(preliminary and incomplete)

Middlemen Interaction and Its Effect on Market Quality [1]

Albert J. Menkveld and Bart Zhou Yueshen

(slides: http://goo.gl/KD35Z; printer friendly: http://goo.gl/gAQHb)

first version: October 22, 2012

current version: November 18, 2012

[1] Albert J. Menkveld, VU University Amsterdam, Tinbergen Institute, and Duisenberg School of Finance; address: FEWEB, De Boelelaan 1105, 1081 HV, Amsterdam, Netherlands; tel: +31 20 598 6130; fax: +31 20 598 6020; albertjmenkveld@gmail.com. Bart Zhou Yueshen, VU University Amsterdam, Tinbergen Institute, and Duisenberg School of Finance; address: FEWEB, De Boelelaan 1105, 1081 HV, Amsterdam, Netherlands; tel: +31 20 598 2895; fax: +31 20 598 6020; yueshenbartzhou@gmail.com. We thank Giovanni Cespa, Carole Comerton-Forde, Sarah Draus, Thierry Foucault, Terrence Hendershott, Bengt Holmström, Albert S. Kyle, and seminar or conference participants at U of Piraeus, the 5th Erasmus Liquidity Conference, U of Luxembourg, Skye Workshop 2012 (U of Aberdeen, U of Bath and Swansea U), Cass Business School, Warwick U, Finance Down Under 2012 (U Melbourne), and Erasmus U for very helpful comments and suggestions. Menkveld gratefully acknowledges VU University Amsterdam for a VU talent grant and NWO for a VIDI grant. Yueshen thanks his MPhil thesis committee members for their helpful comments.

(preliminary and incomplete)

Middlemen Interaction and Its Effect on Market Quality

Abstract

This paper studies a securities market in which high-frequency traders serve as middlemen. Two frictions, inattention and information asymmetry hinder efficient asset reallocation from a low-valuation seller to high-valuation buyers. Middlemen help the uninformed seller find additional buyers (improving welfare). They, however, impair her ability to learn from market activity (reducing welfare). That is, when middlemen trade with one another to find a reselling opportunity, price pressure arises and the seller cannot distinguish it from a fundamental value drop. Overall, only when reselling opportunities are large enough can sufficiently many middlemen improve welfare. The analysis speaks to recent disruptions in electronic markets.

JEL code: G10

keywords: high-frequency trader, electronic market, middleman

1 Introduction

Over the past two decades, securities markets around the world have gradually moved to electronic trading. Market participants responded by automating their trading strategies. High-frequency traders (HFTs) were the most visible group, as they reportedly participated in more than half of U.S. equity trades (SEC, 2010). The SEC characterizes HFTs as professional traders who 1) use extraordinarily high-speed computer programs to generate, route, and execute orders and 2) maintain very short time frames for establishing and liquidating positions (SEC, 2010, p.45). One interpretation is that HFTs are the new middlemen, a perspective maintained throughout this paper.[1]

The role of middlemen in a securities markets has been extensively studied.[2] Most of the literature focuses on the behavior of a representative middleman. Relatively little is known about how middlemen interact with one another and, more importantly, how such interaction feeds back on other participants' trading. This paper develops a theoretical model to study the interaction in modern electronic markets and judges market quality through a welfare criterion.

The engine of the model is trading among middlemen motivated by 1) heterogeneity in reselling opportunities (ROs) and 2) cost of carry. ROs can be defined as the likelihood for a middleman to find an end user (referred to as a fundamental investor), to whom to resell. The term is deliberately broad in order to capture a straightforward resell to not only a buyer in the same market, but also buyers in other markets.[3]

There is heterogeneity in ROs across middlemen because technology choice (both hardware and software) enables them to differentiate and develop their own niche. They are all looking for counterparties. Depending on market conditions some will be more likely than others to find ROs. These new middlemen are subject to substantial cost of carry, since funding liquidity, which is required to maintain margins, is reportedly low at HFT firms (see Easley, López de Prado, and O'Hara, 2012).

To reduce the expected cost of carry, a low-RO middleman resells the asset to a high-RO middleman. Such *inter-middlemen* trades feed back negatively on the quality of trading at large because uninformed investors' ability to learn about fundamental value is impaired. The reason is that the receiving end of such

[1] In the taxonomy of HFT proposed by the SEC, this perspective fits their categories *a* and *b* (market making and arbitrage). The remaining categories *c* and *d* that involve exploitation of structural vulnerabilities, order anticipation strategies, and momentum ignition strategies are beyond the scope of this paper.

[2] See, e.g., Ho and Stoll (1983), Glosten and Milgrom (1985), Easley and O'Hara (1987), Grossman and Miller (1988), Biais (1993), and more recently Dunne, Hau, and Moore (2012).

[3] For example, an equity index ETF can be replicated through a basket of index component stocks or futures can be replicated by trading related options. Menkveld (2011) documents how an HFT actively resells across two equity markets.

a trade–the buying middleman–requires a price discount (price pressure) to compensate for its[4] expected cost of carry (in case its RO fails to materialize). However, regular trades involving investors take off at fundamental value, (unpressured prices). Uninformed investors who observe real-time trade history (through Bloomberg, Reuters, etc.) cannot distinguish between these two types of trade. This hampers their learning about the fundamental value.

The effect of inter-middlemen trades on market quality is examined in the framework developed in Lagos, Rocheteau, and Weill (2011). Two well-known frictions affect the trading process: information asymmetry (as alluded above) and inattention (Duffie, 2010) since investors only infrequently visit markets (Grossman and Miller, 1988). In the model, one investor (the seller) needs to sell a large position, but the two frictions are in the way of an efficient reallocation. She is uninformed relative to others (this is extreme and could be relaxed; see section 6.3), and buyers might not be in the market to which she is connected (it implicitly assumes that connecting to all potential markets is prohibitively costly for investors[5]). In her optimization, she uses middlemen 1) to find buyers beyond the market she connects to and 2) to learn about the fundamental value to minimize adverse selection cost.

Results are developed in terms of a cost-benefit analysis of adding middlemen. The trading process benefits as each additional middleman brings another reselling opportunity to the economy. However, this comes at the cost of a more severe inference problem for the seller, because inter-middlemen trades become more likely. This cost makes the seller *overload* the middleman sector. If the seller reduces her supply a bit, welfare improves. Results show that when each middleman has sufficiently large RO, it is optimal to have an infinite number of middlemen. However, when ROs are small, it is optimal to have only one middleman. As the inference problem hampers trading, one natural policy proposal is to flag each trade as to middlemen involvement. Such a disclosure policy improves welfare.

The model sheds new light on recent disruptions in electronic markets. For example, on May 6, 2010, U.S. equity indices declined by 5 to 6 percent, and then recovered, all in 20 minutes: this event is known as the Flash Crash. Having investigating the event, CFTC and SEC (2010) observe that (p.1-3) against a "backdrop of unusually high volatility and thinning liquidity, a large fundamental trader (a mutual fund complex) initiated a sell program to sell a total of 75,000 E- Mini contracts (valued at approximately $4.1

[4] This paper refers to a middleman as "it" (because, in the context of the model, such middlemen are best thought of as HFT machines). The seller is "she" (see below), and a buyer is "he".

[5] This could be thought of as the outcome of a process whereby middlemen specialize and pay the large fixed cost of connection and investors use middlemen to benefit from their machines.

billion) as a hedge to an existing equity position.” In their view, this triggered a “liquidity crisis” in which, during the one minute of the extreme price drop of 4%, “the same positions were rapidly passed back and forth” by the HFTs.[6] Then, “the large trader responded to the increased volume by increasing the rate at which it was feeding the orders into the market, even though orders that it already sent to the market were arguably not yet fully absorbed by fundamental buyers or cross-market arbitrageurs. In fact, especially in times of significant volatility, high trading volume is not necessarily a reliable indicator of market liquidity.” (p.3, CFTC and SEC, 2010).[7] The model proposed in this paper could be read as a rationalization of this sequence of events. The Flash Crash does not seem to have been a singular event. It has since been followed by disruptions in other markets.[8]

Two new and distinctive features of the proposed model are worth emphasizing. First, investors learn from market activity of middlemen (HFTs) who, contrary to classic market making models (e.g., Glosten and Milgrom, 1985), have an informational advantage relative to (at least some) investors, at least in the short-run (seconds), as their machines are able to quickly parse “hard information”.[9] Second, the inter-middlemen trades are driven by heterogeneity in reselling opportunities, as opposed to the risk-sharing emphasized in the “hot potato” literature (see, e.g. Lyons (1997), Naik, Neuberger, and Viswanathan (1999), and Viswanathan and Wang (2004)). The most notable differences are: RO-driven trades might aggravate price pressure, whereas risk-sharing trades reduce it. Such trades are much harder to explain under pure risk-sharing (unlike risk aversion, ROs are expected to change very frequently).

In addition to the aforementioned HFT literature, the paper also fits into a broader category of algorithmic or automated trading. Foucault and Menkveld (2008) study smart routers that investors use to benefit from liquidity supply in multiple markets. Hendershott, Jones, and Menkveld (2011) show that algorithmic trading (AT) causally improves liquidity and makes price quotes more informative. Chaboud et al. (2009) relate AT to volatility and document a rather weak relationship. Hendershott and Riordan (2010) and find that both AT demanding liquidity and AT supplying liquidity make prices more efficient. Hasbrouck and Saar (2010) study low-latency trading or “market activity in the millisecond environment” in NASDAQ’s

[6] This observation is corroborated in Kirilenko et al. (2011) who study disaggregated data on the E-mini market.

[7] Menkveld and Yueshen (2012) document that *right after* the one minute episode of heavy HFT activity and the 4 percent price drop, the large trader sold heavily in the E-mini market. Coincidentally, price recovery was relatively slow in this market. The seller’s response could be evidence of the inference problem emphasized by the model.

[8] Nanex keeps a list of this type of extreme market events: http://www.nanex.net/flashcrash/OngoingResearch.html.

[9] Empirically, Jovanovic and Menkveld (2011) study HFT trades and evidence in support of a short-run informational advantage. Brogaard (2010) and Hendershott and Riordan (2011) document that HFTs contribute substantially to price discovery. Theoretically, Biais, Foucault, and Moinas (2011) and Jovanovic and Menkveld (2011) also model HFTs as having an informational edge.

electronic limit order book in 2007 and 2008 and find that increased low-latency trading is associated with improved market quality.

Finally, the paper adds to the optimal execution literature through its emphasis on how the seller optimization is affected by the presence of middlemen and their interaction. Keim and Madhavan (1995) and Chan and Lakonishok (1995) document that institutional orders that are broken up. Bertsimas and Lo (1998) model the optimization in the presence of transitory price impacts and a completion deadline. Almgren and Chriss (2001) extend the optimization problem by considering risk. Obizhaeva and Wang (2005) optimize in an environment where liquidity does not replenish instantly. Gârleanu and Pedersen (2012) study optimal execution in the presence of short-lived predictability of returns.

The rest of the paper is structured as follows. A motivating example in section 2 illustrates the intuition for inter-middlemen trades and the associated inference problem. Section 3 sets up the full model. Agents' equilibrium strategies are set up and solved in section 4. Model implications are derived and discussed in section 5. Section 6 discusses the relaxation of several assumptions as well as some extensions of the model. Section 7 concludes. The appendix contains a notation summary (section 7) and proofs (section 7).

2 A motivating example

This section illustrates the engine of our results with a simple example. It shows how inter-middlemen trades hamper investor learning. Note that this example is only a very specific snapshot (period one) of the more general full model to be developed in section 3. Throughout the paper, random variables are in upper case, whereas realizations and deterministic variables are in lower case.

Suppose that at time t, there are two risk-neutral middlemen in the market, **M**1 and **M**2. **M**1 holds one unit of the asset. **M**2 has none of it. In the next instant of time dt, an investor might arrive in the market. This paper will denote the event of investor arrival in this market by random variable E_0, with $E_0 = 1$ for arrival and $E_0 = 0$ for no arrival. At the same time, an investor might arrive privately to **M**2, rather than in the market. This event will be denoted by E_2. These potential investor arrivals constitute the reselling opportunities (ROs) for the two middlemen.

The asset pays off Z (> 0) to investors, but nothing to middlemen. Therefore, there is an opportunity cost to **M**1 and **M**2 if the asset cannot be resold to investors. Both middlemen and the investors (if arriving), observe the realization z. In addition, at t, the two middlemen observe the likelihood of investor arrival.

They observe θ_i which is the realization of $\Theta_i := \mathbb{P}(E_i = 1)$ for $i \in \{0, 2\}$. In particular, fix $\Theta_2 = \theta_2 \in (0, 1)$ and let Θ_0 be Bernoulli distributed with success probability of one half.

Case 1. *(MB trade, $\theta_0 = 1$.)* In this case, **M**1 knows that an investor arrives in dt as $\theta_0 = 1$. It will then post a take-it-or-leave-it offer (a limit order) at price $p = z$ in the market, which maximizes its profit from reselling. In dt, the investor arrives, observes **M**1's offer, and takes it. The paper will henceforth refer to this trade as a "middleman-buyer trade" (MB trade), as the asset is transferred to a buyer from a middleman. The market records a trade at price $p = z$.

Case 2. *(MM trade, $\theta_0 = 0$.)* In this case, **M**1 cannot resell to an investor in the market as the arrival probability in this market is zero. Instead, **M**1 aims to resell the asset to **M**2 and posts a take-it-or-leave-it offer at **M**2's reservation price, which is **M**2's expected revenue of reselling the asset to its private investor: $\theta_2 z$. **M**2 takes the offer and a so-called "inter-middlemen trade" (MM trade) realizes. The market records a trade at **M**2's reservation price $p = \theta_2 z$.

This leads to two important results (part of proposition 1.)

Part of proposition 1. (Inter-middlemen trades.) *Inter-middlemen trades 1) reallocate the asset efficiently among middlemen; and 2) are accompanied by price pressure.*

The intuition for this result is that a transfer of the asset from a low-RO middleman to a high-RO middlemen benefits both. Price pressure arises as reselling is not a certainty for the buying middlemen who requests a compensation of $(1 - \theta_2)z$.

Inference problem. The two types of trade impair learning for an outside investor, who is uninformed about Z and does not observe Θ_0. She cannot back out the fundamental value Z by observing only trade price p, not its type. Instead, she computes its posterior distribution which is $Z = p$ with probability of one half (MB trade) and $Z = p/\theta_2$ with probability of one half (MM trade).

This simple construction is expanded into a full-fledged model in order to analyze how MM trades affect the seller's inference problem and, as a result, her trading strategy. This paper also studies the effect on market quality through a welfare assessment.

3 Primitives

The economic environment is inspired by Lagos and Rocheteau (2007, 2009). See also Duffie, Gârleanu, and Pedersen (2005, 2007), Lagos, Rocheteau, and Weill (2011), Pagnotta and Philippon (2011), and others. A notation summary is kept in section 7. Upper case letters indicate random variables, while their realizations and other deterministic variables are in lower case.

The following part of the economic environment is consistent with Lagos and Rocheteau (2007, 2009).

Goods and assets. There is one general consumption good, defined as numéraire, and one special good ("fruit") for consumption. There is one perfectly divisible asset ("tree"). Each unit of the asset produces one unit of the special good at the end of the time. There is one market, in which only the asset is traded, but not the special good. This paper denotes the amount of the numéraire good as c and the special good as a (which coincides with the amount of the asset).

Agents. There are infinitely many investors, all with preference: $u^{\mathbf{I}}(c, a) = c+Z{\cdot}a$. Z, uniformly distributed on $[0, 1]$, is the random preference shock that the entire investor population is subject to. Z will be referred to as the fundamental value of the asset later in this paper. There are $m \geq 1$ homogeneous middlemen, labeled as $\mathbf{M}1, \ldots, \mathbf{M}m$, who only consume the numéraire good and derive utility from it, but not from the special good: $u^{\mathbf{M}}(c, a) = c(= c + 0 \cdot a)$. The number m is exogenous and known to all.

Incentive to trade. One out of the infinitely many investors suffers a (negative) preference shock, and becomes the seller, $\mathbf{S}$, with shocked preference $u^{\mathbf{S}}(c, a) = c + Z \cdot [a - k(a)]$, where $k(\cdot)$ is $\mathbf{S}$'s cost of the carry of the asset and reduces her (marginal) utility for the special good. For tractability, set $k(a) = a^2/2$ if $a \geq 0$ and $k(a) = 0$ if $a < 0$.[10]

As all other investors are equal and have higher (marginal) utility for the asset than $\mathbf{S}$, they will henceforth be referred to as buyers. This paper focuses on a representative buyer, $\mathbf{B}$.

[10] More generally, one can assume $k(\cdot)$ to be positive, twice differentiable, strictly increasing, and strictly convex. To prevent $\mathbf{S}$ from unlimited selling, it is necessary to impose $k(a) = 0$ for all $a \leq a^*$ for some $a^* > -\infty$ and to let her initial asset position be $a_0 > a^*$. Without loss of generality, one can choose $a^* = 0$.

The environment deviates from Lagos and Rocheteau (2007, 2009) in terms of market structure: Instead of search and bargaining, a stylized limit order market is introduced as described below.[11] Importantly, reselling opportunities are introduced to middlemen.

