_{\substack{\alpha \in \{1,\dots,k\}, \\ q_\alpha > C}} [\lambda_\alpha(q_\alpha) + \frac{\mu_\alpha(q_\alpha)}{z}] z^{q_\alpha - C})
\end{aligned}$$

For $q_\alpha > C$ we can find a z sufficiently close to 1 such that under assumptions 6.1 and 6.2

$$\lambda_\alpha(q_\alpha) + \frac{\mu_\alpha(q_\alpha)}{z} < \frac{-\delta + H(z-1)}{z} < 0.$$

Setting $r := (\delta - H(z-1))/z$ and using assumption 6.2 again we have

$$\begin{aligned}
\mathcal{Q}V(q) &\leq (z-1)\, H - (z-1)\, r \sum_{\substack{\alpha \in \{1,\dots,k\}, \\ q_\alpha > C}} z^{q_\alpha - C} \\
&\leq -(z-1)\, r \sum_{\alpha \in \{1,\dots,k\}} z^{|q_\alpha - C|}) + (z-1)rk + (z-1)\, H \\
&= -(z-1) r\, V(q) + (z-1)(rk + H) \mathbb{1}_\Omega(q)
\end{aligned}$$

Multiplication by $\mathbb{1}_\Omega(q)$ is a tautology since Ω is itself a small set by proposition 6.5. So condition (5) from corollary 6.6 holds after suitable normalisation of the constants.
Therefore the Markov jump process $Q(t)$ is irreducible and admits a unique invariant probability distribution π. By assumption 6.2 the Markov process is non-explosive. So theorem 6.7 ensures that the probability distribution π is obtained by taking the limit $t \to \infty$ of the transition probabilities $P_{p,q}(t)$ and $Q(t)$ is an ergodic Markov process. ⧈

6.2 The Stable Distribution

In order to determine the stable distribution π of ergodic Markov process in section 6.1 we assume that the queues are independent and each queue can be described by a $G/G/1$-queuing model. The process is illustrated in figure 22 for a single queue Q_α. We simplify our notation writing λ_n and

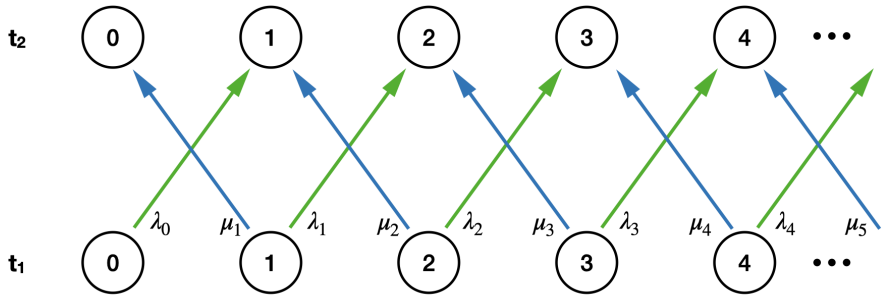


Figure 22: Simple queueing model driven by an ergodic Markov process.

μ_n for the volume increasing and decreasing intensity under the condition that volume in Q_λ is n. Similarly we define π_n to be the n-th component of the invariant distribution at volume n of Q_α. Following [GSTH08], to find the stable distribution π of the Markov process we write down the balance equations for the process.

$$0 = -(\lambda_n + \mu_n)\pi_n + \lambda_{n-1}\pi_{n-1} + \mu_{n+1}\pi_{n+1} \quad \text{for} \quad n \geq 1 \tag{6}$$

$$0 = -\lambda_0\pi_0 + \mu_1\pi_1 \tag{7}$$

$$1 = \sum_{j=0}^{\infty} \pi_j \tag{8}$$

Equations (6) and (7) guarantee that in the limit of the stable distribution π incoming and outgoing volume are equal for each place in the queue. Equation (8) is the normalisation of the probability distribution. The major advantage of this simple queuing model is that it admits a closed analytic solution.

Assuming that $\mu_i > 0$ for all i, re-arranging (6) and (7) gives

$$\pi_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}}\,\pi_n - \frac{\lambda_{n-1}}{\mu_{n+1}}\,\pi_{n-1} \quad \text{for} \quad n \geq 1 \tag{9}$$

$$\pi_1 = = \frac{\lambda_0}{\mu_1}\,\pi_0. \tag{10}$$

For $n = 1$ we find using (10)

$$\begin{aligned}
\pi_2 &= \frac{\lambda_1 + \mu_1}{\mu_2}\,\pi_1 - \frac{\lambda_0}{\mu_2}\,\pi_0 \\
&\overset{(10)}{=} \frac{\lambda_0}{\mu_1}\left(\frac{\lambda_1 + \mu_1}{\mu_2}\right)\pi_0 - \frac{\lambda_0}{\mu_2}\,\pi_0 \\
&= \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\,\pi_0 + \frac{\lambda_0\mu_1}{\mu_1\mu_2}\,\pi_0 - \frac{\lambda_0\mu_1}{\mu_1\mu_2}\,\pi_0 \\
&= \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\,\pi_0
\end{aligned}$$

A straightforward induction argument leads to

$$\boxed{\pi_n = \pi_0 \prod_{j=1}^{n} \frac{\lambda_{j-1}}{\mu_j} \quad \text{for} \quad n \geq 1} \tag{11}$$

From the normalisation condition (8) for probabilities follows

$$\begin{aligned} 1 &= \sum_{i=0}^{\infty} \pi_i \\ &\overset{(11)}{=} \pi_0 + \sum_{i=1}^{\infty} \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \\ &= \pi_0 \left(1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \right) \end{aligned}$$

and therefore

$$\boxed{\pi_0 = \left(1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \right)^{-1}} \tag{12}$$

where we assume that

$$\sum_{i=1}^{\infty} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} < \infty \tag{13}$$

converges. The convergence of (13) poses no problem since only finitely many intensities λ_i and μ_i are non zero.

From the data of the Level 3 updates we can now calculate the intensities λ_j and μ_j for each queue Q_1 to Q_4 as defined in (2). Using equation (11) and (12) we then estimate the invariant distribution π for each queue w.r.t. to bin size given by the average event size AES_1,..,AES_4 of each queue.

6.3 Data Analysis and Model Fitting

To estimate the invariant distributions π for each queue Q_α $\alpha = 1, .., 4$ we follow again closely [HLR15]. First we need to estimate the average event sizes EAS_α as well as the volume increasing intensities $\lambda(Q_\alpha) := (\lambda_1, \lambda_2, ...)$ and the volume decreasing intensities $\mu(Q_\alpha) := (\mu_1, \mu_2, ...)$. The volume increasing intensity corresponds to the intensity of the open order flow while the volume decreasing intensity consists of the intensities of the matched order flow and the cancel order flow. We write $\pi(Q_1) = (\pi_1, \pi_2, ...)$ for the corresponding invariant distributions.We may drop the explicit dependency on the queue in λ, μ and π if it is clear from the context to which queue they belong.

