No. 2265

IMPERFECT MARKET-MONITORING AND SOES TRADING

Thierry Foucault, Ailsa Röell and Patrik Sandås

FINANCIAL ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

IMPERFECT MARKET-MONITORING AND SOES TRADING

Thierry Foucault, Hautes Études Commerciales, Jouy en Jousas and CEPR
Ailsa Röell, Department of Economics, London School of Economics
Patrik Sandås, University of Pennsylvania

Discussion Paper No. 2265
October 1999

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: http://www.cepr.org

CEPR Discussion Paper No. 2265

October 1999

ABSTRACT

Imperfect Market-Monitoring and SOES Trading*

We develop a model of price formation in a dealership market where monitoring of the information flow requires costly effort. The result is imperfect monitoring, which creates profit opportunities for speculators, who do not act as dealers but simply monitor the information flow and quote updates in order to pick off ‘stale quotes’. Externalities associated with monitoring can help to sustain non-competitive spreads. We show that protecting dealers against the execution of stale quotes can result in larger spreads and be detrimental to price discovery due to externalities in monitoring. A reduction in the minimum quoted depth will reduce the spread and speculators’ trading frequency. Our analysis is relevant for the Small Order Execution System (SOES) debate given that the behaviour of speculators in our model is very similar to the alleged behaviour of the real world SOES ‘bandits’.

JEL Classification: G10, G14, G24, L13
Keywords: monitoring, bid-ask spread, automatic execution, SOES

Thierry Foucault
HEC School of Management
1 rue de la Libération
78351 Jouy en Josas
FRANCE
Tel: (33 1) 39 67 94 11
Fax: (33 1) 39 67 70 85
Email: foucault@hec.fr

Ailsa Röell
Department of Economics
Princeton University
Princeton, NJ 08544
Tel: (1 609) 248 4033
Fax: (1 609) 258 6419
Email: a.roell@princeton.edu.

Patrik Sandås
Finance Department
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104
Tel: (1 215) 898 1697
Fax: (1 215) 898 6200
E-mail: sandas@wharton.upenn.edu

* We are grateful for many detailed comments by Dan Bernhardt and Ingrid Werner. We thank Bruno Biais, Michael Fishman, Burton Hollifield, Eugene Kandel, Ken Kavajecz, Chester Spatt, Erik Theissen and seminar participants at the Wharton Lunch Seminar, the 1999 WFA meetings and the 1999 European Summer Symposium in Financial Markets at Gerzensee for useful suggestions and comments. We also thank CEPR for their hospitality at the 1998 European Summer Symposium in Gerzensee, where this research originally began. Financial support from the Rodney L White Center for Financial Research and the Fondation HEC is gratefully acknowledged.

Submitted 20 September 1999

NON-TECHNICAL SUMMARY

Nasdaq's Small Order Execution System (SOES) allows brokerage firms to automatically execute small orders at the best quotes posted by Nasdaq dealers. Participation in SOES is mandatory for all dealers, who must post firm quotes valid up to a maximum quantity, fixed by Nasdaq. Although it was intended for small retail customers, SOES mainly attracted professional day-traders (labelled SOES 'bandits' by market-makers). SOES day-traders quickly place orders when they observe a shift in the value of the asset, either because they become aware of new public information before the dealers or because some dealers are slow to update their quotes. In either case, they increase the adverse selection risk faced by dealers. The role of the SOES and its alleged impact on Nasdaq trading costs, liquidity and volatility has been the subject of a long and heated policy debate. In particular Nasdaq dealers argued that SOES day-traders were the cause of the large spreads observed on Nasdaq. In contrast day-traders claimed that their presence improved price discovery by forcing dealers to closely monitor their quotes.

In this article, we develop a theoretical model in which *imperfect monitoring* by market-makers creates profit opportunities for speculators who compare with SOES day-traders. We show how imperfect monitoring affects price formation and we address, theoretically and empirically, some of the issues raised in the debate regarding the impact of SOES day-traders on market quality. We distinguish between two forms of monitoring: (i) *market-monitoring* and (ii) *quote-monitoring*. Market-monitoring entails monitoring the arrival of new information, e.g. public announcements, whereas quote-monitoring is limited to monitoring quote updates. Traders learn new information through these two forms of monitoring. Monitoring information arrival requires some effort, however. In contrast, quote-monitoring does not require any effort because it can be automated.

In our model, market-makers post firm quotes and select how intensively they monitor information arrival. Since market-monitoring is costly, they never monitor news continuously. Imperfect monitoring of news by the market-makers creates occasional profit opportunities due to 'stale' quotes. A second group of agents, referred to as speculators, seek to exploit these profit opportunities. These speculators behave like the SOES day-traders. They monitor news arrivals and the dealers' quote updates. When the speculators observe new information or a quote revision indicating a change in the asset value, they 'pick off' dealers that fail to adjust their quotes. In equilibrium, consistent with stylized facts, speculators' expected profits are positive.

We find that market-monitoring by one dealer can generate either a *positive* or a *negative externality* for the other dealers. By monitoring quote updates, a dealer can free ride on the efforts that his competitors exert to monitor the market. This is the source of the positive externality. The negative externality stems from the fact that speculators can use quote updates to 'discover' stale quotes. Therefore, more market-monitoring by one dealer increases the likelihood that a speculator will be able to 'pick off' other dealers before they update their quotes.

These externalities influence dealers' bidding behaviour. The positive externality induces dealers to match the best offers in the market rather than undercutting. As a consequence, dealers can capture strictly positive expected profits in equilibrium even if they compete in prices, *à la Bertrand*. In contrast, the negative externality increases the competitive pressure. In fact, by undercutting, a dealer can induce other dealers to monitor less aggressively, which reduces the negative externality.

This interaction between price formation and information externalities has implications for market design. Automatic execution makes it easier for speculators to trade based on quote updates since dealers can not 'back away' from trading at posted prices. We show that this feature of automatic execution weakens dealers' incentive to free ride on information produced by their competitors. As a result, automatic execution encourages dealers to compete in prices and tightens the spread. It also strengthens their incentive to monitor the market, which ultimately is beneficial to price discovery.

Interestingly, the SOES's automatic execution feature has been a major bone of contention between day-traders and Nasdaq dealers. Accordingly, Nasdaq has attempted to eliminate this feature several times. The policy debate has also focused on the effect of SOES on trading costs on Nasdaq. Two intertwined questions have been considered both by regulators and researchers. Namely, whether the spreads are largely due to trading by SOES day-traders and what determines the level of SOES day trading activity. Based on the comparative statistics for the spread and the level of speculator activity in our model we address these two questions empirically. We do not find evidence of higher spreads for stocks with higher levels of SOES activity. This surprising result is inconsistent with the theoretical model and with the claim that Nasdaq spreads are largely because of SOES day-traders. We do find that the level of SOES activity is negatively related to the spread and is positively related to volatility and the minimum quoted depth as predicted by the model.

The implications of our theoretical findings are discussed in the context of the controversy on the effects of SOES trading. The economic insights that our model provides apply more generally, however. For example, investors submitting limit orders are typically exposed to a 'picking off' risk. As the dealers in our model, limit order traders face a trade-off between the risk of being picked off and the costs of monitoring the information flow. Free-riding behaviour induced by the possibility of learning new information from quote changes is also observed in other market structures. For instance, dealers in US regional exchanges have an incentive to match NYSE specialists' quotes and thereby to free ride on their investment in price discovery. Analysing the effect of the externalities induced by market- and quote-monitoring is therefore important to sharpen our understanding of price formation in financial markets.

1 Introduction

Nasdaq's Small Order Execution System (SOES) allows brokerage firms to automatically execute small orders at the best quotes posted by Nasdaq dealers. Participation in SOES is mandatory for all dealers, who must post firm quotes valid up to a maximum quantity, fixed by Nasdaq.[1] Although it was intended for small retail customers, SOES mainly attracted professional day traders (labeled SOES "bandits" by market makers), who account for a large proportion of the SOES trading volume. The bandits trade when they observe a shift in the value of the asset, either because they become aware of new public information before the dealers or because some dealers are slow to update their quotes.[2] The role of the SOES and its alleged impact on Nasdaq trading costs, liquidity, and volatility has been the subject of a long and heated policy debate.[3]

Harris and Schultz (1998) show that bandits on average make positive trading profits, at the expense of dealers. Profitable trading by the bandits is puzzling since they trade on information, which is publicly available, and pay commissions on their trades. Harris and Schultz (1998), p.61, suggest that imperfect monitoring by dealers is a potential explanation:

> *The existence and profitability of SOES bandits raise new questions about the efficiency of different market structures. Bandits do not have any more information than the market makers that they trade against and in many cases they have less information. But bandits still make money. [...]. We believe the answer is that market makers are inherently less efficient at price discovery than are bandits. [...] bandits have a much greater incentives to concentrate on what they are doing, to follow stock prices closely, and to stay in front of their terminals than do market maker employees. Unusually fast or skillful traders may find SOES trading to be more profitable than working for a Nasdaq market maker.*

[1] The minimum quoted depth in the SOES has varied between 100 and 1000 shares since 1987.

[2] See Harris and Schultz (1997) for a detailed description of the SOES "bandits'" trading strategies.

[3] In a Washington Post article, on February 7, 1994, Joseph Hardiman, president of the National Securities Dealers Association said that 'The SOES activists were picking off market makers, who were slow to adjust. The losses to SOES activists made market makers gun shy, causing them to widen their price spreads.' In a testimony before the House Committee on Commerce, David Whitcomb, argued that 'Abolishing SOES would remove the 'market discipline', which keeps market makers on 'their toes' and causes prices to rapidly adjust when news occurs.' See the GAO report on 'The Effects of SOES on the Nasdaq Market' for a summary of the main arguments in the SOES controversy and important SOES-related events.

In this paper, we develop a theoretical model of imperfect monitoring by market makers and show how this friction affects price formation. We distinguish between two forms of monitoring: (i) *market monitoring* and (ii) *quote monitoring*. Market monitoring entails monitoring the arrival of new information, e.g., public announcements, whereas quote monitoring is limited to monitoring quote updates. Traders learn new information through these two forms of monitoring. Monitoring information arrival requires some effort. In contrast, quote monitoring does not require any effort because it can be automated.

In our model, market makers post firm quotes and select how intensively they monitor information arrival. Since market monitoring is costly, they never monitor news continuously. Imperfect monitoring of news by the market makers creates occasional profit opportunities due to "stale" quotes. A second group of agents, referred to as speculators, seek to exploit these profit opportunities. These speculators behave like the SOES bandits. They monitor news arrival and the dealers' quote updates. When the speculators observe new information or a quote revision indicating a change in the asset value, they "pick off" dealers that fail to adjust their quotes. In equilibrium, speculators' expected profits are positive. This provides an explanation for the puzzling stylized fact reported in Harris and Schultz (1998). This explanation does not rely on differences in monitoring costs, differences in ability, or agency problems in trading.[4] The market maker's losses from trading with the speculators are offset by gains from trading with liquidity traders.

Our main results are:

1. Market monitoring by one dealer can generate either *a positive or a negative externality* for the other dealers. By monitoring quote updates, a dealer can free ride on the efforts that his competitors exert to monitor the market. This is the source of the positive externality. The negative externality stems from the fact that speculators can use quote updates to "discover" stale quotes. Therefore, more market monitoring by one dealer increases the likelihood that a speculator will be able to "pick off" other dealers before they update their quotes.

[4]However, the model can be extended to consider agency issues in trading as well as differences in skills. The monitoring decision can be interpreted as an unobservable (and non-contractible) choice of effort made by the trader, who is an employee at a market making firm.

2. These externalities influence dealers' bidding behavior. The positive externality induces dealers to match the best offers in the market rather than undercutting. As a consequence, dealers can capture strictly positive expected profits in equilibrium even if they compete in prices, *à la Bertrand*. In contrast, the negative externality increases the competitive pressure. In fact, by undercutting, a dealer can induce other dealers to monitor less aggressively, which reduces the negative externality.

3. This interaction between price formation and information externalities has implications for market design. Automatic execution makes it easier for speculators to trade based on quote updates since dealers can not "back away" from trading at posted prices. We show that this feature of automatic execution weakens dealers' incentive to free ride on information production by their competitors. As a result, automatic execution encourages dealers to compete in prices and tightens the spread. It also strengthens their incentive to monitor the market, which ultimately is beneficial to price discovery.

Interestingly, the SOES's automatic execution feature has been a major bone of contention between bandits and Nasdaq dealers. Accordingly, Nasdaq has attempted to eliminate this feature several times. The policy debate has also focused on the effect of SOES on trading costs on Nasdaq. Two intertwined questions have been considered both by regulators and researchers. Namely, whether the spreads are larger due to trading by SOES bandits and what determines the level of SOES bandit activity. Based on the comparative statics for the spread and the level of SOES activity in our model we address these two questions empirically. We do not find evidence of higher spreads for stocks with higher levels of SOES activity. This surprising result is inconsistent with the theoretical model and with the claim that Nasdaq spreads are large because of SOES bandits. We do find that the level of SOES activity is negatively related to the spread and is positively related to volatility and the minimum quoted depth as predicted by the model.

Our model is most closely related to Copeland and Galai (1983), who analyze the free-trading option aspect of fixed quotes. We show how the free-trading option problem arises in equilibrium as a result of imperfect monitoring decisions by market makers. Kandel and Marx (1998) develop a theoretical model to study whether odd-eighth avoidance is a rational response by Nasdaq dealers to SOES bandits. In their model the profit opportunities of the SOES bandits are *implicitly* assumed to be due to imperfect monitoring by the dealers. Our contribution is to *explicitly* model imperfect market monitoring and analyze its impact on the spread and the level of SOES activity in equilibrium.

Battalio, Hatch, and Jennings (1997), using vector autoregressions, show that SOES bandits may speed up the price discovery process and that SOES bandits' activity is positively related to price volatility. We obtain results consistent with these empirical findings. Harris and Schultz (1997) analyze trading in the 20 largest Nasdaq stock around a rule change that decreased the maximum SOES trade size from 1000 to 500 shares. They provide evidence suggesting that the rule change had its intended effect of reducing the marker makers' losses to SOES bandits. In our model, a decrease in the minimum quoted depth lead to the entry of fewer speculators and thus to lower losses for the dealers.

The implications of our theoretical findings are discussed in the context of the controversy on the effects of SOES trading. The economic insights that our model provides apply more generally, however. For example, investors submitting limit orders are typically exposed to a "picking off" risk. As the dealers in our model, limit order traders face a trade off between the risk of being picked off and the costs of monitoring the information flow. Free riding behavior induced by the possibility of learning new information from quote changes are also observed in other market structures. For instance, dealers in U.S. regional exchanges have an incentive to match NYSE specialist's quotes and thereby to free ride on his investment in price discovery (see Amihud and Mendelson (1991)). Analyzing the effect of the externalities induced by market and quote monitoring is therefore important to sharpen our understanding of price formation in financial markets.

