To Hop (The Queue) Or Not To Hop (The Queue)

One of the fundamental decisions that every trading algorithm needs to make as it feeds child orders into the market is where to route them. In a previous research note we reported some empirical results demonstrating that, generally, the market has a strong preference for routing aggressive orders to the cheapest venue that displays the best price. In this note, we demonstrate how this market structure can be exploited by a clever algorithm to obtain an all-in performance improvement when posting limit orders by effectively "hopping the queue."

Our previous research note, Inverted-Price Destinations and Smart Order Routing (2011), showed that when the best quote is available on one of the inverted exchanges[1], a disproportionate number of marketable orders are routed to it. For example, when Nasdaq BX (which pays a large 15 mil rebate for taking liquidity) showed the best offer, 35% of orders were routed there—roughly ten times its overall market share. Nasdaq BX is not alone: the same pattern holds at the other inverted exchanges as well. It's obvious why market participants scramble to earn the rebate by taking at inverted venues: the difference in cost between taking at a traditional venue and an inverted venue can be as much as 45 mils—in some cases more than the entire commission paid to the broker for executing the order. But given the apparent cost sensitivity of the market, who pays to post on these inverted venues, and why?

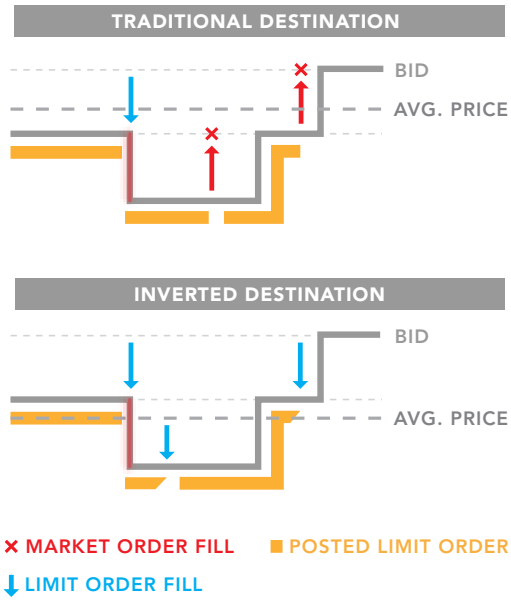

FIGURE 1

COMPARISON OF TWO HYPOTHETICAL DESTINATIONS: TRADITIONAL AND INVERTED.

Suppose our hypothetical algo has to buy several lots every T seconds. We can place a limit at the bid and wait for execution. Figure 1 illustrates this for two hypothetical destinations. If the target quantity has not been executed after T seconds, we are forced to catch up by crossing the spread. This strategy ensures that the best possible price we can get is the current bid. If we post our limit order on an inverted destination, which is more attractive to takers, we are more likely to get executed sooner and thereby to achieve the best price, as show at the bottom of Figure 1. At a conventional destination, we are likely to have to wait longer, and the risk of having to cross the spread is higher, as seen at the top of Figure 1. The average price shortfall is therefore worse.

1 Strictly speaking, an inverted venue is one that pays a rebate for removing liquidity from the book. However, the cost sensitivity of order routing preferences is general—routers generally prefer the cheapest available venue when taking—and in this paper we use the term "inverted" generally to refer to venues that have a significantly lower cost to take than the standard 30 mils/share. As of this writing, the inverted venues are DirectEdge's EDGA, Nasdaq BX and BATS Y.

