Management Science

Execution Risk in High-Frequency Arbitrage

Roman Kozhan, Wing Wah Tham,

To cite this article:
Roman Kozhan, Wing Wah Tham, (2012) Execution Risk in High-Frequency Arbitrage. Management Science 58(11):2131-2149.
http://dx.doi.org/10.1287/mnsc.1120.1541

MANAGEMENT SCIENCE
Vol. 58, No. 11, November 2012, pp. 2131–2149
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)
http://dx.doi.org/10.1287/mnsc.1120.1541

Execution Risk in High-Frequency Arbitrage

Roman Kozhan
Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom, roman.kozhan@wbs.ac.uk

Wing Wah Tham
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam,
3000 DR Rotterdam, The Netherlands, tham@ese.eur.nl

In this paper, we investigate the role of execution risk in high-frequency trading through arbitrage strategies. We show that if rational agents face uncertainty about completing their arbitrage portfolios, then arbitrage is limited even in markets with perfect substitutes and convertibility. Using a simple model, we demonstrate that this risk arises from the crowding effect of competing arbitrageurs entering the same trade and inflicting negative externalities on each other. Our empirical results provide evidence that support the relevance of execution risk in high-frequency arbitrage.

Key words: execution risk; limit to arbitrage; liquidity; high-frequency trading strategies
History: Received March 9, 2011; accepted January 24, 2012, by Wei Xiong, finance. Published online in *Articles in Advance* June 5, 2012.

1. Introduction

Execution risk is fundamental to and important for all investment strategies. Perold (1988) suggests that the failure to accomplish a trade immediately can impose an implicit cost on traders. As a result, supposedly profitable trading strategies can turn within moments into unprofitable ones. This can be especially important when using high-frequency trading strategies. In this paper, we investigate the role of execution risk and the negative externality of congestion on competition in high-frequency trading through arbitrage strategies. High-frequency arbitrage provides an ideal platform to analyze execution risk because it is isolated from any other risks in this setting. Moreover, understanding the role of execution risk in high-frequency trading strategies is of high importance because of increased investor interest and regulatory concerns for this type of trading.

With the recent proliferation of technological advancements in financial markets, algorithmic and high-frequency traders can exploit arbitrage opportunities in a fraction of a millisecond.[1] In these high-speed competitive financial markets, arbitrage opportunities that do not rely on convergence trading, like triangular arbitrage or covered interest parity, should not persist and should be eliminated immediately. Nevertheless, recent empirical literature shows that arbitrage opportunities, such as deviations from triangular arbitrage parity (Marshall et al. 2008) and covered interest parity (Akram et al. 2008, Fong et al. 2010), still prevail in the market. These deviations provide profitable opportunities for algorithmic traders, and their existence and persistence is puzzling. The existing literature on limits to arbitrage does not provide satisfactory explanations, and we attempt to fill this gap.

In this paper, we argue that high-frequency arbitrage opportunities exist because arbitrageurs face uncertainty about completing a profitable arbitrage portfolio due to the crowding effect of competing with others for scarce arbitrage assets. In contrast to the common notion that competition improves price efficiency, we suggest that competition among arbitrageurs limits efficiency because competing arbitrageurs inflict negative externalities on each other. This execution risk in high-frequency arbitrage exploitation depends on the degree of competition among arbitrageurs and the cost of market illiquidity.[2] The intuition is simple: consider two arbitrageurs competing to form a long–short arbitrage

[1] High-frequency trading strategies include cross-asset arbitrage, electronic market making, liquidity detection, short-term statistical arbitrage, and volatility arbitrage (see Tabb et al. 2009). High-frequency trading and these arbitrage strategies are estimated to account for an annual aggregated profit of about 21 billion U.S. dollars according to a report by Tabb et al. (2009). In a study of algorithmic trading in the foreign exchange market, Chaboud et al. (2010) shows that exploiting triangular arbitrage is a commonly used trading strategy by algorithmic traders.

[2] The execution risk in arbitrage exploitation discussed in this paper is different from execution risk in trading. Execution risk in trading is often related to the timing uncertainty about when limit orders will execute, whereas execution risk in arbitrage exploitation is concerned about the uncertainty in executing a profitable arbitrage strategy.

portfolio consisting of two identical but mispriced assets: asset A with two available units and asset B with only one available unit at a profitable price. Exploiting this arbitrage can be risky because one of the arbitrageurs will fail to acquire/short both assets A and B at a profitable price and might incur liquidity costs (walking up the limit order book (LOB)) or inventory costs (unwanted inventory from successfully acquiring asset A but unsuccessfully shorting asset B). The mispricing might remain because competing arbitrageurs might choose not to participate in exploiting this mispricing because of the execution risk. We provide both theoretical and empirical support for this new limit, which is applicable to all asset markets. In contrast to other existing limits to arbitrage, the mechanism we present does not rely on convergence trading, taxes, and regulatory and short-selling constraints. This risk is not only important in explaining the persistence and in exploiting deviations of high-frequency arbitrage opportunities, but also important for any high-frequency investment and trading strategies.

We empirically test the role of execution risk in arbitrage using a set of reliable and detailed limit order book data from a widely traded and liquid electronic trading platform of the spot foreign exchange (FX) market. To isolate execution risk from other existing impediments to arbitrage, our study focuses on triangular arbitrage exploitation in major currency pairs. The main advantage of triangular arbitrage is that it is a non-convergence-trading based arbitrage and is not subject to taxes and regulatory or short-selling constraints.

We find that even after accounting for brokerage fees, bid–ask spread (direct transaction cost), and latency costs, triangular arbitrage opportunities remain in the FX market and these opportunities are not exploited instantly. This finding provides initial support for the existence of execution risk in high-frequency-trading-based arbitrage. We also examine the relationship between the size of arbitrage deviations and market illiquidity. We find that arbitrage deviations increase with market illiquidity, which supports the conclusions of Roll et al. (2007), who argue that market liquidity plays a key role in moving prices to eliminate arbitrage opportunities. We argue that the economic reason behind this relationship could be the presence of execution risk.

In addition, we reinforce the relevance of execution risk by studying the sources and characteristics of submitted orders surrounding arbitrage eliminations. We find that arbitrage opportunities are more likely to end with market orders when arbitrage deviation is large, market liquidity is high, and the arbitrage duration is short. Our finding is consistent with the suggestion that arbitrageurs are more likely to participate in arbitrage exploitation with market orders when the level of execution risk is low. We also perform an economic evaluation of arbitrage strategies and find that losses due to execution risk worsen as the number of competing arbitrageurs increases.

The impact of competition and market illiquidity on the existence of convergence-trading arbitrage opportunities has been studied recently by, among others, Stein (2009), Kondor (2009), and Oehmke (2009). However, these papers focus on convergence trading and the use of wealth constraints that cause arbitrageurs to unwind their positions as prices diverge. The main difference between our model and those of Oehmke (2009) and Kondor (2009) is the effect of crowded trades on the elimination of arbitrage. Efficiency increases with the number of arbitrageurs in their models, although excess competition for the scarce supply of arbitrage assets prevents immediate elimination of mispricing in our model.

The crowding effect in our model is similar to that of Stein (2009), who proposes a limit to arbitrage that stems from the uncertainty about the number of arbitrageurs entering the same trade. His mechanism relies on arbitrage strategies with no fundamental anchor, in contrast to standard models of arbitrage where arbitrageurs know the fundamental value of the asset.[3] We differ from Stein (2009) by showing that overcrowding of arbitrageurs can be risky even if arbitrageurs know the fundamental value of assets. We argue that overcrowding of arbitrageurs competing for a limited supply of arbitrage assets can incur losses because profits can no longer be shared among arbitrageurs in the case of imperfectly divisible assets.[4]

Kleidon (1992), Kumar and Seppi (1994), and Holden (1995) highlight the importance of execution risk in index arbitrage under stressful market conditions (crash of October 1987), where execution risk can be a concern due to trading on stale prices. Their mechanism is different from ours because the "paper environment" of the New York Stock Exchange in 1987 and the software inadequacy to cope with the cancelation or replacement of limit orders, which contribute to stale orders and prices, make exploitation

[3] Trading strategies with no fundamental anchor imply that arbitrageurs do not observe the fundamental value and do not base their demands on an independent estimate of the fundamental value.

[4] It is often observed in practice that financial assets are imperfectly divisible because of minimum trade size restrictions. By accounting for the restriction of minimum trade size in financial markets, there will be insufficient units of assets for all arbitrageurs to have a positive profit. Thus, competition among arbitrageurs for these imperfectly divisible assets creates excess demand for these assets. Under this assumption, it is not guaranteed that all arbitrageurs will always be able to make money.

of index arbitrage risky. In a similar vein, Kamara and Miller (1995) suggest the importance of immediacy (liquidity) risk where arbitrageurs bear the risk of adverse price movements from order submission until order execution. We account for the role of stale prices by controlling for data latency in our empirical exercise.

The current literature on limits to arbitrage focuses on the role of noise trader risk, fundamental risk, synchronization risk, taxes, and regulatory and short-selling constraints. Violations of arbitrage conditions due to these frictions normally last for days and even months.[5] Differently from the current literature, we focus on the limits to high-frequency arbitrage where execution risk plays an important role. With an extremely detailed and reliable limit order and transaction data set, we believe our paper is among the first to provide a high-quality study on high-frequency arbitrage and execution risk. Our paper relates to the fast-growing literature on speed of trading and algorithmic trading. Easley et al. (2009) and Hendershott and Moulton (2011) examine how changes in technology that improve the latency of information transmission and execution affect market quality. Hendershott et al. (2011), Chaboud et al. (2010), and Brogaard (2010) examine the impact of algorithmic and high-frequency trading on market quality. We contribute to this literature by studying a potential risk related to high-frequency trading strategies. More importantly, our paper provides an alternative perspective on how competition might impede rather than improve efficiency. This has important implications for how we think about competition and efficiency in the modeling of financial markets. We also provide an alternative liquidity-based theory for impediments to arbitrage, which supports empirical papers relating high-frequency arbitrage deviations to market illiquidity. From the policy perspective, our work has important implications for regulators who are seeking to better understand mechanisms and characteristics of high-frequency arbitrage trading strategies.[6]

More generally, our paper adds to the literature on the stochastic arrival problem (Lariviere and Mieghem 2004) and congestion pricing (Naor 1969, Mendelson 1985, Dewan and Mendelson 1990). This literature is concerned about how an economic agent strategically chooses when to seek a service while accounting for the impact of congestion externalities on the level of service and price, given of the service provider. Our paper adds to this literature by studying how arbitrageurs select to participate in arbitrage exploitation in the presence of negative congestion externalities in financial markets.[7]

2. A Simple Model of Execution Risk in Arbitrage

In this section, we introduce a simple model to illustrate the idea behind execution risk. We consider a setup, where there are I assets indexed by $i \in \mathcal{I} \equiv \{1, \ldots, I\}$ that are traded in I segmented markets. We assume there exists a portfolio, RP, consisting of all assets from the set $\{2, \ldots, I\}$, which has a payoff structure and dividend stream identical to asset 1. For simplicity, this portfolio includes long and short positions of one unit in each asset denoted by the vector $[w_2, \ldots, w_I]$; w_i takes the value of 1 if it is a long position and -1 if it is a short position in asset i. We assume that there are no short-sale constraints in the market. There is also perfect convertibility between asset 1 and portfolio RP, which is defined as the ability to convert one unit of asset 1 to one unit of portfolio RP. With perfect convertibility, traditional impediments to arbitrage like fundamental risk, noise trader risk, and synchronization risk are absent in our setup.

Following Kondor (2009), there are I groups of local traders, who only trade assets in their own segmented markets for exogenous reasons. Liquidity is offered by these traders in the LOB through posting orders. Asymmetric demands and income shocks to these local traders may cause transient differences in the demand for assets in each market. This captures the idea that similar assets can be traded at different prices until arbitrageurs eliminate the mispricing.

In addition to the local traders, there are k competing risk-neutral arbitrageurs, who can trade across all markets and exploit any existing mispricing.[8] All exploitations are conducted via simultaneous sales and purchases of identical assets with no outlay of personal endowment. Arbitrageurs use market orders to ensure the simultaneity of sales and purchases of mispriced assets. For simplicity, we assume all

[5] See the survey in Gromb and Vayanos (2010) for additional references.

[6] The U.S. Securities and Exchange Commission (SEC) and Committee of European Securities Regulators (CESR) take particular interest in high-frequency trading strategies and call for evidence on microstructure issues in the U.S. and European equity markets (see SEC 2010, CESR 2010).

[7] Other works on congestion externality and competition with stochastic demand include Martínez-de-Albéniz and Talluri (2011), Hosanagar et al. (2008), and Krishnan and Winter (2010), among others.

[8] Allowing for strategic liquidity traders or differential ability, speed of trading and risk aversion of arbitrageurs would generate interesting insights but does not add to the overall understanding of the main issues examined in this paper.

arbitrageurs can only buy one unit of each asset needed to form the arbitrage portfolio.[9]

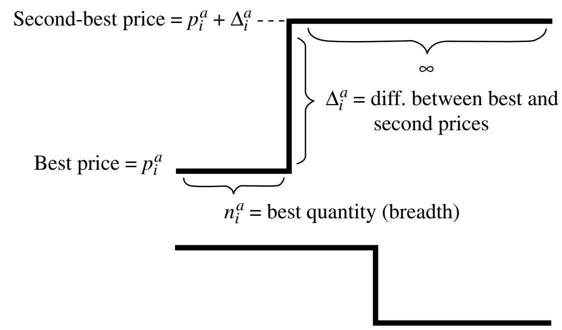
All participants are able to see all prices and depths of the demand and supply schedules of the LOB.[10] There is no cost in posting, in retracting, or in altering any limit orders at any time except in the middle of a trade execution. Many financial exchanges impose a minimum trade size restriction, which implies that traded assets are treated as imperfectly divisible goods in practice. As such, all prices are placed in a discrete grid in the LOB, and the minimum trade size is one unit.