Let us focus on a single queue Q_α. Let κ indicate the type of an event x, i.e. $\kappa = l$ for an open order, $\kappa = m$ for a matched order $\kappa = c$ for a cancel order. $\mathcal{E}(\kappa)$ is defined as the set of all events of type κ arriving at queue Q_α. We denote by κ_n the intensity of events $x \in \mathcal{E}(\kappa)$ given the volume

	AES Q1 (in OMG)	**AES Q2** (in OMG)	**AES Q3** (in OMG)	**AES Q4** (in OMG)
all events	483	691	701	1807
CVBots	603	519	868	3467
CFBots	1762	2422	2477	1433
rem. Bots	132	92	107	174

Figure 23: Average event sizes for queues $Q_1, ..., Q_4$ for all orders and restricted to the three bot types.

of queue Q_α is $q_\alpha(x) = n$. We write $\kappa_n = \lambda^o_n$ for volume increasing open orders. For volume decreasing orders we write $\kappa_n = \mu^m_n$ for matched orders $\kappa_n = \mu^c_n$ for cancel orders.
With $\Delta t(x)$ being the time passed between event x and preceding event arriving at queue Q_α we use the maximum likelihood method to estimate the intensities κ_n

$$\begin{aligned} \Lambda_n &:= (\text{mean}[\Delta t(x)|q_\alpha(x) = n])^{-1} \\ \kappa_n &:= \Lambda_n \frac{|\{x \in \mathcal{E}(\kappa)|q_\alpha(x) = n\}|}{|\{q_\alpha(x) = n\}|}. \end{aligned} \tag{14}$$

We estimate the intensities from the data of the 146 valid LOB intervals in section 4.2. The data is divided into time intervals during which the queues are stable, i.e. the bid price does not change for the bid-side queue and the ask price does not change for the ask-side queue. So we restart recording each time when the bid price or the ask price changes on the respective side of the LOB. We use the data to estimate the intensities κ_n and take the arithmetic mean over all recording intervals.
Let us first estimate the average event sizes for the queues $Q_1, ..., Q_4$ by taking the arithmetic mean of the absolute value of the volume $|v_x|$ of the orders x under consideration. The table in figure 23 shows the AES_α for all events, for constant volume bots $\mathcal{CVB}$, for constant funds bots $\mathcal{CFB}$ and for the remaining bots $\mathcal{RB}$. We see that AES_α increase from Q_1 to Q_4 if we consider all orders. This behaviour is no longer true for the different bot types separately.

All Events: Let us first estimate the intensities and the invariant distribution for all events. In figure 24 the intensities of all events are shown in units of events per second.
The intensity of volume increasing events $\lambda^o(Q_1)$ in Q_1 has an increasing tendency from 1 AES and reaches a maximum at $\sim 16AES \approx 7700$ OMG. So there is an increasing tendency of traders to add volume to Q_1 if the volume increases. We also see a local maximum around $\sim 16AES$ for the intensities $\mu^c(Q_1)$ of the cancel orders and the intensities of matched orders $\mu^m(Q_2)$. We can conclude that traders become more active in Q_1 as the available volume approaches $\sim 16AES$.
We can interpret the maximum at $30AES \approx 14400$ OMG in $\mu^c(Q_1)$ as follows. Events which have been placed into Q_2, Q_3 or Q_4 can be moved into Q_1 if the bid price or ask price changes enough.

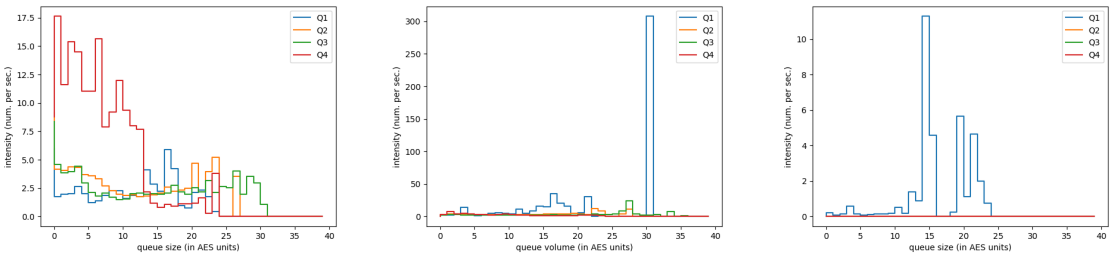


Figure 24: Intensities for limit/open orders (left), cancel orders (middle), matched order (right) of all events.

This unintentional miss-placement of orders triggers the corresponding traders to immediately cancel them. Here we already see that our assumption of independent queues is just an approximation and that the queues interact.

The intensities of the matched orders show clearly that matching only occurs in Q_1 and then predominantly if the queue contains a relatively large volume.

We also observe a relatively high intensity if the volume of Q_1 is zero. Placing an order into Q_1 at volume zero means placing an order at least 7 ticks into the spread and thus moving the queues. The sharp drop $\lambda^o(Q_1)$ from volume zero to $1AES$ indicates that other traders need a certain time to react to the new price levels and become more active as the volume in Q_1 increases again.

The intensities in Q_2 and Q_3 show a similar behaviour to the ones in Q_1 with their maxima shifted to higher queue volume. But Q_3 shows a quite different behaviour. Volume increasing events arrive particularly in Q_3 if its volume is small. This seems to be at odds with the relatively small intensity $\lambda^c(Q_3)$ of cancel orders. But as we saw in section 5.7, Q_4 is largely dominated by high volume CVBs. As we will see below, restricting ourselves to these high volume CVBs provides an explanation for this anomaly.

From the intensity estimates we can now calculate the invariant distributions for each queue using equation (11) and equation (12). Note that the total intensity of volume decreasing order flows is simply $\mu = \mu^c + \mu^m$, see [GSTH08]. We compare the invariant distributions to the average volume distributions we obtained from the LOB snapshots which we collected in one minute intervals from 7pm - 4pm CET, see section 4.2. In this way the two data sets are relatively independent from each other.

In figure 25 we plot the invariant distributions as red lines. The average volume distributions from the LOB snapshots are plotted with bin size adjusted to the corresponding average event size are shown as bar plots. Compare figure 14 for the average volume distribution with smaller bin size.

The invariant distributions fit the average volume distributions quite well for Q_2 and Q_3 apart from the local maxima close to zero volume. The fit of the invariant distribution to the average volume distribution in Q_4 is slightly shifted to larger volumes but still seems reasonable good. In Q_1 the fit does not capture the structure of two maxima for the average volume distribution. The invariant dis-

