Order book spreads, depth, and market efficiency in a general equilibrium model

Scott Condie* Brennan Platt†‡

ABSTRACT

This paper studies equilibrium order book formation in a limit-order market by building a search-theoretic model where the shape of the order book and its spread are determined jointly in equilibrium. The model characterizes liquidity as a function of differences in valuation between sellers and buyers, beliefs about the probability distribution of the arrival of buyers and sellers, exchange fees, and preference parameters like patience and beliefs about the expected lifespan of information. The efficiency of the market is characterized as a function of these parameters. We demonstrate how to extract the model's parameters from order book data and, using a sample of data from Coinbase's Bitcoin/U.S. Dollar exchange, characterize market liquidity as a function of these factors. We derive the optimal timing of frequent batch auctions in our model and show how this timing can be calculated for all limit-order markets.

Keywords: Order book, spread, price impact, equilibrium search, batch auctions

JEL classification: G14, D53

1 Introduction

In most equity markets throughout the world, market makers provide liquidity to those wishing to purchase or sell assets by placing limit orders that remain on the book until either executed or canceled. The liquidity of these markets is often characterized in two specific ways: spread and price impact. The spread is the difference between the highest bid (or buy) limit order and the lowest ask (or sell) limit order. The spread

*Department of Economics, Brigham Young University. Email: ssc@byu.edu. Telephone: +1.801.422.5306.

†Department of Economics, Brigham Young University. Email: bcplatt@byu.edu. Telephone +1.801.422.8904.

‡The authors appreciate insightful feedback from Bradley Larsen, Joey McMurray, John Stovall, Pierre-Olivier Weill and Jonathan Brogaard.

represents the cost associated with obtaining immediate execution of a trader's desired position. More liquid markets are characterized by narrower spreads.

The price impact in a limit-order market is the extent to which the price moves as greater quantities of shares are purchased or sold. One market is more liquid than another if its price impact is smaller. That is, if purchasing 100 shares in market A leads to a change in price of \$0.50 between the price at which the first share is purchased and the price at which the 100th share is purchased and in market B that price change is only \$0.20, then market B is more liquid than market A.

As a market's liquidity increases, the market has greater informational and transactional efficiency. There is evidence that a lack of liquidity is perceived by traders to be risky in the aggregate (Pástor and Stambaugh, 2003), causes higher fees associated with seasoned equity offerings (Butler et al., 2005), and leads to less informative prices (Kerr et al., 2020). These issues suggest that understanding the factors that determine liquidity in equilibrium is an important area of study.

This paper presents a model of liquidity that allows for an equilibrium characterization of both spreads and price impact. It does this without the presence of noise traders; that is, rather than having some traders who must transact regardless of price, the preferences of all participants are fully modeled. This allows us to study how changes in market characteristics (*e.g.* the relative quantity of market makers, the fees associated with providing or removing liquidity, or differences in opinion about the value of the asset) affect the equilibrium prices in the order book and hence measures of market liquidity. This equilibrium model also enables meaningful evaluation of welfare gains associated with the market, including showing the distinction between the liquidity of a market and that market's efficiency.

The model is based on a dynamic search model where buyers arrive at the market looking to immediately purchase an asset for which limit offers have been established.[1] Buyers' valuations of the asset imply that for at least some offered prices they are willing to purchase the asset, although they are not willing to purchase the asset for any offered price. If they do not purchase the asset, they wait for the next period to see if a more attractive offer is available to them.

Sellers in the market are liquidity providers and place limit orders to sell the asset based on their subjective valuation of the asset. They place limit orders at

[1]For ease of exposition, we focus only on the ask side of the market, though the bid side would proceed similarly.

prices that are optimal given the publicly-known demand parameters as well as the understanding that higher-priced limit orders will receive lower priority than lower-priced limit orders at the time of execution.

In this model we characterize steady-state order books, or distributions of offers. Indeed, prices are dispersed continuously, even though all buyers share a common view of the asset's fundamental value. We omit differing valuations from the model to highlight the interaction between order book mechanics and demand in forming the shape of the order book. Variation in the flow of market buy orders means that the probability of a limit offer being executed decreases in the price of that offer, setting up a trade off for sellers between higher markups versus a longer wait to transact. We explore the extent to which order book prices can be explained from variation in participant arrivals rather than variation in participant valuations.

This characterization allows us to answer equilibrium questions which are more difficult to answer in other equilibrium frameworks. We show, for example, that the spread and price impact both decrease when the fee paid to market makers is close to the fee charged to market takers. Intuitively, this occurs because it reduces the gap between the buyer's willingness to pay and the seller's willingness to sell; indeed, the same effect occurs after an equivalent reduction in the gap between fundamental valuations of buyers and sellers for the asset.

In contrast, the ratio of price impact to spread is affected by the distribution of participant arrivals. As it becomes more likely to have as many shares demanded as shares available for sale, this ratio increases, because the spread falls by more than the price impact. We introduce a measure of aggregate liquidity that averages the discounted probability that an entering buyer finds a seller and the probability that an entering seller finds a buyer. This measure contributes to both price impact and spread.

To illustrate the model's potential explanatory power, we apply it to order book data from Coinbase's Bitcoin/U.S. Dollar exchange.[2] Our calibration process interprets the data through the lens of our model. First, we parse the flow of updates to the order book into time periods (regimes) in which the best ask and best bid prices do not cross during the regime. Then we provide an estimation technique that estimates the average order book in each regime, while accounting for bias that naturally

[2]These data are convenient to collect and of independent interest, but the model is equally applicable to any limit-order market.

arises when the lowest price is executed first and thus is frequently unobserved. We can then extract the spread and price impact from the data and use these to estimate measures of market liquidity and implied beliefs about the distribution of market order arrivals. Compared across regimes, this process sheds light on the underlying causes of shifting order book characteristics including liquidity.

With buyer and seller preferences modeled (rather than relying on noise traders), the costs and benefits to all market participants can be quantified in welfare calculations. The theory implies that potential welfare gains are more fully realized as the flows of buyers and sellers are closer to equal. We apply our estimation of model variables to evaluate the efficiency of trading on the Coinbase BTC/USD exchange. We find the exchange typically delivers 27% of the welfare gains that could be obtained from a frictionless market where trade is instantaneous, or 45% of the welfare gains offered by a market where trade takes time but buyers and sellers are always equally balanced.

We also evaluate whether batch auctions, with their ability to reduce variance in the ratio of participants, would improve market welfare. We find that most of the time, variance is small enough that batch auctions are unnecessary, but in 5% of the regimes, a short delay (on average 303 microseconds) would improve welfare (by 3.5% on average). An exchange could utilize this by tailoring the batch auction time delay based on the current mean and variance of the order book.

We proceed in the paper by first reviewing the literature, after which we build a general model and discuss its equilibrium in section 2. In section 3 we illustrate how to apply the general model to observable order book data. Section 4 discusses the welfare implications of order book liquidity, while section 5 concludes.

1.1 Literature

Three canonical models of the order book are Glosten and Milgrom (1985), Kyle (1985) and Easley and O'Hara (1987). Each of these focus on the role of information in determining equilibrium prices — the former two coming from informed traders interacting with noise traders, and the latter from the adverse selection problem inherent in order size. In Glosten and Milgrom (1985), the arrival process of market orders is given exogenously and the best bid and offer prices are then derived optimally in a competitive environment, determining market spread. Kyle (1985) models a game of demand curve submission that can be interpreted as a primitive to forming the order book. Informed traders in the Kyle model understand the impact that their orderflow will have on the price at which their orders are executed and so they shade their bids to maximize the expected profit from their trades. This allows predictions about price impact. Easley and O'Hara (1987) models the order book as a game of adverse selection, where traders with private information about the value of the asset choose order size as part of their strategy. This allows them to uncover the information content of size decisions by traders. Our model limits supply to one share per trader, although changes in the distribution of sellers leads to changes in liquidity provision in a way that can be compared to the adverse selection model.

In contrast to these papers, we model the preferences of both liquidity demanders and liquidity suppliers (who submit limit orders, rather than supply/demand curves). Trade is motivated based on disagreements about asset values (rather than differential information). This allows us to evaluate motivations on both sides of the market and to estimate welfare effects, at the cost of increased complexity. We also generate predictions about spread and price impact from a single model.[3]

A more recent strand of literature has modeled centralized exchange in the presence of search frictions,[4] including Goettler et al. (2005); Foucault et al. (2005) and Roşu (2009). These build dynamic models of limit-order markets where all traders' preferences are fully modeled, as in our paper. In Goettler et al. (2005), traders can only cancel old limit orders with some exogenous probability. All traders agree on the value of the asset at a given time, but that value can change each period, creating a risk that a limit order is executed when it is no longer beneficial to the trader who offered it. Trade in Foucault et al. (2005) is motivated by differing waiting costs, where low cost traders post limit orders that are executed by market orders from high cost traders. By assumption, new limit orders must be price improving (i.e. narrow the spread) by the exogenous tick amount. Roşu (2009) has a similar setting except that neither a minimum tick nor price improving requirement is imposed; rather, both arise endogenously as the most profitable limit order to add to the order book.

[3]Back and Baruch (2004) establishes a connection between Glosten and Milgrom (1985) and Kyle (1985) by nesting each of these in a continuous time model and establishing conditions under which the equilibrium of their model converges to that of Glosten and Milgrom (1985) and Kyle (1985).

[4]Search theory is frequently applied the over-the-counter exchange, where buyers and sellers must find each other through a decentralized process. Weill (2020) provides a recent survey, including the relatively sparse application to centralized trade.

In the models of all three papers, the order book is dynamically built up over time as arriving traders either place limit orders or execute them in market orders. This generates predictions about low-level trends: the correlation between transaction costs and spread, the distribution of market spread and executions times, and bid and ask prices display comovement, respectively. In contrast, our model characterizes the order book in steady state, averaged over short periods where asset fundamentals are steady but arrival of traders may vary. This allows us to examine higher-level relationships between the spread, price impact, and market welfare. Our model also allows free entry and exit from limit orders, and approximates a large market with a continuum of traders. This generates an analytical solution that is particularly useful in deriving empirical implications for the order book that we can take to data.