We assume there are only two layers in our discrete demand and supply schedules. The first layer consists of the best bid and ask prices and the quantities available at this prices. The best bid and ask prices of asset i are denoted by p_i^b and p_i^a, respectively. The corresponding quantities available at the best bid and ask prices of asset i are denoted by n_i^b and n_i^a. The next best available bid (ask) price of the asset is $p_i^b - \Delta_i^b$ ($p_i^a + \Delta_i^a$) at the second layer. As a simplifying assumption, prices of all assets at the second layer are assumed to be available with infinite supply.[11] The modeled structure of the LOB can be visualized in Figure 1.

The best price at which one unit of portfolio RP can then be bought is $P^a = \sum_{i=2}^{I} p_i^a$. The best price at which one unit of portfolio RP can be sold is P^b. Because portfolio RP and asset 1 have identical payoff structures, dividend streams, and risk exposure, they should have the same price. Taking transaction costs into account, a mispricing occurs if $P^a < p_1^b$ or $P^b > p_1^a$ and it can be exploited by arbitrageurs. We define the magnitude of the mispricing then as $A = \max\{0, P^b - p_1^a, p_1^b - P^a\}$.

Professional arbitrageurs frequently compete against each other in exploiting any observable arbitrage opportunities in financial markets. With limited and scarce supply of the required assets available to form an arbitrage portfolio, we assume that there exists an excess demand for these assets among competing arbitrageurs such that $\max\{n_i^a, n_i^b\} < k$ for each $i \in \mathcal{I}$. Because arbitrageurs can only purchase one unit of each required asset, we assume that there are always more arbitrageurs than the maximum number of available units of assets. Arbitrageurs prefer market orders over limit orders because of the advantage of immediacy. With the enormous technological advances in trading tools over recent years, algorithmic trading is widely used in exploiting arbitrage opportunities. These algorithmic trades lead to almost simultaneous exploitation of arbitrage opportunities by large numbers of professional arbitrageurs in the financial market. Thus, arbitrageurs who want to trade upon observing any mispricing are assumed to submit their market orders simultaneously.[12]

Figure 1 **Structure of Limit Order Book**



Notes. This figure represents the structure of the limit order book in the market i. There are only two layers in the discrete demand and supply schedules. The first layer consists of the best bid and ask prices and the quantity available at these prices; p_i^b and p_i^a are the best bid and ask prices of asset i, respectively; n_i^b and n_i^a denote the corresponding quantities available at the best bid and ask prices, respectively. The next best available bid price of the asset is $p_i^b - \Delta_i^b$, and the next best ask price is $p_i^a + \Delta_i^a$ at the second layer. We assume that prices of all assets at the second layer are available in infinite supply.

In our setup, all arbitrageurs have the same probability of execution at the best available price when they submit market orders simultaneously. Arbitrageurs who are unsuccessful in acquiring the required asset at the best available price will execute their market orders at the next best available price ($p_i^a + \Delta_i^a$ for buy trades or $p_i^b - \Delta_i^b$ for sell trades). Thus, the penalty for missing a trade at the best price in one of required assets i is $\Delta_i(w_i)$, where $\Delta_i(-1) = \Delta_i^b$, and $\Delta_i(1) = \Delta_i^a$. In this circumstance, the arbitrageur will be left with a payoff of $A - \Delta_i(w_i)$. The worst situation an arbitrageur could face is one in which she fails to acquire all the required assets at the best available price and receive the payoff

$$A - \sum_{i=1}^{I} \Delta_i(w_i) < 0, \tag{1}$$

which we assume to be negative.

2.1. Equilibrium

All arbitrageurs have two possible strategies upon observing an arbitrage opportunity, "to trade" or "not

[9] Allowing arbitrageurs to buy more than one unit of asset only exacerbates execution risk.

[10] In reality, pretrade transparency is worse than our assumption because not every trader can observe the whole limit order book. The impact of execution risk is likely to worsen with deteriorating pretrade transparency.

[11] Incorporating more than two layers in the LOB would only increase the cost of liquidity.

[12] In reality, actions of arbitrageurs need not take place literally at the same moment. Our assumption of simultaneity is valid as long as arbitrageurs have no knowledge of the queue position of their market orders.

to trade." An arbitrageur who chooses not to trade will have a payoff of zero. We also assume that all information, arbitrageurs' strategies, preferences, and beliefs are common knowledge. We allow arbitrageurs to use mixed strategies in their arbitrage strategies, where they participate in the market but with only a probability of exploiting the mispricing. We denote the probability of participation of arbitrageur j by $\pi_j \in [0, 1]$. For a mixed strategy profile $\Pi = (\pi_1, \dots, \pi_k)$, we denote by $\Pi_{-j} = (\pi_1, \dots, \pi_{j-1}, \pi_{j+1}, \dots, \pi_k)$ a strategy profile of all arbitrageurs other than j.

Proposition 1. *For a given mixed strategy profile* $\Pi = (\pi_1, \dots, \pi_k)$:

(i) *the expected payoff of arbitrageur* j *is* $\pi_j E(U^j \mid \Pi_{-j})$, *where*

$$E(U^j \mid \Pi_{-j}) = A - \sum_{i=1}^{I} \Delta_i(w_i)\big(1 - \mathbf{P}^j_{i\mid k,\Pi_{-j}}\big) \tag{2}$$

is the expected payoff of arbitrageur j *playing pure strategy "trade" while her opponents use mixed strategies* Π_{-j}; $\mathbf{P}^j_{i\mid k,\Pi_{-j}}$ *is the probability of getting the best price for asset* i;

(ii) $\mathbf{P}^j_{i\mid k,\Pi_{-j}}$ *and* $\pi_j E(U^j \mid \Pi_{-j})$ *decline monotonically with the number of arbitrageurs* k.

Proof. See Appendix A.

Equation (2) shows that an arbitrageur's expected payoff is the difference between the observed mispricing A and the expected loss due to execution risk. The severity of the losses or the cost of execution failure increases with market illiquidity $\Delta_i(w_i)$. Furthermore, the expected loss increases with the probability $(1 - \mathbf{P}^j_{i\mid k,\Pi_{-j}})$ of not getting the best price for asset i, which, according to Proposition 1, increases with the number of competing arbitrageurs. This demonstrates that competition for scarce supply of assets and market illiquidity exacerbate execution risk when exploiting arbitrage opportunities.

There exists a unique symmetric mixed strategy Nash equilibrium Π of the above game characterized by the common probability of participation $\pi < 1$. The zero-profit condition implies that in equilibrium the observed arbitrage deviation is a linear function of the differences between the best and the second-best prices in the corresponding markets:

$$A = \sum_{i=1}^{I} \Delta_i(w_i)(1 - \mathbf{P}_{i\mid k,\pi}). \tag{3}$$

Equation (3) shows that the magnitude of the arbitrage deviation is associated with the execution risk for each of the I number of assets in an arbitrage portfolio. The total execution risk compensation or the arbitrage deviation can be seen as the sum of individual compensations for execution risk for each individual asset. Each of these individual components depends on the cost of execution failure, $\Delta_i(w_i)$, and the failure probability of executing the best price market orders, $1 - \mathbf{P}_{i\mid k,\pi}$. Thus, the optimal probability of participation π is also a function the arbitrage deviation, the breadth of the asset supply, and the number of existing arbitrageurs. With k number of arbitrageurs, the probability of participation endogenizes the average number of actively participating arbitrageurs to $\pi \times k$. On the other hand, arbitrageurs in our model behave strategically and use the probability of participation, taking into account actions of their opponents, as a tool to strategically respond to negative externalities from the crowding effect. Thus, our arbitrageurs are not mainly concerned with their own price impact of trade (like in Lyons and Moore 2009) but with the price impact of trades of other competing arbitrageurs.

2.2. Market Efficiency and Competition

In this section we study how the probability of arbitrage elimination changes as the number of competing arbitrageurs increases. To eliminate the observed arbitrage deviation, it is necessary to execute in aggregate all available units of at least one of the assets. We denote the minimum breadth of all assets by $\underline{n} = \min_{i\in\mathcal{I}}\{n_i(w_i)\}$. In the mixed strategy equilibrium, where arbitrageurs participate with probability $\pi < 1$, the probability of arbitrage elimination is equal to the probability that $\underline{n}$ or more arbitrageurs out of k decide to trade. This can be expressed using the binomial distribution as

$$\text{ProbElim} = \sum_{s=\underline{n}}^{k} \binom{k}{s} \pi^s (1-\pi)^{k-s} < 1. \tag{4}$$

Equation (4) shows that under competition for scarce supply of the assets required for the arbitrage portfolio, the arbitrage opportunity might remain in the market for some time. In the face of execution risk, arbitrageurs might all decide not to participate with some positive probability. Although we do not model the duration of the mispricing explicitly, we can see from Equation (4) that the probability of immediate arbitrage elimination is smaller than 1.

The following proposition establishes the probability of immediate arbitrage elimination as the number of competing arbitrageurs goes to infinity. For simplicity, we assume that quantities of available units of assets are equal to 1, and illiquidity costs are the same across all markets.

Proposition 2. *Let* $n_i = 1$ *and* $\Delta_i = \Delta$ *for* $i \in \mathcal{I}$, *and let* $A < I\Delta(1 - 1/k)$ (*not all arbitragers use "trade" strategy*). *Then the probability of immediate arbitrage elimination is*

$$\text{ProbElim}(k) = 1 - (1-\pi)^k, \tag{5}$$

Management Science 58(11), pp. 2131–2149, ©2012 INFORMS

where π solves equation $A = I\Delta(1 - \mathbf{P}_{k,\pi})$ with $\mathbf{P}_{k,\pi} = (1-(1-\pi)^k)/(k\pi)$. Moreover, ProbElim$(k)$ *is a decreasing function of k.*

Proof. See Appendix B.

In contrast to the common notion that competition encourages efficiency, Proposition 2 claims that competition might impede efficiency if arbitrageurs inflict negative externalities on each other.

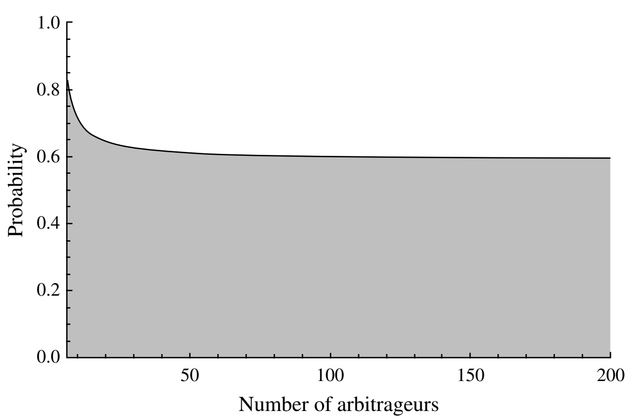
To visualize the result, we compute the probability of arbitrage elimination using sample averages from data as parameters values in Equation (4). The illiquidity costs Δ_i are computed from our sample for three major currency pairs taken from §4: $\Delta_{\text{GBP/USD}} = 2.593$, $\Delta_{\text{EUR/GBP}} = 1.281$, and $\Delta_{\text{EUR/USD}} = 2.856$. We assume the quantity at the best price for all three currency pairs to be $n_{\text{GBP/USD}} = n_{\text{EUR/GBP}} = n_{\text{EUR/USD}} = 3$, and the average arbitrage deviation to be $A = 1.56$. For every value of k (the number of arbitrageurs), the probability of participation π is computed as the solution of equation

$$\begin{aligned} A &= \Delta_{\text{GBP/USD}} \bar{\mathbf{P}}^{\text{GBP/USD}}_{k,\pi} + \Delta_{\text{EUR/GBP}} \bar{\mathbf{P}}^{\text{EUR/GBP}}_{k,\pi} \\ &\quad + \Delta_{\text{EUR/USD}} \bar{\mathbf{P}}^{\text{EUR/USD}}_{k,\pi}, \end{aligned}$$

$$\bar{\mathbf{P}}^i_{k,\pi} = 1 - \mathbf{P}^i_{k,\pi} = \sum_{s=n_i}^{k-1} \binom{k-1}{s} \pi^s (1-\pi)^{k-s-1} \left(1 - \frac{n_i}{s+1}\right),$$

where the last expression is derived from Equation (A1) in Appendix A. Figure 2 shows that the probability of arbitrage elimination is smaller than 1 and declines monotonically as the number of participating arbitrageurs increases.

Figure 2 Probability of Arbitrage Elimination



Notes. The figure presents the probability that the observed arbitrage opportunity will be eliminated instantly (horizonal axis) as a function of number of arbitrageurs k. This probability is computed as $\text{ProbElim}(k) = \sum_{s=\underline{n}}^{k} \binom{k}{s} \cdot \pi^s (1-\pi)^{k-s}$, where π is the solution of equation $A = \Delta_{\text{GBP/USD}} \bar{\mathbf{P}}^{\text{GBP/USD}}_{\pi,k} + \Delta_{\text{EUR/GBP}} \bar{\mathbf{P}}^{\text{EUR/GBP}}_{\pi,k} + \Delta_{\text{EUR/USD}} \bar{\mathbf{P}}^{\text{EUR/USD}}_{\pi,k}$, with $\bar{\mathbf{P}}^i_{\pi,k} = \sum_{s=n_i}^{k-1} \binom{k-1}{s} \pi^s (1-\pi)^{k-s-1} \cdot (1 - n_i/(s+1))$. We consider sample average of parameter values: $A = 1.56$, $\Delta_{\text{GBP/USD}} = 2.593$, $\Delta_{\text{EUR/GBP}} = 1.281$, $\Delta_{\text{EUR/USD}} = 2.856$, and $n_{\text{GBP/USD}} = n_{\text{EUR/GBP}} = n_{\text{EUR/USD}} = \underline{n} = 3$.

2.3. Main Implications

Our model has several implications. In the presence of competition, an arbitrageur faces execution risk in acquiring her arbitrage portfolio at the best price, and the magnitude of this risk depends on the illiquidity costs and the level of the competition. Equation (3) implies that arbitrageurs demand a compensation for execution risk, and the demanded compensation increases with market illiquidity.

Facing execution risk, arbitrageurs participate in exploiting arbitrage only with probability less than 1, which implies that the mispricing might not be exploited immediately (see Equation (4)). The probability of participation depends on the profitability of the arbitrage strategy and market illiquidity.