Reselling opportunities for M. There is a common reselling opportunity (RO), Θ_0, in the market. This common RO is available to all **M**: An early buyer, **EB**, might arrive in the market with probability θ_0 (the realization of Θ_0). **EB** has the same preference as **B**, but simply arrives early (see the paragraph on "time" below). Write the arrival of this **EB** as a random variable E_0 Bernoulli distributed with $\mathbb{P}(E_0 = 1|\theta_0) = \theta_0$. The reason that **EB** arrives early is exogenous.[12]

In addition, each **M***i* has a private RO, Θ_i (for $i \in \{1, \ldots, m\}$). In other words, for each **M***i*, an **EB** *privately* arrives to **M***i* with probability θ_i (the realization of Θ_i). Similarly, denote the arrival of such private **EB** by random variable E_i, which is Bernoulli distributed with $\mathbb{P}(E_i = 1|\theta_i) = \theta_i$. When reselling to its private **EB**, **M***i* privately posts *privately* a supply schedule, and then **EB** chooses how to trade against the supply.[13]

Time and trading procedure. There is no time value. The trading procedure is as follows:

- Period 0: **S** is shocked and then trades with **M**.
  - The preference shock strikes **S**.
  - **S** posts a supply schedule in the market and then leaves.
  - All **M** rush for S's supply, but only one,[14] labeled **M**1, gets it.

[11] Apart from the market structure, two other differences from Lagos and Rocheteau (2007) (and others) are worth mentioning. First, this paper does not assume that all investors form a non-atomic continuum. (See the paragraph on agents above.) This is because, as will become clear soon, one of the main frictions in the model is that not all investors are present in the market at all times (in the spirit of Grossman and Miller, 1988). Second, Lagos and Rocheteau (2007, 2009) model the preference of all agents to be strictly increasing and strictly concave (in the consumption of the special good). This paper's model differs as all agents have linear preference except for **S**. Therefore it is in the tradition of Duffie, Gârleanu, and Pedersen (2005). Linearity in preference is for mathematical convenience, but does not affect the main results of the model (as is discussed in section 6.3).

[12] **EB** might be driven by some signals about the asset value. With this interpretation, such signals are assumed to be the same as those received by **M** (see the time line below) so that information is symmetric between **M** and **EB** (see more discussion on the information structure in section 6.3). A natural distinction between **EB** and **M**i s that middlemen, HFT computers, have much lower latency than investors to access the market.

[13] More generally, one can assume a revenue function $g(a; Z, \Theta_i)$ for **M***i*'s private RO, where a is the amount of the asset (re)sold and Θ_i serves as a random parameter. The specification described here is mathematically equivalent to the (conditionally) constant-return-to-scale form of $g(a; Z, \Theta_i) = \Theta_i Z a$, as will become clear later in section 4.2.1. The functional form of this revenue function can be extended to incorporate market power, information, and other aspects that may affect **M**'s reselling opportunity.

[14] When multiple **M** rush for a single supply schedule in the context of a limit order market honoring time priority, the winning **M** can be, for example, the fastes one. Section 6.3 discusses the consequence of allowing multiple **M** to buy the asset from **S**.

- Period 1: **M** trade with each other and/or with **EB**.
  - Each **M**i observes the realizations of Z, of Θ_0, and of its own Θ_i.
  - **M**1 tries to resell by posting a supply schedule in the market or via its own private RO.
  - All other **M** observe the supply and decide whether to rush for it. If multiple **M** do so, only one gets it.
  - Nature picks the realization of E_i for all $i \in \{0, 1 \ldots, m\}$. If $E_0 = 1$, an **EB** observes Z, arrives in the market, and trades against the residual supply (if there isther e is any). For $i \in \{1, \ldots, m\}$, if $E_i = 1$, an **EB** observes Z, arrives privately to **M**i, and then trades with it.
  - All **EB** and all **M** leave the market.
- Period 2: **S** trades with **B**.
  - S returns to the market and observes the market activity–price quotes, trade prices, and trading volume (if any)–of period 1 (but the private reselling activity of **M**i remains opaque).
  - **S** posts a supply schedule and then leaves.
  - **B** observes Z, arrives, potentially trades against **S**'s supply, and leaves.
- Period 3: The asset pays off and all agents consume.

Two key frictions are embedded in the above time line. The first is investment inattention (Duffie, 2010). At the exact moment (period 0) that the shock strikes **S**, there are no buyers in the market. Only middlemen **M** can provide immediacy (see Grossman and Miller, 1988), because these computers have the technology to continuously monitor the market. **S** does not have such technology. Therefore, she leaves after period 0 and only returns at a later time. **M**'s low latency to access the market provides the opportunity to intermediate between **S** and **EB**.

The second friction is that **S**, uninformed about Z, faces adverse selection by other investors. Because **S** must explicitly offer her supply to indicate her willingness to sell, she is subject to the adverse selection by late buyers, who arrive after observing z. All **M** are also modeled to observe z because in the context of this paper, these **M** are best thought of as powerful computers that have superior information processing technology. Similar assumptions are made in, for example, Biais, Foucault, and Moinas (2011) and Jovanovic and Menkveld (2011).[15]

In a very short term (period 1), some buyers, such as **EB**, might arrive either in the market or privately to **M**. **M** will thus try to resell their position bought from **S**, who, in turn, learns from the market activity of this period about the asset value. To close the economy, period 2 is modeled as a reduced form of a long term where sufficiently many buyers[16], modeled as a representative **B**, arrive in the market and adversely select **S**.

Assumptions. The following two assumptions are made for tractability.

Assumption 1. In period 0, **S**'s posted supply schedule expires at the end of the period for exogenous reasons.[17]

Assumption 2. Θ_0 follows a Bernoulli distribution with success probability of $1/2$. For $i \in \{1, ..., m\}$, $\Theta_i = \phi X_i$, where $0 < \phi < 1$ and X_i is also Bernoulli distributed with success probability of $1/2$. The random variables Θ_i and Z are mutually independent.

Assumption 1 restricts the strategy space of **S** in period 0. Assumption 2 gives a specific form of reselling opportunities. Section 6 discusses the relaxation of these assumptions and argues that the key results and economic intuitions of the paper are not driven by the simplification. Finally, to avoid triviality and to focus on off-corner solutions, the following assumption is needed for **S**'s initial position:

Assumption 3. a_0 is large (in the sense of equations A1 and A4).

S's initial position reflects her willingness to sell immediately after the preference shock: The larger is the position, the larger is the opportunity cost to carry the asset to the final date. As will become clear in the analysis of **S**'s optimal strategy (proof of lemma 5), a trivial solution is obtained if **S**'s initial position is too small. Intuitively, the opportunity cost for **S** to hold the asset becomes negligible if her initial position is too small (in the extreme, zero). To avoid triviality and corner solutions, the above assumption is necessary.

[15] Jovanovic and Menkveld (2011) model the asset value as two parts: a soft information part that can be understood only by humans and a hard information part that only machines can process (see also Petersen, 2004). The current specification that **S** has no private information about Z while **M** observe the true realization can be thought of as an extreme case of their specification, where the soft information component is zero. A discussion on the information structure is provided in section 6.3.

[16] In the presence of infinitely many buyers, **S** no longer has the immediacy problem as she faces in period 0. It is assumed that all **M** leave at the end of period 1. Section 6.3 discusses the consequence of allowing **M** to persist and to participate in the period 2 trades.

[17] A similar assumption is made in Foucault (1999) to simplify the analysis.

4 Equilibrium

The equilibrium is characterized by agents' optimal strategies and the supply and demand schedules in different time periods. These strategies are solved for **B**, **M**, and **S** in sections 4.1, 4.2, 4.3 respectively.

4.1 Buyers (B and EB)

With realization z, **B** arrives at the end of period 2 and trades against the supply schedule posted by **S**. The marginal benefit of buying an additional unit of the asset is $\partial u^B(c,a)/\partial a = z$. The marginal cost is the price, p, paid for that marginal unit. Optimality requires $p = z$. (The same analysis applies to an **EB** who also observes z but arrives early in period 1 and trades against the supply from **M**.) Therefore, given a non-decreasing supply schedule $p = p(q)$ where p is the price of the marginal unit of the asset supplied and q is the cumulative supply amount at price p, **B** (and **EB**) buys all q (≥ 0) units of the asset until $p(q) = z$.[18]

4.2 Middlemen (M)

M's strategies are solved backwardly: In period 1, **M** try to resell the asset, and they trade with **S** in period 0.

4.2.1 Reselling in period 1

In period 1, there are two types of **M**: asset owner **M**1 and $m-1$ non-owners. They all observe z (the asset value) and θ_0 (the common reselling opportunity, RO). In addition, they each observe their own private RO, θ_i for $i \in \{1,\ldots,m\}$.

Consider first a middleman who tries to use an RO. Following the discussion in section 4.1, an **EB** buys all units priced at or below z if he arrives. Conditional on θ_i and z, the **M** expects constant marginal revenue of reselling:

$$\mathbb{E}\left[E_i Z|\theta_i, z\right] = \theta_i z. \tag{1}$$

Under assumption 2, θ_0 can be either 0 or 1 and for $i \neq 0$, θ_i can be either ϕ or 0. The random variable Θ_i can be interpreted as a signal about the reselling channel i ($i = 0$ for the common RO in the market). More

[18] As the paper shows (in sections 4.2 and 4.3.2), the supply schedules that **EB** and **B** face, in periods 1 and 2, respectively, are (weakly) increasing.

generally, the RO can be written as a function $g(a; \Theta, Z)$. For the preferred "early buyer" interpretation, it follows that $\mathbb{E}[g(a; \Theta_i, Z)|\theta_i, z] = \theta_i z a$, for all $i \in \{0, 1, \ldots, m\}$, consistent with equation (1).

Next, consider **M**1, who is trying to resell the asset bought from **S** in period 0. There are three possibilities. First, **M**1 can attempt to resell in the market to an **EB**. Second, it can try its private channel. Finally, it can choose to resell to another **M**. Using equation (1), one can easily show that **M**1's optimal reselling strategy only depends on θ_0 and of θ_1:

Lemma 1. (**M**1's strategy in period 1.) *Suppose* **M**1 *holds q units of the asset and observes the realizations z,* θ_0*, and* θ_1*. Then:* 1) *If* $\theta_0 = 1$*,* **M**1 *posts in the market a take-it-or-leave-it offer of q units at price z.* 2) *If* $\theta_0 = 0$ *and* $\theta_1 = \phi$*,* **M**1 *resells privately (by posting a take-it-or-leave-it offer of q units at price z).* 3) *If* $\theta_0 = \theta_1 = 0$*,* **M**1 *posts in the market a take-it-or-leave-it offer of q units at price* ϕz*.*

When a good signal ($\theta_0 = 1$) is drawn for the common RO, an **EB** arrives with probability 1 and guarantees the success of reselling (by assumption 2, $E_0 = \Theta_0$ almost surely). **M**1 posts in the market a supply of all q units at the **EB**'s reservation price, z, to extract maximal surplus from him. When $\theta_0 = 0$ (a bad draw of the signal for the common RO), **M**1 looks at its private RO and if the signal is a good draw ($\theta_1 = \phi$), **M**1 will try the private channel. If both signals turn out to be bad ($\theta_0 = \theta_1 = 0$), **M**1 attempts to resell to another **M**, who might have a good signal θ_i for its private RO. In this case, **M**1 must lower the price by $1 - \phi z$ to compensate for the buying **M**'s expected cost of carry, $(1 - \phi)z$ (equation 1).

Therefore, the price offered by **M**1 is lower when targeting another **M** ($\theta_0 = \theta_1 = 0$) than when targeting **EB** in the market ($\theta_0 = 1$). This paper refers to case *iii)* of lemma 1 as *inter-middlemen trade* or *MM trade* for short (if such a trade indeed realizes), and to case *ii)* as a *middleman-buyer trade* or MB trade.[19] The difference between the two prices is $(1 - \phi) \times 100\%$. Therefore, $(1 - \phi)$ will be referred to as *price pressure*[20], which is endogeneously determined by the size of **M**'s ROs following assumption 2. Possible generalizations are discussed in section 6.2.

[19] The term middleman-buyer trade is helpful to remind us the trading parties of such a trade. More generally, one can think of such trades as *middleman-investor* trades. That is, the investor is not necessarily a buyer, although in the current model it is so as the paper only focuses on the sell-side of the market. Importantly, such trades occur *at the fundamental value* of the asset (absent frictions between the middleman and the investor).

[20] It should be noted that it is *not* the asymmetry between the distributions of Θ_0 and of Θ_i ($i \in \{1, \ldots, m\}$) that drives the price pressure of MM trades. A detailed discussion on a more general specification of RO is provided in section 6.2.

4.2.2 M's Strategy in period 0

In period 0, no information about Z or Θ_i has been revealed yet. **M** only observe **S**'s supply, which (under assumption 1) in equilibrium will be a simple take-it-or-leave-it offer, as shown below in section 4.3.3. Only one of **M** will get **S**'s offer, and becomes the asset owner in period 1. The others become non-owners. All **M** want to become the owner, because an owner exploits the surplus of **EB** and therefore earns positive expected profit in period 1.[21]

Lemma 2. (**M**'s strategy in period 0.) *In period 0, all* **M** *have the same constant marginal reservation value* $[{}^{1}\!/_{2} + \phi \cdot (\gamma + \beta(m))]\bar{z}$*, where* $\bar{z} := \mathbb{E}Z = {}^{1}\!/_{2}$ *and the probabilities* γ *and* $\beta(m)$ *are defined in equation* (2) *in section 4.3.1. Given a supply schedule from* **S***, each* **M** *wants to buy all the asset supplied at or below this reservation value.*

M's reservation value has three components: First, with probability of $1/2$, Θ_0 realizes to be 1 and in this case, **M**1 can resell to **EB** in the market and get, in expectation, $\bar{z}$ (see equation 1 and recall that Θ_0 and Z are independent). Second, with probability $\gamma = \mathbb{P}(\Theta_0 = 0, \Theta_1 = \phi)$, **M**1 cannot resell in the market but can resell privately, and gets $\phi\bar{z}$ in expectation. Finally, if unable to resell either in the market or via its private channel, **M**1 attempts to resell to the other **M**, the non-owners, at their reservation price $\phi\bar{z}$ (see lemma 1). The MM trade, however, will only succeed if at least one of the $m - 1$ non-owners has a good signal of its private RO; that is, only with probability $\beta(m) = \mathbb{P}\left(\Theta_0 = \Theta_1 = 0, \sum_{i=2}^{m} \Theta_i > 0\right)$. These three components sum up to the reservation value in lemma 2.

4.3 Seller (S)

This section first analyzes how the reselling activity in period 1 might affect **S**'s learning in period 2. **S**'s optimal selling strategy in period 2, given different realizations of market activity, are then derived in section 4.3.2. Finally, **S**'s period 0 supply schedule is backwardly solved in section 4.3.3.

4.3.1 S's learning from market activity

When **S** returns to the market in period 2, she observes the period 1 market activity, in particular **M**1's price quote (if there is any) and whether there was a trade. Note that only the activity occurring *in the market* is observable to **S**, not the activity in the private channels of **M**.

[21] However, such profitability is fully wiped in period 0 because **S** is the first-mover (see section 4.3.3).

[Table 1 about here]

Table 1 summarizes the possible market activity in period 1, together with the underlying events. For clarity of notation, define the following probabilities

$$\begin{aligned}\alpha(m) &= \mathbb{P}\left(\Theta_0 = \Theta_1 = \cdots = \Theta_m = 0\right) = \left(\frac{1}{2}\right)^{1+m},\\ \beta(m) &= \mathbb{P}\left(\Theta_0 = \Theta_1 = 0, \sum_{i=2}^{m} \Theta_i > 0\right) = \frac{1}{4} - \left(\frac{1}{2}\right)^{1+m}, \text{ and}\\ \gamma &= \mathbb{P}\left(\Theta_0 = 0, \Theta_1 = \phi\right) = \frac{1}{4}.\end{aligned} \tag{2}$$

From lemma 1, the supply schedule in period 1, if it exists, is simply a take-it-or-leave-it offer. The informative part to **S** is the price quote P_1, a random variable that depends on Z and on Θ_i. The price quote and potential subsequent trade activity could result in one of the three scenarios:

Fully revealing. The true value of Z is fully revealed to **S** when she observes a trade at price $P_1 > \phi$ or a price quote P_1 but no trade (the first two rows in table 1). A trade at price $P_1 > \phi$ could not be an MM trade because the price pressure accompanying an MM trade caps the trading price at ϕ (= $\sup\{\phi Z\}$). Consequently, **S** perfectly learns that conditional on the observed p_1 ($> \phi$), $Z = p_1$. This happens with probability $\mathbb{P}(\Theta_0 = 1, E_0 = 1, Z > \phi) = (1 - \phi)/2$. A price quote P_1 without a trade is also fully revealing because this situation was only possible when an MM trade was attempted but failed, i.e. when all Θ_i had bad draws ($\Theta_0 = \Theta_1 = \cdots = \Theta_m = 0$), with probability $\alpha(m)$. Seeing p_1, **S** therefore also perfectly learns that Z as it equals p_1/ϕ in this situation.[22]

Partially revealing. When **S** observes a trade at price $P_1 < \phi$, the posterior distribution of Z becomes binomial (the shaded rows in table 1): With probability $\beta(m)$, the observed trade was an MM trade, and with probability $\phi/2$ (= $\mathbb{P}(\Theta_0 = 1, Z \leq \phi)$) the trade was an MB trade. This way, **S** learns, but not perfectly, about Z from period 1 market activity.

[22] The fully revealing result of no trade is a consequence of assumption 2. In particular, the Bernoulli distribution of Θ_0 rules out the no-show of an **EB** in the market given a good draw of Θ_0. More generally, one can assume, for example, that $\Theta_0 = \phi X_0$, where X_0 is Bernoulli distributed, such that all Θ_i are i.i.d. Under this alternative specification, **S**'s learning problem is further complicated as an untraded quote is no longer perfectly revealing because it can be caused by either a failure in MM trade or in MB trade. The key inference problem, as discussed in "partially revealing" below, is not affected.