The rest of the paper is organized as follows. The general features of the model are presented in the next section. In Section 3, we show that market monitoring by one dealer can be a positive or a negative externality for the other dealers. In Section 4, the equilibrium with a fixed number of speculators is analyzed. In Section 5, we study the effect of automatic execution on spreads and price discovery. We analyze the factors that determine the number of speculators and the spread in Section 6. An empirical study of the model's predictions is presented in Section 7. The final section concludes. All proofs are in the appendix.

2 The Model

2.1 Timing, Traders and Market Structure

We consider the market for a risky asset. The liquidation value of the risky asset is uncertain and is denoted $\tilde{V}$. There are three types of traders in this market: (i) $M \geq 2$ *dealers*, who post quotes and monitor the market, (ii) $N \geq 1$ *speculators*, who monitor the market and submit market orders when they perceive profit opportunities, and (iii) *liquidity traders*. Let $\mathcal{M}$ and $\mathcal{N}$ denote the set of all dealers and all speculators, respectively. The expected value of the asset at the beginning of the trading day is denoted v_0.

The trading day consists of a sequence of trading rounds. Figure 1 depicts the different stages in a trading round. Each trading round comprises three stages. In the first stage, *the quoting stage*, the risk neutral dealers *simultaneously* determine their quoted spread. We denote by S_i the spread that is posted by dealer i. We assume that the dealer bid and ask quotes are centered around the asset's expected value. The bid quote posted by dealer i is $b_i = v_0 - \frac{S_i}{2}$ and the ask quote is $a_i = v_0 + \frac{S_i}{2}$. Let $S_b = Min\{S_i\}_{i=1}^{i=M}$ be the *market spread*, i.e., the lowest spread posted in the market and let $\mathcal{M}_b$ be the subset of dealers posting this spread. We assume that the dealers who do not post the market spread (i.e., the best ask and bid quotes) do not participate in the next stages of the trading round.[5] The market organization is such that the dealers posting the market spread are required to honor their quotes for up to Q shares, the minimum quoted depth, of the risky asset and that execution is automatic for all order sizes of Q or smaller. For orders larger than Q, dealers can back away from their quotes.

In the second stage, *after* observing the quotes posted in the market, the dealers who

[5]Alternatively we could assume that, after the quoting stage, the dealers can choose whether or not they wish to maintain their quotes. In this case, the dealers who are not posting the best quotes will optimally decide not to participate in the subsequent stages of the trading round. In practice, a dealer who does not want to trade can post a very large spread so as to deter traders from hitting his quotes.

remain active and the speculators choose their monitoring levels. The monitoring level chosen by a trader determines the probability that she is the first to discover a public announcement regarding the asset value. The monitoring decision is described in more detail below. We refer to the second stage as *the monitoring stage.*

In the third stage, *the trading stage*, one of the three following events occurs. With probability $\alpha < 1$, information arrives indicating an increase or a decrease in the asset value, with equal probabilities. In the case of an increase, the new expected value becomes: $v_1 = v_0 + \frac{\sigma}{2}$, whereas in the case of a decrease the new expected value becomes: $v_1 = v_0 - \frac{\sigma}{2}$. With probability $(1 - \alpha)$, no information arrives. In this case, with probability $\beta > 0$, a buy or a sell market order is submitted by a liquidity trader, with equal probabilities. The expected size of the liquidity trader's order is δQ. Finally with probability $(1 - \beta)$, no order is submitted. In the case of new information, the first trader who reacts to the new information is denoted by f. If f is a dealer then he updates his quotes. If f is a speculator then she can submit a market order. In all the cases, a market order is split equally among the dealers who post *the best quotes*, thus, if M_b dealers are tied at the best price, each dealer buys or sells a fraction $1/M_b$ of the market order.

A trading round ends either when a transaction occurs or when all the dealers have updated their quotes. For brevity, we will just focus on a single trading round since all the trading rounds are replications of the same game.

Our assumptions closely match some of the key features of the Nasdaq's SOES trading system. The quantity, Q, is the minimum quoted depth. The speculators can be thought of as the SOES bandits. A speculator cannot trade more than one time against the dealers. This assumption is consistent with Nasdaq rules, which prohibit SOES traders from initiating more than one position in the same stock within five minutes. Another feature of Nasdaq is that dealers execute, at their posted quotes, orders that are larger than the minimum quoted depth. SOES bandits typically do not take part in these trades since they are negotiated by phone. This slows down the execution process and dealers can back away from their quote upon realizing that the counter-party is a bandit (See Harris and Schultz (1997) and Houtkin (1998)). This suggests that, in our model, there is no reason to restrict our attention to values of δ that are lower than one. SOES bandits make profits by reacting more quickly than dealers to public news announcements.[6] This is possible because dealers do not perfectly monitor the arrival of public information. We will now describe the monitoring activity in more detail.

2.2 Market Monitoring and Quote Monitoring

In each period new information, which will change the asset value, arrives with probability α. Dealers and speculators can become aware of new information by directly monitoring the information flow, an activity that we call *market monitoring*. We model market monitoring as follows. Let $\lambda_i \in [0, +\infty)$ be *the monitoring level* of market-maker $i \in \mathcal{M}$ and let $\gamma_j \in [0, +\infty)$ be the monitoring level of speculator $j \in \mathcal{N}$. If new information arrives, the probability that dealer i is the first trader to observe the new information is denoted by $P(\lambda_i)$. We denote the corresponding probability that speculator j is the first trader to observe the new information by $P(\gamma_j)$. These probabilities are given by :

$$P(\lambda_i) \equiv Prob(f = i) \equiv \frac{\lambda_i}{\lambda_i + \sum_{m \neq i} \lambda_m + \sum_j \gamma_j} \quad \forall i \in \mathcal{M}, \tag{1}$$

$$P(\gamma_j) \equiv Prob(f = j) \equiv \frac{\gamma_j}{\gamma_j + \sum_{k \neq j} \gamma_k + \sum_i \lambda_i} \quad \forall j \in \mathcal{N}. \tag{2}$$

First, note that $P(\cdot)$ increases in the monitoring level chosen by a trader. This captures the intuition that the more closely a trader monitors the market, the larger the chance that he will be able to react to public announcements before the other traders. Second, note that the larger the aggregate monitoring level of the other traders is, the lower is $P(\cdot)$. This captures the intuition that the larger the monitoring levels of other traders are, the lower is the probability that a given trader will be the first to react to a public announcement. Finally note that $P(0) = 0$ and that $P(+\infty) = 1$. This means that a monitoring level equal to zero corresponds to the decision of *not monitoring the market at all*. Conversely, an infinite monitoring level corresponds to a decision of *continuously*

[6] Houtkin (1998) describes the trading strategies followed by SOES bandits. He provides a list of the events that can trigger trades by the SOES bandits. For example, business news, earnings announcements, price movements in related stocks, brokerage firms'upgrades and downgrades of stocks, announcements of economic indicators. Within our model these events would be considered public news announcements.

monitoring the market. For any intermediate monitoring level there is some monitoring but it is imperfect.

Each trader must exert effort to support their chosen level of monitoring. We denote the monetary disutility associated with a given monitoring level by $\Psi_d(\lambda)$ for a dealer and by $\Psi_s(\gamma)$ for a speculator. The monetary disutility is strictly increasing and *strictly convex* both for dealers and speculators. We assume that[7]:

$$\Psi(\lambda_i) = \frac{c_d}{4}\lambda_i^2 \quad \forall i \in \mathcal{M}, \tag{3}$$

and

$$\Psi(\gamma_j) = \frac{c_s}{4}\gamma_j^2 \quad \forall j \in \mathcal{N}. \tag{4}$$

We refer to $c_j > 0$ as the monitoring cost for trader j.

Speculators and dealers *simultaneously* choose their monitoring levels in the second stage of the trading round, after observing the quotes posted in the first stage.[8] We denote by $\lambda(S_b, M_b) = (\lambda_1(S_b, M_b), ..., \lambda_{M_b}(S_b, M_b))$, the vector of monitoring levels chosen by each dealer given the market spread and the number of market makers who post this spread. The dealers who are not posting the market spread do not participate in the trading stage and thus optimally choose not to monitor the market. In the same way, $\gamma(S_b, M_b)$ is the vector of monitoring levels chosen by each speculator.

Dealers and speculators can also monitor quote updates. In this way, traders can also acquire new information. An important feature of *quote monitoring* is that it requires no effort. Traders can invest in software that alerts them to quote updates in different securities.[9] Consequently, the probability of being the first trader to react to a quote update is more likely to be determined by the trading technology used than by the effort

[7] Quadratic monitoring cost functions allow us to derive monitoring levels and equilibrium spreads in closed form. Our qualitative results however only rely on the strict convexity of these functions.

[8] We have assumed that traders choose their monitoring level after observing the market spread for two reasons. First, a trader's monitoring (effort) level is unobservable and therefore quotes cannot be made contingent on monitoring levels. Second, traders can adjust their effort, once the quotes have been posted. For this reason it is natural to assume that the monitoring levels are chosen after the quoting stage, i.e., can be contingent on the market spread.

[9] SOES day trading firms have provided SOES day traders with software to monitor market-makers with stale quotes. Market makers reacted by automating their quote monitoring. See the General Accounting Office 1998 report on "The Effects of SOES on the Nasdaq Market".

exerted. Quote monitoring does not contribute to price discovery. Rather it enables traders to free ride on information production by the dealers (which is reflected in their quote updates). Dealers can then use this information to update their quote and speculators can use it to trade against dealers who are slow to adjust their quotes.

Given that no effort is required for quote monitoring we assume that when a dealer is first to update his quote, there is an exogenous probability Φ that one speculator reacts to this quote update before the $(M_b - 1)$ remaining dealers react. In this case, each speculator has an equal probability $(1/N)$ of being the speculator who first reacts. With probability $(1 - \Phi)$, the $(M_b - 1)$ remaining dealers update their quote before a speculator gets the chance to react to the initial quote update. Note that if $\Phi = 0$, only dealers benefit from quote monitoring and if $\Phi = 1$, only speculators benefit from quote monitoring. Thus Φ can be seen as a measure of speculators' relative advantage in quote monitoring. This parameter depends on the trading rules as well as the technology used. Nasdaq has attempted several times to replace SOES with alternative trading systems (e.g., N*Prove in 1994 and NAqcess in 1995) allowing dealers a delay to accept or to decline an incoming order.[10] It has been correctly pointed out by several commentators that this amounts to eliminating automatic execution. In this case, it becomes more difficult for speculators to "pick off" dealers through quote monitoring since a dealer can decline execution when he realizes that the order arrival is concomitant with his competitors' quote updates. Accordingly, we can analyze the effect of a suppressing automatic execution by comparing the case in which $\Phi = 0$ (automatic execution is not enforced) with the case in which $\Phi > 0$ (automatic execution is enforced). This comparison is performed in Section 4.[11]

2.3 Equilibrium

Assume that the market spread is strictly lower than the size of the revision in the asset's expected value in case of information arrival, i.e., $S_b < \sigma$ (this will always be the case in

[10] In 1991, dealers were allowed a 15-second delay to update quotes before being obliged to execute a second SOES order in the same security.

[11] Nasdaq's Autoquote Policy prohibits software that would automatically update one market maker's quotes as a function of other market makers' quotes. By forcing a dealer to update his quotes manually when he receives an alert, this policy increases his reaction time. In our setting, we can also analyze the effect of suppressing this policy by considering the impact of a decrease in Φ.

equilibrium). Given our previous assumptions, the optimal course of action for the dealers and the speculators in the trading stage is as follows. If a dealer is first to observe the new information, he revises his quotes accordingly. If his competitors react to this quote update before the speculators, they revise their quotes as well. If a speculator is first to react to a quote update by a dealer or to observe new information, she submits a market order of size Q (buy or sell depending on the direction of the quote revision).[12] Tables 1 and 2 list the payoffs to the dealers and the speculators, for different decisions and outcomes in the monitoring and quoting stages.

If the asset volatility, σ, is large then a quote revision can result in a "cross", i.e., a situation in which the best bid price is temporarily above the best ask price. This sometimes occurs in Nasdaq. It is worth stressing, however that our results hold for all values of $\sigma > 0$, i.e., they do not specifically rely on the possibility of crossed markets.

We solve for the perfect equilibrium of the trading round, i.e., a set $\{S_b^*, M_b^*, \lambda^*(.,.), \gamma^*(.,.)\}$ such that **(i)** $\lambda^*(S_b, M_b)$ and $\gamma^*(S_b, M_b)$ form a Nash equilibrium of the monitoring stage for all possible outcomes of the quoting stage and, **(ii)** $\{S_b^*, M_b^*\}$ is a Nash equilibrium of the quoting stage. Although we always consider M as exogenous, the number of dealers (M_b) posting the market spread is endogenous. Note that the case in which $M_b = 1$ and $\Phi > 0$ is subsumed in the case $M_b = 1$ and $\Phi = 0$. Actually if there is only one market maker, then, speculators cannot pick off the market maker posting the market spread by acting upon his quote update. Thus everything is *as if* $\Phi = 0$ and $M_b = 1$.

3 Monitoring Externalities

In this section, we show that market monitoring by one dealer can generate a positive *or* a negative externality for the other dealers, depending on Φ, the probability that a speculator reacts first after a quote update. These externalities play an important role in determining the quotes and we refer to them as monitoring externalities.

[12] Note that in our model quote revisions always imply changes in the asset value. In reality, quote revisions may occur for other reasons, for example, changes in inventories. Uncertainty about the cause of the quote revision will make it harder for speculators to draw inferences from quote revisions and to take advantage of information contained in revisions, which is equivalent to a decrease in Φ in the model.

