Trading speed competition: Can the arms race go too far?*

Dion Bongaerts†, Lingtian Kong‡ and Mark Van Achter§

May 14, 2016

Abstract

We analyze the likelihood of arms race behavior in markets with liquidity provision by HFTs. Liquidity providers (makers) and liquidity consumers (takers) make costly investments in monitoring speed. Competition among makers and takers induces arms race behavior. However, trade success probabilities increase in monitoring speed, giving rise to complementarity externalities between makers and takers. This counters negative arms race effects. Whether arms race effects materialize crucially depends on how marginal gains from trade depend on transaction speed. With the common (often implicit) assumption of constant marginal gains from trade, complementarity effects mostly dominate and market participants tend to under-invest in technology. However, with marginal gains from trade that decline in transaction speed, arms race behavior is much more likely. We provide micro-foundations for declining marginal gains from trade by a dynamic portfolio optimization problem with random rebalancing opportunities.

*We would like to thank Hans Degryse, Joost Driessen, Albert Menkveld, Elvira Sojli, Marti Subrahmanyam, conference participants at the 2015 European Finance Association Meeting (Vienna), the 2015 IFABS meeting (Hangzhou), the 2015 CEMS workshop on Market Liquidity (Brussels), and seminar participants at Erasmus University Rotterdam, VU Amsterdam and the European Securities and Markets Authority (ESMA) for helpful comments and suggestions.

†Rotterdam School of Management, Erasmus University, Department of Finance, Burgemeester Oudlaan 50, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: dbongaerts@rsm.nl.

‡Rotterdam School of Management, Erasmus University, Department of Finance, Burgemeester Oudlaan 50, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: kong@rsm.nl.

§Rotterdam School of Management, Erasmus University, Department of Finance, Burgemeester Oudlaan 50, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: mvanachter@rsm.nl.

1 Introduction

In recent years, financial markets have been completely transformed by a newly-emerging group of market participants: high-frequency traders (HFTs), which provide liquidity using computer algorithms at a millisecond pace. As of 2010, HFTs generate at least 50% of volume, and even more so in terms of order traffic in the US equity market.[1] Facing such radical changes, policy makers in the US and the European Union have called for a welfare assessment of HFTs, in order to design appropriate regulation.[2]

The HFT emergence also induced a fierce academic debate. Early proponents, like Burton Malkiel, argue that "competition among HFTs serves to tighten bid-ask-spreads, reducing transaction costs for all market participants".[3] In contrast, the early opponents, including Paul Krugman, are concerned that HFTs undermine markets and use resources that could have been put to better use.[4] Meanwhile, the vastly growing empirical literature on HFTs has provided evidence consistent with both claims. For example, Brogaard, Hendershott and Riordan (2014) conclude that HFTs contribute to price discovery, Malinova, Park and Riordan (2013) show they improve liqudity, and Carrion (2013) finds they provide liquidity when it is scarce. In contrast, several papers document that HFTs reduce liquidity provision significantly in stressful times, in contrast to traditional market makers (see Anand and Venkataraman (2015) and Korajczyk and Murphy (2015)). Moreover, HFT technology is arguably very expensive for society (see Biais, Foucault and Moinas (2015) for a discussion). For instance, in 2010, Spread Networks installed a new $300 million high-speed fiber optic cable connecting New York and Chicago, to reduce latency from 16 to 13 milliseconds. Meanwhile, that improvement has already become virtually obsolete by the introduction of wireless microwave technology in 2011, which managed to almost shave off an additional 5 milliseconds.[5] Moreover, having some of the brightest minds in the world working on the creation, detection or academic analysis of HFT algorithms implies a large opportunity cost for society.

[1] See the SEC (2010) concept release on equity market structure, and "Casualties mount in high-speed trading arms race", *Financial Times*, Jan 22, 2015.

[2] See "The Morning Risk Report: Future of High Frequency Trading Regulation is Murky", *Wall Street Journal*, January 30, 2014,. and "High-Frequency Traders Get Curbs as EU Reins In Flash Boys", *Bloomberg News*, Apr 14, 2014, respectively.

[3] See "High frequency trading is a natural part of trading evolution", *Financial Times*, Dec 14, 2010.

[4] See "Rewarding bad actors", *New York Times*, Aug 2, 2009.

[5] See Budish, Cramton and Shim (2015) and "Networks Built on Milliseconds", *Wall Street Journal*, May 30, 2012. for a discussion. Other infrastructure-related examples include the cost of co-location services and of individual high-frequency data feeds.

We introduce a model that brings together both opposing views in a unified framework. Our model allows to analyze whether HFTs facilitate allocative efficiency in portfolios sufficiently to justify the incurred (opportunity) costs. In particular, we present two counterbalancing effects HFTs induce on welfare. On the one hand, only the first trader to react to a trading opportunity gains from her investments. As a result, other traders which also invested in trading technology did so in vain (at least, for that trading opportunity), as they arrived (often marginally) later. This negative externality, which we label "substitution effect", materializes both at the liquidity-supply and the liquidity-demand side.[6] Biais, Declerck and Moinas (2015), among others, provide empirical evidence that faster traders indeed obtain larger profits. On the other hand, speedier HFT liquidity provision enlarges opportunities for liquidity demanders to successfully transact, thereby stimulating market participation and investments in trading technology. Hence, the interaction between both market sides entails a positive externality, which we label as the "complementarity effect". This effect incorporates and even goes beyond the competition argument put forward by Malkiel.

To gain intuition on the conditions under which either one of both effects dominates, we need to dig deeper into the model structure. We set up a stochastic monitoring model with two types of agents: market makers which fill the book and market takers emptying the book. When a transaction takes place, trading gains are realized as further explained below. Each agent competes with her own kind for these gains in a winner-take-all fashion.[7] A speed improvement implies makers and takers have better chances to be the first ones to respectively arrive to an empty and filled book. Yet, lowering latency also implies a cost which is quadratically increasing in monitoring intensity.[8] When optimizing their monitoring intensity, both agent types also account for the obtainable marginal gains from an additional trade (labeled "marginal GFT"). The aforementioned substitution and complementarity effects can be identified from the optimization problems for both agent types. Both drive resource allocations away from first best in opposite directions. The marginal GFT essentially function as a weight on

[6]While the early empirical literature mainly focused on the changes in liquidity provision induced by the emergence of high-frequency traders, a similar race is ongoing at the liquidity-demanding side which increasingly applies high-speed algorithmic trading strategies.

[7]Thus, the fastest trader is the only one that profits from a standing trading opportunity. In particular, the first maker to arrive to an empty book can post a sell limit order. Subsequent makers arriving to the book need to wait till the book is empty again. The first taker to arrive to a filled book can transact at the standing sell order. Subsequent takers arriving to the book need to wait till the book is filled again.

[8]In line with reality, this reflects the increasingly costly investments in human capital and IT-infrastructure (which become increasingly costly the more we approach the speed of light).

the complementarity effect. Hence, the proliferation of this positive externality actually hinges on the shape of the GFT: it is more likely to dominate the substitution effect with constant vs declining marginal GFT. Parametrizing our model with realistic participation numbers yields that over-investment in technology is less likely with constant GFT. When assuming the number of takers to be larger than the number of makers (in line with reality), the substitution effect is generally larger among takers than among makers. Meanwhile, there is a substantial complementarity effect from the more numerous takers to the more limited number of makers. Hence, if anything, takers (rather than makers) exhibit arms race behavior under these conditions. However, a different storyline unfolds with declining marginal GFT. The effective weight on the complementarity effect is then lower and we are more likely to find arms race behavior on both sides, implying a net drain on social welfare.

We show that the constant GFT setting has higher tractability, which is a likely reason for its (often implicit) adoption in theoretical models (for example, Biais, Foucault and Moinas (2015), Foucault, Kadan and Kandel (2013) and Hoffmann (2014)). Yet, in the final part of our analysis we provide micro-foundations that marginal GFT are actually declining. In particular, we set up a dynamic portfolio optimization problem with stochastical rebalancing opportunities. In this model, gains of trade arise because risk-averse takers may want to rebalance their portfolios following price innovations. The expected value of doing so increases as the average interval between two trades grows.[9] Thus, the marginal gain of an extra trading opportunity is always positive. Yet, it declines as the interval in between trades shrinks, resulting in declining marginal GFT.

Taking a broader view, our model shows similarities with traditional imperfect competition models such as Cournot (1838). The intensity in our model is (to a large extent) equivalent with quantity in such models. In these models, producers typically face a downward sloping demand curve. The declining marginal gains from trade we provide micro-foundations for are consistent with such a downward sloping demand curve. Yet, there are also key differences with these traditional models. First, our model features competition on both sides of a trade, because makers and takers both compete for profitable trading opportunities. Second, the way individual monitoring intensities translate to transaction intensities generates interesting

[9]Intuitively, volatility and therefore expected portfolio dislocations and rebalancing needs are larger over longer horizons.

patterns. The stochastic winner-takes-all feature of trading induces more over-investment. In addition, the interaction of how buy and sell side monitoring intensities translate to transaction intensities generates the complementarity effect.

Our paper also contributes to the rapidly expanding theoretical literature on the effect HFTs have on welfare. Many recent papers focus on the asymmetric information channel (i.e., the pick-off risk slow traders face) when evaluating the welfare consequences of speed technology (e.g., Biais, Foucault and Moinas (2015), Budish et al. (2015), Cespa and Vives (2015), Hoffmann (2014), Jovanovic and Menkveld (2015), Menkveld and Yueshen (2012), and Rojcek and Ziegler (2016). Other papers, such as Aït-Sahalia and Saglam (2014), explore the welfare impact of the inventory channel (i.e., HFTs are more efficient in optimizing their inventories over time). We document that physical and human capital (opportunity) costs alone suffice to induce a wasteful arms race, and we obtain welfare losses even in the absence of the aforementioned channels. Furthermore, we show that the commonly-made assumptions of constant marginal GFT and risk-neutral investors may induce an underweighting of the negative substitution externality HFTs exert. Adverse selection is then needed to generate arms race effects on the maker side. Our paper provides a further contribution in outlining the micro-foundations for declining marginal GFT with risk-averse takers. Taken together, our model implies that (i) arms race effects are likely to be worse than suggested by models based on these assumptions, and (ii) policy purely targeting information asymmetry or inventory management will be insufficient to prevent arms races from taking place.

2 Setup

The basic setup of our model is based on the one developed by Foucault et al. (2013). We consider a market with two types of participants: market makers and market takers. Each maker $i \in \{1, 2, \ldots, M\}$ monitors the market at discrete points in time and arrives according to a Poisson process with (endogenously chosen) intensity parameter $\mu_i \geq 0$. Similarly, each taker $j \in \{1, 2, \ldots, N\}$ arrives with (endogenously chosen) intensity $\tau_j \geq 0$. The total numbers of makers and takers, M and N, are exogenous.[10]

[10]This setup has a winner-takes-all feature from an ex-post perspective (i.e., the one conducting the trade is the only one benefitting). From an ex-ante perspective, the fastest trader is not guaranteed to always execute. This setup is chosen based on the notion of order processing uncertainty: the fast trader never knows what is

The market we consider operates as a limit order book (LOB) for one security. The LOB can be either empty or filled with a limit order of unit size.[11] When the book is empty (E), the first maker to arrive can post a limit order, thereby changing the LOB status to filled (F). While the LOB is filled, no other makers can post limit orders until the existing quote is hit by a taker (i.e., the filled book becomes empty again). Similarly, once the first taker has seized the trading opportunity in a F book, no other takers can trade anymore until the book becomes F again.[12] The market operates according to the following timeline: before trading begins, each maker i chooses an intensity μ_i to maximize her expected trading profit $\Pi_m(\mu_i)$.[13] Similarly, each taker j chooses τ_j to maximize her expected trading profit $\Pi_t(\tau_j)$. Once the trading begins, each player monitors the market following a Poisson process with the intensity previously chosen. For the moment, we assume that the trading phase of the model repeats itself an infinite number of times. One should notice that while trading happens indefinitely, the model is in essence a one-shot model as arrival intensities are only decided on once at the start.

