Dynamic Trading: Price Inertia and Front-Running*

Yuliy Sannikov
Graduate School of Business
Stanford University
sannikov@stanford.edu

Andrzej Skrzypacz
Graduate School of Business
Stanford University
skrz@stanford.edu

December 7, 2016

Abstract

We build a linear-quadratic model to analyze trading in a market with private information and heterogeneous agents. Agents receive private taste/inventory shocks and trade continuously. Agents differ in their need for trade as well as the cost to hold excessive inventory. In equilibrium, trade is gradual. Trading speed depends on the number and market power of participants, and trade among large market participants is slower than that among small ones. Price has momentum due to the actions of large traders: it drifts down if the sellers have greater market power than buyers, and vice versa. The model can also answer welfare questions, for example about the social costs and benefits of market consolidation. It can also be extended to allow private information about common value.

1 Introduction.

A market with heterogeneous investors – large institutions, small retail investors, liquidity providers and high-frequency traders – presents many puzzles. What determines the speed of trading? What is the link between microstructure and time series properties of prices, such as momentum and excess volatility? What is the price impact of large trades, and how much does it matter for optimal execution of transactions? What about phenomena such as front-running – are they detrimental

*We are grateful to seminar participants at Yale, the New York Fed, Cambridge, Stanford, Latin American Econometric Society Meetings, LBS, Harvard, MIT, Northwestern, Chicago, NYU, Georgetown, Duke, the University of Helsinki, the University of Lausanne, Caltech, Collegio Carlo Alberto and Universidad de Chile for helpful comments. We are also grateful for research assistance of Erik Madsen, Giorgio Martini, and Sergey Vorontsov

to welfare? What about high frequency trading? Is it justified that in practice, certain market-makers are willing to pay to distinguish institutional trading flow from that of retail investors?

We attempt to build a game theoretical framework to analyze these phenomena. Specifically, we model a market in which the price of a risky asset follows a Brownian path as a result of trades of individual market participants. Trade can be motivated by various reasons, such as risk sharing in Vayanos (1999) or because of heterogeneous beliefs as in Kyle, Obizhaeva and Wang (2014). Players have private information about their personal desire to buy or sell at any price, and they trade, optimizing between the price at which they trade and the costs of delayed execution. Thus, from the perspective of an individual, our decision makers are similar to a trader in Almgren and Chriss (2001). This classic financial mathematics paper solves the problem of a trader who decides how quickly it wants to sell a desired quantity, facing the trade-off between price impact and uncertainty in the execution price. Fast sales lead to a low execution price, but waiting exposes the player to price risk that the player is trying to offload. This problem is motivated by basic empirical observations about the price impact of trade.

While individual players in our model face a problem similar to that of Almgren and Chriss (2001), we aim to derive the properties of the market that individuals face from the interaction of individual behaviors. That is, we show that these market properties can be derived in a game theoretic framework and we link the price impact of trades as well as time series properties of prices to the composition of the market.

The price properties we derive can be broadly classified as "on-equilibrium" and "off-equilibrium" (path). On equilibrium, we would like to know how prices behave when everybody follows their optimal strategies. One important on-equilibrium property we derive is momentum, which depends on the relative competitiveness/market power of current buyers and sellers. Off equilibrium, putting ourselves in the shoes of an individual player, we would like to know how prices would respond if the player traded differently, not according to the optimal strategy. Of course, the optimal strategy depends on the off-equilibrium properties of prices, i.e. the price impact that various trades would have. That is, equilibrium strategies have to be optimal given the properties of prices off equilibrium. Off equilibrium, we show that as (assumed) in Almgren and Chriss (2001), trades have "instantaneous" and "permanent" price impacts, i.e. price depends on the trading rate as well as the total amount bought or sold. In addition, depending on market composition, there may also be "transient" price impact, i.e. the temporary impact on the price of the total amount sold. Transient impact exists empirically, and in our model it arises because transactions by an individual player trigger trade among other players.[1]

The dichotomy between on and off-equilibrium phenomena translates to the types of data that one needs to test various hypothesis empirically. On equilibrium, one

[1]For empirical research on price impact in equity markets see for example Almgren et al (2005), Moro et al (2009) and a summary in Bouchaud (2010).

simply needs the time series data of actual market transactions. The data for off-equilibrium phenomena is much less readily available, and ideally would involve experiments by banks that record how prices react depending on the rates of trade.

Game-theoretic justifications of price impact go back to the classic paper of Kyle (1985), where the market maker tries to filter out private information from the combined trades of noise traders and an informed trader. Kyle's "lambda" captures the permanent price impact of trades based on the private information that they carry. A large literature that uses noise traders includes Kyle (1989), Back (1992) and Caldentey and Stacchetti (2010).

Our model builds upon the seminal paper of Vayanos (1999), which models a symmetric market of fully rational traders who have private information about their desire to buy or sell. It is a finite market in which trade takes time, as players signal their supply or demand by the rate of trading.[2] Slow trade leads to inefficiency, as transactions have both instantaneous and permanent price impacts. On equilibrium prices have no momentum when there is symmetry among players. A number of recent papers work within such a symmetric framework. Du and Zhu (2013) study the impact of the frequency of trade on efficiency, and explore how the speed of trade depends on the amount of private information in the market. Kyle, Obizhaeva and Wang (2014) introduce belief heterogeneity, study the speed of trade and observe phenomena related to "flash crashes."[3]

We introduce asymmetry among players into this class of models. Players differ in their risk capacity parameter that determines their impatience to trade. We refer to traders who prefer to trade quickly due to a high holding cost as "small" and players who are willing to wait and tolerate execution risk as "large." As we show, heterogeneity of traders brings many important issues to the table, but also significantly complicates the modeling task. Unlike in symmetric models, price is no longer a sufficient statistic about the private information of others; to optimize, each player wants to know the distribution of trades across large and small players. When this information is unavailable, optimization of individual players involves a filtering problem to infer the distribution of supply and demand from price behavior, and to forecast future price momentum from that distribution. With awareness that such a framework leads to a host of new phenomena that may be difficult to disentangle, we design a trading mechanism based on the assumption that players know and can condition on the trades of all other participants.

Our model of trade for asymmetric environments coincides with that of Vayanos (1999) in symmetric case, but otherwise corresponds to the assumption that trading flow is not anonymous. Our model of trade presents a simple analytical framework - we derive a fully separating equilibrium in which players signal their private in-

[2]That traders signal by choosing how much to trade is familiar from the static models like Leland and Pyle (1977) or Myers and Majluf (1984).

[3]A few papers also study *static* trading with heterogeneous strategic traders; see for example Lambert, Ostrovsky and Panov (2016) and Malamud and Rostek (2014).

formation about their desire to trade through individual trading flows. We think of our model as an important benchmark - through which we can identify a host of phenomena without the complications of belief formation. Also, as in any separating equilibrium, our characterization is invariant to the distribution of shocks to the players' trading needs, and it is even invariant to the correlation among the shocks. A tractable framework for the analysis of markets with heterogeneous participants is a methodological contribution of this paper.

An important implication of heterogeneity is that, when trading is not anonymous, not only trading speed but also price impact depend on player size. Patient/large players are willing to trade slowly, as for any quantity they want to sell they have lower holding costs than small players. Hence, trades of large players are more "toxic:" they generate price momentum that is detrimental to anyone on the other side of the trades. While these observations hold regardless of whether the source of trades is observable, observability implies that the trades of large players have larger instantaneous price impact, i.e. the same quantity generates a bigger change in price. In practice this leads to a variety of behaviors, as players try to obtain information about the source of trade. In the popular press, Patterson (2012) and Lewis (2014) document how high-frequency traders use "latency tables" and "router signatures" to identify the source of flow. Certain market makers, such as Citadel, pay retail brokerages for flow from their investors - the knowledge about the source of the flow seems hugely valuable. Phenomena such as technical front-running are related to identifying large traders.

We show that when the source of trades is observable, instantaneous price impact has a large sensitivity to player size. A trader who is 10 times as large as another trader can have instantaneous price impact that is 3 times as large. Of course, in practice techniques such as splitting of orders allow large traders to hide flow behind that of small traders at least temporarily. However, recent developments in high frequency trading and the general push for transparency in markets move reality closer to our model.[4] The question about the role of transparency is important, but it would be a subject of a separate paper.

Since we explicitly model preferences of all market participants, our model can be a laboratory for studying welfare. Here we find several surprising results. First, with transparency, market power is bad for individual welfare. Market power refers to the ability of a large institution to coordinate its trades, avoiding the competition that would occur if correlated trades were initiated by many dispersed traders. Market power implies slower trades, but it also leads to higher price impact in a transparent market of rational traders who can identify a "whale in the ocean." We present a set of results, demonstrating that large players may prefer to commit to faster trade, and would benefit from being split into smaller traders. These results potentially

[4]For example, NYSE has a set of liquidity programs that allows retail orders to be identified as such, in order to generate greater transparency and price improvement - see www.nyse.com/markets/liquidity-programs.

explain the frustration of large institutions with high frequency traders who try to benefit from identifying large traders, described in Lewis (2014).[5] They suggest that transparent markets may skew the field in favor of small traders, whereas in opaque markets large players are able to hide trades behind those of small traders successfully, and benefit from market power.[6]

This paper is organized as follows. Section 2 presents the linear-quadratic model of traders' preferences, a conditional double auction, and equivalence results for different representations of trading rules. Section 3 describes equilibrium conditions and introduces a model of a competitive fringe. Section 4 provides closed-form characterization of equilibrium in the case of a single large trader and a competitive fringe, discussing on and off path dynamics, as well as welfare properties of the market. Section 5 discusses the general case of N asymmetric traders and includes approximations to equilibrium strategies. It also expands discussion of technical front-running. Section 6 briefly illustrates how the analysis could be extended to common values. Section 7 discusses our findings and Section 8 concludes. Some of the proofs are in the main text of the paper and some of them are relegated to the appendix. The appendix also contains a microfoundation of linear-quadratic preferences with a CARA-utility maximizing traders.

2 The Model.

Consider a market for a single divisible asset. There are either N large traders, or $N-1$ large traders and a "competitive fringe" of infinitely many small players. We model the players' incentives to trade by following financial mathematics literature. While that literature takes the price impact as given, in our model the equilibrium endogenously determines the magnitude and form of the price impact. We start with a brief review of a classic paper from this literature, in order to facilitate the interpretations of the stylized features of the linear-quadratic model that we build later.

Background: Price Impact and Optimal Execution. Almgren and Chriss (2001) build an elegant model of optimal execution of transactions, which takes into account the trade-off between average price and execution risk. The trader would like to liquidate X_0 shares and the execution price is modeled as

$$\hat{p}_t = \hat{p}_0 + \sigma\, z_t - I\, q_t - \Lambda \int_0^t q_s\, ds, \tag{1}$$

[5]Budish, Cramton and Shim (2015) analyze the race for speed in financial markets. Their insights about the effects of high frequency trading are unrelated to ours since the two models feature very different trading frictions (speed vs. private information).

[6]While the impact of transparency on the welfare of large traders *relative* to small traders seems clear, the overall impact may be ambiguous. Certainly, identification of large traders that forces them to trade slower may reduce market liquidity overall.

where $\hat{p}_0$ is the price at time 0 and z_t is a Brownian motion (so σ is the volatility of the price). The selling rate q_t has an instantaneous price impact of I and the permanent price impact of Λ.[7] Once the transaction is completed, i.e. $X_0 = \int_0^T q_t\,dt$, the revenue equals $y = \int_0^T \hat{p}_t q_t\,dt$. The objective is to maximize $E[y] - \gamma\,\mathrm{Var}[y]$. This objective function can be justified as a quadratic approximation of a concave utility function: if the trader consumes his wealth w plus y at time T, receiving utility $u(w+y)$, then

$$E[u(w+y)] = u(w) + u'(w)E[y] + \frac{u''(w)}{2}\,\mathrm{Var}[y] + o(y^2).$$

An equivalent way to write the objective function is in terms of quadratic holding costs, as shown by the following lemma.

Lemma 1 *For deterministic strategies q_s, $s \in [0,T]$ that liquidate the amount $X_0 = \int_0^T q_s\,ds$,*

$$E[y] - \gamma\,\mathrm{Var}[y] = E[y] - \gamma\sigma^2 E\left[\int_0^T X_t^2\,dt\right], \quad \textit{where } y = \int_0^T \hat{p}_t q_t\,dt \tag{2}$$

is the seller's revenue.

Proof. Total revenue equals

$$y = E[y] + \int_0^T X_t\,\sigma dz_t,$$

where the integral measures the unexpected capital gains and losses on the remaining holdings (notice that if we had $\sigma = 0$, then the path of prices would be deterministic for any deterministic strategy, so $y = E[y]$). Hence,

$$\mathrm{Var}[y] = \sigma^2\,\mathrm{Var}\left[\int_0^T X_t\,dz_t\right] = \sigma^2\,E\left[\int_0^T X_t^2\,dt\right],$$

where the last step is just the classic Ito isometry. This implies (2). ■

In our model below, we model the player's preferences via quadratic holding costs. Thus, preferences in our model are similar to those of Almgren and Chriss (2001), and they approximate risk-averse utility. We microfound quadratic costs further in Appendix A.[8]

[7] Parameter Λ is analogous to Kyle (1985)'s lambda, as it measures the sensitivity of price to the total quantity sold.

[8] We should also note that the characterization of optimal trading in Almgren and Chriss (2001) holds in the class of deterministic strategies, but not in the class of stochastic strategies (which allow the traders to improve upon the objective $E[y] - \gamma\mathrm{Var}[y]$ by "burning money"). However, for preferences expressed in terms of quadratic holding costs, the deterministic strategy is optimal even when stochastic strategies are allowed.

Our Model: Preferences and Shocks. Our model is consistent with this framework - each of our N players has private information about their desire to buy or sell and chooses how to execute its trade.[9] Players have quasilinear utilities in money and quadratic holding costs, as in (2). Player i's desire to buy or sell is captured by a private taste shock $\xi_t^i \in \mathbb{R}$ - the holding cost is quadratic in the difference between player i's position $\hat{x}_t^i$ and the "bliss point" $\xi_t^i \in \mathbb{R}$ – i.e. the holding cost is

$$-\frac{b_i}{2}\left(\hat{x}_t^i - \xi_t^i\right)^2. \tag{3}$$

This formulation is the same as in Bank, Soner and Voß (2016) and Almgren and Li (2016), who capture the costs of an imperfect hedge, except that we also assume that players discount payoffs (3) at rate $r > 0$. There are many interpretations. The trader may be a hedge fund manager whose overall risk exposure can affect the optimal holdings of the asset being traded. Likewise, firms can trade to hedge their commodity price or currency risk exposure. Alternatively, as in Kyle, Obizhaeva and Wang (2014), shocks to ξ_t^i could be belief shocks: the trader's beliefs about the asset's "alpha" may change and that would affect his optimal holding of the asset. Finally, if traders are brokers executing trades on the behalf of their clients, then ξ_t^i can reflect the inventory held by the broker.

Parameter b^i reflects the holding cost of player i. We interpret $1/b^i$ as the "risk capacity" of trader i. Players with a lower coefficient b^i are "larger": they can hold larger positions away from their optimal point ξ_t^i at a lower cost. Conversely, players with a higher coefficient b^i are "smaller", and therefore more impatient to trade towards their optimal points.[10]

We call $X_t^i \equiv \hat{x}_t^i - \xi_t^i \in \mathbb{R}$ the inventory or allocation of player i. Inventories change due to taste shocks and trades. Taste shocks have mean 0. For concreteness we take the taste shocks to be Brownian, so that the vector of taste shocks follows

$$d\xi_t = -\Sigma\, dZ_t, \tag{4}$$

where Σ is an $N \times N$ matrix with full rank and Z is a vector of N independent Brownian motions. Denote the selling rate of player i by $-d\hat{x}_t^i/dt = q_t^i \in \mathbb{R}$. The vector of selling flows $q_t = [q_t^1, q_t^2, \ldots, q_t^N]^T$ must satisfy market clearing, i.e. its coefficients have to add up to 0. Due to taste shocks and trades, the vector of inventories $X_t = [X_t^1, X_t^2, \ldots, X_t^N]^T$ follows

$$dX_t = \Sigma\, dZ_t - q_t\, dt. \tag{5}$$

[9] There is a single asset, but our model can be extended to multiple assets with correlated fundamental risk or even segmented markets, following the ideas from Malamud and Rostek (2014).

[10] As we discuss below, our notions of large and small traders represent the fraction of the asset they hold under first-best allocation. In practice, player size can also be measured in terms of the fraction of total volume that the player trades, a measure related to the sizes of inventory shocks in our model. When we say "size" in the paper, we mean the former notion, i.e. risk capacity, as it plays a crucial role in equations that describe equilibria. Shock sizes do not play as big of a role, as separating equilibria do not depend on the distribution of shocks.

If units change hands at price p_t, then the payoff of player i is defined as:

$$E\left[r\int_0^\infty e^{-rt}\left(-\frac{b^i}{2}(X_t^i)^2+p_tq_t^i\right)\,dt\right]. \tag{6}$$

This completes the description of preferences and shocks in our model. Given any history of shocks and trades, equations (5) and (6) give the players' utilities. With these preferences, we can consider various trading mechanisms, and we motivate the particular mechanism we study in this paper in Section 2.1.

It is useful to interpret p_t as the deviation of market price from fundamental value dictated by the current microstructure frictions. With this interpretation, negative values of p_t reflect an overall selling pressure, and positive, a buying pressure. The fundamental value itself may change with public news about future cash flows, and quadratic holding costs in (6) reflect the players' risk aversion to this fundamental risk.[11] With this interpretation in mind, we refer to p_t simply as "price," even though it represents a price discount or premium relative to a benchmark.

Furthermore, while we model inventory shocks to be Brownian and stationary for concreteness, our equilibrium characterization applies to much wider shock structures. The reason is full separation of types in equilibrium: players signal their allocations through their trading rates q_t^i. Hence, trading dynamics remain the same even with Poisson or non-stationary shocks, as long as the support assumptions required for a separating equilibrium hold. Since utilities are quadratic in inventories, welfare depends on the variance of shocks but not other details of their distribution. We use continuous time to simplify some of the algebra, but most of our analysis and all our economic intuitions apply to discrete-time versions of our model as well.