Several search theoretic papers study liquidity in asset markets from a macro perspective. Lagos and Rocheteau (2009) studies liquidity in over-the-counter markets. This model makes predictions about liquidity measures like the bid-ask spread as in our paper, although they do not attempt to model the mechanics of price priority in limit-order books as is done here. Cui and Radde (2016) also abstracts from the specific properties of the limit-order book, embedding a financial sector with search frictions into a dynamic general equilibrium consumption-saving-investment model. Vayanos and Wang (2007) builds a search-based model of trading with the friction that traders can search for only one asset and then studies the equilibria that result. They show that in one equilibrium, short-horizon investors congregate in one market and that this market is more liquid than the other market with long-horizon investors. The question of time horizon does not enter into our model since all traders have the same trading horizon.

2 An order book model

In this paper we will explicitly model the ask (or offer) side of the order book. The bid side can be modeled analogously. The model is populated by a competitive mass of sellers, who place limit orders on the order book, and a flow of buyers who arrive randomly and place market orders to purchase the asset.[5] Buyers always transact

[5]It will be assumed later that buyers have an endogenously-determined maximum price they are willing to pay this period, above which they would prefer to wait for next period before submitting an order. This behavior is consistent with buyers' submission of marketable limit orders—that is, limit buy orders that have a price that crosses the spread. Throughout the paper we use the term

with the lowest priced order on the order book, but if multiple market orders arrive simultaneously, the order of execution is randomized, creating ex-ante uncertainty about the realized transaction price for a given buyer.

2.1 Sellers

Each seller places a limit order to sell one unit of the asset. Sellers' flow utility depends on three components. First, sellers see the relative flow value of holding the asset as d_A. Second, a seller's ask at price a will transact with endogenous probability $G(a)$ over a unit of time. Upon execution, the exchange will provide the seller a rebate f_m for providing liquidity. Finally, at Poisson rate θ, the trading opportunities cease and the seller continues to receive flow value d_A indefinitely.[6] Given this, if $V_S(a)$ denotes the seller's present expected value of placing a limit order with price a, then this trader's Bellman function is:

$$\rho V_S(a) = d_A + G(a)(a + f_m - V_S(a)) - \theta \left(\frac{d_A}{\rho} - V_S(a) \right), \tag{1}$$

which implies that

$$V_S(a) = \frac{(\rho + \theta)d_A + \rho G(a)(a + f_m)}{\rho(\theta + \rho + G(a))}. \tag{2}$$

Note that the rate of time preference, ρ, and the rate of market obsolescence, θ, have similar roles in discounting future opportunities whether they are farther away or less likely to occur. Thus, we can call their sum, $\rho + \theta$, the effective discount rate.

Since sellers are risk-neutral and competitive, they must be indifferent between any prices offered in equilibrium. This implies that for any price a in the support of the order book,

$$\frac{\partial V_S(a)}{\partial a} = 0 \quad \Longrightarrow \quad G'(a) = \frac{\rho(\rho + \theta + G(a))}{(\rho + \theta)(d_A - \rho\,(a + f_m))} G(a) \tag{3}$$

If we assume that the lowest price a_0 will transact with probability 1 each unit of

"market orders" as a short hand for these marketable limit orders.

[6]This shock occurs to all buyers and sellers simultaneously. This shock can be interpreted as buyers' information becoming obsolete, removing their interest in buying the asset. One could further model this as a shock that initiates a new steady state, with transitions between steady state regimes depicted as a Markov process, but this would have minimal impact on the model predictions with significant notational complication.

time, we can solve this differential equation for $G(a)$ to obtain

$$G(a) = \frac{\rho\left(a_0 + f_m\right) - d_A}{\rho\left(\frac{a-a_0}{\rho+\theta} + f_m + a\right) - d_A}. \tag{4}$$

Given this solution for $G(a)$, we can solve for the value function to get

$$V_S(a) = \frac{d_A}{\rho} + \frac{a_0 + f_m - \frac{d_A}{\rho}}{1 + \rho + \theta}. \tag{5}$$

This value function depends on the endogenous value of a_0 which will be determined in what follows.

2.2 Market Execution

One key feature of the order book is its order of execution. Each market buy order is first crossed with the lowest price limit order, moving upward if there is not sufficient liquidity at the initial price to cover the order. Therefore, the probability $G(a)$ of a limit order ask price of a being executed depends on the number of buyers who arrive. Define q as the ratio at any moment of shares desired to be purchased by buyers over shares offered by sellers on the order book. We assume this ratio is randomly distributed according to an exogenous cumulative distribution $F(q)$.

As an example of such a distribution, suppose that the number of shares n_a that sellers offer as limit orders are exponentially distributed with mean $1/\alpha$, while the number of shares n_b that buyers seek as market orders are independently exponentially distributed with mean $1/\beta$. The joint distribution then becomes $f(n_a, n_b) = \alpha e^{-\alpha n_a}\beta e^{-\beta n_b}$. If we define the ratio of market orders to limit orders as $q = \frac{n_b}{n_a}$ and define $\phi = \frac{\beta}{\alpha}$ (the average number of sellers per buyer), this cumulative density can be transformed to become $F(q) = \frac{\phi q}{1+\phi q}$.[7]

If the realized ratio has $q < 1$, then not all limit orders will be executed in equilibrium. If instead $q \geq 1$, then all of the limit orders on the book (below a price to be determined) will be executed, while some marketable limit orders will not be executed. Therefore, in equilibrium under rational expectations, prices a and realized

[7] Our main results do not depend on the parametrization used in this example but we will refer to it where useful for intuition.

participation q must be related by the equation:

$$G(a) = 1 - F(q). \tag{6}$$

That is, beliefs about the likelihood of a limit order being executed should depend on the distribution of buyers relative to sellers. Equation (4) can then be substituted into equation (6), and the resulting solution for a indicates the marginal price (the last share executed) after a realized ratio q of buyers to sellers in the market:

$$a = a_0 + \left(a_0 + f_m - \frac{d_A}{\rho}\right)\left(\frac{\rho+\theta}{1+\rho+\theta}\right)\left(\frac{F(q)}{1-F(q)}\right). \tag{7}$$

From this it can be seen that in equilibrium, each price a is a function of the (endogenous) minimum bid a_0, the difference between the revenue from a sale at the minimum price and the lifetime value of a seller holding the asset $(a_0 + f_m - {}^{d_A}/_{\rho})$, which in equilibrium will be weakly positive (since sellers would not place an offer that has a lower value than their lifetime value of holding the asset, the effective discount rate $\rho + \theta$, and the cumulative hazard function ${}^{F(q)}/_{(1-F(q))}$ of the ratio q. This cumulative hazard function is increasing in q, indicating that the price a is increasing in q, as expected.

In equilibrium, the highest ask price that could be executed corresponds to $q = 1$, since for all $q > 1$ all asks will be executed and some buyers will not desire to purchase the asset. Thus, we can find a_1, the maximum offered price, to be

$$a_1 = a_0 + \left(a_0 + f_m - \frac{d_A}{\rho}\right)\left(\frac{\rho+\theta}{1+\rho+\theta}\right)\left(\frac{F(1)}{1-F(1)}\right). \tag{8}$$

The probability of the price a_1 transacting is given by

$$G(a_1) = 1 - F(1). \tag{9}$$

2.3 Buyers

We now turn to the buyers' side of the market. Buyers, modeled as liquidity takers, are assumed to receive relative flow value d_B from holding the asset, where we assume that $d_B > d_A$ to ensure that the buyers who arrive want to purchase the asset from sellers.[8] If they successfully purchase the asset, buyers must also pay the additional take fee f_t. If buyers do not successfully purchase the asset below their maximum willingness to pay, they continue to search in the next period. At rate θ, the trader's information about the asset become obsolete and the buyer's search ends.

Buyers' outcomes can fall into one of three categories. If the ratio of buyers to sellers q satisfies $q \leq 1$, then all buyers get to transact (at various prices). If $q > 1$, more buyers arrive than there are shares available to sell, so a buyer can either be one of those lucky enough to purchase the asset, or they may be in the set of buyers who wait for the next period to purchase at a price they are willing to pay.

Given these three possible outcomes, the value function for a buyer is

$$\begin{aligned}\rho W_B = &\int_0^1 \left[\int_0^q \frac{1}{q} \left(\frac{d_B}{\rho} - a(s) - f_t - W_B \right) ds \right] F'(q) dq \\ &+ \int_1^\infty \left(\int_0^1 \left(\frac{d_B}{\rho} - a(s) - f_t - W_B \right) ds \right) \frac{1}{q} F'(q) dq + \theta(0 - W_B).\end{aligned} \tag{10}$$

The first term in this expression represents the change in the value that comes from successfully transacting when $q \leq 1$. Notice that while the buyer knows she will transact if $q \leq 1$, she doesn't know what her placement in the order arrival process will be, so the price she will receive is uniformly distributed over the order book. The second term in this expression represents the expected flow value to the buyer when $q > 1$, where there is some probability that the buyer won't transact at all, and if she does transact then it will be at a price that is uniformly distributed over the prices on the order book.

After substituting in the expression for the price given in equation (7), this becomes:

$$W_B = \frac{d_B}{\rho} - (a_0 + f_t) - \frac{\rho + \theta}{\rho + \theta + X_t} \left(\frac{X_m}{1 + \rho + \theta} \left(a_0 + f_m - \frac{d_A}{\rho} \right) + \frac{d_B}{\rho} - a_0 + f_t \right) \tag{11}$$

[8] The difference in flow value between buyers and sellers is fundamental to generating trade. In the analogous bid market, were it to be modeled, those placing limit orders to buy must value the asset more than those placing market orders to sell, but both would have to value the asset less than participants in the ask market.

where

$$\begin{aligned} X_t &= F(1) + \int_1^\infty \frac{1}{q} F'(q) dq \\ X_m &= \int_1^\infty \frac{1}{q} F'(q) \left(\int_0^1 \frac{F(s)}{1-F(s)} ds \right) dq + \int_0^1 \frac{1}{q} F'(q) \left(\int_0^q \frac{F(s)}{1-F(s)} ds \right) dq. \end{aligned} \tag{12}$$

Here, X_t computes the probability that liquidity takers (i.e. buyers) successfully transact and X_m is related to the probability that liquidity makers (i.e. sellers) successfully transact. For the exponential distribution introduced in the previous section, these evaluate to $X_t = \phi \log\left(1 + \frac{1}{\phi}\right)$ and $X_m = \frac{1}{\phi}\log(1+\phi)$.