In our model, monitoring speed does not come for free. In particular, we assume the monitoring costs for both trader types to be quadratically increasing in the monitoring frequency chosen. These costs reflect the required investments in IT-infrastructure and human capital. The marginal cost of technology is increasing, reflecting the observation that as latency approaches zero, the cost for such advancement becomes higher and higher.[14] Monitoring costs can differ between the makers and the takers. This difference allows to assess the impact of heterogeneity in know-how (i.e., a knowledge endowment) among the market participants. We denote the cost per unit of time for maker i by $C_m(\mu_i) = \beta\mu_i^2/2$, while for taker j it equals $C_t(\tau_j) = \gamma\tau_j^2/2$.

By the properties of the Poisson distribution, the aggregate monitoring process of all makers jointly also follows a Poisson distribution with the following intensity: $\bar{\mu} = \sum_{i=1}^{M} \mu_i$. Similarly,

going to happen after she submits the order and before it reaches the server of the exchange. A similar argument can be found in Yueshen (2014).

[11] The unit size assumption is in fact less restrictive than one might think as orders nowadays are often sliced and diced to small standard volumes.

[12] An alternative interpretation of an empty or a filled LOB is whether there is any depth at the best possible quote; not being able to submit a limit order until the book is empty again is then a direct implication of price-time priority rules.

[13] In this sense the agents in our models are the prop traders in Biais, Declerck and Moinas (2015)

[14] For example, to improve latency from a second to a tenth of a second, one would "only" need to automate the trading using algorithms. To get to the millisecond level, however, co-location and super-fast communication lines are required, which are significantly more costly.

the aggregate monitoring intensity of all takers jointly equals $\bar{\tau} = \sum_{j=1}^{N} \tau_j$. Consequently, the expected interval between a transaction and replenishment of the book equals $D_m = 1/\bar{\mu}$. Analogously, the expected interval between the posting of a limit order and transaction is given by $D_t = 1/\bar{\tau}$. Thus on average, the duration between two trades is $D = D_m + D_t$, and the average trading frequency equals $R = (D_m + D_t)^{-1} = \bar{\mu}\bar{\tau}/(\bar{\mu} + \bar{\tau})$. Similarly, one can show that for a given liquidity taker j with monitoring intensity τ_j, the expected trading frequency is given by $\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}$.

Finally, when a transaction takes place, the gains from trade (GFT) are split between between maker and taker according to the fractions π_m and $\pi_t = 1 - \pi_m$, respectively, where $\pi_m \in (0, 1)$.[15] The *expected GFT for a trade* originated by liquidity taker j are given by a (weakly) monotonic and continuously differentiable function $G(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}})$, which hinges on the expected trading frequency.[16] Relatedly, the *expected GFT per unit of time* is given by $\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}} G(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}})$. Concavity of the GFT is equivalent to a $G(\cdot)$ function being uniformly (strictly) decreasing in expected trading speed on its domain.

3 Equilibrium Analysis

We set up and solve the model in this section. First, we do so in general form. Next, we show how market outcomes differ based on specific functional forms for $G(\cdot)$.

3.1 The Most General Case

In this subsection, we derive the equilibrium monitoring intensities of makers and takers. Moreover, we will assess whether these equilibrium monitoring intensities exceed or fall short of first best monitoring intensities. In order to do so, let us first derive the first best outcome by solving the social planner's problem.

Definition 1 ***Social Planner's Problem***

A social planner chooses $\{\mu_i\}_{i=1,2,\ldots,M}$ *and* $\{\tau_j\}_{j=1,2,\ldots,N}$ *to maximize aggregate social welfare:*

[15] Participation incentives dictate that $\pi_m \in (0, 1)$ must hold on average. Because we abstract from informed trading, this must even hold for each individual trade.

[16] We assume that the GFT arise from the portfolio selection and consumption need of the *takers*. This is supported by the reality that in financial markets, the market takers are usually the parties with the intention to hold the security over an horizon exceeding a day, while the makers mainly serve as short term intermediaries. In Section 4, we will provide further micro-foundations for this assumption.

$$\sum_{j=1}^{N} \frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}} G\left(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}\right) - \sum_{i=1}^{M} \frac{\beta \mu_i^2}{2} - \sum_{j=1}^{N} \frac{\gamma \tau_j^2}{2}, \tag{1}$$

such that $\mu_i \geq 0, \forall i \leq M; \tau_j \geq 0, \forall j \leq M$.

As can be gauged from this objective function, the social planner only focuses on the aggregate gains from trade which are realized by the interaction between makers and takers. It does not account for the distribution of these gains between makers and takers (i.e., π_m does not show up in this equation).

Without the explicit shape of the functional form of $G(\cdot)$ with respect to $\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}$, a closed form solution for the first best allocation is impossible.[17] But we do obtain the following result that helps in solving the social planner's problem for specific functions $G(\cdot)$:

Lemma 2 ***First Best Monitoring Ratio***

The first best monitoring intensities' ratio $\frac{\hat{\mu}}{\hat{\tau}}$ *is not affected by the functional form of* $G(\cdot)$.

$$\frac{\hat{\mu}}{\hat{\tau}} = \left(\frac{N^2}{M^2}\frac{\gamma}{\beta}\right)^{\frac{1}{3}} \tag{2}$$

Proof. See Appendix. ■

The intuition behind this result is as follows. The existing expected speed of the trade occurring for the makers does not influence the marginal gain per trade $G(\cdot)$ of the entire economy As such, the social planner can take the trade execution speed of the taker side, thus the gain per trade as given, and independently adjust the maker's frequency, and vice versa. As a result, the first order optimality condition determines a fixed ratio of the makers' optimal monitoring frequency to that of the takers.

Armed with this result, the job to compute the optimal monitoring intensities set by the social planner becomes easier. As can be seen in conjunction with the functional form of her objective function, the social planner only needs to solve a univariate problem instead of a bivariate one. An important feature of our setup contributes to obtaining this result: the monitoring intensities $\bar{\tau}$ and $\bar{\mu}$ enter into the objective function of the social planner only in

[17] When we restrict this function to be linear, like the conventional setup of the current literature, a closed form solution may be obtained (as is shown in the next subsection).

the form of $\frac{\bar{\tau}\bar{\mu}}{\bar{\mu}+\bar{\tau}}$. As such, this result will prove to be an essential intermediary step in solving the linear benchmark in the next subsection.

Next, we define the equilibrium of the general setting in which maker and taker intensities are not set by the social planner.

Definition 3 ***Equilibrium***

The equilibrium that we consider is a Nash equilibrium defined by intensity choices $\{\mu_i\}_{i=1,2,...,M}$ *and* $\{\tau_j\}_{j=1,2,...,N}$*, such that:*

1. *Given all taker intensities,* $\{\tau_j\}_{j=1,2,...,N}$*, as well as all other maker intensities:* $\{\mu_n\}_{n=1,2,...,i-1,i+1,...,M}$ *each maker* i *maximizes her profit after cost:*

$$\pi_m \frac{\mu_i}{\bar{\mu}} \sum_{j=1}^{N} G(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}) \frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}} - \frac{1}{2}\beta \mu_i^2 \tag{3}$$

2. *Given all maker intensities,* $\{\mu_i\}_{i=1,2,...,M}$*, as well as all other taker's choice:* $\{\tau_n\}_{n=1,2,...,j-1,j+1,...,N}$ *each taker* j *maximizes her profit after cost:*

$$G(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}) \pi_t \frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}} - \frac{1}{2}\gamma \tau_j^2 \tag{4}$$

where $\mu_i \geq 0$*, and* $\tau_j \geq 0$*.*

We will now compare the equilibrium allocation with the first best outcome, and explicitly consider the curvature of the functional form of $G(\cdot)$. Doing so will prove to be crucial to determine whether wasteful arms races occur. The current literature mostly assumes that $G(\cdot)$ is linear, irrespective of the trading speed (for example, Biais, Foucault and Moinas (2015), Foucault et al. (2013) and Hoffmann (2014)). We therefore first solve the model for constant $G(\cdot)$ as a benchmark.

3.2 Benchmark Case: Linear Aggregate Gains from Trade

In this case, each transaction that takes place generates the same amount of social welfare, and we have that $G(\cdot) = G_0$. Hence, the aggregate GFT is linear in the expected trading frequency of the taker involved. For this economy, a closed form solution for the allocation and welfare aggregated over all participants is obtainable:

Proposition 1 ***First Best with Constant Marginal GFT***

The first best monitoring intensities are symmetric and given by:

$$\hat{\mu} = \frac{G_0}{\beta}\frac{1}{(\hat{r}+1)^2}, \hat{\tau} = \frac{G_0}{\gamma}\frac{\hat{r}^2}{(\hat{r}+1)^2}; \textit{where } \hat{r} = \left(\frac{N^2}{M^2}\frac{\gamma}{\beta}\right)^{\frac{1}{3}}$$

The resulting aggregate welfare equals:

$$\hat{\Pi} = G_0\frac{M\hat{\mu}\cdot N\hat{\tau}}{M\hat{\mu}+N\hat{\tau}} - \frac{1}{2}\beta M\hat{\mu}^2 - \frac{1}{2}\gamma N\hat{\tau}^2$$

Proof. See Appendix. ■

The economic intuition behind this solution is as follows. First, the optimal maker intensity $\hat{\mu}$ increases in the GFT per trade, G_0, because higher G_0 justifies higher monitoring investments to make trades happen more frequently. Second, $\hat{\mu}$ decreases in the marginal monitoring cost for makers, β, due to the increasing marginal cost of monitoring intensity. Third, it is also intuitive that $\hat{\mu}$ decreases in the number of makers, M. Since it is the aggregate intensity that the social planner cares about and individual costs are quadratic in monitoring intensity. The more makers there are, the less frequently each individual maker should monitor due to the "substitution effect" stemming from competing other makers. In addition to the within-type effects outlined above, cross-type effects can also be observed. First, the optimal maker intensity $\hat{\mu}$ decreases in the marginal cost of the takers, γ. This happens because maker intensity and taker intensity are complementary. After all, high monitoring by takers is only useful if the book is likely to be F and high monitoring intensity by makers is only worthwhile if the book is likely to be E. Hence, makers and takers impose positive externalities on each other which we label "complementarity effect". Second, $\hat{\mu}$ increases in the numbers of takers N due to the same type of complementarity. Third, this complementarity effect can also be seen from the first term of the formula for the aggregate welfare (i.e., $G_0 M\hat{\mu}N\hat{\tau}/(M\hat{\mu}+N\hat{\tau})$). Unilateral increases in maker intensity $\hat{\mu}$ will increase the total welfare not only by a factor of M, but also by N, the number of takers. Due to the symmetry of the results, all six interpretations above apply to the taker intensities $\hat{\tau}$ too.

In reality, however, the first best outcome typically does not materialize. Therefore, we now proceed by solving for the equilibrium of this economy and compare it with the first best

outcome outlined above. While solving in closed form is not possible, following Foucault et al. (2013) we can obtain an implicit solution for the base case equilibrium.