Consider one dealer, say i, posting the market spread S_b and monitoring with intensity λ_i. There are two ways dealer i can be picked off when new information arrives. In the first case, a speculator reacts first to the information. This event occurs with probability $Prob(f \in \mathcal{N})$. Let $\lambda_A \equiv \sum_i \lambda_i$ and $\gamma_A \equiv \sum_j \gamma_j$ be the aggregate monitoring levels of the dealers and the speculators respectively. Using Equation (2), we obtain:

$$Prob(f \in \mathcal{N}) = \frac{\gamma_A}{\lambda_A + \gamma_A}. \tag{5}$$

In the second case, a dealer (different from dealer i) observes the arrival of information, updates his quote and a speculator is first to react to the quote update of this dealer. The probability of this event is $\Phi Prob(f \in \mathcal{M}_b \backslash i)$. Using Equation (1), we obtain:

$$Prob(f \in \mathcal{M}_b \backslash i) = \frac{\sum_{m \neq i} \lambda_m}{\lambda_A + \gamma_A}. \tag{6}$$

Let $\Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)$ be dealer i's expected profit for given levels of monitoring, λ_{-i} and γ, for the other dealers and the speculators respectively. Using the payoff table (Table 1), we obtain:

$$\begin{aligned}\Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b) &= -\alpha\left[Prob(f \in \mathcal{N}) + \frac{\Phi M_b}{M_b - 1} Prob(f \in \mathcal{M}_b \backslash i)\right] \frac{(\sigma - S_b)Q}{2M_b} \\ &\quad + (1-\alpha)\frac{\beta\delta Q S_b}{2M_b} - \Psi_d(\lambda_i) \quad \forall M_b \geq 2. \end{aligned} \tag{7}$$

The first term, which is negative, represents dealer i's expected loss from the risk of being picked off by speculators. The term in brackets is a measure of the total adverse selection risk faced by a dealer. The second term is positive and corresponds to dealer i's expected profit from trading with a liquidity trader. The last term is the monitoring cost incurred by dealer i. The probability of being picked off for dealer i is affected by the monitoring levels chosen by the other dealers. Consequently, dealer i's expected profit depends on the other dealers' monitoring decisions. Thus market monitoring by one dealer is an externality for the other dealers. The direction of this externality is obtained by differentiating one dealer's expected profit, say dealer i, with respect to another dealer's monitoring level:

$$\frac{\partial \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)}{\partial \lambda_m} = -\alpha \left[\frac{\partial Prob(f \in \mathcal{N})}{\partial \lambda_m} + \frac{\Phi M_b}{M_b - 1} \frac{\partial Prob(f \in \mathcal{M}_b \backslash i)}{\partial \lambda_m} \right] \frac{Q(\sigma - S_b)}{2M_b}$$

$$= \frac{\alpha}{(\lambda_A + \gamma_A)^2} \left[(1 - \frac{\Phi M_b}{M_b - 1})\gamma_A - (\frac{\Phi M_b}{M_b - 1})\lambda_i \right] \frac{Q(\sigma - S_b)}{2M_b} \quad \forall m \neq i. \tag{8}$$

Analyzing the sign of this partial first derivative, we obtain the following lemma.

Proposition 1 *: Consider two dealers i and m who are posting the market spread. There exists a constant $\bar{\Phi} \in (0, 1)$ such that:*

1. *If $\Phi \in [0, \bar{\Phi}]$ then market monitoring by dealer m is a positive externality for dealer i, i.e., $\frac{\partial \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)}{\partial \lambda_m} \geq 0$.*

2. *If $\Phi \in [\bar{\Phi}, 1]$ then market monitoring by dealer m is a negative externality for dealer i, i.e., $\frac{\partial \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)}{\partial \lambda_m} \leq 0$.*

where $\bar{\Phi} = \frac{\gamma_A(M_b - 1)}{(\gamma_A + \lambda_i)M_b}$.

The economic intuition for this important property of market monitoring is as follows. An increase in market monitoring by dealer m increases the probability that this dealer will be first to observe new information. This indirectly benefits dealer i since a quote update by dealer m *signals* to dealer i that his own quotes are misaligned. Thus, the increase in market monitoring by dealer m reduces dealer i's probability of being picked off through speculators' market monitoring ($\frac{\partial Prob(f \in \mathcal{N})}{\partial \lambda_m} < 0$). This is the source of the positive externality. However, there is a second effect since speculators monitor quote updates to learn about stale quotes. If $\Phi > 0$, an increase in market monitoring by dealer m results in a greater probability of being picked off through speculators' quote monitoring for dealer i since $\frac{\partial Prob(f \in \mathcal{M}_b \backslash i)}{\partial \lambda_m} > 0$. This is the source of the negative externality. For sufficiently low values of Φ, the reduction in the risk of being picked off through market monitoring is larger than the increase in the risk of being picked off through quote monitoring. When Φ is large, the reverse is true. Accordingly, an increase in market monitoring by one dealer is a double-edged sword: it can amplify or reduce the adverse selection risk faced by his competitors.

4 Bidding Strategies, Spreads and Monitoring Externalities

In this section, we study the implications for price formation of the monitoring externalities. The equilibrium spread depends on the monitoring decisions of the traders, which themselves are affected by the spread. Thus we proceed by backward induction and we analyze first the equilibrium monitoring decisions for a given market spread.

4.1 Equilibrium in the monitoring stage

Given the monitoring levels chosen by the other traders, dealer i chooses the monitoring level which maximizes $\Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)$. Using the expression for dealer i's expected profit, the first order condition is:

$$-\alpha \left[\frac{\partial Prob(f \in \mathcal{N})}{\partial \lambda_i} + (\frac{\Phi M_b}{M_b - 1}) \frac{\partial Prob(f \in \mathcal{M}_b \backslash i)}{\partial \lambda_i} \right] \frac{(\sigma - S_b)Q}{2M_b} = \Psi_d^{'}(\lambda_i).$$

More monitoring reduces the probability of being picked off but requires additional effort. The terms inside the brackets measure the marginal effect of increased monitoring by dealer i on the probability of being picked off. The first order condition sets the marginal benefit of monitoring equal to the marginal cost for dealer i. Using Equations (5) and (6), we can rewrite the first order condition as:

$$\frac{\alpha Q(\sigma - S_b)}{2M_b(\lambda_A + \gamma_A)^2} \left[\gamma_A + (\frac{\Phi M_b}{M_b - 1}) \sum_{m \neq i} \lambda_m \right] = \Psi_d^{'}(\lambda_i). \tag{9}$$

The second order condition is satisfied if $S_b < \sigma$, which will be the case in equilibrium. Let $\Pi_s(\gamma_j, \lambda, \gamma_{-j})$ be the expected profit for speculator j, for given monitoring levels of the other traders. Using Table 2, we obtain:

$$\Pi_s(\gamma_j, \lambda, \gamma_{-j}) = \frac{\alpha Q(\sigma - S_b)}{2} \left[Prob(f = j) + \frac{\Phi Prob(f \in \mathcal{M}_b)}{N} \right] - \Psi_s(\gamma_j). \tag{10}$$

In the case of a change in the asset value, a profit opportunity arises because the dealers' quotes are temporarily mispriced. Recall that a speculator can capture this profit opportunity in two different ways: either (i) she is the first to react to the public announcement of a change in the asset value or (ii) she is the first to react to the quote update of a dealer. The term in bracket is the sum of the probabilities of these two events. Using Equation (1), we obtain:

$$Prob(f \in \mathcal{M}_b) = \frac{\lambda_A}{\lambda_A + \gamma_A}.$$

Speculator j chooses the monitoring level that maximizes $\Pi_s(\gamma_j, \lambda, \gamma_{-j})$. This implies setting the marginal benefit of monitoring equal to the marginal cost for speculator j:

$$\frac{\alpha Q(\sigma - S_b)}{2(\lambda_A + \gamma_A)^2}\left[\lambda_A(\frac{N - \Phi}{N}) + \sum_{s \neq j} \gamma_s\right] = \Psi_s^{'}(\gamma_j). \tag{11}$$

The second order condition is satisfied if $S_b < \sigma$ in this case as well. It follows from the previous analysis that a Nash equilibrium of the monitoring stage is a pair of vectors $(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b))$ that solves Equations (9) and (11), for all the dealers posting the market spread and all the speculators. We say that the Nash equilibrium of the monitoring stage is *symmetric* if all the traders of a given type (e.g., all the dealers) choose the same monitoring level.

Lemma 1 *: If there exists a Nash equilibrium in the monitoring stage, it is symmetric.*

Let λ^* be the monitoring level chosen by each dealer and let γ^* be the monitoring level chosen by each speculator, in equilibrium. Since the equilibrium must be symmetric, the system of equations characterizing the traders' best responses is obtained by rewriting Equations (9) and (11) as:

$$\frac{\alpha Q(\sigma - S_b)}{M_b(M_b\lambda^* + N\gamma^*)^2}\left[N\gamma^* + \Phi M_b\lambda^*\right] = c_d\lambda^*, \tag{12}$$

and

$$\frac{\alpha Q(\sigma - S_b)}{(M_b\lambda^* + N\gamma^*)^2}\left[M_b(\frac{N - \Phi}{N})\lambda^* + (N-1)\gamma^*\right] = c_s\gamma^*. \tag{13}$$

Solving this system equations yields the equilibrium monitoring levels and expected profit functions. The solution is given in the next proposition.

Proposition 2 : *When M_b dealers post a market spread $S_b < \sigma$, the equilibrium of the monitoring stage is unique and is characterized by the following monitoring levels for the speculators and the dealers (with $\Phi = 0$ if $M_b = 1$):*

$$\gamma^*(S_b, M_b) = \sqrt{\frac{\alpha Q(\sigma - S_b)}{c_s}} \frac{\sqrt{(\Upsilon(\frac{N-\Phi}{N}) + (N-1)}}{\Upsilon + N}, \tag{14}$$

$$\lambda^*(S_b, M_b) = \frac{\Upsilon}{M_b}\gamma^*, \tag{15}$$

where $\Upsilon = \frac{N}{2r(N-\Phi)}\left(\sqrt{((r(N-1)-\Phi)^2 + 4r(N-\Phi)} - ((N-1)r - \Phi)\right)$ *and* $r \equiv \frac{c_d}{c_s}$.

For these monitoring levels, the expected profits of the speculators and the dealers when the market spread is S_b are:

$$\Pi_d(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b), M_b) = \frac{Q}{2M_b}[-\alpha(\sigma - S_b)R(M_b, \Phi, r) + (1-\alpha)\beta\delta S_b], \tag{16}$$

$$\text{with} \quad R(M_b, \Phi, r) \equiv \underbrace{\frac{N}{\Upsilon + N} + \frac{\Phi\Upsilon}{\Upsilon + N}}_{Adverse \quad Selection \quad Risk} + \underbrace{(\frac{r\Upsilon^2}{2})\frac{(\Upsilon(\frac{(N-\Phi)}{N}) + N - 1)}{M_b(\Upsilon + N)^2}}_{Monitoring \quad Cost}, \tag{17}$$

and

$$\Pi_s(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b), M_b) = \frac{Q}{2N}[\alpha(\sigma - S_b)R'(N, \Phi, r)], \tag{18}$$

$$\text{with} \quad R'(N, \Phi, r) \equiv \underbrace{\frac{N}{\Upsilon + N} + \frac{\Phi\Upsilon}{\Upsilon + N}}_{Probability\ of\ Hitting\ Stale\ Quote} - \underbrace{\frac{1}{2}\frac{(\Upsilon(N-\Phi) + N^2 - N)}{(\Upsilon + N)^2}}_{Monitoring \quad Cost}. \tag{19}$$

The proposition reveals several interesting properties of the monitoring strategies followed by the traders:

- Speculators and dealers always put some effort in market monitoring ($\gamma^* > 0$ and $\lambda^* > 0$). In particular, it is never optimal for speculators to entirely base their trading strategies on dealers' quote updates.[13]

- The monitoring level of both types of traders decreases with the size of the spread. When the dealers increase their spread, speculators monitor the market less intensively since the profit obtained by picking off dealers is lower. The dealers react by monitoring the market less intensively.

- There is a positive relationship between the monitoring level of the speculators and the monitoring level of the dealers, for a given number of speculators (see Equation (15). This means that securities that are closely monitored by the speculators are also closely monitored by the dealers.

- For a given spread, the aggregate monitoring level of the dealers ($\lambda_A^* = M_b \lambda^*$) is independent of the number of dealers. The reason for this is that the loss conditional on being picked off is shared with the other dealers, thus, if the number of dealers increases each dealer responds by lowering his monitoring effort.

The dealer's expected profit, given by Equation (16), is central to the analysis of the equilibrium spread in the next subsection. Note that the constant R determines the size of the expected loss per share, for a dealer, *i.e., the (per unit) cost of market making.* This expected loss reflects the adverse selection risk borne by the dealer and the monitoring cost that a dealer incurs in order to limit the risk of being picked off. The latter is reflected in the last term of Equation (17). The adverse selection risk is captured by the sum of the first two terms in Equation (17), which together make up the probability of the dealer being picked off in equilibrium.

Equation (18) determines the expected profit for a speculator. The speculator's profit depends on the probability of hitting a stale quote, that is, the adverse selection risk for the dealers, the profit conditional on this event, and the monitoring cost. This equation will be used when we analyze the entry decision of the speculators in Section 6.

[13]This result is consistent with Harris and Schultz (1998) who, empirically, do not find a strong support for the view that bandits only trade after quote updates.

The above results are all conditional on the spread. In the next section we proceed by solving the quoting stage of the game for the equilibrium spread.

4.2 Equilibrium in the quoting stage

The dealer's bidding behavior in the quoting stage crucially depends on whether market monitoring is a positive or a negative externality. In order to pinpoint the causes for this result, we start by considering the case in which market monitoring is unambiguously a positive externality, i.e., the case in which speculators cannot pick off dealers through quote monitoring ($\Phi = 0$). Then we consider the more general case in which market monitoring can be a negative externality ($\Phi > 0$).

4.2.1 The Market Spread in the Absence of Quote Monitoring by the Speculators ($\Phi = 0$)

We first consider the possible equilibria in which *all* the dealers post the market spread ($M_b^* = M$). A market spread equal to S_b^* is a Nash equilibrium of the quoting stage if (i) no dealer has an incentive to widen his spread given that the other dealers post a spread equal to S_b, and (ii) no dealer has an incentive to undercut their competitors' quotes. The first condition ensures that dealers do not expect to incur losses:

$$\Pi_d(\lambda^*(S_b^*, M), \gamma^*(S_b^*, M), M) \geq 0.$$

Let $\hat{S}(M, \Phi, r)$ be the spread such that this equation is binding. Using Equation (16), we get:

$$\hat{S}(M, \Phi, r) = \alpha\sigma(\frac{R(M, \Phi, r)}{\alpha R(M, \Phi, r) + (1-\alpha)\beta\delta}). \tag{20}$$

In equilibrium, the market spread must be at least equal to $\hat{S}$ for the dealers to break even. Now suppose that a dealer improves slightly upon the equilibrium market spread. In this case, only this dealer is exposed to the risk of being picked off and thereby monitors ($M_b = 1$). It follows that the expected profit for the dealer who undercuts is

$\Pi_d(\lambda^*(S_b^*, 1), \gamma^*(S_b^*, 1), 1)$. In equilibrium, the dealer must be better off not undercutting. This requires that the profits earned by matching the quotes of other dealers are at least as high as the profits from undercutting their quotes:

$$G(S_b^*) = \Pi_d(\lambda^*(S_b^*, M), \gamma^*(S_b^*, M), M) - \Pi_d(\lambda(S_b^*, 1), \gamma^*(S_b^*, 1), 1) \geq 0. \tag{21}$$

Computations yield:

$$G(S_b^*) = -\frac{(M-1)Q}{2M}[S_b^*(\alpha\bar{R}(r) + (1-\alpha)\beta\delta) - \alpha\sigma\bar{R}(r)],$$

where

$$\bar{R}(r) = \frac{N}{\Upsilon + N} + \frac{(1+M)}{2M}\frac{\Upsilon N}{(\Upsilon + N)^2}.$$

Let $\bar{S}(r)$ be the spread such that $G(\bar{S}(r)) = 0$. We obtain:

$$\bar{S}(r) = \alpha\sigma(\frac{\bar{R}(r)}{\alpha\bar{R}(r) + (1-\alpha)\beta\delta}).$$

As $G(.)$ decreases with S_b^*, Equation (21) is satisfied if and only if $S_b^* \leq \bar{S}(r)$. It is straightforward to show that $\hat{S} < \bar{S}$ since $R(M, 0, r) < \bar{R}(r)$, for all the values of the parameters. Based on the arguments above we derive the following result.