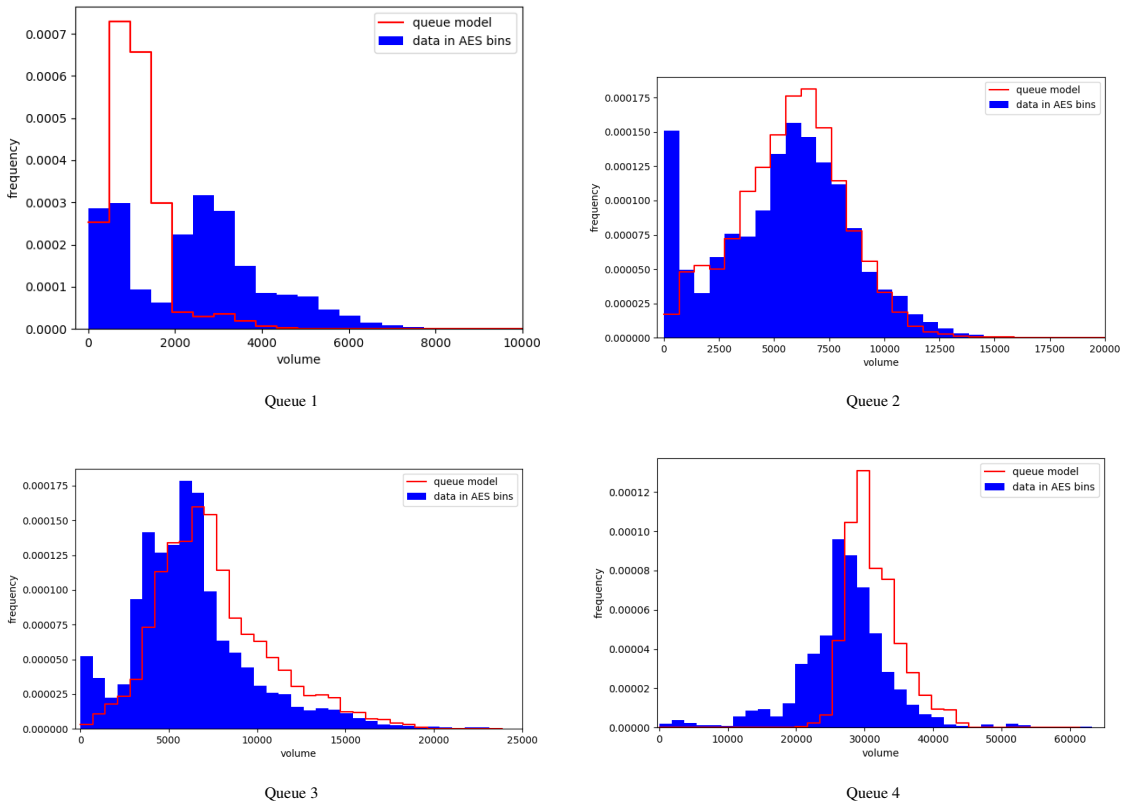


Figure 25: Average volume distribution from LOB snapshots vs. invariant distribution of the queue model estimated from all events. Bin size is given by the average event size AES_α, $\alpha = 1, ..., 4$. The volume axis of the plots are in OMG units for easier comparability.

tribution shows only a single maximum at $\sim 2AES$ and completely misses the second maximum at $\sim 6AES$ of the average volume distribution. This may be due to the fact that the outgoing volume in Q_1 overcompensates the incoming volume for large queue volume due to the moving queue effect described above. Using one single average event size for all events also gives too much weight to events with smaller volume which have a local maxima in their volume increasing intensities at lower queue volume as we will see below in figures 26 and 30.

To get a more detailed view we will next restrict ourselves to events coming from the three bot types, i.e. to events in $\mathcal{CVB}$, $\mathcal{CFB}$ and $\mathcal{RB}$. We ignore those events that we classified as not bots, i.e. those in $\mathcal{NB}$.

Constant Volume Bots: Next we restrict our analysis to CVBs, i.e. events $x \in \mathcal{CVB}$. In figure 26 we plot the intensities of the volume increasing and volume decreasing order flows. Note that we only need to take the limit/open orders and the cancel orders into account, since by $\mathcal{CVB}$ do not contain matched orders by definition 5.7.

The left plot in figure 26 shows the intensities of the limit/open and cancel order flows for Q_1. We

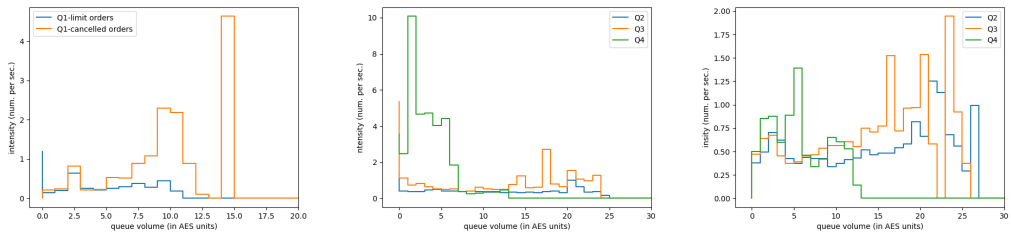


Figure 26: Intensities CVBots. Q_1 open orders and cancel order (left), $Q_2, ..., Q_3$ open orders (middle) and $Q_2, ..., Q_3$ cancel orders (right)

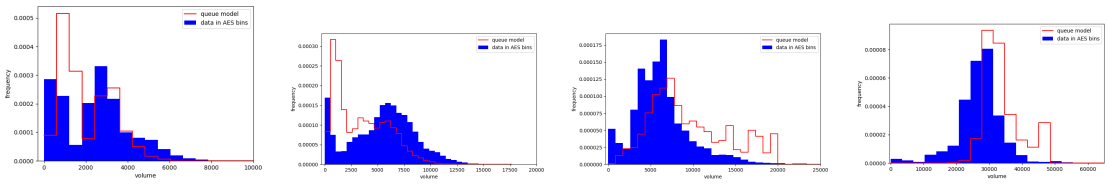


Figure 27: Average volume distribution from LOB snapshots vs. invariant distribution of the queue model estimated from events in $\mathcal{CVB}$. Bin size is given by the average event size AES_α, $\alpha = 1, ..., 4$. The volume axis of the plots are in OMG units for easier comparability.

see that two maxima of the intensities $\lambda^o(Q_1)$ of the limit/open order flow at queue volume zero and at $\sim 2.5AES \approx 1500$ OMG. The intensities $\mu^c(Q_1)$ of the cancel order of have several maxima. The first is at $\sim 2.5AES$ matching the maximum of the limit/open order flow. The second maximum at $\sim 10AES \approx 6000$ OMG and in particular the third maximum at $\sim 15AES \approx 9000$ OMG may again be due to the moving queue effect. The same seems to be true for the intensities volume decreasing order flows in Q_3 and Q_4 as can be seen from the plot in the middle and the plot on the right of 26.

If we focus on Q_4 we see that the intensities $\lambda^o(Q_4)$ of volume increasing and $\mu^c(Q_4)$ volume decreasing order flows appear to be reasonably well balanced. So we suspect that big volumes are placed and canceled predominantly into Q_4. And if Q_1 gets completely depleted and events in Q_4 move into Q_3, Q_2 or even Q_1, they get cancelled immediately. This leads to the structure for the decreasing order flows in the lower queues.

Estimating the invariant distributions using CVBs alone is shows that in particular the two maximum structure of the average volume distribution ofQ_1 can be reproduced, see first plot on the left in figure 27. Also the structure of the invariant distribution $\pi(Q_4)$ of Q_4 is still reasonably well captured.

Constant Funds Bots: Now we restrict to CFBs, i.e. to events $x \in \mathcal{CFB}$, see definition 5.8. In figure 28 we plotted the intensities of the volume increasing and volume decreasing order flows.

We observe that CFBs place their limit/open orders predominantly into Q_2 and Q_3. The volume

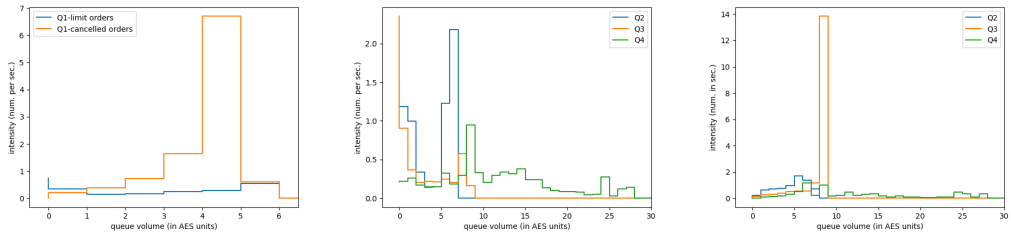


Figure 28: Intensities CFBots. Q_1 open orders and cancel order (left), $Q_2, ..., Q_3$ open orders (middle) and $Q_2, ..., Q_3$ cancel orders (right)