No learning. If **M**1 resold through its own private channel, there would be no market activity in period 1 and **S** does not learn anything (the last row in table 1). This happens with probability γ.

4.3.2 S's selling strategy to B in period 2

Based on her learning from period 1 activity, **S** obtains a posterior distribution of Z. She then optimizes her supply schedule to sell her position of a_2 units of the asset to **B**. Consider the following three scenarios.

Fully revealing. **S** observes the true value z. From section 4.1, her maximal marginal revenue of selling the asset to **B** is z. On the other hand, her marginal utility of retaining some positive units of the asset is $z \cdot (1 - k_a(a))$, always weakly lower than z for $a \geq 0$. Therefore, when Z is fully revealed, **S** posts all a_2 units of the asset at price z, a take-it-or-leave-it offer, and gets

$$u^S_{2,\mathrm{fr}}(a_2; z) = z \cdot a_2, \tag{3}$$

where the superscript "S" and subscript "2,fr" indicate this utility is for **S** in period 2 with Z fully revealed.

Partially revealing. **S** observes a trade at price $p_1 \leq \phi$. Unconditionally, an MM trade happens with probability $\beta(m)$ while an MB trade at price $P_1 \leq \phi$ happens with probability $\phi/2$ (see table 1). Therefore, the posterior distribution of Z is

$$Z|_{p_1 \leq \phi, \mathrm{trade}} = \begin{cases} p_1, & \text{(low) with probability } 1 - \hat{\beta} \\ p_1/\phi, & \text{(high) with probability } \hat{\beta} \end{cases} \tag{4}$$

where $\hat{\beta} := \beta/[\beta + \phi/2]$ is the conditional probability of a period 1 MM trade.

Lemma 3. (**S**'s period 2 supply when Z is partially revealed.) *With a_2 units of the asset and Z conditionally distributed as in equation* (4)*,* **S** *posts a supply schedule of*

$$s_2(p) = \begin{cases} a_2, & \textit{if } p \geq p_1/\phi \\ \max\{0, a_2 - \hat{a}(m)\}, & \textit{if } p_1 \leq p < p_1/\phi \\ 0, & \textit{if } p < p_1 \end{cases} \tag{5}$$

where $\hat{a}(m) = 2(1-\phi)\beta(m)/\phi^2$ and $\beta(m)$ defined as in equation (2).

[Figure 1 about here]

Figure 1 plots the supply schedule. Panels (a) and (b) depict the cases of $\hat{a}(m) > a_2$ and of $\hat{a}(m) \leq a_2$, respectively. **S** trades off two effects: the expected cost of carry and the expected adverse selection cost. The cost of carry can be reduced by transferring more units of the asset to **B**, at the cost of more severe adverse selection (the shaded area in panel (a)). She chooses her optimal supply at the low price such that, in expectation, the marginal adverse selection cost is equal to the marginal cost of carry. In the current model, this trade-off has a simple solution where **S** chooses to retain a fixed amount of the asset, $\hat{a}(m)$, at price p_1 (see the proof of lemma 3 in appendix 7).

Note that the threshold $\hat{a}(m)$ strictly increases in $\beta(m)$, the probability of an MM trade, hence also in m, the number of middlemen. Intuitively, this threshold measures how costly the adverse selection is: A larger m implies a higher $\beta(m)$ and a higher $\hat{\beta}(m)$. In other words, **S** is more likely to be adversely-selected. To avoid that, she chooses to sell fewer units at the low price to retain a larger amount–$\hat{a}(m)$–of the asset. (Discussion on the properties of the supply schedule is deferred to section 5.2.)

Next, evaluate the (conditionally) expected period 2 utility of **S**, using the supply schedule described by lemma 3 above. First, when the observed period 1 trade was indeed an MM trade, $z = (p_1/\phi) \in [0,1]$ and **S** sells everything and suffers adverse-selection cost. For $z \in [0,1]$:

$$u^S_{2,\mathrm{MM}}(a_2; z) = z \cdot a_2 - \phi z \cdot s_2(p1) = \begin{cases} z \cdot [a_2 - (1-\phi)(a_2 - \hat{a}(m))], & \text{if } a_2 \geq \hat{a}(m) \\ z \cdot a_2, & \text{if } a_2 < \hat{a}(m) \end{cases}. \tag{6}$$

Second, if the observed period 1 trade was in fact an MB trade, $z = p_1 \in [0, \phi]$, and **S** will not be able to transfer all her position to **B**. She will suffer the cost of carry $k(a_2 - s_2(p_1))$. For $z \in [0, \phi]$:

$$u^S_{2,\mathrm{MB}}(a_2; z) = z \cdot [a_2 - k(a_2 - s_2(p_1))] = \begin{cases} z \cdot [a_2 - k(\hat{a}(m))], & \text{if } a_2 \geq \hat{a}(m) \\ z \cdot [a_2 - k(a_2)], & \text{if } a_2 < \hat{a}(m) \end{cases}. \tag{7}$$

In equations (6) and (7) above, subscripts "MM" and "MB" respectively indicate the states of the world in which these expressions apply.

No learning. Without learning anything from period 1, **S** holds the prior belief that Z is uniform on $[0, 1]$. With a_2 units of the asset, she tries to solve an optimal supply schedule $s(\cdot)$ that maximizes her expected utility:

$$\max_{s(\cdot)\geq 0} \mathbb{E}\left[\int_0^Z (s(Z) - s(p))\mathrm{d}p + Z \cdot ((a_2 - s(Z)) - k((a_2 - s(Z)))\right], \tag{8}$$

from which it can be seen that she is subject to both the cost of carry, $k(a_2 - s(Z))$, and the adverse selection, $\int_0^Z s(p)\mathrm{d}p$. The solution is given by the following lemma.

Lemma 4. (S's period 2 supply when there is no learning.) *With a_2 units of the asset and Z uniformly distributed on* $[0, 1]$*,* **S** *posts a supply schedule of*

$$s_2(p) = \begin{cases} a_2, & \textit{if } p > 1 \\ a_2 - (1/p - 1), & \textit{if } z^*(a_2) \leq p \leq 1 \\ 0, & \textit{if } p < z^*(a_2) \end{cases} \tag{9}$$

where $z^*(a) = 1/(1 + a)$ (< 1). **S***'s expected utility in this case is*

$$\mathbb{E}u^S_{2,nl}(a_2; Z) = \frac{1}{2} \cdot \ln(1 + a_2), \tag{10}$$

where the superscript "S" refers to **S** *and the subscript "2,nl" to period 2 and "no learning".*

Intuitively, optimality requires that the marginal increase in being adversely-selected equal the marginal reduction in the cost of carry.

4.3.3 S's selling strategy to M in period 0

With the solutions of **S**'s period 2 expected utility (section 4.3.2) and **M**'s period 0 strategy (section 4.2.2), **S**'s optimal supply schedule in period 0 can now be solved. By lemma 2, **S** knows that all **M** have reservation value $[1/2+\phi\cdot(\gamma+\beta(m))]/2$ in period 0. To maximize her expected utility, therefore, **S** will not sell anything below this reservation value. Under the simplifying assumption 1, it is (weakly) dominating for **S** to post nothing at prices above this reservation value, because no **M** in period 0 will touch such supply and because all untouched supply schedules expire at the end of period 0. (The consequence of allowing the supply schedule to persist across time periods is discussed later in section 6.1.) Therefore, **S**'s strategy space reduces to a simple take-it-or-leave-it offer of a pair (p_0, q), where

$$p_0 = \left[\frac{1}{2} + \phi \cdot (\gamma + \beta(m))\right] \cdot \bar{z} =: p_0(m) \tag{11}$$

(with $\bar{z} = \mathbb{E}Z = 1/2$) is all **M**'s reservation value in period 0. It only remains to solve the optimal number of assets to supply, q.

S trades off between selling early in period 0, getting constant marginal revenue p_0, and late in period 2, getting $\mathbb{E}u_2^S(a_0 - q; Z)$:

$$\begin{aligned}
u_0^S = \max_q\ & p_0(m)q && \text{//selling early} \\
& + \frac{1}{2}(1-\phi)\mathbb{E}\left[u_{2,\text{fr}}^S(a_0 - q; Z)|Z > \phi\right] + \alpha(m)\mathbb{E}\left[u_{2,\text{fr}}^S(a_0 - q; Z)\right] && \text{//late, fully revealing} \\
& + \frac{1}{2}\phi\mathbb{E}\left[u_{2,\text{MB}}^S(a_0 - q; Z)|Z \le \phi\right] + \beta(m)\mathbb{E}\left[u_{2,\text{MM}}^S(a_0 - q; Z)\right] && \text{//late, partially revealing} \\
& + \gamma\mathbb{E}\left[u_{2,\text{nl}}^S(a_0 - q; Z)\right]. && \text{//late, no learning}
\end{aligned}$$

The corresponding probabilities $\alpha(m), \beta(m)$, and γ for **S**'s period 2 learning are defined in equation (2). The expected utility expressions $u_{2,\text{fr}}^S(\cdot)$, $u_{2,\text{MM}}^S(\cdot)$, $u_{2,\text{MB}}^S(\cdot)$, and $u_{2,\text{nl}}^S(\cdot)$ can be found by evaluating equations (3), (6), (7), and (10) respectively. Lemma 5 solves S's optimization problem.

Lemma 5. (**S**'s supply in period 0.) *There exists an m_0 and an $\hat{m}$ such that $1 < m_0 < \hat{m} \le \infty$.*[23] *In particular, $\hat{m} < \infty$ if $\phi < 1/2$ and $\hat{m} = \infty$ if $\phi \ge 1/2$. Let* **S** *hold a_0 units of the asset. Then the optimal supply function of* **S** *in period 0 is $q(m)$, a function of the number of middlemen, such that*

$$q(m) = \begin{cases} 0^+, & \text{if } m < m_0 \\ q_I(m), & \text{if } m_0 \le m \le \hat{m} \\ q_{II}(m), & \text{if } m > \hat{m} \end{cases}, \tag{12}$$

where 0^+ denotes an infinitesimally small yet positive number, $0 \le q_I(m) \le a_0 - \hat{a}(m)$, and $a_0 - \hat{a}(m) < q_{II}(m) < a_0$. The expressions of $q_I(m)$ and $q_{II}(m)$ are given by equations (A2) *and* (A3)*, respectively, in the*

[23] Here, the threshold m_0 and $\hat{m}$ are positive real numbers, though in reality, the exact number of middlemen can only be (positive) integers. However, the existence of real-numbered thresholds m_0 (> 1) and $\hat{m}$ (> 1) implies the existence of integer thresholds. We stick to real numbers here to avoid further notational burden.

appendix.

[Figure 2 about here]

Figure 2 illustrates **S**'s trade-off between selling early and late. Three cases are illustrated for different values of m. Panel (a) shows the corner solution (when $m < m_0$) where the marginal utility of selling late is strictly higher than that of selling early for all $q \in [0, a_0]$. In this case, **S** only sells an infinitesimally small amount, i.e. 0^+, of the asset at price p_0 to **M**.[24] Panel (b) illustrates the second case of $m_0 \leq m \leq \hat{m}$, where there is a unique intersection between the marginal utilities of selling early and late. In particular, the resulting optimal $q(m)$ falls in $[0^+, a_0 - \hat{a}(m)]$. Finally, panel (c) illustrates the last case of $m > \hat{m}$, in which the unique intersection between marginal utilities falls in $(a_0 - \hat{a}(m), a_0)$.

[Figure 3 about here]

Figure 3 plots the shape of the optimal supply $q(m)$ as a function of the number of middlemen, m. The supplied amount is (weakly) increasing in m. Depending on the RO parameter ϕ, there are two regimes: $q = q_{\mathrm{I}}(m)$ for $m \leq \hat{m}$ (panel (a)) and $q = q_{\mathrm{II}}(m)$ for $m > \hat{m}$ (panel (b)). Further discussion on the properties of $q(m)$ is deferred to proposition 4 and corollary 1.

5 Model implications

This section discusses the model implications. Subsection 5.1 focuses on inter-middlemen (MM) trades. Subsection 5.2 studies how MM trades create an inference problem and how **S** responds. It is then shown that **S** overloads **M** in period 0 in terms of welfare (section 5.3), which is characterized as a function of the number of middlemen (section 5.4). Finally, section 5.5 shows how trade type disclosure can improve welfare.

5.1 Inter-middlemen trades

As analyzed in section 4.2.1, an inter-middlemen (MM) trade occurs when **M**1 (the asset owner **M** in period 1) cannot resell by itself ($\theta_0 = \theta_1 = 0$), seeks help from other **M**, and fortunately finds another **M** to help.

[24] Selling nothing is dominated by selling 0^+. Without selling anything, the **M** will have no asset in period 1 and **S** is not able to learn anything. Instead, by selling 0^+, **S** incurs (almost) no opportunity cost and yet obtains the non-zero probability to learn from period 1 market activity.

The asset transfers from a low-RO middleman to a high-RO middleman. Price pressure arises, because the buying **M** requests compensation for its expected cost of carry: $(1-\phi)z$. Ex ante, an MM trade happens with probability $\beta(m)$ (lemma 2), which increases with the number of middlemen, m. This is because when an MM trade is needed, an additional **M** adds to the probability that at least some **M** can help.[25] The following proposition summarizes the above discussion about MM trades. (The first two results are repeated from proposition 2 on page 5.)

Proposition 1. (Summary of MM trades.) *Inter-middlemen trades, or MM trades, arise when two middlemen trade with each other for the best reselling opportunity. It* 1 *improves the allocational efficiency within the middlemen sector,* 2) *is accompanied by price pressure, and* 3) *is more likely to happen with more middlemen in the market.*

Perhaps the most interesting implication from the above proposition is the accompanying price pressure. Empirical evidence from the Flash Crash suggests that there was huge inter-HFT trading volume when the market price was falling most rapidly. See, for example, CFTC and SEC (2010)[26] and figure 8 of Kirilenko et al. (2011), who metaphorically refer to the inter-HFT trading volume as a "hot potato" game.[27] CFTC and SEC (2010) further note that there were two liquidity crises during the Flash Crash[28]: One was in the E-Mini future and the other in individual stocks, which could serve as the RO for the index future. Consistent with the model prediction, the price pressure of MM trades was very large when RO was low. MM trades could be a significant source of recent extreme price movements in electronic markets.

5.2 S's inference problem

MM trades are not distinguishable from MB trades: An investor does not know whether or not the observed trade price is pressured. This creates an inference problem for **S** as she cannot perfectly learn Z, the fundamental value.

[25] The two extreme cases illustrate this clearly: When there is only one **M**, $\beta(1) = 0$, and there is no possible counterparty with whom to trade. When there are infinitely many **M**, $\beta(\infty) = 1/4$, i.e. whenever an MM trade is needed (when $\theta_0 = \theta_1 = 0$), there is (almost surely) always a buyer **M** ready to trade (recall that each Θ_i is independently drawn).

[26] On page 3, the SEC notes that MM trades triggered a liquidity crisis: "Still lacking sufficient demand from fundamental buyers or cross-market arbitrageurs, HFTs *began to quickly buy and then resell contracts to each other*–generating a 'hot-potato' volume effect as *the same positions were rapidly passed back and forth*." (emphasis added).

[27] "Hot potato" trades have been studied in the literature with the main motivation of risk-sharing among intermediaries. See, for example, Lyons (1997), Naik, Neuberger, and Viswanathan (1999), and Viswanathan and Wang (2004). However, what "inter-middlemen trades" in this paper are instead motivated by heterogeneous reselling opportunities in middlemen.

[28] Cespa and Foucault (2012) provide a theoretical model of such liquidity crises (and the spillover), driven by imperfect learning across different but correlated assets.

The inference problem exposes **S** to the adverse selection by **B**, as she faces residual uncertainty about Z whereas **B** knows Z. **S** optimizes her supply $s_2(p)$ to trade off the adverse selection cost with the cost of carry (see lemma 3 and its discussion). The proposition below points out two situations inwhich **S** negatively affects market quality due to the inference problem.

Proposition 2. (Two consequences of **S**'s inference problem.) *Fix* **S***'s position a_2 in period 2.* 1) *When the adverse-selection is moderate ($a_2 > \hat{a}(m)$),* **S** *reinforces the price pressure of an MM trade.* 2) *When the adverse-selection is severe ($a_2 \leq \hat{a}(m)$),* **S** *refrains from selling at low price and the market breaks down following an MB trade.*

S reinforces the price pressure of an MM trade when her position is relatively large and she posts part of her position at the pressured price, as shown in panel (a) of figure 1. If **S**'s position is small, or if the adverse selection problem is severe, she would rather stick to her position, at the observed price, and bear the cost of carry than to sell and to suffer adverse selection. See panel (b) of figure 1. The market then breaks down if the observed trade was the fundamental price. The breakdown is a standard result of Akerlof (1970).

5.3 S overloading M in period 0

Lemma 5 outlines **S**'s optimal supply schedule in period 0. Essentially, she "loads" **M** and then lets them help her find buyers in period 1. When she returns to the market, she learns from the activity in period 1. This subsection asks whether **S** appropriately loads **M**, in terms of social welfare.