Proposition 2 ***Equilibrium with Constant Marginal GFT***

The equilibrium monitoring intensity for makers and takers is given by:

$$\mu^* = \frac{M + (M-1)\tau^*}{M(1+r^*)^2}\frac{G_0\pi_m}{\beta}, \tag{5}$$

$$\tau^* = \frac{r^*((1+r*)N-1)}{N(1+r^*)^2}\frac{G_0\pi_t}{\gamma}, \tag{6}$$

respectively, where $r^* = \gamma G_0\pi_m/\beta G_0\pi_t$ *is the positive real solution to:*

$$Nr^3 + (N-1)r^2 - (M-1)zr - Mz = 0$$

with $z \equiv \gamma G_0\pi_m/\beta G_0\pi_t$.

The resulting aggregate welfare is given by:

$$\Pi^* = G_0 M\mu^* N\tau^*/(M\mu^* + N\tau^*) - \beta M\mu^{*2}/2 - \gamma N\tau^{*2}/2$$

Proof. See Appendix. ■

An easy comparison between the equilibrium and first best intensities is not possible because r^* is implicitly defined. However, we can observe some interesting regularities. First, we can compare the first order conditions with respect to μ_i of the social planner outcome (SPP) and the unrestrained equilibrium outcome (EQ):[18]

$$G_0\bar{\tau}^2/(\bar{\mu}+\bar{\tau})^2 = \beta\mu_i \quad \textit{(FOC-SPP)} \tag{7}$$

$$G_0\pi_m(\bar{\tau}^2 + \mu_{-i}\bar{\tau})/(\bar{\mu}+\bar{\tau})^2 = \beta\mu_i \quad \textit{(FOC-EQ)} \tag{8}$$

where $\mu_{-i} \equiv \sum_{j\neq i}^{j\leq M}\mu_j$. This way, we can see whether the complementarity effect and the substitution effect work differently in the EQ than in the SPP. First, observe that the LHS of the FOC-SPP has a multiplier G_0, yet in the FOC-EQ it is $\pi_m < 1$. This is a demonstration that in equilibrium, the individual maker fails to fully endogenize the positive externality her

[18] For the derivation of these FOCs, please refer to the appendix.

monitoring imposes on the takers. In other words, she fails to take into account the complementarity effect her activity induces on her counter-parties. As a result, makers are inclined to under-monitor as compared to the first best. Secondly, the numerators of the LHSs of the two FOCs differ by a term $\mu_{-i}\bar{\tau}$, which increases the EQ intensities relative to the SPP intensities. This over-monitoring in EQ happens because each maker i suffers from the competition from all other makers. Makers tend to individually over-invest in speed to keep up and do not fully endogenize the negative substitution externality on the other makers.

The two effects described above are (partially) offsetting. Which of the two dominates depends on parameters. Due to the complexity of the model closed form solutions are not available and we need to rely on numerical analysis. A result which can be shown more generally, however, is that the smaller the cost coefficients β, γ become, the closer monitoring intensities are to first best. This can be seen by comparing the FOCs of the equilibrium (for example, $\mu^* = G_0\left(M + (M-1)r^*\right)\pi_m / M(1+r^*)^2\beta)$ with those of the SPP: $\hat{\mu} = G_0/\beta(\hat{r}+1)^2$. We can see that the difference is inversely related to β and γ. The intuition here is that γ and β measure the opportunity costs of investment in trading technology. With a lower opportunity cost, the distortion imposed by the substitution effect also shrinks.

In order to obtain insights under which parameter ranges over- and under monitoring take place, we visualize the maker and taker intensities relative to first best in Figure 1. To enable 3D-plotting, we reduce the number of free parameters by restricting our model parameters as follows: $G_0 = 1$ and $\beta = \gamma$. Then we make plots of the over-monitoring of either type as a function of the profit split ratio π_m and the cost coefficient β, for several combinations of (M, N). In particular, we plot the following combinations of M and N: $(2,2); (2,20); (50,250)$. We set the number of takers larger or equal to the number of makers as, in reality, there are usually more liquidity demanders than suppliers in any particular market (think of any one stock). Even if these parameterizations do not hold in some markets, due to the symmetry of the way we model the makers and takers, symmetrical conclusions can be drawn if the maker/taker ratio is reversed.[19] This is because the marginal costs of technology of both groups are assumed to be equal in the plotting. This assumption can be easily relaxed to consider specific markets.

[19]Note that this reversion is only possible for the linear case.

Some preliminary features of the plots are consistent with our intuition. First, as argued above, the smaller the cost coefficient (β, γ), the more monitoring diverges from first best (i.e., over- and under-monitoring are generated). Second, over-monitoring is more likely to happen when there are more makers and/or takers competing. In this case there are more competitors, and hence more sources of negative substitution externalities.

Our main observation from this figure is that for relatively small values of M and N, neither side over-monitors severely. The first column in Figure 1, which corresponds to the case when $M = 2$ and $N = 2$, shows no over-monitoring at all. The arms race only begins gradually from $N = 20$. Even with $(M, N) = (2, 20)$ over-monitoring is still very limited as shown in the second column of Figure 1. When we look at the third column of the figure, we notice that as the market size grows, the substitution effect starts to dominate the complementarity effect more and more. Yet, especially for the makers, we see large parameter ranges where there is under- rather than over-monitoring. Only for relatively high values of π_m, do we see over-monitoring among makers. This also intuitively makes sense as a higher value for π_m gives more value to being the first one to execute a trade and lowers the benefit of the takers to invest in trading technology in response to an upgrade in maker technology and speed.

These results tell us that when the GFT is linear, the complementary effect dominates the substitution effect for sizable parameter ranges. In other words, the higher frequency of monitoring of either side benefits the other by providing trading opportunities to them, which more than compensates for any arms race behavior. If an arms race is going on, it is on the taker rather than on the maker side, due to the relatively higher presence of the takers in the market.

Some additional observations are also interesting and have policy implications. It seems that the over-monitoring of the takers is more prevalent in the M, N combinations that we deem realistic, namely where the number of takers exceeds the number of makers. This is in contrast with the mainstream view on the arms race that it is more likely to occur among makers. In addition, it stands out from the figure that the over-monitoring of the takers only occurs when π_m is close to 0.5, that is, the two sides have similar bargaining power. An intuitive explanation is that when π_m is too small, there are not enough makers to generate enough positive externality to motivate sufficient monitoring of the takers, let alone the over-

monitoring. On the other hand, when π_m is large, takers benefit little from trading and will invest too little. This observation suggests that the relative market power of the liquidity suppliers and demanders, in addition to their speed and their sheer numbers, can also be relevant factors to take into account when designing regulatory measures to ensure efficient monitoring.

The above results yield interesting insights. However, they crucially depend on the assumption of constant marginal GFT. Therefore, in the next subsection, we relax this assumption, and fall back on our previous intuition that the faster the trading speed becomes, the less benefit there is left to be gained from each trade. In other words, when the GFT is concave in the expected trading speed. We then show that if aggregate GFT are concave in trading frequency, over-monitoring by either side of the market is much more likely.

3.3 Concave Aggregate Gains from Trade

While in the current literature, the gains from trade are usually assumed to be linear in the aggregate expected trading speed, this may not be true in practice. After all, if trading is motivated by portfolio rebalancing or hedging demands, the need for trading is likely lower shortly after the previous trade, as market prices are unlikely to have moved much (see also Section 4 where we work out such micro-foundations). In other words, the expected gains from a trade are likely to be declining in trading frequency and the expected aggregate gains from trade are likely to be concave in trading frequency. It would be interesting to know whether the results obtained under the traditional linearity assumption would still hold in this more realistic setting. Therefore in this subsection we impose on the model that the expected aggregate gains from trade are positive but strictly decreasing in trading frequency.

We start by analyzing how the relationship between equilibrium and first best differs from their counterparts in the previous section. When $G(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}})$ is decreasing in $\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}$, the gains from trade are declining in trading frequency. As a result, the complementarity effect declines as $\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}$ increases, and over-monitoring is more likely to take place. To illustrate the effect of concavity of GFT in $\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}$ we repeat the previous plots for one of the simplest concave functions of $G(.)$: $G(\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}}) = -k\frac{\tau_j \bar{\mu}}{\bar{\mu}+\bar{\tau}} + G_0$. Furthermore, we let $G_0 = 1$ to make this setting comparable to the previous plots. From a conceptual point of view, this linear form for the expected marginal gains

from trade is interesting to analyze as it is also the result of a second-order Taylor approximation around 0 of any monotonically increasing and differentiable function that crosses the origin. Unfortunately, even a simple function as a quadratic polynomial leads to loss of tractability as the FOC involves solving for the root of a fifth order polynomial. This is a problem that has been proven to have no closed form solution by Abel (1881). Therefore, we can only analyze this setting numerically. A graphical representation of this numerical analysis can be found in Figure 2. We first plot the case when $k = 1$. Compared to Figure 1, we observe that the tendency for makers to over-monitor is higher, holding constant the number of makers and takers. Moreover, even for small values of M, makers only over-monitor in this concave setup, and hardly ever in the linear setup.

Our results are robust to different choices of the quadratic parameters. Allowing the coefficient on the linear term, k, to take values of 0.2 (Figure 3) and 5 (Figure 4), we obtain qualitatively similar results. Moreover, as the concavity increases, the over-monitoring in general becomes more severe, demonstrating that the nature of the marginal gains from trade plays an important role in determining the occurrence of the arms race. This can be seen by a cross-sectional comparison of the corresponding panels of the Figures 1, 3 , 2, 4.[20] As the concavity increases, the parametric ranges over which the over-monitoring occurs seem to expand. When M and N are large, for example, when $M = 50, N = 250$, the domain of the β, π_m over which over-monitoring occurs is strictly expanding in the concavity. Though such large numbers of makers or takers may not be feasible in reality, this result demonstrates that when the marginal effect of one particular maker (taker) on the entire maker side (taker side) is negligible, the concavity has a monotonic effect on the over-monitoring.

While the assumption of concave GFT yields results that conform more to the "Krugman view" on HFTs, one could wonder whether such a shape is likely to arise. In the next section, we extend our model by introducing investors with portfolio rebalancing needs and show that the concavity assumption is actually supported by asset pricing theory. As such, we provide micro foundations for an expected aggregate gains from trade function that is concave in transaction speed.

[20] In this particular order, the concavity increases.

4 Gains from Trade: Micro-Foundations

4.1 An Extended Setup

In our extended framework, portfolio optimization is the main motivation for trading. We build our analysis on the Merton (1969) style portfolio selection problem in which risk-averse takers trade in response to market price changes due to a re-balancing need. One could alternatively look at a dynamic hedging problem for a non-linear derivative position where the taker is risk averse to hedging error and can only trade at random points in time; the intuition in such a problem would run along the same lines. In turn, the makers in our extended model are risk neutral, are rewarded a share of the gains from trade from the takers and serve to provide liquidity. Overall, this new setup extends from the one in Subsection 3.1, integrating the mechanism generating the need for trading.

We now assume that in an economy of infinite time horizon, there are two assets available for trade. There is a risk free asset with constantly compounded rate of r:

$$dX_0(t) = rX_0(t)dt, \tag{9}$$

and a risky asset whose price change follows a geometric Brownian motion:

$$dX_1(t) = \alpha X_1(t)dt + \sigma X_1(t)dB(t), \tag{10}$$

where $B(t)$ is a Brownian motion. Only the risky asset is traded in the LOB that was introduced earlier.