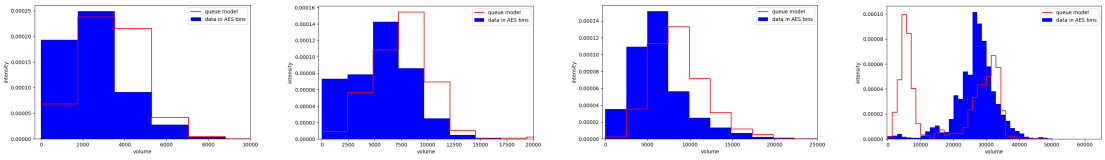


Figure 29: Average volume distribution from LOB snapshots vs. invariant distribution of the queue model estimated from events in $\mathcal{CFB}$. Bin size is given by the average event size AES_α, $\alpha = 1, ..., 4$. The volume axis of the plots are in OMG units for easier comparability.

increasing intensity $\lambda^o(Q_2)$ has a local maximum if the volume in the queue is small $\sim 1AES \approx 2500$ OMG and a global maximum at $\sim 6AES \approx 15000$ OMG. Yet, as can be seen in the right plot, the maximum of the intensity of cancel orders $\mu^c(Q_2)$ is increasing for larger queue volume reaching its maximum also also at $\sim 6AES \approx 15000$ OMG and decreasing from there more slowly than λ^o. So volume is placed into and removed from Q_2 by CFBs at a high frequency if there is a lot of volume in the queue. If there is only a small volume in Q_2, CFBs place volume into the queue but is removed at a lower frequency.

In Q_3 CFBs tend to place volume into the queue at high frequency if the volume in the queue is close to zero (in AES units) and a local maximum at $\sim 8AES \approx 19000$ OMG. There is a sharp peak removing volume from Q_3 if the volume in the queue is $\sim 8AES \approx 19700$ OMG largely overcompensating the corresponding local maximum of $\lambda^o(Q_3)$. We interpret this maximum of $\mu^c(Q_3)$ again as being mainly an effect due to moving queues.

The moving queue effect is probably also the reason for the maximum of the volume decreasing intensity $\mu^c(Q_1)$ at $\sim 5AES \approx 8800$ OMG. We interpret this maximum of $\mu^c(Q_1)$ as volume being cancelled immediately from Q_1 as it is moved from higher queues by a sufficiently large change in price.

CFBs play a negligible role in Q_4 as they contribute few events and little volume, see figure 21.

Estimating the invariant distributions from CFBs alone gives a relatively good agreement with the average volume distributions from the LOB snapshots for Q_2 and Q_3, see figure 29. This is to be expected since CFBs dominate the volume placed into and removed from these two queues as is

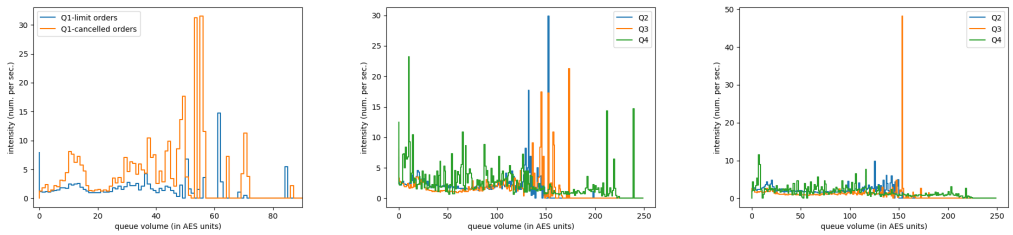


Figure 30: Intensities remaining Bots. Q_1 open orders and cancel order (left), $Q_2, ..., Q_3$ open orders (middle) and $Q_2, ..., Q_3$ cancel orders (right)

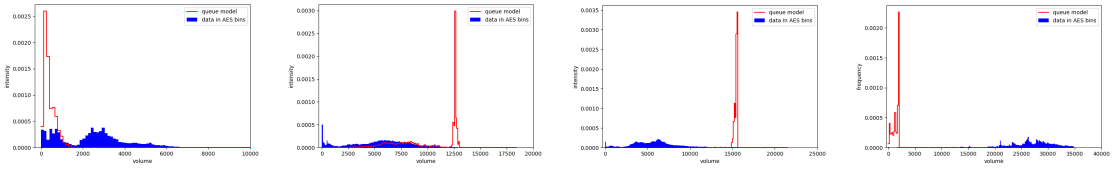


Figure 31: Average volume distribution from LOB snapshots vs. invariant distribution of the queue model estimated from events in $\mathcal{RB}$. Bin size is given by the average event size AES_α, $\alpha = 1, ..., 4$. The volume axis of the plots are in OMG units for easier comparability.

shown in figure 21.

The invariant distributions for Q_1 and Q_4 are hard to interpret. The distribution $\pi(Q_1)$ has little information since the bin size of one $AES = 1762$ OMG obscures the details of the average volume distribution from the LOB snapshots. And the invariant distribution for Q_4 is insignificant due to the negligible contribution the CFB order flow to this queue.

Remaining Bots For those events which originate from the remaining bots $x \in \mathcal{RB}$ we observe in general similar features as for CVBs and CFBs. Figure 30 shows the intensities of the volume increasing order flows $\lambda^o(Q_\alpha)$ and the volume decreasing order flows $\mu^c(Q_\alpha)$ for the four queues. The invariant distributions $\pi(Q_\alpha)$ restricted to the remaining bots match the average volume distribution of the LOB snapshots rather poorly as can be seen in figure 31. This is not surprising, since the volume of the events is not significant compared to CVBs and CFBs, see figure 21.

Let us focus on the intensity of the volume increasing order flows for the case that the relevant price moved by more than 6 ticks, i.e. the case where Q_1 moves by more than 6 ticks into the spread and the volume of Q_1 is zero. Compare $\lambda_0(Q_1)$ with zero volume for the CVBs, the CFBs and the remaining bots:

$\lambda_0(Q_1)$	**CVBots**	**CFBots**	**rem. Bots**
intensity (num. per sec.)	1.19	0.74	7.90

We see that $\lambda_0(Q_1)$ for $x \in \mathcal{RB}$ is by a factor of 6.6 larger than $\lambda_0(Q_1)$ for $x \in \mathcal{CVB}$. Furthermore $\lambda_0(Q_1)$ for $x \in \mathcal{RB}$ is by a factor of 10.6 larger than $\lambda_0(Q_1)$ for $x \in \mathcal{CFB}$. We conclude that the remaining bots are the predominant drivers of price change and therefore volatility and we conjecture that CVBs with small volume also fall into this category. The CFBs on the other hand follow the price in the second row (or queue).

7 Concluding Remarks and Outlook

In this thesis we show that the market for OMG tokens on the crypto exchange Coinbase Pro shares many of standard features and stylised facts with classical assets traded via limit order books. We use the information provided by Level 3 updates to partially identify the structure of algorithmic trading traders. These traders or trading bots dominate the OMG market and their existence is strongly incentivised by the fee structure of Coinbase Pro. A striking feature of the relative depth profiles of the limit order book is the dynamical formation of queue-like structures close to the bid price and ask price.

It is possible to identify two major bot types, constant volume bots and constant funds bots. We analyse some statistical properties of these bots and in particular their relevance in a simple queuing model in order to understand the average volume distribution of the queues. Nevertheless, we only scratched the surface of possibilities to analyse these crypto markets offered by Level 3 updates of the LOB.

The fact that the market is dominated by algorithmic traders, incentivised the by fee structure and easily accessible APIs is certainly also due to the prospect of substantial returns in the hyped market for crypto assets.

Our analysis provides some insights into the structure of the OMG market on Coinbase Pro. But it also left many questions unanswered and generated new questions and ideas. So we end this thesis with some of these open questions and ideas.