In order to find the maximum amount the buyer would be willing to pay, and hence the price at which they would place their marketable limit order, we find the point at which the buyer is indifferent between a price a on the book and waiting to purchase the asset. This price is the solution a_1 to

$$\frac{d_B}{\rho} - (a_1 + f_t) = W_B. \tag{13}$$

Combining Equations (8), (11) and (13) allows us to solve for the best ask that will be placed in equilibrium. That is, we are requiring that the maximum willingness to pay of buyers be consistent with the maximum price that sellers expect to be able to charge.

$$a_0 = \frac{d_A}{\rho} - f_m + \frac{(1+\rho+\theta)}{\frac{\rho+\theta+X_t}{1-F(1)} + 1 - X_m - X_t} \left(\frac{d_B - d_A}{\rho} - f_t + f_m \right). \tag{14}$$

The denominator in Equation (14) appears frequently, which we define as:

$$L \equiv \frac{\rho + \theta + X_t}{1 - F(1)} + 1 - X_m - X_t \tag{15}$$

The term L reflects the aggregate liquidity of the market, which we explore in the next subsection. In our exponential example, this evaluates to $L = 1 + 0.5(\phi - \ln(1+\phi) + \phi^2 \ln(\frac{1+\phi}{\phi})$.

Using the solution for a_0, we can substitute into equation (7) to get the equilibrium order book relationship between the transaction price if fraction $q \leq 1$ of the available shares are transacted.

$$\begin{aligned} a(q) &= \frac{d_A}{\rho} - f_m + \frac{\left(\frac{d_B - d_A}{\rho} - f_t + f_m\right)}{L}\left(1 + \rho + \theta + \frac{(\rho+\theta)F(q)}{1-F(q)}\right) \\ &= \frac{d_A}{\rho} - f_m + \frac{\Delta V}{L}\left(1 + \frac{\rho+\theta}{1-F(q)}\right) \end{aligned} \tag{16}$$

where $\Delta V \equiv \frac{d_B - d_A}{\rho} - f_t + f_m$ is the difference in net valuations between buyers and sellers. Note that the order book can be constructed from the price function $a(q)$ on the horizontal axis and q times the order book volume on the vertical axis. With our example of exponentially distributed arrivals, $a(q)$ is a linear function of q, but in general, $a(q)$ has the same shape as $\frac{1}{1-F(q)}$. Notice that if there were Bertrand competition among sellers, each would receive $\frac{d_A}{\rho} - f_m$ for their asset, exactly compensating them for their lowest willingness to sell. As such, the term

$$\frac{\Delta V}{L}\left(1 + \frac{\rho+\theta}{1-F(q)}\right) \tag{17}$$

depicts how trading in an order book market will raise the price above this competitive outcome. The order book enables a larger markup when there is a larger difference in net valuations between buyers and sellers, ΔV, or when there is a larger the effective discount rate $\rho+\theta$ (indicating a higher cost of waiting another period to match). The markup in the order book is also affected by the unbalanced arrival of participants in two ways. Participants care about the particular realization of q as it determines transaction prices now (seen in $F(q)$). They also care about the full distribution of participant arrivals (used in computing X_m and X_t and therefore L) so they can anticipate future prices if they wait to transact in a future period.

2.4 Market liquidity

In limit-order markets, the liquidity of an asset is often defined by the spread of the asset and the price impact. From equation (16), we can calculate the spread and price impact in this market as a function of the model parameters. We define the spread to be

$$\begin{aligned} S &= a_0 - \left(\frac{d_A}{\rho} - f_m\right) \\ &= (1+\rho+\theta)\frac{\Delta V}{L} \end{aligned} \tag{18}$$

which is the difference between the best price that the order book ever offers, a_0, and the price that would come from a competitive market.

The price impact for an order that absorbs the fraction q of the order book in our model is defined to be

$$\begin{aligned} PI(q) &= a(q) - a_0 \\ &= \left(\left(\frac{d_B}{\rho} - f_t\right) - \left(\frac{d_A}{\rho} - f_m\right)\right) \frac{\rho+\theta}{\frac{\rho+\theta+X_t}{1-F(1)} + 1 - X_m - X_t} \frac{F(q)}{1-F(q)} \\ &= (\rho+\theta)\frac{\Delta V}{L}\frac{F(q)}{1-F(q)}. \end{aligned} \tag{19}$$

Equations (18) and (19) show that market liquidity as characterized by spread and price impact depend on market characteristics as summarized in the following result.

Proposition 1. *Equilibrium spread and price impact decrease, when:*

1. *the difference between buyer and seller valuations ($d_B - d_A$, in ΔV) decreases,*
2. *the fee spread ($f_m - f_t$ in ΔV) decreases,*
3. *aggregate liquidity L increases, and*
4. *traders become more patient (ρ declines) and/or information is longer-lasting (θ declines).*

These equilibrium characterizations of market liquidity demonstrate several important factors in understanding variations in liquidity across exchanges or assets. Differences in valuations between buyers and sellers can arise through tailored marketing campaigns, substantive differences in the interpretation of available public data, differences in information gathering technologies or preferences amongst market participants, differences in hedging demands due to alternative portfolio goals, or alternative tax or regulatory constraints among market participants. In situations where this heterogeneity is more extreme, we would expect markets to be less liquid.

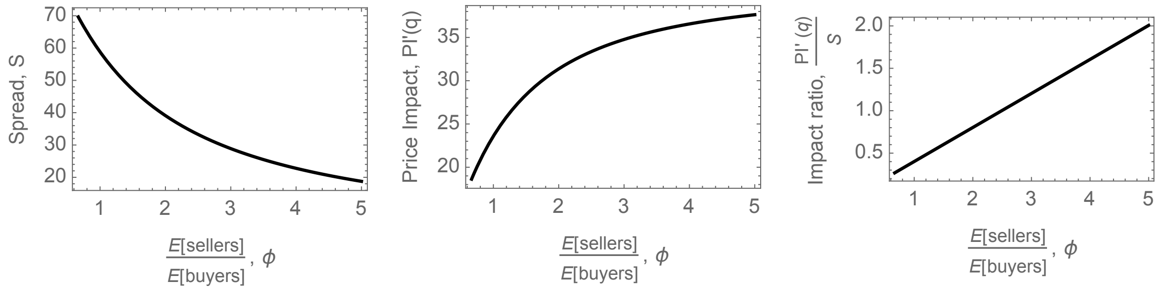


Figure 1: The equilibrium spread, price impact, and their ratio as a function of ϕ

Additionally, ΔV depends on the fee spread $f_m - f_t$ provided to market participants by the exchange. Exchanges where the fee spread is high will be less liquid, *ceteris paribus*.[9] The increased wedge between seller valuations and buyer valuations that arises from increases in the fee spread leads to less liquid markets.

Given the similar terms involved in price impact and spread, it is natural to consider the ratio of these, which is:

$$\frac{PI(q)}{S} = \frac{\rho + \theta}{1 + \rho + \theta} \cdot \frac{F(q)}{1 - F(q)} \tag{20}$$

This reveals that this ratio is more stable than price impact or spread on their own.

Proposition 2. *The ratio of price impact to spread, $PI(q)/S$:*

1. *is constant with respect to $d_B - d_A$, $f_m - f_t$, and L*
2. *decreases with more patient traders or longer-lasting information (smaller ρ or θ)*
3. *increases when buyers and sellers are more likely to be equal (smaller $F(q)$ for $q < 1$)*

Changes in the distribution of participant arrivals $F(q)$ have a more subtle effect on S or PI, in part because it also appears in aggregate liquidity L. In our exponential arrival example for $F(q)$, a decrease in ϕ causes $F(q)$ to fall at each q, raising the expected number of buyers to sellers. Changes in $F(\cdot)$ could occur due to extensive

[9] Colliard and Foucault (2012) build a model that focuses specifically on the efficiency of changes in exchange fees. They find that although smaller fee spreads increase gains from trade, when executions happen, they do not necessarily make markets more efficient.

marketing, "pump and dump" schemes, or the narrow release of insider information. We illustrate the impact of changes in $F(\cdot)$ for our parametric example in Figure 1. Note that the liquidity measures presented here need not move in tandem. The spread increases as ϕ falls, but the price impact decreases, as does the ratio of price impact to spread.

For a clearer connection between $F(q)$ and L in the general case we consider the effect of compressing the distribution F around $q = 1$. In the following result, we show that aggregate liquidity L is higher when a distribution places more weight (in a first-order stochastic dominance sense) near $q = 1$ (indicating more balance in the arrival of buyers and sellers). We show this in two parts, comparing the effect from the range where $q > 1$ and then the range where $q < 1$.

Proposition 3. *Consider two distributions of participants, F and $\hat{F}$. Let L and $\hat{L}$ be aggregate liquidity under the distribution F and $\hat{F}$, respectively. $\hat{L}$ is larger than L if:*

1. *$F(q) = \hat{F}(q)$ for all $q \leq 1$, and $F(q) < \hat{F}(q)$ for all $q > 1$, or*
2. *$F(q) > \hat{F}(q)$ for all $q < 1$, and $F(q) = \hat{F}(q)$ for all $q \geq 1$.*

Proof. **Part 1**: With rearrangement of the components of X_t and X_m, we get:

$$\begin{aligned} L &= \frac{\rho+\theta}{1-F(1)} + \frac{X_t}{1-F(1)} + 1 - X_m - X_t \\ &= 1 + \frac{\rho+\theta}{1-F(1)} - \frac{F(1)^2}{1-F(1)} - \int_0^1 \frac{F'(q)}{q} \left(\int_0^q \frac{F(s)}{1-F(s)} ds \right) dq \\ &\quad + \left(\int_1^\infty \frac{F'(q)}{q}\, dq \right) \left(\frac{F(1)}{1-F(1)} - \int_0^1 \frac{F(s)}{1-F(s)} ds \right) \end{aligned}$$

Note that the first line will be the same under F or $\hat{F}$ since they only involve $q \leq 1$. In the second line, the first integral will be larger under $\hat{F}$ as it places more weight where $1/q$ is bigger. The last parenthetical element is the same under F and $\hat{F}$, and is positive because $F(1) > F(s)$ for all $s < 1$, so $\frac{F(1)}{1-F(1)} > \frac{F(s)}{1-F(s)}$. Thus, $L < \hat{L}$.