We allow continuous quantities of limit orders in the order book to be traded on, after which the rest of the limit order is withdrawn by assumption. This way, the book accommodates the desire to continuously rebalance as a result of price changes. By assumption, if the LOB is filled, there is always enough volume available to accommodate any trading demand of a single trader.[21] Whenever a maker arrives to an empty LOB, she can fill it to "full", meaning that she puts in a limit buy order as well as a sell order, both of infinite size. The taker has an

[21] Restricting the available volume in the LOB would only strengthen our results. To see this, with infrequent trading, the required trading size will on average be larger than with frequent trading as the volatility of realized distortions from optimal portfolio weights since the last trade is larger. As a result, the constraint on available depth becomes less and less binding, implying declining marginal GFT.

opportunity to trade when she arrives to the LOB and sees that the LOB is full (i.e., it has been filled by a maker and not yet been hit by another taker). Whenever she sees this opportunity, the taker will trade almost surely in equilibrium, since the price of the risky asset is continuously changing and therefore, her portfolio deviates from the optimal portfolio with probability 1.

We maintain the assumptions in Section 2 about the monitoring intensities of the makers and the takers. Thus for any taker j, the occurrences of the joint event that "she reaches the book *and* the book is filled at the time" follow a Poisson Point Process (PPP): $(s_1, s_2, \dots)$, henceforth referred to as this taker's "trading times". Note that a certain taker's trading time is discrete and stochastic.

Let $P(t)$ denote the Poisson process from which the aforementioned PPP is generated. Thus the differential:

$$dP(t) = \begin{cases} 1 & \text{when} \quad \exists n, \text{ such that:} \quad s_n \in [t, t+dt), \\ 0 & \text{otherwise.} \end{cases}$$

And the intensity parameter of this parameter, henceforth referred to as her "trading frequency" can be computed:

$$\lambda_j = \frac{\tau_j \bar{\mu}}{\bar{\mu} + \bar{\tau}} \tag{11}$$

In this new model, we abstract from the multitude of agents of the previous model, and focus on one of the takers, say j, since the relationship of her GFT with her trading depends only on the portfolio selection problem she privately faces, rather than the interaction with other agents. For this representative taker, let $W_0(t)$ and $W_1(t)$ be the amount of her wealth invested in the risk free and the risky securities respectively at time t. $Q(t)$ is the amount of wealth that she transfers from the risky asset to the risk free asset at her trading time $t = s_n$, $n = 1, 2, \dots$. Let $C(t)$ be the process of her consumption flow. The consumption of the taker is allowed to happen continuously.

We assume that there is a cost for each transaction paid by the taker to the maker involved. This fee equals a fixed proportion π_m of the GFT $g(P(t))$ of this trade with the trade size $P(t)$. We further assume that whenever trade execution happens, the rest of the depth (i.e., the remaining limit orders) is automatically canceled. This simplifying assumption implies that the maker responsible for these orders does not have to monitor again to cancel them.

To generate a need for portfolio rebalancing, we assume the takers to be risk averse with

CRRA utility with relative risk aversion parameter δ.[22]

4.2 Endogenizing Gains from Trade

In order to endogenize the gains from trade, we first define the problem of taker j, the solution of which pins down all the quantities in this model.

We assume that each taker maximizes the expected sum of her time-discounted consumption trajectory under inter-temporal budget constraints and the no-bankruptcy constraint.

Definition 4 *Given her trading frequency λ_j, the* ***"taker's problem"*** *is each taker j's problem to choose the quantity of portfolio re-balancing $\{Q(t)\}_{t=s_n}$[23] and the continuous consumption process $\{C(t)\}_{t\geq 0}$ to solve:*

$$\max_{\{C(t),Q(t)\}_t} \mathbb{E}\int_0^\infty e^{-\rho t}U(C(t))dt \equiv J(x,y,t=0|\lambda_j,\pi_m) \tag{12}$$

subject to:

$$W_0(0) = x, \tag{13}$$

$$W_1(0) = y, \tag{14}$$

$$dW_0(t) = W_0(t)\frac{dX_0(t)}{X_0(t)} - C(t)dt + (Q(t) - \pi_m g(Q(t)))dP(t), \tag{15}$$

$$dW_1(t) = W_1(t)\frac{dX_1(t)}{X_1(t)} - Q(t)dP(t), \tag{16}$$

$$0 \leq W_0(t) + W_1(t), \tag{17}$$

where

$$g(Q(t)|\lambda_j,\pi_m) \equiv J(W_0(t)+Q(t), W_1(t)-Q(t), t|\lambda_j,\pi_m) - J(W_0(t),W_1(t),t|\lambda_j,\pi_m) \tag{18}$$

Equations 13 and 14 are initial conditions. Equations 15 and 16 are the laws of motions for the risk free security and the risky security, respectively. Finally, equation 17 is the no-

[22] Our results can be generated for any utility function with risk aversion. The log utility function happens to provide us with a high degree of tractability.

[23] Sometimes we write $\{Q(t)\}_{t=s_n}$ simply as $\{Q(t)\}_{t\geq 0}$, since the value of $Q(t)$ doesn't matter when $dP(t) = 0$.

bankruptcy condition for taker j.

The term $-Q(t)dP(t)$ in equation 16 and the term $Q(t)dP(t)$ in equation 15 represent that wealth of amount $Q(t)$[24] is transferred from the risky security to the risk free security at time $t = s_n$, $n = 1, 2, \ldots$.

In equation 18, the gain from trade function $g(Q(t))$ is defined as the value function (defined in 12) when this trade is carried out at quantity $Q(t)$, subtracted by the value function when this trade is not carried out. Correspondingly, term $\pi_m g(Q(t))dP(t)$ in equation 15 represents that this transaction $Q(t)$ costs the taker $\pi_m g(Q(t)))$, which is a proportion π_m of the gain from this trade: $g(Q(t)))$. As will become clear later, this assumption of the transaction cost being proportional to the gain from a trade is crucial for the tractability of our model.

Next, we define the "expected GFT for a trade", the variable that we endogenize in this section.

Definition 5 *The* ***"expected GFT for a trade"****, is defined as:*

$$G(\lambda_j, \pi_m) \equiv \mathbb{E}_{t=s_n} g(Q^*(W_0(s_n), W_1(s_n)), s_n | \lambda_j, \pi_m) \tag{19}$$

where $Q^*(W_0(s_n), W_1(s_n))$ is the second argument of the solution to the problem 4.

By the properties of the conditional expectations, for a given taker j, every one of her trade has the identical *ex ante* expected gain. It equals the expectation of the gain from trade function $g(Q(t))$ (defined in equation 18), evaluated at the optimal trade size: $Q^*(W_0(s_n), W_1(s_n), t)$.

The goal of this section is to identify, for any taker j, the monotonicity of her expected GFT for a trade: $G(\lambda_j, \pi_m)$ with respect to her trading frequency λ_j. [25]

4.3 Asymptotically Diminishing Expected GFT for a Trade

As shown by the review of Pham (2009), a closed-form solution is not currently possible for the type of stochastic control problem as in Problem 4.[26] But we are able to bypass the need

[24] $Q(t)$ is allowed to be negative in which case the transfer is the other way around.

[25] Note that the concavity of the *aggregate* gains from trade, the nomenclature we used in Subsection 3.3, is equivalent to the decreasing expected GFT for a trade.

[26] Our problem 4 differs from the classic portfolio re-balancing models, such as Merton (1969), in two important ways. First, the trade can only occur at discrete random time points; second, there is transaction costs when the trade happens.In this aspect, our model is therefore related to the models of dynamic equilibria with transaction costs, such as Constantinides (1986), but they do not have the restriction that the trades can only happen at discrete random time points.

for an analytic solution of this problem, and instead identify the monotonicity of the expected GFT for a trade directly.

To achieve this, we first simplify the problem by showing that it is to some extent equivalent with the same problem, but *without* transaction costs.

Proposition 3 ***(sufficient condition of the monotonicity of the "expected GFT for a trade")*** *To show that $G(\lambda_j, \pi_m)$ defined in Definition 5 decreases in λ_j, it suffices to show that the following quantity is concave in λ_j:*

$$J(W_0(0), W_1(0), 0|\lambda_j, \pi_m = 0)$$

Proof. See appendix. ■

Simply put, this quantity is the value function of problem 4 when there is no transaction cost. Intuitively, this result crucially follows from the assumption that for any trade, the transaction cost is proportional to the gain from this trade. In this case, at any of the trading times s_n, $n = 1, 2, \ldots$, the actual objective function of the taker is simply a linear transformation of the objective function of the corresponding transaction-cost-free problem.

Based on the Proposition 3, it remains only to show that $J(W_0(0), W_1(0), 0|\lambda_j, \pi_m = 0)$ [27] is concave in λ_j. To achieve this, we first write down the Bellman Equation of the problem without transaction cost.

Equation 12 can be restated in dynamic programming form to apply the Bellman Principle of Optimality:

$$\begin{aligned} & J[W_0(t_0), W_1(t_0), t_0|\lambda_j] \\ = & \max_{C(s), W_1(s)} E_{t_0}\left[\int_{t_0}^{t} e^{-\rho s} U[C(s)]ds) + J[W_0(t), W_1(t), t|\lambda_j]|W_0(t_0), W_1(t_0)\right] \end{aligned} \tag{20}$$

Then by Taylor's theorem and the mean value theorem for integrals, there exist $\bar{t} \in [t_0, t]$,

[27] Henceforth written as: $J(W_0(0), W_1(0), 0|\lambda_j)$

such that the above equation can be restated as:

$$J[W_0(t_0), W_1(t_0), t_0|\lambda_j] = \max_{C(t), W_1(t)} E(t_0) \Big[e^{-\rho \bar{t}} U[C(\bar{t})] l + J + \frac{\partial J}{\partial W_0(t_0)} (W_0(t) - W_0(t_0)) + \frac{\partial J}{\partial W_1(t_0)} (W_1(t) - W_1(t_0)) + \frac{1}{2} \frac{\partial^2 J}{\partial W_1(t_0)^2} (W_1(t) - W_1(t_0))^2 + \lambda_j [J(W_0(t_0) + Q(t_0), W_1(t_0) - Q(t_0), t_0|\lambda_j) - J(W_0(t_0), W_1(t_0))] + o(l) \Big] \tag{21}$$

where $l \equiv t - t_0$ is the latency at time t_0. The last term is due to the possibility of the jump of the portfolio processes.

We now take the $\mathbb{E}_{t_0}$ operator onto each term, and eliminate $J[W_0(t_0), W_1(t_0), t_0|\lambda_j] = \mathbb{E}_{t_0} J[W_0(t_0), W_1(t_0), t_0|\lambda_j]$ from both sides, then evaluate $\mathbb{E}_{t_0}[W_0(t) - W_0(t_0), t_0|\lambda_j]$, $\mathbb{E}_{t_0}[W_1(t) - W_1(t_0)]$ and $\mathbb{E}_{t_0}[W_1(t) - W_1(t_0)]^2$, and finally take the limit as $l \to 0$ we get:

$$0 = \max_{C(t), W_1(t)} \Big[U[C(t)] - \rho J[W_0(t), W_1(t), t_0|\lambda_j] + \frac{\partial J}{\partial W_0(t)} (r W_0(t) - C(t)) + \frac{\partial J}{\partial W_1(t)} \alpha W_1(t) + \frac{1}{2} \sigma^2 W_1^2(t) \frac{\partial^2 J}{\partial W_1^2(t)} + \lambda_j [J(W_0(t) + Q(t), W_1(t) - Q(t), t_0|\lambda_j) - I(W_0(t), W_1(t))] \Big] \tag{22}$$

Using the technique of asymptotic expansion of the solution as in Roger and Zane (2002), we are able to show the asymptotic concavity of the expected gain from a trade in the trading speed, as $l \to 0$.