The authors of [PRH20] assert that high-frequency trading (5-minute intervals) of may be mainly human based or at least human initiated. They base their conclusion on intraday patterns in the CRIX (CRyptocurrency IndeX). In contrast to this the authors of [SRK19] find no such patterns for the Bitcoin market. Whether these results are at odds with our findings is a priori not clear. The trading bots we found could be initiated by humans on relatively short time scales so that we actually see bot assisted trading. Or there could be a significant difference between cryptocurrencies such as Bitcoin and the OMG token. The precise relation of human and algorithmic traders certainly deserves further research.

We focus our analysis on a single asset, the OMG token, on a single exchange, Coinbase Pro. The next step would be to include other assets (tokens, coins) traded on Coinbase pro and compare the results. This would have the advantage of having also the information of Level 3 updates at our disposal. Another step would be to compare assets across exchanges. Here the effect of different

fee structures would be interesting to investigate.

Concerning the queue structure we were not able to clarify how the queues are actually formed. What is the dynamical feature responsible for their emergence? Can the traders or bot types be identified which produce such queue patterns? And and important question for the statistical analysis and the application of queuing models is the stability of the queues over time. The analysis of these questions will require larger datasets collected over longer time periods.

We assume that the queues are independent. From our analysis of the intensities of the order flows we already saw that this is certainly not the case. The queues interact at least when queues move due to changes in the bid price and the ask price. So what are the correlations between the queues? And can we track the life cycle of individual orders through the queues? Into which queue is an order place and in which queue is it cancelled or matched? Such an analysis should be possible using the order ID provided by Level 3 updates.

Our bot classification is rather straight forward and certainly oversimplifies the structure of the algorithmic traders active on the OMG market. The next step would certainly be to include market orders into the bot classification. The classification could also be improved using clustering techniques based on artificial intelligence to categorise traders more accurately and more efficiently.

A risk analysis to what would happen if (private) bots malfunction would also an interesting line of research. Many of the trading bots we encountered were probably written by amateurs. How disruptive could a malfunctioning bot be for the market?

We see that the simple queue model we employed is certainly not good enough to successfully explain the data. Yet, the model looks promising and may be significantly improved.

In order to estimate intensities of order flows, we restarted data collection after every price change that moved the queue. As a result we also collected data from short time intervals, i.e. during times of high volatility. These time intervals contain only very few events and the queues do not stay stable in the LOB. But in order to estimate an invariant distribution we should at least come reasonably close to the $t \to \infty$ limit.

To improve the situation one could restrict the analysis to time intervals which contain at least certain minimal number of events. On such time intervals it may be better justified that the invariant distribution can be estimated from the order flows since the queues had some time to approach an equilibrium state.

This approach could be extended in the spirit of the queue reactive model from [HLR15]. Apart from the stable regime described above there would be a volatile regime where prices change, the queues move and do not settle into the invariant distribution.

One of the major short comings of the simple queue model seems to be the restriction to orders of one single size given by the average event size. This would be appropriate if the average order volume would at least be close to the AES. But in contrast to this assumption we can clearly see in figure 21 that order volumes of the different bot types have complicated distributions which are poorly modeled by a single (average) value.

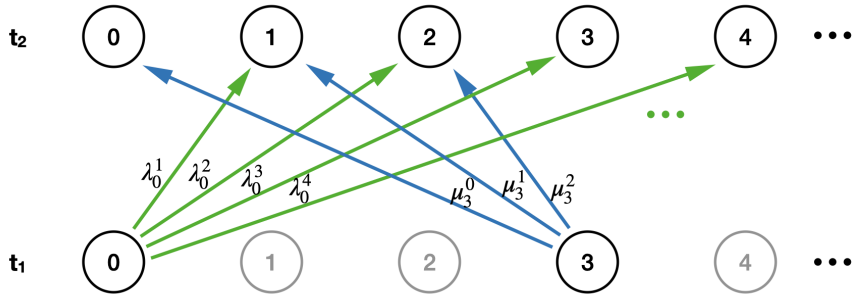


Figure 32: Extended queue model for order flows of variable volume.

We would therefore propose to extend the simple queuing model by generalising to intensities corresponding to order flows of variable volume. A tractable model would be to allow for order volumes that are multiples of a suitable order unit. Figure 32 shows a graphical depiction of such a queuing model where order volume can be added or subtracted in multiples of a unit order volume. This model has the advantage that the balance equation can be written down explicitly as follows:

$$
\begin{aligned}
0 &= -\sum_{j=1}^{\infty} \lambda_n^j \, p_n - \sum_{j=0}^{n-1} \mu_n^j \, p_n + \sum_{j=0}^{n-1} \lambda_j^n p_j + \sum_{j=n+1}^{\infty} \mu_j^n p_j \quad \text{for} \quad n \geq 1 \\
0 &= -\sum_{j=1}^{\infty} \lambda_0^j \, p_0 + \sum_{j=1}^{\infty} \mu_j^0 p_j \\
1 &= \sum_{j=0}^{\infty} p_j
\end{aligned}
$$

This system of linear equations will in general not have a simple closed solution for the invariant distribution such as equation (11) and equation (12) for the simple model. But numerical solutions may be possible. Furthermore it is not a priori clear whether this solution is indeed the invariant distribution of a Markov process. This needs to be shown for the process sketched in figure 32.
One could also extend the model including the reaction time of traders to the state of the LOB. The present models assume that the traders react immediately to the state of the LOB. In reality there is always a finite reaction time due to the exchange broadcasting with a delay, the finite transmission speed of the internet and the finite reaction time of the trader. As a consequence traders necessarily react to outdated LOB states.
In this thesis we focus on high frequency trading and do not consider that open orders have an a priori infinite life time. This peculiarity of Coinbase Pro and other crypto exchanges stands in contrast to classical exchanges. Even FX exchanges with continuous 24/7 trading usually limit the life time of orders. So it would be interesting to investigate the impact of extremely long lived orders on crypto exchanges in contrast to classical exchanges.
Another problem we do not address in this theses is an error analysis for our results. This is on the

one hand due to lack of data points. But the main point is the difficulty performing a reliable error analysis with data that is far from being normally distributed. We expect that the errors themselves are not normally distributed and standard techniques will fail. A possible solution may be to resort to warping methods or to numerical fitting of the distributions.

Finally, we do not investigate different approaches to modeling LOBs which are dominated by algorithmic traders. There are many models on the market that could be adapted and tested, see for example [C11] and [PRH20]. In particular agent based models and behavioral models may be promising due to the availability of Level 3 updates that allow to investigate the fine structure of these agents (or bots). See for example [CTM19] for an agent based model for the Bitcoin market and [MF08] for behavioral models.

Appendix

A Example for LOB Order Placement

Figure 33 schematically illustrates six consecutive time steps of a LOB. At each time step the bid-price, ask-price, mid-price and spread are given as well as the incoming order type and the resulting events:

$t = 0$: At $t = 0$ we have the bid-price $b(t) = 3$, the ask-price $a(t) = 7$, the mid-price $m(t) = 5$ and the spread $s(t) = 4$.

An incoming incoming limit sell order at price $p = 6$ with volume $v = 1$ is received and generates an event (open order) $x = (+1, 6, 1, 0)$. The event is placed on the ask side of the LOB.

$t = 1$: We have bid-price $b(t) = 3$, ask-price $a(t) = 6$, mid-price $m(t) = 4.5$ and spread $s(t) = 3$.