Welfare is measured in terms of the sum of all agents' expected losses.[29] In this economy, all buyers (**B** and all potential **EB**) have the highest valuation for the asset. Under the first-best allocation, all a_0 units of the asset should be transferred from **S** to **B** (or **EB**). However, not all units of the asset can be transferred without friction. Some units might still end up with **S**, while some others with **M**. Both **S** and **M** have lower marginal utility than **B** (and **EB**) for the asset. Ex ante, the social loss can be written as

$$l(q) = \mathbb{E}\overbrace{\sum_{i=1}^{m}\left[u^B(0,A_3^{\mathbf{M}i}(q)) - u^M(0,A_3^{\mathbf{M}i}(q))\right]}^{\text{losses for } \mathbf{M}} + \mathbb{E}\overbrace{\left[u^B(0,A_3^S(q)) - u^S(0,A_3^{\mathbf{S}}(q))\right]}^{\text{loss for } \mathbf{S}}, \tag{13}$$

[29] An alternative, mathematically equivalent way is to directly evaluate the sum of expected utility of all agents–**S**, **EB**, **B**, and all **M**–in this economy. Social welfare, therefore, is simply the first-best expected total gains from trade minus the expected social loss: $w = \bar{z} \cdot a_0 - l(q)$.

where the asset positions in period 3, $A_3^{\mathbf{M}i}$ for $i \in \{1, \ldots, m\}$ and $A_3^{\mathbf{S}}$, are stochastic functions of the initial loading q.[30]

Proposition 3. (Overloading in period 0.) *If $m \geq m_0$,* **S** *"overloads"* **M** *in period 0 in the sense that social loss can be decreased by reducing her supply $q(m)$ (as defined in lemma 5).*

To understand the intuition behind overloading, consider again **S**'s optimization problem in period 0 (see section 4.3.3): a tradeoff between selling early (in period 0) and late (in period 2). Selling early allows **S** to transfer some of her position to **M** who are more likely to find **EB**. Selling late exposes **S** not only to her expected cost of carry, *but also to the expected adverse selection cost*. Effectively, **S** trades off the expected losses for **M** (in case of failure to resell to **EB**) with the expected losses for herself. This looks very much like the trade-off for a social planner, except that a social planner only cares about the cost of carry on **S**'s loss, while **S** also dislikes the adverse selection. Therefore, **S** sees a larger expected loss in selling late than does a social planner. Consequently, **S** loads more than would a social planner to **M** in period 0, implying higher allocational inefficiency.

5.4 How many middlemen?

How does market quality depend on the number of middlemen? Proposition 1 shows that the price pressure associated with MM trades appears more frequently with more middlemen. To this extent, market volatility increases with the number of middlemen. This section explores how an additional **M** changes **S**'s expected utility and, more importantly, social welfare.[31]

5.4.1 The number of M on S's expected utility

Consider **S**'s expected utility in period 0 (section 4.3.3 and lemma 5). There are two effects of an additional **M** on **S**'s utility, as outlined in the proposition below (see also equation A5 for an example).

Proposition 4. (Two effects of an additional **M** on **S**'s utility.) *An additional middleman* 1) *increases the adverse-selection cost of* **S** *and* 2) *increases* **M***'s reservation value in period 0.*

[30] Note that in computing the losses, the transfer of numéraire good is set to 0 in the above utility functions. This is because all agents equally value the numéraire good.

[31] The analysis only focuses on the effect of an additional **M** on **S**'s expected utility and on welfare, but not on that of **M** or **B** (**EB**). This is because in the current model, **M**'s expected utility is always zero (as **S** has market power in period 0). Consequently, the difference between welfare and **S**'s expected utility becomes **B**'s (and **EB**'s) expected utility, whose analysis is omitted to avoid repetition.

[Figure 2 and 3 about here]

First, an additional **M** makes MM trades more likely (proposition 1), worsens the learning of **S**, and exposes **S** to a higher probability of being adversely selected in period 2. Second, additional **M** increase ROs of all middlemen. More **M** means that it is easier (larger $\beta(m)$) to find a counterparty with whom to trade when needed. This increase of RO translates to the increase of $p_0(m)$ as seen from equation (11), shifting up the marginal utility of selling early. Graphically, the marginal utility of selling late shifts downward while $p(0)$ shifts up in figure 2, from panel (a) to (c). (A third effect sketched in figure 2 is that the threshold $\hat{a}(m)$ is also increased by an additional **M**; see the discussion in "partial revealing" of section 4.3.2.)

These two effects together push the intersection point of the marginal utility of selling early and late rightwards, implying a (weakly) increasing function $q(m)$:

Corollary 1. (**S**'s supply in period 0 and the number of middlemen.) *Ceteris paribus, the more middlemen, the more* **S** *sells to* **M** *in period 0. Mathematically,* $q(m)$*, as defined in equation* (12)*, is weakly increasing in* m*.*

Figure 3 sketches q against the number of middlemen, m: panel (a) for the case of $\hat{m} < \infty$ and panel (b) for $\hat{m} = \infty$.

5.4.2 The number of M on welfare

We now turn to welfare and the number of middlemen. There are three effects on welfare of an additional **M** (see also equation A6):

Proposition 5. (Three effects of an additional middleman on welfare.) *When there is an additional middleman,* 1) **S** *further overloads all middlemen (in the sense of proposition 3),* 2) **S** *faces (weakly) higher expected cost of carry, and 3) the overall reselling opportunity of all middlemen increases. The first two effects reduce welfare and the third effect improves welfare.*

The first is the "overloading" effect: As m increases, **S** loads more to the middleman sector in period 0 (corollary 1), though she has already "overloaded" them (proposition 3). The second effect results from **S**'s worsened inference problem as MM trades become more likely (proposition 1): **S** retains a larger amount of the asset due to her increased adverse selection cost in period 2 (the first part of proposition 4). Finally, the third effect benefits **M** (and also welfare) by making reselling more likely as a means of efficient allocation among **M** (proposition 1). This is the second part of proposition 4. The net effect, however, cannot be signed in general. Nevertheless, the shape of welfare, as a function of m, can be characterized as follows.

Corollary 2. (Social welfare and the number of middlemen.) *Welfare is strictly decreasing on* $m \in [1, m_0)$ *and is quasi-convex on* $m \in [m_0, \infty)$ *for some* $m_0 \in \mathbb{R}^+$.

This corollary implies that to determine the socially optimal number of middlemen, one only needs to compare the polar cases of $w(m = 1)$ and $w(m = \infty)$. The comparison reveals that the result critically depends on ϕ, the abundance of ROs.

Corollary 3. (Optimal number of middlemen in terms of welfare.) *There exists a threshold* $\phi^* \in (1/2, 1)$ *such that* $w(m = 1; \phi^*) = w(m = \infty; \phi^*)$ *and for all* $\phi \in (\phi^*, 1)$*,* $w(m = 1; \phi) < w(m = \infty; \phi)$*.*

This corollary says that when each **M** has relatively abundant RO ($\phi > \phi^*$), society is better off with as many **M** as possible. Intuitively, ϕ captures the severity of **S**'s inference problem: In the extreme of $\phi \to 1$, the first two negative effects of proposition 5 go away and an additional **M** only adds to the RO and benefits welfare (the last effect of proposition 5). Contrarily, when each **M** contributes very small RO ($\phi < \phi^*$), it is optimal to have only one **M**. This is because in such case, the social loss due to the inference problem is so large that it cannot be compensated for by the increased RO.

[Figure 4 about here]

Figure 4 illustrates $w(m)$ (and $u_0^{\mathbf{S}}(m)$) for two cases.[32] Panel (a) plots welfare and **S**'s expected utility with $\phi < 1/2 (< \phi^*)$. (Recall from lemma 5 that $\phi < 1/2$ implies $\hat{m} < \infty$.) Panel (b) plots the case of $\phi > \phi^* (> 1/2)$. The kink at $m = m_0$ is due to the lower bound of $q = 0$ that binds for $m < m_0$. In both cases, the shape of $w(m)$ is quasi-convex. In panel (a), $w(m = 1) > w(m = \infty)$ because ROs are relatively scarce, and in panel (b), $w(m = 1) < w(m = \infty)$ because ROs are relatively abundant.

5.5 Flagging inter-middlemen (MM) trades

The above analysis shows that **S**'s inference problem impedes efficient reallocation. One way to resolve this problem is to flag MM trades so that **S** is able to distinguish pressured trades from unpressured ones. This section studies what such disclosure policy would mean for welfare.

[32] Note that the illustration in figure 4 is not meant to be exhaustive. Among others, there is also a case where $\phi \in (1/2, \phi^*)$. Only the general patterns of $w(m)$ and $u_0^{\mathbf{S}}(m)$ are depicted.

S's learning (from period 1 market activity). When MM trades get flagged, **S**'s inference problem in period 2 is solved. This is most clearly seen in table 2. A period 1 trade with price $P_1 \leq \phi$ is an MM trade if it is flagged. Otherwise, it is an MB trade (c.f. table 1). This way, the period 1 market activity (if any) is always fully revealing, and **S** learns nothing only with probability γ.

[Table 2 about here]

S's supply in period 2. With probability $(1-\gamma)$, **S** perfectly learns about Z when she returns to the market in period 2. Following the "fully revealing" analysis in section 4.3.2, **S** posts all her a_2 units of the asset at z and gets expected utility $z \cdot a_2$ (equation 3). With probability γ, **S** learns nothing, applies lemma 4, and in expectation gets $\ln(1+a_2)/2$ (equation 10).

S's supply in period 0. For clarity notation, variables under the disclosure policy are overlined with a "~". As in the no-disclosure case, **S** considers a take-it-or-leave-it offer at the price $p_0(m)$ of equation (11). Only the optimal amount to supply, q, remains to be solved (see the analysis in section 4.3.3). **S** chooses $\tilde{q} \in [0, a_0]$ to optimally trade off selling early and late:

$$\begin{aligned} \tilde{u}_0^S = \max_{\tilde{q}}\ & p_0(m)\tilde{q} && \text{//selling early} \\ & +(1-\gamma)\mathbb{E}\left[u_{2,\text{fr}}^S(a_0-\tilde{q};Z)\right] && \text{//late, fully revealing} \\ & +\gamma\mathbb{E}\left[u_{2,\text{nl}}^S(a_0-\tilde{q};Z)\right]. && \text{//late, no learning} \end{aligned}$$

For the purpose of welfare comparison, instead of fully solving this optimization problem, it suffices to characterize the solution $\tilde{q}(m)$ and compare it with the optimal supply $q(m)$ when there is no disclosure.

Lemma 6. (Optimal supply in period 0 under disclosure and no disclosure.) *In period 0,* **S** *sells (weakly) fewer units under the disclosure policy than she does without disclosure. Mathematically,* $\tilde{q}(m) \leq q(m)$.

To understand the intuition behind the above lemma, compare the above tradeoff between selling early and selling late with the same tradeoff but without disclosure (as stated in section 4.3.3). The marginal utility of selling early is the same: p_0. However, the marginal utility of selling late is higher with disclosure than without, because flagging solves the inference problem and **S** no longer faces adverse selection. Figure 5 graphically illustrates that $\tilde{q} \leq q$.

[Figure 5 about here]

Welfare. The welfare loss $\tilde{l}(q)$ consists of two components (see also section 5.3): the (expected) loss for **M** and the (expected) loss for **S**. Compared to $l(q)$, the loss for **M** is unchanged. However, the loss for **S** is reduced. Intuitively, this is because the disclosure resolves **S**'s inference problem and more units of the asset can be transferred to **B** than in the no-disclosure case. Therefore, for the same amount q_0, $\tilde{l}(q_0) \leq l(q_0)$ (this result is proved formally in the proof of proposition 6 in appendix 7). Together with lemma 6, the following proposition can be proved:

Proposition 6. (Welfare and disclosure.) *With the disclosure of trade types, welfare improves. In particular,* $\tilde{w}(m = 1) = w(m = 1)$ *and* $\tilde{w}(m = \infty) = w(m = \infty)$.

Finally, the disclosure might also switch the optimal number of **M** from 1 to ∞.

Corollary 4. (Switch of the optimal number of middlemen.) *There exists* $\tilde{\phi}^* \in (0, \phi^*)$ *such that for all* $\phi \in (\tilde{\phi}^*, \phi^*)$, $w(m = 1, \phi) > w(m = \infty, \phi)$ *(without disclosure) but* $\tilde{w}(m = 1, \phi) > \tilde{w}(m = \infty, \phi)$ *(with disclosure).*

6 Discussion

This section relaxes some of the model's assumptions (see section 3) and argues that the main results and economic intuitions, in particular, MM trades and **S**'s inference problem from market activity, are robust to alternative specifications or model extensions.

6.1 Slow cancellation of S's supply schedule in period 0

Assumption 1 restricts **S**'s strategy space in period 0 by forcing any untraded supply schedule to expire at the end of period 0. One possible motive is that **S** does not want to expose her willingness to sell so long that it starts to hurt her (see, for example, Brunnermeier and Pedersen (2005) and Attari, Mello, and Ruckes (2005) who explain how predatory trading hurts large investors). How is the analysis affected when **S**'s supply schedule cannot be canceled quickly enough? This relaxation complicates the analysis for two reasons. First, the remaining supply schedule in period 1 enriches the set of possible market activity. For example, both **M** and potential **EB** can trade against the remaining supply, creating quotes and trades that are different from those outlined in table 1. Also, when some **M** trades against the remaining supply of **S** in period 1, this **M** becomes another asset owner and also has an incentive to further resell (via the common RO, its private RO, or reselling to another **M**). This reselling creates new types of market activity. The expanded set of market activity is surely more realistic than the five rows listed in table 1, but is also harder to analyze in a compact and orderly way.

Second, the uncanceled supply schedule exposes **S** to adverse selection by both **M** and the potential **EB**, who are informed of the fundamental value Z in period 1. **S** will then also take into account the expected adverse selection cost when choosing her supply in period 0. The period 0 optimization problem then requires **S** solve a supply function $s_0(\cdot)$ (such as the one currently solved in section 4.3.2, "no learning"), whose analytical solution is not guaranteed.

Nevertheless, under the relaxation, the asset owner **M** (there can be a multiple of them) in period 1 is not affected by the relaxation: It still seeks to resell to another **M** when necessary. Further, **S**'s inference problem is not resolved by the additional (more complicated) market activity. Although lacking tractable algebraic derivation, the intuitions behind the model implications in section 5 remain valid.

6.2 Generalization of reselling opportunities

Assumption 2 imposes a very specific form of reselling opportunities (RO). Three possible generalizations are discussed below. The first one relates to possible market activity in period 1. The other two allow for more general characterizations of price pressures associated with MM trades.

First, the common RO and private ROs are not identically distributed under assumption 2. Specifically, Θ_0 follows a Bernoulli distribution while all other ROs follow an i.i.d. *scaled* Bernoulli distribution that yields ϕ on a successful draw. (As explained in footnote 20, it is not this asymmetry that drives the price pressure of MM trades.) Setting the successful draw of the common RO to be 1 only simplifies the analysis by reducing the set of possible period 1 market activity. Consider an alternative specification under which all Θ_i have the same i.i.d. distribution with realizations 0 or ϕ equally likely. Conditional on a good draw of the common RO, **EB** arrives in the market with probability $\mathbb{P}(E_0 = 1|\theta_0 = \phi) = \phi < 1$. **M**1 still has the (weak) incentive to target **EB** in the market. However, there is non-zero probability $(1 - \phi)$ that such an **EB** does not arrive, and this adds to the possible market activity (c.f. table 1), complicating the analysis.

Second, the very simple Bernoulli distribution of private ROs permits only discrete realizations. They are either low at zero, or high at ϕ, such that $1 - \phi$ describes the price pressure of MM trades, as shown in section 4.2.1. One can generalize price pressure by assuming a continuous distribution for Θ_i, $i \in \{1, \dots, m\}$, with differentiable C.D.F. $F(\theta)$ defined on $[0, 1]$. As in section 4.2.1, conditional on θ_i and z, $\mathbf{M}i$ has the constant expected marginal revenue, $\theta_i z$, of privately reselling. To illustrate, consider when $\mathbf{M}1$, the asset owner in period 1, seeks to trade with another $\mathbf{M}$. Given the linear preference $u^{\mathbf{M}}(\cdot)$, $\mathbf{M}1$ chooses the MM-trade price p_1 to maximize its expected utility (per unit of the asset):

$$\max_{p_1} \left[1 - \mathbb{P}(\Theta_2 z < p_1, \dots, \Theta_m z < p_1)\right] p_1,$$

where the term in the square brackets is the probability that at least one $\mathbf{M}$ (other than $\mathbf{M}1$) has a large enough draw of Θ_i such that it is willing to buy the asset at price p_1 from $\mathbf{M}1$. The optimal solution of p_1 (off corner) must satisfy the first order condition:

$$1 - F\left(\frac{p_1}{z}\right)^{m-1} = (m-1)F\left(\frac{p_1}{z}\right)^{m-2} \cdot f\left(\frac{p_1}{z}\right) \cdot \frac{p_1}{z}$$

for $m \geq 2$. The price pressure of MM trades, defined as $(1 - p_1/z) \times 100\%$, becomes a function of m, parametrized by variables shaping the distribution $F(\cdot)$. Such generalization allows tracking of how price pressure varies with the number of middlemen (at the cost of additional complexity and losing closed-form solution). This feature is ignored in the current model due to the simple Bernoulli distribution of reselling opportunities.

Third, private ROs are assumed to be independent of each other. This ensures deterministic price pressure associated with MM trades. To allow for stochastic price pressure, consider a more general joint C.D.F. $F(\theta_1, \dots, \theta_m)$. In period 1, the asset owner $\mathbf{M}1$ chooses the MM-trade price p_1 to maximize its expected utility (per unit of the asset), *conditional on its own private RO* θ_1:

$$\max_{p_1} \mathbb{E}\left[1 - \mathbb{P}(\Theta_2 z < p_1, \dots, \Theta_m z < p_1) | \Theta_1 = \theta_1\right] p_1.$$

The solution p_1 is, in general, dependent on $\mathbf{M}1$'s own private RO, θ_1, and therefore becomes stochastic ex ante. Stochastic price pressure further complicates $\mathbf{S}$'s inference problem. Conditional on a period 1 trade, $\mathbf{S}$ estimates the fundamental value in two steps: 1) What is the probability that the observed trade is an MM trade? 2) How large is the price pressure if it is an MM trade? Under the current framework where price pressure is deterministic, the second step is trivial. However, the extension heavily burdens the solution of **S**'s optimal supply schedule in period 2, where she faces a more complicated inference problem.