Proposition 4 ***(expected aggregate GFT without transaction cost)*** *For each taker, her value function $J[W_0(0), W_1(0), 0|\lambda_j]$ is a concave function of λ_j to the 2nd order of the inverse of the trading frequency λ_j, [28], that is, as the trading frequency λ_j is high enough that $\frac{1}{\lambda_j^2}$ is negligible.*

Proof. See appendix. ■

Our reasoning in the proof is based on asymptotic expansion. The very high speed of the HFTs warrants our assumptions of λ_j getting very large in the current trading environment.

The concavity comes from the nature of the Geometric Brownian price change. The price change in infinitesimal dt is to the higher order of dt itself, due to the continuity of the sample path of this type of random processes. As a result, the deviation of the portfolio should the

[28] While fixing the main arguments W_0, W_1.

trade not have occurred is also to the higher order of dt heuristically. This translates into the marginal gains from trade shrinks quicker than t, that is, being concave.

5 Conclusion

In this paper, we have explored whether competition on speed among stock market participants is likely to trigger arms races, leading to socially wasteful investments. We highlight two opposing economic channels that influence such effect in opposing and partially offsetting ways. Competition among makers as a group and among takers as a group may indeed trigger arms races in the classical sense. However, a complementarity between the two sides, the increased success rate of trading, may offset this competition effect if the gains from trade are large enough. Therefore, the likelihood of arms races depends on how gains from trade depend on transaction frequency. Using a portfolio rebalancing model, we show that gains from trade are likely to be concave in transaction frequency. Intuitively, the gains realized in a trade shrink the smaller the time interval in between subsequent trades. Under this new more realistic specification, arms races are more likely to occur than under the standard paradigm in the literature (featuring gains from trade that are independent on the time interval in between subsequent trades).

While providing important insights, our model does make some concessions to reality. A potential concern is that it does not allow for the dual role of participants in modern limit order markets. Yet, in fact this concern is not as grave as one would think. After all, there is a group of market participants that are likely to show a net demand for liquidity. Moreover there is a group that on net will be providing liquidity. This is what in the end generates the welfare gains. Trades among makers, which currently are very common, are zero sum within the group of makers (one maker could have been the only intermediary rather than a whole chain). The single role assumption massively simplifies our problem, leading to better tractability.

Another comment to make on the results of the model is that competition in speed may have positive externalities in the sense that it boosts technological progress and knowledge. Other industries may benefit from this progress in unanticipated ways. As an example, the internet was developed for internal and largely military purposes, but in an unanticipated way massively improved productivity and living standards across the globe. Unfortunately, incorporating such

effects is notoriously difficult due to the unanticipated and uncertain nature of such effects. The model could be adjusted for such effects in reduced form, but conclusions would depend crucially on parameters and probability distributions that are very hard to calibrate.

References

Abel, N. H.: 1881, *Oeuvres Completes*, Grondahl and Son. editors: L. Sylow and S. Lie.

Aït-Sahalia, Y. and Saglam, M.: 2014, High frequency traders: Taking advantage of speed. NBER Working Paper.

Anand, A. and Venkataraman, K.: 2015, Market conditions, fragility and the economics of market making., *Journal of Financial Economics* . Forthcoming.

Biais, B., Declerck, F. and Moinas, S.: 2015, Who supplies liquidity, how and when? Working Paper.

Biais, B., Foucault, T. and Moinas, S.: 2015, Equilibrium fast trading, *Journal of Financial Economics* **116**(2), 292–313.

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

Budish, E., Cramton, P. and Shim, J.: 2015, The high-frequency trading arms race: Frequent batch auctions as a market design response, *The Quarterly Journal of Economics* **130**(4), 1621.

Carrion, A.: 2013, Very fast money: High-frequency trading on the nasdaq, *Journal of Financial Markets* **16**(4), 680–711.

Cespa, G. and Vives, X.: 2015, The beauty contest and shortterm trading., *The Journal of Finance* **70**(5), 2099–2154.

Constantinides, G. M.: 1986, Capital market equilibrium with transaction costs., *The Journal of Political Economy* pp. 842–862.

Cournot, A.-A.: 1838, *Recherches sur les principes mathmatiques de la thorie des richesses par Augustin Cournot.*, chez L. Hachette.

Foucault, T., Kadan, O. and Kandel, E.: 2013, Liquidity cycles and make/take fees in electronic markets, *The Journal of Finance* **68**(1), 299–341.

Hoffmann, P.: 2014, A dynamic limit order market with fast and slow traders, *Journal of Financial Economics* **113**(1), 156–169.

Jovanovic, B. and Menkveld, A. J.: 2015, Middlemen in limit order markets. Working Paper.

Korajczyk, R. A. and Murphy, D.: 2015, High frequency market making to large institutional trades. Working Paper.

Malinova, K., Park, A. and Riordan, R.: 2013, Do retail traders suffer from high frequency traders? Working Paper.

Menkveld, A. J. and Yueshen, B. Z.: 2012, Middlemen interaction and its effect on market quality. Working Paper.

Merton, R.: 1969, Lifetime portfolio selection under uncertainty: The continuous-time case., *Lifetime portfolio selection under uncertainty: The continuous-time case.* pp. 247–257.

Pham, H.: 2009, *Continuous-time stochastic control and optimization with financial applications*, Springer Science and Business Media.

Roger, L.-C.-G. and Zane, O.: 2002, A simple model of liquidity effects, *Advances in finance and stochastics* pp. 161–176.

Rojcek, J. and Ziegler, A.: 2016, High-frequency trading in limit order markets: Equilibrium impact and regulation. Working Paper.

SEC: 2010, Concept release on equity market structure. No. 34–61358; File No. S7–02–10;.

Yueshen, B. Z.: 2014, Queuing uncertainty in limit order market. Working paper.

A Figures

Figure 1: The relationship between the over-monitoring and the numbers of makers and takers. We restrict the values of the parameters, such that the marginal cost coefficient of the makers equals that of the takers: $\beta = \gamma$ and that the expected gain from a trade is normalized to 1: $G \equiv \pi_m + \pi_n = 1$. Then we make plots of the over-monitoring, that is the equilibrium intensity minus the first best intensity, of either side (row 1 for the maker intensity and row 2 for taker intensity), as a function of the maker's share of the total gains from trade: $\frac{\pi_m}{\pi_t}$ and the marginal cost coefficient $\beta = \gamma$. The flat surface is the 0 plane. All the part above 0 is thus over-monitoring, while that under 0 is under-monitoring. Each column of the figure is for one of the combinations of (M, N).

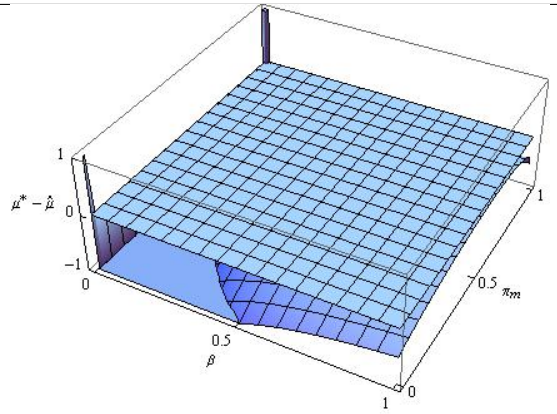
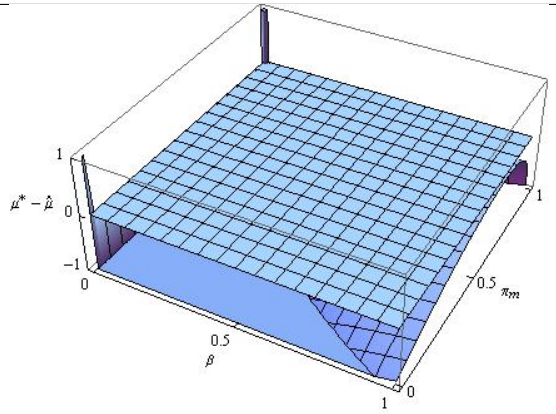
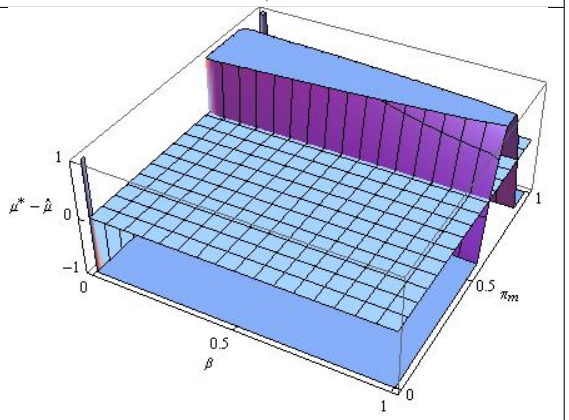
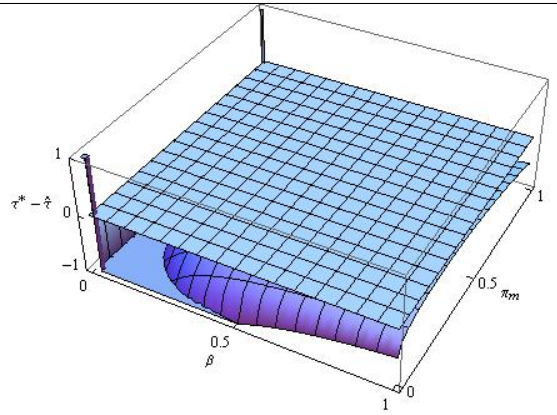
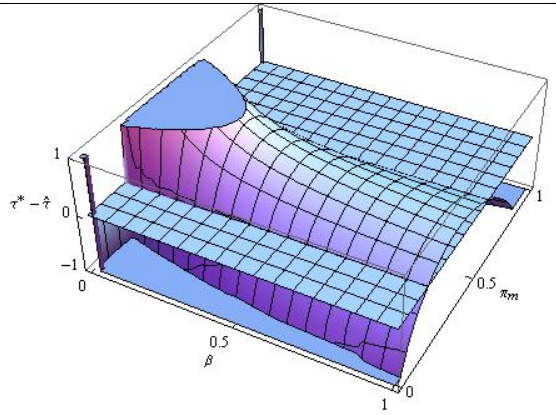
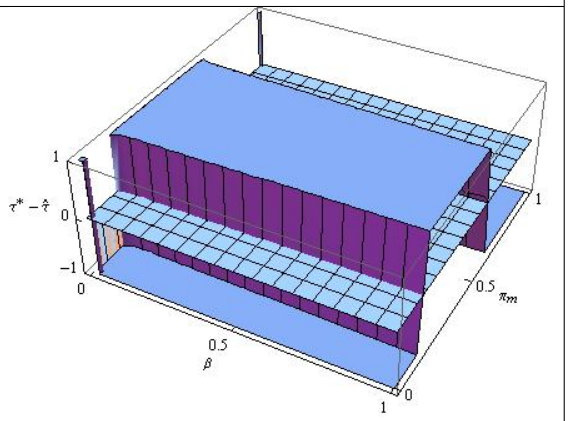
side of market	combinations of the numbers of the makers (M) and that of the takers (N)		
	$M = 2,\ N = 2$	$M = 2,\ N = 20$	$M = 50,\ N = 250$
Maker side:			
Taker side:			

Figure 2: The relationship between the over-monitoring and the numbers of makers and takers, when the expected gain from a trade $G(R) = -R + G_0$ is decreasing in R. Compared with Figure 1, the over-monitoring appears more severe.