Proposition 3 : *In the absence of quote monitoring by the speculators ($\Phi = 0$), the situation in which all the dealers post a market spread equal to S_b is a Nash equilibrium of the quoting stage if and only if $S_b \in [\hat{S}(M, \Phi, r), \bar{S}(r)]$. Furthermore, for all the Nash equilibria in which the market spread is strictly larger than $\hat{S}(M, 0, r)$, the dealers obtain strictly positive expected profits.*

In the quoting stage dealers compete in prices, *à la Bertrand*. One would therefore expect the equilibrium to feature zero expected profits for the dealers. But this is not necessarily the case. The intuition for this result is as follows. When a dealer undercuts his competitors, two effects are at work. On the one hand, the dealer captures a larger part of the order flow. On the other hand, the dealer monitors the market more intensively. This follows because his competitors lose price priority but, for this reason, face no risk of being picked off. As a consequence, they do not monitor the market at all. It follows that the dealer who undercuts loses the benefit of the positive externality associated with market monitoring by the other dealers. He therefore must devote more effort to market monitoring. These two effects have opposite impacts on the expected profit of a dealer. The first effect (larger share of the order flow) increases a dealer's profit, *gross* of the monitoring cost, but the second effect increases his market monitoring cost. It turns out that for all spreads below $\bar{S}(r)$, the second effect is larger than the first. Dealers are better off not undercutting their competitors in order to benefit indirectly from their monitoring. Consequently, the positive externality associated with market monitoring by the dealers helps to sustain spreads that are larger than the competitive level.

The possibility for dealers to free-ride on market monitoring by other dealers has another implication: it makes quote matching strategies profitable. It follows that there is no equilibrium in which some dealers prefer to bid themselves out of the trading round. To formally establish these points, let $\Pi_J(M_b, S_b) \equiv M_b\Pi_d(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b), M_b)$ be the dealers' joint expected profit when $M_b < M$ dealers quote a market spread S_b.

Proposition 4 : *In the absence of quote monitoring by speculators,*

- *If* $\Pi_J(M_b, S_b) \geq 0$ *then* $\Pi_J(M_b+1, S_b) > 0$. *It follows that quote matching is always a profitable strategy since each dealer obtains an equal share of the joint expected profit.*
- *This implies that there is no equilibrium in which a dealer chooses not to post the market spread or* $M_b^* = M$ *is the unique possibility in equilibrium.*

The intuition is as follows. Consider a dealer, say i, who does not post the market spread when $M_b < M$ dealers already post this spread. If dealer i does not match or improve upon the market spread, he obtains zero profits. Suppose that instead dealer i matches this spread. Now $(M_b + 1)$ dealers monitor the market. Each dealer monitors the market less but the aggregate monitoring level of the dealers is unchanged in equilibrium (λ_A^* does not depend on M_b). Consequently, quote matching by dealer i leaves unchanged the probability of being picked off for each dealer and thereby leaves unchanged the *gross* joint expected profit for the dealers. However each dealer can free-ride on a larger number of dealers and the total monitoring cost is now “shared” among $(M_b + 1)$ dealers. As $\Psi_d(.)$ is strictly convex, this cost sharing results in a lower total monitoring cost $((M_b+1)\Psi_d(\frac{\lambda_A}{(M_b+1)}) < M_b\Psi_d(\frac{\lambda_A}{M_b}))$. Eventually dealers' joint expected profit, net of the monitoring cost, is larger.

Note that the set of possible equilibrium spreads is determined by the ratio of dealers' and speculators' monitoring costs (r) and not by the absolute costs. Thus, our results hold even if market monitoring costs are small.

Corollary 1 *:* **(spread and monitoring cost)** *The lower bound and the upper bound of the set of possible equilibrium spreads have the following properties:*

- *They go to zero when dealers' monitoring cost becomes negligible relative to speculators' monitoring cost (*$\lim_{r\to 0}\hat{S} = \lim_{r\to 0}\bar{S} = 0$*),*
- *They increase when dealers' monitoring cost increases relative to speculators' monitoring cost (*$\frac{\partial \hat{S}}{\partial r} > 0$ *and* $\frac{\partial \bar{S}}{\partial r} > 0$*),*
- *They go to* $\frac{\alpha\sigma}{\alpha+(1-\alpha)\beta\delta}$ *when speculators' monitoring cost becomes negligible relative to dealers' monitoring cost (*$\lim_{r\to +\infty}\hat{S} = \lim_{r\to +\infty}\bar{S} = \frac{\alpha\sigma}{\alpha+(1-\alpha)\beta\delta}$*).*

The first part of the corollary shows that imperfect market monitoring is the source of trading cost. In fact, when dealers' monitoring cost becomes *relatively* small, their monitoring increases (and the speculators' monitoring decreases). As a consequence, the risk of being picked off vanishes and the spread becomes zero. The second part of the corollary shows that the spread widens when the monitoring cost for the dealers increases relative to the monitoring cost for the speculators. The intuition is as follows. Other things equal, when dealers' monitoring costs increase, λ^* decreases, i.e., dealers put less effort in market monitoring. It follows that the probability of being picked off increases. Consequently, dealers must increase their spreads in order to break-even. This means that differences in monitoring ability or monitoring costs between dealers and speculators affect the size of the spread.[14] However, these differences are not necessary for dealers to be

[14]There is another way to interpret a change in r in the model. Suppose that the marginal cost of effort is identical to speculators and dealers ($c_d = c_s = c$) but that dealers recover only a fraction $\tau < 1$ of their

exposed to the risk of being picked off in case of public information arrival. Thus, from now on, we assume that $r = 1$. This assumption does not affect the qualitative results derived below but simplifies the computations. Furthermore, it ensures that the results do not come from *ad hoc* assumptions about differences in monitoring skills between the dealers and the speculators.

4.2.2 The Market Spread in the Presence of Quote Monitoring by Speculators ($\Phi > 0$)

We will now solve the quoting stage of the game for the case of $\Phi > 0$. First, we note that when $r = 1$, $\Upsilon = \frac{N}{N-\phi}$. In this case, the per unit cost of market making, given by Equation (17), simplifies as:

$$R(M_b, \Phi, 1) = \underbrace{\frac{N}{N+1-\Phi}}_{Adverse\ Selection\ Risk} + \underbrace{\frac{N}{2M_b(N+1-\Phi)^2}}_{Monitoring\ Cost}, \quad \forall M_b \geq 2. \tag{22}$$

Recall that having $M_b = 1$ and $\Phi > 0$ is identical to $M_b = 1$ and $\Phi = 0$. It follows that:

$$R(1, \Phi, 1) = R(1, 0, 1) = \frac{N}{N+1} + \frac{N}{2(N+1)^2}. \tag{23}$$

From now on, we will write $R(M_b, \Phi, 1)$ simply as $R(M_b, \Phi)$. We can proceed exactly as in the previous section to show that for $\{S_b^*, M\}$ to be a Nash equilibrium of the quoting stage, it is a necessary condition that $S_b^* \in [\hat{S}(M, \Phi), \bar{S}(\Phi)]$ where $\hat{S}(M, \Phi)$ is defined as in Equation (20) and :

$$\bar{S}(\Phi) = \alpha\sigma(\frac{\bar{R}(\Phi)}{\alpha\bar{R}(\Phi) + (1-\alpha)\beta\delta}),$$

with:

$$\bar{R}(\Phi) = \frac{\frac{MN}{N+1} + \frac{MN}{2(N+1)^2} - \frac{N}{N+1-\Phi} - \frac{N}{2M(N+1-\Phi)^2}}{M-1}. \tag{24}$$

trading profits (but bear in full the disutility due to market monitoring). This assumption is consistent with the fact that dealers act as *agents* of market-making firms. In this case, we can derive exactly the same results as those derived until this point, substituting c_d by $\frac{c}{\tau}$. Thus a decrease in τ has the same effect as an increase in r.

Lemma 2 *:*

1. *$\hat{S}(M, \Phi)$ increases with Φ whereas $\bar{S}(\Phi)$ decreases with Φ.*

2. *There exists $\hat{\Phi}(M, N) \in (0, 1)$ such that (i) if $\Phi \in [0, \hat{\Phi}]$ then $\hat{S}(M, \Phi) \leq \bar{S}(\Phi)$ and (ii) if $\Phi \in (\hat{\Phi}, 1]$ then $\hat{S}(M, \Phi) > \bar{S}(\Phi)$.*

($\hat{\Phi}$ is characterized in the proof of this lemma).

An increase in Φ has two effects. On the one hand, it increases the adverse selection risk. This explains why the zero expected profit spread, $\hat{S}$, increases with Φ. On the other hand, it makes the negative externality stronger (see Proposition 1). The negative externality has an effect *opposite* to the effect of the positive externality: it encourages dealer i to improve upon the quotes of his competitors. If dealer i turns out to be the sole dealer posting the market spread in the quoting stage, his competitors will not monitor the market in the subsequent stages. Consequently if he undercuts, dealer i can eliminate the risk of being picked off because of the signal sent by his competitors' quote updates. Thus, when $\Phi > 0$, undercutting acts as a defensive strategy against the risk of being picked off through quote monitoring. For this reason it is more difficult to sustain a non-competitive spread as an equilibrium when Φ increases. This explains why the largest possible equilibrium spread, $\bar{S}$, decreases with Φ.

For an equilibrium with all the dealers ($M_b^* = M$) quoting the inside spread, it is necessary that: $\hat{S}(M, \Phi) \leq \bar{S}(\Phi)$. The lemma above ($2^{nd}$ part) states that this occurs if and only if Φ is sufficiently small. When $\Phi > \hat{\Phi}$, the negative externality is so strong that one dealer is always better off undercutting the zero expected profit spread with M dealers (i.e., $\bar{S}(M, \Phi) < \hat{S}(\Phi)$). In this case there is no equilibrium with all the dealers posting the market spread.

Proposition 5 *:*

1. *When $0 \leq \Phi \leq \hat{\Phi}$, the situation in which all the dealers post a market spread equal to S_b is a Nash equilibrium of the quoting stage if and only if $S_b \in [\hat{S}(M, \Phi), \bar{S}(\Phi)]$. The*

expected profits of the dealers posting the market spread are strictly positive when the equilibrium market spread is strictly larger than $\hat{S}(M, \Phi)$.

2. *When* $\hat{\Phi} < \Phi \leq 1$, *the unique Nash equilibrium of the quoting stage is such that only one dealer* ($M_b^* = 1$) *posts the market spread which is* $S_b^* = \hat{S}(1, 0)$. *The expected profit of the dealer posting the market spread is zero.*

Figure 2 represents the set of possible equilibrium spreads as a function of Φ. The explanations for the first part of the proposition are the same as the explanations we gave for Proposition 3. The claim in second part of Proposition 5 stands in striking contrast with our result regarding the profitability of quote matching strategies in the previous section. In order to avoid the negative externality associated with quote monitoring, dealers undercut each other until the point where a single dealer has no incentive to undercut the market spread. For this reason, the dealer posting this spread just breaks even. The equilibrium spread in this case is also such that if another dealer were to match the spread, then the two dealers posting the best quotes would incur losses. This is due to the negative externality they inflict on each other. Quote matching exacerbates the negative externality and thus becomes an unprofitable strategy when speculators react relatively quickly to quote updates.

When $\Phi < \hat{\Phi}$, there is a multiplicity of possible equilibrium spreads.[15] In order to sharpen our prediction regarding the outcome of the quoting stage, we use the concept of *Pareto-Dominance*. Pareto-Dominant equilibria are the natural outcomes when players can communicate before the game takes place (See Fudenberg and Tirole (1991)). A Nash equilibrium of the quoting stage is Pareto-Dominant if there is no other equilibrium outcome that improves or leaves unchanged each dealers' expected profits.

Proposition 6 : *When there is a multiplicity of equilibria in the quoting stage, the unique Pareto-Dominant equilibrium is such that dealers post the largest possible equilibrium spread,* $S_b^* = \bar{S}(\Phi)$.

[15] In addition to the equilibria described in the first part of Proposition 5, there is another equilibrium with $M_b^* = 1$ when $\Phi \leq \hat{\Phi}$ but Φ sufficiently greater than zero. In this equilibrium, the equilibrium spread is $S_b^* = \hat{S}(1, 0)$, which belongs to the set of possible equilibrium spreads described in the first part of Proposition 5.

The result is immediate since each dealers' expected profit increases with the market spread (see Equation (16)). We will refer to the equilibrium in which dealer posts a spread equal to $\bar{S}(\Phi)$ as the Pareto-Dominant equilibrium and to the equilibrium in which dealers post a spread equal to $\hat{S}(M, \Phi)$ as the zero expected profit equilibrium.[16]

To sum up, in this section, we have shown how externalities associated with market monitoring influence the price formation process. The possibility for dealers to free-ride on market monitoring by other dealers (i) reduces the incentive to undercut and (ii) makes quote matching strategies profitable. Free-riding is more dangerous when speculators can pick off dealers based on quote updates, however. In fact, if speculators react sufficiently quickly to quote updates, a dealer can be hurt by other dealers' quote updates. This negative externality (i) encourages dealers to undercut and (ii) can make quote matching strategies unprofitable. In the next section, we derive the implications of these results for market design.

5 Market Quality and Automatic Execution

Should dealers be protected against the automatic execution of stale quotes or not? This question has been central to the controversy between Nasdaq dealers and SOES bandits. Nasdaq dealers have argued that automatic execution made it easier for bandits to pick off dealers who were slow to adjust their quotes. Accordingly dealers were obliged to widen their spreads. In response, SOES bandits have argued that their presence has strengthened price competition among dealers. They also argued that they contributed to price discovery by forcing dealers to monitor more closely the stocks in which they were making the market. As explained previously (Section 2.2), we can formally examine these arguments regarding the effects of automatic execution by comparing market spreads and monitoring levels in the cases $\Phi = 0$, *automatic execution is not enforced* and $\Phi > 0$, *automatic execution is*

[16]Kandel and Marx (1997) show that multiple equilibrium spreads can be obtained when a financial market features a positive tick size. Interestingly we obtain a multiplicity of equilibrium spreads even if the tick size is zero. A positive tick size would clearly not change this result but could help dealers to coordinate on the Pareto-Dominant equilibrium.

enforced.[17]

Corollary 2 *:*

1. *When the equilibrium of the quoting stage is the Pareto-Dominant equilibrium, the market spread is smaller when automatic execution is enforced.*

2. *When the equilibrium of the quoting stage is the zero expected profit equilibrium, the market spread is larger when automatic execution is enforced.*

Consider Figure 2. If the dealers post the zero expected profit spread then the equilibrium spread is clearly larger when $\Phi > 0$ than when $\Phi = 0$. This reflects the fact that the adverse selection risk for the dealers is larger when speculators can use the information revealed by quote updates to pick off dealers. This supports the standard argument that the presence of SOES bandits and automatic execution increase the spread. On the other hand, if dealers post the Pareto-Dominant equilibrium spread, the conclusion is reversed: the equilibrium spread is smaller when $\Phi > 0$ than when $\Phi = 0$. Recall that the dealers' incentive to undercut are stronger when speculators can hit dealers who are slow to adjust their quotes than when they cannot do so. As a result, non-competitive spreads are more difficult to sustain when dealers can not decline or delay execution of trades. This observation supports the SOES bandits' claim that they have increased price competition among dealers. This discussion points to one interesting effect of automatic execution: it makes free-riding on other dealers' market monitoring less profitable for a dealer. In this way automatic execution is conducive to price competition and can indeed result in a lower spread.