Finally, execution risk in arbitrage worsens with an increasing number of competing arbitrageurs. This is because the failure probability of acquiring the arbitrage portfolio at a profitable price increases with the number of competing arbitrageurs. Increasing competition for limited numbers of exploitable arbitrage opportunities prevents efficient elimination of asset mispricing.

3. Triangular Arbitrage in the FX Market

We test the relevance of execution risk in high-frequency trading by examining the triangular arbitrage parity in the FX market. Triangular arbitrage involves one exchange rate traded at two different prices, a direct price and an indirect price (vis-à-vis other currencies). Arbitrageurs are often assumed to eliminate any price discrepancy if the inferred cross-rate between currencies A and B is known through the two currencies' quotes vis-à-vis the third currency C. Let us denote by $P^b(A/B)$ and $P^a(A/B)$ bid and ask prices of one unit currency B in terms currency A, respectively. Taking transaction costs into account, the triangular no-arbitrage conditions are

$$\begin{aligned} &P^a(A/B) \ge P^b(A/C) \times P^b(C/B) \quad \text{and} \\ &P^b(A/B) \le P^a(A/C) \times P^a(C/B). \end{aligned} \tag{6}$$

Any violation of inequalities in (6) represents an arbitrage opportunity.

The literature on triangular arbitrage supports the existence and persistence of triangular arbitrage opportunities. Marshall et al. (2008) and Lyons and Moore (2009) find exploitable and persistent arbitrage opportunities in the FX spot market. Our paper adds to the existing work by further elaborating on how competition, cost of inventory, and liquidity might deter exploitation of arbitrage deviations.

Table 1 Summary Statistics

	Mean	Std. dev.	Min	Q1	Median	Q3	Max
A	1.560	1.921	1	1.153	1.352	1.648	94.18
$S_{\text{GBP/USD}}$	1.839	1.839	1	1	1.428	2.307	115
$S_{\text{EUR/GBP}}$	1.038	1.153	0.5	0.5	0.958	1.125	63
$S_{\text{EUR/USD}}$	1.904	2.279	1	1	1	2	162
$\Delta_{\text{GBP/USD}}$	2.593	2.771	1	1	2	3	106.33
$\Delta_{\text{EUR/GBP}}$	1.281	1.384	0.5	0.5	1	1.5	59
$\Delta_{\text{EUR/USD}}$	2.856	5.053	1	1	1.818	3	160
$\lambda_{\text{GBP/USD}}$	1.596	2.114	0.003	0.5	1	2	106.33
$\lambda_{\text{EUR/GBP}}$	0.742	1.065	0.002	0.25	0.5	1	59
$\lambda_{\text{EUR/USD}}$	1.614	3.609	0.009	0.425	1	1.667	123
Duration of arbitrage cluster	0.771	1.542	0.01	0.04	0.35	1.01	96.46
Number of arbitrage in a cluster	4.358	4.958	1	1	3	6	88
t-stat. (arbitrage deviation = 0)				162.75			
t-stat. (arbitrage duration = 0.200)				71.61			
Number of profitable clusters				40,166			

Notes. This table presents summary statistics on the deviation, duration, and illiquidity measures from triangular arbitrage parity. We identify triangular arbitrage opportunities by comparing the bid and ask prices for each set of three currencies. The table reports the mean and the standard deviation of average arbitrage deviation A (in pips), duration of arbitrage cluster (in seconds), the numbers of arbitrage opportunities in a cluster, and three different illiquidity measures. Variable Δ_i is defined as the time-series average of the difference between the best and second-best prices of the corresponding exchange rates on the relevant side within each cluster; λ_i is the time-series average of slopes of the corresponding exchange rate on the relevant side within an arbitrage cluster (slope of the demand or supply schedules is calculated as difference between the best and the second-best prices divided by the quantity of the best price); S_i is the time-series average bid–ask spread within a cluster of the corresponding exchange rate. We also present results for the t-test statistics that we use to determine if the mean is statistically different from zero. The sample period is from January 2, 2003, to December 30, 2004.

4. Data Sources and Preliminary Analysis

Our sample includes tick by tick data from the Reuters trading system Dealing 3000 for three currency pairs: U.S. dollar per euro, U.S. dollar per pound sterling, and pound sterling per euro (hereafter, EUR/USD, GBP/USD, and EUR/GBP, respectively) and the sample period runs from January 2, 2003, to December 30, 2004. The Bank for International Settlements (2004) estimates that trades in these currencies constitute up to 60% of the FX spot transactions, 53% of which are interdealer trades during the sample period. Thus, our data represent a substantial part of the FX market.

For each limit order, the data set reports the currency pair, unique order identifier, price, order quantity, hidden quantity (D3000 function), quantity traded, order type, transaction identifier of order entered or removed, status of market order, entry type of orders, removal reason, and times of orders entered and removed. The data time stamps are accurate to one hundredth of a second. This extremely detailed data set makes it easier for us to track all types of orders submitted throughout the day and to update the limit order book for all entries, removals, amendments, and trade executions. The data we analyze exclude all weekends and holidays. An important fact is that the minimum trade size in Reuters trading system Dealing 3000 is 1 million units of the base currency. This is consistent with our earlier theoretical assumption. The currency pairs of interest are traded on a highly liquid market, as highlighted by Tham (2009).

4.1. Summary Statistics

We report the preliminary statistics of arbitrage deviations, clusters (sequences) of profitable triangular arbitrage deviations, and three illiquidity measures within these arbitrage clusters for each currency pair i, where i is GBP/USD, EUR/GBP, and EUR/USD in Table 1. We define a cluster as at least one consecutive deviation from triangular arbitrage given in Equation (6). The duration of a cluster is the elapsed time required for exchange rates to revert to no-arbitrage values after a deviation has been identified. We drop all arbitrage opportunities with a deviation that is less than one pip to account for unexploitable opportunities due to trade fees.[13]

We use three measures of market illiquidity for all analyses in our paper. Consistent with much of the recent literature and Roll et al. (2007), we use the bid–ask spread as one of illiquidity measures and construct our measure, S_i, in their spirit by averaging the bid–ask spreads across each arbitrage cluster for each market i. Our second proxy for illiquidity, slope of the LOB demand and supply schedule, is motivated by Biais et al. (1995) and Næs and Skjeltorp (2006). They

[13] There are costs involved in obtaining a Reuters trading system, but given that market participants are bank dealers who participate in the FX market for purposes other than arbitrage, these costs are sunk costs to a bank who wishes to also pursue arbitrage. We estimate the cost to be about 0.2 pip trade fees.

Table 2 Correlations

	$S_{\text{GBP/USD}}$	$S_{\text{EUR/GBP}}$	$S_{\text{EUR/USD}}$	$\Delta_{\text{GBP/USD}}$	$\Delta_{\text{EUR/GBP}}$	$\Delta_{\text{EUR/USD}}$	$\lambda_{\text{GBP/USD}}$	$\lambda_{\text{EUR/GBP}}$	$\lambda_{\text{EUR/USD}}$
A	**0.059**	**0.125**	**0.044**	**0.295**	**0.090**	**0.382**	**0.241**	**0.086**	**0.395**
$S_{\text{GBP/USD}}$	1.000	**0.080**	**0.040**	**0.157**	**0.107**	**0.122**	0.007	**0.088**	**0.101**
$S_{\text{EUR/GBP}}$		1.000	**0.094**	**0.311**	**0.098**	**0.193**	**0.139**	**0.046**	**0.165**
$S_{\text{EUR/USD}}$			1.000	**0.132**	**0.109**	**0.126**	**0.103**	**0.085**	**0.091**
$\Delta_{\text{GBP/USD}}$				1.000	**0.071**	**0.109**	**0.783**	**0.046**	**0.091**
$\Delta_{\text{EUR/GBP}}$					1.000	**0.082**	**0.057**	**0.779**	**0.073**
$\Delta_{\text{EUR/USD}}$						1.000	**0.091**	**0.071**	**0.864**
$\lambda_{\text{GBP/USD}}$							1.000	**0.031**	**0.076**
$\lambda_{\text{EUR/GBP}}$								1.000	**0.066**

Notes. This table presents the correlation matrix for variables used in the analysis. The variable A denotes the average arbitrage deviation size within the arbitrage cluster. Variable Δ_i is defined as the time-series average of the difference between the best and second-best prices of the corresponding exchange rates on the relevant side within each cluster; λ_i is the time-series average of slopes of the corresponding exchange rate on the relevant side within an arbitrage cluster (slope of the demand or supply schedules is calculated as difference between the best and the second-best prices divided by the quantity of the best price); S_i is the time-series average bid–ask spread within a cluster of the corresponding exchange rate. The sample period is from January 2, 2003, to December 30, 2004. Bold values are significant at 1% level.

argue that there is a close relation between the illiquidity of a security and order book elasticity. Slopes of the demand or supply schedule are calculated as the difference between the best and the second-best prices divided by the quantity of the best price. We define λ_i, the slope illiquidity measure, as the average slopes within an arbitrage cluster of the relevant side for each market i.[14] Motivated by our model, we use Δ_i, the time-series average of the difference between the best and the second-best bid or ask prices within the arbitrage cluster on the relevant side as our third proxy.

Table 1 shows that the mean of the average arbitrage profit, A, within a cluster is about 1.560 pips with a standard deviation of 1.921. Thus, *on average*, triangular arbitrage is profit making after accounting for transaction costs. Furthermore, the associated t-statistics suggest that the deviations are statistically significant. The average duration of a cluster of arbitrage opportunities is 0.771 seconds, indicating that the market eliminates profitable deviations rather quickly. The standard deviation of the duration is about 1.542 seconds. The associated t-statistics in Table 1 suggest that the duration of the arbitrage clusters is statistically significant even when taking into account the potential latency of 200 milliseconds. The average number of limit order messages within a cluster is 4.358. Overall, the preliminary evidence shows that there are potential profitable arbitrage opportunities. These opportunities are small in relative number to the total number of limit orders, but they are sizeable.

[14] For example, if the no-arbitrage condition $P^a(A/B) \geq P^b(A/C) \times P^b(C/B)$ in Equation (6) is violated, the slope measures of the relevant side will be the slope of supply schedule (the ask curve) of currency A/B and the slope of demand schedules (bid curves) of currency pair A/C and C/B. The slope measure λ_i is the time-series average of the slopes on the relevant side across a arbitrage cluster.

Table 2 presents a simple correlation matrix of the average arbitrage deviation size and various liquidity measures. The illiquidity measures are positively correlated to each other, and there are significant positive correlations between the deviation size and various illiquidity measures.

5. Empirical Results

The main goal of this section is to study the main implications of §2.3 empirically. First, we assess whether there is an economically significant arbitrage profit, and then we investigate the relationship among arbitrage profit, market illiquidity, probability of participation, and probability of arbitrage elimination.

5.1. Data Latency

A natural question to ask regarding the positive profit from exploiting arbitrage opportunities is if it comes entirely from data latency and stale quotes. Data latency is defined as the time delay experienced by an investor who submits an order to receive feedback on the status of the order. Controlling for latency in our sample, we will study and account for latency cost up to 200 milliseconds. We compute the total profit without latency costs as the sum of all arbitrage deviations at the beginning of each cluster across our sample from 2003–2004. This is analogous to the exploitation of 40,166 arbitrage opportunities upon observing them. The total profit with latency cost is computed as follows:

- observe an arbitrage opportunity at time t;
- form a portfolio with three currency pairs to exploit triangular arbitrage at time t;
- record the round trip cost (profit/loss) at $t + \tau$, where τ is the latency.

Although exploiting an arbitrage at time t (no latency) without competition is always profitable, accounting for latency can potentially lead to a loss

Table 3 Latency Costs

Latency (in milliseconds)	Total profit	Mean	Std. dev.	t-stat.
0 (Without latency)	6,265,896.07	1.560	1.921	162.70
50	4,321,095.32	1.075	2.979	72.32
100	4,135,887.95	1.029	3.035	67.94
150	3,968,667.98	0.988	3.101	63.85
200	3,783,992.24	0.942	3.146	60.00

Notes. This table presents descriptive statistics and t-statistics of the profits from exploiting triangular arbitrage deviations by a monopolistic arbitrageur (no competition). The "Without latency" row reports statistics of arbitrage profits observed at the beginning of each arbitrage cluster. The rest of the rows provide statistics of arbitrage profits/losses with latencies from 50 to 200 milliseconds. This is tabulated as the round trip cost 50 to 200 milliseconds from the beginning of each cluster. The total profit is the sum of all profits/losses of all arbitrage clusters. The sample period is from January 2, 2003, to December 30, 2004.

especially for arbitrage opportunities that are resolved within τ seconds.[15] We investigate the total arbitrage profit with latencies from 50 to 200 milliseconds.

Table 3 reports the means, standard deviations and t-statistics of profit from exploiting arbitrage opportunities with and without latency costs. The result indicates the importance of accounting for latency cost as the total arbitrage profit decreases with latency. However, the remaining arbitrage profit is significantly large, which demonstrates that the existence of triangular arbitrage is not only due to data latency. The t-statistics indicate that arbitrage profits net of latency cost are still positive and statistically significant.

5.2. Residual Inventory

Because trades can only be carried out in multiples of one million units of the base currency, exploiting triangular arbitrage will cause arbitrageurs to have some residual position exposures. Because the residual position from an arbitrage trade is always smaller than one million units, the arbitrageur will only be able to clear her accumulated residual position when it exceeds a million units. Thus, the arbitrageur is subjected to transaction costs and the risk of price fluctuations when clearing her residual position. We study the role of this residual cost by assuming that the arbitrageur manages her inventory every 60 minutes (strategy 1); that is, she checks her accumulated residual every 60 minutes and trades it away if it exceeds one million on the Reuters platform. There will still be some residual inventory that will remain at the end of the day, which she will trade at the prevailing price

[15] For example, an arbitrageur observes a five pip arbitrage deviation and forms her arbitrage portfolio instantly to exploit the mispricing. If there is no latency, she (without competition) will make a profit of five pips. However, if there is a latency of 200 milliseconds (her portfolio will only be formed 200 ms later) and the arbitrage opportunity is resolved within 50 milliseconds, the arbitrageur will not only miss out on the arbitrage opportunity but also incur a loss because of the transaction cost of her arbitrage portfolio.