To summarize, the very simple (scaled) Bernoulli specification of ROs assumed in this paper ignores some important and realistic aspects of **S**'s learning and on price pressure associated with MM trades. This sacrifice is made in exchange for model tractability and clear interpretations of economic forces as discussed in section 5. Importantly, the extension to more generalized RO structures only adds to, but does not affect the existing effects of ROs on **M**'s reselling, MM trades, price pressure, and **S**'s learning. We therefore view the simplification a first step in addressing the interaction among middlemen in, as well as its impact on, electronic securities markets.

6.3 Assumptions embedded in the model structure

Information structure. A key friction in the model is the information asymmetry between **S** and **B**. Such friction is motivated by that the late agent observes more than does the early one. It is necessary for the early one, **S**, to stick her neck out to explicitly show her interest to sell, thereby suffering adverse selection. In addition, it is assumed that **M** also perfectly observe the fundamental value, due to their superior information processing technology. More generally, one can impose an information structure similar to that of Jovanovic and Menkveld (2011) where all human investors (**S** and **B**) observe the soft component of the asset value, while all machines (**M**) observe the hard component. This way, the reselling activity of **M** helps **S** learn about the hard component but not perfectly. The intuition developed in the current model about MM trades and **S**'s inference problem still holds true. In fact, the current information structure is an extreme case of Jovanovic and Menkveld (2011), with the soft component being zero and the late **B** observing the hard component with certainty.

Linear preference of M. Linear preferences have been used extensively in the literature to model middlemen (see, for example, Glosten and Milgrom (1985), Kyle (1985), Easley and O'Hara (1987), and more recently Duffie, Gârleanu, and Pedersen (2005)). The classic inventory management problem of middlemen in this setting is very reduced-form, i.e. the marginal cost of carry is strictly positive but constant. (c.f. Stoll (1978) and Ho and Stoll (1983)). One could introduce curvature in **M**'s utility function, such as risk aversion or convex cost of carry. One can then study how **M** optimally manage inventory when reselling to **EB** or to one another, and how such inventory control problem affects **S**'s decision in period 0. The cost, however, is mainly in model tractability. Notably, with nonlinear preference, reselling in period 1 involves decisions on both price and quantity. In contrast, with linear preference, **M**'s reselling decision only involves price, while the quantity is always as many as possible due to linearity. The analysis of **S**'s inference problem in period 2 is complicated by the new quantity dimension.

M's non-linear preference also incentivizes trades between two middlemen. Instead of ROs, such trades are driven by risk-sharing motivation (see, for example, Lyons (1997), Naik, Neuberger, and Viswanathan (1999), and Viswanathan and Wang (2004)). The linear preference adopted in this paper, therefore, captures the new motivation of heterogeneous ROs. The associated price pressure and inference problem are not affected by more complicated and more realistic formulations of **M**'s preference.

No middlemen in period 2. **S** uses **M** in the short run (periods 0 and 1) but not in the long run (period 2). Few (attentive) investors are in the market in a short time period, so **M** arise to provide immediacy for **S**. In the long run, many investors participate in trading, so there is no need for **M** to present and provide immediacy. Nevertheless, this is admittedly a simplifying assumption, which essentially rules out intertemporal strategic behavior of **M**. When **M** participate in trading in period 2, they intermediate between **S** and **B** and profit from adversely selecting **S** whenever **S** fails to learn Z perfectly from period 1 market activity. **M** have an incentive to prevent **S** from learning Z in period 1. Switching on this channel, therefore, will only worsen **S**'s inference problem in equilibrium.

S trading with only one M in period 0. It is assumed that after **S** posts her take-it-or-leave-it offer in period 0 (see section 4.3.3), all **M** will rush for her supply. However, only one **M**, chosen randomly, gets the offer. Such allocation within **M** is typical in electronic limit order markets, where the fastest trader who submits a marketable order gets the trade. An alternative allocation is to let some **M** share **S**'s offer, for example, equally. Such split does not affect the intuition behind MM trades, the associated price pressure, and the inference problem. In period 1, all **M**-owners who hold non-zero positions seek to resell the asset, either in the market, privately, or to another **M**. Each of them makes its decision based on its signals of E_0 and E_i (to achieve this, one needs to generalize reselling opportunities such that each **M** draws private signals about E_0). This way, some **M**-owners might target the **EB** in the market, while others might target at other **M** (MM trades). Multiple trades will occur, with some prices pressured and some not. Realizations of period 1 market activity will become more complicated than what table 1 illustrates, but the inference problem still exists. When the period 0 allocation is not equally split across **M**, the inference problem is further complicated because **S** must also take into account trade size when parsing period 1 activity.

S moving first in period 0. An alternative formulation of period 0 is to let **M** move first and post demand schedule to **S**. This, however, does not change the analysis if one maintains the assumptions that information is symmetric in period 0 and that all **M** are identical and perfectly competitive. **M** will then Bertrand compete to bid the asset by posting at their common reservation value (lemma 2). Observing such a demand schedule, **S** will choose the optimal quantity to sell (section 4.3.3).

6.4 Voluntary disclosure

Section 5.5 studies a disclosure policy to flag MM trades, in order to improve welfare. An alternative, milder policy is to allow **M** voluntarily flag their orders, for example, in a limit order market. This way, an inter-middlemen trades will carry two flags, while a middleman-buyer/investor trade will have one. The analysis in section 5.5 can be thought of as an extreme case where **M** always voluntarily flag their orders. The other extreme case is where no middlemen flags voluntarily, as in the benchmark case studied in section 4. Proposition 6 shows that welfare is improved by the disclosure (the first polar case). Therefore, voluntary flagging also improves welfare, as it is a convex combination of the two extreme cases.

It remains to understand why **M** will, at least in some states of the world, have incentive to voluntarily flag orders. In the current version of the model, **M** are indifferent between flagging or not, because they always get zero expected profit ex ante. To formally understand why voluntary flagging might be incentive compatible, one needs to build model extensions beyond the scope of this paper, such as to allow curvature in **M**'s preference, or to construct a recursive formulation of the reallocation problem. We suggest the following intuition for future research. When reselling opportunities are scarce in the market, **M** suffer from **S**'s overloading, and therefore, there is incentive to flag MM trades so that **S** refrains from further loading them (see lemma 6). A similar intuition applies to buyers. When **M** have trouble reselling, flagging MM trades reveals the price pressure and attracts potential buyers to help alleviate **M**'s inventory positions.

7 Conclusion

High frequency traders (HFT), to the extent that they serve as middlemen (**M**) in electronic securities markets, not only intermediate between fundamental investors but also trade with each other to resell the asset. When an **M** has a small reselling opportunity, it seeks to trade with other **M** for better reselling opportunity. This trade is referred to as "inter-middlemen (MM) trade". Such MM trades are motivated by middlemen's (ex post) heterogeneous reselling opportunities, and reallocate the asset efficiently within the middleman sector.

Price pressure arises with MM trades because the buying **M** needs compensation for its expected cost of carry. In contrast, when **M** successfully resells the asset to an (investor) buyer, the price is unpressured. Hence, uninformed investors face an inference problem about the type of the observed trade.

A theoretical model characterizes the above features. A low valuation seller (**S**) holds a large position of the asset and seeks to trade with high valuation buyers (**B**). Two frictions stand in the way of an efficient transfer: investment inattention and information asymmetry. More **M** help the seller find more (attentive) **B**, but also worsen the learning of the uninformed **S** due to the potential inference problem described above. The model provides rich implications on market conditions and welfare. As the size of the middleman sector increases, MM trades become more likely and **S**'s learning ability worsens. To avoid her expected cost of carry and the expected adverse selection cost, **S** overloads **M** in terms of social welfare. When each **M** has a very small reselling opportunity, it is optimal to have as few **M** as possible for welfare. Under a disclosure policy that solves the inference problem, welfare improves.

Appendix A: Notation summary

- a, units of the asset. When subscripted, a_t refers to the amount of the asset in period t.
- **B**, the (representative) buyer.
- c, units of the numéraire good.
- **EB**, a buyer who arrives early (in period 1) and has the same preference as the (late) **B**.
- E_i, indicator of whether an early buyer (**EB**) arrives. It is a Bernoulli random variable with realization 1 for arrival and 0 for no arrival. For $i = 0$, it refers to the **EB** arriving in the market. For $i \in \{1, \dots, m\}$, it refers to the **EB** that privately arrives to middleman **M**i.
- $k(a)$, the cost for **S** to carry a units of the asset to period 3. $k(a) = a^2/2$ for $a \geq 0$ and $k(a) = 0$ for $a < 0$.
- $l(q)$, the expected social loss, as a function of q, the amount of the asset that **S** sells to **M** in period 0.
- m, number of middlemen.
- **M**i, middleman i, $i \in \{1, \dots, m\}$. In particular,

M1 refers to the asset owning middlemen in period 1.

- p_t, price in period t. It can refer to either a price quote or a transaction price.
- q, quantity of the asset supplied by **S** in period 0.
- RO, acronym for reselling opportunity.
- $s_t(p)$, the supply schedule posted by **S** in period t.
- **S**, the seller.
- $u(c, a)$, utility function of an agent who consumes c units of the numéraire good and a units of the special good.
- $w(m)$, social welfare as a function of the number of middlemen.
- Z, the preference shock that strikes all investors equally in the economy. Unconditionally it is uniformly distributed on $[0, 1]$.
- $\bar{z}$, the unconditional expectation of Z, i.e. $\bar{z} = \mathbb{E}Z$.

Probabilities:

- $\alpha(m)$, the probability of all reselling opportunities being low. That is, $\alpha(m) = \mathbb{P}(\Theta_i = 0, \forall i \in \{0, 1, \ldots, m\})$.
- $\beta(m)$, the probability of an MM trade. That is, $\beta(m) = \mathbb{P}(\Theta_0 = \Theta_1 = 0, \sum_{i=2}^{m} \Theta_i > 0)$.
- γ, the probability of no market activity in period 1. That is, $\gamma = \mathbb{P}(\Theta_0 = 0, \Theta_1 = \phi)$.
- ϕ, the abundance of middlemen's reselling opportunity, such that $\Theta_i = \phi$ or 0 with equal probability ex ante for all $i \geq 1$.
- Θ_i, the probability that **EB** i arrives, such that $\mathbb{P}(\mathbb{E}_i = 1|\Theta_i) = \Theta_i$.

Appendix B: Proofs

Lemma 1

Proof. If $\theta_0 = 1$, $\mathbb{E}(E_0 = 1|\theta_0) = 1$, i.e. an **EB** is to arrive in the market with certainty. **M**1 can then sell all q units to **EB** by posting a take-it-or-leave-it offer at the **EB**'s reservation value, z, for all q units (because **EB** has linear preference), getting revenue qz. If $\theta_0 = 0$ and $\theta_1 = 1$, no **EB** is to arrive in the market, but if **M**1 tries its private channel to resell, with probability $\mathbb{E}(E_1 = 1|\theta_1) = \phi$ an **EB** will arrive. In expectation, therefore, **M**1 gets $(1 - \phi) \cdot 0 + \phi qz = \phi qz$. Finally, if $\theta_0 = \theta_1 = 0$, **M**1 can only try reselling to another **M**. The best **M**1 can do is to post a take-it-or-leave-it offer in the market at the potential buyer **M**'s reservation value (per unit of the asset), which is ϕz, the same as that of **M**1 when **M**1 has θ_1 (setting $q = 1$). In this last case, the expected revenue is $\mathbb{P}(\sum_{i=2}^{m} \theta_i > 0)\phi qz$, where the probability is strictly less than 1 because there is always non-zero probability for all other **M** to draw a bad signal about their ROs. (If $m = 1$, **M**1 is indifferent between posting or not such a take-it-or-leave-it offer because there is no other **M**, i.e. $\mathbb{P}(\sum_{i=2}^{1} \theta_i > 0) = 0$.)

Note that the expected reselling revenue conditional on respective realizations of θ_0 and θ_1 can be ranked as $qz > \phi qz > \mathbb{P}(\sum_{i=2}^{m} \theta_i > 0)\phi qz$. Therefore, **M**1's decision described in the lemma is incentive-compatible. □

Lemma 2

Proof. From the proof of lemma 1 above, one can see that in period 1, a non-owner **M** has zero expected profit (because the owner, **M**1, sets price at their reservation value when MM trading). On the other hand, **M**1 has positive expected profit from exploiting either **EB** or other **M**'s surplus. Therefore, in period 0, all **M** has the incentive to become the asset owner later in period 1. Conditional on being selected by nature to be the asset owner, the **M** gets the sum of expected reselling revenues (per unit of the asset) conditional on the three sets of realizations of Θ_0 and Θ_1: $1/2 \cdot z + \gamma\phi z + \beta(m)\phi z$, where the probability weights γ and $\beta(m)$ are defined as in the gamma. Simplifying the expression, one gets the reservation value of all **M** in period 0, and the strategy to trade against **S**'s supply immediately follows. □

Lemma 3

Proof. **S** observes a trade at price p_1 from period 1 market activity. The updated distribution of Z is binomial as stated in equation (4). Selling anything at prices other than the high or the low prices is dominated, because **B** observes Z and will be willing to buy any amount of the asset at or below Z (see section 4.1). Therefore, **S** only needs to choose the amount to supply at the two prices:

$$\max_{s_l,s_h}(1-\hat{\beta})\left[p_1 a_2 - p_1 k(a_2 - s_l)\right] + \hat{\beta}\left[\frac{p_1}{\phi}a_2 - \left(\frac{p_1}{\phi} - p_1\right)s_l - \frac{p_1}{\phi}k(a_2 - s_h)\right],$$

where the choice variables s_l and s_h are her supply at p_1 and p_1/ϕ respectively. The optimization problem has unique solution given by the first order conditions, and the second order condition is satisfied by the convexity of $k(\cdot)$. Off the corners, we have $s_h = a_2$ and $s_l = a_2 - k_a^{-1}(\hat{a}(m))$ $(< s_h)$, where $k_a^{-1}(\cdot)$ denotes the inverse function of the first order derivative of $k(\cdot)$, and $\hat{a}(m)$ is as defined in the lemma. There is, however, one implicit bound: $s_l \geq 0$, because when **S** supplies negatively, she essentially becomes a buyer and wants to buy at the low price, at which no one is selling the asset. Therefore, **S** sets $s_l = \max\{0, a_2 - k_a^{-1}(\hat{a}(m))\}$. By the quadratic functional form of $k(\cdot)$, one can easily simplify the expression and get **S**'s supply schedule as stated in equation (5). □

Lemma 4

Proof. Without learning anything, **S** keeps the prior distribution of Z that it is uniform on $[0, 1]$. Holding a_2 units of the asset, **S** solves a supply function to maximize her expected utility subject to **B**'s adverse selection:

$$\max_{s(\cdot)}\mathbb{E}\left[\int_0^Z (s(Z) - s(p))\mathrm{d}p + Z \cdot ((a_2 - s(Z)) - k(a_2 - s(Z)))\right],$$

subject to the boundary condition $s(1) = a_2$, which economically says that **S** sells all her position at price $\sup Z = 1$. (**S** will not keep anything unsold because $k(a) > 0$ for all $a > 0$.) Note that the integral term gives the adverse selection cost for a realization of Z. Write $T(Z) := \int_0^Z s(p)\mathrm{d}p$, and use the uniform density of Z to rewrite the optimization as (after some simplification) $\max_{s(\cdot)}\int_0^1 L(Z, s(Z), T(Z)) \cdot 1\mathrm{d}Z$, where $L(Z, s(Z), T(Z)) := Z \cdot a_2 - T(Z) - Z \cdot k(a_2 - s(Z))$. Apply Euler-Lagrange method to get the differential equation $\partial L_s/\partial Z = L_T$, which together with the boundary condition above solves $s(Z) = a_2 - k_a^{-1}(1/Z - 1)$. The second order condition (Legendre condition) is also satisfied by the convexity of $k(\cdot)$. One can also easily verify that $s(Z)$ is increasing in Z on $[0, 1]$. There is, however, one bound on non-negativity: $s(Z) \geq 0$, because even if **S** wants to supply negatively (to buy), there is no other sellers in the market. Let $s(z) \geq 0$, for a given a_2, one can get $z \geq 1/(1 + k_a(a_2)) = 1/(1 + a_2) =: z^*(a) \in (0, 1)$, such that for $z^*(a_2) \leq Z \leq 1$, $s(Z) = a_2 - (1/Z - 1)$ and for $0 \leq Z < z^*(a_2)$, $s(Z) = 0$.

Finally, one can compute the expected utility of **S** by plugging the solved supply function above into the maximization problem and evaluate the integrals. This gives $\mathbb{E}u^{\mathbf{S}}_{2,\mathrm{nl}}(a_2; Z) = \ln(1 + a_2)/2$. □

Lemma 5

Proof. To begin with, observe that the optimization problem has two different expressions, depending on whether $a_2 = a_0 - q \lesseqgtr \hat{a}(m)$ (see equations 6 and 7). To avoid the triviality where a_2 is always less than $\hat{a}(m)$ for all m, assume that a_0 is sufficiently large:

$$a_0 > \lim_{m\to\infty} \hat{a}(m) = \hat{a}(\infty) = \frac{1-\phi}{2\phi^2}, \tag{A1}$$

which also implies $a_0 > \hat{a}(m)$ for all m because $\hat{a}(m)$ is strictly increasing in m.