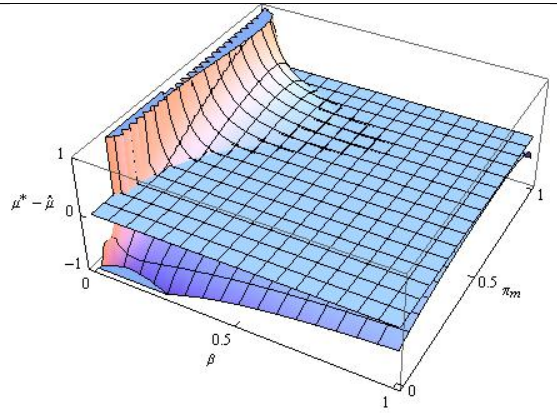
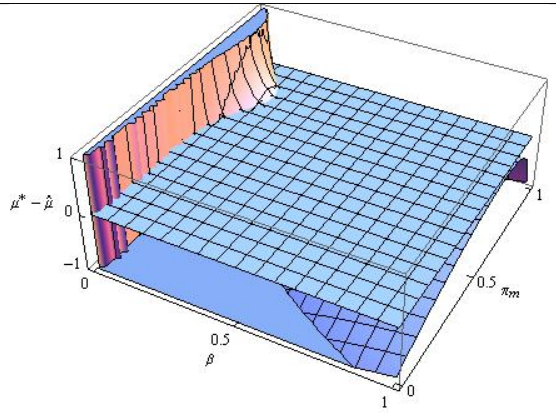
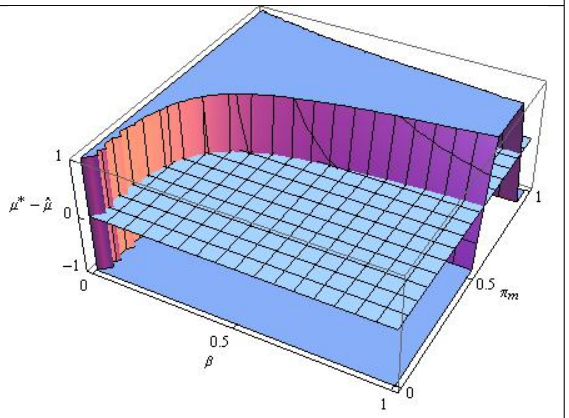
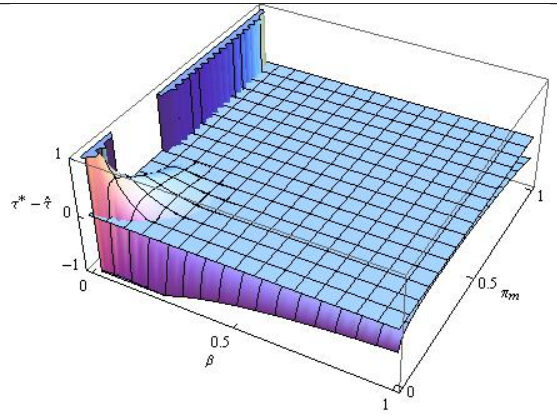
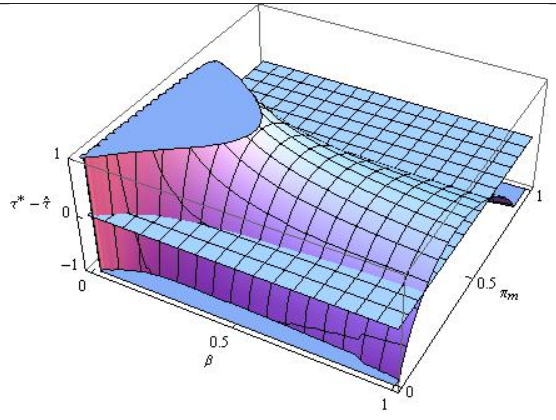
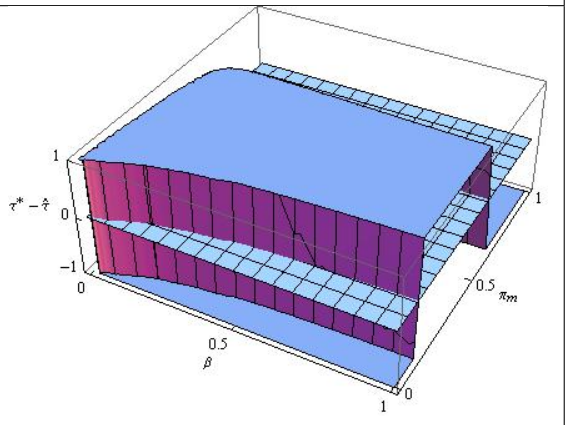
side of market	combinations of the numbers of the makers (M) and that of the takers (N)		
	$M = 2,\ N = 2$	$M = 2,\ N = 20$	$M = 50,\ N = 250$
Maker side:			
Taker side:			

Figure 3: The relationship between the over-monitoring and the numbers of makers and takers, when the expected gain from a trade $G(R) = -0.2R + G_0$ is decreasing but less steep than 2)

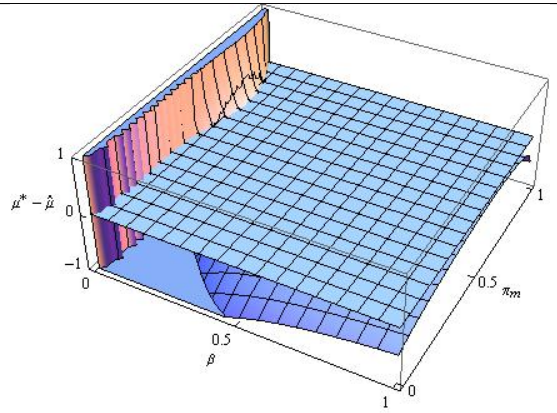
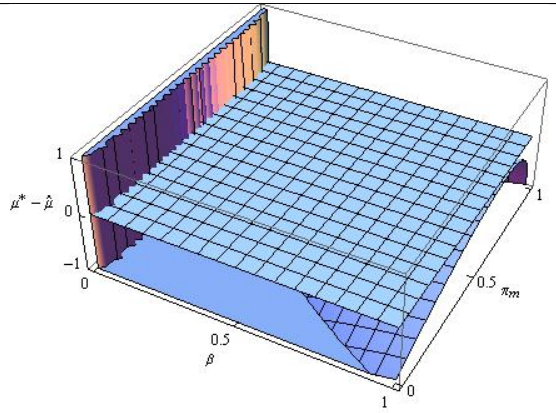
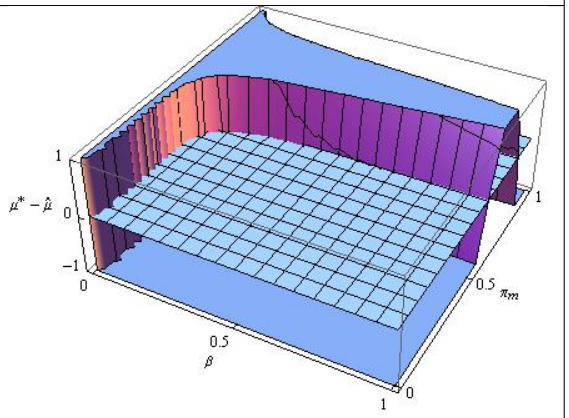
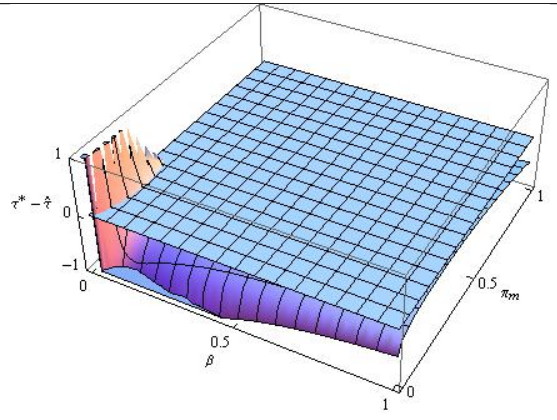
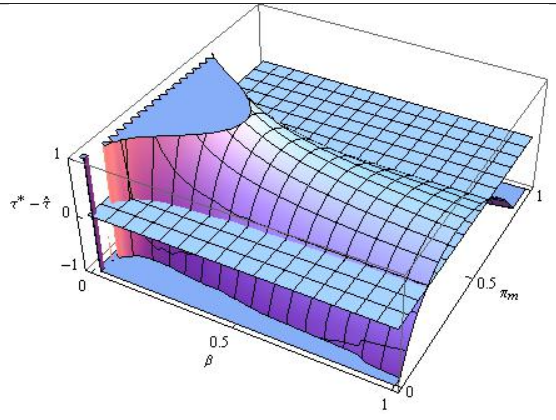
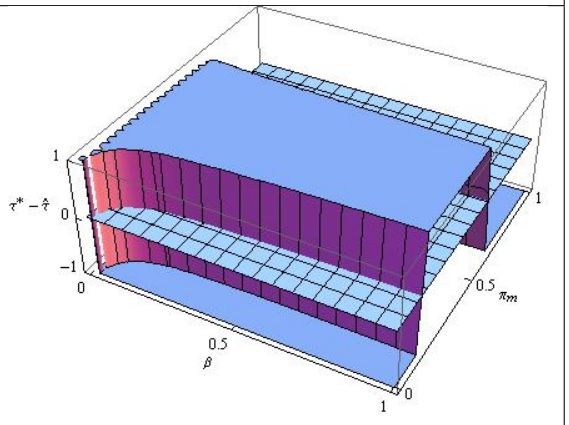
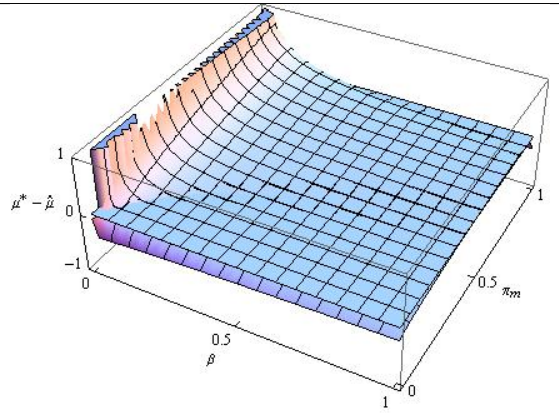
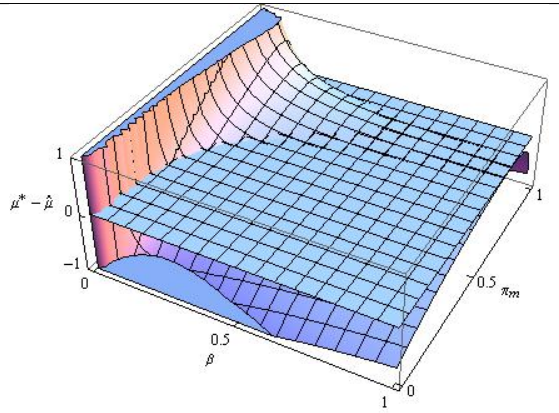
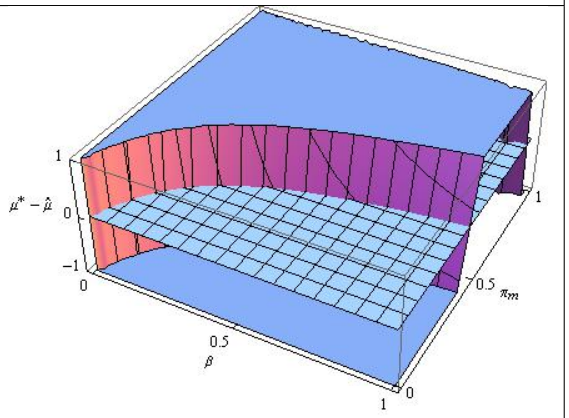
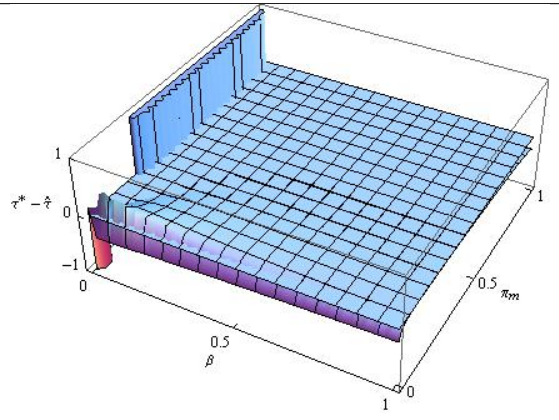
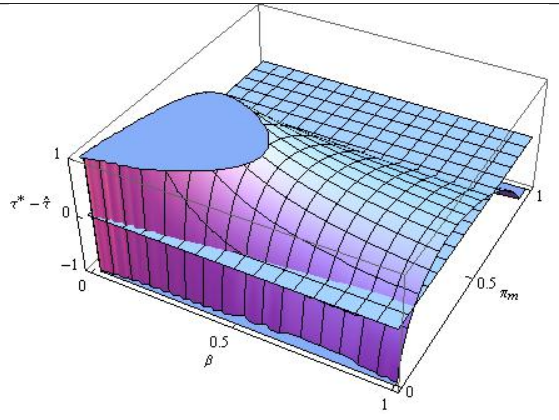
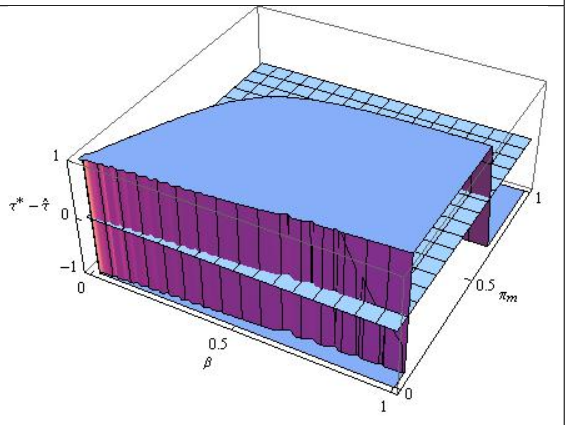
side of market	combinations of the numbers of the makers (M) and that of the takers (N)		
	$M = 2,\ N = 2$	$M = 2,\ N = 20$	$M = 50,\ N = 250$
Maker side:			
Taker side:			