Preferences and shocks in this model are quite similar to the models in Vayanos (1999), Du and Zhu (2013), and Kyle, Obizhaeva and Wang (2014). The main distinction is that we allow for asymmetry, i.e. we are mainly interested in the case when the risk coefficients b^i are not identical. As we show, the asymmetry leads to new price and trade dynamics in equilibrium.

First Best. The efficient allocation of assets is proportional to risk capacities. First best requires that any vector of inventories X_t should be immediately reallocated so that each player i holds the fraction of total supply that is proportional to his/her risk capacity, i.e.

$$\tilde{X}_t^i=\frac{1/b_i}{\bar{\beta}}\bar{X}_t,\quad \text{where } \bar{X}_t=\sum_{i=1}^N X_t^i \text{ and } \bar{\beta}\equiv\sum_{i=1}^N 1/b_i \text{ (market risk capacity)} \tag{7}$$

[11] In Appendix A we make this interpretation precise using a model with exponential utility. Of course, players care about price changes due to both fundamental shocks and microstructure risk. From the exponential utility model, we get the same set of equilibrium equations in the limit when fundamental risk overwhelms microstructure risk, and slightly more complicated but similar equations when endogenous microstructure risk that depends on trades is significant. Coefficients b^i in our model correspond to the traders' risk aversion in the exponential model, and holding costs here correspond to the costs of exposure to fundamental/dividend risk in the exponential model.

If inventories were publicly observable and price were set to the marginal disutility of holding a marginal unit of the asset at the efficient allocation (which is the same across all agents), then the traders would be able to trade to the efficient allocation immediately. Price at the efficient allocation is given by

$$\tilde{p}_t = \frac{d}{dy} E\left[\int_t^\infty e^{-r(s-t)} \frac{-b_i}{2} (\tilde{X}_s^i + y)^2 \, ds\right] = -\frac{b_i \tilde{X}_t^i}{r} = -\frac{\bar{X}_t}{r\bar{\beta}}. \tag{8}$$

Common Values. Note that in our model preference shocks can be correlated across players (the off-diagonal elements of Σ do not have to be zero). Since all players know their private shocks, this is a model of correlated private values. It can be generalized to allow for common values as follows. Suppose each player observes a signal ξ_t^i and his preferences depend on a linear combination of the signals of others. Let F be an $N \times N$ matrix with rows $\underline{F}_i$ and 1's on the diagonal. Let the agent's utility flow (including net revenue from trade) be

$$-\frac{b_i}{2}\left(\hat{x}_i^t - \underline{F}_i \xi_t\right)^2 + q_t^i p_t, \tag{9}$$

If $\underline{F}_i$ has non-zero entries in positions other than i, trader i cares directly about the signals of others because they inform his preferences, for example because other players may have information about the fundamental value of the asset.[12]

A simple example of an F matrix is

$$F = \begin{bmatrix} \underline{F}_1 \\ \underline{F}_2 \\ \vdots \\ \underline{F}_N \end{bmatrix} = \begin{bmatrix} 1 & \phi & \cdots & \phi \\ \phi & 1 & & \vdots \\ \vdots & & \ddots & \phi \\ \phi & & \phi & 1 \end{bmatrix} \tag{10}$$

When $\phi = 0$, it simplifies to our pure-private-values model. If $\phi = 1$ it implies a pure common value model (all players have the same taste equal to the sum of all signals). We discuss at the end of the paper how our analysis could be extended to these richer preferences (at least for ϕ small enough so that there is trade in equilibrium), but for simplicity in this paper we analyze the (correlated) private values case.

If F were asymmetric, traders would care about who they trade with to infer the shocks of particular players: analogously to the Kyle (1985) model, where traders would like to know if their counterparty is an informed or a noise trader. Perhaps more surprisingly, we show that with strategic heterogeneous traders, even in a pure private values setup agents want to know who they trade with.

[12] As in our private values model, the trader also cares about the signals of others *indirectly* because they are informative about future prices.

2.1 Mechanisms for trading.

We now turn to the determination of prices and trade flows. Existing papers with symmetric rational traders model trading through a double uniform-price auction, as in Vayanos (1999) and Du and Zhu (2013), following the tradition of Kyle (1989). However, in our setting, since players are not symmetric, such a mechanism leads to a complicated fixed point problem which involves filtering. For reasons that will become clear later, players would want to know not only the price but also information about who else is buying and selling. They would be making inferences about the distribution of supply and demand, across players of different sizes, from the dynamic properties of prices and through other means.

We instead propose an alternative trading mechanism in which players observe the flows currently being traded, with the goal of gaining the most insights about trading dynamics in asymmetric markets in as simple a framework as possible. The benefit of this framework is that we obtain a separating equilibrium, in which dynamics depend only on the vector of risk coefficients b^i, and not the nature, correlation or variance of the shocks. While the dynamics are simple, they are quite rich and capture many of the phenomena that we set out to study.

The mechanism we propose is a uniform-price *conditional* double auction. We assume that players observe the flows of all other players, or can condition their demand functions on these flows. While we make this assumption out of necessity, there is evidence that market participants in practice spend a considerable amount of effort identifying the sources of trades. For example, brokers call each other to find out who traded, and some market-makers pay discount brokerages for the flow specifically from retail investors. Moreover, recently NYSE began allowing orders from retail investors to be marked as such through the Retail Liquidity Program (RLP). Finally, recent popular books like Lewis "Flash Boys" or Patterson's "Dark Pools" describe strategies used by high-frequency traders to determine likely sources of trades. Therefore, we think that the conditional auction model not only helps us with tractability but also helps us capture important real-life phenomena. Another interpretation of our model is that it allows us to understand why traders care about the counterparty even if they do not believe that the counterparty has private information about fundamental value.

Formally, our conditional double auction format is defined as:

Conditional Double-Auction. *At each moment of time t, each player i announces a supply-demand function*

$$p = \bar{\pi}^i - \sum_{j \neq i} \pi^{ij} q^j$$

that gives the price at which the player is willing to trade, as a function of the selling rates of all other players (with player i buying the net residual supply). The market

maker then determines the price p and the selling rates q^j from the system of equations

$$\sum_{i=1}^{N} q^i = 0, \qquad \forall\, i, \quad p = \bar{\pi}^i - \sum_{j \neq i} \pi^{ij} q^j. \tag{11}$$

A profile of strategies is *stationary* if the slopes of the demand functions π^{ij} remain constant over time, while the intercepts $\bar{\pi}^i$ may depend on the players' inventories. Furthermore, a stationary profile of strategies is *linear* if $\bar{\pi}^i = \hat{\pi}^i X^i$ for an appropriate constant $\hat{\pi}^i$, where X^i is player i's inventory. Obviously, a linear stationary profile such that $\hat{\pi}^i \neq 0$ for all i is revealing (i.e. fully separating). We are interested in characterizing equilibria in revealing linear stationary strategies.[13]

We would like to comment on the determination the price-flow vector pair (p, q) from linear stationary strategy profiles in a conditional double auction. There is, unfortunately, a (non-generic) set of slopes $\{\pi^{ij},\ i \neq j\}$ such that there is no solution (p, q) to (11) (or multiple solutions exist) for some intercepts $\{\bar{\pi}^i\}$. This leads to indeterminacy. The solution must be unique at least for the intercepts $\bar{\pi}^i = 0$, which correspond to $X = 0$. As the following proposition shows, if the solution is unique for $\bar{\pi}^i = 0$, then it is unique for all intercepts $\{\bar{\pi}^i\}$ (this property amounts to certain matrix being nonsingular). We call profiles with this property, and also the property that $\hat{\pi}^i \neq 0$ for all i (so that each player's allocation has effect on trade), *acceptable*.

Proposition 1 *Given a set of stationary slopes $\{\pi^{ij},\ j \neq i\}$, the following two statements are equivalent*

1. *equations (11) have a unique solution (p, q) for $\bar{\pi}^i = 0$,*
2. *equations (11) have a solution for all intercepts $\{\bar{\pi}^i\}$,*

and imply that equations (11) have a unique solution for all intercepts $\{\bar{\pi}^i\}$.

Proof. See Appendix. ■

While the conditional double auctions provide an intuitive way to model price formation in the market, it is easier to analyze price formation and trade dynamics using a direct revelation mechanism that is strategically equivalent to the auction, as we show below.

Direct Revelation Mechanism. *A stationary linear mechanism is a pair (P, Q) that consists of an N-dimensional vector P and an $N \times N$ matrix Q, whose*

[13] While we focus on fully separating stationary equilibria, we have investigated and are aware of other possibilities. There are also non-stationary equilibria, under a broader set of strategies, with periods of no trade, periods of continuous trade, and time points where a strictly positive number of shares are traded. The players' bids may not always reveal information about their inventories, and some players may be excluded from trade. While a full characterization of equilibria is interesting, it is beyond the scope of this paper.

columns add up to 0. In this mechanism, at each moment of time t the market maker asks every trader to announce his inventory X_t^i. The vector of announcements determines the price $p_t = PX_t$ and the vector of trading rates $q_t = QX_t$. We call the mechanism truthtelling if telling the truth is an equilibrium of the mechanism when announcements are observable.[14]

Let us call a stationary linear mechanism (P, Q) *acceptable* if for any X in the null space of Q (so that $QX = 0$), price $PX = 0$ only if $X = 0$. For an acceptable mechanism, there is a one-dimensional space of allocations X that result in no trade (i.e. it is the null space of Q) and all of these allocations result in different prices.

Proposition 2 *The following statements about a mechanism (P, Q) are equivalent:*

- *it is acceptable;*
- *the matrix Q^P obtained by replacing the first (or any other) row of Q with the vector P is invertible; and*
- *if everybody tells the truth, the allocation X can be inferred by the outside observer from the price p and the vector of flows q.*

Proof. See Appendix. ■

The last property implies that for acceptable mechanisms, the requirement that announcements are observable in the definition of a truthtelling mechanism can be replaced with the requirement that the price and all trading flows are observable. That is, players can fully infer the inventories of others from the price and the trading flows.

The following theorem provides a result about equivalence between acceptable mechanisms and (linear stationary) strategy profiles in a conditional double auction.[15]

Theorem 1 *There is a one-to-one map between*

- *acceptable profiles of linear stationary strategies and*
- *acceptable mechanisms, such that for any nonzero allocation X that results in no trade, every element of X is nonzero*

[14] That is, a truthtelling mechanism has to be ex-post incentive compatible.

[15] The one-to-one map implies that mechanisms that are not acceptable do not generate dynamics that correspond to any profile of a conditional double auction. Conversely, for a profile in a conditional double auction that is not acceptable, the map $X \to (p, q)$ is not well-defined.

that lead to the same map $X \to (p, q)$.

Moreover, consider a profile $\{(\hat{\pi}^i, \pi^{ij}),\ j \neq i\}$ *and a corresponding mechanism* (P, Q). *Then for each player* i, *for any allocation* X^{-i} *of other players, player* i *can attain the same one-dimensional sets of price-trading flow pairs* (p, q) *by making a report in a mechanism or by submitting a supply-demand function in a strategy profile of a conditional double auction.*

Proof. See Appendix. ■

An immediate corollary of this theorem is that, since each player has the same degree of control over prices and flows in corresponding mechanism and profile of a double auction, a profile of a double auction is an equilibrium if and only if the corresponding mechanism is truth telling.

Corollary 1 *If an acceptable profile of a conditional double auction is an equilibrium, then the corresponding direct revelation mechanism is truthtelling, and vice versa.*

In all the equilibria that we construct below for every vector of reports of others, trader i can find a report that implies he does not trade (because his trade is linearly increasing in his report). For such equilibria ex-post incentive compatibility implies that (ex-post) individual rationality holds.

From now on we will focus on truth telling direct revelation mechanisms, as the representation in terms of P and Q provides a convenient direct map from the players' allocations to prices and flows.

3 Equilibrium Characterization.

We now derive equations that characterize trading dynamics under stationary linear equilibria in our model. We cover the case of N large players first. At the end we describe what happens when player N represents a continuum of players with total risk capacity of $1/b^N$, representing a competitive fringe.

Under mechanism (P, Q), on the equilibrium path the vector of trading flows is given by $q_t = QX_t$ and the price is $p_t = PX_t$. If player i deviates and reports inventory $y + X_t^i$ instead of X_t^i, then the resulting price is $PX_t + p^i y$ and the vector of trading flows is $QX_t + Q^i y$, where Q^i is the ith column of Q, and p^i is the ith element of P. Hence, from (5), the inventory vector follows

$$dX_t = \Sigma\, dZ_t - QX_t\, dt - Q^i y\, dt. \tag{12}$$

Under this law of motion, the value function $f^i(X)$ of player i must satisfy the HJB equation

$$r f^i(X) = \max_y\ \frac{-b^i}{2}(X^i)^2 + (PX + p^i y)(\underline{Q}^i X + q^{ii} y) - \tag{13}$$

$$f_x^i(X)(QX + Q^i y) + \frac{1}{2}\text{tr}\,[\Sigma\Sigma^T f_{xx}^i(X)],$$

where $\underline{Q^i}$ is the i-th row of Q, q^{ii} is the i-th diagonal entry of Q, f_x^i is the gradient of f^i and f_{xx}^i is the Hessian. In a truth-telling mechanism, $y = 0$ must solve the maximization problem in (13) for all inventory vectors X.

Since (13) is a quadratic optimization problem, the value function must be of quadratic form $f^i(X) = X^T A^i X + k^i$, where A^i is a symmetric $N \times N$ matrix and k^i is a constant. Then the HJB equation (13) becomes

$$r(X^T A^i X + k^i) = \max_y \frac{-b^i}{2}(X^i)^2 + (PX + p^i y)(\underline{Q^i} X + q^{ii} y) \tag{14}$$
$$-2X^T A^i (QX + Q^i y) + \text{tr}\,[\Sigma\Sigma^T A^i].$$

Taking the first-order condition at $y = 0$, the HJB equation reduces to the following system of equations

$$p^i \underline{Q^i} + q^{ii} P = 2(A^i Q^i)^T, \tag{15}$$

$$rA^i + A^i Q + Q^T A^i = \frac{P^T \underline{Q^i} + (\underline{Q^i})^T P}{2} - \frac{b^i}{2} 1^{ii} \quad \text{and} \quad k^i = \frac{\text{tr}\,[\Sigma\Sigma^T A^i]}{r}, \tag{16}$$

where 1^{ii} denotes the square N-by-N matrix that has 1 in the i-th diagonal position and zeros everywhere else.

We call matrix Q stable if it has no eigenvalues with positive real parts. Stability implies that the transversality condition $E[e^{-rt} X_t^2] \to 0$ holds on the equilibrium path. The following proposition formally registers the fact that appropriate solutions of equations (15) and (16) indeed lead to equilibria.

Proposition 3 *Consider any solution $(P, Q, k^i, A^i, i = 1, \ldots N)$ of the system (15) and (16) such that $p^i < 0$ and $q^{ii} \geq 0$ for all $i = 1, \ldots N$, and the matrix Q is stable. Then, for all i if all other players tell the truth in mechanism (P, Q), it is better to follow the truthtelling strategy than any other strategy that satisfies the no-Ponzi condition $E[e^{-rt} X_t^2] \to 0$.[16] That is, (P, Q) is a stationary linear equilibrium.*

Proof. See Appendix. ■

The trading game has degenerate stationary equilibria, in which some or all of the traders are excluded from the market (i.e. the matrix Q consists of zeros in several rows and columns). Other degenerate equilibria involve a splitting of the market in which subsets of players trade among themselves, without trade across the subsets (so that Q has zeros in some positions). We are interested primarily in non-degenerate equilibria (in which Q has no zeros, so all players trade with each other), and would like to understand their properties such as the speed of trade, price momentum, and inefficiencies.

[16] If player i is allowed to violate the no-Ponzi condition, he can get infinite utility.

Equilibrium with a Competitive Fringe. We define a competitive fringe as a continuum of traders with a given finite risk capacity $1/b^F$. A group of m identical traders has risk capacity $1/b^F$ if each trader has utility function of the form (6) with risk coefficient mb^F. Taking $m \to \infty$, we obtain a competitive fringe. We can include a competitive fringe into our model and, if so, we designate trader N as the fringe.[17]

The HJB equation for the total utility $f^N(X)$ of the fringe is given by the same equation as the equation (16) for large traders.[18] However, the first-order condition differs from (15), since fringe members can no longer affect the price with their individual actions.

Proposition 4 *If player N represents a competitive fringe, then prices and flows must satisfy the first-order condition*

$$rP + PQ + b^F 1^N = 0, \tag{17}$$

where 1^N denotes a row vector with 1 in the N-th position and zeros everywhere else.

Proof. See Appendix. ■

If we right-multiply (17) by X_t it becomes $rp_t = -b^F X_t^F + E[\frac{dp_t}{dt}|X_t]$. This optimality condition has a simple economic interpretation: since any member of the fringe has no price impact, he has to be indifferent between holding a marginal unit of inventory and selling it to buy back a moment later. Selling allows him to collect interest flow on the cash; holding implies extra inventory cost and expected capital gains (if price change in expectation).

Unfortunately, the system of (15) and (16) (together with (17) if player N represents the fringe) in general cannot be solved in closed form. We can solve the equations numerically and provide several computed examples in Sections 4-6. We also present one useful (and surprisingly accurate) approximation of the solution.

We should note also that we do not necessarily view equations (15) and (16) as the end goal, as there are many meaningful modifications of this set of equations. We can replace a single first-order condition (15) to model the behavior of a competitive fringe. Appendix A presents a model of risk-averse agents that care about both fundamental and microstructure risk, and allows agents to have private information about fundamentals. In addition, if someone believes that our model may be too

[17] When analyzing a model with a fringe we restrict attention to equilibria that are symmetric with respect to the fringe members, so that the fringe can be treated as one (as we mentioned above, our game has also degenerate equilibria in which some players are excluded altogether while others trade only within subgroups).

[18] For clarity, we maintain the assumption that all fringe members get identical shocks, but total utility of the fringe is given by $f^N(X)$ even if individual fringe members get idiosyncratic shocks of bounded volatility driven by *finitely* many Brownian motions. Any misallocation among fringe members gets traded to efficiency infinitely fast.

sophisticated to capture the behavior of various market participants, it is possible to accommodate various forms of bounded rationality. For example, it is possible to study what happens when some or all of the players ignore price momentum that occurs as a result of asset allocation, or make incorrect assumptions about their price impact.