We examine now the effect of automatic execution on the equilibrium monitoring level chosen by a dealer. From Proposition 2, we obtain:

$$\begin{cases} \lambda^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - S_b^*)}{cM^2(1+N-\Phi)^2}} & for \quad \Phi \leq \hat{\Phi} \\ \lambda^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - \hat{S}(1,0))}{c(1+N)^2}} & for \quad \Phi > \hat{\Phi} \end{cases}$$

[17] Stoll (1992) discusses the impact of automatic execution on the values of the "free trading options" in limit order markets and in dealer markets. He points out that automatic execution can be detrimental to market quality because it increases the risk of being picked off for traders with stale quotes.

Corollary 3 : *The monitoring level chosen by a dealer in equilibrium is always larger when automatic execution is enforced, both in the zero expected profit and in the Pareto-Dominant equilibria.*

The direct effect of automatic execution is to strengthen dealers' incentive to be first to discover new information. Actually free riding on monitoring by other dealers is more dangerous since speculators can use the information contained in a quote update at the dealers' expense. This effect is present whatever the nature of the equilibrium in the quoting stage. Automatic execution has also has an indirect effect on market monitoring because it affects the equilibrium spread. The direction of the indirect effect depends on the equilibrium in the quoting stage. In the Pareto-Dominant equilibrium, enforcing automatic execution reduces the spread and in this way further enlarges dealers' market monitoring. In contrast, in the zero expected profit equilibrium, enforcing automatic execution widens the spread and in this way reduces dealers' need to monitor. Still this is insufficient for their equilibrium monitoring level to be smaller than when automatic execution is not enforced.

The previous result shows that automatic execution forces dealers to monitor the market more closely. However, for price discovery, it is the *total* effort, $\lambda_A + \gamma_A$, exerted by all the traders in monitoring the market that matters.[18] Thus we compare now the aggregate monitoring level when $\Phi = 0$ and when $\Phi > 0$.

Corollary 4 :

1. *When the equilibrium of the quoting stage is the Pareto-dominant equilibrium, the aggregate monitoring level, $\lambda_A^* + \gamma_A^*$, of all the traders is larger when automatic execution is enforced.*

2. *When the equilibrium of the quoting stage is the zero expected profit equilibrium, the aggregate monitoring level, $\lambda_A^* + \gamma_A^*$, of all the traders is smaller when automatic*

[18] In the model, the probability that one trader will discover whether an informational event has taken place is always equal to one. However, it is easy to modify the model in such a way that this probability is less than one, by adding a constant p in the denominators of $P(\lambda_i)$ and $P(\gamma_j)$. The probability that the informational event will *not* be discovered is then $\frac{p}{\lambda_A+\gamma_A+p}$. It decreases with $(\lambda_A + \gamma_A)$. Thus the speed of price discovery increases with $(\lambda_A + \gamma_A)$.

execution is enforced.

Automatic execution may or may not improve price discovery. On the one hand, it strengthens dealers' incentives to monitor the market. On the other hand, it alters the monitoring strategy followed by the speculators. They have less incentive to directly monitor the information flow since they can use the costless information contained in quote updates to pick off dealers. In the zero expected profit equilibrium, this effect is reinforced by the fact that the spread is larger when automatic execution is enforced (γ^* decreases with the spread). It follows that in this case the aggregate monitoring is lowest when automatic execution is enforced. On the contrary, in the Pareto Dominant equilibrium, the spread is smaller with automatic execution. In this case the increase in dealers' aggregate monitoring is larger than the reduction in speculators' monitoring level and price discovery is improved.

6 Empirical Implications

Much of the debate about the possible effects of SOES on Nasdaq trading has focused on two related empirical questions. The first question is whether or not trading by SOES bandits have caused market makers to post wider spreads. A second question is what determines the level of SOES bandit activity. Researchers have noted that the spread and the level of SOES bandit activity are likely to be interdependent. Thus, in order to formally address these key questions in the SOES debate using our model it is necessary to solve for the level of speculators activity. This is one goal of this section. A second goal is to derive comparative statics for the model. All through the analysis, we treat the number of dealers, M, as exogenous. The number of dealers on Nasdaq is likely to be determined by all the trades they execute, not just those going through SOES. In contrast, the number of SOES bandits is certainly determined by the profitability of the trades they execute on SOES.

We first establish that the market spread depends on the number of speculators.

Proposition 7 : *An increase in the number of speculators results in an increase in the equilibrium spread, both in the zero expected profit and in the Pareto Dominant equilibria.*

When the number of speculators increases, the risk of being picked off is larger for the dealers. Accordingly the expected loss to speculators is greater. It follows that dealers must widen their spreads in order to break-even.

We assume that a trader decides to become a speculator, or not, before the quoting stage of the trading game. A speculator bears a fixed cost, $K > 0$, that is sunk at the entry stage. The number of speculators is then determined in such a way that the expected trading profit of a speculator is just equal to the fixed cost borne by the speculator.[19] For brevity, from now on, we assume that $\Phi < \hat{\Phi}(M, 1)$. As $\hat{\Phi}(M, N)$ increases with N, this assumption guarantees that, in equilibrium, all the dealers post the market spread (See Proposition 5). This assumption simplifies the presentation of the results in this section but does not affect them qualitatively.

Using Proposition 2, the speculator's expected profit when $\Upsilon = \frac{N}{N-\Phi}$ is given by:

$$\Pi_s(\lambda^*(S_b^*), \gamma^*(S_b^*), S_b^*, N) = \alpha Q(\sigma - S_b^*)\left[\frac{2N(N+1-\Phi) - (N-\Phi)^2)}{4N(N+1-\Phi)^2}\right]. \qquad (25)$$

The expected profit decreases in the equilibrium market spread. The latter increases with the number of speculators. It is therefore straightforward to show that a speculator's expected profit decreases with the number of speculators. Assume that $K \leq \Pi_s(\lambda^*(S_b^*), \gamma^*(S_b^*), S_b^*, 1)$ (otherwise no speculator would find it profitable to enter the market). The equilibrium number of speculators, N^*, solves:

$$\Pi_s(\lambda^*(S_b^*), \gamma^*(S_b^*), S_b^*, N^*) = K. \qquad (26)$$

Using the fact that $S_b^* < \sigma$, we obtain:

$$\lim_{N \to +\infty} \Pi_s(\lambda^*(S_b^*), \gamma^*(S_b^*), S_b^*, N) = 0 < K.$$

[19]There may not be an integer solution to the equality. In order to avoid this technical problem, we treat N as a real number, as it is usual in market entry analysis.

Thus, as Π_s decreases with N, the equilibrium number of speculators is uniquely defined whereby $1 \leq N^* < +\infty$.

We illustrate the effect of a change in the exogenous parameters on the market spread and the number of speculators assuming that dealers post the zero expected profit spread in equilibrium. The case of the Pareto-Dominant equilibrium is discussed at the end of the section. Recall that the zero expected profit spread is:

$$\hat{S}(M, \Phi, N) = \alpha\sigma(\frac{R(M, \Phi)}{\alpha R(M, \Phi) + (1-\alpha)\beta\delta}). \tag{27}$$

Note that, *for a given number of speculators*, an increase in σ, the size of the innovations in the asset value, widens the market spread whereas an increase in β or δ, the probability of or expected size of liquidity trades, reduces the market spread. By replacing S_b^* by $\hat{S}(M, \Phi, N^*)$ in Equation (25), we obtain that:

$$(\frac{\alpha(1-\alpha)\beta\delta Q\sigma}{\alpha R(M, \Phi) + (1-\alpha)\beta\delta})\left[\frac{2N^*(N^*+1-\Phi) - (N^*-\Phi)^2)}{4N^*(N^*+1-\Phi)^2}\right] = K. \tag{28}$$

This equation defines implicitly the equilibrium number of speculators. We use this equilibrium condition to derive the next proposition.

Proposition 8 : *When the dealers post the zero expected profit market spread, the following comparative static results are obtained in equilibrium:*

- *When the average order size (δ) submitted by liquidity traders increases, the number of speculators increases and the spread decreases.*
- *When the probability of a liquidity trade (β) increases, the number of speculators increases and the spread decreases.*
- *When the size of the innovation in the asset value (σ) increases, the number of speculators increases and the spread increases.*
- *When the number of dealers (M) increases, the number of speculators increases and the spread decreases.*

Consider an increase in the average order size of a liquidity trader, i.e., δ. The direct effect of such an increase is that the spread posted by the dealers decreases, other things equal. But for this reason, more speculators are active (See Equation (25)). This counterbalances the initial effect of the change in the average order size on the spread but never sufficiently to ultimately result in an increase of the spread. Exactly the same argument can be used to explain the impact of a change in β on the spread and the number of speculators. Note that both δ and β measure the trading activity (respectively the trading volume and the trading frequency) that is independent of speculators' activity.

When σ increases, the price revisions generated by news arrival are larger. Thus σ is a measure of price volatility. The third part of the proposition shows that the number of speculators is greater when price volatility is high. The speculator's expected profit increases with the size of the price revision, other things equal. Dealers' expected loss when they are picked off also increases with the size of the price revision. Combined with the increase in the number of speculators, this effect explains why the spread increases with price volatility. Battalio *et al.* (1997) find empirical evidence consistent with this prediction. Finally, when the number of dealers increases, each dealer can free-ride on larger number of dealers for market monitoring. Each dealer monitors less and the monitoring cost incurred by each dealer is lower. Thus, one component of the cost of market making is lower (the adverse selection component is not affected by M), which explains why the zero expected profit spread decreases. This triggers an increase in a speculator's expected profit, which explains why the number of speculators is positively related to the number of dealers.

The minimum quoted depth, Q, is of interest since it has been changed several times on Nasdaq; it was reduced from 1000 shares to 500 shares in January 1994, for most stocks, on a trial basis; it was restored to 1000 shares in March 1995 and eventually it has been reduced to 100 shares starting in January 1997. Nasdaq argued that the reduction of the minimum quoted depth would lessen SOES bandits activity and in this way would help to narrow spreads. We obtain the following result.

Proposition 9 *: When the dealer's minimum quoted depth (Q) decreases, the number of speculators decreases and the spread decreases.*

The direct effect of a decrease in Q is to reduce speculators' expected profit (See Equation (25)). For this reason, other things equal, the number of speculators decreases when the minimum quoted depth is reduced. But it follows that the risk of being picked off for the dealers is lower. This explains why the spread narrows when the minimum quoted depth is reduced.[20] In line with our prediction, Harris and Schultz (1997) find a decline in the number of trades initiated by SOES bandits after the reduction in the minimum quoted depth in 1994.[21]

Note that all the previous results have been obtained considering the zero expected profit equilibrium. It is straightforward to show that changes in parameters $\{Q, \delta, \sigma, \beta\}$ have a similar effect in the Pareto Dominant equilibrium. The impact of a change in the number of dealers on the spread depends on the nature of the equilibrium, however. This is the next result.

Proposition 10 *: In the Pareto-Dominant equilibrium,*

1. *For a given number of speculators, N, there exists $\Phi^*(M, N) \in (0, \hat{\Phi}(M, N))$ such that the market spread decreases with the number of dealers if $\Phi \in [0, \Phi^*(M, N)]$ and increases with the number of dealers if $\Phi \in [\Phi^*, \hat{\Phi}(M, N)]$ (where $\Phi^*(M, N)$ is characterized in the proof).*

2. *This implies that an increase in the number of dealers can result in an increase in the spread and a decrease in the number of speculators.*

Recall that each dealer monitoring is reduced when the number of dealers increases. Consequently an increase in the number of dealers posting the market spread results in lower monitoring cost for each dealer posting this spread. But this also means that a

[20] Interestingly it is possible to show that both dealers and speculators' aggregate monitoring decreases when the minimum quoted depth is reduced. Thus such a reduction has an adverse effect on price efficiency.

[21] Barclay *et al.* (1998) observe the same phenomenon after this quantity was reduced to 100 shares in 1997.

dealer who indercuts faces a larger increase in its monitoring cost, making undercutting less attractive as the number of dealers increases. On the other hand, each dealer executes a decreasing fraction of the order flow when the number of dealers increases. This effect encourages undercutting when spreads are above the zero expected profit spread. As claimed in the proposition, which of the two effects dominates, depends on Φ. If Φ is sufficiently large, an increase in the number of dealers help dealers in sustaining non-competitive spreads.

7 Empirical Analysis

Based on our model and the comparative static results presented in the previous section it is possible to address empirically the two main questions in the debate on the effect of SOES trading. Namely, does an increase in the SOES bandit activity lead to higher spreads and what factors determine the level of SOES bandit activity? At first glance, this might seem impossible given the multiplicity of equilibria. However, note that most of the comparative statics, except the effect of a change in the number of market makers, were identical in the Pareto-Dominant and the zero-profit equilibrium. Furthermore, the probability Φ can be interpreted as being constant across the stocks as it is mainly driven by the trading rules and the technology used. We will make this assumption here. This is clearly not the only possibility, but a more detailed analysis is beyond the scope of this study.