An incoming limit buy order at price $p = 3$ with volume $v = 3$ is received and generates an event (open order) $x = (-1, 3, 3, 1)$. The event is placed on the bid side of the LOB. Note that the incoming event is added according to the FIFO principle.

$t = 2$: We have bid-price $b(t) = 3$, ask-price $a(t) = 6$, mid-price $m(t) = 4.5$ and spread $s(t) = 3$.

An order is cancelled on the ask-side at price $p = 9$ volume $v = 4$ is cancelled by the event (cancel order) $x = (+1, 9, -4, 2)$. The event is placed on the ask side of the LOB and removes the corresponding event.

$t = 3$: We have bid-price $b(t) = 3$, ask-price $a(t) = 6$, mid-price $m(t) = 4.5$ and spread $s(t) = 3$.

An incoming market buy order with funds for volume $v = 4$ is received and generates three events (matched orders) $x_1 = (+1, 6, -1, 3)$, $x_2 = (+1, 7, -2, 3 + \varepsilon_1)$ and $x_3 = (+1, 8, -1, 3 + \varepsilon_2)$. Here $\varepsilon_i < 1$ are small time intervals with $\varepsilon_i < \varepsilon_j$ for $i < j$ so the events are placed into the LOB in the order with the FIFO principle. The events are placed on the ask side of the LOB and remove the events at price $p = 6$, $p = 7$ and (partially) match with the ask-side event at price $p = 8$ with volume $v = 2$ also according to the FIFO principle.

$t = 4$: We have bid-price $b(t) = 3$, ask-price $a(t) = 8$, mid-price $m(t) = 5.5$ and spread $s(t) = 5$.

An incoming limit sell order at price $p = 3$ and volume $v = 10$ is received and generates four events (three matched orders, one open order) $x_1 = (-1, 3, -3, 4)$, $x_2 = (-1, 3, -1, 4 + \varepsilon_1)$, $x_4 = (-1, 3, -5, 4 + \varepsilon_2)$ and $x_4 = (+1, 3, 1, 4 + \varepsilon_3)$. First, the three matched orders remove the total volume at price $p = 3$ on the bid side. Then the open order is placed into the LOB at price $p = 3$ with volume $v = 1$, now on the ask side.

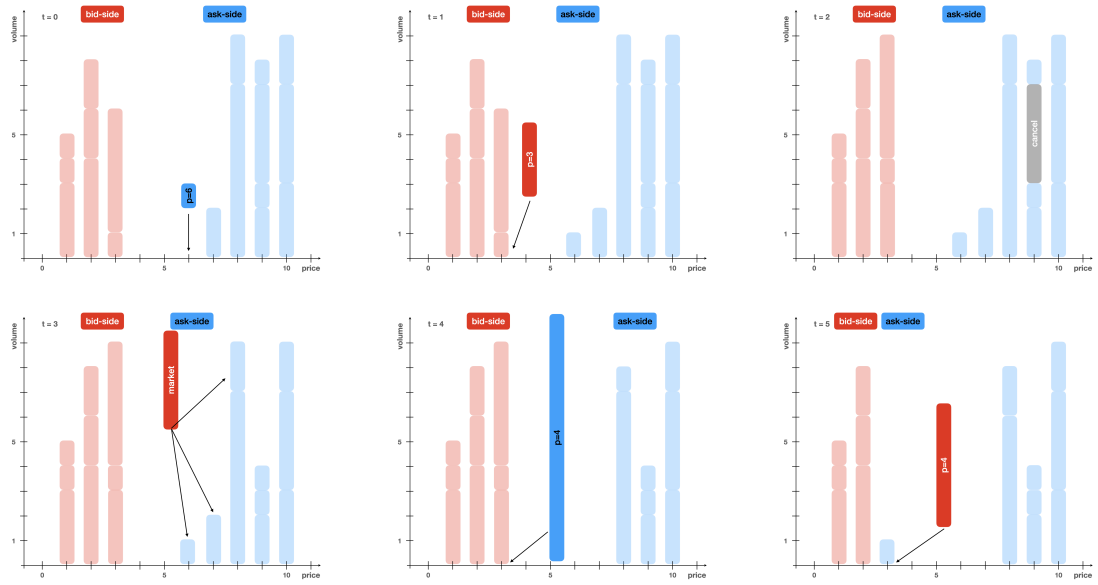


Figure 33: Six consecutive time steps illustrating the dynamics of a LOB.

$t = 5$: We have bid-price $b(t) = 2$, ask-price $a(t) = 3$, mid-price $m(t) = 2.5$ and spread $s(t) = 1$.

An incoming limit buy order at price $p = 4$ and volume $v = 4$ is received and generates two events (one matched order and one open order) $x_1 = (+1, 3, -1, 5)$, $x_2 = (-1, 3, 3, 5 + \varepsilon_1)$. First, the matched order remove the total volume at price $p = 3$ on the ask side. Then the open order is placed into the LOB at price $p = 3$ with volume $v = 3$, now on the bid side.

B Data Collection and Technical Setup

The public API of Coinbase Pro, see [CBPapi] for complete specifications, can be used to manage trading accounts, access historical data as well as live data of all currency pairs traded on the exchange.

For this thesis we are interested in real time LOB data which are broadcasted the websocket of the Coinbase Pro API. The technical setup to connect to the exchange consists of a Raspberry Pi 4 with a Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC 1.5GHz processor and 8GB SDRAM. The data stream is saved on a 3 TB raid hard disk using couchDB as a no-SQL open source database. Since the data of the API is broadcasted in JSON format the flexibility of the no-SQL database technology allows to store the data without any time consuming preprocessing. For a schematic view of the technical setup see figure 34.

The program to access the Coinbase Pro websocket is written in JavaScript and runs in a Node.js runtime environment. We use the JavaScript library CryptoCurrency eXchange WebSockets [CCXWS]

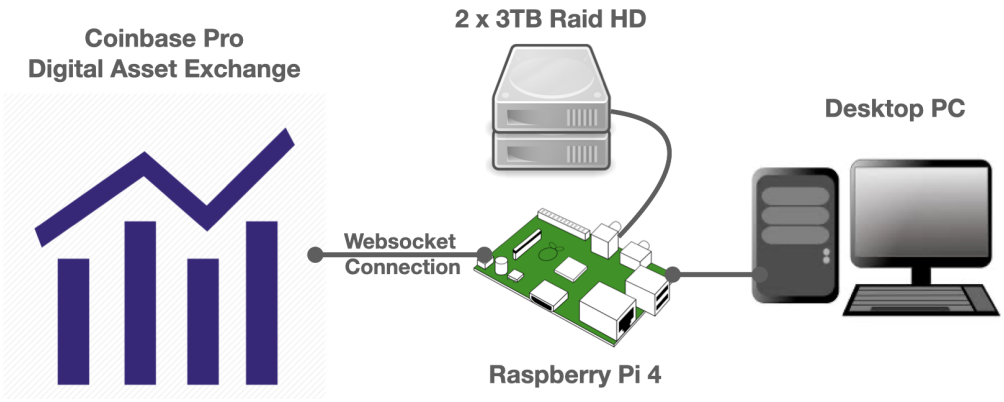


Figure 34: Technical setup for data collection from Coinbase Pro digital asset exchange via the API websocket connection.

which takes care of most of the technical issues concerning the connection to the websocket and stable retrieval of the data. The data is written into the couchDB using the JavaScript library couchdb-nano.