We can solve two special cases explicitly. First is the case when one large trader faces a competitive fringe. This case is particularly useful because it illustrates most of the important properties of trading in asymmetric markets, including the existence of price momentum, the form and heterogeneity of price impact across players, and welfare implications of market power. The next section describes trading in a market with a single large trader and competitive fringe. The second case we can solve explicitly is when all players have identical risk coefficients - we present this solution in Proposition 12.

4 Trading between a Large Player and a Fringe.

The following proposition characterizes in closed form equilibrium trading between one large player with risk capacity $1/b^L$ and a competitive fringe with risk capacity $1/b^F$.

Proposition 5 *Consider a market with $N = 2$, in which player 1 is an individual large player and player 2 is a competitive fringe. Then in the unique nondegenerate linear stationary equilibrium, equilibrium prices and the players' trading rates are characterized by vectors*

$$P = -\frac{1}{r}\frac{b^F}{3b^F + b^L}\left[b^L,\ b^L + 2b^F\right], \quad Q = \frac{r}{2}\begin{bmatrix} b^L/b^F & -1 \\ -b^L/b^F & 1 \end{bmatrix}. \tag{18}$$

The welfare of the large trader and the fringe are characterized by matrices

$$A^L = \frac{b^F}{2r(3b^F + b^L)}\begin{bmatrix} -3b^L & -b^L \\ -b^L & b^F \end{bmatrix} \quad \textit{and}$$

$$A^F = \frac{b^F}{2r\left(3b^F + b^L\right)\left(2b^F + b^L\right)}\begin{bmatrix} (b^L)^2 & -b^L b^F \\ -b^L b^F & -\left((b^L)^2 + 5b^L b^F + 5(b^F)^2\right) \end{bmatrix}$$

Proof. See Appendix. ■

From this closed-form solution, we can analyze several salient properties of equilibria. We can divide properties into two groups: on and off equilibrium. On equilibrium properties refer to what is seen by an independent observer - properties such as the

speed of trade and price momentum. Off-equilibrium we can analyze what happens when a single player deviates and changes the trading rate, i.e. we can analyze price impact.

Speed of Trade and Price Momentum. Let us discuss the properties of P and Q. Matrix Q has two eigenvalues, 0 and

$$\kappa = \frac{r}{2}\frac{b^L + b^F}{b^F} \tag{19}$$

with the corresponding eigenvectors being the efficient allocation and $[1, -1]$. That is, there is no trade if players are already at the efficient allocation, and *any misallocation gets traded to efficiency at the rate* (19). To understand why this is, we have to look at player incentives by considering what would happen off equilibrium - we do this in a bit. For now, let us observe that the speed of trade depends on how large trader 1 is relative to the rest of the market. When trader 1 is small, as the risk capacity $1/b^L \to 0$, so the market gets closer to being fully competitive, trading speed (19) converges to infinity. It is interesting, however, that the trading speed here is always bounded from below by $r/2$, a limit approached as the large player gets large, i.e. $b^L \to 0$.

The relationship between the speed of trade and market competitiveness is similar to the result of Vayanos (1999) that in symmetric markets the speed of trade increases with the number of players N, and is proportional to $r(N-2)/2$. Here, when the large player is $1/N$ of market size, e.g. with $1/b^L = 1$ and $1/b^F = N - 1$, then (19) becomes equal to $rN/2$. That is, the speed of trade is on the order of the inverse of player size, and it is slightly greater when the player faces a competitive market than players of equal market power. The property that the speed of trade increases as the market power of individual players falls extends to asymmetric markets that we consider in the next section, with the new result that there may be different trading speeds in different market segments (i.e. matrix Q has several positive eigenvalues).

The price is first best whenever players are at an efficient allocation X, i.e. in that case $PX = \bar{P}X$, where

$$\bar{P} = -\frac{1}{r}\frac{b^L b^F}{b^F + b^L}[1,\ 1]$$

is the first-best pricing vector. However, since $P \neq \bar{P}$, price differs from first best whenever players are not at an efficient allocation. The immediate implication of this is *price momentum*: in the absence of further inventory shocks, price converges to the first-best price at rate given by (19) as players trade to the efficient allocation. With shocks, this is the expected price path.

The direction of price momentum is connected with the property that the price is skewed away from first best in favor of the large players. For example, at allocation $X_0 = [1, -1]$ the price is

$$p_0 = \frac{1}{r}\ \frac{2(b^F)^2}{3b^F + b^L} > 0.$$

That is, the large player starts selling at a positive price, while the first-best price is 0.[19] The sales of the large player produce a downward price momentum, in this example directed from p_0 to 0. Conversely, when the large player is buying, the price has an upward momentum. In contrast, in the symmetric-trader environment of Vayanos (1999), there is no price momentum: the price immediately adjusts to first best following any shock. In general, as we explain in the next section, price momentum is driven by the unequal market power of current buyers and sellers. The side with greater market power gets a temporary price advantage.

Price Impact. With our trading model, we can study how a given sequence of trades affects the price. Effectively, we provide a game-theoretic way of deriving parameters of a model such as that of Almgren and Chriss (2001). It is also a result, rather than an assumption, that price impact of the large player takes the form given by (1) when he is trading against a fringe.[20] In general, however, our model implies price impact of a form that is more general than (1). Trades also have a *transient* price impact, which has an empirical counterpart (see discussion in the next section).

We should emphasize that the study of price impact is essentially an off-equilibrium exercise. On equilibrium, a player has a uniquely optimal trading strategy. In principle, however, a player can choose to trade in any way he/she desires. Given an arbitrary set of trades, the market responds by forming incorrect beliefs about the player's inventory. The price that the trader receives is a function of market beliefs, which are formed based on the trading speed. To solve for price impact following a given path of sales by the large player q^L_s, $s \in [0,t]$, we have to calculate market beliefs following these sales, i.e. infer the shocks of the large player that would have resulted in this path of trades. The following proposition provides the form of the large player's price impact in the model with a single large player and the fringe.

Proposition 6 *The price impact of the large trader takes the form (1), where*

$$\Lambda = \frac{b^F}{r}, \qquad I = \frac{\Lambda}{r+\kappa}, \qquad p_0 = -\Lambda X^F_0, \quad \textit{and} \quad \sigma = \Pi\,|\underline{\Sigma}^F|, \tag{20}$$

where κ given by (19) is the speed of trade and $\underline{\Sigma}^F$ represents the last row of Σ. Given a sequence of sales q^L_s, $s \in [0,t]$, by the large trader, the fringe forms the belief about

[19] The deviation of the price from first best depends on the market power of the large trader. Suppose $b^F = 1$, so that the marginal value of a unit to the fringe at allocation $X = [1,-1]$ is $1/r$. The actual price is less than this: it can get as high as $2/(3r)$ when $b^L \to 0$ (i.e. the large player has tremendous market power) but it approaches first best when the market power of the large player diminishes (i.e. as $b^L \to \infty$).

[20] In the broader interpretation that p_t is the deviation (due to microstructure forces) of the actual price $\hat{p}_t$ from the fundamental value, the volatility of the actual price $\hat{p}_t$ would be greater than that given by Proposition 10 due to fundamental shocks. Thus, representation (20) captures only the microstructure shocks.

the large trader's inventory to be

$$X_t^L = \frac{b^F}{b^L}\left(\frac{2}{r}q_t^L + X_t^F\right). \tag{21}$$

Proof. The volatility of inventory shocks to the fringe is $|\underline{\Sigma}^F|$, so we can express

$$X_t^F = X_0^F + |\underline{\Sigma}^F|\, z_t + \int_0^t q_s^L\, ds, \tag{22}$$

where z is a Brownian motion. Now, given this allocation of the fringe, the relationship between the allocation of the large player and the selling rate, from the first row of Q, is

$$q_t^L = \frac{r}{2}\left(\frac{b^L}{b^F}X_t^L - X_t^F\right).$$

Then from the selling rate q_t^L, the fringe infers that X_t^L is given by (21).

Now, the price is given by PX, where (X_t^L, X_t^F) is given by (21) and (22). Therefore,

$$p_t = -\frac{1}{r}\frac{b^F}{3b^F + b^L}\left(b^F\left(\frac{2}{r}q_t^L + X_t^F\right) + (b^L + 2b^F)X_t^F\right) =$$

$$-\frac{b^F}{r}\underbrace{\left(X_0^F + |\underline{\Sigma}^F|\, z_t + \int_0^t q_s^L\, ds\right)}_{X_t^F} - \frac{1}{r}\frac{2b^F}{r}\frac{1}{3 + b^L/b^F}q_t^L.$$

This implies (20). ■

Proposition 6 confirms the intuitive guess that the permanent price impact Λ of large player's trades (i.e. Kyle's lambda) is proportional to the risk capacity of the fringe. On the other hand, the instantaneous price impact I depends also on the risk capacity of the large player relative to the fringe. A player of larger risk capacity trades more slowly to minimize price impact, hence any selling rate q_t^L signals a larger inventory X_t^L of the large player, according to (21). Hence, as the speed of trade κ slows down, the price impact rises.

Large risk capacity or "patience" gives the large player market power to obtain better prices by selling slowly, but the expectations of the fringe of this behavior raise the instantaneous price impact of the large player. This creates illiquidity in the market, which is detrimental to the large player. Thus, "market power" bites back by reducing liquidity. There is ample anecdotal evidence that large traders try to hide information about their desire to buy or sell, e.g. by splitting orders into small portions or waiting for the flow to trade against. In practice, therefore, to some extent large traders are able to hide their flow behind that of small retail investors, although other market participants, especially the high-frequency traders, seek to identify large traders - see Patterson (2012).

Our model of trade presents an extreme benchmark in which the entire market knows immediately when a large player is trading, and the price reacts accordingly. In this case, is market power beneficial to have? With transparency and rational expectations, the cost of market power is illiquidity - the large player is forced to trade slowly, even though he can take advantage of the current price in the short run. We can ask several questions to explore market power. Would several small players benefit by merging into a single large unit to coordinate trades? If the large player could commit to a rate of trade in order to affect the expectations of the fringe and the price impact, would the optimal rate be faster or slower than the equilibrium one? We address these and other questions next.

4.1 Market Power and Welfare.

Our general conclusion, which is somewhat counterintuitive, is that in transparent markets market power is detrimental. That is the cost of illiquidity that goes with market power outweighs the ability to control the price by trading slowly. We illustrate this message via a sequence of three results, which we informally call "fractions" as they hold when the large player has less than 1/2, 1/3 and 1/4 of market risk capacity, respectively.

First, we address the question of mergers. Suppose that in a fully competitive market a portion of the fringe merges to form a single large player in order to coordinate trades - would their welfare improve? To keep the experiment clean, we assume that the fringe members who merge have identical shocks. Before the merger, these fringe members trade in the same direction without taking into account price impact, i.e. the effect of their trades on the price that everybody else is facing.[21] As a result, the price adjusts immediately to first best, and fringe members trade to efficiency instantaneously at that price. After the merger, the price adjusts gradually to first best, i.e. the newly formed large player trades at prices better than the first-best price, but slowly.

Proposition 7 *In a fully competitive market, suppose that a mass of agents facing identical inventory shocks, and with risk capacity less than 1/2 of the market, merge to form a single large player. Then for any distribution of shocks, the total welfare of the merging players falls relative to the level before merger.*

Proof. See Appendix. ■

Next, we observe that the price impact of the large player is connected with the equilibrium trading rate κ. What would happen if the large player could commit to any trading rate and the price impact is determined from the first-order condition

[21] The phenomenon in which agents to not take into account the effect that their trades have on the price that other agents are facing is sometimes referred to as the "firesale externality."

(17) given that rate? What is the optimal selling rate, and is it slower or faster than the equilibrium trading rate given by (19)?

Proposition 8 *Consider a game in which the large player (through commitment) determines the trading matrix*

$$Q = \frac{\tilde{\kappa}}{\kappa}\frac{r}{2}\begin{bmatrix} b^L/b^F & -1 \\ -b^L/b^F & 1 \end{bmatrix}$$

from a class of matrices that lead to convergence to the efficient allocation at rate $\tilde{\kappa}$. Then the trading rate $\tilde{\kappa}$ that maximizes the utility of the large trader is infinity when the large player's risk capacity is 1/3 of the market or less.

Proof. See Appendix. ■

Our third result is connected with the phenomenon of (technical) *front-running*. If new players enter the market, they take away market power from the large player. Specifically, any entrant would want to make profit from the momentum generated by the sales of the large player, by selling short early when the price is high and buying back later when the price is low. This is what we define as front-running.

For example, given an initial allocation between the large player and the fringe of $[1, -1]$, imagine that a new portion of the fringe with a neutral allocation of 0 enters (so that the efficient allocation is for all players to hold zero inventory). How does trade proceed? The new fringe members will not stay at their bliss points. Rather the fringe (both groups) trades at time 0 to redistribute its inventory uniformly, and then fringe members buy from the large trader at proportionate rates. Effectively, the fringe members who start at their bliss point of zero inventory front-run the large trader who wants to sell. These players will sell assets to other fringe members ahead of the large trader, while the price is high, and then buy back later from the large trader at a lower price.

Entry increases the equilibrium speed of trade, κ (recall it rises with the risk capacity of the fringe). At the same time, the price P becomes less skewed away from first best. Does the benefit of faster trade offset the large player's loss of market power? This question is tough to answer, because entry of new fringe members adds to the risk capacity of the market overall, and thus improves opportunities for risk sharing. However, we propose a way of disentangling the risk-sharing and liquidity effects.

Specifically, assume that new fringe entrants arrive with shocks that are proportionate to the average shock of the existing market, so that first-best allocation does not change. Then the entrants cannot help to share the risk, as their risk capacity is already absorbed by the shocks they receive. Thus, we capture purely the effect of rising trading speed on the welfare of the large player. The following proposition shows that the welfare of the large player increases with entry as long as the risk capacity share of the large player is less than 1/4 (which is still an unrealistically high number).

Proposition 9 *Suppose that the risk capacity share of the large player is less than* $1/4$. *If new fringe members enter without adding risk capacity (i.e. with shocks such that first best allocation does not change) then the welfare of the large player goes up.*

Proof. See Appendix. ■

We finish this section with a general observation. With transparency about the source of flow, the market anticipates large players to split their trades into small portions. As a result, the price impact of even small orders becomes large, which harms the welfare of large players, even though by exercising their market power they are able to extract a better price. As we show in Section 5, with a heterogeneous market of N strategic players, larger traders trade more slowly and have a greater price impact. It is not surprising then that in practice, large traders want to hide their flow behind that of small traders, in order to take advantage of the private information about their supply or demand for as long as possible.

5 Equilibria with N Large Traders.

In this section we describe the properties of equilibria in a market with N large traders. We follow the same classification of results as in Section 4, by distinguishing on and off-equilibrium phenomena.

On equilibrium, the speed of trade depends on the overall competitiveness of the market and the market power of individual players. Small players trade faster, and misallocations among small players get traded to efficiency faster than those among large players, as captured by the eigenvector decomposition of Q. Equilibrium price deviates from first best, depending on the allocation among players and their size/market power. While first-best price depends only on total inventory, i.e. it is equally sensitive to allocations of all players, equilibrium price is more sensitive to the inventories of small players than those of large players. That is, as the market converges towards the efficient allocation, market price is skewed in favor of the large players. In equilibrium price drifts to first best. This price momentum depends on the relative market power of buyers and sellers. For example, if there is a small number of large buyers and many small sellers at a given moment of time, the price drifts up.

Off equilibrium, trades of an individual player have *permanent* price impact (i.e. Kyle's λ), which depends on the risk capacity of the rest of the market, and *instantaneous* impact, as in the model with a single large player and a fringe. The instantaneous impact, the sensitivity of price to flow, is different for players of different size. Instantaneous impact is greater for larger players, as they trade more slowly and retain a greater portion of their inventory for longer. In addition, trades have *transient* impact: i.e. price reaction to total sales that decays over time. Transient impact exists because players do not absorb flow proportionately to risk capacity and

so, when sales by a single player stop, trade among other players continues, affecting the price.

To sum up, while most properties of equilibria are apparent already in the closed-form solution with a single large player and a fringe, we observe a more complete picture and several new phenomena in the general model. In general, the speed of trade varies by the competitiveness of a market segment. Also the general model illustrates in greater depth how price momentum depends on the relative competitiveness of buyers and sellers. Moreover, instantaneous price impact depends on player identity, providing a rationale for why the knowledge of the origin of trade flow matters a great deal to sophisticated market participants in practice. Finally, the general model gives rise to the transient price impact: while this effect is absent from Almgren and Chriss (2001), it exists empirically.

Unfortunately, in general our model does not have exact analytical solutions. Hence, we illustrate these results, first, using computed examples and, second, analytically using an approximation of true solutions that is valid near the symmetric case. That approximate solution provides explanation and analytical basis for the patterns we observe in the numerical examples.

5.1 Equilibrium Trade.

We start by discussing trade on the equilibrium path, from the perspective of an outside observer. Players trade towards the efficient allocation with speeds that depend on their size. Price momentum exists, and depends on the relative risk capacity of buyers and sellers.

Example. Let us start with a simple example which captures broadly the properties we observed from exploring numerically solutions to many examples. We discuss economic intuition here, and at the end of this section we provide analytical results that affirm these properties.

Consider a game with five traders, whose holding cost coefficients are

$$[b^1, b^2, b^3, b^4, b^5] = [1, 1.5, 2, 2.5, 3],$$

and the discount rate is normalized to $r = 1$. Solving the equilibrium conditions numerically, we get that any allocation X is priced by vector

$$P = [-.254, -.329, -.387, -.435, -.476],$$

and the rates of trading flows are given by matrix

$$Q = \begin{bmatrix} 0.630 & -0.244 & -0.319 & -0.389 & -0.455 \\ -0.163 & 0.965 & -0.326 & -0.401 & -0.473 \\ -0.160 & -0.244 & 1.289 & -0.405 & -0.481 \\ -0.156 & -0.241 & -0.324 & 1.598 & -0.483 \\ -0.152 & -0.236 & -0.320 & -0.403 & 1.892 \end{bmatrix}.$$

When players have different costs of holding inventory, the equilibrium pricing vector P does not assign the same weights to the inventories of different players (even though the first-best pricing vector still assigns the same weight to all players). The reason is that players with large risk capacity $1/b^i$ exercise market power by selling their inventories more slowly. They do it in order to get a more favorable price from smaller players: it is less costly for large players than for small players to hold interim excessive inventory in order to reduce price impact and hence improve execution price.