Table 4 Residual Costs

Panel A: Residual costs	Total costs	Mean	Std. dev.	t-stat.
Strategy 1 (Hourly)	982,053.57	0.244	15.70	3.12
Strategy 2 (20 minutes)	1,275,294.50	0.318	11.40	5.58
Strategy 3 (Daily + force liquidation)	1,130,160.21	0.281	63.98	0.88
Strategy 4 (Hourly + force liquidation)	1,270,156.47	0.316	12.72	4.98
Strategy 5 (20 minutes + force liquidation)	1,678,800.05	0.418	6.109	13.70

Panel B: Latency and residual costs using strategy 5				
Latency (in milliseconds)	Total profit	Mean	Std. dev.	t-stat.
0 (Without latency)	4,587,096.02	1.142	8.029	28.5
50	2,642,295.27	0.657	9.088	14.4
100	2,457,087.90	0.611	9.144	13.3
150	2,289,867.93	0.570	9.210	12.4
200	2,105,192.19	0.524	9.255	11.3

Notes. This table presents descriptive statistics of the costs of residual inventory management from exploiting triangular arbitrage deviations by a monopolistic arbitrageur (panel A) and the total profit from exploiting arbitrage taking into account both latency and residual costs. We allow the arbitrageur to clear the inventory at different horizons: daily (5:00 P.M. GMT every day), hourly, and every 20 minutes. We also either force the liquidation of the residual inventory at the end of horizon (strategies 3 to 5) or allow the arbitrageur to clear the accumulated residual position only when it exceeds one million units (strategies 1 and 2). In the latter case, the residual inventory only clears at the end of the trading day (5:00 P.M. GMT) using the best prices. In panel B, the "Without latency" row reports statistics of arbitrage profits observed at the beginning of each arbitrage cluster minus the residual costs (using strategy 5). The rest of the rows provide statistics of arbitrage profits/losses with latencies from 50 to 200 ms less the residual costs. The total profit is the sum of all profits/losses of all arbitrage clusters. The sample period is from January 2, 2003, to December 30, 2004.

even when it is below one million units.[16] In addition, we also report the result for 20 minute inventory management (strategy 2).[17] For comparison purposes, we also report results where we force the arbitrageur to clear the inventory at the prevailing prices on Reuters every day (strategy 3), hourly (strategy 4), and every 20 minutes (strategy 5), even though this is not possible in practice because of the trade size limitation.

Panel A in Table 4 reports the total residual cost for various inventory management strategies.

[16] With this procedure, we are likely to overestimate the residual cost because inventory can be either cleared internally within the bank or using alternative trading platforms like OANDA, where one can trade away any small amount of currency at a cheaper cost. We use Reuters best prices at 5:00 P.M. Greenwich mean time (GMT) to clear the end-of-the-day inventory.

[17] This is motivated by Lyons (1998) who shows that the half-life for an FX dealer in inventory management is about 10 minutes, which indicates that a dealer clears her excess inventory approximately every 20 minutes.

Table 5 The Effect of Execution Risk on Arbitrage Deviation

	Contemporaneous			Lagged		
	Δ	λ	S	$lag\Delta$	$lag\lambda$	$lagS$
$illiq_{\text{GBP/USD}}$	0.173 (64.4)	0.195 (54.9)	0.044 (9.70)	0.047 (8.52)	0.006 (0.90)	0.039 (4.70)
$illiq_{\text{EUR/GBP}}$	0.058 (10.6)	0.093 (13.1)	0.135 (18.2)	0.050 (9.37)	0.008 (1.18)	0.009 (0.85)
$illiq_{\text{EUR/USD}}$	0.125 (77.8)	0.185 (80.4)	0.013 (3.49)	0.023 (2.96)	0.030 (3.15)	0.022 (3.60)
$trvol_{\text{GBP/USD}}$	0.052 (1.43)	0.027 (0.73)	−0.054 (−1.30)	0.022 (0.58)	−0.005 (−0.15)	0.001 (0.03)
$trvol_{\text{EUR/GBP}}$	0.018 (0.38)	0.024 (0.50)	−0.021 (−0.39)	−0.110 (−2.05)	−0.130 (−2.42)	−0.117 (−2.16)
$trvol_{\text{EUR/USD}}$	0.034 (0.58)	0.031 (0.51)	−0.195 (−2.85)	−0.481 (−7.69)	−0.480 (−7.68)	−0.459 (−7.23)
$vol_{\text{GBP/USD}}$	−0.185 (−1.18)	−0.110 (−0.70)	0.228 (1.29)	−0.037 (−0.22)	0.173 (0.97)	0.108 (0.63)
$vol_{\text{EUR/GBP}}$	−0.470 (−2.77)	−0.366 (−2.13)	0.104 (0.55)	0.197 (1.05)	0.275 (1.42)	0.182 (0.97)
$vol_{\text{EUR/USD}}$	0.005 (0.12)	0.025 (0.53)	0.067 (1.27)	0.136 (2.59)	0.147 (2.71)	0.125 (2.38)
$quant_{\text{GBP/USD}}$	−0.129 (−1.35)	−0.049 (−0.51)	0.258 (2.39)	0.298 (2.98)	0.313 (3.13)	0.332 (3.32)
$quant_{\text{EUR/GBP}}$	0.101 (1.60)	0.076 (2.01)	0.082 (1.16)	0.122 (1.74)	0.144 (2.04)	0.130 (1.85)
$quant_{\text{EUR/USD}}$	0.016 (0.21)	0.057 (0.74)	−0.104 (−1.21)	0.182 (2.18)	0.183 (2.20)	0.199 (2.39)
ted	0.336 (1.43)	0.156 (0.66)	−0.109 (−0.41)	−0.381 (−1.45)	−0.439 (−1.67)	−0.402 (−1.53)
$p_{\text{GBP/USD}}$	1.272 (0.24)	3.138 (0.58)	8.469 (1.41)	12.620 (2.13)	12.540 (2.12)	13.880 (2.31)
$p_{\text{EUR/GBP}}$	4.756 (0.36)	8.608 (0.64)	20.880 (1.39)	32.040 (2.17)	31.700 (2.14)	34.880 (2.33)
$p_{\text{EUR/USD}}$	−2.930 (−0.37)	−5.333 (−0.67)	−12.530 (−1.42)	−18.770 (−2.16)	−18.550 (−2.13)	−20.550 (−2.33)
Rescosts	4.416 (0.79)	8.069 (1.42)	1.341 (0.21)	7.034 (1.13)	7.609 (1.22)	3.459 (0.55)
$\bar{R}^2$, %	25.56	23.67	5.69	0.62	0.40	0.47

Notes. This table lists coefficient estimates from regression of arbitrage deviation (A) on three measures of illiquidity, $illiq_i$ for each market i, where i is EUR/USD, GBP/USD, and EUR/GBP. The sample consists of 40,166 triangular arbitrage clusters that appeared between January 2, 2003, and December 30, 2004. We use every arbitrage opportunity with deviation larger than one basis point. The variable A denotes the average arbitrage deviation size within the arbitrage cluster. We consider the following proxies for illiquidity: Δ_i is the time-series average of the difference between the best and second-best prices of exchange rates i on the relevant side within each cluster; λ_i is the time-series average of slopes of the corresponding exchange rate on the relevant side within an arbitrage cluster (slope of the demand or supply schedules is calculated as difference between the best and the second-best prices divided by the quantity of the best price); S_i is the time-series average bid–ask spread within a cluster of exchange rate i. "Contemporaneous" columns contain the estimation results when the illiquidity measures are contemporaneous with the profit, whereas "Lagged" columns contain results for the predetermined lagged illiquidity variables. The variable $lag\Delta_i$ is the average across one hour prior to each arbitrage cluster difference between the best and second-best prices on the relevant side of exchange rate i; $lag\lambda_i$ is the average across one hour prior to the arbitrage cluster of the slope of the relevant side of exchange rate i; $lagS_i$ is the average across one hour prior to the arbitrage cluster of the bid–ask spread of exchange rate i; $trvol_i$ is the total trading volume for market i during one hour prior to each arbitrage cluster; vol_i is the volatility for market i constructed using hourly realized volatility of exchange rate returns sampled at a five-minute frequency one hour prior to the arbitrage cluster; $quant_i$ is the average trade size for market i one hour prior to each arbitrage cluster; *ted* is the previous-day TED spread defined as a measure of the premium on Eurodollar deposit rates in London relative to the U.S. Treasury; p_i is the level of the exchange rate for each market. The residual cost, *Rescosts*, is computed as the hourly profit or loss of managing the inventory of the residual one hour prior to each corresponding arbitrage cluster. t-statistics are given in parentheses and are adjusted for autocorrelation.

The result shows the importance of adopting an optimal inventory management strategy as suggested by Lyons (1998).

Panel B in Table 4 reports the combined profit after accounting for latency and residual cost using strategy 5. Although it is likely that we overestimate the residual cost, we use strategy 5 because it is the largest. Taking only residual costs into account, the total profit for a monopolistic arbitrageur is about 4.6 million British pounds. The total profit decreases to 2.1 million when we take a latency cost of 200 milliseconds into account. Table 4 shows the importance of both residual and latency costs as the total profit decreases with latency and residual cost. Nonetheless, we observe that the total profit remains statistically and economically significant.

5.3. Arbitrage Deviations and Market Illiquidity

In equilibrium, there is a positive relationship between execution risk and market illiquidity. Execution risk is more severe in illiquid markets as the impact of trade on prices increases. We test the relation between execution risk and market illiquidity by first estimating the following linear regression:

$$A = \alpha_0 + \alpha_1 illiq_{\text{GBP/USD}} + \alpha_2 illiq_{\text{EUR/USD}} + \alpha_3 illiq_{\text{EUR/GBP}} + \gamma' X + \varepsilon, \tag{7}$$

where A denotes the average arbitrage deviation within the arbitrage cluster, and $illiq_i$ is a measure of illiquidity in the market i. We use three different measures of market illiquidity, S_i, λ_i, and Δ_i as defined in §4.1, and X is a vector of control variables consisting of time dummies, trading volume, volatility, average trade size, the level of each exchange rate, residual costs, and TED spread.

Because an arbitrage will involve all three currency pairs, we cannot have fixed effects for each currency, but we include the level of exchange rate to control for currency pair effects. We control for the effect of the trading volume for market i using the total trading volume one hour prior to the corresponding arbitrage cluster, denoted by $trvol_i$. Volatility for market i is constructed using hourly realized volatility of exchange rate returns sampled at five-minute frequency one hour prior to the corresponding

Table 6 The Effect of Execution Risk on Arbitrage Deviation: Nonlatency Arbitrage

	Contemporaneous			Lagged		
	Δ	λ	S	$lag\Delta$	$lag\lambda$	$lagS$
$illiq_{GBP/USD}$	0.123 (43.5)	0.130 (34.0)	0.048 (10.7)	0.038 (6.58)	0.010 (1.37)	0.042 (5.08)
$illiq_{EUR/GBP}$	0.060 (10.6)	0.100 (12.4)	0.116 (16.0)	0.006 (0.75)	0.004 (0.43)	0.001 (0.12)
$illiq_{EUR/USD}$	0.072 (43.6)	0.102 (42.3)	0.025 (6.94)	0.002 (1.06)	0.004 (1.93)	0.021 (5.73)
$trvol_{GBP/USD}$	−0.001 (−0.05)	−0.011 (−0.95)	−0.037 (−0.71)	0.064 (1.51)	0.048 (1.12)	0.054 (1.28)
$trvol_{EUR/GBP}$	0.015 (0.31)	0.002 (0.47)	−0.024 (0.34)	−0.238 (−4.03)	−0.253 (−4.29)	−0.221 (−3.73)
$trvol_{EUR/USD}$	0.010 (0.18)	−0.006 (−1.87)	−0.122 (−1.42)	−0.439 (−6.43)	−0.439 (−6.42)	−0.403 (−5.87)
$vol_{GBP/USD}$	−0.049 (−0.33)	0.038 (0.25)	0.184 (1.13)	0.011 (0.06)	0.034 (0.83)	0.019 (0.11)
$vol_{EUR/GBP}$	−0.227 (−1.36)	−0.148 (−0.87)	0.053 (0.29)	0.353 (1.73)	0.443 (1.88)	0.251 (1.23)
$vol_{EUR/USD}$	−0.019 (−0.48)	−0.006 (−0.16)	0.019 (0.45)	0.047 (0.97)	0.011 (1.04)	0.034 (0.71)
$quant_{GBP/USD}$	−0.001 (−0.02)	0.023 (0.24)	0.238 (2.29)	0.224 (2.03)	0.233 (2.11)	0.254 (2.31)
$quant_{EUR/GBP}$	0.078 (1.25)	0.069 (1.09)	0.078 (1.17)	0.371 (4.87)	0.384 (5.03)	0.382 (5.02)
$quant_{EUR/USD}$	0.023 (0.32)	0.052 (0.69)	−0.052 (−0.65)	0.326 (3.60)	0.337 (3.73)	0.360 (3.98)
ted	0.015 (0.06)	−0.151 (−0.62)	−0.226 (−0.87)	−0.610 (−2.08)	−0.645 (−2.20)	−0.611 (−2.08)
$p_{GBP/USD}$	4.698 (0.88)	6.149 (1.13)	8.638 (1.50)	10.040 (1.54)	9.600 (1.47)	11.810 (1.81)
$p_{EUR/GBP}$	12.350 (0.93)	15.290 (1.12)	21.300 (1.48)	25.860 (1.58)	24.590 (1.50)	29.860 (1.83)
$p_{EUR/USD}$	−7.407 (−0.94)	−9.260 (−1.16)	−12.660 (−1.49)	−14.820 (−1.54)	−14.090 (−1.46)	−17.360 (−1.80)
Rescosts	3.885 (0.74)	5.426 (1.01)	−1.885 (−0.33)	2.497 (0.39)	3.180 (0.49)	13.700 (1.10)
$\bar{R}^2$, %	21.25	18.16	8.08	0.72	0.55	0.80