C LOB Data and Level 3 LOB Updates

We use two types of data types broadcasted by Coinbase Pro to build the full history of the LOB in a given time interval. For a detailed exposition of the data structure see appendix D.
The first type of data is the LOB snapshot. It consists of the full LOB at the instance in time when data collection starts. Each LOB snapshot is divided into bid side and ask side and consists of all price levels with available volume at the respective price. Since the LOB snapshots contain the complete LOB at a given instance in time they provide a considerable amount of information. But for us their main purpose will be to serve as initial conditions for the level 3 LOB updates to obtain the dynamics of the LOB.
The second type of data we are interested in are level 3 LOB updates. Each level 3 update contains a single order which belongs to a set of six order types. There are two types of order being broadcasted by the exchange upon receipt are limit orders and market orders. These orders are processed by the exchange before they appear on the LOB itself.
These processed order is then placed on the LOB in the form of one or several orders of the following order types: open orders, matched orders, cancel orders. These orders actually change the available volume at a given price level of the LOB and we will call them *events*.
The last type of order is the filled order. It does not change the LOB but indicates the end of order life cycle.
An order life cycle starts either with a limit order or a market order being received by the exchange. Let us focus first on the life cycle and information contained in the level 3 update for a limit order.

As in the case of level 2 limit orders are either buy or sell orders at a given price and size and come with the timestamp at which the order is received. For details on price increments and tick size of the OMG token see 4.2. Furthermore the level 3 limit orders contain a unique order ID which allows to track the order over its entire life cycle and a unique ongoing sequence ID which allows to ensure the completeness of the incoming orders and is thus central to ensure data quality.
Before being placed into the LOB the limit order is preprocessed by the exchange. It is checked whether the limit order can be fully or partially matched with existing orders on the opposite side of the LOB. If this is the case, the exchange will generate corresponding orders of type matched and will place them into the LOB on with a negative volume on the LOB side opposite to the limit order. The remaining positive volume will be placed as an order of type open on the LOB side of the limit order. If no matching can take place the full volume of the limit order is placed into the LOB as a limit order. All of the orders following the limit order are given the order ID of the limit order. Matched orders contain in addition the order ID of the opposing order. The life cycle of a limit order ends if either the whole volume is consumed due to matching or if, at a later time, the an order of type cancelled is received. If the volume is consumed due to matching an order of type filled is broadcasted.
A market order is built similarly to a limit order, the most notable difference being that it does not come at a fixed price. Instead a sell order states the volume to be sold and a buy order the funds for which assets shall be bought. This asymmetry ensures that the seller owns enough volume to sell if the price drops and the buyer does not exceed its funds if the price rises. After broadcasting the market order corresponding matched orders are generated and placed into the LOB with negative volume on the opposite side of the LOB. Its life cycle ends with an order of type filled. As in the case of limit orders the order ID is given to all orders following the market order.
It is clear that only events, i.e. orders of types open, matched or cancelled result in a change of the LOB state. The other order types are provide additional information but do not directly contribute to the dynamics of the LOB. Yet, they are vital to maintain a valid order book since their sequence ID is needed to ensure that no orders have been lost in transmission.
A very rare order type which we will only briefly mention is the change order. On Coinbase Pro the volume of limit orders and the price or volume of market orders can be changed by the trader before the order is placed in the order book. The time span between the broadcasting of the limit or market order and its placement in the order book is usually $\sim$1 millisecond, so change orders require an extremely short reaction time. Out of the $\sim 13 \cdot 10^6$ orders collected for this thesis, there were only 28 change orders.
From the incoming data we build the order book $\mathcal{L}(t)$ according to the rules laid out in section 2. The initial condition is given by the LOB snapshot broadcasted by the exchange when the collection of Level 3 updates is started. The updates are then added to the LOB in the order indicated by the sequence ID. If the data is complete, i.e. if no orders were lost, the resulting LOB is valid for the time period of data collection.

D API and Data Structure

The Coinbase Pro API [CBPapi] broadcasts data in JavaScript Object Notation [JSON], in short json. It a lightweight format which easily read by machines and humans and fits the needs of broadcasting LOB data which has in general no fixed length or data structure. Storing data in json format in a data base requires either a mapping to a fixed SQL compatible format or a non-SQL database. Since writing speed is an issue when receiving LOB data and parsing large amounts of data to an SQL format consumes a lot of time an computational power, we chose the open source non-SQL database couchDB [CDB]. It is compatible with Raspbian, the native Linux distribution of the Raspberry Pi.
To access the real time data stream the Coinbase Pro API provides a public websocket. We connect to the websocket, collect the data and write it into the database using a JavaScript code using the CrypotCurrency eXchange WebSockets library running on a node.js kernel. This proves to be a very stable and fast environment resulting in very little data loss.
The json files can be easily exported from the database for further analysis in Python using the json.py package to parse json files to Python dictionaries.
We do not display the different order types in json format as they are broadcasted by the Coinbase Pro websocket. Instead we give for the LOB snapshot and for LOB updates in each order type a schematic example of its main data content. In view of their scarcity we do not cover change orders.

LOB Snapshot: A LOB snapshot is broadcasted directly after connecting to the websocket. It contains the full LOB at this instant in time in terms of price-volume pairs on the bid and ask side. Furthermore it contains basic data on the currency pair:

base: OMG, quote: EUR,
bids:
(price: 3.1971, volume: 1989.2), (price: 3.1959, volume: 400.0), (price: 3.1955, volume: 1994.1), ...
asks:
(price: 3.1982, volume: 1.0), (price: 3.1991, volume: 1.0), (price: 3.2000, volume: 1926.4), ...

Note that there is no reference to the sequence ID of the first LOB update. This makes it sometimes hard to determine whether no updates have been lost and further measures of validity of the LOB have to be applied. We required that a valid LOB has never a negative volume and that the spread is always positive.

Limit Order: The main content of the limit order is its price, its volume and the side of the order book where the order is placed. It is broadcasted with an ongoing (unique) sequence ID and a timestamp in milliseconds since 1970-01-01 00:00:00 UTC. Furthermore it has a unique order ID

which will be carried by all subsequent orders. The meta data specifies the order further.

base: OMG, quote: EUR,
sequenceId: 390619818, timestampMs:1600096532629,
asks:
orderId: 6f1702a396cd4f54b19500f54a82bfbb, price: 3.2684, volume: 2.2,
meta: type: received, side: sell, order_type: limit
bids:

A limit order is pre-processed by the exchange and can generate an open order as well as matched orders if it can be executed immediately.

Market Order: Data of market orders is dependent whether it is a buy or a sell order. Apart from the sequence ID, the timestamp and the order ID a buy market order contains the total funds (in terms of the quote) for which the base currency shall be purchased. A sell market order contains the volume (in terms of the quote) which shall be sold.

base: OMG, quote: EUR,
sequenceId: 390619982, timestampMs: 1600096553785,
asks:,
bids:
orderId: 31b5144ecab74963b824eb4a32b69a44,
meta: type: received, side: buy, order_type: market, funds: 57.2162

base: OMG, quote: EUR,
sequenceId:458781526, timestampMs: 1602947732898,
asks:
orderId: 85524af7a5544d5baae4752d3a096f5e, volume: 916.2,
meta: type: received, side: sell, order_type: market,
bids:,

The asymmetry between buy and sell market orders guaranties that the oder can be executed since the available funds for purchase or volume for sale of the trader are known in advance.