We now solve the optimization problem for two cases: 1) $a_0 - q \geq \hat{a}(m)$ and 2) $a_0 - q < \hat{a}(m)$ respectively, as if no bound binds.

Case 1. ($a_0 - q \geq \hat{a}(m)$.) After substituting in the various expressions of $u_2^{\mathbf{S}}(\cdot)$, especially equations (6) and (7) for $a_2 \geq \hat{a}(m)$, the optimization problem can be rewritten as: $u_{0,\mathrm{I}}^{\mathbf{S}}(q) := \max_q p_0(m)q + \bar{z} \cdot (a_0 - q) - \gamma\left(a_0 - q - \ln(a_0 - q + 1)\right)/2 - \phi^2 \bar{z} k(\hat{a}(m))/2 - \beta(m)(1-\phi)\bar{z} \cdot (a_0 - q - \hat{a}(m))$. From the first order condition, one can get

$$q_{\mathrm{I}}(m) = a_0 - \left(\frac{1}{\phi - 2^{-(m-1)}} - 1\right). \tag{A2}$$

It is easy to show that $u_{0,\mathrm{I}}^{\mathbf{S}}(q)$ is strictly concave in q, and the second order condition is satisfied.

Case 2. ($a_0 - q < \hat{a}(m)$.) As in the above case, substitute in the expressions of $u_2^{\mathbf{S}}(\cdot)$, this time equations (6) and (7) for $a_2 < \hat{a}(m)$. The optimization problem becomes: $u_{0,\mathrm{II}}^{\mathbf{S}}(q) := \max qp_0(m)q + \bar{z} \cdot (a_0 - q) - \gamma\left(a_0 - q - \ln(a_0 - q + 1)\right)/2 - \phi^2 \bar{z} \cdot k(a_0 - q)/2$. From the first order condition, one can get

$$q_{\mathrm{II}}(m) = a_0 - \left(\frac{\sqrt{b(m)^2 + 2\gamma\phi^2} - b(m)}{\phi^2} - 1\right), \tag{A3}$$

where $b(m) := p_0(m)/\bar{z} + \gamma - 1 - \phi^2/2$. The second order condition can be easily verified that $u_{0,\mathrm{II}}^{\mathbf{S}}(q)$ is strictly concave in q.

We next check the corners and boundaries of the above solutions. First, at $q = a_0$, case 2 applies. Evaluate the derivative: $\partial u_{0,\mathrm{II}}^{\mathbf{S}}(q = a_0)/\partial q = p_0(m) - \bar{z}$. Observe from equation 11 that $p_0 m < \bar{z}$. Therefore, by strict concavity, the upper bound $q = a_0$ never binds.

Second, at $q = 0^+$, case 1 applies.[33] Evaluate the derivative to get: $\partial u_{0,\mathrm{I}}^{\mathbf{S}}(q = 0^+)/\partial q = [\phi - 1/(1 + a_0) - 2^{-(m-1)}]/4$, which is strictly increasing in m, the number of middlemen. If $\phi \leq 1/(1 + a_0)$, i.e. if $a_0 \leq (1-\phi)/\phi$, the above derivative is always negative, and by the strict concavity of $u_{0,\mathrm{I}}^{\mathbf{S}}(q)$, the lower bound $q = 0^+$ always binds. That is, if the initial position a_0 is too small, **S** always sells virtually nothing in period 0. To ensure that **S** sells more than virtually nothing (0^+) of the asset to **M** in period 0, therefore, assume:

$$a_0 > \frac{1-\phi}{\phi}, \tag{A4}$$

[33] Note that in order for case 1 to apply, **S** must sell strictly more than 0 units of the asset to **M** in period 0, because otherwise no market activity in period 1 will occur and **S** will learn nothing from it. We consider an infinitesimally small, yet positive amount, denoted by 0^+. Selling strictly nothing makes **S** worse off than selling 0^+ because the latter gives **S** an informational advantage: There is non-zero probability that the fundamental value can be fully or partially revealed from period 1 market activity. In practice, this 0^+ units can be the smallest amount, e.g. 1 share/contract of the traded asset, allowed by the exchange to trade.

which constitutes part of assumption 3. Under this assumption, it is easy to show that there always exists $m_0 = 1 + \ln(\phi - 1/(1 + a_0))/\ln(1/2) > 1$ such that $u^{\mathbf{S}}_{0,\mathrm{I}}(q = 0^+; m_0) = 0$. By monotonicity in m, the lower bound $q = 0^+$ binds if and only if $m < m_0$.[34]

Finally, let us turn to the interim bound that switches the solution from $q_{\mathrm{I}}(m)$ to $q_{\mathrm{II}}(m)$. At $q = a_0 - \hat{a}(m)$, **S**'s optimized expected utility is continuous and first order differentiable: $u^{\mathbf{S}}_{0,\mathrm{I}}(q) = u^{\mathbf{S}}_{0,\mathrm{II}}(q)$ and $\partial u^{\mathbf{S}}_0(q)/\partial q = y(m)/4$, where $y(m) := \phi - 1/(1 + \hat{a}(m)) - 2^{-(m-1)}$, defined for $m \geq m_0$. It is easy to show that the auxiliary function $y(m)$ is strictly increasing in m. At the left-end, $y(m = m_0) = 1/(1 + a_0) - 1/(1 + \hat{a}(m)) < 0$ because $a_0 > \hat{a}(m)$ by equation (A1). At the right-end, $y(\infty) = \phi \cdot (1 - \phi)(1/2 - \phi)/(1 - \phi + 2\phi^2)$, which is positive if and only if $\phi < 1/2$. Therefore, if $\phi < 1/2$, by continuity, there exists $\hat{m}$ ($> m_0$) such that $\partial u^{\mathbf{S}}_0(a0 - \hat{a}(\hat{m})) = y(\hat{m})/4 = 0$. Then because $y(m)$ is strictly increasing in m, we have $\partial u^{\mathbf{S}}_0(q = a_0 - \hat{a}(m))/\partial q > 0$ for all $m > \hat{m}$ (and vice versa), i.e **S** will choose a q larger than $a_0 - \hat{a}(m)$ (case 2). By the strict concavity of $u^{\mathbf{S}}_0(q)$, therefore, for all $m > \hat{m}$ it is optimal to have $q = q_{\mathrm{II}}(m)$ and for all $m_0 \leq m \leq \hat{m}$, it is optimal to switch to $q = q_{\mathrm{I}}(m)$.

For $\phi \geq 1/2$, however, such a threshold $\hat{m}$ does not exist. For clarity of notation, however, define $\hat{m} = \infty$ if $\phi \geq 1/2$. That is, $q = q_{\mathrm{I}}(m)$ is optimal for all $m \geq m_0$ if $\phi \geq 1/2$. □

Lemma 6

Proof. Compare **S**'s two problems in period 0, with (section 5.5) and without disclosure (section 4.3.3). In both situations, the marginal utility of selling early is the same, $p_0(m)$. The marginal utility of selling late under disclosure is $(a_0 - q)\bar{z} - \gamma \cdot (a_0 - q + \ln(a_0 - q))/2$. From the first order conditions in the proof of lemma 5 above, for both cases and for all q, the marginal utility of selling late without disclosure is smaller than the above expression. As optimality requires q to equate the marginal utilities of selling early and late, off the corner, the solution q under disclosure is always strictly smaller than that without disclosure. When the lower bound $q = 0^+$ binds in the no-disclosure situation, it must also bind for the with disclosure situation. □

Proposition 1 (on page 5 and continued on page 19)

Proof. By lemma 1, $\theta_0 = \theta_1 = 0$ and for at least one $i \in \{2, \ldots, m\}$ ($m \geq 2$), $\theta_i = \phi$, because an MM trade can happen. Without MM trade, **M**1 holds q units of the asset and cannot resell, and the aggregate expected utility for all **M** is simply zero. After an MM trade, **M**1 gets the buying **M**'s surplus, which is $\phi z q$. The aggregate expected utility for all **M** becomes also $\phi z q$, which is strictly larger than zero. Hence, an MM trade improves the allocational efficiency in middlemen.

From lemma 1, conditional on a realization z, an MM trade occurs at price $\phi z < z$, hence, a price pressure.

Ex ante, an MM trade happens with probability $\beta(m) = \mathbb{P}(\Theta_0 = \Theta_1 = 0, \sum_{i=2}^{m})$. From the expression of $\beta(m)$ it can be seen that it is strictly increasing in m. □

Proposition 2

Proof. This proposition immediately follows lemma 3. In particular, when $a_2 > \hat{a}(m)$, **S** posts $a_2 - \hat{a}(m)$ units of the asset at the observed price, p_1, which is the MM trading price ϕz. **S** reinforces the price pressure

[34] It should be noted that the solution m_0 is not necessarily an integer. In practice, however, the number of middlemen can only take integer values. There always exists a positive integer that is nearest to m_0 because $m_0 > 1$. To avoid burdening the notation further, we do not introduce additional symbols to differentiate between integer and non-integer m. Instead, the number of middlemen m should always be understood to refer to one of the nearest positive integers of m.

for $a_2 - \hat{a}(m)$ units of the asset. When $a_2 \leq \hat{a}(m)$, **S** sells nothing at p_1 (but all at p_1/ϕ). Should the observed trade be an MB trade ($z = p_1$), the market breaks down. □

Proposition 3

Proof. To prove this proposition, we evaluate the first order derivative of social loss function $l(q)$ (equation 13) with respect to q at the optimal q as solved in lemma 5. Only the cases where $m \geq m_0$ where **S** actually loads things to middlemen are considered.

If $m_0 \leq m \leq \hat{m}$, equation (13) simplifies to $l(q) = (1 - p_0/\bar{z})q - \gamma \cdot [\ln(1 + a_0 - q) + 1/(1 + a_0 - q) -] - \phi^2 k(\hat{a}(m))/2$ and $l_q(q) = (1 - p_0/\bar{z}) - \gamma \cdot [1/(1 + a_0 - q) - 1/(1 + a_0 - q)^2]$. Substitute $q = q_{\mathrm{I}}(m)$ into the first order derivative to get, after some simplification, $l_q(q_{\mathrm{I}}(m)) = (1 - p_0/\bar{z}) - (\phi - 2^{-(m-1)}) \cdot (1/4 - \phi/4 + 2^{-(m+1)}) > 1 - p_0/\bar{z} - (1/4 - \phi/4 + 2^{-(m+1)})$, because $(1/4 - \phi/4 + 2^{-(m+1)}) > 0$ and $(\phi - a^{-(m-1)}) \in (0, 1)$. After expanding $p_0/\bar{z}$, we have $l_q(q_{\mathrm{I}}(m)) > \phi\beta(m) > 0$.

If $m > \hat{m}$ (and $\hat{m} < \infty$), equation (13) simplifies to $l(q) = (1 - p_0/\bar{z})q - \gamma \cdot [\ln(1 + a_0 - q) + 1/(1 + a_0 - q) -] - \phi^2 k(a_0 - q)/2$ and $l_q(q) = (1 - p_0/\bar{z}) - \gamma \cdot [1/(1 + a_0 - q) - 1/(1 + a_0 - q)^2] - \phi^2 (a_0 - q)/2$. Substitute the first order condition from case 2 of the proof of lemma 5 into $l_q(q)$ above to get, after some simplification, $l_q(q_{\mathrm{II}}(m)) = \gamma \cdot [1 - 1/(1 + a_0 - q_{\mathrm{II}}(m))]^2 > 0$

Therefore, in either case, the social loss is strictly increasing at **S**'s optimal period 0 supply; social welfare can be improved if **S** reduces the supply slightly. □

Proposition 4 and corollary 1

Proof. We begin with $m_0 \leq m \leq \hat{m}$, such that case 1 of the proof of lemma 5 applies. Compute the partial derivative of **S**'s expected utility with respect to m using envelope theorem[35]:

$$\frac{\partial u^{\mathbf{S}}_{0,\mathrm{I}}(q_{\mathrm{I}}(m))}{\partial m} = \frac{\partial p_0(m)}{\partial m} q_{\mathrm{I}}(m)\bar{z} - \frac{\partial \beta(m)}{\partial m}(1 - \phi) \cdot (a_0 - q_{\mathrm{I}}(m) - \hat{a}(m))\,\bar{z}, \tag{A5}$$

where the first term on the right hand side represents the marginal increase in middlemen's reservation value, and the second term the marginal increase in **S**'s adverse selection as MM trades become more likely (see proposition 1). (A useful trick in the above computation is to note $\hat{a}(m)\beta'(m) = \hat{a}'(m)\beta(m)$.)

Similarly, when $m < m_0$, $q = 0^+$ and one can show that only the increase in the adverse selection remains, because, naturally, **S** does not care about **M**'s reservation value when selling virtually nothing to them. When $m > \hat{m}$, applying envelope theorem to compute the partial derivative of $u^{\mathbf{S}}_{0,\mathrm{II}}(q_{\mathrm{II}}(m))$ with respect to m results in only $\frac{\partial p_0(m)}{\partial m} q_{\mathrm{I}}(m)$ because $a_0 - q = q_2 < \hat{a}(m)$ and **S** is no longer subject to the adverse selection.

Corollary 1 immediately follows the economic consequences from the above two effects: an additional **M** (weakly) increases the marginal utility of selling early and (weakly) reduces the marginal utility of selling late. Mathematically, one can compute the derivatives of the optimal q with respect to m using the expressions (A2) and (A3). □

Proposition 5 and corollary 2

Proof. If ($1 \leq$) $m < m_0$, $q = 0^+$ and $l(q(m))/\bar{z} = \gamma \cdot (\ln(1 + a_0) + 1/(1 + a_0) - 1) - \phi^2 k(\hat{a}(m))/2$. The marginal effect of m is $\partial l(q(m))/\partial m = \phi^2 \partial k(\hat{a}(m))/\partial m \bar{z}/2 = \beta'(m)(1 - \phi)\hat{a}(m)\bar{z} > 0$, which is the second effect of

[35] Again, m is treated as if it has continuous support on $[m_0, \hat{m}]$, although in practice, the number of middlemen can only admit integer values. See also footnote 34.

proposition 5. The other two effects in this case are both zero because **S** loads $q = 0^+$ to **M** in period 0. As the (expected) social loss increases with m, social welfare reduces in this range.

From the proof of proposition 2, when $m_0 \leq m \leq \hat{m}$, $l(q) = (1 - p_0/\bar{z})q - \gamma \cdot [\ln(1 + a_0 - q) + 1/(1 + a_0 - q) -] - \phi^2 k(\hat{a}(m))/2$. The marginal effect of m, by chain rule, is

$$\frac{\partial l(q_{\mathrm{I}}(m), m)}{\partial m} = \frac{\partial l}{\partial q}\Big|_{q_{\mathrm{I}}(m)} \cdot \frac{\partial q_{\mathrm{I}}(m)}{\partial m} + \frac{1}{2}\phi^2 \frac{\partial k(\hat{a}(m))}{\partial m} - \frac{\partial p_0(m)}{\partial m} \cdot q_{\mathrm{I}}(m) \tag{A6}$$

The first term is the overloading effect, the second is the increase in **S**'s expected cost of carry, and the third is the reduction in social loss due to increased overall reselling opportunities. Using equation (A2), one can evaluate equation (A6) as $\partial l(q_{\mathrm{I}}(m))/\partial m = \bar{z}\beta'(m)y(m)$, where an auxiliary function is defined as $y(m) := (a_0 - q_{\mathrm{I}}(m))^2 - \phi q_{\mathrm{I}}(m) + (1 - \phi)\hat{a}(m) + (1 - \phi)\gamma\beta(m)/(\gamma\phi - \alpha(m))^2$. In particular, this auxiliary function is strictly decreasing in m:

$$y'(m) = -\underbrace{[2(a_0 - q_{\mathrm{I}}(m)) - (1 - \phi)]}_{>2(1-\phi)/\phi} \overbrace{q_{\mathrm{I}}'(m)}^{>0} + (1 - \phi)\hat{a}'(m) + (1 - \phi)\gamma\beta'(m) \underbrace{\frac{-\gamma \cdot (1 - \phi) - \beta(m)}{(\gamma\phi - \alpha(m))^3}}_{<0}$$

$$< -\left(\frac{2(1 - \phi)}{\phi} - (1 - \phi)\right) \underbrace{\frac{\gamma\beta'(m)}{(\gamma\phi - \alpha(m))^2}}_{>\beta'(m)/(\gamma\phi^2)} + \frac{2(1 - \phi)^2}{\phi^2}\beta'(m) < -\frac{(1 - \phi)\beta'(m)}{\gamma\phi^3}\left[1 + \frac{1 - \phi}{2} + \frac{(1 - \phi)^2}{2}\right] < 0.$$

With this, the following property about the second order derivative of $l(q_{\mathrm{I}}(m))$ with respect to m is obtained: $\partial^2 l/\partial m^2 = \bar{z}\beta''(m)y(m) + \bar{z}\beta'(m)y'(m) > \bar{z}\beta''(m)y(m)$. Suppose there exists a (local) extreme point such that $\partial l/\partial m = \bar{z}\beta'(m)y(m) = 0$, implying $y(m) = 0$ because $\beta'(m) > 0$ for all m. By the above property, at such an extreme point, the second order derivative of $l(q_{\mathrm{I}}(m))$ is negative, implying a (local) maximum. That is, there can be no (local) minimum of $l(q_{\mathrm{I}}(m))$.