Notes. This table lists coefficient estimates from regression of arbitrage deviation (A) on three measures of illiquidity, $illiq_i$ for each market i, where i is EUR/USD, GBP/USD, and EUR/GBP. The sample consists of triangular arbitrage clusters that appeared between January 2, 2003, and December 30, 2004, with a duration greater than 200 milliseconds. We use every arbitrage opportunity with deviation larger than one basis point. The variable A denotes the average arbitrage deviation size within the arbitrage cluster. We consider the following three different proxies for illiquidity: Δ_i is the time-series average of the difference between the best and second-best prices of exchange rates i on the relevant side within each cluster; λ_i is the time-series average of slopes of the corresponding exchange rate on the relevant side within an arbitrage cluster (slope of the demand or supply schedules is calculated as difference between the best and the second-best prices divided by the quantity of the best price); S_i is the time-series average bid–ask spread within a cluster of exchange rate i. "Contemporaneous" columns contains the estimation results when the illiquidity measures are contemporaneous with the profit, whereas "Lagged" columns contain results for the predetermined lagged illiquidity variables. The variable $lag\Delta_i$ is the average across one hour prior to each arbitrage cluster difference between the best and second-best prices on the relevant side of exchange rate i; $lag\lambda_i$ is the average across one hour prior to the arbitrage cluster of the slope of the relevant side of exchange rate i; $lagS_i$ is the average across one hour prior to the arbitrage cluster of the bid–ask spread of exchange rate i; $trvol_i$ is the total trading volume for market i for one hour prior to each arbitrage cluster; vol_i is volatility for market i is constructed using hourly realized volatility of exchange rate returns sampled at a five-minute frequency one hour prior to the corresponding arbitrage cluster; $quant_i$ is the average trade size for market i one hour prior to each arbitrage cluster; ted is the previous-day TED spread define as a measure of the premium on Eurodollar deposit rates in London relative to the U.S. Treasury; p_i is the level of the exchange rate for each market. The residual cost, *Rescosts*, is computed as the hourly profit or loss of managing the inventory of the residual one hour prior to each corresponding arbitrage cluster. t-statistics are given in parentheses and are adjusted for autocorrelation.

arbitrage cluster, denoted by vol_i. We also control for trade size, $quant_i$, using the average trade size for market i one hour prior to the corresponding arbitrage cluster. The residual cost, *Rescosts*, is computed as the hourly profit or loss of managing the inventory of the residual one hour prior to the corresponding arbitrage cluster (we use inventory management strategy 5 described in the previous section). We control for counterparty risk in our regression using lag daily TED spread, *ted*, defined as a measure of the premium on Eurodollar deposit rates in London relative to the U.S. Treasury. We use predetermined control variables to reduce the potential effect of variations in arbitrage deviation on contemporaneous explanatory variables. Thus, we study the contemporaneous relation between liquidity and arbitrage deviation with lagged control variables.[18] We estimate the parameters using the generalized method of moments with a Newey–West correction for autocorrelation and heteroscedasticity.

[18] The result remains qualitatively similar when we use contemporaneous control variables.

The results shown in Table 5 establish positive statistically significant relations between market liquidity and arbitrage deviations. The central results hold for all three different measures of illiquidity. Our results are robust to different choices of lag control variables, i.e., 30 minutes or three hours prior to each cluster. For robustness and to address any potential simultaneity bias, we further test the relation between market illiquidity and arbitrage deviations using predetermined illiquidity variables. We study the relation between measures of illiquidity one hour prior to the arbitrage cluster because it is reasonable to suppose that current arbitrage deviations do not affect lagged illiquidity measures. The results are fairly robust, all signs of the variables remain unchanged and are in most cases statistically significant.

We have highlighted the importance of latency cost in the earlier section; thus, we carry out subsample regression analysis with contemporaneous and lagged illiquidity variables by excluding arbitrage clusters with durations that are less than 200 milliseconds. Results are presented in Table 6. In every instance, the regression coefficients of illiquidity variables are positive and in most cases statistically significant.

The result shows that as the price impact of trade increases with market illiquidity, the cost of execution failure due to crowded trade increases as well. Overall, the result demonstrates that crowded trade is an important limit of arbitrage if it imposes a negative externality on other competing arbitrageurs.

5.4. EBS Data

In our sample, the FX interdealer market is dominated by two trading venues, Electronic Broking Services (EBS) and Reuters. EBS is the primary trading venue for euro and yen (EUR/USD, USD/JPY, EUR/JPY, USD/CHF, and EUR/CHF), and Reuters dominates all other interbank currency pairs. With Reuters being less dominant in EUR/USD than EBS, it is important to understand how this might affect our results because our sample only comes from Reuters.

To study this in more detail, we investigate the impact of EBS by combining the data from both Reuters and EBS. Our EBS sample, acquired from ICAP, consists of 10 best layers of the limit order book at one-second frequency for three currency pairs, EUR/USD, GBP/USD, and EUR/GBP, starting from January 2, 2004, to December 30, 2004.[19] We match each arbitrage opportunity in the Reuters platform with the best EBS quotes and liquidity measures at the last second preceding the arbitrage. We record the generalized profit ($\tilde{A}$) by using the best prices out of the two platforms. Clearly, $\tilde{A} \geq A$.

To account for differences in liquidity across the two platform, we compute a generalized illiquidity measure. For each of the three legs of the arbitrage portfolio, if Reuters provides a better price, we use the corresponding illiquidity measure from Reuters data. Otherwise, it comes from EBS.[20]

Table 7 presents the summary statistics of the generalized profit and illiquidity measures. Comparing $\tilde{A}$ and A in 2004, we observe that the profit has increased significantly. Of the 18,813 profitable Reuters arbitrage opportunities during 2004, 47.2% experience a profit improvement when we account for the EBS data. The average arbitrage deviation in Reuters during 2004 is 1.540 pips, whereas the average profit on the combined data set is significantly larger, with a magnitude of 5.799 basis points. The magnitude of arbitrage deviations is substantial, and it highlights the economic importance of execution risk. The difference between Reuters and the combined profit is highly statistically significant, with a t-statistic of 50.39.

Table 7 Summary Statistics: Reuters Liquidity vs. Generalized Liquidity

	Mean	Std. dev.	Mean	Std. dev.
	Reuters		*Generalized*	
A	1.540	1.280	5.799	11.520
$S_{GBP/USD}$	1.799	1.905	2.580	3.232
$S_{EUR/GBP}$	1.029	1.219	1.182	1.489
$S_{EUR/USD}$	1.925	2.242	1.602	1.968
$\Delta_{GBP/USD}$	2.779	3.056	4.278	10.050
$\Delta_{EUR/GBP}$	1.203	1.037	1.732	4.211
$\Delta_{EUR/USD}$	3.089	5.729	1.896	2.404
$\lambda_{GBP/USD}$	1.690	2.203	2.814	7.413
$\lambda_{EUR/GBP}$	0.691	0.812	1.060	3.397
$\lambda_{EUR/USD}$	1.727	3.951	0.870	1.829
No. of obs.		18,813		

Notes. This table presents summary statistics on the deviations from triangular arbitrage parity and market illiquidity during arbitrage times for Reuters-only data and the combined data (Reuters and EBS). For each arbitrage opportunity in the Reuters platform, we match it with the EBS best prices and liquidity measures at the last second preceding the arbitrage. The column *Reuters* reports the mean and standard deviation of average arbitrage deviation A and three illiquidity measures only on the Reuters platform, whereas the *Generalized* column presents summary statistics for variables defined on the merged data. The variable Δ_i is the time-series average of the difference between the best and second-best prices of exchange rates i on the relevant side within each cluster; λ_i is the time-series average of slopes of the corresponding exchange rate on the relevant side within an arbitrage cluster (slope of the demand or supply schedules is calculated as difference between the best and the second-best prices divided by the quantity of the best price); S_i is the time-series average bid–ask spread within a cluster of exchange rate i. Generalized liquidity measures are taken from the platform that provides the best price. The sample period is from January 2, 2004, to December 30, 2004.

The statistics also reflect EBS's dominance in the EUR/USD and Reuter's dominance the GBP/USD and EUR/GBP market. The dominance of EBS can be seen in all the illiquidity measures of EUR/USD where the generalized illiquidity measures are lower than those in Reuters. For example, the spread for EUR/USD decreases from 1.925 to 1.602 when we include EBS data into the Reuters data set. On the other hand, the generalized illiquidty measures for GBP/USD and EUR/GBP, where Reuters is the major player, are substantially higher.

We study the impact of EBS on our result by reestimating Equation (7) in the following manner:

$$\tilde{A} = \alpha_0 + \alpha_1 \widetilde{illiq}_{\mathrm{GBP/USD}} + \alpha_2 \widetilde{illiq}_{\mathrm{EUR/USD}} + \alpha_3 \widetilde{illiq}_{EUR/GBP} + \gamma' X + \varepsilon, \tag{8}$$

where $\widetilde{illiq}_i$ is a generalized measure of illiquidity in the market i, and X is a vector of control variables. In addition to the control variables defined in the previous section, we include a dummy variable, *ebs*, that is equal to 1 when at least one of the three best

[19] EBS does not provide or sell data with a more accurate time stamp. The choice of 2004 data is arbitrary, and the reason for acquiring only one year of data is cost related.

[20] If an arbitrageur is exploiting $\tilde{A}$ that can involve securities from both EBS and Reuters in her arbitrage portfolio, then she will be concerned about the liquidity of these securities for the venue where she trades these securities.

Table 8 The Effect of Execution Risk on Arbitrage Deviation: EBS Combined Data Sets

	Contemporaneous			Lagged		
	$\tilde{\Delta}$	$\tilde{\lambda}$	$\tilde{S}$	$lag\tilde{\Delta}$	$lag\tilde{\lambda}$	$lag\tilde{S}$
$\widetilde{illiq}_{GBP/USD}$	0.149 (19.4)	0.212 (20.7)	0.517 (21.7)	0.118 (12.8)	0.254 (17.9)	0.856 (29.7)
$\widetilde{illiq}_{EUR/GBP}$	0.191 (10.8)	0.159 (7.31)	0.879 (17.3)	0.077 (2.41)	0.091 (1.85)	0.753 (9.81)
$\widetilde{illiq}_{EUR/USD}$	0.533 (15.8)	0.565 (12.7)	0.303 (7.80)	0.085 (3.18)	0.091 (2.55)	0.231 (3.44)
ebs	7.468 (48.8)	7.560 (49.4)	6.674 (41.9)	7.792 (49.9)	7.654 (49.2)	6.138 (36.9)
$trvol_{GBP/USD}$	−2.273 (−6.38)	−2.452 (−6.85)	−2.252 (−6.33)	−2.512 (−6.93)	−2.542 (−7.06)	−2.360 (−6.67)
$trvol_{EUR/GBP}$	1.252 (2.51)	1.215 (2.43)	1.461 (2.94)	1.133 (2.24)	1.193 (2.36)	1.382 (2.78)
$trvol_{EUR/USD}$	−6.871 (−9.98)	−6.676 (−9.67)	−7.144 (−10.4)	−7.111 (−10.1)	−6.831 (−9.81)	−7.115 (−10.3)
$vol_{GBP/USD}$	9.327 (5.87)	9.599 (6.03)	9.658 (6.10)	8.935 (5.52)	8.601 (5.34)	7.892 (4.97)
$vol_{EUR/GBP}$	−6.658 (−3.30)	−6.225 (−3.08)	−5.788 (−2.88)	−5.653 (−2.74)	−5.908 (−2.88)	−6.584 (−3.29)
$vol_{EUR/USD}$	−0.787 (−2.23)	−0.776 (−2.19)	−0.656 (−1.87)	−0.730 (−2.04)	−0.742 (−2.08)	−0.773 (−2.20)
$quant_{GBP/USD}$	14.160 (14.2)	14.260 (14.2)	14.230 (14.2)	14.290 (14.0)	14.400 (14.2)	13.720 (13.8)
$quant_{EUR/GBP}$	−1.773 (−2.56)	−1.852 (−2.66)	−2.150 (−3.10)	−1.605 (−2.28)	−1.496 (−2.13)	−1.464 (−2.12)
$quant_{EUR/USD}$	1.894 (2.60)	1.850 (2.53)	1.476 (2.03)	1.502 (2.03)	1.627 (2.21)	1.344 (1.86)
ted	−5.705 (−2.62)	−5.487 (−2.51)	−6.230 (−2.87)	−7.483 (−3.39)	−7.383 (−3.36)	−6.915 (−3.20)
$p_{GBP/USD}$	3.726 (2.15)	3.646 (2.10)	3.721 (2.15)	2.649 (1.51)	2.261 (1.29)	3.119 (1.81)
$p_{EUR/GBP}$	11.110 (2.37)	10.850 (2.31)	11.040 (2.36)	8.056 (1.69)	7.029 (1.48)	9.250 (1.98)
$p_{EUR/USD}$	−5.428 (−2.14)	−5.329 (−2.09)	−5.441 (−2.15)	−3.873 (−1.50)	−3.310 (−1.29)	−4.624 (−1.83)
Rescosts	8.407 (12.4)	8.527 (12.6)	7.771 (11.5)	8.478 (12.3)	8.288 (12.1)	7.001 (10.4)
$\bar{R}^2$, %	26.07	25.65	26.44	23.74	24.36	26.95

Notes. This table lists coefficient estimates from regression of arbitrage deviation on three measures of illiquidity for each market i, where i is EUR/USD, GBP/USD, and EUR/GBP. The sample consists of triangular arbitrage clusters appeared between January 2, 2004, and December 30, 2004, drawn from the combined Reuters D3000 and EBS data set. For each arbitrage opportunity in the Reuters platform, we match it with the EBS best quotes and liquidity measures at the last second preceding the arbitrage. Dependent variable $\bar{A}$ is the average arbitrage deviation within the arbitrage cluster. We construct three illiquidity measures, $\tilde{\Delta}$, $\tilde{\lambda}$, and $\tilde{S}$, taken from the corresponding platform that provides the best price. The variable $\tilde{\Delta}_i$ denotes the time-series average of the difference between the best and second-best prices in the market i on the relevant side within each cluster; $\tilde{\lambda}_i$ is the time-series average of slopes of the demand or supply schedules calculated as difference between the best and the second-best prices divided by the quantity of the best price in the market i on the relevant side within an arbitrage cluster; $\tilde{S}_i$ is the time-series average bid–ask spread within a cluster in the market i. "Contemporaneous" columns contain the results when the illiquidity measures are contemporaneous with the profit, whereas "Lagged" columns contain results for the predetermined lagged illiquidity variables. The variables $lag\tilde{\Delta}_i$, $lag\tilde{\lambda}_i$, and $lag\tilde{S}_i$ are the time-series averages of the illiquidity variables across one hour prior to each arbitrage cluster of corresponding variables in market i; $trvol_i$ is the total trading volume for market i for one hour prior to each arbitrage cluster; vol_i is volatility for market i constructed using hourly realized volatility of exchange rate returns sampled at a five-minute frequency one hour prior to the corresponding arbitrage cluster; $quant_i$ is the average trade size for market i one hour prior to each arbitrage cluster; ted is the previous-day TED spread; ebs is a dummy variable that is equal to 1 when at least one of the three best prices comes from EBS and 0 otherwise; p_i is the level of the exchange rate for each market. The residual cost, *Rescosts*, is computed as the hourly profit or loss of managing the inventory of the residual one hour prior of each corresponding arbitrage cluster. t-statistics are given in parentheses and are adjusted for autocorrelation.

prices comes from EBS data. Otherwise, it is 0. Table 8 presents our results. The *ebs* variable is positive and significant, suggesting that arbitrage deviations where at least one of the best prices come from EBS are larger than those from Reuters only. Estimated coefficients of illiquidity variables have the correct sign and remain highly statistically significant.