As a result, prices are not martingales. The price is skewed from first best if players hold inventories away from the first best. In the absence of further shocks, trade towards the efficient allocation results in price momentum. For example, if large players are net sellers and small players are net buyers at any moment of time, then the price has downward drift. New shocks result in changes in prices that have mean 0 at the time of the shock, but which generate momentum in the future.

The diagonal of Q consists of positive numbers that capture the rates at which each player sells its inventory. Consistent with the explanation above, and what we have seen in Section 3, larger players sell at slower rates.

The off-diagonal terms of Q are negative. The numbers in each column i indicate how the sales of trader i become absorbed by other traders. Interestingly, these entries are much closer to each other than the risk coefficients b^i or the diagonal terms of Q. As a result, in equilibrium flows are not absorbed proportionately to risk capacity. Rather, smaller traders absorb a disproportionately large portion of the flows, and eventually re-trade with large traders who are slower.

Smaller players trade faster, larger players trade slower, and it is useful to think about the overall convergence to the efficient allocation through the eigenvector decomposition of Q.

Eigenvalue Decomposition of Equilibrium Dynamics. To understand the dynamics of equilibrium allocations, it is useful to consider "impulse responses," that is, how allocations evolve over time for any given starting allocation and in the absence of additional shocks. Then full trade, with many shocks, can be thought of as integration of impulse responses of individual shocks.

If no further shocks occur, then the allocation converges towards the efficient allocation according to the equation

$$dX_t = -QX_t\,dt.$$

We can see how quickly different misallocations get traded to efficiency from the eigenvector decomposition of the matrix Q,

$$Q = UKU^{-1}.$$

The columns of U are eigenvectors and K is a diagonal matrix of eigenvalues.

The efficient allocation $[1/b^1, \ldots 1/b^N]^T$ is always an eigenvector of Q with the corresponding eigenvalue 0. In equilibrium, players do not trade if they are at the

efficient allocation: if they traded then at least one player would be worse off than if he had not traded at all. If matrix Q has rank $N-1$ (a requirement for the mechanism (P,Q) to be acceptable), then the remaining eigenvectors of Q have nonzero eigenvalues. Market clearing implies that the coefficients of those eigenvectors add up to 0, i.e. eigenvectors are misallocations away from efficiency, and the corresponding eigenvalues are the rates at which those misallocations get traded away. If in equilibrium players eventually trade to efficiency, then the remaining $N-1$ eigenvalues of Q must be positive.

The following lemma transforms the equilibrium conditions into a simpler form using the eigenvector decomposition of Q.

Lemma 2 *Consider a mechanism characterized by a vector P and a trading matrix Q that has only real eigenvectors, with decomposition $Q = UKU^{-1}$. Then the HJB equations (16) that determine the value functions of the players under truthtelling are equivalent to $\hat{A}^i = U^T A^i U$, $i = 1 \dots N$, where the coefficients of $\hat{A}^i$ are given by*

$$\hat{a}^i_{jk} = \frac{-b^i u^{ij} u^{ik} + (PU^j)u^{ik}\kappa_k + (PU^k)u^{ij}\kappa_j}{2(r+\kappa_k+\kappa_j)}, \tag{23}$$

and κ_k is the k-th diagonal element of K.

Given (23), the first-order conditions (15) can be written as

$$u^{ik}\left(b^i + \kappa_k(r+\kappa_k)\frac{PU\frac{1}{r+\kappa_k+K}(U^{-1})^i}{\underline{U}^i\frac{K}{r+\kappa_k+K}(U^{-1})^i}\right) = -(r+\kappa_k)PU^k. \tag{24}$$

Proof. See Appendix. ■

Equations (23) provide a convenient direct formula to compute the players' value functions from the pair (P,Q). Otherwise, to obtain the matrices A^i from (16), one has to solve a more complicated system of equations, or obtain A^i via an iterative procedure. Equation system (24) is useful because it expresses the equilibrium conditions purely in terms of the vector P and the eigenvector decomposition UKU^{-1} of Q, i.e. it provides a simpler system of equations than the system (15) and (16).

Example continued. Returning to our numerical example, the eigenvector decomposition of Q is given by

$$U = \begin{bmatrix} 1 & 1 & .218 & .118 & .081 \\ .666 & -.530 & .782 & .214 & .123 \\ .5 & -.221 & -.619 & .668 & .213 \\ .4 & -.143 & -.234 & -.748 & .583 \\ .333 & -.106 & -.147 & -.252 & -1 \end{bmatrix}, \quad \text{diag}\, K = [0,\ .93,\ 1.38,\ 1.82,\ 2.24]. \tag{25}$$

Note the sign pattern in the eigenvectors: they break the market into two sides according to risk capacity, with one side of the market selling and the other, buying. The eigenvectors, the columns of Q, are determined only up to a constant. We normalized eigenvectors 2 through N so that one unit in total is misallocated between large sellers and small buyers. Misallocations among the smallest players get traded away faster than those among the largest players in the market: for example, the second eigenvector represents the largest trader having excess inventory while the last vector represents the smallest trader having excess inventory; the latter is traded at more than twice the speed of the former.

5.2 Off-Equilibrium Market Properties: The Price Impact.

In our model, we can study what would happen when a player does not choose the rate of trade optimally, but instead chooses a specific alternative trading rate. Any equilibrium trading rate of a player corresponds to some allocation of that player (given the allocations of others). If a player deviates with respect to the speed of trade, other players would incorrectly infer the deviating player's allocation, and market would react as if the allocation of the deviating player were different. The optimal equilibrium trading rates of players depend on the hypothetical possibility of what would happen if they deviated.

Let us discuss the implications of our model on the price impact of trades in general, building upon our discussion of price impact with a single large trader and fringe in Section 4. The following proposition characterizes the path of prices and allocations under a deviation of player i, as functions of player i's trading rate and allocations of other players.

Proposition 10 *Player i's selling rate q_t^i, given the allocation vector of other players X_t^{-i}, leads other players to believe that player i's allocation is given by*

$$\hat{X}_t^i = \frac{q_t^i - \underline{Q}^{i,-i} X_t^{-i}}{q^{ii}}, \tag{26}$$

where $\underline{Q}^{i,-i}$ represents row i of Q excluding entry i. Consequently, the allocation of players excluding i follows

$$dX_t^{-i} = \left(-Q^{-i,-i} + \frac{Q^{-i,i} \underline{Q}^{i,-i}}{q^{ii}} \right) X_t^{-i}\, dt - q_t^i \frac{Q^{-i,i}}{q^{ii}}\, dt + \underline{\Sigma}^{-i}\, dZ_t, \tag{27}$$

where $Q^{-i,i}$ is column i of Q excluding entry i, and $\underline{\Sigma}^{-i}$ is the matrix constructed from Σ by removing row i. The price is given by

$$p_t = P^{-i} X_t^{-i} + p^i \underbrace{\frac{q_t^i - \underline{Q}^{i,-i} X_t^{-i}}{q^{ii}}}_{\hat{X}_t^i}, \tag{28}$$

where P^{-i} represents vector P excluding entry i.

Proof. Notice that if player i acted as if he had allocation $\hat{X}^i_t$, then the selling rate of player i would be given by

$$q^i_t = q^{ii}\hat{X}^i_t + \underline{Q}^{i,-i}X^{-i}_t.$$

Solving for $\hat{X}^i_t$ given q^i_t, we find (26). Plugging in the allocation $\hat{X}^i_t$ of the type that player i is imitating into the law of motion of X_t and into the price formation equation, we obtain (27) and (28). ■

From (28), we identify several components of the price impact of player i's trades. First, the *instantaneous* price impact measures the sensitivity of price to flow q^i_t, given by

$$I^i = -p^i/q^{ii}. \tag{29}$$

As large traders trade more slowly and exercise market power, they also have a greater instantaneous price impact. That is, from the same selling rate, the market infers that a large player has a greater inventory to sell than a small player, hence the price reacts by a greater amount. This observation is a consequence of our assumption that players observe the identity of traders.

In our example, the vector of instantaneous price impacts is given by

$$[0.403,\ \ 0.341,\ \ 0.301,\ \ 0.272,\ \ 0.251]\,.$$

The price depends on the current flow of a given trader (instantaneous impact), as well as the past trades. If player i sells the total amount of $x = \int_0^t q^i_s\,ds$ and then stops trading, the price will eventually reflect this total volume absorbed by the rest of the market. Given the risk capacity of the market excluding player i, total sales of x causes the price to drop by

$$\Lambda^i\,x, \quad \text{where} \quad \Lambda^i = \frac{1}{\sum_{j\neq i}\ 1/b^j} \tag{30}$$

is the total risk capacity of the market excluding player i. In our example, the vector of permanent price impacts is

$$[0.526,\ \ 0.447,\ \ 0.416,\ \ 0.4,\ \ 0.389]\,.$$

Smaller players face a slightly lower permanent price impact, as the market excluding them has a slightly larger risk capacity.

The impact of sales of x by player i becomes $\Lambda^i x$ only eventually, unless player i faces a market consisting of traders with identical risk capacity. The reason is that players do not absorb the flow of player i proportionately to their risk capacity: the off-diagonal elements of Q in column i are not proportional to risk capacities of the

players. Smaller players absorb a disproportionate share of flow, and if player i stops trading, other players continue trading among themselves. Since smaller players also have a greater price impact, their disproportionate absorption of player i's trades causes the price to overshoot the permanent price impact level. Hence, trades have a transient price impact.

Denote by $T^i + \Lambda^i$ the sensitivity of price to sales of player i, when the sales of x by player i occur over a short interval of time $[0, t]$, as $t \to 0$. Then the vector of transient price impacts T^i in our example is

$$[0.020,\ \ 0.055,\ \ 0.054,\ \ 0.045,\ \ 0.034]\,,$$

and these coefficients can be computed from the following lemma.

Lemma 3 *If sales of x by player i occur over a small interval of time $[0, t]$ and then stop, then as $t \to 0$, total (permanent and transient) impact of trades is given by*

$$(T^i + \Lambda^i)x = \frac{P + I^i \underline{\underline{Q^i}}}{q^{ii}} Q^i x. \tag{31}$$

Proof. The sales of x are initially absorbed by players $j \neq i$ according to $X^{-i}_{0+} - X^{-i}_0 = -Q^{-i,i}/q^{ii}\, x$. Thus, from (28), the price immediately after sales stop is given by

$$p_{0+} = p_0 \underbrace{-P^{-i}\frac{Q^{-i,i}}{q^{ii}}x + p^i \frac{\underline{\underline{Q^{i,-i}}}}{q^{ii}}\frac{Q^{-i,i}}{q^{ii}}x}_{-\frac{P+I^iQ^i}{q^{ii}}Q^i x}\,.$$

Hence, the impact of sales is given by (31). ■

To get a better sense for price impact, suppose that in our example player 3 (with risk capacity of $b^3 = 2$) sells at rate $q^3_t = 1$ over the time interval $[0, 0.5]$ and then stops trading. Assume that other players do not receive any shocks. Then, integrating equations (27) and (28), we get the path of prices shown in Figure 1. To recap, transient impact (which is quantitatively small here) exists because sales of player 3 are not absorbed by others proportionately to risk-capacity, hence after player 3 stops selling, other players continue trading among each other.

The term "transient impact" is also used to capture the empirical pattern that the price, following an interval of sales, overshoots and retracts a bit after sales cease.

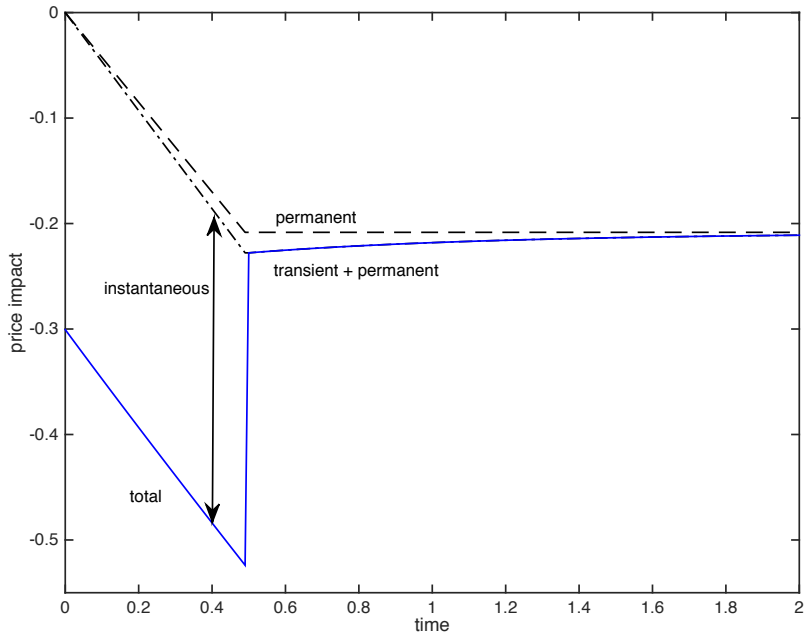

Figure 1: The decomposition of price impact.

5.3 Optimal Trading without Transient Price Impact.

Optimal trading with only instantaneous and permanent price impact has a simple characterization, which we derive here. This case is of special interest, because we can use it to quickly derive equilibrium in the case of a symmetric market, where players have no transient price impact, and approximate equilibria in nearly symmetric markets, where transient impact is second order.

In general, this case leads to a simple approximately optimal rule for trading, which is helpful because (1) transient impact in our model is small and (2) it is complex, since the rates of decay of transient impact depend on the eigenvalues of trading dynamics among players other than player i.

Proposition 11 *Suppose player i's trades have only instantaneous and permanent price impact characterized by coefficients I^i and Λ^i. Then, in order for player i's strategy to be optimal, the following condition has to hold*

$$r(p_t - I^i q_t^i) = -b^i X_t^i + \Lambda^i q_t^i + E_t\left[d(p_t - I^i q_t^i)/dt\right]. \tag{32}$$

Equation (32) generalizes condition (17) of Proposition 4 that characterizes optimal trading of a fringe. Indeed, fringe members act as if their trades have no price impact at all (instantaneous or permanent). In equation (32), $p_t - I^i q_t^i$ is the marginal

revenue of player i from an extra unit sold. Thus, unlike (17), which can be written as

$$rp_t = -b^i X_t^i + E_t\left[dp_t/dt\right],$$

equation (32) is a condition on the level and rate of change of not the price, but rather marginal revenue. In vector form, equation (32) can be written as

$$(P - I^i\underline{Q}^i)(rI - Q) = -b^i\underline{1}^i + \Lambda^i\underline{Q}^i. \tag{33}$$

Notice, in particular, that if player i trades according to (32), player i does not ignore the equilibrium-path price momentum. Rather, this condition requires a high level of sophistication - player i recognizes who the buyers and the sellers in the market are, and how their behavior pushes the price. Player i is only biased about the hypothetical scenario of what would happen if he traded at an off-equilibrium rate, potentially underestimating the impact on price a little bit. When transient price impact exists, equation (32) leads to a slight bias to more aggressive trading relative to a fully rational trader.[22]

Proof. Consider the optimal strategy q_t^i and a deviation $q_t^i + y_t$. The deviation results in the allocation

$$\tilde{X}_t^i = X_t^i - \int_0^t y_s\, ds.$$

If p_t is the price process for the strategy q_t^i, then the price after a deviation, given only the instantaneous and permanent price impacts, becomes

$$p_t - \Lambda^i \int_0^t y_s\, ds - I^i y_t = p_t + \Lambda^i(\tilde{X}_t^i - X_t^i) - I^i y_t.$$

Denote by v_t the player's value function under the strategy q_t^i, and conjecture the following quadratic form

$$a(\tilde{X}_t^i - X_t^i)^2 + b_t(\tilde{X}_t^i - X_t^i) + v_t \tag{34}$$

for the value function after the deviation, where a is an appropriate constant and b_t depends on the stochastic environment of the price process. The HJB equation is

$$r\left(a(\tilde{X}_t^i - X_t^i)^2 + b_t(\tilde{X}_t^i - X_t^i) + v_t\right) = \max_{y_t} \underbrace{-\frac{b^i(\tilde{X}_t^i - X_t^i + X_t^i)^2}{2}}_{-b^i(\tilde{X}_t^i)^2/2,\text{ payoff flow}} +$$

$$(p_t + \Lambda^i(\tilde{X}_t^i - X_t^i) - I^i y_t)(q_t^i + y_t) - 2y_t a(\tilde{X}_t^i - X_t^i) - b_t y_t + \frac{E_t[db_t]}{dt}(\tilde{X}_t^i - X_t^i) + \frac{E_t[dv_t]}{dt}.$$

[22] In general, we believe that there can be huge insights from exploring the dynamics of this model with players of various levels of sophistication. Equations such as (32) can be used to formulate trading rules for particular subgroups of players.

If we solve this equation with $y_t = 0$, then we obtain a lower bound on player i's value function after a deviation (since $y_t = 0$ is not necessarily optimal following a deviation). Matching coefficients on $(\tilde{X}^i_t - X^i_t)^2$ and $\tilde{X}^i_t - X^i_t$, we obtain a lower bound if

$$ra = -\frac{b^i}{2} \quad \text{and} \quad rb_t = -b^i X^i_t + \Lambda^i q^i_t + \frac{E_t[db_t]}{dt} \tag{35}$$

(notice that the coefficient on the constant matches automatically from the definition of v_t as the value function). Given this bound, in order for player i to find it unattractive to deviate when $\tilde{X}^i_t = X^i_t$, the first-order condition with respect to y_t must hold at $y_t = 0$, i.e. we must have

$$b_t = p_t - I^i q^i_t. \tag{36}$$

Combining this with the condition for coefficient b_t we obtain (32).[23] ■

5.4 N Traders with Symmetric Risk Capacities.

Let us use Proposition 11 to characterize dynamics in a market with N traders of identical risk capacities.

