"The New Realities of Market Structure and Liquidity: Where Have We Been? Where Are We Going?"[1]

By Chester Spatt[2]

May 21, 2016

1. Introduction

Market structure has change dramatically in the last decade and we are likely to see continuing change to the financial system in light of changes to both technology and the regulatory environment. Regulation in trading is inherent due to underlying externalities, such as the liquidity externality in trading (the posting of orders and the broader availability of liquidity in a platform attracts more liquidity and activity to that platform).

Besides the liquidity externality the potential need and importance of regulation emerges due to the agency relationship in brokerage and the importance of delegated decision-making in trading. Currently, there are about 60 platforms that trade equities in the United States and there is no longer a dominant platform. The structure of equity trading has changed dramatically over the last decade, moving to electronic markets and away from a dominant market that was manually oriented. This leads to the pair of questions that I highlight in the sub-title of my paper: where have we been? and where are we going? My focus on equity trading, rather than bond trading or

[1] An earlier version of this paper was prepared for presentation at the Federal Reserve Bank of Atlanta's May 2016 Financial Markets Conference, "Getting a Grip on Liquidity: Markets, Institutions and Central Banks."

[2] Tepper School of Business, Carnegie Mellon University and National Bureau of Economic Research.

trading in other markets, reflects the substantial changes to our equity markets a decade ago and the much greater transparency of the equity markets, making knowledge of these markets more readily available and apparent. Of course, much of what we can learn from the equity markets is potentially relevant for understanding liquidity in other market contexts. Because the equity markets provide permanent capital, these are arguably especially important to capital formation.

In Section 2 I'll offer some perspective on the evolution of our equity market structure through the lens of Regulation NMS (National Market System), which had been adopted by the Securities and Exchange Commission in 2005 (and fully effective in 2007), highlighting the evolution of competition and fragmentation in these markets. Then I'll turn to the more microeconomic aspects of the trading process in Section 3 to emphasize the routing of orders to platforms and the role of incentives offered by various platforms under the "maker-taker" and "taker-maker" pricing models. I plan to offer perspectives on the changes to the speed of our markets and high-frequency trading and potential manipulation of the markets in Section 4. One of the core features of policy-making for our equity markets in recent years is the central role of pilot empirical analyses, which is discussed in Section 5. As we conclude in Section 6, I offer some perspective on the evolution of liquidity in the bond market.

2. Regulation NMS

The nature of competition in security market trading is ambiguous. On the one hand, competition for liquidity at a point in time is the competition for liquidity to face individual orders (better pricing for the customer who is engaged in trading); on the other hand, another

crucial aspect of competition is the competition among platforms (the intermediaries operating trading businesses) in which innovation plays a central role. One way to frame this tension is that between a "CLOB" (central limit order book) vs. fragmentation. I view Regulation NMS as largely promoting fragmentation, though with limited elements of a central limit order book as well. In particular, the "Order Protection" or "Trade-through" Rule (Rule 611) protects orders at the top (bottom) of the book of each platform, requiring that those orders be filled prior to execution at inferior prices on other platforms (of course, this does not mandate that the orders on competing platforms be filled, but instead that the pricing established by these must be respected to avoid a "trade-through").[3] However, NMS does not provide for order protection going down the book. In a sense NMS integrates the order books at the top of the book, but not away from the top (down the book). Of course, there is a modest element of integration and centralization in this, but fundamentally NMS focuses upon competition among platforms (the order protection rule provides for a limited degree of integration).

This raises the question how did Regulation NMS promote fragmentation? I would emphasize first a number of empirical observations. In the aftermath of Regulation NMS we observed the end of the specialist system on the New York Stock Exchange (NYSE), accompanied by a decline in its market share from 80% to 20% on stocks for which it was the "listing" exchange. We also observed dramatic proliferation in trading venues with a huge increase in the number of platforms to about 60. One key feature of NMS is that "order protection" would only be available to "fast markets," but participating in the benefits of order protection was viewed as essential to being able to attract order flow in the new framework. In this sense NMS led to the

[3] An overview of Regulation NMS is given by Securities and Exchange Commission (2015).

demise of the specialist system because of the inability of a manual market (or one with dominant manual elements) to become a "fast market." In the pre-NMS world the dominance of the NYSE reflected the liquidity externality—it was an attractive location to place orders because of the extent of its activity (orders attract orders, trades attract trades). NMS provided a way for other platforms to attract market share (since the various platforms would have the orders at the top of their respective books protected), leading to more competition for the NYSE and the decline of its once dominant trading venue.