Part 2: Using the same rearrangement, in the second line, the first integral is the same under F or $\hat{F}$. Moreover, $\frac{F(s)}{1-F(s)} > \frac{\hat{F}(s)}{1-\hat{F}(s)}$, so the second parenthetical term is larger under $\hat{F}$. In the first line, the second integral is smaller for the same reason for each q, and furthermore $\hat{F}'$ will place more weight where there is a smaller $1/q$.

Thus the double integral is smaller, and with the negative, the whole term is larger under $\hat{F}$. Thus, $L < \hat{L}$. $\square$

Of course, Proposition 3 only applies to changes in $F(q)$ that shift weight towards $q = 1$. If the distribution of participants shifts weight towards 1 in some regions and away from 1 in other regions, the impact on the aggregate measure L would be ambiguous. For example, in our exponential example, an increase in ϕ will increase $F(q)$ at each q, which mixes the first part of the proposition with the opposite of the second part. In that example, the competing effects net out such that L is strictly increasing in ϕ. That is, a greater relative abundance of sellers shows up as greater aggregate liquidity.

2.5 Comparison with the Glosten and Milgrom (1985) and Kyle (1985) models

One of the key differences between the model in this paper and the workhorse model of Glosten and Milgrom (1985) and Kyle (1985) is the absence of noise traders in the present model. This absence means that variables like the fraction of traders who are believed to be informed (as in Glosten and Milgrom (1985)) or the relative size of noise trader demand relative to fundamental uncertainty (as in Kyle (1985)) have no analog in our paper. However, the issue of spread size addressed in Glosten and Milgrom (1985) and price impact, addressed in Kyle (1985), can be compared.

Several key pieces of intuition about the formulation of the spread arise from Glosten and Milgrom (1985). These include the idea that competition amongst liquidity providers narrows spreads and that increases in the fraction of traders who are thought to be informed increases spreads. We see a similar result in terms of spread. An increase in the competitiveness of the offer side (as parameterized by an increase in ϕ in the example), leads to narrower spreads. Glosten and Milgrom (1985) also contains the result that increases in the variability of the underlying asset value (as characterized by the difference in the value of the asset in the good state of the world vs. the bad state of the world) lead to increases in the spread. While not directly analogous, in our model, differences in the valuation of market makers vs. liquidity demanders also lead to increases in the spread and a reduction in market liquidity.

Kyle (1985) characterizes the equilibrium price in the market as an increasing function of the ratio of the underlying asset variation relative to the variance of noise trader demand. If one interprets the difference in values between buyers and sellers in our model as being positively related to the underlying fundamental uncertainty about asset values, then our model and Kyle's model give results that are consistent in that price impact is increasing in this measure. Our model goes further in relating more competition among buyers to a larger price impact.

3 Estimating model parameters with order book data

In this section, we apply the model to order book data from Coinbase's Bitcoin/USD exchange, which we parse into 2480 segments of time (which we call *regimes*). We first describe key features of the data, then develop a method for extracting analogs of the model parameters from the data. We then analyze each of the regimes, first by inspecting four representative examples, then by describing the aggregate fit of the model. This allows the model to decompose the underlying factors driving the observed variation in prices and liquidity across regimes.

3.1 Data Description

The Coinbase Bitcoin/USD exchange provides a live feed of order book updates through their Websocket API. These updates provide all changes to the order book in near real-time. In order to demonstrate the technique used here, we limit our analysis to one hour,[10] which yielded 1,321,084 order book updates and 14,807 order executions. These updates occur at the rate of approximately 353 updates per second and 3.95 executions per second.

Our steady-state model is most applicable during periods when the order book is relatively stable for long enough that participants can fully adjust to market conditions. In settings with algorithmic traders, that stability can plausibly be achieved in seconds rather than (as in macro models) months or years. To identify potential steady state regimes, we begin a regime at the start of trading and track the highest best bid and the lowest best ask over time. We end a regime and start a new one when the best ask and bid cross, repeating this process until the end of the sample.

[10] Data was collected on June 20, 2023, from approximately 22:25:51.04 UTC to 23:28:18.46 UTC.

Table 1: Regime summary statistics

Regimes ($N = 2480$)	Duration (sec)	Updates (#)
Mean	1.51	533
St. Dev.	3.07	836
Min	0.00	2
25%	0.04	97
50%	0.35	244
75%	1.60	664
Max	35.93	12820

In so doing, we generate 2480 regimes. Some of these regimes are very short, lasting only microseconds, while others last for several seconds.

Table 1 gives statistics on the duration and the number of updates for these regimes. Regime duration varies with the median regime lasting about 0.35 seconds. The longest steady-state regime lasts about 36 seconds and the mean is 1.51 seconds. The median number of order book updates performed during a regime is 244 updates, with a mean of 533 updates and a maximum of 12,820 updates.[11] We anticipate that our model is most applicable to regimes with longer durations; those with extremely short durations occur when both bid and ask prices are rapidly rising or falling, which our model would interpret as a change in other fundamental parameters.

Next we report price statistics from the data period. Figure 2 shows the evolution of the best ask (price), the spread, and the price impact (computed for a purchase of 2 BTC) for the steady-state averaged order books, while Table 2 provides summary statistics of the data.

During this period, the best available ask price varied from a low of \$28,167.07 to a high of \$28,284.36, which represents a one-hour swing of about 0.38%. The mean and median price were essentially centered in that range. The spread averaged \$0.16, but over half of the regimes had a spread of \$0.01, making the distribution highly skewed. The price impact averaged a \$2.12 increase from a purchase of 2 bitcoins, but this distribution is also left-skewed, with a median price impact of \$1.81. As seen in Figure 2, the price shows a minor trend over the hour, where spread and price

[11] Note that the duration quartiles do not necessarily correspond to the same regime as the corresponding quartile in the distribution of number of updates, since the rate of updates per second is not constant.

Figure 2: Evolution of Price, Spread and Price Impact for Regime-Averaged Order books

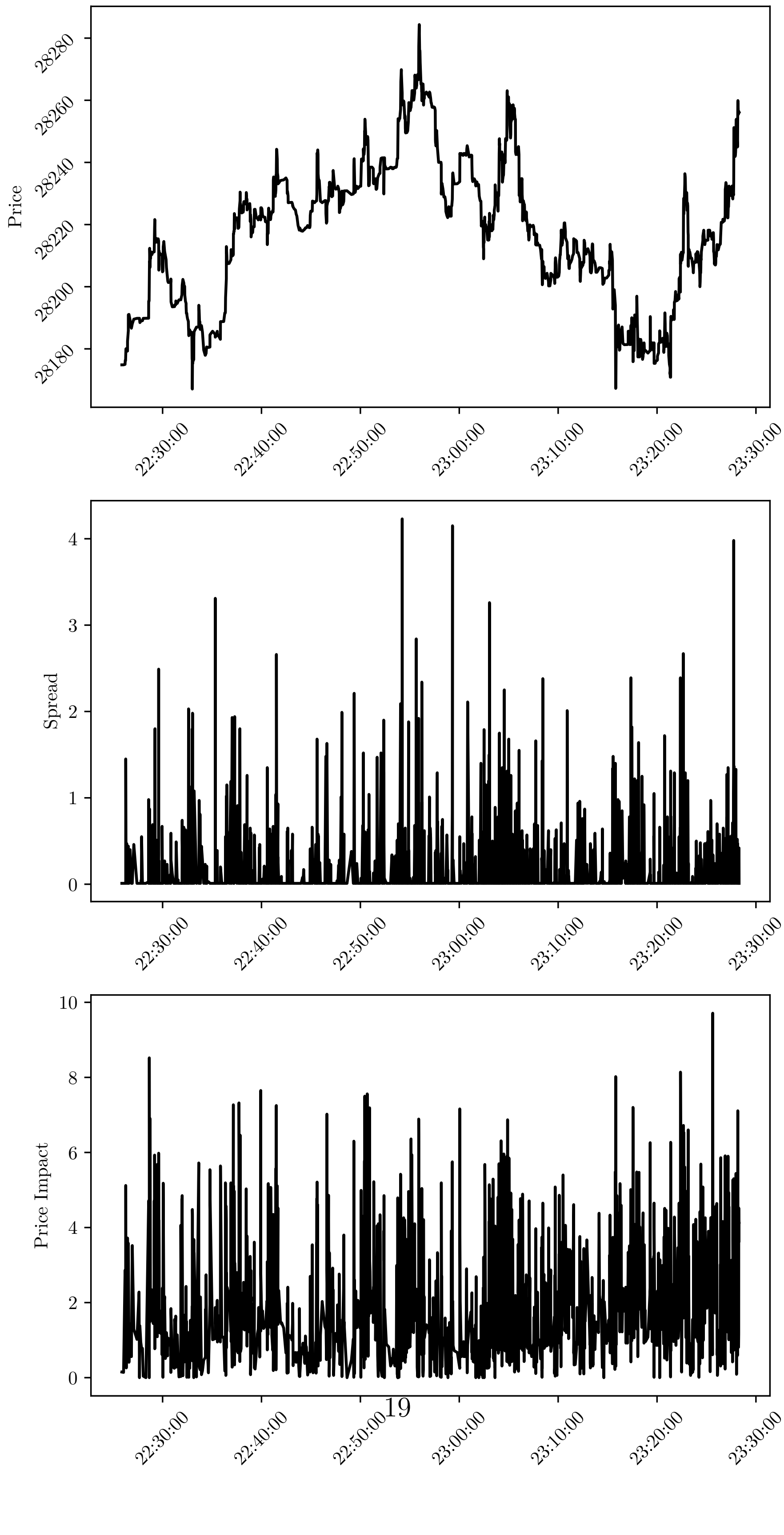


Table 2: Summary statistics of Price, Spread and Price Impact for Averaged Order books

($N = 2480$)	Price	Spread	Price Impact
Mean	\$28,218.92	\$0.16	\$2.12
St. Dev.	\$23.92	\$0.38	\$1.48
Min	\$28,167.07	\$0.01	\$0.00
Median	\$28,219.10	\$0.01	\$1.81
Max	\$28,284.36	\$4.23	\$9.71

impact are more noisy.