5.5. Probability of Participation and Order Characteristics Surrounding Arbitrage Eliminations

Although we have established the relationship between arbitrage deviations and market liquidity in the earlier sections, it is still unclear how liquidity and arbitrage deviation affect the probability of arbitrageurs' participation and the duration an arbitrage opportunity takes to be eliminated. In this section, we focus on better understanding this through investigating the relationship among the probability of arbitrage participation, arbitrage deviation, and liquidity. If the probability of participation among arbitrageurs is high, it is more likely that an arbitrage opportunity will be eliminated by a market order. In the presence of execution risk, we expect the probability of participation to be positively related to the size of the arbitrage deviation and the liquidity of arbitrage securities.

To study this, we first split the arbitrage clusters into two groups. The first group consists of arbitrage clusters that are consistent with a textbook arbitrage in that arbitrage opportunities in this group are eliminated by any next incoming order (market order or cancelation), i.e., arbitrages that are executed immediately in the event time scale. The remaining clusters fall into the second group where market participants deliberate on their participation in the market to exploit the observed arbitrage opportunity. We call this the risky arbitrage.

The *Duration* column of Table 9 reports the means and standard deviations of the durations for the

Table 9 Textbook vs. Risky Arbitrage: Summary Statistics

	No. of obs.	*Duration*	*ArbMO*	*Profit*
Textbook arbitrage	15,001	0.12 [0.36]	0.90 [0.30]	1.59 [2.36]
Risky arbitrage	25,165	1.15 [1.82]	0.69 [0.46]	1.54 [1.52]
All	40,166	0.77 [1.54]	0.77 [0.42]	1.56 [1.92]
t-stat.		−86.91	54.21	2.16

Notes. This table presents the means and standard deviations of duration, percentage of arbitrage opportunities terminated by market order, and profit of textbook versus risky arbitrage opportunities. The "Textbook arbitrage" row reports statistics of arbitrage clusters that are eliminated by any next incoming order (market order or cancelation), indicating that clusters in this group have only one profitable triangular arbitrage deviation. This comprises arbitrage clusters that are consistent with a textbook example of an efficient elimination of arbitrage opportunities. The remaining clusters fall into the risky arbitrage group, in which market participants deliberate on the participation of exploiting the observed arbitrage opportunity. *Duration* is the time between the start and end of an arbitrage cluster. *ArbMO* is the fraction of arbitrage opportunities that are terminated by market order. *Profit* is the arbitrage deviation. Standard deviations are reported in squared brackets. The "*t*-stat." row contains the values of the statistic to test if the means of textbook arbitrage duration, profit, and frequency of market order terminations are statistically different from the mean and of risky arbitrage duration. The sample period is from January 2, 2003, to December 30, 2004.

textbook and risky arbitrage. A typical text book arbitrage has an average duration of about 0.12 second, whereas the risky arbitrage takes an average of 1.15 seconds to be eliminated from the market. The results suggest a statistical difference between the duration of textbook and risky arbitrage. This is hardly surprising given that textbook arbitrage, by definition, is eliminated by the next incoming order. More interestingly, we observe that about 90% of the orders that eliminate textbook arbitrage are market orders. In comparison, only 70% of risky arbitrage opportunities are eliminated by market orders. If arbitrage opportunities that are not eliminated immediately (risky arbitrage in our case) are considered riskier than those that are eliminated immediately (textbook arbitrage), then it appears that arbitrageurs are more willing to participate in arbitrage exploitation when it is less risky. This preliminary finding supports our suggestion that arbitrageurs' probability of participation decreases with execution risk. We can also see that textbook arbitrage opportunities are more profitable on average.

We further investigate the relevance of execution risk by examining whether the probability of participation is related to the magnitude of the deviation and the liquidity of arbitrage securities. To investigate this, we estimate the following probit regression:

$$Hit = \alpha_0 + \alpha_1 illiq_{\mathrm{GBP/USD}} + \alpha_2 illiq_{\mathrm{EUR/USD}} + \alpha_3 illiq_{\mathrm{EUR/GBP}} + \alpha_4 A + \alpha_5 dur + \gamma' X + \varepsilon, \quad (9)$$

where *Hit* takes 1 if arbitrage is eliminated by a market order and 0 otherwise. This variable serves as a proxy for probability of participation. The variable A is the average arbitrage deviation during the cluster, dur is the duration of the arbitrage opportunity, $illiq_i$ is the average illiquidity within the corresponding arbitrage cluster in the market i, and X is a vector of control variables. We include time dummies, trading volume, volatility, average trade size, the level of each exchange rate, and TED spread, where control variables are defined in §5.3. We also include dummy variables end_i that takes a value of 1 if the arbitrage ends with orders from market i to control for a specific exchange rate effect.

For robustness, we also investigate the determinants of the after-arbitrage trading activity by estimating the following model:

$$MO = \alpha_0 + \alpha_1 illiq_{\mathrm{GBP/USD}} + \alpha_2 illiq_{\mathrm{EUR/USD}} + \alpha_3 illiq_{\mathrm{EUR/GBP}} + \alpha_4 A + \alpha_5 dur + \gamma' X + \varepsilon, \quad (10)$$

where MO is the total number of market orders posted on the relevant arbitrage exploitation side of the limit order book within 200 milliseconds after an arbitrage ends. This specification considers an alternative measure of the probability of participation by estimating the number of competing arbitrageurs in exploiting the observed arbitrage. For example, if there are five competing arbitrageurs pressing the "button" just before the arbitrage terminates, we should observe a cascade of five market orders after the arbitrage termination. The selection of the 200 millisecond window captures the role of latency.

Table 10 reports the results of the regression. We find that *duration* has a negative and statistically significant coefficient. It implies that the longer the duration of the arbitrage cluster, the less likely an arbitrage will end with a market order. The estimated coefficient of A is positive and statistically significant, which suggests that it is more likely that an arbitrage ends with a market order if the arbitrage deviation is large. More importantly, we find that the probability that an arbitrage ends with a market order decreases with illiquidity.

We find similar results for the number of market orders in the after-arbitrage cascade. Expected number of market orders decreases with the duration of arbitrage opportunity, increases with the magnitude of arbitrage deviation, and decreases with market illiquidity. This suggests that the less risky the arbitrage opportunity is, the more aggressively arbitrageurs try to exploit it. These results are consistent with our model and support the importance of execution risk in high-frequency trading strategies.

5.6. Economic Costs of Execution Risk and Competition

One of the theoretical implications is that execution risk in arbitrage worsens with increasing number

Table 10 Determinants of Probability of Participation

Dependent variable:	*Hit*			*MO*		
Illiquidity:	Δ	λ	*S*	Δ	λ	*S*
dur	−0.024 (−18.9)	−0.024 (−19.3)	−0.020 (−16.0)	−0.007 (−5.53)	−0.007 (−5.47)	−0.007 (−5.15)
A	0.006 (5.15)	0.001 (0.11)	0.001 (1.66)	0.001 (1.12)	0.003 (2.62)	0.001 (0.05)
$illiq_{GBP/USD}$	−0.006 (−9.16)	−0.002 (−2.95)	−0.012 (−11.7)	−0.001 (−1.31)	−0.003 (−3.80)	−0.004 (−3.90)
$illiq_{EUR/GBP}$	−0.009 (−7.15)	−0.006 (−3.38)	−0.007 (−4.14)	−0.002 (−1.46)	−0.004 (−2.41)	0.001 (0.20)
$illiq_{EUR/USD}$	−0.002 (−7.03)	0.001 (0.99)	−0.023 (−26.8)	−0.001 (−2.30)	−0.002 (−4.68)	−0.002 (−2.16)
$trvol_{GBP/USD}$	−0.030 (−3.39)	−0.030 (−3.42)	−0.026 (−3.00)	−0.011 (−1.13)	−0.010 (−1.11)	−0.012 (−1.23)
$trvol_{EUR/GBP}$	0.026 (1.83)	0.043 (3.00)	0.001 (0.01)	0.047 (3.02)	0.042 (2.68)	0.048 (3.10)
$trvol_{EUR/USD}$	−0.003 (−0.24)	0.002 (0.18)	−0.006 (−0.51)	0.039 (2.90)	0.037 (2.76)	0.039 (2.93)
$vol_{GBP/USD}$	0.043 (1.11)	0.035 (0.90)	0.045 (1.17)	−0.041 (−0.97)	−0.039 (−0.92)	−0.037 (−0.88)
$vol_{EUR/GBP}$	−0.074 (−1.72)	−0.104 (−2.41)	−0.036 (−0.84)	−0.017 (−0.38)	−0.010 (−0.22)	−0.017 (−0.38)
$vol_{EUR/USD}$	0.001 (0.11)	−0.002 (−0.24)	0.008 (0.72)	−0.010 (−0.83)	−0.009 (−0.75)	−0.010 (−0.82)
$quant_{GBP/USD}$	−0.016 (−0.85)	−0.021 (−1.12)	−0.007 (−0.39)	−0.018 (−0.90)	−0.018 (−0.89)	−0.019 (−0.93)
$quant_{EUR/GBP}$	−0.009 (−0.43)	−0.028 (−1.22)	0.010 (0.45)	0.010 (0.43)	0.014 (0.59)	0.010 (0.41)
$quant_{EUR/USD}$	0.001 (0.03)	0.001 (0.05)	0.004 (0.30)	−0.022 (−1.29)	−0.023 (−1.30)	−0.020 (−1.18)
ted	0.124 (2.05)	0.149 (2.46)	0.107 (1.78)	−0.099 (−1.51)	−0.104 (−1.58)	−0.099 (−1.50)
$p_{GBP/USD}$	−1.054 (−0.77)	−1.324 (−0.97)	−0.250 (−0.19)	0.134 (0.09)	0.172 (0.12)	0.123 (0.08)
$p_{EUR/GBP}$	1.465 (0.73)	1.840 (0.91)	0.291 (0.15)	0.048 (0.02)	−0.003 (−0.01)	0.065 (0.03)
$p_{EUR/USD}$	−2.692 (−0.79)	−3.355 (−0.98)	−0.486 (−0.14)	−0.560 (−0.15)	−0.459 (−0.12)	−0.567 (−0.15)
$end_{GBP/USD}$	0.379 (80.5)	0.380 (79.6)	0.383 (79.0)	−0.034 (−6.63)	−0.032 (−6.29)	−0.037 (−7.01)
$end_{EUR/GBP}$	0.385 (80.4)	0.387 (80.1)	0.399 (82.0)	−0.038 (−7.36)	−0.038 (−7.38)	−0.036 (−6.87)
$\bar{R}^2$	18.73	18.38	20.23	0.72	0.80	0.75

Notes. This table presents the estimated coefficient from the probit regression of the nature of an arbitrage termination and the number of subsequent market orders within 200 milliseconds on three measures of liquidity: $illiq_i$ for each market i, where i is EUR/USD, GBP/USD, and EUR/GBP; the average arbitrage deviation; and the duration of the cluster. The dependent variable, *Hit*, takes a value of 1 if the arbitrage cluster is terminated be a market order and 0 otherwise. The dependent variable *MO* is the total number of market orders after 200 milliseconds from the termination of arbitrage cluster; *dur* is the duration of the arbitrage cluster; *A* is the average arbitrage deviation within the arbitrage cluster; Δ_i is the time-series average of the difference between the best and second-best prices of exchange rates i on the relevant side within each cluster; λ_i is the time-series average of slopes of the corresponding exchange rate on the relevant side within an arbitrage cluster (slope of the demand or supply schedules is calculated as difference between the best and the second-best prices divided by the quantity of the best price); S_i is the time-series average bid–ask spread within a cluster in the market i; $trvol_i$ is the total trading volume for market i one hour prior to each arbitrage cluster. Volatility for market i is constructed using hourly realized volatility of exchange rate returns sampled at a five-minute frequency one hour prior to the corresponding arbitrage cluster, denoted by vol_i; $quant_i$ is the average trade size for market i one hour prior to each arbitrage cluster; *ted* is the previous-day TED spread defined as a measure of the premium on Eurodollar deposit rates in London relative to the U.S. Treasury; p_i is the level of the exchange rate for each market; end_i is a dummy variable that takes 1 if the arbitrage ends with orders from market i for GBP/USD and EUR/GBP. The sample period is from January 2, 2003, to December 30, 2004.

of competing arbitrageurs. However, it is nontrivial establishing this relation because measuring the number of competing arbitrageurs empirically is challenging. Simulation provides a platform to study the impact of the number of arbitrageurs on execution risk. Moreover, simulation also allows us to approximate the economic cost of execution risk.