The structure of Regulation NMS by protecting and to a degree rewarding the top of the respective order books provided a direct regulatory incentive that promoted the proliferation of trading platforms (see Spatt (2014)). This protection of the top of the book of different platforms does not protect the best set of prices in the overall market available for a given quantity of shares (quotes below the best set of prices on each platform aren't protected) and in that sense it appears inconsistent; instead it just protects the best individual prices that each respective platform offered (so splitting a platform into components in which each obtained protection at different prices would be beneficial). In contrast, if the regulatory structure protected all the way down the book, then the degree of protection would not be enhanced by splitting up some platforms (the protection would be determined solely by the overall supply curve), so there would not be a direct incentive induced by the regulatory structure for additional platforms to enter. In addition to the direct regulatory incentive that encourages proliferation of platforms, the structure of the order protection rule requires platforms to access better prices available at the other platforms prior to filling on one's own platform; consequently on larger orders there is a focus on filling in small pieces across many platforms. These fragmented fills and the related

focus on filling the next piece of the overall execution are further manifestations of how NMS promoted fragmentation. This also illustrates that NMS is highly prescriptive in mandating how executions must occur; indeed, the rise of trading in dark pools may in part be a response to the highly prescriptive nature of NMS and to avoid the import of other features of it. Indeed, one of the concerns in Michael Lewis's book, *Flash Boys*, is that once executions start to occur in response to a particular order or set of orders that traders respond to the initial fill by backing off (widening spreads); this is an important reason why the institutional "buy-side" might prefer a less fragmented system in which the investor or his broker could more directly manage the overall execution.

When the SEC formulated its re-proposal of Regulation NMS at the end of 2004 it included an alternative in which prices would be protected all the way down the book. However, there was very strong industry opposition to that approach due to the complexity and costs of the implementation, including technological challenges. For example, the NYSE, though somewhat surprisingly a supporter of NMS (I presume because they feared the alternatives), was not sympathetic to protecting the full book.

Finally, I think it is helpful to reflect on the relationship between Regulation NMS and Best Execution responsibilities. I do so in part because some of the decision-makers at the time of the adoption of NMS were motivated by concerns about execution quality (as in the form of "trade-throughs"). The SEC has had a long-standing requirement requiring the broker-dealers obtain "best execution" on behalf of their customers. Note that Best Execution is a responsibility of the

broker-dealer rather than the trading platforms. Reg NMS transfers some mechanics of order routing to the platforms via NMS linkages. Of course, Best Execution is much more germane when there is a serious "routing" decision. To some degree the platforms and broker-dealers might be viewed as at least partial substitutes, but despite this there has been a fair amount of debate about Best Execution in recent years as routing can be distorted by incentive payments to the broker-dealer, such as embedded in "make-or-take" pricing.

3. 'Make-take' or "Take-make" Pricing: Equilibrium and Incentives

The nature of pricing by trading platforms has received considerable attention over the years. Many platforms offer rebates to attract certain orders and under some conditions charge fees on other orders. The array of pricing models raises some important issues about the nature of the equilibrium. For example, how does the structure of fees and rebates relate to which markets offer the fastest and most favorable executions, as well as what are the incentive of brokers routing orders to platforms. The "maker-taker" model involves subsidies (rebates) to the "maker" of a transaction (the side that provides the limit order) and charges fees to the "taker" (the side "taking" liquidity via a market order). The underlying motivation of this approach is to encourage market participants to provide liquidity (limit orders) rather than to consume it. In recent years the "maker-taker" approach has been reversed by some platforms, which instead follow the "taker-maker" model under which the "taker" (market order) receives rebates and the "maker" (limit order) pays fees (this framework is sometimes referred to as an "inverted" model).