3.2 Interpreting data through a steady-state lens

We begin our estimation procedure by calibrating values of ρ and θ. With the rapid pace of transactions in this exchange (with time in units of seconds), we set the rate of time preference to $\rho = 0$. Meanwhile, the average regime duration of 1.5 seconds is used to calibrate the rate of information obsolescence as $\theta = 0.67$.

We next turn to parameterizing the maximum transaction rate, from which we can derive q. The median number of shares executed per second over the regimes is 0.10. Choosing the maximum transaction size $Q = 10$ implies that $q \leq 1$ for all but 255 of the 2480 regimes. We assume that larger execution rates, while possible, would be deferred to a future period so as to avoid the high price impact (as in the model). This quantity $Q = 10$ corresponds to $q = 1$ in equation (16). We then convert the cumulative quantity of shares x available on the order book (at or below a given price a) to a fraction $q = x/Q$.

Next, after each order book update, we compute the price impact for various transaction sizes and divide them by the current spread to obtain the ratio $\frac{PI(q)}{S}$ for all q. When this data is combined with ρ and θ, equation (20) allows us to solve for $F(q)$, the distribution of the ratio of buyers to sellers.

This procedure only allows us to estimate $F(q)$ for $q \leq 1$. Our model allows for the possibility that the realized q is greater than 1, meaning that the number of shares demanded by buyers exceeds the number of shares available from sellers (at an acceptable price). These excess buyers will not be observable because they do not generate a transaction. However, we do obtain an estimate for $F(1)$, and $1 - F(1)$ indicates the fraction of realizations with excess buyers.

The next step is to use the estimates of $F(q)$ to calculate X_m and X_t in equation (12), which in turn appear in the liquidity measure L from equation (15). We note that both X_m and X_t include the expression:

$$\chi \equiv \int_1^\infty \frac{1}{q} F'(q) dq. \tag{21}$$

For an interpretation, note that $X_t = F(1) + \chi$ is the average fraction of buyers who complete their purchase in a period, so χ is the portion of the average where buyers are in excess. Realizations of $q > 1$ are not observable, but would be censored at $q = 1$, thus χ is a free parameter. However, it necessarily must lie between 0 (if $F(q)$ were concentrated on very high qs) and $1 - F(1)$ (if $F(q)$ were concentrated on qs just above 1). We proceed by setting $\chi = (1 - F(1))/2$.

After calculating L, we can decompose the spread into the components ΔV and L using the spread in equation (18). In so doing, our estimation of L leads to 47 out of the 2480 regimes having a negative value of L. These regimes represent situations where the assumptions of the model are violated. Specifically, in each case they occur when there is a large gap between the lowest price and the next available price on the order book. When this second-best price has a small amount of liquidity, the price impact becomes very large for a small change in q. In such cases, the estimation of X_m is very large which leads to a negative value of L. We exclude these regimes from our analysis.

3.3 A correction for order book bias

The order book predicted by the model depicts the steady state distribution of available prices, representing the order book as pairs (a, q) of prices a and cumulative fraction of available shares q. At any moment, however, the observed order book will temporarily deviate from this steady state while some trades are transacted before being replenished by new limit orders. Indeed, market orders always cross with the most favorable limit order, so at any random moment of time, a snapshot of the ask order book will be biased towards higher prices, relative to the underlying steady state distribution. The observed distribution is missing the weight that has already been executed and has not yet been refilled.

To correct this bias in our estimate of the steady-state order book, we first take our snapshots of the order book at each point in time t, expressing the order book at price a as a cumulative distribution $q_t(a)$ by dividing the volume of limit orders at or below price a by the total volume of limit orders offered on the book up to $Q = 10$. In practice, each distribution occurs on a discrete grid of prices. Let $[a^1, a^2, \ldots, a^I]$ denote the grid of all possible ask prices in this steady state. For each order book t, let a_{0t} denote the lowest price such that $q_t(a) > 0$.

Start with the lowest price on the grid, a^1. There is no (observable) missing weight for any q_t whose lowest price $a_{0t} = a^1$, so let $\hat{q}_t(a) = q_t(a)$ for all a, and then let $\bar{q}(a^1) = mean_t\{\hat{q}_t(a^1)\}$.

Proceeding across the grid of prices, for all order books t whose lowest price $a_{0t} = a^i$, we then adjust for the missing weight below a^i by letting $\hat{q}_t(a) = \bar{q}(a^{i-1}) + (1 - \bar{q}(a^{i-1}))q_t(a)$, and compute the average at price a^i as $\bar{q}(a^i) = mean_t\{\hat{q}_t(a^i)\}$.

This procedure assumes that each of the observed snapshots come from the steady state order book, only truncated up to whichever limit orders have been executed. Averaging the snapshots is desirable because of the inherent lumpiness of actual limit orders, which can be smoothed out in this procedure. At the same time, the reweighting for missing truncated weight ensures that the resulting steady state order book $\bar{q}(a)$ is an unbiased estimate.[12]

3.4 Four Case Studies

To explore more carefully the empirical content of the model, we select four regimes that lasted at least 15 seconds from the sample. Characteristics for these regimes are given in table 3. From equation (19) we see that price impact can be decomposed into the terms $(\rho + \theta)\frac{\Delta V}{L}$ and $\frac{F(q)}{1-F(q)}$.

These four regimes have similar durations (between 17 and 22 seconds), but differ in the average spread over the period as well as the number of executions observed and the quantity of BTC transacted per second. The model calculated probability of 2 BTC or less being transacted per second (which is derived from the equilibrium conditions and the observed order book, as opposed to observed executions) varies from between 68.4% and 99.7%.

Consider the two regimes that have the highest average spreads (1 and 2). These two regimes have approximately the same duration, while regime 1 is more tradition-

[12]This process is similar to the Kaplan-Meier (KM) adjustment for left-truncated data. The key difference is that KM would take each limit order as the unit of observation, rather than each order book. If the number of limit orders in each book were the same, the processes would be identical.

Table 3: Descriptive statistics of selected regimes

	Avg. Spread	Duration	Number of executions	BTC executed per second	PI ($q = 0.2$)	ΔV	L	F ($q = 0.2$)
1	0.46	17.90	66	0.496	1.29	6.30	27.39	0.830
2	0.38	17.66	26	0.016	0.47	2.42	12.76	0.684
3	0.03	18.52	66	0.129	0.13	3.56	237.05	0.883
4	0.01	21.31	32	0.015	2.03	4.91	982.39	0.997

ally illiquid in that it has a large spread (\$0.46) and a large price impact (\$1.29). While regime 2 has nearly as large an average spread, the price impact is smaller than in regime 1 (\$0.47). The model attributes this difference to regime 1's higher ratio of valuation component (ΔV) to liquidity component L.

In contrast, regimes 3 and 4 have smaller spreads than regimes 1 or 2, leading to relatively large market liquidity measures L. Interestingly, regime 4 has the largest price impact, resulting in a large $F(0.2)$ estimate. This suggests that sellers were expecting a low flow of executions. Indeed, the difference between regimes 3 and 4 was largely a function of changes in the liquidity measure L, which is derived from changes in the shape of the order book (i.e. price impact), as opposed to changes in the spread. In contrast, regimes 1 and 2 differ mainly because of the difference in valuations of the market participants. Figure 3 shows the full estimated distribution of order flows, $F(q)$, for each of these regimes.

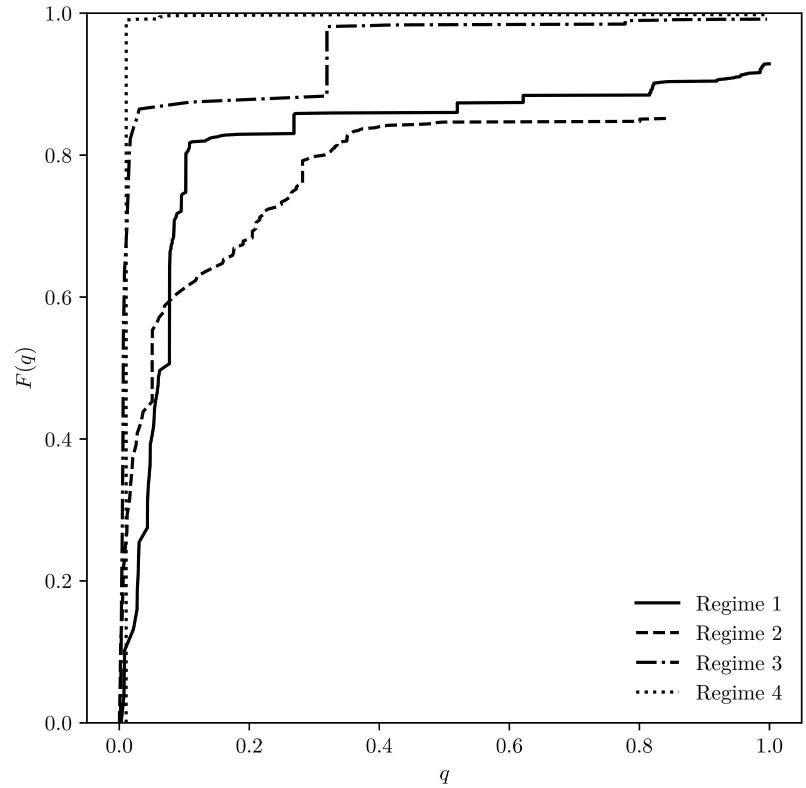
These examples demonstrate that the components that determine liquidity, ΔV, L and $F(\cdot)$ can vary across steady-state regimes in an asset market and a greater understanding of liquidity can be obtained by understanding each of these components.

3.5 The contribution of valuation differences and market frictions in explaining spread and price impact

Moving beyond these case studies, we now compare the estimated parameters across all the regimes. We primarily focus on distributional information the estimates for valuation differences ΔV and liquidity measure L, and the correlation of each with the observed spread.

Figure 4 shows the distribution of the valuation component and the liquidity component, respectively, across all of the regimes. Note that ΔV follows a roughly

Figure 3: Estimated distribution of order flows, $F(q)$, for selected regimes



normal distribution, with a mean of 7.15 and standard deviation of 3.13. In contrast, L has a large concentration with low aggregate liquidity, with 23.6% of regimes having $L < 100$. Higher levels of aggregate liquidity are more normally distributed, with a mean of 1317.95 and standard deviation of 678.9 for regimes with $L > 100$. Figure 5 depicts how these same components vary over the hour in our data. Most noteworthy there is that neither ΔV nor L has any noticeable trend over time.