We simulate a trading game of artificial arbitrageurs using real limit order book data. This exercise allows us to record the economic value of exploiting triangular arbitrage with different numbers of competing arbitrageurs using market data. It also demonstrates the relevance of execution risk in arbitraging by closely mimicking the trading behavior of arbitrageurs and studying the risk and return of their trading strategies. The setup of the simulation can be found in Appendix C.

The sample size of our data is two years, and we repeat the trading exercise across the sample 1,000 times with numbers of arbitrageurs varying from 2 to 16. Arbitrageurs only attempt to eliminate arbitrage opportunities when they observe deviations from the parity condition of more than one basis point. We also allow the arbitrageurs to employ mixed strategies where they participate with some probability. Arbitrage opportunities that are not immediately eliminated from the market are only assumed to be exploited once at the very first moment the arbitrage conditions are violated.

Table 11 presents the means and standard deviations of profits and losses of arbitrageurs. It shows that arbitrageurs have positive profits when there are only two participating players in the market. As the probability of participation drops, the profit of arbitrageurs in the case of two participants decreases.

To emphasize the importance of execution risk, we increase the number of competing arbitrageurs in the market. The results clearly demonstrate that arbitrageurs can incur losses in the presence of other competing arbitrageurs, as suggested by our model. The loss of arbitraging increases as it becomes more difficult to complete the arbitrage portfolio at the desired

Table 11 Arbitrage Strategy Average Profits

No. of traders	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.4$	$\pi = 0.5$	$\pi = 0.6$	$\pi = 0.7$	$\pi = 0.8$	$\pi = 0.9$	$\pi = 1.0$
2	0.323	0.619	0.881	1.120	1.328	1.498	1.647	1.763	1.846	1.902
	[0.056]	[0.075]	[0.085]	[0.093]	[0.093]	[0.092]	[0.088]	[0.077]	[0.058]	[0.022]
4	0.295	0.496	0.598	0.598	0.490	0.271	−0.064	−0.523	−1.109	−1.828
	[0.056]	[0.076]	[0.084]	[0.093]	[0.096]	[0.093]	[0.094]	[0.084]	[0.073]	[0.056]
6	0.263	0.364	0.284	0.015	−0.448	−1.128	−2.023	−3.140	−4.495	−6.108
	[0.055]	[0.075]	[0.088]	[0.095]	[0.098]	[0.101]	[0.100]	[0.099]	[0.096]	[0.092]
8	0.229	0.223	−0.049	−0.610	−1.472	−2.657	−4.178	−6.046	−8.271	−10.850
	[0.056]	[0.075]	[0.088]	[0.097]	[0.104]	[0.107]	[0.114]	[0.118]	[0.126]	[0.128]
10	0.197	0.077	−0.401	−1.271	−2.558	−4.288	−6.477	−9.125	−12.250	−15.860
	[0.056]	[0.076]	[0.090]	[0.101]	[0.110]	[0.119]	[0.128]	[0.141]	[0.155]	[0.171]
12	0.164	−0.102	−0.865	−2.168	−4.055	−6.549	−9.675	−13.440	−17.890	−23.040
	[0.253]	[0.338]	[0.386]	[0.414]	[0.425]	[0.422]	[0.402]	[0.368]	[0.313]	[0.228]
14	0.127	−0.231	−1.154	−2.693	−4.892	−7.781	−11.38	−15.710	−20.800	−26.640
	[0.057]	[0.077]	[0.095]	[0.111]	[0.129]	[0.146]	[0.170]	[0.197]	[0.226]	[0.257]
16	0.091	−0.394	−1.551	−3.446	−6.123	−9.619	−13.96	−19.160	−25.230	−32.140
	[0.056]	[0.078]	[0.097]	[0.116]	[0.138]	[0.165]	[0.192]	[0.224]	[0.260]	[0.298]

Notes. This table presents the means and standard deviations (in squared brackets) of arbitrageurs' profits and losses in the presence of execution risk. Profits are in million GBP. The arbitrage strategy is based on the triangular arbitrage condition. Arbitrageurs can participate in exploiting arbitrage opportunities when they observe any price discrepancy higher than one basis point. We allow each arbitrageur to trade one million GBP. We repeat the exercise with different numbers of participating arbitrageurs in the market. Arbitrageurs use a mixed strategy, with which they participate with probability π. We compute the values in the table based on 1,000 simulations. All depth and breadth of the market are based on the reconstructed limit order book. The sample period is from January 2, 2003, to December 30, 2004.

price. However, they still have positive profits if they adopt a mixed strategy with a probability of participation of equal to or less than 20%. The results illustrate that a mixed strategy ($\pi < 1$) might be preferred over a pure strategy ($\pi = 1$) in some circumstances.

Table 11 also demonstrates how execution risk increases as the number of competing arbitrageurs increases. This can be seen by the increase in magnitude of losses in a strategy with a probability of participation of one. With the increasing number of competing arbitrageurs and use of algorithmic trading in recent years, our results show that arbitrage can be a very risky business because of execution risk.

6. Conclusion

The paper investigates the role of execution risk in high-frequency trading. To cleanly understand the role of execution risk controlling for other potential risks, we investigate high-frequency arbitrage strategies. We present a new impediment to high-frequency arbitrage with the distinctive feature that it is associated with the crowding effect of arbitrageurs competing for imperfectly divisible assets necessary to form their arbitrage portfolio. In contrast to the common notion that competition improves price efficiency, we argue that competition among arbitrageurs can limit efficiency if they inflict negative externalities on each other. With a detailed data set, we provide empirical evidence of the relevance of execution risk in high-frequency trading. With the recent proliferation in high-frequency trading, our work provides useful insights for practitioners, regulators, and researchers on one of the risks related to high-frequency trading strategies.

An interesting extension for future research is to include in the model a cost for technology that allows arbitrageurs to enter at the head of the queue, like for algorithmic traders. Doing that raises an important question about the determinants of the speed race among market participants and how this might affect execution risk with the increasing use of algorithmic trading. This would allow one to study the effect of algorithmic trading on execution risk and the duration of arbitrage opportunities.

Acknowledgments

For comments and suggestions, the authors thank three anonymous referees, an anonymous associate editor, Wei Xiong (departmental editor), Kees Bouwman, Alain Chaboud, Ba Chu, Dick van Dijk, Mathijs van Dijk, Ingolf Dittman, Jana Fidrmuc, Matthijs Fleischer, Thierry Foucault, Gordon Gemmill, Allaudeen Hameed, Lawrence Harris, Joel Hasbrouck, Kim P. Huynh, Marcin Kacperczyk, Kostas Koufopoulos, Lyubov Kozhan, Rostyslav Kozhan, Bruce Lehmann, Francis Longstaff, Albert Menkveld, Michael Moore, Anthony Neuberger, Stijn Van Nieuwerburgh, Ingmar Nolte, Carol Osler, Andrew Oswald, Christine Parlour, Paolo Pasquariello, Richard Payne, Dagfinn Rime, Mark Salmon, Asani Sarkar, Lucio Sarno, Peter Swan, Ilias Tsiakas, Giorgio Valente, Clara Vega, Michel van de Wel, Ingrid Werner, Avi Wohl, Gunther Wuyts, Bernt Arne Ødegaard, and participants at the 2009 European Finance Association Conference in Bergen, 2009 European Economic Association Meeting in Barcelona, 2009 NYSE Euronext and

Tinbergen Institute Workshop on Liquidity and Volatility in Amsterdam, 2009 Second Erasmus Liquidity Conference in Rotterdam, and 2010 Asian Finance Association Meeting in Hong Kong, and 2010 Annual Central Bank Workshop on the Microstructure of Financial Markets in New York. Special thanks go to Elvira Sojli for her detailed and helpful comments, which led to considerable improvements in this paper. All errors remain the authors' responsibility.

Appendix A. Proof of Proposition 1

(i) We derive Equation (3) by induction. Let us first check that the statement is true for $I=2$. The expected profit of the arbitrageur using "trade" strategy is

$$\begin{aligned}E(U^j) &= A\mathbf{P}_1^j\mathbf{P}_2^j + (A-\Delta_1(w_1))\mathbf{P}_2^j(1-\mathbf{P}_1^j)\\ &\quad + (A-\Delta_2(w_2))\mathbf{P}_1^j(1-\mathbf{P}_2^j)\\ &\quad + (A-\Delta_1(w_1)-\Delta_2(w_2))(1-\mathbf{P}_1^j)(1-\mathbf{P}_2^j)\\ &= A-\Delta_1(w_1)(1-\mathbf{P}_1^j)-\Delta_2(w_2)(1-\mathbf{P}_2^j).\end{aligned}$$

Suppose that $E(U^j)=A-\sum_{i=1}^{I-1}\Delta_i(w_i)(1-\mathbf{P}_i^j)$ for $I-1$ markets. Let us show that the corresponding statement is also true in the case of I markets.

Let J be a nonempty subset of $\mathcal{I}$. If arbitrageur fails to get the best prices in each market from J and gets the best prices in all the rest $\mathcal{I}\backslash J$ markets, her payoff will be $A-\sum_{i\in J}\Delta_i(w_i)$, which occurs with probability $\prod_{i\in J}\bar{\mathbf{P}}_i^j\cdot\prod_{i\in\mathcal{I}\backslash J}\mathbf{P}_i^j$. The expected payoff is

$$\begin{aligned}E(U^j) &= \sum_{J\in 2^{\mathcal{I}}}\left(\left(A-\sum_{i\in J}\Delta_i(w_i)\right)\prod_{i\in J}(1-\mathbf{P}_i^j)\prod_{\iota\in\mathcal{I}\backslash J}\mathbf{P}_\iota^j\right)\\ &= A-\sum_{J\in 2^{\mathcal{I}}}\left(\left(\sum_{i\in J}\Delta_i(w_i)\right)\prod_{i\in J}(1-\mathbf{P}_i^j)\prod_{\iota\in\mathcal{I}\backslash J}\mathbf{P}_\iota^j\right)\\ &= A-(1-\mathbf{P}_I^j)\sum_{J\in 2^{\mathcal{I}\backslash\{I\}}}\left(\left(\Delta_I(w_I)+\sum_{i\in J}\Delta_i(w_i)\right)\right.\\ &\qquad\left.\cdot\prod_{i\in J}(1-\mathbf{P}_i^j)\prod_{\iota\in\mathcal{I}\backslash J}\mathbf{P}_\iota^j\right)\\ &\quad -\mathbf{P}_I^j\sum_{J\in 2^{\mathcal{I}\backslash\{I\}}}\left(\left(\sum_{i\in J}\Delta_i(w_i)\right)\prod_{i\in J}(1-\mathbf{P}_i^j)\prod_{\iota\in\mathcal{I}\backslash J}\mathbf{P}_\iota^j\right)\\ &= A-\Delta_I(w_I)(1-\mathbf{P}_I^j)\\ &\quad -(1-\mathbf{P}_I^j)\left(\sum_{i=1}^{I-1}\Delta_i(w_i)(1-\mathbf{P}_i^j)\right)-\mathbf{P}_I^j\left(\sum_{i=1}^{I-1}\Delta_i(w_i)(1-\mathbf{P}_i^j)\right)\\ &= A-\sum_{i=1}^{I}\Delta_i(w_i)(1-\mathbf{P}_i^j).\end{aligned}$$

(ii) Let $2^{\mathcal{K}_{-j}}$ denotes a family of all subsets of the set $\mathcal{K}_{-j}$ of all opponents of arbitrager j. For any $S\in 2^{\mathcal{K}_{-j}}$, denote by $|S|$ the number of elements in S. Consider the set of mutually exclusive and exhaustive events X_S with $S\in 2^{\mathcal{K}_{-j}}$, each of which means that all opponents of trader j from S participate in the market with certainty and the rest $\mathcal{K}_{-j}\backslash S$ opponents do not. The probability of the event X_S occurring is equal to $\mathbf{P}(X_S)=\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}_{-j}\backslash S}(1-\pi_s)$, and the probability of failing to get the best price in the market i for trader j *conditional* on X_S is

$$\bar{\mathbf{P}}_{i,|X_S|}^j=\begin{cases}0 & |S|\le n_i(w_i)-1,\\ 1-\dfrac{n_i(w_i)}{|S|+1} & |S|>n_i(w_i)-1,\end{cases}$$

because there are only $|S|$ opponents are in the market. By the law of total probability, we get

$$\bar{\mathbf{P}}_{i\,|\,k,\Pi_{-j}}^j=\sum_{S\in 2^{\mathcal{K}_{-j}}}\bar{\mathbf{P}}_{i,|S|}^j\cdot\mathbf{P}(X_S)=\sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}_{-j}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|}^j. \tag{A1}$$

Let us now add one more arbitrageur $k+1$ into the market who plays her mixed strategy π_{k+1}. The new set of arbitrageurs is now denoted by $\mathcal{K}'=\{1,\dots,k+1\}$, and let the new mixed strategy profile be $\Pi'=\{\pi_1,\dots,\pi_k,\pi_{k+1}\}$. We need to show that $\bar{\mathbf{P}}_{i\,|\,k+1,\Pi'_{-j}}^j>\bar{\mathbf{P}}_{i\,|\,k,\Pi_{-j}}^j$ for each $j\in\mathcal{K}$ and $i\in\mathcal{I}$. Decomposing the sum in (A1) into the part with subsets S containing the arbitrageur $k+1$ and not yields

$$\begin{aligned}\bar{\mathbf{P}}_{i\,|\,k+1,\Pi'_{-j}}^j &= \sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S\cup\{k+1\}}\pi_s\prod_{s\in\mathcal{K}_{-j}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|+1}^j\\ &\quad+\sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}'_{-j}\cup\{k+1\}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|}^j\\ &= \pi_{k+1}\sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}_{-j}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|+1}^j+(1-\pi_{k+1})\\ &\quad\cdot\sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}'_{-j}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|}^j\\ &> \pi_{k+1}\sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}_{-j}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|}^j+(1-\pi_{k+1})\\ &\quad\cdot\sum_{S\in 2^{\mathcal{K}_{-j}}}\prod_{s\in S}\pi_s\prod_{s\in\mathcal{K}'_{-j}\backslash S}(1-\pi_s)\bar{\mathbf{P}}_{i,|S|}^j=\bar{\mathbf{P}}_{i\,|\,k,\Pi_{-j}}^j.\end{aligned}$$

This also implies the monotonicity of the expected payoff (see Equation (2)). Q.E.D.