Proposition 12 *Consider a market of N traders with identical risk coefficients b. Then the equilibrium price and trading dynamics in the unique non-degenerate symmetric equilibrium are characterized by the pair*

$$P = -\frac{b}{Nr}\underline{1} \quad \textit{and} \quad Q = \underbrace{\frac{N-2}{2}r}_{\kappa}(I - S/N), \tag{37}$$

where $\underline{1}$ is a row vector of ones and S is an $N \times N$ matrix of ones. That is, the allocation converges towards efficiency at the exponential rate κ.

The players' welfare is characterized by the matrices A^i with entries

$$a^i_{ii} = -\frac{rb}{2}\frac{3N-2}{N^2}, \quad a^i_{ij} = -\frac{rb}{2}\frac{N-2}{N^2} \quad \text{and} \quad a^i_{jk} = \frac{rb}{2}\frac{N-2}{(N-1)N^2} \tag{38}$$

[23] We only showed that condition (32) is necessary. For sufficiency, we can show that the actual value function also takes the form (34). Indeed, the first-order condition for y_t in the HJB equation, given (36), is

$$y_t = \frac{(\Lambda^i - 2a)(\tilde{X}^i_t - X^i_t)}{2I^i}.$$

Plugging this into the HJB equation and matching coefficients, we obtain

$$ra = -\frac{b^i}{2} + \frac{(\Lambda^i - 2a)^2}{4I^i}$$

and the same equation for b_t as in (35). That is, the true value function and the upper bound we used have the same slope at $\tilde{X}^i_t = X^i_t$ (obviously, if the first-order condition with respect to y_t is to hold) but the true function is less concave, i.e. a is higher.

for $j, k \neq i$.

Proof. From symmetry, matrix Q must be of the form $\kappa(I - S/N)$ and vector P must have identical coefficients p. Player i's trades have no transient impact (since flow of any player is absorbed efficiently) so price impact of player i is characterized by (we use (31) to get Λ^i)

$$I^i = -\frac{p}{q^{ii}} = -\frac{pN}{(N-1)\kappa} \quad \text{and} \quad \Lambda^i = -\frac{p}{q^{ii}}\frac{\underline{Q}^i Q^i}{q^{ii}} = -\frac{pN}{N-1}.$$

We now use optimal trading condition (32). If $\sum_j X^j = Nx$, then price is $p_t = Npx$. Since expected change in total supply is 0, $E_t[dp_t] = 0$. The allocations converge to the efficiency at rate κ, i.e. $q_t^i = \kappa(X_t^i - x)$, and player's trading rates also decay at expected rate κ, i.e. $E_t[dq_t^i] = -\kappa^2(X_t^i - x)$. Hence, (32) becomes

$$r(pNx - I^i\kappa(X_t^i - x)) = -bX_t^i + \Lambda^i\kappa(X_t^i - x) + I^i\kappa^2(X_t^i - x). \tag{39}$$

Matching coefficients in (32), we obtain

$$x: \quad rp(N-2) = 2p\kappa \quad \Rightarrow \quad \kappa = \frac{N-2}{2}r$$

$$X_t^i: \quad (r+2\kappa)\frac{pN}{N-1} = -b \quad \Rightarrow \quad p = -\frac{b}{Nr}.$$

In the Appendix we provide an alternative proof, using the system of equilibrium equations (15) and (16) directly, and derive the players' value functions. ■

Intuitively, the price vector P must have all the same elements from symmetry, and must correspond to the first best price (otherwise, if players are at the efficient allocation already, they would still want to trade). Since inventory shocks have zero mean, it follows that prices are martingales in equilibrium of the symmetric benchmark. Off-equilibrium, trades have instantaneous and permanent price impact. As a result, players take time to trade to the efficient allocation, and so the equilibrium is inefficient (even though prices are first-best on the equilibrium path).[24] Not surprisingly, the speed of trade is increasing in the number of players in the market, suggesting that market power could be bad for efficiency.[25]

[24] As we mentioned before, an analogous result has already been shown in a slightly different context in Vayanos (1999).

[25] The full effect is more subtle though. If traders merged in pairs (so that the market would move from N to $N/2$ traders) and the merged traders shared their inventories efficiently, for a symmetric Σ the overall welfare would actually remain unchanged. Such a merger would result in two opposite effects that exactly cancel each other out in the symmetric model: the market would converge to efficiency at a slower rate, which hurts efficiency, but the merged players would share their shocks more efficiently, which helps welfare.

It may seem surprising at first that value function matrices A^i have positive coefficients everywhere outside column and row i (see equation (38)). The reason is that, while players experience disutility from shocks to their own allocations - which they have to hold or pay other traders to offload - players actually earn utility by taking excess inventory off of other players. Therefore, in equilibrium players who help other players trade - liquidity providers - may have positive expected utility.

5.5 Approximation near Symmetric Risk Capacities.

While in general we cannot characterize equilibrium in closed form, we can approximate equilibria, up to terms of order ϵ^2, near symmetric risk coefficients, perturbed by terms of order ϵ. As we show in a bit, the approximate formulas we obtain are extremely useful. First, in practice the approximate formulas generate answers that are quite precise even in cases where risk coefficients b^i are not nearly symmetric. Second, we can use the approximate formulas to illustrate analytically the important economic properties of asymmetric markets.

It is possible to derive an approximation of equilibria near the symmetric case using formulas (15) and (16) for the perturbation of risk coefficients

$$b\,(\underline{1} + \epsilon), \tag{40}$$

directly by keeping only terms of order $|\epsilon|$. In (40), $\underline{1}$ is a row vector of ones and ϵ is a small vector whose coefficients add up to 0.

Proposition 13 *Consider the perturbation of symmetric risk coefficients (40). Then, equilibrium equations hold, up to terms of order $|\epsilon|^2$, for the pair $(\hat{P}, \hat{Q})$ given by*

$$\hat{P} = P(I + x^P D_\epsilon), \quad x^P = \frac{3N-4}{N(N-1)} \quad \textit{and} \quad \hat{Q} = Q(I + D_\epsilon), \tag{41}$$

where P and Q are the pair that solves the symmetric case (37), and D_ϵ is a diagonal matrix with vector ϵ on the diagonal. The transient price impact is of order $|\epsilon|^2$.

Recall that post-multiplication by a diagonal matrix $I + D_\epsilon$ in (41) results in the multiplication of the columns of the preceding matrix (in this case Q) by the entries of the diagonal elements.

We take a shortcut to prove Proposition 13, using the property that transient price impacts are of order $|\epsilon|^2$ near the symmetric case.[26] Then the the characterization of Proposition 11 applies. As a result, we bypass the algebra that involves value function matrices A^i, and focus only on equations that contain P and Q.

[26]Transient impact is of second-order because absorption of sales by any player is up to order ϵ proportionate to risk capacities and price is, up to order ϵ, first best. Taking these effects together, transient price impact is of second order. See the proof of Proposition 13.

Proof. We have from (29) and (30), using expressions for $\hat{P}, \hat{Q}$ in (41),

$$I^i = \frac{b/r}{(N-1)\kappa}(1 + (x^P - 1)\epsilon^i) + O(|\epsilon|^2), \tag{42}$$

$$\text{and} \quad \Lambda^i = \frac{b/r}{N-1}\left(1 - \frac{\epsilon^i}{N-1}\right) + O(|\epsilon|^2).$$

Furthermore, (31) yields

$$T^i + \Lambda^i = \frac{\overbrace{\frac{b\kappa}{rN}\left(1 + \frac{N-2}{N-1}\epsilon^i\right) + O(|\epsilon|^2)}^{(\hat{P}+I^i\hat{Q}^i)\hat{Q}^i}}{\underbrace{\frac{N-1}{N}\kappa\left(1+\epsilon^i\right) + O(|\epsilon|^2)}_{q^{ii}}} = \Lambda^i + O(|\epsilon|^2).$$

Hence, transient impact is of order $|\epsilon|^2$, and we can use equations (33) to characterize (the approximate) equilibrium. In matrix form, these equations can be written as

$$D_\Lambda \hat{Q} = D_B + (1^T\hat{P} - D_I\hat{Q})(rI + \hat{Q}), \tag{43}$$

where D_Λ, D_I and D_B are diagonal matrices of permanent and instantaneous price impacts, and the players' risk coefficients, respectively. Plugging in (41), we obtain[27]

$$\frac{b/r}{N-1}\left(I - \frac{D_\epsilon}{N-1}\right)Q(I + D_\epsilon) = b(I + D_\epsilon) -$$

$$\left(\frac{b/r}{N}S(I + x^P D_\epsilon) + \frac{b/r}{(N-1)\kappa}(I + (x^P - 1)D_\epsilon)Q(I + D_\epsilon)\right)(rI + Q(I + D_\epsilon)).$$

Collecting terms of order ϵ (and multiplying by r/b), we obtain

$$-\frac{D_\epsilon Q}{(N-1)^2} + \frac{QD_\epsilon}{N-1} = rD_\epsilon -$$

$$\left(\frac{x^P}{N}SD_\epsilon + \frac{x^P - 1}{N-1}D_\epsilon(I - S/N) + \frac{I - S/N}{N-1}D_\epsilon\right)(rI + Q) - \left(\frac{S}{N} + \frac{I - S/N}{N-1}\right)QD_\epsilon.$$

To make sure that (43) holds up to terms of order $|\epsilon|^2$, we need to make sure that coefficients in front of D_ϵ, $D_\epsilon S$ and SD_ϵ match (notice that $SD_\epsilon S = 0$). Collecting terms that multiply D_ϵ, $D_\epsilon S$ and SD_ϵ, respectively, we get

$$-\frac{\kappa}{(N-1)^2} + \frac{\kappa}{N-1} = r - \frac{x^P}{N-1}(r + \kappa) - \frac{\kappa}{N-1},$$

[27]Recall that S is an $N \times N$ matrix of ones.

$$\frac{\kappa}{N(N-1)^2} = \frac{1}{N-1}\frac{\kappa}{N} + \frac{x^P - 1}{N-1}\frac{r+\kappa}{N} \quad \text{and}$$

$$-\frac{\kappa/N}{N-1} = -\left(\frac{x^P}{N} - \frac{1}{N(N-1)}\right)(r+\kappa) + \frac{\kappa}{N(N-1)}.$$

It is easy to verify that each of these three equations holds with x^P given by (41). ■

The approximation turns out to be quite precise numerically even for risk capacities away from symmetric. For example, consider an example with five players, with risk coefficients [1, 1.5, 2, 2.5, 3]. Then the approximation gives

$$\hat{P} = -\left[.29\ .345\ .4\ .455\ .51\right], \quad \hat{Q} = \begin{bmatrix} .6 & -.225 & -.3 & -.38 & -.45 \\ -.15 & .9 & -.3 & -.38 & -.45 \\ -.15 & -.225 & 1.2 & -.38 & -.45 \\ -.15 & -.225 & -.3 & 1.5 & -.45 \\ -.15 & -.225 & -.3 & -.38 & 1.8 \end{bmatrix}. \tag{44}$$

The true solution, computed from equations (15) and (16) directly, is

$$-\left[.254\ .3294\ .3875\ .4351\ .4757\right], \quad \begin{bmatrix} .6298 & -.2441 & -.3192 & -.3894 & -.4554 \\ -.1628 & .9653 & -.3256 & -.4010 & -.4727 \\ -.1596 & -.2442 & 1.2893 & -.4050 & -.4808 \\ -.1557 & -.2406 & -.3240 & 1.5983 & -.4835 \\ -.1518 & -.2364 & -.3205 & -.4029 & 1.8924 \end{bmatrix}.$$

The approximation is surprisingly good in practice for a vector of risk coefficients quite away from symmetric (but without coefficients differing on the order of a magnitude). There is a slight bias in magnitude, but the approximation reveals coefficients of P that look "right" relative to each other. Moreover, the approximation highlights important properties of the true Q, specifically that diagonal coefficients are roughly proportional to risk capacities, and off-diagonal coefficients in each column are roughly identical. Finally, it is a nice property of $\hat{Q}$ that even if it is an approximation, it results in trade to the true efficient allocation (and not an approximately efficient allocation).

Overall, we think it is fair to say that (while it is unfortunate that our general equations do not have a closed-form solution) the approximation of Proposition 13 is of huge value: it conveys the economic properties of the model and does it through simple formulas. For example, the columns of (44) indicate that when player i sells, the flow is absorbed approximately equally by all other traders despite the differences in their risk capacities. That is, small players absorb flow more quickly (and if player i stops trading, they resell some of what they bought to other larger traders). This gives rise to transient price impact (away from the symmetric case), because price is more sensitive to allocations of small players than to those of large traders.

Let us reiterate some of the most important on and off-equilibrium properties of our model in light of the approximation. First, let us discuss equilibrium price momentum.

From (41) we see that the price is less sensitive to the inventories of large players than those of small players (recall the players with the most negative ϵ^i are the largest). As we explained above, the intuition is that large players have market power and they control their trading rates better, while the small players compete with each other and pay much less attention to the impact of their trades on the price. Hence, equilibrium price is skewed in favor of large players, relative to first-best, and it tends to first best over time as the allocation converges. Second, let us discuss the speed of trade. The diagonal elements of $\hat{Q}$ are proportional to risk coefficients. Hence, in this approximation the speed of trade of individual players is inversely proportional to their risk capacities. The following proposition characterizes eigenvectors and eigenvalues of the approximating matrix $\hat{Q}$.

Proposition 14 *Suppose that the coefficients of vector ϵ are strictly increasing and $1+\epsilon^1 > 0$ (that is, players are ordered from largest to smallest, and the risk coefficient of the largest player is $b^1 = b(1+\epsilon^1) > 0$). Then the eigenvalues of matrix $\hat{Q} = \kappa(I - S/N)(I + D_\epsilon)$ satisfy*

$$\kappa_1 = 0, \;\; \kappa_2 \in (\kappa(1+\epsilon^1),\, \kappa(1+\epsilon^2)), \;\; \ldots \;\; \kappa_N \in (\kappa(1+\epsilon^{N-1}),\, \kappa(1+\epsilon^N))$$

and corresponding eigenvectors take the form (for $\kappa' = \kappa_i$)

$$v = [v_1, v_2, \ldots v_N]^T \quad with \quad v_i = \frac{1}{1+\epsilon^i - \kappa'/\kappa}. \tag{45}$$

The decomposition confirms the curious sign pattern we observed in our example (25). Notice that the first $i-1$ coefficients of vector v_i are positive, in increasing order, while the remaining coefficients are negative (also in increasing order). Hence, in each eigenvector, large players are on one side of the market while small players are on the other side. The fewer players there are in the group of small traders, the larger the corresponding eigenvalue, which corresponds to the speed of trade. Notice that the first eigenvector is the efficient allocation.

Proof. Consider a vector $v = [v_1, v_2, \ldots v_N]^T$, and let us normalize $\sum_i (1+\epsilon^i)v_i = N$. Then $\hat{Q}v = \kappa[(1+\epsilon^1)v_1 - 1, \ldots (1+\epsilon^N)v_N - 1]$. If v is an eigenvector with eigenvalue κ', then the coefficients v_i satisfy $\kappa((1+\epsilon^i)v_i - 1) = \kappa' v_i$ or (45).

For the normalization to hold, we must have

$$\sum_i \frac{1+\epsilon^i}{1+\epsilon^i - \kappa'/\kappa} = N. \tag{46}$$

Equation (46) has solution $\kappa' = 0$. In addition, for each $i = 2, \dots N$, as κ'/κ goes through the interval $(1 + \epsilon^{i-1}, 1 + \epsilon^i)$, the sum

$$\frac{1 + \epsilon^{i-1}}{1 + \epsilon^{i-1} - \kappa'/\kappa} + \frac{1 + \epsilon^i}{1 + \epsilon^i - \kappa'/\kappa}$$

goes from $-\infty$ to ∞, while the remaining terms of (46) stay finite. Hence, equation (46) has a root $\kappa' \in \kappa(1 + \epsilon^{i-1}, 1 + \epsilon^i)$. Corresponding to each of these roots, there is an eigenvector with coefficients of the form (45). Since we have found N distinct eigenvalues κ_i, and $\hat{Q}$ is $N \times N$, we have found all eigenvalues. ■

Finally, let us turn to off-equilibrium properties, i.e. price impact. From (42), instantaneous price impact (the sensitivity of price to trading flow) of player i is given by

$$I^i = \frac{b/r}{(N-1)\kappa} \left(1 - \frac{(N-2)^2}{N(N-1)} \epsilon^i \right) + O(|\epsilon|^2).$$

Large players (with small ϵ^i) have greater price impact than small players. This property stems from differences in equilibrium strategies - large players trade slower and hence their flow signals that they have a larger amount waiting to be traded.

Permanent impact reflects simply with the risk capacity of other traders. Finally, as we discussed earlier, the existence of transient impact has to do with the off-diagonal entries of the matrix Q - smaller players, whose allocations have a greater impact on price, initially absorb a greater portion of flow than their risk capacity dictates.

5.6 Application: Technical Front-Running.

In Section 4, we saw a basic example of front-running of large trader's sales. Fringe members, who start with first-best inventories, pick up supply from other fringe members (in an instant) and then sell to the large trader. In this particular pattern individual fringe members have no price impact and hence do not optimize the timing of their buying. We now illustrate the process of technical front-running in more detail using a model of multiple symmetric large players and a fringe. We also illustrate how properties of prices as well as the speed of trade towards the efficient allocation can be understood through the eigenvector decomposition of the matrix Q.

Consider a market with $N-1$ large players with identical risk capacities $1/b^L$ and a competitive fringe with risk capacity $1/b^F$. In this case trading is characterized by the matrix $Q = UKU^{-1}$, where

$$U = \begin{bmatrix} 1/b^L & 1 & \dots & 1 & 1 \\ 1/b^L & -1 & \dots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1/b^L & 0 & \dots & -1 & 1 \\ 1/b^F & 0 & \dots & 0 & 1-N \end{bmatrix} \quad \textit{and} \quad K = \text{diag}\,[0,\ \kappa,\ \dots,\ \kappa,\ \bar{\kappa}], \tag{47}$$

with $\kappa < \bar{\kappa}$. The price is given by $P = \Pi U^{-1}$, where

$$\Pi = \left[-\frac{1}{r}, \quad 0, \quad \ldots, \quad 0, \quad \pi^B \right], \quad where \;\; \pi^B = \frac{(N-1)b^F}{r + \bar{\kappa}}.$$

The eigenvector decomposition of trading dynamics can be understood as follows. The entries of Π correspond to the prices that result from allocations in the corresponding columns of U. The first column is the efficient allocation. Columns 2 through N are misallocations, with coefficients that add up to 0. These eigenvector-misallocations get traded to efficiency at rates given by the eigenvalues κ and $\bar{\kappa}$. Any allocation X can be represented as a linear combination of eigenvectors, to see how it is traded as well as the resulting price.