7.1 Empirical Predictions

Equations (26) and (27) determine the equilibrium number of speculators and the equilibrium bid-ask spread respectively. The bid-ask spread, S, and the speculator activity, N, are determined jointly and therefore Equations (26) and (27) can be estimated as a system of simultaneous non-linear equations. Rather than directly estimating this non-linear system, we consider the following system of linear simultaneous equations for stock i:

$$\begin{cases} S_i = a_1 + a_2 N_i + a_3 \sigma_i + a_4 M_i + a_5 \beta_i + a_6 \delta_i \\ N_i = a_1 + b_2 S_i + b_3 \sigma_i + b_4 Q_i \end{cases}$$

We do not observe the actual number of speculators. A natural measure of speculator activity is the unconditional probability of observing a trade initiated by a speculator. In our model, this probability is given by:

$$\alpha(Prob(f \in \mathcal{N}) + \Phi Prob(f \in \mathcal{M})) = \frac{\alpha N}{N + 1 - \Phi},$$

in equilibrium. Note that it is strictly increasing in the number of speculators N. The qualitative effects of a change in the exogenous parameters on the number of speculators and the probability of observing a trade initiated by a speculator are identical. Furthermore in our model speculators always place orders equal to the minimum quoted depth. Consequently we will use the probability of a maximum size SOES trade as our proxy for speculator or SOES bandit activity. Studies by Harris and Schultz (1997) and others suggest that the frequency of maximum size SOES trades provide a good proxy for SOES bandit activity.[22]

We use the volatility of the hourly mid-quote returns as a proxy for the size of a price revision, σ. Recall that δ determines the average size of liquidity demand in the theoretical model. The average size of all Nasdaq trades except maximum size SOES trades is a proxy for this theoretical variable. We will also use an indicator for stocks with different minimum quoted depth. For the period we consider there are stocks traded with a minimum quoted depth of 1000 and 500 shares. The bid-ask spread may, of course, depend on other variables outside our model. We will include two such variables, the price level and the market capitalization of the stock. The price level may affect the bid-ask spread if there is a "tick size" effect (See Kandel and Marx (1999)). The market capitalization may proxy for the number of analysts following a given stock and thereby affect the adverse selection component of the spread. For both variables we would therefore expect a negative

[22] Harris and Schultz (1997) document a dramatic shift in the frequency of trade size from 1000 to 500 shares when Nasdaq reduced the minimum quoted depth from 1000 shares to 500 shares in 1994. They also find that maximum size SOES trades are more likely to be motivated by information than other trades. Thus the number of maximum size SOES trades is a proxy for SOES bandits activity.

coefficient. In addition, we include the lagged probability of maximum SOES trades as an independent variable. This variable can potentially capture clustering effects in SOES bandit activity. If there is a clustering we expect this variable to have a positive effect on the level of SOES activity. The system of equations that we estimate is given by

$$\begin{cases} \text{SPREAD}_{ti} = & a_1 + a_2\text{SOES}_{ti} + a_3\text{VOLTY}_{ti} + a_4\text{NDLRS}_{ti} \\ & +a_5\text{LIQDEM}_{ti} + a_6 log(\text{AVGP}_{ti}) + a_7 log(\text{MKTCP}_{ti}), \\ & \\ \text{SOES}_{ti} = & b_1 + b_2\text{SPREAD}_{ti} + b_3\text{SOES}_{t-1,i} + b_4\text{VOLTY}_{ti} + b_5\text{MAXQ}_{ti}, \end{cases} \tag{29}$$

where $i = 1, \ldots, N$ index the stocks and $t = 1, \ldots, T$ index the trading days. The list below defines the variables used in the empirical analysis. The variables are daily averages for each stock i, $i = 1, \ldots, N$.

List of Variables

$SOES_{ti}$	The probability of a maximum size SOES trades (N) (the ratio of the number of maximum size SOES trades to the total number of trades)
$SPREAD_{ti}$	The percentage bid-ask spread (S)
$NDLRS_{ti}$	The number of market makers (M)
$VOLTY_{ti}$	The volatility of the hourly mid-quote returns (σ)
$LIQDEM_{ti}$	The average size of all Nasdaq trades excluding the maximum size SOES trades (δ)
$AVGP_{ti}$	The average price of the stock computed based on the mid-quotes
$MKTCP_{ti}$	The market capitalization of stock i [million dollars]
$MAXQ_{ti}$	Indicator, which is one if the minimum quoted depth is 1000 shares (Q)

The first equation determines the daily relative bid-ask spread as a function of the probability of a maximum size SOES trade, the volatility of the stock, the number of market makers, the liquidity demand, the logarithm of the price level, and the logarithm of the market capitalization. The probability of a maximum SOES order is expressed as a function of the bid-ask spread, the lagged value of the probability of a maximum size SOES trade, the volatility, and an indicator for the maximum SOES trade size.

The following exclusion restrictions imply that the system is overidentified. In the model, the bid-ask spread is determined by the per share costs of market making and is thus not directly dependent on the minimum quoted (SOES) depth (MAXQ). We assume that the current bid-ask spread is not affected by the lagged level of SOES activity (SOES(t-1)). This provides two restrictions for the spread equation. According to the theoretical model the number of market makers (NDLRS) and the liquidity demand (LIQDEM) do not affect the level of SOES activity directly. These variables only affect the SOES activity through their effect on the bid-ask spread. This provides three restrictions for the second equation. The model predicts that $a_2 > 0$ and $a_3 > 0$ whereas $a_5 < 0$. It also predicts that $b_2 < 0$, $b_4 > 0$, and $b_5 > 0$. Finally the sign of a_4 can be positive or negative depending on the nature of the equilibrium. Thus, the last prediction is the only one that is dependent on whether the relevant equilibrium is the Pareto-Dominant or the zero-profit one. Our estimation strategy can not deal with stocks that are traded in different types of equilibria.

7.2 The Data

We use data provided by NASDAQ on transactions and dealer quotes for a sample of NASDAQ stocks in December 1996. The sample consists of the 50 most actively traded stocks on NASDAQ for 1996 (measured by number of stock traded), for which the maximum order size was 1000 shares and a random sample of 50 stocks that were traded using a maximum SOES trading size of 500 shares.

Table 3 provides summary statistics for the variables we use in the analysis. Each variable is averaged for each stock over the one-month sample and the cross-sectional means, medians, standard deviations, minimum, and maximum values are reported (N=50 stocks and T=21 trading days). For Panel A, the average daily number of maximum size SOES trades is 437. There is however large cross-sectional variation with a minimum of 4 and a maximum of 2916. The bid-ask spread is on average 0.71% for the sample stocks. The standard deviation of 0.58% indicates that there is also large variation in the spread across the stock in our sample. The number of market makers varies between a low of 18 and a high of about 63. There is also variation across time for individual stocks. The volatility and trading activity variables indicate that there are relatively large fluctuations in prices and active trading for stocks in Panel A. For stocks in Panel B, the maximum order size is 500 shares. Those stocks are small capitalization stocks. Their spreads are on average larger than for stocks in Panel A. These stocks are also less actively traded and have a lower number of dealers on average. Interestingly, the average number of maximum orders size SOES trades is much lower for Panel B than for Panel A. According to the model, this can be due both to the lower maximum order size and to the larger spread for the stocks in Panel B.

7.3 Empirical Results

The system of equations presented above was estimated using GMM with all the predetermined variables as instruments. The results are reported in Table 4. The results for the stocks with a maximum SOES size of 1000 (500) are reported in Panel A (B). Panel C reports the results for the combined sample of stocks.

Based on the model we expect increased SOES activity to be associated with larger spreads. While this coefficient is positive in Panel B it is not significant. Thus, we do not find evidence that higher levels of SOES activity lead to higher spread for these stocks. Even more surprising, in the second sample (Panel B), the SOES activity coefficient is negative and significant. This result is not consistent with the model predictions. For the combined sample the effect found in the second sub-sample seems to dominate producing a statistically significant negative coefficient. The coefficient on the number of market makers is negative in all three samples. This result is not consistent with the Pareto-Dominant equilibrium where we would expect a positive coefficient for the number of market makers. It is, of course, not possible to conclude that we are in the zero-profit equilibrium based on these results. Within the model, this finding only suggests the data is consistent with an equilibrium where the negative externality is not too strong (Φ is small). The other model variables, volatility and liquidity demand, have the expected signs in all three equations.

The level of SOES activity is in all three equations negatively related to the spread and positively related to volatility as predicted by the model. The coefficient on the minimum quoted depth is positive and statistically significant, which is also consistent with our model.

The system of equations is overidentified. The test statistics for a test of overidentifying restrictions is reported at the bottom of each panel. The tests do not reject the model.

8 Conclusion

We present a model of price formation with costly monitoring of the information flow. In particular we consider a trading environment where speculators, different from dealers, monitor the arrival of news and quote updates. We show how information externalities associated with monitoring arise and how they affect the price formation process. In particular we find that quote matching is a way for dealers to free ride on monitoring by the other dealers, which leads to equilibria in which posted spreads are above the competitive level. Allowing dealers to decline the execution of an incoming order strengthens dealers incentive to free ride on information production by other dealers. For this reason, such a policy can result in wider spreads and be detrimental to price discovery.

Our model captures key features of Nasdaq's Small Order Execution System (SOES). The alleged effects of this trading systems on trading costs has generated intense regulatory interest. Using our model predictions we revisit the main empirical questions regarding the effects of SOES. We find that the level of SOES bandit activity is positively related to the minimum quoted depth in SOES, the volatility, and negatively related to the spread as predicted by our model. Surprisingly we find either no effect or a negative effect of SOES bandit activity on the bid-ask spread. This empirical finding contradicts the claim that SOES bandits can be held responsible for large spreads on Nasdaq.

In our model, it is always optimal for the market makers to adjust their quotes as they discover new information. This is consistent with the fact that dealers can not automatically execute trades against dealers in Nasdaq.[23] In other market structures, the relevant choice may be to either update the quotes or to conceal the information and attempt to

[23]Note that even if they had the choice they may choose not to do so for reputation reasons.

trade against other market participants. This is an interesting question for future research.

9 Appendix

Proof of Proposition 1

The R.H.S. of Equation (8) is positive **if and only if** the expression in square brackets is, i.e., as long as $(1-\frac{\Phi M_b}{M_b-1})\gamma_A - (\frac{\Phi M_b}{M_b-1}\lambda_i) \geq 0$. Hence $\bar{\Phi}$ follows directly. *Q.E.D*

Proof of Lemma 1.

Suppose (to be contradicted) that there exists a Nash equilibrium in which some dealers do not choose the same monitoring levels. Consider two dealers i and i' such that $\lambda_i^* > \lambda_{i'}^*$. Using the fact that Equation (9) must hold for these two dealers, we obtain the following equality:

$$\frac{\alpha Q(\sigma - S_b)}{2M_b(\lambda_A+\gamma_A)^2}\left[(\frac{\Phi M_b}{M_b-1})(\lambda_{i'}^* - \lambda_i^*)\right] = \Psi_d^{'}(\lambda_i^*) - \Psi_d^{'}(\lambda_{i'}^*).$$

Since (a) $\lambda_i^* > \lambda_{i'}^*$ and (b) $\Psi_d(.)$ is strictly convex, the L.H.S of this inequality is strictly negative whereas the R.H.S is strictly positive, which is impossible. This implies that in equilibrium all the dealers choose the same monitoring level. In the same way we can prove that in equilibrium all the speculators must choose the same monitoring level. This proves that the Nash equilibrium in the monitoring stage must be symmetric.*Q.E.D.*

Proof of Proposition 2.

Dividing Equation (12) by Equation (13), we find that λ^* and γ^* must satisfy:

$$\frac{N\gamma^* + \Phi M_b \lambda^*}{\frac{(N-\Phi)}{N}M_b\lambda^* + (N-1)\gamma^*} = r(\frac{M_b\lambda^*}{\gamma^*}).$$

This equation can be written as an equation with unknown $\Upsilon \equiv \frac{M_b\lambda^*}{\gamma^*}$. Since the monitoring levels must be positive, it must be the case that $\Upsilon \geq 0$. The previous equation imposes:

$$\frac{N + \Phi\Upsilon}{\frac{(N-\Phi)}{N}\Upsilon + (N-1)} = r\Upsilon.$$

This equation has two solutions but only one is positive. This solution is:

$$\Upsilon = \frac{N}{2r(N-\Phi)}\left(\sqrt{(r(N-1)-\Phi)^2+4r(N-\Phi)}-((N-1)r-\Phi)\right).$$

Substituting λ^* by $\frac{\Upsilon\gamma^*}{M_b}$ in Equation (13), we find that γ^* solves:

$$\frac{\alpha Q(\sigma-S_b)(\Upsilon(\frac{N-\Phi}{N})+(N-1))}{\gamma^*(\Upsilon+N)^2}=c_s\gamma^*.$$

There is a unique positive solution to this equation, which yields the closed form solution for γ^*. Since Υ and γ^* are uniquely defined, there is a unique Nash equilibrium in the monitoring stage.

Substituting the expressions for λ^* and γ^* in Equations (5) and (6), we obtain that in equilibrium:

$$Prob(f \in \mathcal{N}) = \frac{N}{\Upsilon+N}, \tag{30}$$

and

$$Prob(f \in \mathcal{M}_b \backslash i) = \frac{(M_b-1)\Upsilon}{M_b(\Upsilon+N)}. \tag{31}$$

Direct substitution of these probabilities in Equations (7) and (10) yield the dealer and speculator profit functions. *Q.E.D.*

Proof of Proposition 3. Immediate from arguments in the text.

Proof of Proposition 4.

Suppose that the outcome of the quoting stage is $\{S_b, M_b\}$. The expected profit of a dealer posting the market spread is:

$$\Pi_d(\lambda^*(S_b,M_b),\gamma^*(S_b,M_b),M_b) = \frac{Q}{2M_b}[-\alpha(\sigma-S_b)R(M_b,0,r)+(1-\alpha)\beta\delta S_b]$$

Note that $R(M_b,0,r)$ decreases with M_b. This means that if S_b is such that M_b dealers obtain positive expected profits then M_b+1 dealers must obtain strictly positive expected profit when they post this spread. This proves the first part of the proposition. But this means that we cannot construct an equilibrium in which $M_b < M$ dealers are active at the best offers and the remaining dealers are better off not matching the best offers. *Q.E.D*

Proof of Corollary 1.

As $\Phi = 0$, Υ can be written:

$$\Upsilon = \sqrt{\frac{(N-1)^2}{4} + \frac{N}{r}} - \frac{(N-1)}{2}.$$

It follows that $\frac{\partial \Upsilon}{\partial r} < 0$. Furthermore it can be checked that: $r\Upsilon(\Upsilon + (N-1)) = N$. Consequently, $R(M_b, \Phi, r) = \frac{N}{\Upsilon + N} + \frac{\Upsilon N}{2M_b(\Upsilon+N)^2}$. Thus we obtain:

$$\frac{\partial R}{\partial r} = -\frac{\partial \Upsilon}{\partial r}(\Upsilon + N)^{(-2)} \left(N - \frac{N}{2M} + \frac{N\Upsilon}{M(\Upsilon + N)} \right) > 0.$$

Since $\hat{S}$ increases with R, it follows that $\frac{\partial \hat{S}}{\partial r} > 0$. Υ becomes infinite when r goes to zero and goes to zero when r becomes infinite. Thus $lim_{r\to 0} R(M, 0, r) = 0$ and $lim_{r\to +\infty} R(M, 0, r) = 1$. Consequently:

$$lim_{r\to 0} \hat{S}(M, \Phi, r) = 0 \quad and \quad lim_{r\to +\infty} \hat{S}(M, \Phi, r) = \frac{\alpha\sigma}{\alpha + (1-\alpha)\beta\delta}.$$

Similar computations yield the results for $\bar{S}.Q.E.D$

Proof of Lemma 2.