This latter model bears some similarities to the "payment for order flow" framework from the 1990s. Like the taker-maker model, the payment for order flow framework involved rebates to those brokers providing market orders. In the case of the payment for order flow model the broker would attempt to purchase relatively uninformed orders (e.g., screening characteristics such as the broader activity in the stock, the size of the order and screening out informationally informed orders by not accepting program trades or orders on deal stocks, for example etc.) rather than paying for all market orders, as in the taker-maker approach. In the various models the rebate is often received by the *broker* and the fees are paid by the *broker*. These payments and rebates change the effective tick size as they are typically a fraction of a tick. Because the rebates and payments are received by the broker rather than the investor, these also raise the potential of an agency conflict leading to distorted incentives. It also is important to recognize that the investor often is unaware of the payment or does not appreciate its significance, such as the possible indirect impact on the quality of his execution. Disclosures pointing to or resolving the agency problem either are not made or they are not internalized by the customer.

The connection between the "make-take" approach and Regulation NMS is potentially significant. Because the "maker-taker" model predates Regulation NMS and prior to NMS the brokers were allowed to route orders to platforms that offered rebates to the brokers, it would not be accurate to suggest that Reg NMS was first to allow the broker to receive rebates from routing market orders (indeed, that even arose under the "payment for order flow" framework). However, NMS capped the permissible fee that could be imposed upon brokers and customers when the linkages are utilized at $.003/share in light of the order protection provided under Reg

NMS.[4] The order protection rule requires protection of orders at the top of the book without adjusting for fees, provided that the fees do not exceed $.003/share. The situation is somewhat analogous to being forced to accept the "best" price on E-Bay, but not considering the shipping fees in "ranking" the costs—up to a threshold on the shipping fees. This leads to distortions among firms with different models of handling the costs of shipping. For example, the ranking is not based upon the "net price" after adjusting for the shipping fees, but instead the "gross price" without fully adjusting.

Absent frictions (including the absence of agency conflicts so that any fees are paid by the customer and rebates are received by the customer) and regulatory impediments the maker-taker and taker-maker models produce equivalent net trading costs. Analogously, in the presence of frictionless monetary transfers between the two sides of a market whether buyers or sellers are taxed is irrelevant (and similarly, whether makers or takers are taxed is irrelevant). In effect, only net trading costs matter without frictions. Of course, this "neutrality theorem" can fail in the presence of various frictions such as transaction costs, fixed costs, etc. For concreteness, consider different platforms re-selling sports tickets, where the platforms employ different pricing models. For example, imagine hypothetically that one platform charges the buyer, while another charges the seller (or alternatively, one platform charges the maker and the other charges the taker—as the potential seller posts limit orders on these ticket platforms). Of course, the nominal/notional prices would differ in the two situations by the differential fees. (In a sense this is somewhat like the Modigliani-Miller irrelevancy theorem for capital structure.) This is

[4] This restriction together with the economic relationship between the allowed fees and rebates helps determine the prevailing rebates.

analogous to the idea that in some market contexts it does not matter whether buyer or sellers are assessed a tax; under certain conditions the essence and incidence of the tax is identical.

What the neutrality theorem highlights is the significant potential effect of various frictions. For example, if an agency conflict were present (so fees are paid by the broker and the broker is collecting the rebates), then the neutrality theorem would fail. The neutrality characterization points to the limitation of claims that the maker-taker model encourages liquidity provision, because of the rebates being paid to those providing liquidity through limit orders. Initially, we will assume that there is no agency problem so that the fees and rebates flow back to the customer.

Not all platforms are equivalent just because the notional (nominal) prices are the same. Indeed, a platform is more attractive if it provides relatively quicker execution for limit orders at the same price. Speed (faster execution) is significant because the underlying order would be much less exposed to adverse selection the faster it would fill (faster execution implies that the order would be less exposed to execution in more adverse states of the world). This raises the question as to which platform will first receive the orders on the opposite side of the market—in particular, the market that pays rebates (and particularly the highest rebates) to the other side will execute first at a price level because it is more attractive to the counterparty. This leads to predictions about the equilibrium routing of orders across platforms (ignoring agency) under an NMS style regime in which the notional price must be respected. We can view this as a special case of a two-sided market (see Rochet and Tirole (2003)) in which there are strong

complementarities between the two sides of the market as each side contributes to the surplus of the other.