When $m > \hat{m}$ (and $\hat{m} < \infty$), $l(q) = (1 - p_0/\bar{z})q - \gamma \cdot [\ln(1 + a_0 - q) + 1/(1 + a_0 - q) -] - \phi^2 k(a_0 - q)/2$. One can similarly compute the marginal effect of m on social loss using chain rule to derive an equation similar to equation (A6) above. The difference in this case is that there will be no marginal increase in **S**'s expected cost of carry, because when $m > \hat{m}$, **S**'s period 2 supply does not change with m. The effects 1) and 3) as stated in proposition 5 still remain. With equation (A3), one can evaluate the marginal effect of m on social loss as $\partial l(q_{\mathrm{II}}(m))/\partial m = \phi\bar{z}\beta'(m)y(m)$, where a new auxiliary function is defined as $y(m) := -q_{\mathrm{II}}(m) + \gamma \cdot (a_0 - q_{\mathrm{II}}(m))^2(b(m) + 2\gamma\phi^2)^{-1/2}$ with $b(m)$ defined as in equation (A3). Again, $y(m)$ is strictly decreasing in m: $y'(m) = -q_{\mathrm{II}}'(m) - (a_0 - q_{\mathrm{II}}(m))^2\gamma \cdot b(m)b'(m)/(1 + a_0 - q_{\mathrm{II}}(m))/(b(m)^2 - 2\gamma\phi^2)^{3/2} - 2\gamma q_{\mathrm{II}}'(m)(a_0 - q_{\mathrm{II}}'(m))/(1 + a_0 - q_{\mathrm{II}}(m))^2/\sqrt{b(m)^2 + 2\gamma\phi^2} > 0$. By the same argument as before, any (local) extrema of $l(q_{\mathrm{II}}(m))$ on $m > \hat{m}$ cannot be minimum.

Combining the above two cases of $m \geq m_0$ and noting that $l(q(m))$ is continuous and twice-differentiable at $m = \hat{m}$, one can conclude that there is no local minimum points on $m \geq m_0$. There can be at most one local maximum. Put alternatively, $l(q(m))$ is a quasi-concave in m. Recall the identity between social welfare and the social loss $w(m) = a_0\bar{z} - l(q(m))$. Therefore, to sum up, social welfare $w(m)$ is strictly decreasing on $1 \leq m < m_0$, and is quasi-convex on $[m_0, \infty)$. □

Corollary 3

Proof. Compare the welfare difference between two polar cases of $m = 1$ and $m = \infty$. At $m = 1$, social welfare is $w(m = 1) = a_0\bar{z} - \gamma(\ln(1 + a_0) + 1/(1 + a_0) - 1)/2$ (note that $\hat{a}(m = 1) = 0$). Consider the following two cases for $w(m = \infty)$: $\phi < 1/2$ and $\phi \geq 1/2$.

If $\phi < 1/2$, by lemma 5, $\hat{m} < \infty$, and $w(m = \infty) = a_0\bar{z} - l(q_{\text{II}}(\infty))$. After some calculation, the welfare change from $m = 1$ to $m = \infty$ can be written as

$$\begin{aligned} w(m = 1) - w(m = \infty) &= (\bar{z} - p_0(\infty))q_{\text{II}}(\infty) + \frac{1}{2}\overbrace{\bar{z}\phi^2 k(a_0 - q_{\text{II}}(\infty))}^{>0} - \frac{\gamma}{2}y(q_{\text{II}}(\infty)) \\ &> (1 - \phi)q_{\text{II}}(\infty) - \frac{1}{2}y(q_{\text{II}}(\infty)), \end{aligned}$$

where an auxiliary function is defined as $y(q) := \ln(1 + a_0) - \ln(1 + a_0 - q) + 1/(1 + a_0) - 1/(1 + a_0 - q)$. A useful property of the auxiliary function $y(q)$ is that $y(q) \leq q$ for all $q \geq 0$ because $y(0) = 0$ and $y'(q) < 1$ for all $q \geq 0$. Using this property and recalling that $\phi < 1/2$, we have $w(m = 1) - w(m = \infty) > [q_{\text{II}}(\infty) - y(q_{\text{II}}(\infty))]/2 > 0$. That is, in the case of $\phi < 1/2$, $w(m = 1) > w(m = \infty)$.

If $\phi \geq 1/2$, by lemma 5, $\hat{m} = \infty$ and $w(m = \infty) = a_0\bar{z} - l(q_{\text{I}}(\infty))$. After some calculation, the welfare change can be written as

$$w(m = 1; \phi) - w(m = \infty; \phi) = (1 - \phi)q_{\text{I}}(\infty; \phi) - \frac{1}{2}y(q_{\text{I}}(\infty; \phi)) + \frac{1}{8}\frac{(1 - \phi)^2}{\phi^2}.$$

At $\phi \to 1$ (from below), $w(m = 1) - w(m = \infty) \to -y(q_{\text{I}}(\infty; 0^+)) < 0$. At $\phi = 1/2$, $w(m = 1) - w(m = \infty) = 1/8 + [(a_0 - 1) - y(a_0 - 1)]/2 > 1/8 > 0$, because $a_0 > (1 - \phi)/\phi \geq 1$ by equation A4 and by $\phi = 1/2$. Therefore, by continuity, there exists some $\phi^* \in (1/2, 1)$ such that $w(m = 1; \phi^*) - w(m = \infty; \phi^*) = 0$ and for all $\phi > \phi^*$, $w(m = 1; \phi) < w(m = \infty)$. (The uniqueness of ϕ^* can be guaranteed by assuming a sufficiently large a_0. We do not prove this here.) □

Proposition 6 and corollary 4

Proof. To begin with, consider the social loss function under the disclosure policy: $\tilde{l}(\tilde{q}) = (\bar{z} - p_0)\tilde{q} + \gamma \cdot (\ln(1 + a_0 - \tilde{q}) + 1/(1 + a_0 - \tilde{q}) - 1)/2$. It is strictly increasing in the period 0 supply of **S**: $\partial\tilde{l}/\partial\tilde{q}/\bar{z} = (1 - p_0/\bar{z}) - \gamma \cdot (1 - 1/(1 + a_0 - \tilde{q}))/(1 + a_0 - \tilde{q})$. By optimality, as stated in the proof of lemma 6 above, $\tilde{q}$ should equate **S**'s marginal utility of selling early and late, i.e. $p_0 = (1 - \gamma)\bar{z} + \gamma\bar{z}/(a_0 - \tilde{q} + 1)$. Using this equation, one can easily show that $\partial\tilde{l}/\partial\tilde{q} > 0$ by noting $1 + a_0 - \tilde{q} > 1$.

Next, compare the social losses with and without the disclosure. For $m < m_0$, $q = 0^+$ by lemma 5, and hence $\tilde{q} = 0^+$ by lemma 6. Then with some calculation, $l(q) = \tilde{l}(\tilde{q}) + \bar{z} \cdot \phi^2 k(\hat{a}(m))/2 \geq \tilde{l}(\tilde{q})$, where the equality holds if and only if $m = 1$ (recall that $\hat{a}(m = 1) = 0$). For $m_0 \leq m \leq \hat{m}$, similarly, $l(q) = \tilde{l}(q) + \bar{z} \cdot \phi^2 k(\hat{a}(m))/2 > \tilde{l}(q) > \tilde{l}(\tilde{q})$. For $m > \hat{m}$, $l(q) = \tilde{l}(q) + \bar{z} \cdot \phi^2 k(a_0 - q)/2 > \tilde{l}(q) > \tilde{l}(\tilde{q})$.

Therefore, to sum up, the social loss is larger without disclosure. Equivalently, social welfare is improved by the disclosure. In particular, $w(m = 1) = \tilde{w}(m = 1)$.

Corollary 4 immediately follows proposition 6 and corollary 3 by the continuity of social welfare functions $w(\cdot)$ and $\tilde{w}(\cdot)$ in ϕ. □

References

Akerlof, George A. 1970. "The Market for Lemons: Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84:488–500. 20

Almgren, Robert and Neil Chriss. 2001. "Optimal Execution of Portfolio Transactions." *Journal of Risk* 3:5–39. 4

Attari, Mukarram, Antonio S. Mello, and Martin E. Ruckes. 2005. "Arbitraging Arbitrageurs." *The Journal of Finance* 60:2471–2511. 25

Bertsimas, Dimitris and Andrew W. Lo. 1998. "Optimal Control of Execution Costs." *Journal of Financial Markets* 1:1–50. 4

Biais, Bruno. 1993. "Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets." *The Journal of Finance* 48:157–185. 1

Biais, Bruno, Thierry Foucault, and Sophie Moinas. 2011. "Equilibrium High Frequency Trading." manuscript. Retrieved August 17, 2012 from http://ssrn.com/abstract=1834344. 3, 8

Brogaard, Jonathan A. 2010. "High frequency trading and its impact on market quality." manuscript. Working papger. Retrieved June 16, 2011 from http://ssrn.com/abstract=1641387. 3

Brunnermeier, Markus K. and Lasse Heje Pedersen. 2005. "Predatory Trading." *The Journal of Finance* 60:1825–1863. 25

Cespa, Giovanni and Thierry Foucault. 2012. "Illiquidity Contagion and Liquidity Crashes." Manuscript, HEC Paris. 19

CFTC and SEC. 2010. "Findings regarding the market events of May 6, 2010." Tech. rep., CFTC and SEC. Retrieved June 16, 2011 from http://www.sec.gov/news/studies/2010/marketevents-report.pdf. 2, 3, 19

Chaboud, Alain, Ben Chiquoine, Erik Hjalmarsson, and Clara Vega. 2009. "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market." Manuscript, Federal Reserve Board. 3

Chan, Louis K.C. and Josef Lakonishok. 1995. "The Behavior of Stock Prices Around Institutional Trades." *The Journal of Finance* 50:1147–1174. 4

Duffie, Darrell. 2010. "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *The Journal of Finance* 65:1237–1267. 2, 8

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica* 73:1815–1847. 6, 7, 28

———. 2007. "Valuation in Over-the-Counter Markets." *The Review of Financial Studies* 20:1865–1900. 6

Dunne, Peter, Herauld Hau, and Michael Moore. 2012. "Dealer Intermediation Between Markets." Manuscript. 1

Easley, David, Marcos M. López de Prado, and Maureen O'Hara. 2012. "Flow Toxicity and Liquidity in a High Frequency World." *The Review of Financial Studies* 25(5):1457–93. 1

Easley, David and Maureen O'Hara. 1987. "Price, Trade Size, and Information in Securities Markets." *Journal of Financial Economics* 19:69–90. 1, 28

Foucault, Thierry. 1999. "Order Flow Composition and Trading Costs in a Dynamic Limit Order Market." *Journal of Financial Markets* 2:99–134. 9

Foucault, Thierry and Albert J. Menkveld. 2008. "Competition for Order Flow and Smart Order Routing Systems." *The Journal of Finance* 63:119–158. 3

Gârleanu, Nicolae and Lasse Heje Pedersen. 2012. "Dynamic Trading with Predictable Returns and Transaction Costs." *The Journal of Finance* . 4

Glosten, Lawrence R. and Paul R. Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents." *Journal of Financial Economics* 14:71–100. 1, 3, 28

Grossman, Sanford J. and Merton H. Miller. 1988. "Liquidity and Market Structure." *The Journal of Finance* 43:617–633. 1, 2, 7, 8

Hasbrouck, Joel and Gideon Saar. 2010. "Low-Latency Trading." Manuscript, Cornell University. 3

Hendershott, Terrance and Ryan Riordan. 2011. "High Frequency Trading and Price Discovery." Manuscript, University of California, Berkeley. 3

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance* 66:1–33. 3

Hendershott, Terrence and Ryan Riordan. 2010. "Algorithmic Trading and Information." Manuscript, University of California, Berkeley. 3

Ho, Thomas and Hans R. Stoll. 1983. "The Dynamics of Dealer Markets Under Competition." *The Journal of Finance* 38:1053–1074. 1, 28

Jovanovic, Boyan and Albert J. Menkveld. 2011. "Middlemen in limit-order markets." Tech. rep. Working paper. Retrieved July 17, 2011 from http://ssrn.com/abstract=1624329. 3, 8, 9, 28

Keim, Dnoald B. and Ananth Madhavan. 1995. "Anatomy of the trading process Empirical evidence on the behavior of institutional traders." *Journal of Financial Economics* 37:371–398. 4

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2011. "The flash crash: the impact of high frequency trading on an electronic market." manuscript. Working paper. Retrieved June 12, 2011 from http://ssrn.com/abstract=1686004. 3, 19

Kyle, A. S. 1985. "Continuous auctions and insider trading." *Econometrica* 53:1315–36. 28

Lagos, Ricardo and Guillaume Rocheteau. 2007. "Search in asset markets: market structure, liquidity, and welfare." *The American Economic Review* 97:198–202. 6, 7

———. 2009. "Liquidity in asset markets with search frictions." *Econometrica* 77:403–26. 6, 7

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill. 2011. "Crises and Liquidity in Over-the-Counter Markets." *Journal of Economic Theory* 146:2169–2205. 2, 6

Lyons, Richard K. 1997. "A simultaneous trade model of the foreign exchange hot potato." *Journal of International Economics* 42:275–98. 3, 19, 29

Menkveld, Albert J. 2011. "High frequency trading and the new-market makers." Tech. rep., VU University, Amsterdam. Working paper. Retrieved June 17, 2011 from http://ssrn.com/abstract=1722924. 1

Menkveld, Albert J. and Bart Zhou Yueshen. 2012. "Anatomy of the Flash Crash." Work in progress, VU University Amsterdam. 3

Naik, Narayan Y., Anthony Neuberger, and S. Viswanathan. 1999. "Trade Disclosure Regulations in Markets with Negotiated Trades." *Review of Financial Studies* 12(4):873–900. 3, 19, 29

Obizhaeva, Anna and Jiang Wang. 2005. "Optimal Trading Strategy and Supply/Demand Dynamics." Manuscript, MIT. 4

Pagnotta, Emiliano and Thomas Philippon. 2011. "Competing on Speed." Manuscript, New York University. 6

Petersen, Mitchell A. 2004. "Information: hard and soft." Tech. rep. Working paper. Retrieved July 18, 2011 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.8246&rep=rep1&type=pdf. 9

SEC. 2010. "Concept release on equity market structure." Tech. rep., SEC. Release No. 34-61358; File No. S7-02-10. Retrieved July 17, 2011 from http://www.sec.gov/rules/concept/2010/34-61358.pdf. 1

Stoll, Hans R. 1978. "The Supply of Dealer Services in Securities Markets." *The Journal of Finance* 33(4):1133–1151. 28

Viswanathan, S. and James J.D. Wang. 2004. "Inter-Dealer Trading in Financial Markets." *Journal of Business* 77:987–1040. 3, 19, 29

Table 1: S’s inference from period 1 market activity

This table summarizes the connection between possible events and the market activity in period 1. Row 1 corresponds to the event of a good draw on the common RO ($\theta_0 = 1$, an **EB** is to arrive almost surely) and of a relatively high value of Z ($> \phi$). The market activity resulted from these events is a trade at price $p_1 > \phi$, fully revealing to **S** that $Z = p_1$. Similarly, row 2 has the event of bad draws of all ROs ($\theta_i = 0$, for all $i \in \{0, \ldots, m\}$). The corresponding market activity is a price quote of $p_1 \le \phi$ but no trade, fully revealing to **S** that $Z = p_1/\phi$. The two shaded rows share the same market activity: a trade at price $p_1 \le \phi$, but with different underlying events. It could either be an MM trade (row 3) where $\theta_0 = \theta_1 = 0$ and at least one $\theta_i > 0$, or an MB trade (row 4) where $\theta_0 = 1$ and $Z \le \phi$. The market activity partially reveals Z to **S**. Finally, row 5 corresponds to a draw of $\theta_0 = 0$ and $\theta_1 = \phi$, i.e. **M**1 resells privately. There is no market activity observable to **S**, who does not learn anything in this case.