Using Equations (22) and (24), we get:

$$\frac{\partial R(M, \Phi)}{\partial \Phi} > 0 \qquad and \qquad \frac{\partial \bar{R}(\Phi)}{\partial \Phi} < 0.$$

Since $\hat{S}$ ($\bar{S}$) increases with $R(\Phi)$ ($\bar{R}(\Phi)$), we obtain the first part of the lemma. The condition $\hat{S}(M, \Phi) \leq \bar{S}(\Phi)$ is equivalent to $R(M, \Phi) \leq \bar{R}(\Phi)$. This condition turns to be equivalent to:

$$(1 + N - \Phi)\left[(1+N)(1-2\Phi) - \Phi\right] - \frac{(N+1)^2}{M} \geq 0.$$

The L.H.S of this equation is decreasing with Φ. It is equal to $\frac{(1+N)^2(M-1)}{M}$ when $\Phi = 0$ and it is strictly negative for $\Phi \geq \frac{1}{2}$. Thus there exists $\hat{\Phi}(M, N) \in (0, 1/2)$ such that the previous condition is satisfied if and only if $\Phi \leq \hat{\Phi}(M, N)$. The threshold $\hat{\Phi}(M, N)$ solves:

$$(1+N-\Phi)\left[(1+N)(1-2\Phi)-\Phi\right]-\frac{(N+1)^2}{M}=0$$

This equation has two solutions, but only one is lower than 1. This solution is:

$$\hat{\Phi}(M,N)=\frac{(1+N)(2+N)}{3+N}[1-\sqrt{(1-\frac{(M-1)(3+N)}{M(2+N)^2}}].$$

Note that $\hat{\Phi}(.,.)$ increases with M and $N.Q.E.D$

Proof of Proposition 5

It is direct that $R(M_b, \Phi)$ decreases with M_b for $M_b \geq 2$. Thus, using exactly the same argument as in the proof of Proposition 4, we can show that there cannot be an equilibrium with $1 < M_b^* < M$. Recall that $R(1,\Phi) = R(1,0)$. As there exists values of Φ such that $R(1,0) < R(2,\Phi)$, we cannot exclude in this case the possibility of equilibria with only one dealer posting the best offers.

Equilibria in which $M_b^* = M$ exist **if and only if** $\Phi \leq \hat{\Phi}$ since $\bar{S}(\Phi) \geq \hat{S}(M,\Phi)$ if and only if $\Phi \leq \hat{\Phi}$. In this case if a dealer expects his competitors to post a market spread $S_b^* \in [\hat{S}(M,\Phi), \bar{S}(\Phi)]$ then he optimally matches their offers, which proves that $\{M, S_b^*\}$ is a Nash equilibrium.

When $\Phi > \hat{\Phi}$, the previous arguments impose that $M_b^* = 1$ if an equilibrium exists. For an equilibrium with only one dealer posting the best offers to exist, three conditions must be satisfied. First the active dealer should not widen his spread in the quoting stage. This requires $S_b^* \geq \hat{S}(1,0)$ (recall that when $M_b = 1$, everything is as if $\Phi = 0$). Second among the dealers who do not post the market spread, none should be better off undercutting. This requires that the dealer posting the market spread obtains zero expected profit, i.e., $S_b^* = \hat{S}(1,0)$. Third among the dealers who do not post the market spread, none should be better off matching the best offers. Using Equation (16), this imposes:

$$R(2,\Phi) > R(1,0)$$

We show that this is the case if $\Phi > \hat{\Phi}$. First note that when $\Phi = \hat{\Phi}$, $\hat{S}(M,\hat{\Phi}) = \bar{S}(\hat{\Phi})$. This implies that the expected profit of a dealer is zero in equilibrium when $\Phi = \hat{\Phi}$. In this case a dealer is just indifferent between undercutting the zero expected profit market spread with M dealers or matching this spread. Therefore $\hat{S}(M,\hat{\Phi}) = \hat{S}(1,0)$ which is equivalent to $R(M,\hat{\Phi}) = R(1,0)$. Since $R(M,.)$ increases with Φ, it is the case that $R(M,\Phi) > R(M,\hat{\Phi})$ when $\Phi > \hat{\Phi}(M,N)$. Furthermore $R(.,\Phi)$ decreases with M_b for $M_b \geq 2$. It follows that:

$$R(2,\Phi) > R(M,\Phi) > R(M,\hat{\Phi}) \quad for \quad \Phi > \hat{\Phi},$$

But since $R(M,\hat{\Phi}) = R(1,0)$, then $R(2,\Phi) > R(1,0)$ for $\Phi > \hat{\Phi}$. It follows that no dealer can profitably match the offers of the dealer posting $\hat{S}(1,0)$. *Q.E.D*

Proof of Proposition 6. Each dealer's expected profit increases with the market spread (see Equation 16). Thus the equilibrium that features the largest spread is Pareto-Dominant. *Q.E.D*

Proof of Corollary 2. Immediate using Lemma 2 and Proposition 5

Proof of Corollary 3.

Suppose $\Phi \leq \hat{\Phi}$. In the zero expected profit equilibrium, the monitoring level of a dealer is:

$$\lambda^*(\Phi) = \sqrt{\frac{N(\alpha(1-\alpha)\sigma Q\beta\delta)}{cM^2[\alpha R(M,\Phi) + (1-\alpha)\beta\delta](1+N-\Phi)^2}}.$$

This can be written as:

$$\lambda^*(\Phi) = \sqrt{\frac{N(\alpha(1-\alpha)\sigma Q\beta\delta)}{cM^2\left[\alpha\left(N(1+N-\Phi) + \frac{N}{2M_b}\right) + ((1-\alpha)\beta\delta)(1+N-\Phi)^2\right]}}.$$

It follows that $\frac{\partial\lambda^*}{\partial\Phi} > 0$ in this case. In the Pareto Dominant equilibrium, we obtain the same expression for λ^*, but $R(M,\Phi)$ is replaced by $\bar{R}(\Phi)$. As $\bar{R}(\Phi)$ decreases with Φ, it is direct that dealers' monitoring level increases with Φ. Thus, independtly of the equilibrium we consider in the quoting stage, we obtain:

$$\lambda^*(0) < \lambda^*(\Phi) \quad \forall\Phi \leq \hat{\Phi}. \tag{32}$$

For $\Phi > \hat{\Phi}$, $S_b^* = S(1,0)$. Thus we obtain:

$$\lambda^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - S(1,0))}{c(1+N)^2}} \quad \forall \Phi > \hat{\Phi}$$

Note that in this case $\lambda^*(\Phi)$ does not depend on Φ. Let $\bar{\lambda} \equiv \lambda^*(\Phi)$ for $\Phi > \hat{\Phi}$. Recall that $\hat{S}(M, \hat{\Phi}) = \bar{S}(M, \hat{\Phi}) = S(1,0)$. This implies (i) that $\lambda^*(\hat{\Phi})$ takes the same value in the zero expected equilibrium and in the Pareto Dominant equilibrium and (ii) that this value is:

$$\lambda^*(\hat{\Phi}) = \sqrt{\frac{N\alpha Q(\sigma - S(1,0))}{cM^2(1+N-\Phi)^2}}$$

Using the fact that $\hat{\Phi} < \frac{1}{2}$, computations show that $\lambda^*(\hat{\Phi}) < \bar{\lambda}$. Combining this inequality with Equation (32), we obtain:

$$\lambda^*(0) < \lambda^*(\hat{\Phi}) \quad \forall \Phi.$$

Q.E.D.

Proof of Corollary 4.

Suppose $\Phi \leq \hat{\Phi}$. In this case, $M_b^* = M$. Using Proposition 2, we obtain that the aggregate monitoring level is:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) = M\lambda^* + N\gamma^* = \Upsilon(N+1-\Phi)\gamma^*.$$

which yields:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - S_b^*)}{c}}$$

For $\Phi > \hat{\Phi}$, $M_b^* = 1$. It follows that the aggregate monitoring level is:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) = \lambda^* + N\gamma^* = (N+1)\gamma^*$$

which is:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - S(1,0))}{c}}$$

Now consider the zero expected profit equilibrium. In this case, $S_b^* = \hat{S}(M,\Phi)$ for $\Phi \leq \hat{\Phi}$ and $\hat{S}(M,\hat{\Phi}) = S(1,0)$. As $\hat{S}(M,\Phi)$ increases with Φ, it follows from the previous equations that:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) < \lambda_A^*(0) + \gamma_A^*(0) \quad \forall \Phi > 0.$$

Now consider the Pareto-Dominant equilibrium. In this case, $S_b^* = \bar{S}(\Phi)$ for $\Phi \leq \hat{\Phi}$ and $\bar{S}(M,\hat{\Phi}) = S(1,0)$. As $\bar{S}(M,\Phi)$ decreases with Φ, we now obtain that:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) > \lambda_A^*(0) + \gamma_A^*(0) \quad \forall \Phi > 0.$$

Q.E.D

Proof of Corollary 7.

Case 1: $\Phi \leq \hat{\Phi}$.

Computations yield

$$\frac{\partial R(M,\Phi)}{\partial N} = \frac{(N+1-\Phi)[(1-\Phi)2M+1] - 2N}{2M(1+N-\Phi)^3}.$$

As $\Phi \leq \hat{\Phi} < \frac{1}{2}$, we obtain $\frac{\partial R}{\partial N} > 0$ which implies that $\frac{\partial \hat{S}}{N} > 0$. Furthermore, we obtain:

$$\frac{\partial \bar{R}(\Phi)}{\partial N} = \frac{\frac{M(N+3)}{2(N+1)^3} - \frac{\partial R(M,\Phi)}{\partial N}}{M-1}.$$

Computations show that $\frac{\partial^2 R(M,\Phi)}{\partial \Phi \partial N} < 0$. It follows that $\frac{\partial^2 \bar{R}(\Phi)}{\partial \Phi \partial N} > 0$. Now for $\Phi = 0$, we get:

$$\bar{R}(0) = \frac{N}{N+1} + \frac{(M+1)N}{2M(N+1)^2},$$

which increases with N. This finally yields:

$$\frac{\partial \bar{R}(\Phi)}{\partial N} > \frac{\partial \bar{R}(0)}{\partial N} > 0. \tag{33}$$

It follows that $\bar{S}(\Phi)$ increases with N.

Case 2: $\Phi > \hat{\Phi}$**.** In this case the equilibrium spread is:

$$\hat{S}(1,0) = \alpha\sigma(\frac{R(1,0)}{\alpha R(1,0) + (1-\alpha)\beta\delta}) \tag{34}$$

$R(1,0)$ increases with N which implies that $\hat{S}(1,0)$ increases with $N.Q.E.D$

Proof of Proposition 8.

Equation (28) implicitly defines N^* in term of the exogenous parameters. The L.H.S of this equation is a speculator's expected profit in equilibrium. We denote it by Π_s^*. Consider the effect of a change in δ. Using Equation (28), we obtain:

$$\frac{dN^*}{d\delta} = -\frac{\frac{\partial \Pi_s^*}{\partial \delta}}{\frac{\partial \Pi_s^*}{\partial N^*}}$$

It is straightforward that $\frac{\partial \Pi_s^*}{\partial \delta} > 0$ and we know that $\frac{\partial \Pi_s^*}{\partial N^*} < 0$. It follows that $\frac{dN^*}{d\delta} > 0$. Then recall that:

$$\Pi_s(\lambda^*(\hat{S}), \gamma^*(\hat{S}), \hat{S}, N^*) = K = (\sigma - \hat{S})Q\alpha \left[\frac{2N^*(N^*+1-\Phi) - (N^*-\Phi)^2)}{4N^*(N^*+1-\Phi)^2}\right].$$

As the number of speculators increases with δ, the term in bracket in the R.H.S. of the previous equation decreases with δ. But we have found that a speculator's expected profit (the L.H.S. of the equation) increases with δ, in equilibrium. Consequently the market spread $\hat{S}$ must decrease when δ increases. The same type of argument can be used to derive the impact of σ, β and M on the zero expected profit market spread and the number of speculators. *Q.E.D.*

Proof of Proposition 9.

Using Equation (28), we obtain:

$$\frac{dN^*}{dQ} = -\frac{\frac{\partial \Pi_s^*}{\partial Q}}{\frac{\partial \Pi_s^*}{\partial N^*}}.$$

It is straightforward that $\frac{\partial \Pi_s^*}{\partial Q} > 0$. It follows that $\frac{dN^*}{dQ} > 0$. Since $\hat{S}$ increases with the number of speculators, it follows that $\hat{S}$ increases with Q.

Proof of Proposition 10.

The market spread in the Pareto-Dominant equilibrium increases with $\bar{R}(\Phi)$, which is given in Equation (24). Computations yield:

$$\frac{\partial \bar{R}(\Phi)}{\partial M} = \frac{1}{(M-1)^2}\left[R(M,\Phi) - R(1,0) + \frac{N(M-1)}{2M^2(1+N-\Phi)^2}\right]$$

The term in bracket increases with Φ. It is strictly negative for $\Phi = 0$ and strictly positive for $\Phi = \hat{\Phi}$ (because $R(1,0) = R(M,\hat{\Phi})$). Thus there exists $\Phi^* \in (0,\hat{\Phi})$ such that $\frac{\partial \bar{R}(\Phi^*)}{\partial M} = 0$. For $\Phi < \Phi^*$, $\frac{\partial \bar{R}(\Phi)}{\partial M} < 0$ and for $\Phi > \Phi^*$, $\frac{\partial \bar{R}(\Phi)}{\partial M} > 0$. This proves the first part of the corollary.

For the Pareto-Dominant equilibrium, the equilibrium number of speculators is determined by the following equation:

$$\left(\frac{\alpha(1-\alpha)\beta\delta Q\sigma}{\alpha\bar{R}(\Phi) + (1-\alpha)\beta\delta}\right)\left[\frac{2N^*(N^*+1-\Phi) - (N^*-\Phi)^2)}{4N^*(N^*+1-\Phi)^2}\right] = K. \tag{35}$$

which is identical to Equation (28), except that $\bar{R}(\Phi)$ replaces $R(M,\Phi)$. Now consider an equilibrium in which the number of dealers, the equilibrium number of speculators and Φ are such that: $\Phi \in (\Phi^*(M,N^*), \hat{\Phi}(M,N^*)$. Using the fact that in this case $\frac{\partial \bar{R}(\Phi)}{\partial M} > 0$ and Equation (35), we can proceed as in the proof of Proposition 8 to prove the second part of Proposition 10.$Q.E.D$

References

[1] Amihud, Y., and Mendelson, H., 1991, How (Not) to Integrate the European Capital Markets, in A. Giovannini and L. Mayer (eds) *European Financial Integration*, Cambridge University Press 1991, 73–111.

[2] Barclay, M., Christie, W., Harris, J., Kandel, E., and Schultz, P., 1999, The Effect of Market Reform on the Trading Costs and Depths of Nasdaq Stocks, *Journal of Finance.*, 54, 1-34.