Figure 4: The relationship between the over-monitoring and the numbers of makers and takers, when the expected gain from a trade $G(R) = -5R + G_0$ is decreasing, and steeper than Figure 2)

side of market	combinations of the numbers of the makers (M) and that of the takers (N)		
	$M = 2,\ N = 2$	$M = 2,\ N = 20$	$M = 50,\ N = 250$
Maker side:			
Taker side:			

B Proofs

Before analyzing the social planner's or the equilibria problem, we derive some mathematical conclusions:

$$\frac{\partial\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)}{\partial\mu_i}=\frac{\tau_j\bar{\tau}}{(\bar{\mu}+\bar{\tau})^2} \tag{23}$$

$$\frac{\partial\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)}{\partial\tau_j}=\frac{\bar{\mu}^2+\bar{\mu}(\bar{\tau}-\tau_j)}{(\bar{\mu}+\bar{\tau})^2} \tag{24}$$

and for any $k\neq j$:

$$\frac{\partial\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)}{\partial\tau_k}=-\frac{\tau_j\bar{\mu}}{(\bar{\mu}+\bar{\tau})^2} \tag{25}$$

Proof of Lemma 2:

For the SPP defined in 1, we take the first order conditions with respect to $\mu_i, i=1,2,\ldots,M$ and $\tau_j, j=1,2,\ldots,N$, respectively:

FOC–μ_i:

$$\beta\mu_i=\sum_{j=1}^{N}G'\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\frac{\tau_j^2\bar{\mu}\bar{\tau}}{(\bar{\mu}+\bar{\tau})^3}+\sum_{j=1}^{N}G\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\frac{\tau_j\bar{\tau}}{(\bar{\mu}+\bar{\tau})^2} \tag{26}$$

FOC:–τ_j:

$$\begin{aligned}\gamma\tau_j=&G'\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\frac{\bar{\mu}^2+\bar{\mu}(\bar{\tau}-\tau_j)}{(\bar{\mu}+\bar{\tau})^2}\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}+G\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\frac{\bar{\mu}^2+\bar{\mu}(\bar{\tau}-\tau_j)}{(\bar{\mu}+\bar{\tau})^2}\\&-\frac{\tau_j\bar{\mu}}{(\bar{\mu}+\bar{\tau})^2}\sum_{k\neq j}G'\left(\frac{\tau_k\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\frac{\tau_k\bar{\mu}}{\bar{\mu}+\bar{\tau}}-\frac{\tau_j\bar{\mu}}{(\bar{\mu}+\bar{\tau})^2}\sum_{k\neq j}G\left(\frac{\tau_k\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\end{aligned} \tag{27}$$

Due to the symmetry of the problem, the optimal monitoring intensity for all makers should

be equal, so are those of the takers. Thus we plug into 26 and 27 the symmetry conditions:

$$\begin{aligned}\mu_i = \hat{\mu} = \frac{\bar{\mu}}{M}, \quad \text{for all } i = 1, 2, \ldots, M; \\ \tau_j = \hat{\tau} = \frac{\bar{\tau}}{N}, \quad \text{for all } j = 1, 2, \ldots, N.\end{aligned} \tag{28}$$

FOC–μ_i:

$$\beta\hat{\mu} = NG'\left(\frac{M\hat{\mu}\hat{\tau}}{M\hat{\mu}+N\hat{\tau}}\right)\frac{M\hat{\mu}N\hat{\tau}^3}{(M\hat{\mu}+N\hat{\tau})^3} + NG\left(\frac{M\hat{\mu}\hat{\tau}}{M\hat{\mu}+N\hat{\tau}}\right)\frac{N\hat{\tau}^2}{(M\hat{\mu}+N\hat{\tau})^2} \tag{29}$$

FOC–τ_j:

$$\begin{aligned}\gamma\hat{\tau} = G'\left(\frac{M\hat{\mu}\hat{\tau}}{M\hat{\mu}+N\hat{\tau}}\right)\frac{M^2\hat{\mu}^2}{(M\hat{\mu}+N\hat{\tau})^2}\frac{M\hat{\mu}\hat{\tau}}{M\hat{\mu}+N\hat{\tau}} \\ + G\left(\frac{\hat{\tau}M\hat{\mu}}{M\hat{\mu}+N\hat{\tau}}\right)\frac{M^2\hat{\mu}^2}{(M\hat{\mu}+N\hat{\tau})^2}\end{aligned} \tag{30}$$

Although not in a closed form, given the functional form of G the first best allocation $(\hat{\mu}, \hat{\tau})$ can be simulated from the above simultaneous equations.

Similar to Foucault et al. (2013), we divide (29) by (30), and we get:

$$\frac{\hat{\mu}}{\hat{\tau}} = \left(\frac{\gamma}{\beta}\frac{N^2}{M^2}\right)^{\frac{1}{3}} \tag{31}$$

QED

Derivation of the equilibrium under the most general setup:

From the definition in 3, we obtain the FOCs:

FOC–maker i:

$$\beta\mu_i = \frac{\bar{\mu}+\bar{\tau}-\mu_i}{(\bar{\mu}+\bar{\tau})^2}\pi_m\sum_{j=1}^{N} G\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\tau_j + \frac{\mu_i\bar{\tau}}{(\bar{\mu}+\bar{\tau})^3}\pi_m\sum_{j=1}^{N} G'\left(\frac{\tau_j\bar{\mu}}{\bar{\mu}+\bar{\tau}}\right)\tau_j^2 \tag{32}$$

FOC–taker j:

$$\gamma\tau_j = \pi_t \frac{\bar{\mu}^2 + \bar{\mu}(\bar{\tau} - \tau_j)}{(\bar{\mu} + \bar{\tau})^2} \left[G'\left(\frac{\tau_j \bar{\mu}}{\bar{\mu} + \bar{\tau}}\right) \frac{\tau_j \bar{\mu}}{\bar{\mu} + \bar{\tau}} + G\left(\frac{\tau_j \bar{\mu}}{\bar{\mu} + \bar{\tau}}\right) \right] \tag{33}$$

To solve for symmetric equilibrium, plug 28 into 32 and 33, we get:

FOC–maker i:

$$\beta\hat{\mu} = \frac{M\hat{\mu} + N\hat{\tau} - \hat{\mu}}{(M\hat{\mu} + N\hat{\tau})^2} \pi_m N G\left(\frac{\hat{\tau} M\hat{\mu}}{M\hat{\mu} + N\hat{\tau}}\right)\hat{\tau} + \frac{\hat{\mu} N\hat{\tau}}{(M\hat{\mu} + N\hat{\tau})^3} \pi_m N G'\left(\frac{\hat{\tau} M\hat{\mu}}{M\hat{\mu} + N\hat{\tau}}\right)\hat{\tau}^2 \tag{34}$$

FOC–taker j:

$$\gamma\hat{\tau} = \pi_t \frac{M^2\hat{\mu}^2 + M\hat{\mu}(N\hat{\tau} - \hat{\tau})}{(M\hat{\mu} + N\hat{\tau})^2} \left[G'\left(\frac{\hat{\tau} M\hat{\mu}}{M\hat{\mu} + N\hat{\tau}}\right) \frac{\hat{\tau} M\hat{\mu}}{M\hat{\mu} + N\hat{\tau}} + G\left(\frac{\hat{\tau} M\hat{\mu}}{M\hat{\mu} + N\hat{\tau}}\right) \right] \tag{35}$$

Given the functional form of G, the symmetric equilibrium $(\hat{\mu}, \hat{\tau})$ can be numerically simulated with the above simultaneous equations.

The constant marginal GFT setup here is a special case of the general setup; with:

$$G\left(\frac{\bar{\mu}\bar{\tau}}{\bar{\mu} + \bar{\tau}}\right) = G_0 \tag{36}$$

$$G'\left(\frac{\bar{\mu}\bar{\tau}}{\bar{\mu} + \bar{\tau}}\right) = 0 \tag{37}$$

Proof of proposition 1:

Apply the above conditions 36 and 37 to the 29 and 30, we get:

$$\hat{\mu} = \frac{N^2}{(M\hat{r} + N)^2} \frac{G_0}{\beta} \tag{38}$$

$$\hat{\tau} = \frac{M^2 r^2}{(M\hat{r} + N)^2} \frac{G_0}{\gamma} \tag{39}$$

, where $\hat{r}$ can be computed from parameters as instructed in 31.