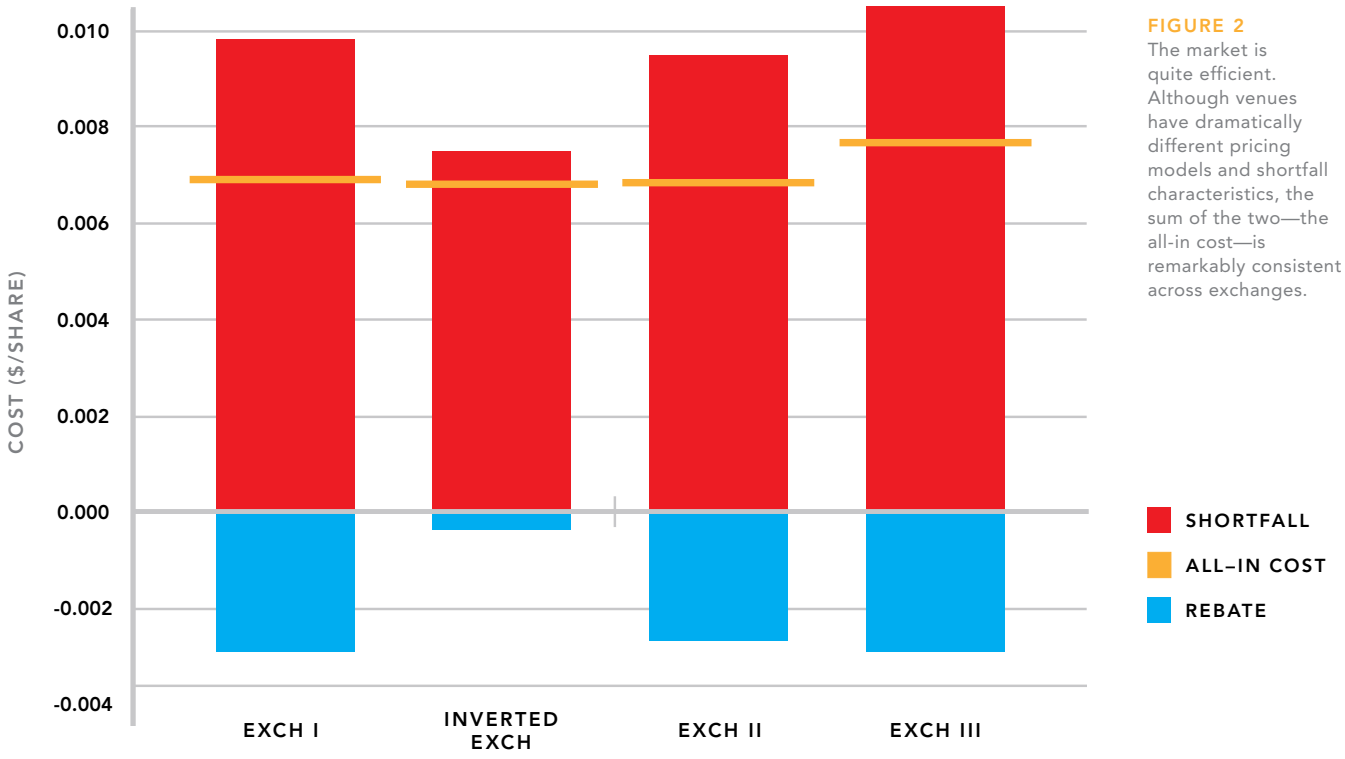

FIGURE 2
The market is quite efficient. Although venues have dramatically different pricing models and shortfall characteristics, the sum of the two—the all-in cost—is remarkably consistent across exchanges.

One answer, in view of the routing preferences described above, may be that posting on an inverted exchange effectively allows one to "hop the queue"—to effectively get execution priority over orders posted at other exchanges that are more expensive for the taker. Theoretically, hopping the queue this way could provide an increased probability of fill before an adverse price move, with a resulting improvement in execution shortfall (see sidebar.)

Thus, we see that the execution cost has two components: the explicit cost (fees and rebates) of executing at a given destination, and the implicit cost paid in the form of price shortfall. Buy-side participants with a pass-through model pay the net of the two. Most institutional clients on the other hand, do not or cannot operate on a pass-through model, and realize only the implicit costs, while their brokers take the explicit cost (or rebate). The real question for pass-through participants, therefore, is whether the performance improvement of hopping the queue actually exceeds the explicit cost one pays to do so. Accepted wisdom is that markets are efficient, and a cursory look at the data suggests there is not much to be gained here. Figure 2 shows empirical results from randomized order routing experiments conducted by Pragma, demonstrating no significant difference in all-in costs among three exchanges with traditional pricing models and one inverted exchange. Based on this data, it appears that there is an opportunity through routing choices to shift the type of costs—from explicit cost to implicit cost—but not to improve the overall result.

However, the efficient market hypothesis notwithstanding, a deeper look reveals that there are inefficiencies that can be exploited by a clever algorithm. Pragma has developed a dynamic pricing model that expresses the value of hopping the queue. This model identifies transient opportunities when the value of hopping the queue exceeds the cost to do so, resulting in an improvement to all-in performance. It also indicates when the benefit of hopping the queue is negligible compared with the foregone provide rebate, so that posting on a traditional destination is optimal.

Figure 3 shows the predictive power of the model. The X axis shows the model's predicted benefit in cents/share for hopping the queue, and the Y axis shows the average realized shortfall in cents/share conditioned on the predicted value, from actual trading results. As the figure shows, the benefit of hopping the queue ranges from a few mils/share—much less than the cost differential of using an inverted venue—to well over half a penny, significantly more than the roughly 25 mil cost differential between the two venues in the figure. The model is able to predict the shortfall savings of hopping the queue quite well.