Given the distinction between the valuation component of spread and the liquidity component of the spread, we now turn to the question of understanding the extent to which variation in the spread is driven by changes in ΔV or changes in L. Figure 6 plots the histograms of ΔV and L for regimes where the spread is \$0.02 or less vs. regimes where the observed spread is greater than \$0.02. The distributions for ΔV appear similar and a Kolmogorov-Smirnov test of the null hypothesis that the two

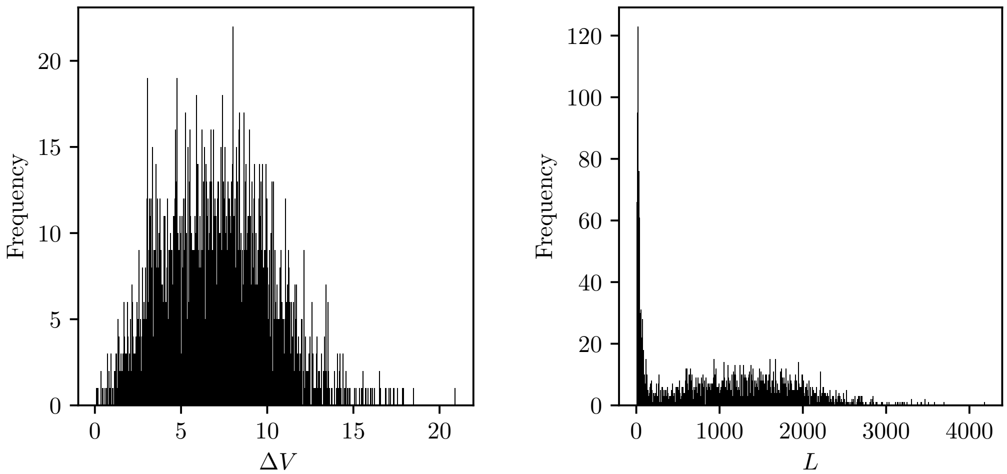


Figure 4: Histograms of ΔV and L

distributions are the same cannot be rejected at the 5% level. This suggests that the valuation component of the spread is not driving the variation in the spread.

The distributions of L are substantially different. The distribution of L for regimes where the spread is greater than \$0.02 is much more concentrated around zero than the distribution of L for regimes where the spread is small. This suggests that the liquidity component of the spread is driving the variation in the spread. A Kolmogorov-Smirnov test of the null hypothesis that the two distributions are the same yields a p-value of 2.08×10^{-322}.

Since $S = (1 + \rho + \theta)\Delta V / L$ from equation (18), and L is calculated directly, while ΔV is the residual of S not explained by L, ρ and θ, a natural question to ask is to what extent is variation in the spread explained by variation in the order book shape as encoded in the value L. A regression of the log spread on the log liquidity component (plus a constant) yields a significant coefficient of -0.922 and an R^2 of 0.905. Thus, the model-suggested summary statistic of order book shape given in L explains nearly 91% of the variation in spread in our sample. In contrast, a regression of the log price impact (at $q = 0.2$) on $\log S$ has an R^2 of 0.0002. That is, price impact alone for a particular q does not have explanatory power on its own.

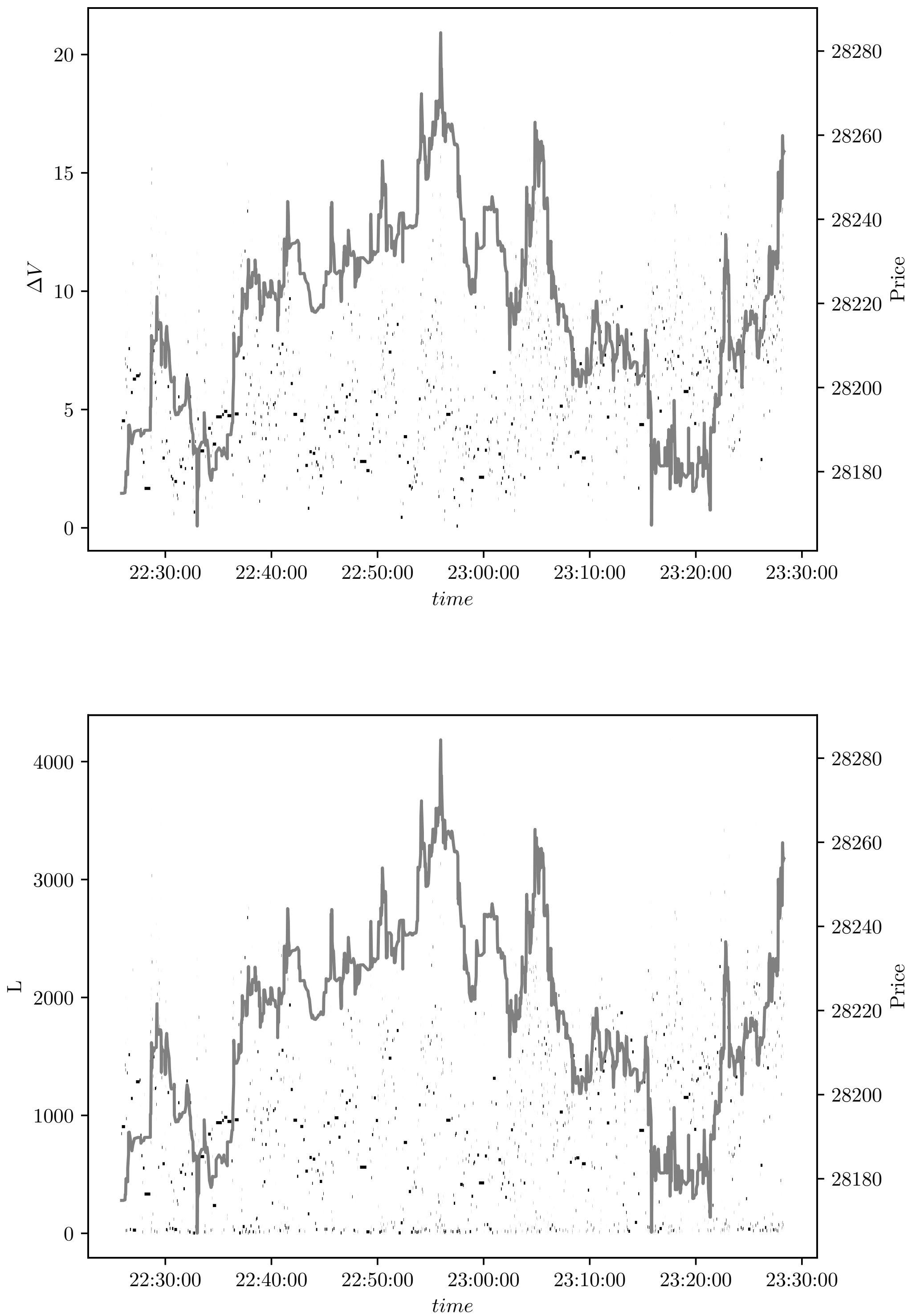


Figure 5: ΔV and price (top) and L and price (bottom) over time

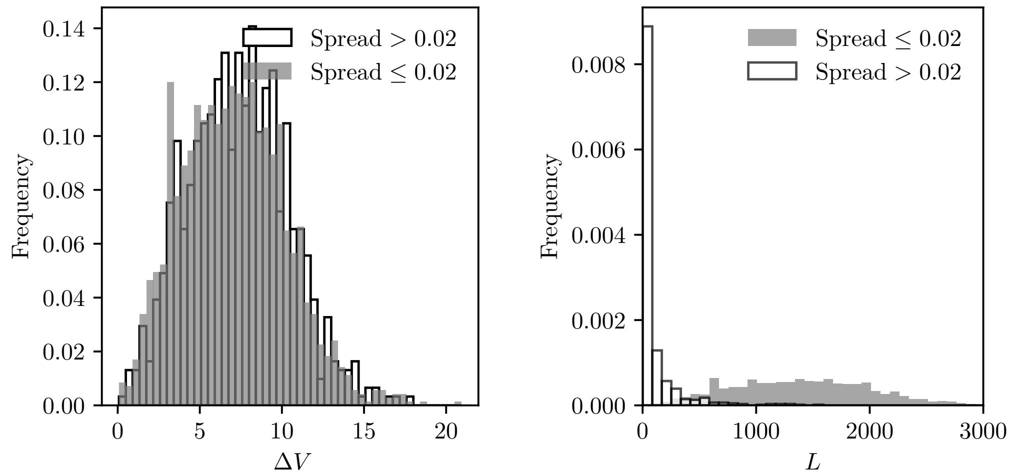


Figure 6: Histograms of ΔV (left) and L (right) for spreads less than \$0.02 (gray) and greater than \$0.02 (white)

3.6 Consistency of model predictions with data

One way to verify external validity of the model is to use executions data for the period in question and correlate it with the model-implied distribution of executions found in F. As is often the case in models that involve investor expectations, checks on the consistency of the model with observed data are a joint test of many model assumptions, including the correctness of individuals' expectations. If individuals' expectations are correct, then in regimes where $F(q)$ is high for a given q, we would expect to see a low number of executions, since $F(q)$ gives the implied probability that the fraction of the order book executed will be less than q.

A regression of $F(0.2)$ on BTC executed per second yields a coefficient of -0.265 with a t-statistic of 4.103 and an R^2 of 0.078 when the sample includes all regimes that last at least 5 seconds. This result is consistent with the model's predictions.

4 Welfare

We now consider the gains from trade in an order book market. First, we establish two extremes—one with no trade and the other with frictionless trade. Autarky would leave the asset with sellers indefinitely, generating a total welfare of $\frac{d_A}{\rho}$ for the seller. At the other extreme, if there were no frictions from the timing of trades, each seller who enters the market would immediately find a buyer, generating a total welfare of $\frac{d_B}{\rho} - f_t - f_m$. The welfare gain from this frictionless trade is the difference between this quantity and the welfare of autarky, which we have already defined as ΔV.