[3] Battalio, R., Hatch, B., and Jennings, R., 1997, SOES Trading and Market Volatility, *Journal of Financial and Quantitative Analysis,* 32, 225-238.

[4] Copeland, T., and Galai, D., 1983, Information Effects on the Bid-Ask Spread, *Journal of Finance,* 38, 1457-1469.

[5] Fudenberg, D., and Tirole, J. 1991, Game Theory, MIT Press.

[6] General Accounting Office, 1998, The Effects of SOES on the Nasdaq Market, United States General Accounting Office Report 98-194.

[7] Harris, J., and Schultz, P., 1997, The Importance of Firm Quotes and Rapid Executions: Evidence from the January 1994 SOES rules change, *Journal of Financial Economics,* 45, 135-166.

[8] Harris, J., and Schultz, P., 1998, The Trading Profits of SOES Bandits, *Journal of Financial Economics,* 50, 39-62.

[9] Hinden, S., 1994, Nasdaq's Big Guns Send Trading Bandits Packing, Washington Post, February 7, 1994.

[10] Houtkin, H., 1998, Secrets of the SOES Bandit, McGraw-Hill.

[11] Kandel, E., and Marx, L., 1999, Odd-eight Avoidance as a Defense Against SOES Bandits, *Journal of Financial Economics.*, 51, 85-102.

[12] Kandel, E., and Marx, L., 1997, Nasdaq Market Structure and Spread Patterns, *Journal of Financial Economics*, 35, 61-90.

[13] Stoll, H., 1992, Principles of trading market structure, *Journal of Financial Services Research*, 6, 75-107.

[14] Whitcomb, D., 1998, The NASDAQ Small Order Execution System: Myth and Reality, testimony before the House Committee on Commerce, Subcommittee on Finance, August 3, 1998.

Event	Action	Probability	Dealer i's Payoff
Liquidity Trader	Submits Buy Order	$\frac{(1-\alpha)\beta}{2}$	$\frac{\delta Q S_b}{2M_b} - \Psi_d(\lambda_i)$
	Submits Sell Order	$\frac{(1-\alpha)\beta}{2}$	$\frac{\delta Q S_b}{2M_b} - \Psi_d(\lambda_i)$
Good News	A Speculator Observes News First	$\frac{\alpha}{2} Prob(f \in \mathcal{N})$	$-\frac{Q(\sigma - S_b)}{2M_b} - \Psi_d(\lambda_i)$
$(v_1 = v_0 + \frac{\sigma}{2})$	Dealer i Observes News First	$\frac{\alpha}{2} Prob(f = i)$	$-\Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Speculator is Second	$\frac{\alpha}{2} \Phi Prob(f \in \mathcal{M}_b \backslash i)$	$-\frac{Q(\sigma - S_b)}{2(M_b - 1)} - \Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Dealers Follow	$(1 - \Phi) \frac{\alpha}{2} Prob(f \in \mathcal{M}_b \backslash i)$	$-\Psi_d(\lambda_i)$
Bad News	A Speculator Observes News First	$\frac{\alpha}{2} Prob(f \in \mathcal{N})$	$-\frac{Q(\sigma - S_b)}{2M_b} - \Psi_d(\lambda_i)$
$(v_1 = v_0 - \frac{\sigma}{2})$	Dealer i Observes News First	$\frac{\alpha}{2} Prob(f = i)$	$-\Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Speculator is Second	$\frac{\alpha}{2} \Phi Prob(f \in \mathcal{M}_b \backslash i)$	$-\frac{Q(\sigma - S_b)}{2(M_b - 1)} - \Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Dealers Follow	$(1 - \Phi) \frac{\alpha}{2} Prob(f \in \mathcal{M}_{\rfloor} \backslash i)$	$-\Psi_d(\lambda_i)$
No Order/No Information		$(1 - \beta)(1 - \alpha)$	$-\Psi_d(\lambda_i)$

Table 1: Dealer i's Payoffs

Event	Action	Probability	Speculator j's Payoff
Liquidity Trader	Submits Buy Order	$\frac{(1-\alpha)\beta}{2}$	$-\Psi_s(\gamma_j)$
	Submits Sell Order	$\frac{(1-\alpha)\beta}{2}$	$-\Psi_s(\gamma_j)$
Good News	Speculator j observes the news first	$\frac{\alpha}{2}Prob(f=j)$	$\frac{Q}{2}(\sigma - S_b) - \Psi_s(\gamma_j)$
$(v_1 = v_0 + \frac{\sigma}{2})$	Speculator k, $k \neq j$, is first to observe news	$\frac{\alpha}{2}Prob(f \in \mathcal{N}\backslash j)$	$-\Psi_s(\gamma_j)$
	Dealer Updates Quote and Speculator j Reacts First	$\frac{\alpha}{2N}\Phi Prob(f \in \mathcal{M}_b)$	$\frac{Q}{2}(\sigma - S_b) - \Psi_s(\gamma_j)$
	Dealer Updates Quotes and Speculator j is Not First to React	$(1-\Phi)\frac{\alpha}{2}Prob(f \in \mathcal{M}_b)$	$-\Psi_s(\gamma_j)$
Bad News	Speculator j observes the news first	$\frac{\alpha}{2}Prob(f=j)$	$\frac{Q}{2}(\sigma - S_b) - \Psi_s(\gamma_j)$
$(v_1 = v_0 - \frac{\sigma}{2})$	Speculator k, $k \neq j$, is first to observe news	$\frac{\alpha}{2}Prob(f \in \mathcal{N}\backslash j)$	$-\Psi_s(\gamma_j)$
	Dealer Updates Quote and Speculator j Reacts First	$\frac{\alpha}{2N}\Phi Prob(f \in \mathcal{M}_b)$	$\frac{Q}{2}(\sigma - S_b) - \Psi_s(\gamma_j)$
	Dealer Updates Quotes and Speculator j is Not First to React	$(1-\Phi)\frac{\alpha}{2}Prob(f \in \mathcal{M}_b)$	$-\Psi_s(\gamma_j)$
No Order/No Information		$(1-\beta)(1-\alpha)$	$-\Psi_s(\gamma_j)$

Table 2: Speculator j's Payoffs

Table 3: Summary Statistics

Panel A: Maximum SOES quantity of 1000 (N=50, T=21)

Variable	Mean	Median	Std. Dev.	Min	Max
SOES	437	202	601	4	2916
SPREAD	0.0071	0.0058	0.0051	0.0010	0.0239
NDLRS	37	36	9	18	64
VOLTY	0.0087	0.0083	0.0036	0.0029	0.0240
NTRDS	1743	1183	1613	189	7846
LIQDEM	1617	1347	761	821	4101
AVGP	38.71	33.25	27.49	3.22	132.30
MKTCP	12000	3487	24905	90	108619

Panel B: Maximum SOES quantity of 500 (N=50, T=21)

Variable	Mean	Median	Std. Dev.	Min	Max
SOES	10	9	8	0	41
SPREAD	0.0197	0.0188	0.0068	0.0078	0.0474
NDLRS	10	10	5	3	23
VOLTY	0.0110	0.0100	0.0038	0.0039	0.0192
NTRDS	140	106	119	3	509
LIQDEM	1489	1362	659	373	3816
AVGP	24.12	21.44	13.12	7.95	74.51
MKTCP	539	353	551	30	3109

Table 3 reports summary statistics on the daily number of maximum size SOES trades (SOES), the bid-ask spread (SPREAD), the number of market makers (NDLRS), the hourly mid-quote volatility (VOLTY) , the number trades (excluding maximum size SOES trades) (NTRDS), the average trade size (LIQDEM), and the average mid-quote (AVGP), and the market capitalization (MKTCP)in million dollars for the stocks in our samples. Panel A presents the statistics for a sample of 50 actively traded NASDAQ stocks for December 1996. These stocks traded with a maximum SOES trade size of 1000 shares. Panel B presents the statistics for a sample of 50 NASDAQ stocks, which traded with a maximum SOES size of 500 units, in December 1996. All figures are computed by first computing a cross-section of stock specific time-series averages for all variables and then by computing the different statistics based on this cross-section.

Table 4: Estimation Results

	Panel A: Max SOES Size 1000 (N=50,T=20)				Panel B: Max SOES Size 500 (N=50,T=20)				Panel C: Full Sample (N=100,T=20)			
	Spread Eq.		SOES Eq.		Spread Eq.		SOES Eq.		Spread Eq.		SOES Eq.	
Variable	Coeff.	t-stat.	Coeff.	t-stat.	Coeff.	t-stat.	Coeff.	t-stat.	Coeff.	t-stat.	Coeff.	t-stat.
Const.	2.99×10^{-2}	70.200	3.39×10^{-2}	9.430	5.65×10^{-2}	38.600	1.85×10^{-2}	2.880	4.07×10^{-2}	44.500	3.26×10^{-2}	7.270
SOES(t)	5.07×10^{-4}	0.595			-5.02×10^{-2}	-8.370			-1.69×10^{-2}	-8.460		
SPREAD			-4.46×10^{0}	-14.100			-3.88×10^{-1}	-1.300			-2.12×10^{0}	-9.080
SOES(t-1)			3.20×10^{0}	16.300			1.99×10^{0}	14.400			2.16×10^{0}	13.100
VOLTY	2.86×10^{-2}	2.910	7.95×10^{-1}	68.800	1.11×10^{-1}	4.830	4.63×10^{-1}	22.900	5.06×10^{-2}	2.840	7.45×10^{-1}	50.200
NDLRS	-7.75×10^{-5}	-10.600			-8.03×10^{-4}	-19.900			-2.83×10^{-4}	-18.500		
LIQDEM	-1.91×10^{-7}	-3.240			-2.91×10^{-7}	-2.100			-5.65×10^{-8}	-0.470		
AVGP)	-6.44×10^{-3}	-33.900			-6.63×10^{-3}	-13.300			-4.57×10^{-3}	-11.300		
MKTCP	2.03×10^{-4}	2.520			-1.02×10^{-3}	-4.450			-6.35×10^{-4}	-3.530		
MAXQ											2.18×10^{-7}	2.830
J-test	1.299	(0.245))			2.099	(0.471)			2.31	(0.196)		

Table 4 presents the estimation results for the system of equations describing the joint determination of the bid-ask spread (SPREAD) and the SOES activity (SOES). The exogenous variables are the lagged SOES activity (SOES(t-1)), volatility (VOLTY), numbre of market makers (NDLRS), the average order size for non-maximum SOES trades (LIQDEM), the logarithm of the average price (AVGP), the logarithm of the market capitalization (MKTCP), and the indicator for a maximum SOES order size of 1000 shares (MAXQ). Panel A presents the results for a sample of 50 actively traded NASDAQ issues for December 1996. Panel B presents the results for a sample of 50 NASDAQ issues traded using a maximum SOES size of 500 stocks. Panel C presents the results for the full sample of 100 stocks. The system of equations presented in Equation (29) is estimated as a system using GMM with all the exogenous variables as instruments. The coefficient estimates are reported together with the t-statistics in parenthesis. The J-test of overidentifying restrictions is reported on the last line with the corresponding p-values in parenthesis.

Figure 1: Timing of the trading game

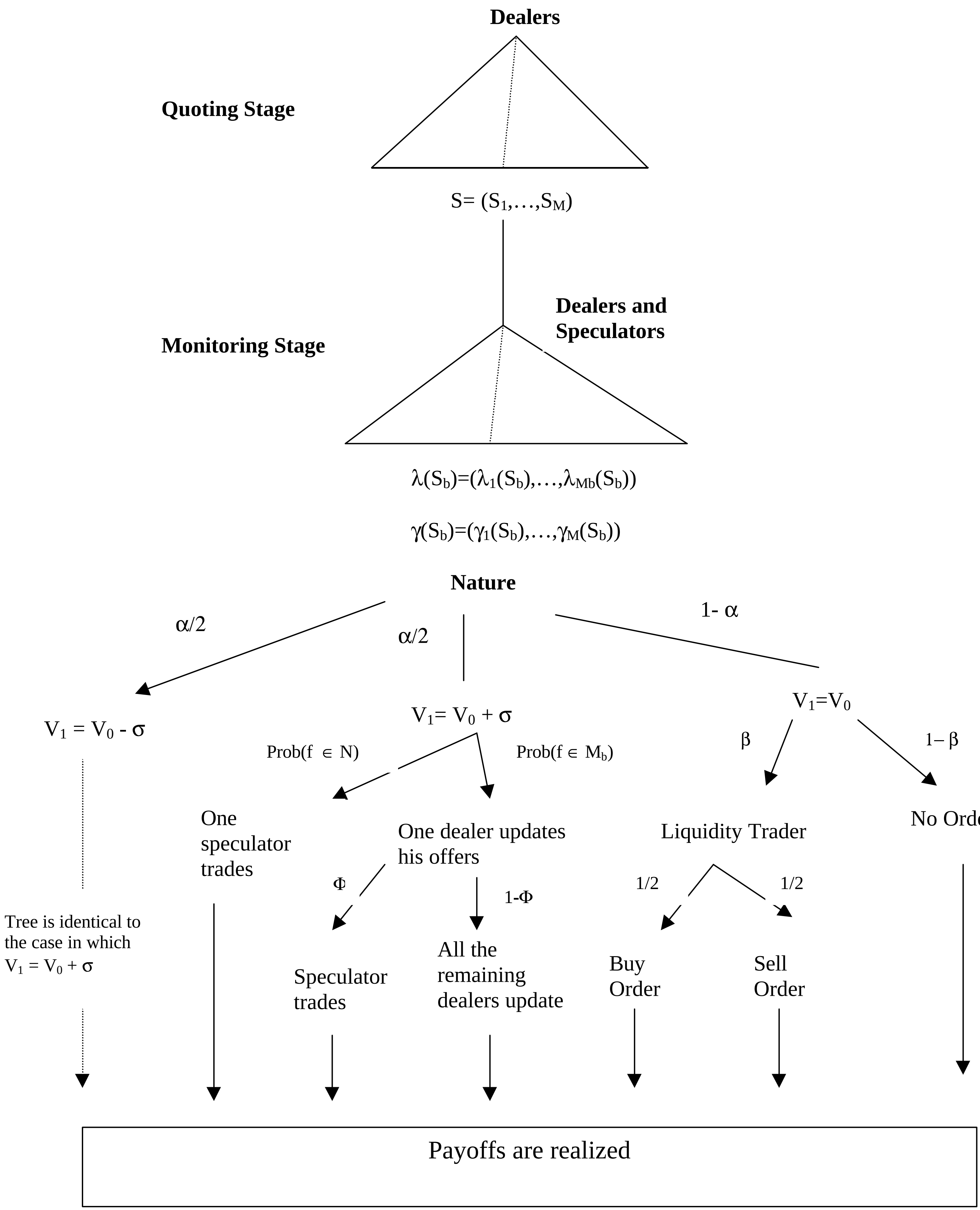

Figure 2: Equilibrium Spreads