From symmetry, any misallocation among the large traders has no effect on the price. This is illustrated by the zero coefficients 2 through $N-1$ of Π; the corresponding columns of U span the space of various misallocations among large players. In contrast, any imbalance between the large traders and the fringe leads to a price that is different from first best. When the large traders are net sellers, the price is above first best, as illustrated by the positive value of π^B. That is, the price is π^B for the allocation X that corresponds to the last column of U.

Let us illustrate how front-running can happen in this model. With two large players and a fringe, suppose that player 1 wants to sell one unit, while the fringe wants to buy one unit. Player 2 wants neither to buy nor sell, so that $X_0 = [1 \; 0 \; -1]^T$. Because player 1 has market power against the fringe, he sells slowly and charges a price above first best. This can be seen by the positive last component of Π, which prices the excess inventory of the large players relative to the fringe. Over time, the price drifts down and converges to first best.

Knowing this, of course, player 2 does not stay at his bliss point. Rather, he sells short initially to the fringe, and buys back from player 1 later. In other words, player 2 front-runs player 1. The mere presence of player 2 in the market changes the dynamics between player 1 and the fringe. Player 1 trades much faster knowing that there is a front runner. The trading speed $\bar{\kappa}$ between the two large players and the fringe is faster than the trading speed κ given by (19), which would have resulted had player 2 been absent.

In this example, the initial allocation can be decomposed into eigenvectors in the following way

$$\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \frac{U^2 + U^3}{2} = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 1/2 \\ -1 \end{bmatrix}.$$

That is, the allocation $X_0 = [1 \;\; 0 \;\; -1]^T$ consists of an imbalance between the large players, and between large players and the fringe. The former gets traded to efficiency much more slowly than the latter, with the corresponding convergence rates $\kappa < \bar{\kappa}$. Thus, at the beginning player 2 will be primarily selling to the fringe, and buying

only a little from player 1. The misallocation between the two large players persists a lot longer than the misallocation between the large traders and the fringe.

6 Common Values.

Our model has correlated *private* values. In Section 2 we also provided an extension of the model to incorporate *common* values, with payoff flows given by (9). While a complete analysis of this case is beyond the scope of this paper, we conjecture that when the common value components are not too large, the fully separating linear equilibrium of the conditional double auction still exists. The corresponding direct revelation mechanism asks players to report their private signals ξ_t^i, and sets prices and trade flows according to

$$p_t = P(\underbrace{\hat{x}_t - F\xi_t}_{X_t}), \quad q_t = Q(\hat{x}_t - F\xi_t),$$

where holdings $\hat{x}_t$ are computed directly from the past history of trades. If player i reports signal $\xi_t^i + y$ rather than ξ_t^i, the report affects the entire vector X_t that feeds into the mechanism, and not only component X_t^i. As a result, the first-order condition for player i to report truthfully becomes

$$(PF^i)\underline{Q}^i + P(\underline{Q}^i F^i) = 2(A^i Q F^i)^T,$$

rather than (15), where F^i is the ith column of F. Matrix A^i satisfies the same HJB equation (16), while the coefficient k^i of the value function is now

$$k^i = \frac{\operatorname{tr}\left[\Sigma F F^T \Sigma^T A^i\right]}{r}.$$

We use these equations to illustrate how common values affect trade via two examples. First, we extend the symmetric result of Proposition 12 to the case of common values. Second, we explore what happens when one particular player has private information about fundamentals.

In the symmetric case with $b^i = b$ for all i and matrix F as in (10), we can replicate our calculations from the proof of Proposition 12 to obtain

$$P = -\frac{b}{Nr}\underline{1} \quad \text{and} \quad Q = \kappa(I - S/N), \quad \text{where} \quad \kappa = \frac{(N-2)(1-\phi) - \phi N}{2(1-\phi) + 2\phi N}.$$

If $\phi = 0$, $\kappa = (N-2)/2$, as in (37). As ϕ goes up, κ goes down. Intuitively, as the common value concerns increase, trade slows down.[28] As before, κ increases in N, so

[28] Du and Zhu (2013) present an analogous result.

more competition speeds up trade. However, in order for the equilibrium with trade to exist, we need $\kappa > 0$, or

$$\phi < \frac{N-2}{2(N-1)}.$$

The common value component of signals cannot be too large. As $N \to \infty$ this critical value of ϕ converges to 1/2.

For the asymmetric case, we extend the numerical example in Section 5.1 to provide some casual intuition. To that end, suppose that

$$F = \begin{bmatrix} 1 & 0.2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0.2 & 1 & 0 & 0 \\ 0 & 0.2 & 0 & 1 & 0 \\ 0 & 0.2 & 0 & 0 & 1 \end{bmatrix}$$

rather than the identity matrix, and the rest of the parameters remain unchanged. According to matrix F, traders 1, 3, 4 and 5 have signals about only their private value, while trader 2's signal contains information about common value that impacts other traders symmetrically. In this case the (P, Q) change to

$$P = -\left[0.276, \quad 0.180, \quad 0.439, \quad 0.500, \quad 0.553\right],$$

$$Q = \begin{bmatrix} 0.503 & -0.089 & -0.303 & -0.370 & -0.433 \\ -0.062 & 0.335 & -0.113 & -0.134 & -0.154 \\ -0.151 & -0.085 & 1.018 & -0.380 & -0.450 \\ -0.147 & -0.082 & -0.303 & 1.259 & -0.451 \\ -0.143 & -0.079 & -0.299 & -0.375 & 1.487 \end{bmatrix}.$$

Comparison to the example in Section 5.1 reveals the following three effects of common-value information in the signals of trader 2. First, trade gets slower overall. This can be seen by noticing that all entries in Q get closer to 0.[29] Second, common value can create a reversal in the speed of trade: trader 2 now trades slower than trader 1 even though he has higher inventory costs. This is because the common-value information significantly increases player 2's price impact.[30] Finally, the off-diagonal elements of Q change in the following way: when trader 2 sells, his flow is still absorbed approximately equally by all other traders. But when others trade, trader 2

[29] A general pattern we observe when we introduce positive off-diagonal elements in F is that the diagonal entries of Q get smaller; the exact pattern for off-diagonal terms is more complicated.

[30] If player i changes the announcement of his shock by y, the mechanisms' perception of the vector X changes by $F^i y$, and hence player i's flow changes by $\underline{Q}^i F^i y$. At the same time, price changes by $PF^i y$. Hence player i's instantaneous price impact with common-value information is $-PF^i/(\underline{Q}^i F^i)$. For player 2 in our example, instantaneous price impact is 2.2 and significantly higher than those for other players.

absorbs a disproportionately small portion of their flow. Again, trades of 2 are slowed down by the high price impact. Intuitively, other players may be worried that the willingness to trade by player 2 is a signal of unfavorable common-value information.

7 Discussion

There is a large literature in financial mathematics that focuses on optimal trading under various transaction costs and microstructure frictions. The focus of this literature is largely on optimal behavior of an individual trader given a realistic price process, and a realistic model of how trades affect prices - as in Almgren and Chriss (2001). Prices are often modeled as an arithmetic Brownian motion (so they may be negative) - as Almgren and Li (2016) argue, while geometric Brownian motion is a more realistic representation of prices, arithmetic Brownian motion is an adequate and tractable representation of prices for the time horizons over which trades take place (e.g. over a day). Trades have instantaneous and permanent price impact, and possibly transient price impact characterized by an appropriate decay kernel, as in Gatheral, Schied and Slynko (2011). While in practice price impact may not be linear in flow - e.g. Almgren, Thum, Hauptmann, and Li (2005) document that it may be concave for small trades - linearity is a common assumption. Concavity may reflect transaction costs due to discreteness of micro-level microstructure, but linearity leads to greater tractability, and certain results suggest that only linearity is consistent with absence of arbitrage (e.g. see Gatheral (2010) and Huberman and Stanzl (2004)).

Our paper follows this literature's focus on realistic price behavior. However, instead of assuming that prices are Brownian and that trades have instantaneous and permanent impact, we derive these properties in equilibrium. That is, we model a market in which individual participants have objectives similar to those of individual traders in Bank, Soner and Voß (2016) - they have evolving hedging motives for trading individual assets - but we derive price behavior in these markets, and characterize price impact faced by individual participants. We obtain price behavior consistent with the assumptions of this literature, and because we model the interaction of all traders, we tie the properties of price impact to market characteristics (i.e. the number of market participants and their market power). Thus, our model can be thought of as an explanation for why price impact has a particular form, and how it is tied to market characteristics.

In order to characterize equilibrium in a tractable manner, we assume that trading is not anonymous, but rather players can condition their behavior on who they are trading against. This assumption is not completely realistic, even though it gives players information that traders would like to know, since flow of large investors tends to be much more persistent than that of small traders. However, with this assumption we get important benefits. First, we obtain a stationary characterization of equilibria - market liquidity does not change stochastically depending on the active players in the market. Second, we get a characterization of price impact that depends only on

the allocation of market power among players, and not on trading needs or activity. Finally, in solving the model, we avoid the complex issue of filtering that players would otherwise need to perform to infer the source of trades from price behavior. Moreover, our paper makes a methodological contribution – we model trading as a conditional double auction to allow players to take into account who they are trading against.

By modeling the entire market, we are able to answer questions not just about optimal execution of transactions, but also about welfare, determinants of liquidity, and strategic interactions among traders. Of course, we are not the first ones to model the whole market - Vayanos (1999), Du and Zhu (2013) and Kyle, Obizhaeva and Wang (2014) model symmetric markets, while Malamud and Rostek (2014) and Lambert, Ostrovsky and Panov (2016) model asymmetric markets in a one-shot setting - we model dynamics of an asymmetric market. Large traders trade slower, and so their flow is highly persistent, consistently with empirical literature. This leads to price momentum, which is absent from symmetric models. Momentum from the flow of large traders gives rise to the front-running phenomenon in our model. For example, some players may sell short together with a large trader early on, and buy back later after the price drifts down. The front-running phenomenon helps reconcile the persistence of institutional flow and price momentum with market efficiency - a question addressed also in Farmer et al (2006). Our answer is that while flow of large traders is persistent, market price reacts right away fully to the news of the entire supply of a large trader, as the front runners trade with the large trader. That is, the market recognizes the entire quantity that the large player would like to sell. However, momentum still exists, reflecting limits to arbitrage. Front runners recognize momentum, but they are already exploiting it to the fullest given their risk capacity. Overall, our model suggests that the front-running phenomenon not only improves efficiency, as prices deviate from efficiency less and converge to efficiency faster in more competitive markets, but that it can be good for large traders to be front run. That is, while front runners take away large players' profits, they also make it easier to execute transactions with less price impact.

There are a number of questions that are at the center of the debate about financial market regulation, but which cannot be addressed definitively given existing theory. While our model is too simple for many of these questions, with the focus on the entire market and with explicit modeling of players' utilities, it can be used to think about a subset of them. One question is about opaque versus transparent markets. Our model is transparent, because traders observe flows of other traders, but how would the equilibrium change if players did not see that? Presumably, large traders would have lower price impact, and so they could trade faster, as they could hide their flow behind that of small traders. If so, then what is the effect of separation of flows of large and small traders, e.g. when some market makers pay retail brokerages to execute retail flow. Does this improve execution prices of retail traders, as these market makers claim? It seems quite likely, since our model suggests that trades of

small players should have lower price impact. However, another implication is that the residual flow has a greater fraction of trades of institutional investors, and is therefore more persistent. This leads to lower market liquidity overall, and this equilibrium effect could harm retail investors who seem to be getting better execution.

Our model approximates flow as continuous, whereas in practice transactions are discrete. Is there any relationship, then, between our results and bid-ask spreads? Potentially there is a useful link. Suppose bid-ask spreads are set by liquidity providers who take into account tradeoff between risk and profit that they can make from facilitating transactions. Then the size of the bid-ask spread should depend on competition among liquidity providers, as well as properties of the flow, e.g. persistence, and our model speaks to those properties.

8 Conclusions

We would like to summarize the main contribution of the paper. First, the paper makes a methodological contribution. We introduce a linear-quadratic model of conditional double auctions in a market where traders have hedging motives to trade. This model provides a tractable way to analyze the market with heterogeneous participants. Second, we characterize important on and off-equilibrium properties of trade. On equilibrium players trade slowly to reduce price impact, and flow of larger traders is more persistent than that of small traders. This heterogeneous behavior leads to price momentum, and creates incentives for players to know the source of flow for technical front-running. Off equilibrium, our characterization of price impact matches the classical assumptions of financial mathematics literature - trades have instantaneous, permanent and transient impact. Our results have welfare implications - for example, we show that market power, characterized as the ability to coordinate the execution of a large number of trades, may be detrimental to players' welfare in a fully transparent market, i.e. when large traders cannot hide flow behind those of small traders.

A Appendix A: A Microfoundation of Quadratic Preferences.

In this section we microfound our model with quadratic preferences by laying out a more natural model with exponential utilities, in which players trade to hedge private shocks that expose them to a common risk factor. We show that the equilibrium equations of the linear-quadratic model match those of the exponential model in the special case when the shocks that expose players to the common risk factor become small. In this sense, the exponential model is more general, but the linear-quadratic model provides a clean special case as the equilibrium dynamics, characterized by the pair (P, Q), depend only on the players' risk capacities and not the sizes of shocks that individual players receive, or the correlation among shocks.

We then extend the model to also allow shocks to carry information about a common component of value. We confirm the result of Du and Zhu (2013) that in symmetric markets, as the players get more information about common fundamentals, the speed of trade slows down. In general asymmetric markets, equilibrium in this more general setting is characterized by a system of equations analogous to what we described in the text, with only one extra term.

A.1 The Exponential Model.

Consider a model in which all players $i = 1, \dots N$ have exponential utility

$$-\exp(-\alpha^i c_t),$$

where $\alpha^i > 0$ is the coefficient of absolute risk aversion. Players consume continuously and have a common discount rate r, which is also the risk-free rate in the market.

Players have private information about their risk exposure X_t^i to a common Brownian risk factor dW_t. Risk exposure changes due to shocks driven by Brownian motions $Z = [Z_t^1, \dots Z_t^N]$, independent among each other and of W_t with a vector of volatilities σ. Risk exposures can be traded in the market. We consider a (linear, stationary) equilibrium, in which players announce their risk exposures, and given a vector of announcements $\tilde{X}_t$, the trading flows are given by $Q\tilde{X}_t$, and the market price is given by $P\tilde{X}_t$. Then the risk exposures follow

$$dX_t = \Sigma\, dZ_t - Q\tilde{X}_t\, dt$$

where Σ is the diagonal matrix with the elements of σ on the diagonal. The wealth of agent i follows

$$dw_t^i = (rw_t^i - c_t^i)\, dt + (P\tilde{X}_t)(\underline{Q}^i\tilde{X}_t)\, dt + X_t^i\, dW_t,$$

where c_t^i is the consumption of player i.

Conjecture that the equilibrium value function of player i takes the form

$$-\frac{1}{r}\exp(-r\alpha^i \underbrace{(w_t^i + X_t^T A^i X_t + k^i)}_{v_t^i}). \tag{48}$$

Then

$$dv_t^i = (rw_t^i - c_t^i)\,dt + (P\tilde{X}_t)(\underline{Q}^i\tilde{X}_t)\,dt + X_t^i\,dW_t$$
$$+2X_t^T A^i(\Sigma\,dZ_t - Q\tilde{X}_t\,dt) + \operatorname{tr}\left[\Sigma\Sigma^T A^i\right]dt.$$

In order to write down the HJB equation for player i, we must consider $\tilde{X}_t^i$ of the form $X_t + 1^i y$, where 1^i is the i-th coordinate vector and y is the amount by which player i lies.

Then the HJB equation of player i is

$$-\exp(-r\alpha^i v^i) = \max_{c,\,\hat{X}=X+1^i y} -\exp(-\alpha^i c)$$

$$+\alpha^i\exp(-r\alpha^i v^i)\left(rw^i - c^i + (P\tilde{X})(\underline{Q}^i\tilde{X}) - 2X^T A^i Q\tilde{X} + \operatorname{tr}\left[\Sigma\Sigma^T A^i\right]\right)$$
$$-\frac{r(\alpha^i)^2}{2}\exp(-r\alpha^i v^i)\left(4X^T A^i\Sigma\Sigma^T A^i X + (X^i)^2\right).$$

The term $4X^T A^i\Sigma\Sigma^T A^i X$ is the incremental variance of v_t^i from the volatility of the entire vector X_t.[31]

The first-order condition with respect to c is

$$\exp(-\alpha^i c) = \exp(-r\alpha^i v^i) \quad \Leftrightarrow \quad -c = -r(w^i + X^T A^i X + k^i).$$

Given this, the HJB equation simplifies to

$$0 = \max_{\hat{X}=X+1^i y} -r(X^T A^i X + k^i) + (P\tilde{X})(\underline{Q}^i\tilde{X}) - 2X^T A^i Q\tilde{X} + \operatorname{tr}\left[\Sigma\Sigma^T A^i\right]$$

$$-\frac{r\alpha^i}{2}\left(4X^T A^i\Sigma\Sigma^T A^i X + (X^i)^2\right). \tag{49}$$

Separating the first-order condition, we obtain matrix equations that characterize stationary linear equilibria in this model. We summarize them in the following proposition.

[31] This expression assumes that A^i is symmetric, otherwise the second instance of A^i would need to be replaced with $(A^i)^T$.