We compare this frictionless benchmark to the gains in an order book market. Since the realization of q changes the proportion of buyers and sellers in the market, we compute the average gains per participant at each moment in time. If the realized $q \leq 1$, then sellers are in excess and only fraction q of them will be able to transact, where each transaction yields a welfare increase of ΔV. This occurs among measure $q + 1$ participants (buyers and sellers), so the average welfare gain is $\frac{q\Delta V}{q+1}$. If instead $q > 1$, then buyers are in excess and only fraction $1/q$ of them will be able to transact. The average welfare gain is $\frac{\frac{1}{q}\Delta V}{1+\frac{1}{q}} = \frac{\Delta V}{1+q}$. Since these opportunities occur over time, we must discount them by the effective discount rate of $1 + \theta + \rho$, and we divide by ΔV to express them as a percentage of potential welfare gains, generating:

$$\%Gain(q) = \begin{cases} \frac{2q}{(1+q)(1+\theta+\rho)} & \text{if } q \leq 1 \\ \frac{2}{(1+q)(1+\theta+\rho)} & \text{if } q > 1 \end{cases} \tag{22}$$

The average gains from trade via the order book would be 0 when $q = 0$ or $q = \infty$. Effectively, there is no market because there are no participants on the other side. These gains are maximized when $q = 1$ at a level of $1/(1 + \theta + \rho)$. This falls strictly below frictionless trade when either $\rho > 0$ or $\theta > 0$ — that is, the impatience of actors or the possible expiration of information can prevent the market from fully realizing the gains from trade because transactions are delayed, possibly beyond the time when trade is useful. Moreover, welfare falls further when participants are imbalanced in either direction, since the excess buyers or sellers will not be able to transact.

The expected welfare gain $E[\%Gain]$ takes the average welfare gain across all possible realizations of q:

$$E[\%Gain] = \frac{1}{1+\theta+\rho}\left(\int_0^1 \frac{2q}{1+q}F'(q)dq + \int_1^\infty \frac{2}{1+q}F'(q)dq\right). \tag{23}$$

As with welfare gains for any specific q, the expected welfare gain is reduced by the effective discount rate, since transactions will take time. However, the variance in the arrival of buyers and sellers further limits the gains from trade in an order book market. This is illustrated in our exponential arrival example, where the gains from trade evaluate to: $E[\%Gain] = \frac{2\phi}{(1+\theta+\rho)(1-\phi)^2} \ln\left(\frac{(1+\phi)^2}{4\phi}\right)$. This is maximized at $\phi = 1$ with a value of $\frac{1}{2(1+\theta+\rho)}$. Note that this is half of the gains from trade when $q = 1$ for sure.

4.1 Comparative Statics on Welfare

The theoretical metric of welfare gains also readily lends itself to comparative statics on welfare, as reported in the following proposition. The first two results are a form of first-order stochastic dominance: a distribution that shifts weight towards $q = 1$ is always better for welfare. The second two results are a form of second-order stochastic dominance. The first says that a distribution that reduces variance among $q < 1$ is better for welfare. The second says that a distribution that increases variance among $q > 1$ is better for welfare.

Proposition 4. *Consider two distributions of participants, F and $\hat{F}$. Welfare is higher under $\hat{F}$ if:*

1. *$F(q) = \hat{F}(q)$ for all $q \leq 1$, and $F(q) \leq \hat{F}(q)$ for all $q > 1$, or*

2. *$F(q) = \hat{F}(q)$ for all $q \geq 1$, and $F(q) \geq \hat{F}(q)$ for all $q < 1$, or*

3. *$F(q) = \hat{F}(q)$ for all $q \leq 1$, and $\int_1^Q (F(q) - \hat{F}(q))dq \leq 0$ for all $Q > 1$, or*

4. *$F(q) = \hat{F}(q)$ for all $q \geq 1$, and $\int_0^Q (F(q) - \hat{F}(q))dq \geq 0$ for all $Q < 1$.*

Proof. The first two results follow directly from the fact that $\%Gain'(q)$ is positive for $q < 1$, and is negative for $q > 1$. For $q < 1$, $\%Gain''(q) = -\frac{4}{(q+1)^3(\theta+\rho+1)} < 0$, while for $q > 1$, $\%Gain''(q) = \frac{4}{(q+1)^3(\theta+\rho+1)} > 0$. Thus, in either case, $\hat{F}$ second-order stochastically dominates F. □

While the first two results are intuitive, the second two warrant further examination. For result three, this says that variance in the ratio of buyers to sellers is harmful when buyers are the limiting factor. This is because of diminishing marginal returns from raising the number of buyers: each additional buyer allows for an extra transaction (and the fixed amount of gains per transaction), but it also increases the number of participants and thereby dilutes the average. Thus, the biggest welfare gains occur when there are very few buyers.

In result four, additional buyers will not increase the number of transactions but will increase the number of unmatched buyers, thereby reducing the average; but here, that average utility falls by less the larger q gets. Thus, welfare is greater when there is greater variance among $q > 1$. For instance, suppose that there was a 20% chance of getting exactly $q = 2$, with all other outcomes occurring for $q < 1$ (e.g. $F(1) = 0.8$). A mean-preserving spread that places 0.1 at $q = 1$ and 0.1 at $q = 3$ will increase the expected welfare by $0.1 * \frac{2}{2(1+\theta+\rho)} + 0.1 * \frac{2}{4(1+\theta+\rho)} - 0.2 * \frac{2}{3(1+\theta+\rho)} = 0.2 * \frac{1}{12(1+\theta+\rho)} > 0$.

It is worth recalling that Proposition 3 had a similar result as the first two claims of Proposition 4. That is, both aggregate liquidity, L, and welfare, $E[\%Gain]$, are higher when the distribution of participants are more concentrated near $q = 1$. This makes it tempting to conflate liquidity with welfare, assuming they always move in tandem, but this is not the case. Welfare is fully driven by balance in market participants, which ensure that both buyers and sellers will quickly transact. Aggregate liquidity is derived in part from the prices at which they transact, so having relatively more sellers results in lower prices and a larger L. This can be seen in our exponential example: welfare is maximized when $\phi = 1$, but L is strictly increasing in ϕ, even when ϕ is very large so that very few sellers encounter a buyer.

4.2 Estimation of welfare gains

We now utilize this measure of welfare applied to our order-book data. Table 4 shows statistics for the distribution of welfare gains across regimes in our sample and Figure 7 plots this histogram. For the data collected, these gains are centered around 27% of the welfare that would be possible in perfectly frictionless markets. A positive effective discount rate (parameterized at $\rho + \theta = 0.67$) introduces a time friction; if the flow of buyers and sellers were always equated ($q = 1$), this balanced frictional market could realize $1/(1 + 0.67) = 59.9\%$ of the frictionless market welfare. This table and figure demonstrate that, for these data, the realized welfare gains are tightly packed between 25% and 30% of a frictionless market, or between 40% and 50% of a frictional but balanced market.

The first column of Table 5 gives point estimates and standard errors for a regression of the realized welfare gains on the spread and price impact. Common intuition says that more liquid markets are more economically efficient. This regression shows that by our measure of welfare, this is partially true. Controlling for spread, markets

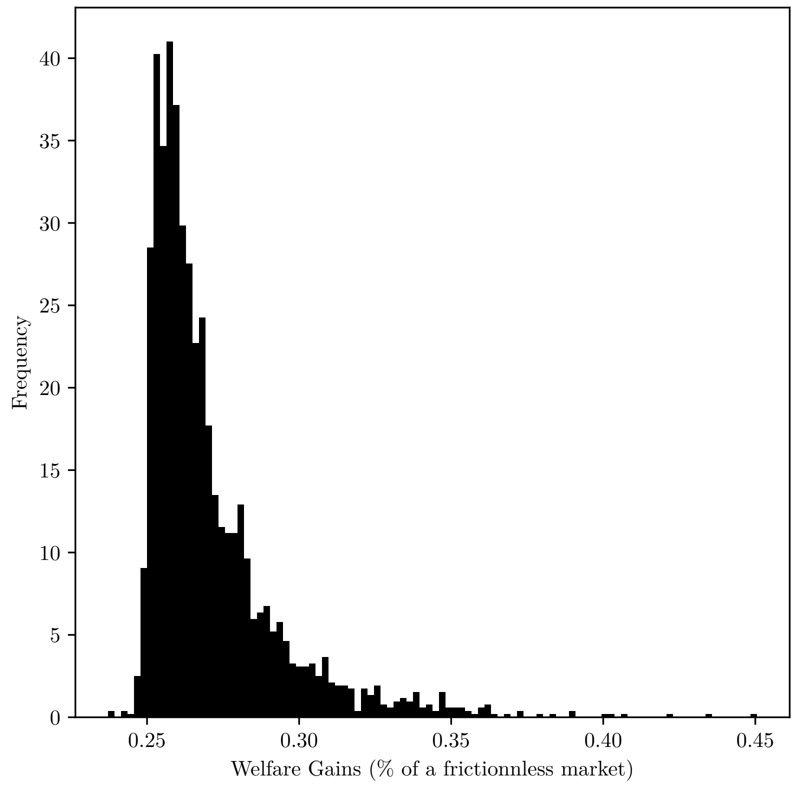


Figure 7: Estimated distribution across steady-state regimes of the realized fraction of potential welfare gains.

with lower price impacts will achieve more of their potential welfare gains. However, in this sample, the welfare impact of spread is positive. That is, when controlling for price impact, markets with larger spreads realize a larger fraction of potential welfare gains.

To understand why this happens, we decompose the spread into its components ΔV and L in column 2, as well as other metrics used in fitting the model to the data. When these covariates are added, the coefficient on spread becomes insignificant. Instead, more aggregate liquidity (which Section 3.5 showed is associated with a smaller spread) reduces the realized fraction of potential welfare gains by a small but significant amount. This underscores that liquidity L can increase without creating more balance in the market, which ultimately drives welfare. This can also be seen

Table 4: Fraction of welfare gains across regimes

	Percent of potential welfare gains
Mean	0.271
St. Dev.	0.023
Min	0.237
Median	0.264
Max	0.451

in the significant negative coefficient on $F(0.2)$; having many sellers who only match with buyers 20% of the time is inefficient. In contrast, a larger $F(1)$ has a significant positive coefficient, indicating greater balance in the flow of sellers and buyers.