Events				Probability	Market activity observed by **S**		**S**’s inference
Θ_0 (= E_0)	Θ_1	$\sum_{i=2}^{m} \Theta_i$	Z		price quote	trade?	
1	$\ldots$	$\ldots$	$> \phi$	$(1-\phi)/2$	$P_1\,[= Z] > \phi$	yes	full
0	0	0	$\ldots$	$\alpha(m)$	$P_1\,[= \phi Z] \le \phi$	no	full
0	0	> 0	$\ldots$	$\beta(m)$	$P_1\,[= \phi Z]$ $\Big\} \le \phi$	yes	partial
1	$\ldots$	$\ldots$	$\le \phi$	$\phi/2$	$P_1\,[= Z]$		
0	ϕ	$\ldots$	$\ldots$	γ	(no activity)		no learning

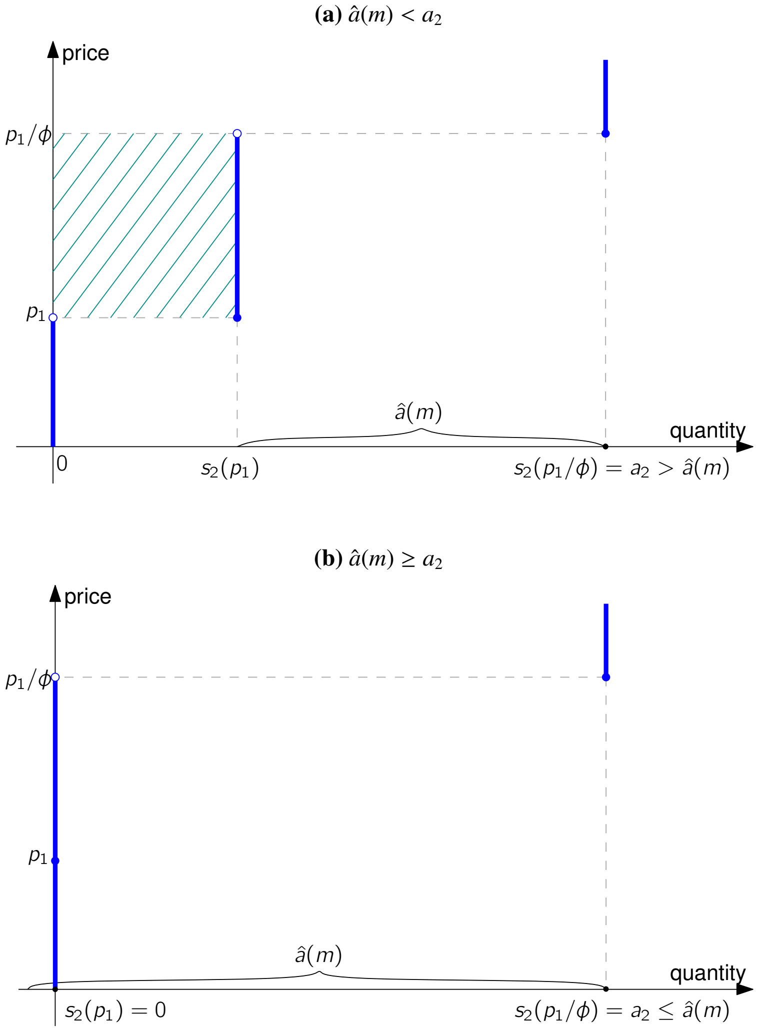
Table 2: S's inference from period 1 market activity, with inter-middlemen flags

This table summarizes the connection between the underlying events and the observed market activity in period 1, under the disclosure policy that flags MM trades. It is similar to table 1, but now inter-middlemen trades are flagged and published in the market along with all other activity. Compared with table 1, the first two rows and the last row have the same respective results for **S**. The inference problem, shaded in table 1, is solved by the flag.

Events				Probability	Market activity observed by **S**			**S**'s inference
$\Theta_0\ (= E_0)$	Θ_1	$\sum_{i=2}^{m} \Theta_i$	Z		price quote	trade?	flag?	
1	$\ldots$	$\ldots$	$> \phi$	$(1-\phi)/2$	$P_1\ [= Z] > \phi$	yes	no	full
0	0	0	$\ldots$	$\alpha(m)$	$P_1\ [= \phi Z] \leq \phi$	no	no	full
0	0	> 0	$\ldots$	$\beta(m)$	$P_1\ [= \phi Z] \leq \phi$	yes	yes	full
1	$\ldots$	$\ldots$	$\leq \phi$	$\phi/2$	$P_1\ [= Z] \leq \phi$	yes	no	full
0	ϕ	$\ldots$	$\ldots$	γ	(no activity)			no learning

Figure 1: S's supply in period 2 with Z partially revealed

The blue, bald lines illustrate **S**'s supply in period 2 after she observes a trade at price $p_1 \leq \phi$. Conditional on the observed market activity, **S** partially learns about Z and infers that Z is binomially distributed: It is either high (p_1/ϕ) or low (p_1). Whether or not **S** supplies anything at all at the low price depends on the severity of adverse-selection; if her position a_2 exceeds a threshold $\hat{a}(m)$, she does (panel (a)); if it does not, she supplies nothing at the low price (panel (b)). The shaded area in panel (a) represents the adverse selection cost of **S** when the true price is high: $Z = p_1/\phi$.

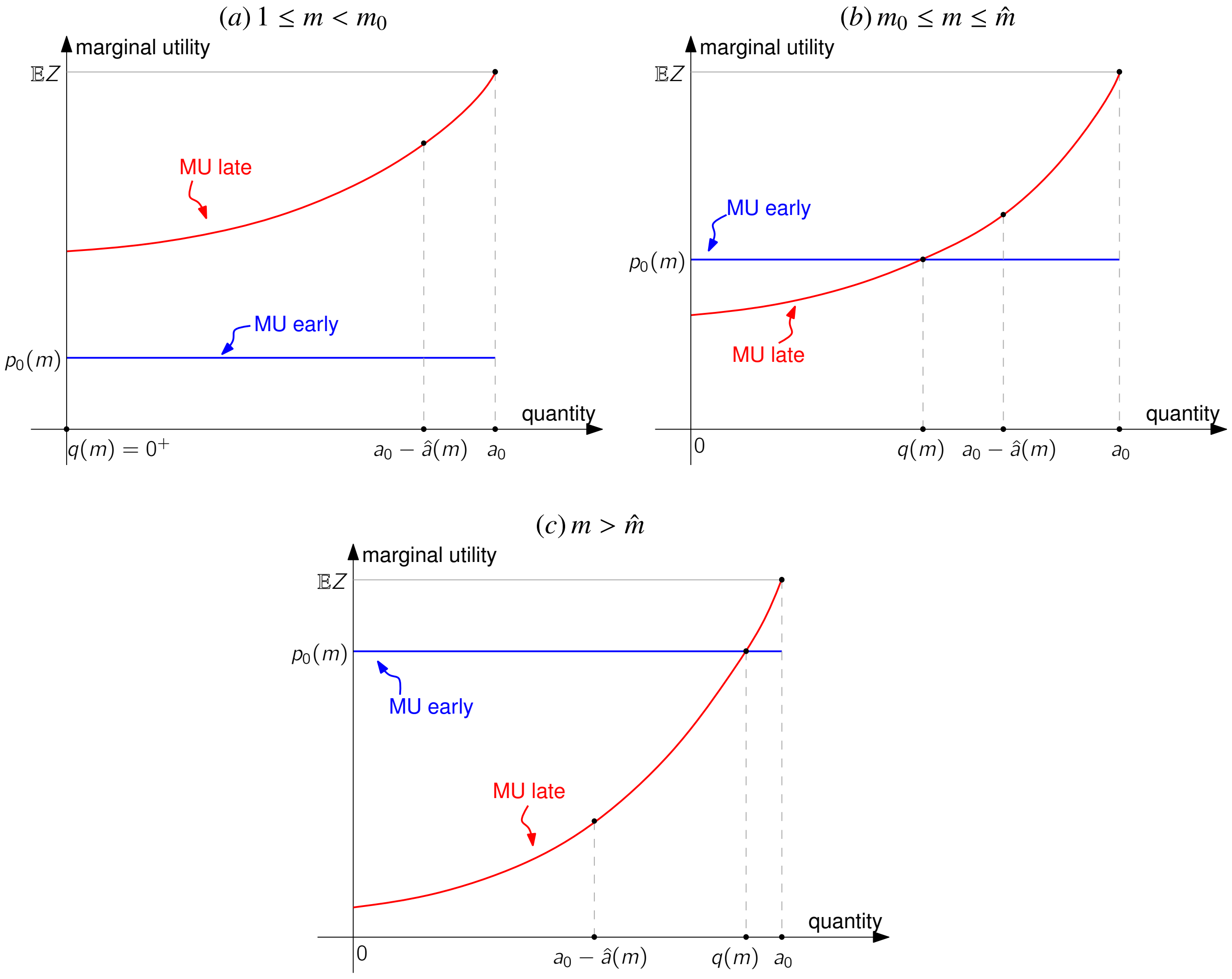


(a) $\hat{a}(m) < a_2$

(b) $\hat{a}(m) \geq a_2$

Figure 2: S's tradeoff between selling early (in period 0) and selling late (in period 2)

In period 0, **S** chooses her optimal supply q by trading off between selling early and selling late. This figure illustrates this tradeoff for three cases. In all panels, the horizontal, blue line shows the marginal utility (MU) of selling early, while the convex red curve represents the MU of selling late. Fixing a large amount of a_0 (assumption 3), the relative positions of "MU early" and "MU late" depend on the number of middlemen, m. In panel (a), "MU late" is always strictly larger than "MU early", and therefore, **S** refrains from selling early, posting only $q(m) = 0^+$ units of the asset, where 0^+ denotes an infinitesimally small, strictly positive number. In panel (b), "MU late" and "MU early" have a unique intersection which yields $q(m) < a_0 - \hat{a}(m)$. In this case, after selling $q(m)$ units in period 0, **S** still has $a_0 - q(m) > \hat{a}(m)$ units of the asset in period 2. Finally, in panel (c), "MU late" and "MU early" uniquely intersect at $q(m) > a_0 - \hat{a}(m)$. In this case, after selling $q(m)$ units in period 0, **S** still has $a_0 - q(m) < \hat{a}(m)$ units of the asset in period 2.



$(a)\ 1 \leq m < m_0$

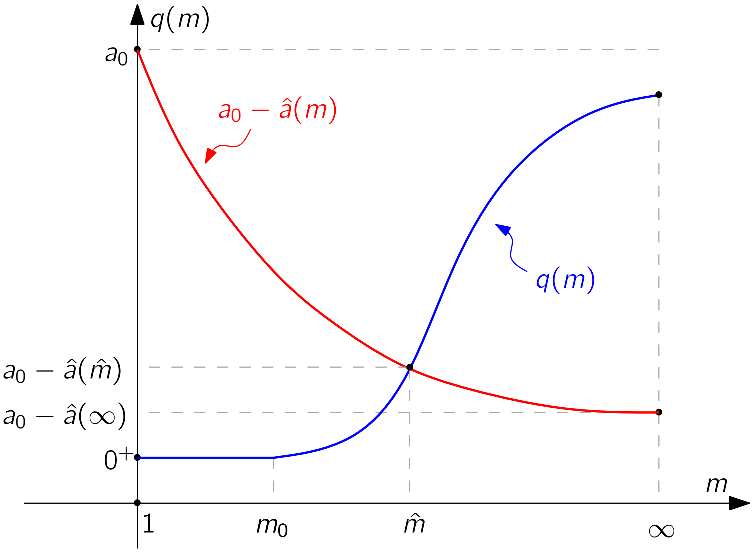
$(b)\ m_0 \leq m \leq \hat{m}$

$(c)\ m > \hat{m}$

Figure 3: S's supply in period 0 and the number of middlemen

This figure illustrates **S**'s optimal supply in period 0 as a function of the number of m. Panel (a) shows the case of $\phi < 1/2$, where $q(m)$ and $a_0 - \hat{a}(m)$ intersect at $m = \hat{m}$. Panel (b) shows the case of $\phi \geq 1/2$, where there is no intersection of $q(m)$ and $a_0 - \hat{a}(m)$. In this latter case, $\hat{m}$ is defined to be ∞.

(a) $\hat{m} < \infty$ **(**$\phi < 1/2$**)**



(b) $\hat{m} = \infty$ **(**$\phi \geq 1/2$**)**

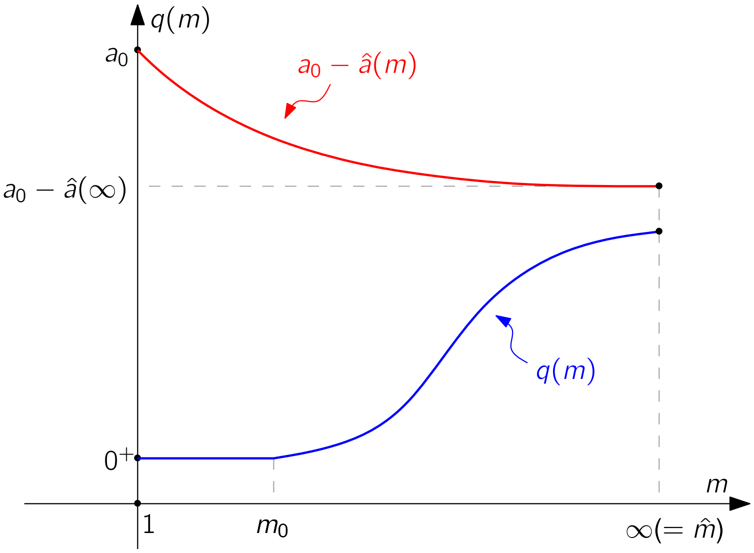
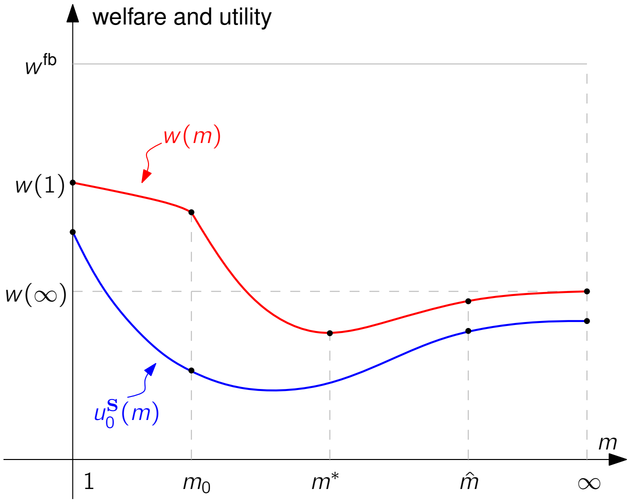


Figure 4: Welfare as a function of the number of middlemen

This figure plots welfare as a function of the number of middlemen, m, for two cases: Panel (a) plots the case of a relatively small ROs ($\phi < 1/2 < \phi^*$), while panel (b) plots the case of a large ROs ($\phi > \phi^* > 1/2$). By lemma 5, $\hat{m} < \infty$ in panel (a) and $\hat{m} = \infty$ in panel (b) (see also figure 3). By corollary 3, $w(m = 1) > w(\infty)$ in panel (a) but $w(m = 1) < w(\infty)$ in panel (b). The welfare values are always lower than the first-best welfare (w^{fb}), because there is always non-zero expected loss when some of the asset ends up with either **S** or **M**. There is a kink at $m = m_0$, as can be seen from the graph. This is because for $m < m_0$, **S** sells almost nothing ($q(m) = 0^+$) and the only social loss comes from **S**'s cost of carry when some units of the asset ended up with **S**, while as soon as **S** starts selling strictly more than 0^+ ($m \geq m_0$) to **M**, additional expected social loss arises due to **M** in the case they cannot resell the position. Also plotted in the figure is **S**'s expected utility, $u_0^{\mathbf{S}}(m)$. The difference between $w(m)$ and $u_0^{\mathbf{S}}(m)$ is **B**'s expected utility, $u_0^{\mathbf{B}}(m)$. Panels (a) and (b) are not meant to be exhaustive. For example, the case of $\phi \in (1/2, \phi^*)$, where $\hat{m} = \infty$ and $w(1) > w(\infty)$, is not included. They only serve to illustrate the general pattern of $w(m)$ (as well as $u_0^{\mathbf{S}}(m)$).

(a) $\phi < {}^1\!/\!_2 < \phi^*$ ($\hat{m} < \infty$)



(b) $\phi > \phi^* (> {}^1\!/\!_2)$ ($\hat{m} = \infty$)

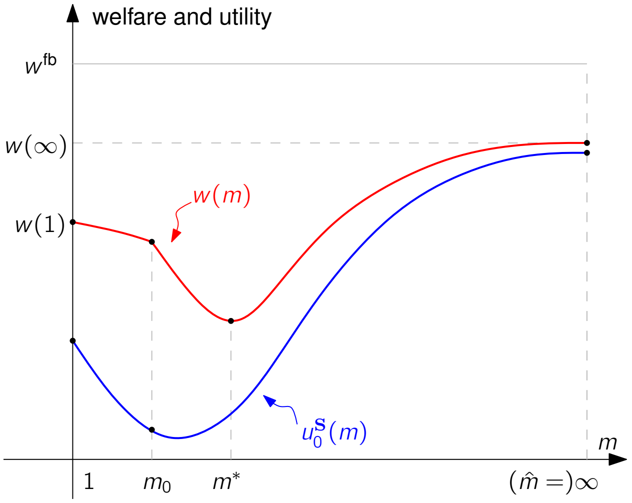
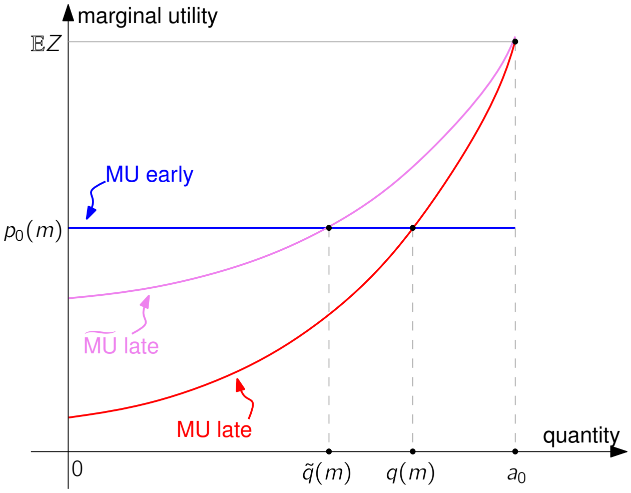


Figure 5: S's supply in period 0, with and without disclosure

Similar to figure 2, this figure shows **S**'s tradeoff between selling early and late in period 0, but adds the comparison between the no-disclosure (baseline) and disclosure regimes (in the latter case, MM trades are flagged). "∼" indicate variables with disclosure. The marginal utility (MU) of selling early is the same in both regimes. The MU of selling late, however, is higher in the disclosure regime. Consequently, with disclosure, **S** sells fewer in period 0.


marginal utility
$\mathbb{E}Z$
MU early
$p_0(m)$
$\widetilde{\text{MU}}$ late
MU late
quantity
$\tilde{q}(m)$
$q(m)$
a_0