Introducing the agency distortion under which the broker pays the make-take and take-make fees and receives the corresponding rebates would lead to distortions in the routing practices of the broker-dealer because these cash flows would go to the broker, while the conventional pricing would flow through to the customer (the distinct buckets for the broker vs. the customer lead to the agency problem in routing of orders). Empirical evidence points to routing to platforms that offer poor/slow execution has emerged as a byproduct of the payment of rebates to brokers (see Battalio, Corwin and Jennings (2015)). We would expect theoretically that platforms that offer high rebates finance these by high fees on the opposite side of the market. If the broker obtains the rebates they would be anticipated to first route to the platforms that offer high rebates (and charge high fees on the opposite side), but these would be least attractive on the opposite side so the potential execution would be worse.

This leaves open the question of how can we solve the agency problem (see related discussion in Angel, Harris and Spatt (2011, 2015)). At a high level, we can try to impose a coercive solution to eliminate the make-take problem. One approach to do so would be an outright ban on make-take pricing, e.g., under NMS not allow any fees to be included under the umbrella of NMS order protection or perhaps allow a nominal amount of fees that might be reflective of the underlying economic costs (e.g., such as a nominal fee of two basis points). This changes the effective "tick." Indeed, Chao, Yao and Ye (2015) argue that the effective tick is reduced by the

make-take pricing structure. In effect, in a setting with discrete ticks Chao, Yao and Ye (2015) argue that "make-take" reduces frictions by reducing the effective tick size.[5] Alternatively, another solution to the routing conflict with make-take (or take-make) pricing would be to ban the broker from using a side pocket, so all rebates and fees would flow directly back to the customer. Since the customer would then be the marginal beneficiary of the fees and rebates as well as the execution costs, this would eliminate the agency conflict—at least conceptually. However, many market participants cite a practical problem with the customer being credited the rebates and fees, which is that these may not be fully known at the time of the implementation of the transaction because the rate of these fees or payments might reflect the overall volume on the platform for a longer period, such as a month (one obvious exception is that if unit fees and rebates were constant over the period and even perhaps required to be constant, though for economic reasons it could be reasonable to allow volume discounts). On the other hand, the unknown nature of the fees and rebates would seem to reinforce the significance of the agency conflict.

An alternative approach to resolving the agency problem is by disclosure. Then the contract between the broker-dealer and his client can reflect directly the distortion in the routing decision. For example, if the rebate to the broker or the fee received by the broker can be conditioned on in the commission, then the consequences would flow through to the client.[6] In principle this information could be disclosed through the "confirmation slip" sent by the broker, though some

[5] However, this may not be a compelling rationale for permitting "maker-taker" pricing as a tighter trading grid and the resulting benefits could be obtained by direct regulation of the permissible tick size.
[6] The prior discussion suggests that these may not be as yet fully known.

clients (such as many retail clients) would not understand the import of the disclosure.[7] A second alternative approach to disclosure would arise by requiring the brokerage firm to disclose information about the performance of its executions at various platforms and its order routing algorithm, so the client could adjust for the expected costs associated with the broker's routing choice.[8]

Whether or not the agency problem can be resolved contractually, we would expect theoretically that competition among brokers would limit the brokers to a competitive return and the adverse consequence of the rating agency distortion would be borne by the client. This is analogous to the agent receiving his reservation utility in the generic agency problem in satisfying the "individual rationality" constraint at equality and the principal bearing the agency distortion. Despite potential frictions associated with the agency conflict, the rebates received by brokers indirectly flowed through to the clients and indeed, commissions have been surprisingly low as the brokers compete for customers and rebate opportunities.