Regimes with a larger ΔV had a small but significant negative impact on welfare. The theory does not provide any hypothesis connecting ΔV and $E[\%Gain]$, though we emphasize that this is the effect on percentage of welfare gains relative to a frictionless market. In terms of dollar levels, ΔV *is* the maximum potential gains, and the realized level of frictional gains are proportional to ΔV. We also note that, after including these other variables, the realized duration of the regime is not significant. That is, long regimes are not more efficient than short ones. All together, these observables explain roughly 62% of the variation in welfare gains across regimes.

4.3 Welfare impact of batch auctions

Proposition 4 indicates above that exchanges offering order book markets have a natural incentive to keep parity between the flows of buyers and sellers. For instance, advertising or discounts could be targeted towards potential participants on the binding side of the market. Even if the average participant ratio cannot be significantly shifted, the third and fourth claims suggest that the platform could try to reduce variance when buyers are binding. For instance, rather than continuously executing market orders, they could be held to accumulate for a small period of time.

This insight then offers possible welfare gains from implementing frequent batch auctions of the type discussed in Budish et al. (2015), under some circumstances. If q is centered around something less than 1 (as it is in all our regimes), then allowing traders to accumulate before matching decreases the variance in the number of buyers

Table 5: The determinants of welfare gains

	(1)	(2)
Spread	0.018***	0.0017
	(0.0010)	(0.001868)
$PI(0.2)$	-0.0069***	-0.0046***
	(0.00026)	(0.00031)
ΔV		-0.00072***
		(0.000168)
L		-0.000002***
		(0.000001)
$F(1)$		0.10***
		(0.013)
$F(0.2)$		-0.12***
		(0.0029)
Duration		0.00017
		(0.000095)
Constant	0.28***	0.29***
	(0.00068)	(0.012)
R^2	0.306	0.616

which increases welfare, given the concavity in the $\%Gain$ to the left of $q = 1$. However, this also slows down the rate at which transactions occur, running the risk that information becomes obsolete (the regime ends) before the transaction occurs. This sets up a tradeoff between gains from lower variance versus lost opportunities to transact.

For a simple illustration of this, if transactions currently occur at Poisson rate λ, consider a batch auction held every α/λ units of time, where $\alpha > 1$. Over that time span, q will be realized an average of α times, so the aggregation of arriving buyers and sellers generates a batch auction with a q_α that is the average of the realized qs. Note that $E[q_\alpha] = E[q]$, while $Var[q_\alpha] = Var[q]/\alpha$. At the same time, there are fewer opportunities to trade before the regime ends. We represent this by scaling up the effective discount rate to $\alpha(\theta + \rho)$.

We incorporate this batch auction rate into welfare gains in Equation (23) by taking a second-order Taylor series approximation of the integrand, evaluated around the mean of q. This means $\frac{2q}{1+q} \approx \frac{2E[q]}{1+E[q]} + \frac{2}{(1+E[q])^2}(q - E[q]) - \frac{2}{(1+E[q])^3}(q - E[q])^2$. When the integral is evaluated, we obtain expected welfare as a function of the batch auction delay rate α:

$$\frac{2}{(1 + \alpha(\theta + \rho))(1 + E[q])} \left(E[q] - \frac{Var[q]}{\alpha(1 + E[q])^2} \right) \tag{24}$$

This welfare approximation has the advantage of relying on easily computable moments of the order book, allowing an exchange to potentially adjust the batch auction to fit current conditions. For instance, if $\frac{\theta}{1+2(\theta+\rho)}E[q](1 + E[q])^2 > Var[q]$, then welfare is strictly decreasing in α, so it is optimal to not impose any batch auction (set $\alpha = 1$). Otherwise, the welfare-maximizing batch auction time period is:

$$\frac{\alpha^*}{\lambda} = \frac{Var[q]}{E[q](1 + E[q])^2} \left(1 + \sqrt{1 + \frac{E[q](1 + E[q])^2}{Var[q](\theta + \rho)}} \right), \tag{25}$$

where λ is the average rate of executions, which is 3.95 per second in the data (or an average of 253 ms between each transaction).

When applied to our data, we find that batch auctions are unnecessary in 95% of regimes. This occurs when variance in participant arrival q is quite small, usually because the distribution $F(q)$ is highly concentrated near $q = 0$, as with Regime 4 in our case studies. Among the 5% of regimes where batch auctions are beneficial, the left panel of Figure 8 plots the distribution of the optimal batch auction length, while the right panel plots the distribution of welfare gains from implementing an optimal batch auction (in percentage terms relative to welfare without a batch auction). For those regimes where batch auctions are beneficial, the average batch auction delay is modest at 303 ms (which is 20% longer than the average rate of transactions), but boosts welfare by an average of 3.5%, though some regimes could improve welfare by over 45%.

A practical implementation could trigger a batch auction delay whenever conditions of the order book warrant it. By extracting the mean and variance of the order book $F(q)$, the exchange could instantaneously evaluate whether bath auctions are

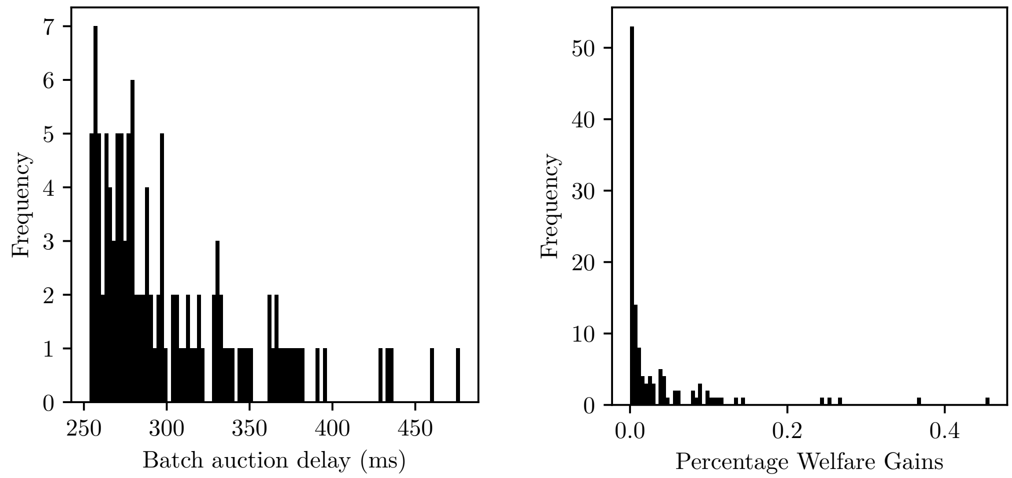


Figure 8: Histograms of optimal batch auction delay α^*/λ (left) and percentage welfare gains from using batch auctions (right, relative to welfare without batch auctions), for regimes where batch auctions are welfare improving.

needed and for how long, using equation (25).

5 Conclusion

This paper builds a model of limit order books that characterizes both spreads and price impact in general equilibrium. The model predicts that market liquidity increases as asset valuations converge, fee spreads shrink, and the number of buyers and sellers become more equal. We derive a welfare formula that measure the efficiency of the market as the fraction of potential gains from trade that are realized, then use this formula to derive conditions under which instituting frequent batch auctions will be welfare improving. Additionally, we build techniques for estimating model parameters from order book data and use them to estimate efficiency in our sample.

Applying the model to data from the Coinbase BTC/USD exchange, we find plausible time periods (regimes) during which the market appears to be in steady state, and use the model to interpret differences among them. We find that changes in market liquidity as measured by the liquidity measure L, calculated as a summary statistic of the whole order book, account for about 91% of the variation in the spread. In terms of welfare, we find that the most liquid steady-state regimes capture just over one fourth of the welfare gains possible in a completely frictionless market, with relatively little variance. We also find that in our sample about 5% of regimes would benefit from instituting frequent batch auctions.

Further work in this area could investigate the liquidity of alternative exchanges as a function of the aggregate liquidity measure L, derived from the shape of beliefs given in F. This work would allow us to characterize the pricing and welfare efficiency across exchanges and allow for the characterization of the optimal horizon for frequent batch auctions. Such an effort could yield insights into how exchanges compete, the welfare gains and/or losses associated with this competition, and the role of market regulation in promoting such competition.

References

Back, Kerry and Shmuel Baruch (2004): "Information in securities markets: Kyle meets Glosten and Milgrom," *Econometrica*, 72 (2), 433–465.

Budish, Eric, Peter Cramton, and John Shim (2015): "The high-frequency trading arms race: Frequent batch auctions as a market design response," *Quarterly Journal of Economics*, 130 (4), 1547–1621.

Butler, Alexander W, Gustavo Grullon, and James P Weston (2005): "Stock market liquidity and the cost of issuing equity," *Journal of Financial and Quantitative Analysis*, 40 (2), 331–348.

Colliard, Jean-Edouard and Thierry Foucault (2012): "Trading Fees and Efficiency in Limit Order Markets," *Review of Financial Studies*, 25 (11), 3389–3421.

Cui, Wei and Sören Radde (2016): "Money and asset liquidity in frictional capital markets," *American Economic Review*, 106 (5), 496–502.

Easley, David and Maureen O'Hara (1987): "Price, trade size, and information in securities markets," *Journal of Financial Economics*, 19 (1), 69–90.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel (2005): "Limit order book as a market for liquidity," *Review of Financial Studies*, 18 (4), 1171–1217.

Glosten, Lawrence R and Paul R Milgrom (1985): "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of Financial Economics*, 14 (1), 71–100.

Goettler, Ronald L, Christine A Parlour, and Uday Rajan (2005): "Equilibrium in a dynamic limit order market," *Journal of Finance*, 60 (5), 2149–2192.

Kerr, Jon, Gil Sadka, and Ronnie Sadka (2020): "Illiquidity and price informativeness," *Management Science*, 66 (1), 334–351.

Kyle, Albert S (1985): "Continuous auctions and insider trading," *Econometrica*, 1315–1335.

Lagos, Ricardo and Guillaume Rocheteau (2009): "Liquidity in asset markets with search frictions," *Econometrica*, 77 (2), 403–426.

Pástor, L'uboš and Robert F Stambaugh (2003): "Liquidity risk and expected stock returns," *Journal of Political Economy*, 111 (3), 642–685.

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

Vayanos, Dimitri and Tan Wang (2007): "Search and endogenous concentration of liquidity in asset markets," *Journal of Economic Theory*, 136 (1), 66–104.

Weill, Pierre-Olivier (2020): "The Search Theory of Over-the-Counter Markets," *Annual Review of Economics*, 12, 747–773.