And the SPP total welfare:

$$\hat{\Pi} = G_0 \frac{M\hat{\mu} \cdot N\hat{\tau}}{M\hat{\mu} + N\hat{\tau}} - \frac{1}{2}\beta M\hat{\mu}^2 - \frac{1}{2}\gamma N\hat{\tau}^2 \tag{40}$$

QED

Proof of proposition 2:

Apply the linear GFT conditions 36 and 37 to the 34 and 35, we get:

$$\mu^* = \frac{M\mu^* + N\tau^* - \mu^*}{(M\mu^* + N\tau^*)^2} N\tau^* \pi_m \frac{G_0}{\beta} \tag{41}$$

$$\tau^* = \frac{M^2(\mu^*)^2 + M\mu^*(N-1)\tau^*}{(M\mu^* + N\tau^*)^2} \pi_t \frac{G_0}{\gamma} \tag{42}$$

Similar to Foucault et al 2013 A6-A11, all equilibria (EQ) are symmetric, and the EQ allocations are:

EQ monitoring intensity for each maker: $\mu^* = \frac{M+(M-1)r^*}{M(1+r^*)^2} \frac{\pi_m}{\beta}$;

EQ monitoring intensity for each taaker: $\tau^* = \frac{r^*((1+r^*)N-1)}{N(1+r^*)^2} \frac{\pi_t}{\gamma}$;

where: r^* (intepreted as $r^* \equiv \frac{M\mu^*}{N\tau^*}$) is the positive real solution to:

$$Nr^3 + (N-1)r^2 - (M-1)zr - Mz = 0 \tag{43}$$

where $z \equiv \frac{\gamma}{\beta} \frac{\pi_m}{\pi_t}$. EQ "total welfare", defined as the sum of all agents' profit:

$$\Pi^* = \Gamma \frac{M\mu^* \cdot N\tau^*}{M\mu^* + N\tau^*} - \frac{1}{2}\beta M(\mu^*)^2 - \frac{1}{2}\gamma N(\tau^*)^2 \tag{44}$$

QED

Proof of proposition 3:

We prove this proposition by proving two Lemmas, then apply the logic of deduction and prove the proposition.

Lemma 6 *If* $J(W_0(0), W_1(0), 0|\lambda_j, \pi_m = 0)$ *is concave in* λ_j, *then so is* $J(W_0(0), W_1(0), 0|\lambda_j, \pi_m)$, *for any* $\pi_m > 0$

Lemma 7 *If* $J(W_0(0), W_1(0), 0|\lambda_j, \pi_m)$ *is concave in* λ_j *for any* $\pi_m > 0$, *then:* $G(\lambda_j, \pi_m)$ *decreases in* λ_j, *for any* $\pi_m > 0$.

Before proving these two lemmas, we first express $J(W_0(0), W_1(0), 0|\lambda_j, \pi_m)$ recursively.

By definition,

$$
\begin{aligned}
& J(W_0(0), W_1(0), 0|\lambda_j, \pi_m) \\
= & \max_{\{C(s)\}_{0\le s\le s_1}} \mathbb{E}_{t=0} \int_0^{s_1} e^{-\rho s} U(C(s)) ds \\
& + \mathbb{E}_{t=0} e^{-\rho s_1} \left[(1-\pi_m) g(Q^*(s_1)) + J(W_0(s_1), W_1(s_1), s_1|\lambda_j, \pi_m) \right] \\
= & \max_{\{C(s)\}_{0\le s\le s_1}} \mathbb{E}_{t=0} \int_0^{s_1} e^{-\rho s} U(C(s)) ds \\
& + \mathbb{E}_{t=0} e^{-\rho s_1} (1-\pi_m) g(Q^*(s_1)) \\
& + \mathbb{E}_{t=0} \max_{\{C(s)\}_{s_1\le s\le s_2}} \mathbb{E}_{t=s_1} \int_{s_1}^{s_2} e^{-\rho s_1} e^{-\rho s} U(C(s)) ds \\
& + \mathbb{E}_{t=0} \mathbb{E}_{t=s_1} e^{-\rho s_2} J(W_0(s_2), W_1(s_2), s_2|\lambda_j, \pi_m) \\
& + \mathbb{E}_{t=0} \mathbb{E}_{t=s_1} e^{-\rho s_2} (1-\pi_m) g(Q^*(s_2)) \\
& \text{(Iterate the above process by } N-1 \text{ times...;)} \\
& \xrightarrow{n\to+\infty} \\
& \sum_{n=1,2,...} \max_{\{C(s)\}_{s_{n-1}\le s\le s_n}} \mathbb{E}_{t=0} \int_{s_{n-1}}^{s_n} e^{-\rho s_{n-1}} e^{-\rho s} U(C(s)) ds \\
& + 0 \cdot \mathbb{E}_{t=0} J(W_0(s_N), W_1(s_N), s_N|\lambda_j, \pi_m) \\
& + \mathbb{E}_{t=0} \sum_{n=1,2,...} e^{-\rho s_n} (1-\pi_m) g(Q^*(s_n)) \\
= & J(W_0(0), W_1(0), 0|0, \pi_m) \\
& + (1-\pi_m) G(\lambda_j, \pi_m) \mathbb{E}_{t=0} \sum_{n=1,2,...} e^{-\rho s_n} \\
& \text{(by definition: } G(\lambda_j, \pi_m) \equiv \mathbb{E}_{t=0} g(Q^*(s_n))\text{)}
\end{aligned}
$$

Evaluate the second term:

$$
\begin{aligned}
&(1-\pi_m)G(\lambda_j,\pi_m)\mathbb{E}_{t=0}\sum_{n=1,2,...} e^{-\rho s_n} \\
=&(1-\pi_m)G(\lambda_j,\pi_m)\sum_{n=1,2,...}\mathbb{E}_{t=0}e^{-\rho s_n} \\
=&(1-\pi_m)G(\lambda_j,\pi_m)\sum_{n=1,2,...}\int_0^\infty e^{-\rho s_n}\frac{\lambda_j}{\Gamma(n)}s_n^{n-1}e^{-\lambda_j s_n}ds_n \\
&\text{(We plug in the pdf of } Gamma(n,\lambda) \text{ distribution)} \\
=&(1-\pi_m)G(\lambda_j,\pi_m)\sum_{n=1,2,...}\left(\frac{\lambda_j}{\rho+\lambda_j}\right)^n \\
=&\frac{1-\pi_m}{\rho}G(\lambda_j,\pi_m)\lambda_j
\end{aligned}
$$

Thus:

$$J(W_0(0),W_1(0),0|\lambda_j,\pi_m) = J(W_0(0),W_1(0),0|0,\pi_m) + \frac{1-\pi_m}{\rho}G(\lambda_j,\pi_m)\lambda_j \tag{45}$$

Now we prove the two lemmas.

Proof of Lemma 6:

Since 45 applies for any $1 \geq \pi_m \geq 0$, Then:

$$J(W_0(0),W_1(0),0|\lambda_j,0) = J(W_0(0),W_1(0),0|0,\pi_m) + \frac{1}{\rho}G(\lambda_j,\pi_m)\lambda_j \tag{46}$$

Notice that $J(W_0(0),W_1(0),0|0,\pi_m)$ is flat with respect to λ_j, thus the concavity of $J(W_0(0),W_1(0),0|\lambda_j,0)$ implies the concavity of $\frac{1}{\rho}G(\lambda_j,\pi_m)\lambda_j$, which in turn implies the concavity of $J(W_0(0),W_1(0),0|\lambda_j,\pi_m)$. **QED for Lemma 6**

Proof of Lemma 7:

We prove it by contradiction. Suppose that $G(\lambda_j,\pi_m)$ does not monotonically decrease in λ_j, then by the continuity of $G(\lambda_j,\pi_m)$ in λ_j, there must exist an interval of λ_j inside $[0,+\infty)$ on which $G(\lambda_j,\pi_m)\lambda_j$ is (locally) convex in λ_j, a contradiction to the concavity of $J(W_0(0),W_1(0),0|\lambda_j,\pi_m)$. **QED for Lemma 7**

QED for proposition 3

Proof of proposition 4:

Closed-form solution is not currently possible even for the problem without transaction cost Pham (2009), but the lack of the transaction cost allows for[29] asymptotic expansion of the solution, same as Roger and Zane (2002). The value function of our portfolio choice and consumption problem, $J[x, y, 0|\lambda_j, 0]$, can be expressed by that of the Merton problem Merton (1969), adjusted by a series of small numbers:

$$J[x, y, 0|\lambda_j, 0] = a^{-\delta}U(x + y) + \sum_{n=1}^{\infty} k_n \frac{l^n}{n!} \tag{47}$$

, where$a^{-\delta}U(x + y)$ is the value function of the Merton problem, where $a^{-\delta} \equiv \frac{1}{\delta}\left(\rho + (\delta - 1)\left[r + \frac{(\alpha-r)^2}{2\delta\sigma^2}\right]\right)$ is a coefficient entirely composed of exogenous parameters. In addition, $l \equiv \lambda_j$ is the latency of the trades for taker j, where the subscript is suppressed in the left hand side. $b_n, n \geq 1$ are the coefficients that we try to determine in the expansion practice.

Following Roger and Zane (2002), we can solve for:

$$b_1 = -\frac{\sigma^6\delta^3 w^2(1 - w)^2}{2\sigma^2\delta\rho + (\delta - 1)[2\delta\sigma^2 r + (\alpha - r)^2]} \tag{48}$$

where $w \equiv \frac{\alpha-r}{\sigma^2\delta}$ is the Merton proportion, which is entirely composed of exogenous parameters of this problem. Thus it's obvious that always we have $b_1 < 0$.

Assuming that latency l is very small, a fact very much true in the context of HFTs, terms $b_2, b_3, \ldots$ vanishes leaving the value function influenced only by the first order of the latency:

$$J[x, y, 0|\lambda_j, 0] \approx a^{-\delta}U(x + y) + b_1 l = a^{-\delta}U(x + y) + b_1\lambda_j \tag{49}$$

Then it can be seen by taking derivative that the value function is asymptotically concave in trading speed:

$$\frac{\partial^2 J[x, y, 0|\lambda_j, 0]}{\partial(\lambda_j)^2} < 0, \text{for} \lambda_j \text{big enough} \tag{50}$$

QED

[29]By the way we construct the transaction cost, asymptotic expansion is not possible for the problem with transaction cost.

C Notation Summary

Symbol	*Support*	*Description*
	Parameters	
M	$[1, \infty)$	the total number of makers
N	$[1, \infty)$	the total number of takers
G_0	$[0, \infty)$	marginal gain per trade in linear GFT setting
π_m	$[0, 1]$	share of gains from trade that is captured by the maker
π_t	$[0, 1]$	share of gains from trade that is captured by the taker
β	$[0, \infty)$	technological cost parameter for each maker
γ	$[0, \infty)$	technological cost parameter for each taker
r	$[0, \infty)$	the risk free rate
α, σ	$[0, \infty)$	the parameters of the Brownian motion
δ	$[0, 1) \cup (1, \infty)$	the representative taker's relative risk aversion.
	States of nature	
$X_0(t)$	$[0, \infty)$	the price trajectory of the riskless asset accumulation
$X_1(t)$	$[0, \infty)$	the price trajectory of the stock price
$B(t)$	$[0, \infty)$	Geometric Brownian Motion serving as control process of $S(t)$.
$P(t)$	$[0, \infty)$	representative taker's Poisson process recording transactions
	Indices	
i	$\{1, 2, .., M)\}$	maker i
j	$\{1, 2, .., N)\}$	taker j
t	$\{0, .., \infty\}$	time
s_n	$\{1, 2, .., \infty\}$	representative taker's trading time
	Decision variables	
μ_i	$[0, \infty)$	the intensity parameter of maker i.
τ_j	$[0, \infty)$	the intensity parameter of taker j.
$W_0(t)$	$[0, \infty)$	the wealth invested in the riskless asset at time t.
$W_1(t)$	$[0, \infty)$	the wealth invested in the stock at time t.
$Q(t)$	$(-\infty, \infty)$	the amount transferred from the stock to the riskless asset at time t (when trading is possible).
$C(t)$	$[0, \infty)$	the amount of consumption at time t.