The overall discussion of the maker-taker framework raises a variety of questions about the agency conflict. Can we quantify the importance of the agency conflict and distortion in practice? How should policy be altered to mitigate routing distortions, such as a ban on make-take pricing, a ban on side pockets (by directing rebates to the client?) or enhanced disclosure

[7] The actual confirmation slips disclosure points to the possibility that the broker received compensation.
[8] The related disclosures under Rules 605 and 606 are viewed as complex and not a strong fit along these lines. Additionally, the broad approach is a more complex route to achieve the benefits of disclosure than disclosure at the transaction level.

policies? To what extent are current practices consistent with best execution standards? How can we sort out the empirical consequence of these pricing regimes by a potential pilot analysis?[9]

4. Speed and Trading

There has been much focus in recent years on speed in equity trading. In fact, speed is considered so important that some market participants engage in an "arms race," making substantial investments in technology, as orders are prioritized at platforms by the timing of their arrival. This discussion in turn points to the importance of locating near the underlying platform at which the trades would be executed, "co-location," so that one's order reaches the market quickest. The "arms race" and co-location emerge in response to the incentives to obtain a relative advantage via time priority in the competition for intermediary rents. Indeed, this suggests that not only is there is competition for economic rents, but that such rents are present.[10] The nature of the competition suggests that the desired outcome of market participants is to establish the relative priority of their orders, which would require small relative time advantages.

The theme of the value of co-location and differential access is not a new one. Whether monitoring the extent of goods on trade ships returning to Europe from Asia four centuries ago or messages via the Pony Express or telegraph in the 19th century or even practices on the floor of the New York Stock Exchange when the floor was more active illustrate the importance of co-location and differential access. Indeed, trying to capitalize on the value of co-location the

[9] The Securities and Exchange Commission is currently examining the possibility of a pilot analysis of changes to the make-take framework, e.g., SEC Equity Market Structure Advisory Committee Regulation NMS Subcommittee (2016), including the possibility of substantially lower allowed access fees within the NMS framework.

[10] Analogously, in some other contexts advertising emerges as part of the competition for rents. Despite the dissipative value of some advertising, relatively few would favor a partial ban on advertising.

NYSE banned cellphones on the floor for many years, enhancing the value of the booths that it rented out around the periphery of the trading floor. Of course, the value of the "time and place advantage" of NYSE floor participants in an earlier era under the specialist system was reflected in the pricing of "NYSE seats" and even in the extent of nepotism in the specialist firms.[11] Of course, the time scale of the differential access is completely different now than in the past (it is currently measured in milliseconds or even microseconds), but the inherent possibility and importance of differential access did not suddenly emerge simply because crucial trading decisions now are made at speeds much faster than "human" decision-making and are highly automated through electronic trading engines. Why should we consider this an "arms race" now, but not in earlier eras? Objectively, trading decisions and responses have become much more rapid over time and are much faster than previously as reflected in a range of timing statistics, the degree of concern about leaving unfilled orders exposed with the trading platforms (resulting in much higher cancellation statistics over time) and the time profile of correlations at very high frequency across related markets.[12]

Much of the attention to speed and the interest of investors in "fast" execution and not leaving orders open in the book (rapid cancellation and high quote/trade ratios) reflect the importance of avoiding staleness in one's quotes and controlling the situations in which one's orders are filled. Cancellations also reflect the nature of our modern interconnected platforms in which executions

[11] Limits to the direct marketability of the specialist franchise and limited formal education requirements compared to some professions (such as doctors and lawyers) promoted the extent of succession of family members, relative to other professions.

[12] See Angel, Harris and Spatt (2011), Angel, Harris and Spatt (2015) and Budish, Cramton and Shim (2015).

of modest size are followed by cancellations as investors and traders fear that the initial fill is just the start of a much larger execution (and so pricing backs off).[13]

It is sometimes suggested that extremely high cancellation rates and quote-to-fill ratios are indicative of an attempt to mislead or even manipulate the market, but such statistics need to be interpreted in light of the specific conditions and strategies of the investor. Manipulation involves the establishment of an artificial price; an important consideration would be the "intent" of the trader, which in many situations can be difficult to establish. At the same time it can be challenging to demonstrate manipulation—after all, it would seem legitimate for market participants not to telegraph their intentions and the extent of their interest (e.g., as they could in establishing or liquidating a position by trading on one side repeatedly in a predictable fashion).