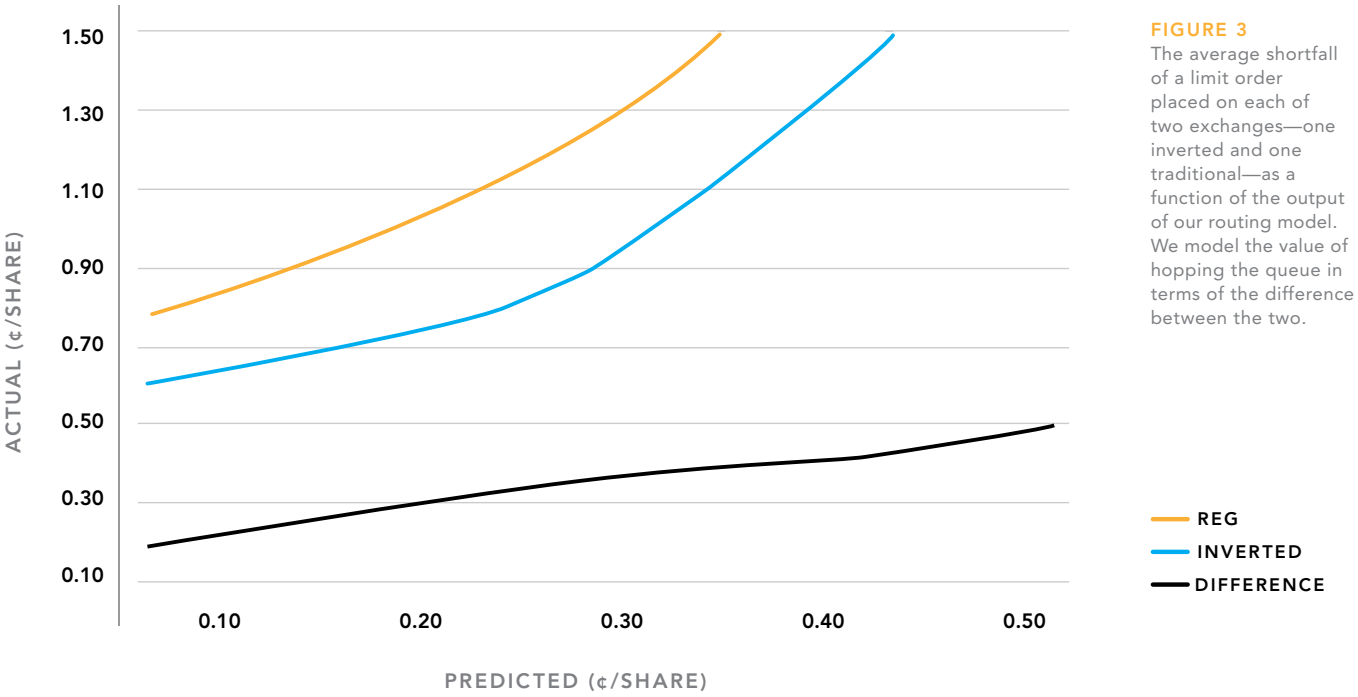

FIGURE 3
The average shortfall of a limit order placed on each of two exchanges—one inverted and one traditional—as a function of the output of our routing model. We model the value of hopping the queue in terms of the difference between the two.

Moreover, using the cost differential as a threshold, the model provides a systematic method for truly smart order routing to maximize all-in performance.

The model makes some interesting and intuitive predictions. The higher the volatility, the higher the incentive to hop the queue, as waiting even a short period of time without a fill might result in a large adverse price move. Similarly, the lower the arrival rate of trades in the market, the higher the benefit of hopping the queue—a queue of a given length takes longer to work through when the arrival rate of opposing orders is low. These effects demonstrate that while easily measured and understood metrics like latency get a lot of attention, mundane decisions like where limit orders are posted can have an enormous effect on execution quality.

In conclusion, the variety of pricing models offered by exchanges, coupled with the cost sensitivity of smart order routers, creates an interesting market structure in which participants can effectively choose their priority in a virtual inter-market priority queue by paying an explicit fee, bypassing the priority rules of the individual venues. This has interesting consequences from a couple of perspectives.

From the perspective of institutional clients who don't and in some cases can't receive pass-through fees and rebates, this research raises the question of what to expect from a broker. Institutional clients should understand how their brokers approach the tradeoff between the client's shortfall and the broker's execution expenses. An algorithmic broker who has no policy and has done no research on this tradeoff may be optimizing their own economics at an unknown cost to the client in execution quality. Similarly, brokers who outsource their algorithms to other brokers also delegate the management of this tradeoff. Our suggestion is for brokers to optimize the all-in performance of the algorithm even when fees and rebates are *not* passed through to the client. Regardless of the approach a broker takes, it should explain its reasoning and routing practices to its clients – transparency should be the keystone of the broker-client relationship.

Finally, from the perspective of the practitioner with a focus on all-in performance, this market structure leads to an opportunity to significantly improve the execution quality of an algorithm beyond that allowed by traditional, cost-sensitive smart order routing systems.

For questions or comments, please email Dr. Eran Fishler, Director of Research (technotes@pragmatrading.com).