Appendix B. Proof of Proposition 2

(i) We get Equation (5) by directly substituting $\Delta_i=\Delta$ and $n_i=1$. Also note that probability of failing to get the best price in this case is

$$\begin{aligned}\bar{\mathbf{P}}_{k,\pi}^i &= \sum_{s=1}^{k-1}\binom{k-1}{s}\pi^s(1-\pi)^{k-s-1}\left(1-\frac{1}{s+1}\right)\\ &= \sum_{s=1}^{k-1}\binom{k-1}{s}\pi^s(1-\pi)^{k-s-1}\\ &\quad-\sum_{s=1}^{k-1}\binom{k-1}{s}\pi^s(1-\pi)^{k-s-1}\left(\frac{1}{s+1}\right)\\ &= 1-(1-\pi)^{k-1}-\frac{1}{k\pi}\sum_{s=2}^{k}\binom{k}{s}\pi^s(1-\pi)^{k-s}\\ &= 1-(1-\pi)^{k-1}-\frac{1}{k\pi}(1-(1-\pi)^k-k\pi(1-\pi)^{k-1})\\ &= 1-\frac{1-(1-\pi)^k}{k\pi}.\end{aligned} \tag{B1}$$

(ii) Throughout the proof we write $\pi(k)$ to emphasize that the probability of participation π is a function of number of arbitrageurs k. Note that Equations (5) and (B1) can be rewritten as

$$\text{ProbElim}(k) = D\pi(k)k, \tag{B2}$$

$$D = \frac{1-(1-\pi(k))^k}{\pi(k)k}, \tag{B3}$$

where $D = 1 - A/(I\Delta)$. Because we assume that $A < I\Delta(1-1/k)$, we have $1 > D > 1/k$. Under such assumption there is always a unique solution of Equation (B3) on interval $(0,1)$. To see this, for a given $k>1$, define a function $\Psi(p) = 1-(1-p)^k - Dkp$. Note that $\Psi(0) = 0$, and $\Psi(1) = 1 - Dk < 0$. The first derivative $\Psi'(p) = k((1-p)^{k-1} - D)$ equals $k(1-D) > 0$ at point 0, and its second derivative $\Psi''(p) = -k(k-1)(1-p)^{k-2}$ is negative on $(0,1)$. Therefore, function Ψ has the only zero on interval $(0,1)$, which implies that probability of participation $\pi(k)$ is correctly defined.

Claim 1. $\pi(k) \to 0$ *as* $k \to \infty$.

Assume by contradiction that there exists $\varepsilon > 0$ such that for any integer k_0 there exists $k_1 > k_0$ satisfying $\pi(k_1) \geq \varepsilon$. Choose any $k_0 > 1/(\varepsilon D)$. Then there exists $k_1 > k_0$ such that

$$\frac{1-(1-\pi(k_1))^{k_1}}{\pi(k_1)k_1} < \frac{1}{\pi(k_1)k_1} \leq \frac{1}{k_1\varepsilon} < D,$$

which contradicts Equation (B3).

Claim 2. $\pi(k)$ *is decreasing with* k.

To show this we extend function π onto $(1,+\infty)$. For any fixed $t \in (1,+\infty)$, let $\tilde{\pi}(t)$ be a solution of equation

$$D = \frac{1-(1-\tilde{\pi}(t))^t}{\tilde{\pi}(t)t}. \tag{B4}$$

Using the same argument as before, one can show that for each t there exists a unique solution of this equation on interval $(0,1)$. Moreover, $\tilde{\pi}(k) = \pi(k)$ for any integer $k > 1$.

Using the implicit function theorem, we can compute the derivative $\partial\tilde{\pi}(t)/\partial t$ using the implicit function $\Phi(\tilde{\pi}, t) = 1 - (1-\tilde{\pi})^t - Dt\tilde{\pi} = 0$ as follows:

$$\frac{\partial\tilde{\pi}(t)}{\partial t} = -\frac{\partial\Phi/\partial t}{\partial\Phi/\partial\tilde{\pi}} = -\frac{1}{t}\cdot\frac{(1-\tilde{\pi}(t))^t\ln(1-\tilde{\pi}(t)) + D\tilde{\pi}(t)}{D-(1-\tilde{\pi}(t))^{t-1}}. \tag{B5}$$

Note that $D-(1-\tilde{\pi}(t))^{t-1} > 0$ (use the fact that $(1-p)^k < 1/(1+pk)$ for any $p\in(0,1)$ and $k>1$). Given the inequality $\ln(1-p) > -p/1-p$ for any $p \in (0,1)$ we have

$$\frac{\partial\tilde{\pi}(t)}{\partial t} < -\frac{1}{t}\cdot\frac{(1-\tilde{\pi}(t))^t(-(\tilde{\pi}(t)/(1-\tilde{\pi}(t)))) + D\tilde{\pi}(t)}{D-(1-\tilde{\pi}(t))^{t-1}} = -\frac{\tilde{\pi}(t)}{t} < 0.$$

Hence, $\tilde{\pi}(t)$ is monotonically decreasing on $(1,+\infty)$, which implies that for any integer k we have $\tilde{\pi}(k+1) - \tilde{\pi}(k) = \pi(k+1) - \pi(k) < 0$.

Claim 3. ProbElim(k) *is decreasing with* k.

Define $\widetilde{\text{ProbElim}}(t) = Dt\tilde{\pi}(t)$ for any $t \in (1,+\infty)$. Note that for any integer k we have $\widetilde{\text{ProbElim}}(k) = \text{ProbElim}(k)$. Let us show first that $\partial\widetilde{\text{ProbElim}}(t)/\partial t < 0$. Indeed,

$$\begin{aligned}\frac{\partial\widetilde{\text{ProbElim}}(t)}{\partial t} &= D\left(\tilde{\pi}(t) + t\frac{\partial\tilde{\pi}(t)}{\partial t}\right)\\ &= D\left(\tilde{\pi}(t) - \frac{(1-\tilde{\pi}(t))^t\ln(1-\tilde{\pi}(t)) + D\tilde{\pi}(t)}{D-(1-\tilde{\pi}(t))^{t-1}}\right)\\ &= -\frac{D}{t}\cdot\frac{(1-\tilde{\pi}(t))^{t-1}(\tilde{\pi} + (1-\tilde{\pi}(t))\ln(1-\tilde{\pi}(t)))}{D-(1-\tilde{\pi}(t))^{t-1}} < 0.\end{aligned}$$

Thus, $\widetilde{\text{ProbElim}}(t)$ is monotonically decreasing on $(1,+\infty)$, so for any integer k we have $\widetilde{\text{ProbElim}}(k+1) - \widetilde{\text{ProbElim}}(k) = \text{ProbElim}(k+1) - \text{ProbElim}(k) < 0$. Q.E.D.

Appendix C. Simulation Setup

The exercise is set up with k arbitrageurs competing for limited supplies of three currency pairs ($I = 3$) required to construct a profitable arbitrage portfolio. These competing arbitrageurs trade on three currency pairs in the spot FX market and are assumed to be able to see the whole limit order book. When an arbitrage opportunity arises, all arbitrageurs observe it and compete to earn the arbitrage profit. To do this, they need to complete a full round of buying and selling of the three currencies in the three different markets. Arbitrageurs can trade at most one unit of each currency by placing all three market orders simultaneously. They trade with probability π. Whether or not their demand for a particular currency is fulfilled at the best available price depends on the demand of the competitors and the supply at the best price. For all arbitrageurs to walk away with a profit, the minimum quantity available at the best price for each currency in the arbitrage portfolio has to be at least k. If there are more arbitrageurs than the quantity available for one of the currencies and the probability of participation is one, each arbitrageur has a probability of $\mathbf{P} = n_1^a/k$ to get the currency at the best price, where n_1^a is the quantity available at the best ask price.

To simulate the trading game when probability of participation $\pi < 1$, we first generate the number of arbitrageurs $|S|+1$ participating at current instant using Bernoulli distribution. We then randomly determine the order in which the currently participating arbitrageurs arrive in each market. Those arbitrageurs who are unsuccessful in acquiring all the required currencies are assumed to complete the remaining legs of the arbitrage portfolio at the next best price or sell their excess inventory at the best available price, whichever has the least loss. Excess inventory exists when an arbitrageur fails to buy all three currencies at the best prices. For example, if an arbitrageur only manages to buy or sell currencies 1 and 2 at the best available prices but misses out on currency 3 because of excess demand, she is left with an open position consisting of currencies 1 and 2. She can either complete the third leg (currency 3) at the next available price or resell and rebuy currencies 1 and 2 (losing out on transaction costs) to close her position. We assume she closes her position using the strategy with the best payoff (the payoff can still be positive if the next best price of currency 3 still yields a positive profit). Our starting and base currency is GBP. There will be some residual position exposures in the exercise because we assume that trades can only be carried out in multiples of one million units of the base currency. We trade out these residual positions at the market prevailing prices and convert them back to GBP at the end of the day.

References

Akram F, Rime D, Sarno L (2008) Arbitrage in the foreign exchange market: Turning on the microscope. *J. Internat. Econom.* 76(2):237–253.

Bank for International Settlements (2004) Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity in April 2004. Press release, September 28. Accessed May 28, 2012, http://www.bis.org/publ/rpfx04.pdf.

Biais B, Hillion P, Spatt C (1995) An empirical analysis of the limit order book and the order flow in the Paris Bourse. *J. Finance* 50(5):1655–1689.

Brogaard JA (2010) High frequency trading and its impact on market quality. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Chaboud A, Chiquoine B, Hjalmarsson E, Vega C (2010) Rise of the machines: Algorithmic trading in the foreign exchange market. International Finance Discussion Papers, The Federal Reserve Board, Washington, DC.

Committee of European Securities Regulators (CESR) (2010) Call for evidence: Micro-structural issues of the European equity markets. CESR/10-142. Accessed May 28, 2012, http://www.esma.europa.eu/system/files/10_142.pdf.

Dewan S, Mendelson H (1990) User delay costs and internal pricing for service facility. *Management Sci.* 36(12):1502–1517.

Easley D, Hendershott T, Ramadorai T (2009) Levelling the trading field. Working paper, Cornell University, Ithaca, NY.

Fong WM, Valente G, Fung J (2010) Covered interest arbitrage profits: The role of market liquidity and credit risk. *J. Banking Finance* 34(5):1098–1107.

Gromb D, Vayanos D (2010) Limits of arbitrage: The state of the theory. NBER Working Paper 15821, National Bureau of Economic Research, Cambridge, MA.

Hendershott T, Moulton P (2011) Automation, speed, and stock market quality: The NYSE's hybrid. *J. Financial Markets* 14(4):568–604.

Hendershott T, Jones C, Menkveld A (2011) Does algorithmic trading improve liquidity? *J. Finance* 66(1):1–33.

Holden C (1995) Index arbitrage as cross-sectional market making. *J. Futures Markets* 15(4):423–455.

Hosanagar K, Chuang J, Krishnan R, Smith MD (2008) Service adoption and pricing of content delivery network (CDN) services. *Management Sci.* 54(9):1579–1593.

Kamara A, Miller TW (1995) Daily and intradaily tests of European put-call parity. *J. Financial Quant. Anal.* 30(4):519–539.

Kleidon A (1992) Arbitrage, nontrading, and stale prices: October 1987. *J. Bus.* 65(4):483–507.

Kondor P (2009) Risk in dynamic arbitrage: Price effects of convergence trading. *J. Finance* 64(2):631–655.

Krishnan H, Winter RA (2010) Inventory dynamics and supply chain coordination. *Management Sci.* 56(1):141–147.

Kumar P, Seppi D (1994) Information and index arbitrage. *J. Bus.* 67(4):481–509.

Lariviere MA, Mieghem JAV (2004) Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing Service Oper. Management* 6(1):23–40.

Lyons R (1998) Profits and position control: A week of FX dealing. *J. Internat. Money Finance* 17(1):97–115.

Lyons R, Moore M (2009) An information approach to international currencies. *J. Internat. Econom.* 79(2):211–221.

Marshall B, Treepongkaruna S, Young M (2008) Exploitable arbitrage opportunities exist in the foreign exchange market. *American Finance Association Annual Meeting, New Orleans*.

Martínez-de-Albéniz V, Talluri K (2011) Dynamic price competition with fixed capacities. *Management Sci.* 57(6):1078–1093.

Mendelson H (1985) Pricing computer services: Queueing effects. *Comm. Assoc. Comput. Machinery* 28(3):312–321.

Næs R, Skjeltorp JA (2006) Order book characteristics and the volume-volatility relation: Empirical evidence from a limit order market. *J. Financial Markets* 9(4):408–432.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Oehmke M (2009) Gradual arbitrage. Working paper, Columbia University, New York.

Perold A (1988) The implementation shortfall: Paper vs. reality. *J. Portfolio Management* 14(3):4–9.

Roll R, Schwartz E, Subrahmanyam A (2007) Liquidity and the law of one price: The case of future-cash basis. *J. Finance* 62(5):2201–2234.

Stein J (2009) Sophisticated investors and market efficiency. *J. Finance* 64(4):1517–1548.

Tabb L, Iati R, Sussman A (2009) US equity high frequency trading: Strategies, sizing, and market structure. Research paper, TABB Group, Westborough, MA.

Tham WW (2009) Macroeconomic announcements, price discovery, liquidity in FX market. Working paper, Erasmus School of Economics, Rotterdam, The Netherlands.

U.S. Securities and Exchange Commission (SEC) (2010) Concept release on equity market structure. Release No. 34-61358; File No. S7-02-10. Accessed May 28, 2012, http://www.sec.gov/rules/concept/2010/34-61358.pdf.