5. Pilot Analyses and Policy

An increasingly important approach in recent years for understanding liquidity issues and enhancing the design of markets is to undertake pilot analyses to assess the impact of potential regulatory changes. In trading contexts there is the possibility of conducting controlled experiments in which a portion of the market is treated, but a control sample is used as well (which thereby facilitates the ability to control for time effects that would emerge in a "before" and "after" analysis). The presence of high frequency trading data from a thoughtfully designed setting facilitates the possibility of an informative statistical analysis. Random assignment rather

[13] Among market microstructure theorists the resulting pricing is referred to as upper (lower)--tail expectations, e.g., see Glosten (1994).

than voluntary assignment to the control and treatment groups and careful consideration of spillover effects would be valuable for enhancing the design of experiments for the evaluation of the liquidity consequences of policy alternatives (also see discussion in Spatt (2015)). The SEC successfully undertook such an approach as part of its efforts to repeal up-tick restrictions on short sales a decade ago. More recently, the SEC is planning to re-evaluate tick size through a controlled experiment and in light of the current concerns about the role of fees and rebates in the equity pricing framework the SEC also has begun to consider the possibility of undertaking a pilot study on that front (e.g., SEC Equity Market Structure Advisory Committee Regulation NMS Subcommittee (2016)). Of course, such methods are relevant to other market structure settings as well, such as the bond market, and indeed the phase-in (roll-out) of moves towards greater transparency have been used there to considerable advantage, though perhaps without as much attention to random assignment.

6. Concluding Comments

In this paper I have focused upon the impact of equity market structure upon trading and market liquidity in the post--Regulation NMS era. The structure of equity trading has changed dramatically during the last decade. We have moved from market architecture with a dominant platform with significant manual elements (including a monopolist market-maker) to a trading system with a large number of electronic platforms that are linked together. In this sense the current architecture is highly fragmented, actually facilitating ease of execution of small investors. Larger executions are more complex to complete in the current context, especially given the prescriptive nature of Regulation NMS. While there are certainly important frictions

and distortions remaining in our system of equity trading, the evolution of our trading system has resulted in substantial improvements in the cost of trading in at least some contexts.

While not the focus of the paper, there also have been substantial changes to the structure of trading in other contexts over the last decade. For example, the bond markets are very different than equity, but these also have emphasized electronic trading to a greater degree as well and have moved towards much greater degrees of trade reporting (post-trade transparency) over time rather than an opaque architecture. Given the diffusion of trading across so many instruments and the limited number of trades in most instruments, the design of the bond markets is very different than equity markets (not as prescriptive, not the potential for linkages across platforms and not the potential for the same type of pre-trade transparency as in equity). Recently, it has become clear that many of the traditional dealers have become much less willing to commit capital to trading than previously (so bonds actually stay in inventory less time), but the empirical evidence about changes in trading costs is not clear-cut, perhaps in part because of the response of hedge funds to fill some of the void. Of course, the regulation of bond trading and the markets themselves are very different than for equity, but are likely to continue to evolve substantially.

References

Angel, J., L. Harris and C. Spatt, 2011, "Equity Trading in the 21st Century," *Quarterly Journal of Finance*, 1, 1-53.

Angel, J., L. Harris and C. Spatt, 2015, "Equity Trading in the 21st Century: An Update," *Quarterly Journal of Finance*, 5.

Battalio, R., S. Corwin and R, Jennings, 2015, "Can Brokers Have it All? On the Relationship between Make-Take Fees and Limit Order Execution Quality," working paper.

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

Chao, Y., C. Yao and M. Ye, 2015, "Tick Size Constraints, Two-Sided Markets and Competition Between Stock Exchanges," working paper.

Glosten, L., 1994, "Is Electronic Limit Order Book Inevitable?" *Journal of Finance*, 49, 1127-1161.

Lewis, M., 2014, *Flash Boys*, W.W. Norton and Company, New York.

Rochet, J. and J. Tirole, 2003, "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association* 1, 990-1029.

SEC Equity Market Structure Advisory Committee Regulation NMS Subcommittee, April 19, 2016, "Framework for Potential Access Fee Pilot."

SEC Division of Trading and Markets, April 30, 2015, "Memorandum on Rule 611 of Regulation NMS."

Spatt, C., February 28, 2014, *Statement for House Subcommittee on Capital Markets and Government Sponsored Enterprises (GSEs)* hearing on *"Equity Market Structure: A Review of SEC Regulation NMS,"* February 28, 2014.

Spatt, C., revised October 1, 2015, "Measurement and Policy Formulation," unpublished manuscript.