Proposition 15 *Stationary linear equilibria of the exponential model are characterized by the equations*

$$P^i\underline{Q}^i + Q^{ii}P = 2(A^iQ^i)^T, \quad rk^i = \text{tr}\,[\Sigma\Sigma^T A^i], \tag{50}$$

$$\textit{and} \quad rA^i + A^iQ + QA^i = \frac{P^T\underline{Q}^i + (Q^i)^TP}{2} - \frac{r\alpha^i}{2}1^{ii} - 2r\alpha^iA^i\Sigma\Sigma^TA^i. \tag{51}$$

Proof. Equation (49) must hold for all vectors $X \in \mathbb{R}^N$. To ensure that, the coefficients on the constant term as well as the terms of the form X^jX^k must match, and the first-order condition with respect to y must hold at $y = 0$. From those conditions, we obtain (50) and (51). ■

The system of (50) and (51) is different from equations (15) and (16) in the linear-quadratic model only in the term $2r\alpha^iA^i\Sigma\Sigma^TA^i$. Parameter b^i in the linear-quadratic model corresponds to $r\alpha^i$ in the exponential model, i.e. it reflects the players' capacities to wait and absorb risk waiting for a better price to hedge at. In the limit as $\sigma \to 0$, the equations in the exponential model become identical to those in the linear-quadratic model. Thus, the linear-quadratic model is a special case of the exponential model. We summarize this finding in the following proposition.

Proposition 16 *Any solution of the linear-quadratic model solves equations (50) and (51) in the limit as* $|\Sigma| \to 0$.

Proof. The conclusion follows immediately, since the term that distinguishes the two sets of equations converges to 0 as $|\Sigma| \to 0$. ■

Even though the exponential case is more general, the linear-quadratic model provides a much cleaner picture of equilibrium dynamics, as the equilibrium equations depend only on the players' risk capacities and not the distribution of shocks. This makes our benchmark case particularly attractive. Nevertheless, in order to provide a more complete picture, we present a couple of computed examples for the general case at the end of this section.

A.2 Examples.

We return to the example from Section 5 to explore how the extra term in (51) affects prices and the rates of trading. Recall that the risk coefficients are $[b^1, b^2, b^3, b^4, b^5] = [1, 1.5, 2, 2.5, 3]$ and $r = 1$ in that example. Then the coefficients of absolute risk aversion are also $[\alpha^1, \alpha^2, \alpha^3, \alpha^4, \alpha^5] = [1, 1.5, 2, 2.5, 3]$. Assume that $\Sigma = 0.1I$, i.e. shocks to individual players are uncorrelated and have volatility 0.1.

Then, we have

$$P = [-.257,\ -.334,\ -.394,\ -.443,\ -.485].$$

The price sensitivities to the allocations of all players increase slightly. Trading dynamics are now characterized by

$$Q = \begin{bmatrix} 0.619 & -0.243 & -0.318 & -0.388 & -0.454 \\ -0.161 & 0.953 & -0.324 & -0.400 & -0.472 \\ -0.157 & -0.242 & 1.278 & -0.403 & -0.479 \\ -0.153 & -0.237 & -0.320 & 1.589 & -0.481 \\ -0.147 & -0.231 & -0.315 & -0.398 & 1.886 \end{bmatrix}.$$

The speed of trading slows down somewhat, but qualitatively and quantitatively the solution looks similar to our baseline model.

Now, consider $\Sigma = \text{diag}\,[0.1, 0.3, 0.3, 0.1, 0.1]$. We raise the fundamental needs to trade of players 2 and 3, while keeping shocks to everyone else the same. Now, players 1, 4 and 5 can provide liquidity to players 2 and 3, and help them share risks. Then the trading dynamics are characterized by the price vector

$$P = [-.265,\ -.358,\ -.426,\ -.463,\ -.507]$$

and the trading matrix

$$Q = \begin{bmatrix} 0.572 & -0.246 & -0.326 & -0.374 & -0.437 \\ -0.146 & 0.951 & -0.322 & -0.374 & -0.445 \\ -0.131 & -0.222 & 1.312 & -0.364 & -0.439 \\ -0.150 & -0.244 & -0.334 & 1.500 & -0.467 \\ -0.145 & -0.239 & -0.330 & -0.387 & 1.787 \end{bmatrix}.$$

The price impact of shocks rises and trade tends to slow down, especially for players who are hit by relatively smaller shocks.

These examples seem to imply that the more general model with exponential utility does not add much intuition about market dynamics on top of what the baseline linear-quadratic model already tells us. Of course, there may be interesting effects that we are overlooking.

B Appendix B: Omitted proofs

B.1 Mechanisms for Trading.

Proof of Proposition 1. Equations (11) can be represented as

$$\bar{\pi} = p1 + \Pi q, \tag{52}$$

where $\bar{\pi}$ is a vector of intercepts, 1 is a vector of ones, and Π is a matrix with coefficients π^{ij} and zeros on the diagonal. Equation (52) has a solution for all vectors $\bar{\pi}$ if and only if the image of the set of all flow vectors (i.e. vectors whose entries

add up to 0) q under Π has dimension $N-1$ and does not contain vector 1. In this case, a unique pair (p,q) corresponds to each vector $\bar{\pi}$, and statements 1 and 2 follow immediately.

Let us show that if either the image of flow vectors under Π has dimension less than $N-1$, of if the image contains 1, then statements 1 and 2 are both false. If the dimension is less than $N-1$, then the pre-image of 0 has dimension of at least 1. Then 1 is false as $(0,q)$ is a solution for any q in the pre-image, and 2 is also false, as the span of the image together with vector 1 has dimension at most $N-1$.

If vector 1 is in the image of flow vectors under Π, e.g. $1=\Pi\tilde{q}$, then any pair $(p,-p\tilde{q})$ is a solution to $0=p1+\Pi q$, so statement 1 is false. Likewise, statement 2 is false since the span of the image (which does not change if we include vector 1) has dimension at most $N-1$. ■

Proof of Proposition 2. First, let us demonstrate equivalence of the first two statements. If the mechanism (P,Q) is not acceptable, i.e. there exists a nonzero vector X in the null space of Q such that $PX=0$, then $Q^PX=0$, so Q^P is not invertible. Conversely, if Q^P is not invertible, i.e. $Q^PX=0$ for $X\neq 0$, then $QX=0$ also since if all flows except for that of player 1 are zero, then the flow of player 1 must also be 0. Thus, X is in the null space of Q and $PX=0$, so (P,Q) is not acceptable. This shows that the first two statements are equivalent.

Second, let us demonstrate the equivalence of the last two statements. If Q^P is invertible, then $X=(Q^P)^{-1}(p,q^2,\dots q^N)$, where $p=PX$ and $q=QX$, so the allocation X can be inferred from the pair (p,q). If Q^P is not invertible, then the set of $(p,q^2,\dots q^N)$ that corresponds to all X has dimension of at most $N-1$. Since q^1 is a linear combination of $q^2,\dots q^N$, the set of (p,q) has the same dimension. Hence, we cannot infer X from (p,q). This shows that the last two statements are equivalent. ■

Proof of Theorem 1. Let us first identify the one-to-one map between acceptable profiles and mechanisms. Consider any acceptable stationary linear revealing profile of strategies of the conditional double auction. Then equation (52) has a unique solution. From market clearing, (52) can be written as

$$\underbrace{\begin{bmatrix} 1 & \pi^{12} & \dots & \pi^{1N} \\ 1 & -\pi^{21} & \dots & \pi^{2N}-\pi^{21} \\ 1 & \vdots & & \vdots \\ 1 & \pi^{N2}-\pi^{N1} & \dots & -\pi^{N1} \end{bmatrix}}_{\hat{\Pi}} \begin{bmatrix} p \\ q^2 \\ \vdots \\ q^N \end{bmatrix} = \begin{bmatrix} \bar{\pi}^1 \\ \bar{\pi}^2 \\ \vdots \\ \bar{\pi}^N \end{bmatrix}, \tag{53}$$

where matrix $\hat{\Pi}$ is obtained by placing 1 in the first column and by subtracting the first column of Π from each remaining column. The profile is acceptable if and only if $\hat{\Pi}$ is invertible and all $\hat{\pi}^i$ are nonzero. Hence, $Q^P=\hat{\Pi}^{-1}\operatorname{diag}(\hat{\pi}^i)$, i.e. the pair (P,Q) is uniquely defined from the acceptable profile $\{\hat{\pi}^i,\ \pi^{ij},\ i\neq j\}$. The pair (P,Q)

must also be acceptable. Indeed, if X is in the null space of Q and $PX = 0$, then $\hat{\Pi}^{-1}\,\mathrm{diag}(1/\hat{\pi}^i)\,X = 0$ implies that $X = 0$.

Let us show that the mechanism we just constructed has no stationary allocations $X \neq 0$ that have $X^i = 0$ for one of the players. Notice that $(Q^P)^{-1} = \mathrm{diag}(1/\hat{\pi}^i)\hat{\Pi}$ has no zeros in column 1. If so, then for any stationary allocation, $X \neq 0$, the corresponding price is $p \neq 0$, so $X = (Q^P)^{-1}(p, 0 \ldots 0)^T$, and all elements of X are nonzero since column 1 of $(Q^P)^{-1}$ has no zeros. The converse is also true: if $X \neq 0$ is a stationary allocation of an acceptable mechanism (P, Q), and if $X^i \neq 0$ for all i, then the first column of $(Q^P)^{-1}$ equals $(1/p)X$, and so it has no zeros.

Conversely, the pair (P, Q) is acceptable if and only if Q^P is invertible by Proposition 2. Since the corresponding profile must satisfy $\mathrm{diag}(\hat{\pi}^i)(Q^P)^{-1} = \hat{\Pi}$, and since the first column of $\hat{\Pi}$ must contain ones, we can obtain the corresponding profile as follows. We must set $\hat{\pi}^i$ to the inverse of the i-th entry in column 1 of $(Q^P)^{-1}$, and then divide each row i of $(Q^P)^{-1}$ by $\hat{\pi}^i$ to obtain $\hat{\Pi}$. Then the first column of $\hat{\Pi}$ contains ones, and we can infer the slopes π^{i1} from the diagonal of $\hat{\Pi}$, and infer the remaining ones from the remaining entries. In order for this procedure to work, it is necessary (and sufficient) for the first column of $(Q^P)^{-1}$ to have no zeros. We know that this is the case, because for any stationary allocation $X \neq 0$ of Q, all elements X^i are nonzero.

Now, let us prove strategic equivalence of the corresponding mechanism and profile. In a mechanism, given the allocation X^{-i} of other players, player i has the choice between flow and price pairs given by $(P\tilde{X}, Q\tilde{X} \mid \tilde{X} = X + 1^i y)$. In the corresponding profile $(\hat{\pi}^i, \pi^{ij},\ j \neq i)$, player i also has the same one-dimensional set of price-flow pairs, which he can attain by changing the intercept from $\bar{\pi}^i = \hat{\pi}^i X^i$ to $\bar{\pi}^i = \hat{\pi}^i(X^i + y)$. Let us prove that player i does not have any more choices than that (i.e. he cannot achieve any price-flow pair outside this set by also changing the slopes π^{ij} of his supply-demand function).

Any price-flow pair (p, q) that can be attained by a deviation of player i must satisfy the $N-1$ equations from (53), with equation i excluded. Since these are independent equations (when the matrix $\hat{\Pi}$ is invertible), they define a set of pairs (p, q) of dimensions 1. Hence, any price-flow pair that player i can generate, can be attained by changing only the intercept $\hat{\pi}^i$. We conclude that corresponding mechanism and profile must be strategically equivalent.[32] ■

B.2 Equilibrium Characterization.

Proof of Proposition 3. We have to prove that the truth-telling strategy maximizes the utility of any player i. For an arbitrary strategy $\{y_t,\, t \geq 0\}$, which specifies

[32] Of course, player i may be able to choose a supply-demand function that does not lead to any allocation at all.

the misrepresentation y_t of player i's allocation for any history $\{X_s,\ s \in [0,t]\}$ of allocations, consider the process

$$G_t = \int_0^t e^{-rs}\left((PX_s + p^i y_s)(\underline{Q}^i X_s + q^{ii} y_s) - \frac{b^i}{2}(X_s^i)^2\right)\, ds + e^{-rt} f^i(X_t).$$

Then the conditions $p^i < 0$ and $q^{ii} > 0$ ensure that $y_t = 0$ maximizes the drift of G_t, and (13) ensures that the maximal drift of G_t equals 0. That is, the process G_t is always a supermartingale, and a martingale under the truth-telling strategy.

Now, since the process X defined by (12) is nonexplosive, it follows that when all players follow the truthtelling strategies

$$E[e^{-rt} f^i(X_t)] \to 0$$

as $t \to \infty$. Therefore, player i's expected payoff under the truthtelling strategy is

$$E\left[\int_0^\infty e^{-rs}\left((PX_s)(\underline{Q}^i X_s) - \frac{b^i}{2}(X_s^i)^2\right)\, ds\right] = E[G_\infty] = G_0 = f^i(X_0).$$

Consider any alternative strategy $\{y_t,\ t \geq 0\}$ that satisfies the no-Ponzi condition $E[e^{-rt}X_t^2] \to 0$ as $t \to \infty$. Then for any quadratic value function $f^i(X)$, $E[e^{-rt}f^i(X_t)] \to 0$ as $t \to \infty$. It follows then that player i's payoff under this strategy is

$$E\left[\int_0^\infty e^{-rs}\left((PX_s + p^i y_s)(\underline{Q}^i X_s + q^{ii} y_s) - \frac{b^i}{2}(X_s^i)^2\right)\, ds\right] = E[G_\infty] \leq G_0 = f^i(X_0).$$

Thus, truth-telling is optimal. This completes the proof of Proposition 3. ■

Proof of Proposition 4. Individual fringe members do not have price impact, and an individual may choose to hold allocation x different from that of the rest of the fringe, X^N. If $x = X^N$, then individual continuation utility is $f^N(X) = X^T A^N X + k^N$ and in general the value function takes the form $X^T A^N X + (x - X^N)PX + k^N$, since trading $x - X^N$ generates the income of $(x - X^N)PX$.

Since there is no price impact, we treat dynamic choice of an individual fringe member simply as choosing x at all times. Then the HJB equation is written as

$$\max_x\ \frac{-b^F}{2}x^2 - r(X^T A^N X + (x - X^N)PX + k^N) + (PX)(\underline{Q}^N X)$$
$$-2X^T A^N Q X - (x - X^N)PQX + \text{tr}\,[\Sigma\Sigma^T A^N] = 0.$$

The first-order condition

$$-b^F x - P(rX + QX) = 0$$

must hold at $x = X^N$, so it follows that (17) must hold. Notice also that (k^N, A^N) must satisfy the same equations as before, (16). ■

B.3 Large Player and Fringe.

Proof of Proposition 5. With one large trader and the fringe, the mechanism is described by the four parameters:

$$Q = \begin{bmatrix} q_L & -q_F \\ -q_L & q_F \end{bmatrix}, P = [p_L, p_F]$$

First, fringe's optimality condition (17) is

$$rp_L + p_L q_L - p_F q_L = 0 \quad \text{and} \quad rp_F - p_L q_F + p_F q_F + b^F = 0. \tag{54}$$

Second, for the large player, equation (16) implies that

$$\begin{aligned} ra_{11} - 2a_{12}q_L + 2a_{11}q_L &= -\frac{1}{2}b^L + p_L q_L & (55)\\ ra_{12} - a_{11}q_F + a_{12}(q_L + q_F) - a_{22}q_L &= \frac{p_F q_L - p_L q_F}{2} & (56)\\ ra_{22} - 2a_{12}q_F + 2a_{22}q_F &= -p_F q_F & (57) \end{aligned}$$

and the first-order conditions (15) are

$$p_L q_L = q_L a_{11} - q_L a_{12} \quad \text{and} \quad p_F q_L - p_L q_F = 2a_{12}q_L - 2a_{22}q_L. \tag{58}$$

Using (58), we can transform (55), (56) and (57) to

$$\begin{aligned} ra_{11} &= -b^L/2 - p_L q_L \\ ra_{12} &= p_L q_F \\ ra_{22}q_L &= -p_L q_F^2 \end{aligned}$$

and plugging coefficients of large player's value function into (58), we get

$$rp_L q_L = -(b^L/2 + p_L q_L + p_L q_F)q_L \quad \text{and} \quad rp_F q_L - rp_L q_F = 2(q_L + q_F)p_L q_F.$$

Notice that (54) cannot hold with $p_L = p_F$, so

$$q_L = \frac{rp_L}{p_F - p_L} \quad \text{and} \quad q_F = \frac{rp_F + b^F}{p_L - p_F}.$$

Plugging these into the first-order conditions for the large player, we obtain (if $q_L \neq 0$)[33]

$$\frac{b^L}{2b^F} = \frac{p_L}{p_F - p_L} \quad \text{and} \quad 1/2 = \frac{p_F + b^F/r}{p_L - p_F}.$$

[33] If $q_L = 0$, then the corresponding solution is the degenerate equilibrium,

$$p_L = 0, \quad p_F = -\frac{b^F}{r}, \quad q_L = q_F = 0 \quad \text{and} \quad A^L = \frac{1}{r}\begin{bmatrix} -b^L & 0 \\ 0 & 0 \end{bmatrix}.$$

This leads to a system of linear equations in p_F and p_L, which have a unique solution

$$p_L = -\frac{b^L b^F}{r(3b^F + b^L)} \quad \text{and} \quad p_F = -\frac{(2b^F + b^L)b^F}{r(3b^F + b^L)}.$$

Then

$$q_F = \frac{r}{2}, \quad q_L = \frac{r}{2}\frac{b^L}{b^F}, \quad \text{and} \quad A^L = \frac{b^F}{2r\,(3b^F + b^L)} \begin{bmatrix} -3b^L & -b^L \\ -b^L & b^F \end{bmatrix}.$$

The welfare of the fringe can found from equation (16). We have

$$\begin{aligned} ra^F_{22} - 2a^F_{12}q_F + 2a^F_{22}q_F &= -\frac{1}{2}b^F + p_F q_F \\ ra^F_{12} - a^F_{22}q_F + a_{12}\,(q_L + q_F) - a_{11}q_L &= \frac{p_L q_F - p_F q_L}{2} \\ ra^F_{11} - 2a^F_{12}q_L + 2a^F_{11}q_L &= -p_L q_L, \end{aligned}$$

hence

$$A^F = \frac{1}{2}\frac{b^F}{r\,(3b^F + b^L)\,(2b^F + b^L)} \begin{bmatrix} (b^L)^2 & -b^L b^F \\ -b^L b^F & -\left((b^L)^2 + 5b^L b^F + 5(b^F)^2\right) \end{bmatrix}.$$

■

B.4 Welfare Results.