Open Order: A limit order that cannot or only partially be matched is placed into the order books as an open order. The order ID is the one of the preceding limit order. A limit order actually changes the LOB.

base: OMG, quote: EUR,
sequenceId: 390619819, timestampMs: 1600096532629,
asks:
orderId: 6f1702a396cd4f54b19500f54a82bfbb, price: 3.2684, volume: 2.2,
meta: type: open,
bids:,

Matched Order: A market order will be executed immediately and will generate one or several matched orders according to its volume. A limit order can be partially executed and may also generate one or several matched orders. The order of type matched is put into the order book on the opposite side of the incoming market or limit order. Apart from the standard data such as sequence ID and timestamp the matched order contains an ongoing trade ID. Furthermore the order IDs of both orders which have been matched are submitted in the meta data section. The order ID of the matched order corresponds to the order which was placed first in the order book.

base: OMG, quote: EUR,
sequenceId: 390619983, timestampMs: 1600096553785,
asks:
orderId: 6f1702a396cd4f54b19500f54a82bfbb, price: 3.2684, volume: 2.2,
meta: type: match, trade_id: 670995, maker_order_id: 6f1702a396cd4f54b19500f54a82bfbb,
taker_order_id: 31b5144ecab74963b824eb4a32b69a44,
bids:,

Filled Order: A market or limit (open) order that ends its life cycle by being completely executed will generate an order of type done with the reason filled stated in the meta data.

base: OMG, quote: EUR,
sequenceId: 390619992, timestampMs: 1600096553785,
asks:,
bids:
orderId: 31b5144ecab74963b824eb4a32b69a44,
meta: type: done, reason: filled

Cancel Order: An open order can in principle life infinitely on the order book until it is either executed with an incoming order (see filled order) or until it is cancelled by the trader. In this case an order of type done with the reason cancelled is placed into the order book and the life cycle of the order ends.

base: OMG, quote: EUR,
sequenceId: 390613652, timestampMs: 1600095922821,
asks:
orderId: 51fe81a2019b4cb78729df23e77932d8, price: 3.2668, volume: 1940.1,
meta: type: done, reason: canceled,
bids:,

References

[AD20] A critical investigation of cryptocurrency data and analysis, C. Alexander & M. Dakos, *Quantitative Finance* 20.2: 173-188, 2020

[BBDG18] Trades, Quotes and Prices, J.-P. Bouchand, J. Bonart, J. Donier & M. Gould, Cambridge University Press, 2018

[BBHN17] Some stylized facts of the Bitcoin market, A.F. Bariviera, M.J. Basgall, W. Hasperué, & M. Naiouf, *Physica A: Statistical Mechanics and its Applications* 484: 82-90, 2017

[BTC08] Bitcoin White Paper, S. Nakamoto, https://bitcoin.org/bitcoin.pdf, 2008

[CTM19] An Agent-Based Artificial Market Model for Studying the Bitcoin Trading, L. Cocco, R. Tonelli & M. Marchesi, *IEEE Access 7*: 42908-42920, 2019

[C01] Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues, R. Cont, *Quantitative Finance*: 1 (2), 223-236, 2001

[C11] Statistical modeling of high-frequency financial data, R. Cont, *IEEE Signal Processing Magazine* 28.5: 16-25, 2011

[CBPapi] Coinbase Pro API: https://docs.pro.coinbase.com

[CBPfees] Coinbase Pro fee structure: https://pro.coinbase.com/fees

[CBPomg] Coinbase Pro OMG-EUR trading dashboard: https://pro.coinbase.com/trade/OMG-EUR

[CBPrules] Coinbase Pro Trading Rules: https://www.coinbase.com/legal/trading_rules

[CCXWS] CryptoCurrency eXchange WebSockets, github, https://github.com/altangent/ccxws

[CDB] Apache CouchDB: https://couchdb.apache.org/

[CS01] Analyzing and Modelling 1+1d Markets, D. Challet & R. Stinchcombe, *Physica A: Statistical Mechanics and its Applications* 300.1-2: 285-299, 2001

[CST10] A stochastic model for order book dynamics, R. Cont, S. Stoikov & R. Talreja. *Operations research* 58.3: 549-563, 2010

[ETH18] Ethereum White Paper A Next-Generation Smart Contract and Decentralized Application Platform, V. Buterin, https://whitepaper.io/coin/ethereum, 2018

[FVBKKMW20] Cryptocurrency Trading: A Comprehensive Survey, F. Fang, C. Ventrea, M. Basiosb, H. Kongb, L. Kanthan, D. Martinez-Regob, F. Wuband & L. Li, arXiv preprint arXiv:2003.11352, 2020

[GPWMFH13] Limit Order Books, M.D. Gould, M.A. Porter, S. Williams, M. McDonald, D.J. Fenn & S.D. Howison, *Quantitative Finance* 13.11: 179-1742, 2013

[GSTH08] Fundamentals of Queuing Theory - 4th ed., D. Gross, J.F. Shortle, J.M. Thompson & C.M. Harris, Wiley, 2008

[HHR19] Understanding cryptocurrencies, W.K. Härdle, C.R. Harvey, and R.C.G. Reule, *SSRN* 3360304, 2019

[HLR15] Simulating and analyzing order book data: The queue-reactive model, W. Huang, C.-A. Lehalle & M. Rosenbaum, *Journal of the American Statistical Association* 110.509: 107-122, 2015

[HPR19] Cryptocurrencies: Stylized facts on a new investible instrument. A.S. Hu, C.A. Parlour & U. Rajan, *Financial Management* 48(4): 1049-1068, 2019

[JSON] JavaScript Object Notation, ECMA-404 The JSON Data Interchange Standard, ECMA-404: https://www.json.org

[LF04] The long Memory of the efficient Market, F. Lillo & J.D. Farmer, *Studies in Nonlinear Dynamics & Econometrics* 8: 1-33, 2004

[M66] The Variation of certain speculative Prices, B. Mandelbrot, *Journal of Business* XXXVI:392-417, 1966

[MF08] An empirical Behavioral Model of Liquidity and Volatility, S. Mike & J.D. Farmer, *Journal of Economic Dynamics and Control* 32.1: 200-234, 2008

[MT93] Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes, S.P. Meyn & R.L. Tweedie, *Advances in Applied Probability*: 518-548, 1993

[N98] Markov chains, J.R. Norris, *Cambridge University Press*, 1998

[OMG17] OmiseGO Decentralized Exchange and Payments Platform, J. Poon & OmiseGO Team, https://whitepaper.io/coin/omisego, 2017
OMG Network: https://omg.network/

[PCP18] A new look at cryptocurrencies, A. Phillip, J.S. Chan & S. Peiris *Economics Letters* 163: 6-9, 2018

[PB03] More statistical Properties of Order Books and Price Impact, M. Potters, J.P. Bouchaud, *Physica A: Statistical Mechanics and its Applications* 324.1-2: 133-140, 2003

[PRH20] Rise of the machines? Intraday high-frequency trading patterns of cryptocurrencies, A.A. Petukhina, R.C.G. Reule & W. K. Härdle *The European Journal of Finance*: 1-23, 2020

[RR01] Small and pseudo-small sets for Markov chains, G.O. Roberts & J.S. Rosenthal *Stochastic Models* 17(2): 121-145, 2001

[RR04] General state space Markov chains and MCMC algorithms, G.O. Roberts & J.S. Rosenthal *Probability surveys 1*: 20-71, 2004

[SRK19] Testing Stylized Facts of Bitcoin Limit Order Books, M. Schnaubelt, J. Rende & C. Krauss, *Journal of Risk and Financial Management* 12.1: 25-55, 2019

[ZWLS18] Some stylized facts of the cryptocurrency market, W. Zhang, P. Wang, X. Li, & D. Shen *Applied Economics* 50(55): 5950-5965, 2018