In order to prove Proposition 7, we need to be able to evaluate the welfare of a portion of the fringe when the fringe is heterogeneous. Consider a market with $N-1$ large players and the fringe, in which the joint welfare of the fringe is characterized by the N-by-N matrix A^N. Consider two parts of the fringe with risk capacities $\hat{\beta}_1/b^N$ and $\hat{\beta}_2/b^N$, respectively, with $\hat{\beta}_1 + \hat{\beta}_2 = 1$. Let us determine the value function matrices $\hat{A}^1$ and $\hat{A}^2$ for the two parts of the fringe (these are $(N+1) \times (N+1)$ matrices). How would an initial allocation $\hat{X} = (X^1, \ldots X^{N-1}, \hat{X}^1, \hat{X}^2)$ get traded? We know that part 1 of the fringe instantaneously sells

$$\hat{X}^1 - \hat{\beta}_1(\hat{X}^1 + \hat{X}^2) \tag{59}$$

at price PX, where $X = (X^1, \ldots X^{N-1}, \hat{X}^1 + \hat{X}^2)$. After the initial trade, conditional on the absence of future shocks, the two parts of the fringe get utilities $\hat{\beta}_1 X^T A^N X$ and $\hat{\beta}_2 X^T A^N X$, respectively. Thus, prior to the trade, the welfare of the first part of the fringe is given by

$$\hat{X}^T \hat{A}^1 \hat{X} = \hat{\beta}_1 X^T A^N X + (PX)((1-\hat{\beta}_1)\hat{X}^1 - \hat{\beta}_1 \hat{X}^2). \tag{60}$$

Proof of Proposition 7. Let us normalize market risk capacity to $1 = \beta_1 + \beta_2$, where $\beta_1 = 1/b^L$ and $\beta_2 = 1/b^F$ are risk capacities of the large player and the fringe, respectively. Then welfare of the large player is given by

$$A^L = \frac{1}{2r(2\beta_1 + 1)} \begin{bmatrix} -3 & -1 \\ -1 & \beta_1/\beta_2 \end{bmatrix}. \tag{61}$$

Let us compare this with the welfare of this group of players under perfect competition. The welfare of a fringe of risk capacity 1 is expressed by the matrix $A^1 = -1/(2r)$ and the price is given by $P = -1/r$, thus the welfare of the first part of the fringe (with risk capacity $\beta_1 = \beta_L$) by (60) is

$$-\beta_1 \frac{(\hat{X}^1 + \hat{X}^2)^2}{2r} - \frac{\hat{X}^1 + \hat{X}^2}{r}((1-\beta_1)\hat{X}^1 - \beta_1 \hat{X}^2),$$

i.e. it is expressed by the matrix

$$\hat{A}_1 = \frac{1}{2r} \begin{bmatrix} \beta_1 - 2 & \beta_1 - 1 \\ \beta_1 - 1 & \beta_1 \end{bmatrix}$$

Compare with (61). Given shocks of standard deviation (σ_1, σ_2) with correlation ρ, $2r(2\beta_1 + 1)$ times the difference in welfare between the fringe and the large player in these two games is

$$(2\beta_1 + 1)\left((\beta_1 - 2)\sigma_1^2 + 2(\beta_1 - 1)\sigma_1\sigma_2\rho + \beta_1\sigma_2^2\right) + 3\sigma_1^2 + 2\sigma_1\sigma_2\rho - \frac{\beta_1}{1-\beta_1}\sigma_2^2 =$$

$$(\beta_1 - 1)(2\beta_1 - 1)\sigma_1^2 + 2(2\beta_1 - 1)\beta_1\sigma_1\sigma_2\rho - \frac{\beta_1^2(2\beta_1 - 1)}{1-\beta_1}\sigma_2^2 =$$

$$\frac{2\beta_1 - 1}{1-\beta_1} \underbrace{\left(-(1-\beta_1)^2\sigma_1^2 + 2\beta_1(1-\beta_1)\sigma_1\sigma_2\rho - \beta_1^2\sigma_2^2\right)}_{\leq 0}$$

The quantity in parentheses is minus the variance of

$$(1-\beta_1)\sigma_1\, dZ_t^1 - \beta_1\sigma_2\, dZ_t^2,$$

which is always non-positive (zero only if $\rho = 1$ and $\sigma_1/\beta_1 = \sigma_2/(1-\beta_1)$, i.e. shocks are such that no trade is required to achieve first best).

Therefore, if $\beta_1 < 1/2$ then the welfare of the fringe before merger is greater than the utility of this segment after merger. ■

Proof of Proposition 8. Denote $q = \tilde{\kappa}/\kappa(r/2)$. Then fringe optimality implies that

$$P + b^F \underline{1}^F + PQ = 0 \quad \Rightarrow \quad [p_L,\ p_F] = -\frac{[qb^L,\ b^F + qb^L]}{1 + qb^L/b^F + q}. \tag{62}$$

Given this, the large player's value function can be found by solving (16). We find that the large player's value function is characterized by the matrix

$$A^L = \frac{1}{1+2q(1+b^L/b^F)} \left[\begin{array}{cc} -b^L/2 - qb^L + qp_L b^L/b^F & -q(b^L+p_L) \\ -q(b^L+p_L) & qb^F/b^L(b^L+p_L) \end{array} \right].$$

We would like to find the selling rate q that maximizes

$$[1 \ -1] A^L \left[\begin{array}{c} 1 \\ -1 \end{array} \right] = \frac{-b^L/2(1+qb^L/b^F+q) + q(b^L+b^F)}{(1+qb^L/b^F+q)(1+2q(1+b^L/b^F))},$$

an expression obtained by plugging in p_L from (62). The first-order condition is given by the quadratic equation

$$b^F + 2qb^L - q^2(b^L - (b^L)^2/b^F + 2b^F) = 0,$$

For the equilibrium value of $q = r/2$, the left-hand side is positive, $(2b^F + b^L)(b^F + b^L)/4 > 0$, so the large player would prefer to commit to a faster selling rate. If $2b^F > b^L$ (i.e. the large player is more than 1/3 of the market), then $b^L - (b^L)^2/b^F + 2b^F > 0$ so the left-hand side goes to minus infinity as $q \to \infty$. In this case, the optimal selling rate corresponds to a finite root on $[r/2, \infty)$. If $2b^F < b^L$ (i.e. the large player is less than 1/3 of the market), then the optimal selling rate is infinity. ■

Proof of Proposition 9. Consider a market in which aggregate shock is $\sigma\, dZ_t$ per unit of risk capacity, and the large player with $b^L = 1$ (without loss of generality) gets shock $\sigma_L\, dZ_t + \tilde{\sigma}\, d\tilde{Z}_t$. Then the fringe, with risk coefficient b^F, gets the shocks $((1+1/b^F)\sigma - \sigma_L)\, dZ_t - \tilde{\sigma}\, d\tilde{Z}_t$. Then

$$A^L = \frac{b^F}{2r(3b^F+1)} \left[\begin{array}{cc} -3 & -1 \\ -1 & b^F \end{array} \right],$$

so the large player's welfare is

$$\frac{b^F}{2r(3b^F+1)} \left(-3(\sigma_L^2 + \tilde{\sigma}^2) - 2(\sigma_L \left(\frac{b^F+1}{b^F}\sigma - \sigma_L \right) + \tilde{\sigma}^2) + b^F(\left(\frac{b^F+1}{b^F}\sigma - \sigma_L \right)^2 + \tilde{\sigma}^2) \right)$$

$$= \frac{(b^F)^2 - b^F}{2r(3b^F+1)} \left(\tilde{\sigma}^2 + (\sigma_L - \sigma)^2 \right) + \frac{\sigma^2 - 2\sigma\sigma_L}{2r}.$$

Differentiating $\frac{(b^F)^2 - b^F}{3b^F+1}$ with respect to b^F, we obtain

$$\frac{(2b^F-1)(3b^F+1) - 3((b^F)^2 - b^F)}{(3b^F+1)^2} = \frac{3(b^F)^2 + 2b^F - 1}{(3b^F+1)^2} = \frac{(3b^F-1)(b^F+1)}{(3b^F+1)^2},$$

which is less than 0 if and only if $b^F < 1/3$. Hence, as the size of the fringe increases (i.e. b^F falls), the welfare of the large player rises if $b^F < 1/3$, i.e. the large player is less than 1/4 of the market. ■

B.5 General Characterization.

Proof of Lemma 2. First, let us first justify that the coefficients of the matrix $\hat{A}^i = U^T A^i U$ are given by (23). Multiplying (16) by U^T on the left and U on the right, we obtain

$$r\hat{A}^i + \hat{A}^i K + K\hat{A}^i = \frac{(PU)^T \underline{U}^i K + (\underline{U}^i K)^T PU}{2} - \frac{b^i}{2}(\underline{U}^i)^T \underline{U}^i.$$

In position jk, the left-hand side has $(r + \kappa_k + \kappa_j)\hat{a}^i_{jk}$ and the right-hand side, $((PU^k)u^{ij}\kappa_j + (PU^j)u^{ik}\kappa_k - b^i u^{ij}u^{ik})/2$. Hence,

$$\hat{a}^i_{jk} = \frac{-b^i u^{ij}u^{ik} + (PU^j)u^{ik}\kappa_k + (PU^k)u^{ij}\kappa_j}{2(r + \kappa_k + \kappa_j)}.$$

To express the first-order conditions using $\hat{A}^i$, notice that

$$A^i Q^i = (U^{-1})^T \hat{A}^i \underbrace{\underline{U}^{-1}\underline{U}}_{I} K (U^{-1})^i \quad \Rightarrow \quad (A^i Q^i)^T U = ((U^{-1})^i)^T K \hat{A}^i.$$

Multiplying the first-order condition (15) by U on the right-hand side, we obtain

$$p^i \underline{U}^i K + \underbrace{(\underline{U}^i K (U^{-1})^i)}_{q^{ii}} PU = 2((U^{-1})^i)^T K \hat{A}^i.$$

Entry k of this vector equation is given by

$$\underbrace{p^i \kappa_k u^{ik}}_{u^{ik}\kappa_k PU\frac{r+\kappa_k+K}{r+\kappa_k+K}(U^{-1})^i} + \underbrace{(\underline{U}^i K (U^{-1})^i)}_{q^{ii},\text{ or }(\underline{U}^i \frac{K(r+\kappa_k+K)}{r+\kappa_k+K}(U^{-1})^i)} PU^k =$$

$$-u^{ik}b^i \underline{U}^i \frac{K}{r+\kappa_k+K}(U^{-1})^i + u^{ik}\kappa_k PU \frac{K}{r+\kappa_k+K}(U^{-1})^i + (PU^k)\underline{U}^i \frac{K^2}{r+\kappa_k+K}(U^{-1})^i,$$

where $\frac{K}{r+\kappa_k+K}$ is the matrix that has entries $\frac{\kappa_j}{r+\kappa_k+\kappa_j}$, $j = 1 \ldots N$, on the diagonal. Performing some cancellations, we obtain

$$u^{ik}\kappa_k PU \frac{r+\kappa_k}{r+\kappa_k+K}(U^{-1})^i + u^{ik}b^i \left(\underline{U}^i \frac{K}{r+\kappa_k+K}(U^{-1})^i\right) =$$

$$-(r+\kappa_k)\left(\underline{U}^i \frac{K}{r+\kappa_k+K}(U^{-1})^i\right)(PU^k)$$

or

$$u^{ik}\left(b^i + \kappa_k(r+\kappa_k)\frac{PU\frac{1}{r+\kappa_k+K}(U^{-1})^i}{\underline{U}^i \frac{K}{r+\kappa_k+K}(U^{-1})^i}\right) = -(r+\kappa_k)PU^k$$

■

Proof of Proposition 12. From symmetry, equilibrium pair (P, Q) is characterized by two parameters p and κ, and takes the form

$$P = p\underline{1}, \quad Q = \kappa(I - S/N).$$

Then, from symmetry player 1's value function is characterized by the matrix

$$A^1 = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{12} \\ a_{12} & a_{22} & \dots & a_{22} \\ \vdots & \vdots & & \vdots \\ a_{12} & a_{22} & \dots & a_{22} \end{bmatrix}$$

From (16), the coefficients of matrix A^1 satisfy the equations

$$\begin{aligned} ra_{11} + 2\kappa\frac{N-1}{N}a_{11} - 2\kappa\frac{N-1}{N}a_{12} &= -\frac{b}{2} + p\kappa\frac{N-1}{N} \\ ra_{12} + \kappa a_{12} - \kappa\frac{N-1}{N}a_{22} - \frac{\kappa}{N}a_{11} &= p\kappa\frac{N-2}{2N} \\ ra_{22} + 2\frac{\kappa}{N}a_{22} - 2\frac{\kappa}{N}a_{12} &= -p\frac{\kappa}{N}. \end{aligned}$$

From (15), the first-order conditions are

$$\begin{aligned} 2\kappa\frac{N-1}{N}a_{11} - 2\kappa\frac{N-1}{N}a_{12} &= 2p\kappa\frac{N-1}{N} \\ 2\kappa\frac{N-1}{N}a_{12} - 2\kappa\frac{N-1}{N}a_{22} &= p\kappa\frac{N-2}{N}. \end{aligned}$$

Using the first-order conditions to eliminate the differences $a_{11} - a_{12}$ and $a_{12} - a_{22}$ from the value function equations, we obtain

$$\begin{aligned} ra_{11} &= -\frac{b}{2} - p\kappa\frac{N-1}{N} \\ ra_{12} &= \frac{p\kappa}{N} \\ ra_{22} &= -\frac{p\kappa}{N(N-1)}. \end{aligned}$$

After substitutions, the first-order conditions become

$$\kappa\left(-\frac{b}{2} - p\kappa\right) = rp\kappa \quad \text{and} \quad p\kappa^2 = rp\kappa\frac{N-2}{2}.$$

If $\kappa \neq 0$, then we obtain[34]

$$p = -\frac{b}{Nr} \quad \text{and} \quad \kappa = r\frac{N-2}{2}.$$

[34]If $\kappa = 0$, then the solution is degenerate with $a_{11} = -b/2$, $a_{12} = a_{22} = 0$, and p can be any number.

Notice that for any vector X whose coefficients add up to 0, $SX = 0$, so $QX = \kappa X$, and hence any misallocation gets traded to efficiency exponentially at rate κ. This completes the proof. ■

Bibliography.

Almgren, R. and N. Chriss (2001) "Optimal Execution of Portfolio Transactions," *Journal of Risk.* 3(2), 5–39.

Almgren, R. and T. M. Li, (2016) "Option hedging with smooth market impact," forthcoming in *Market Microstructure and Liquidity.*

Almgren, R., C. Thum, E. Hauptmann and H. Li (2005) "Equity Market Impact," *Risk* July, 57–62.

Back, K. (1992) "Insider Trading in Continuous Time," *Review of Financial Studies*, 5, 387–409.

Back, K., C. H. Cao, and G. A. Willard (2000) "Imperfect Competition among Informed Traders," *Journal of Finance*, 55, 2117–2155.

Bank, P., M. Soner, M. Voß (2016) "Hedging with Temporary Price Impact," Mathematical Finance.

Bouchaud, J.-P. (2010). "Price Impact." Encyclopedia of Quantitative Finance. DOI: 10.1002/9780470061602.eqf18006

Budish, E., P. Cramton and J. Shim, (2015) "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics* 130(4), 1547–1621.

Caldentey, R. and E. Stacchetti (2010) "Insider Trading with a Random Deadline," *Econometrica*, 78(1), 245–283.

Du, Songzi and Haoxiang Zhu (2013) "Dynamic Ex Post Equilibrium, Welfare and Optimal Trading Frequency in Double Auctions," working paper, MIT.

Farmer, J. D., A. Gerig, F. Lillo, and S. Mike (2006) "Market Efficiency and the Long-memory of Supply and Demand: is Price Impact Variable and Permanent or Fixed and Temporary?" *Quantitative Finance* 6, 107–112.

Gatheral, J. (2010) "No-Dynamic-Arbitrage and Market Impact." *Quantitative Finance* 10(7), 749–759

Gatheral, J., A. Schied and A. Slynko (2011) "Transient linear price impact and Fredholm integral equations," *Mathematical Finance*, 22(3), 445–474.

Glosten, L. R. and P. R. Milgrom (1985) "Bid, Ask and Transaction Prices in a Specialist Market with Heterogenously Informed Traders," *Journal of Financial Economics*, 14, 71–100.

Huberman, Gur, and Werner Stanzl (2004) "Price Manipulation and Quasi-arbitrage," *Econometrica* 72, 1247–1275.

Kyle, A. S. (1985) "Continuous Auctions and Insider Trading," *Econometrica*, 15, 1315–1335.

Kyle, A. S. (1989) "Informed Speculation with Imperfect Competition," *Review of Economic Studies*, 56, 317-355.

Kyle, A. S., A. A. Obizhaeva, and Y. Wang (2014) "Smooth Trading with Overconfidence and Market Power," working paper.

Lambert, N., M. Ostrovsky and M. Panov (2016) "Strategic Trading in Informationally Complex Environments," working paper, Stanford GSB.

Leland, Hayne and Pyle, David H, (1977), Informational Asymmetries, Financial Structure, and Financial Intermediation, Journal of Finance, 32, issue 2, p. 371–87.

Lewis, M. (2014) *Flash Boys.*

Malamud, S. and M. Rostek (2014) "Decentralized Exchange," working paper.

Moro, E., Vicente, J., Moyano, L. G., Gerig, A., Farmer, J. D., Vaglica, G., Lillo, F., and Mantegna, R. N. (2009) "Market Impact and Trading Profile of Hidden Orders in Stock Markets," Physical Review E, 80(6); DOI: 10.1103/PhysRevE.80.066102.

Myers, Stewart C. and Majluf, Nicholas S., (1984), Corporate financing and investment decisions when firms have information that investors do not have, Journal of Financial Economics, 13, issue 2, p. 187–221.

Patterson, (2012) *Dark Pools.*

Vayanos, D. (1999) "Strategic Trading and Welfare in a Dynamic Market," *Review of Economic Studies,* 66 (2), 219–254.