Hautsch, Nikolaus; Scheuch, Christoph; Voigt, Stefan

Working Paper

Limits to arbitrage in markets with stochastic settlement latency

CFS Working Paper Series, No. 616

Provided in Cooperation with:
Center for Financial Studies (CFS), Goethe University Frankfurt

Suggested Citation: Hautsch, Nikolaus; Scheuch, Christoph; Voigt, Stefan (2018) : Limits to arbitrage in markets with stochastic settlement latency, CFS Working Paper Series, No. 616, Goethe University Frankfurt, Center for Financial Studies (CFS), Frankfurt a. M., https://nbn-resolving.de/urn:nbn:de:hebis:30:3-480536

This Version is available at:
https://hdl.handle.net/10419/190979

CENTER FOR FINANCIAL STUDIES

CFS Working Paper Series

No. 616

Nikolaus Hautsch, Christoph Scheuch, and Stefan Voigt

Limits to Arbitrage in Markets with Stochastic Settlement Latency

Center for Financial Studies | Goethe University | Theodor-W.-Adorno-Platz 3 | 60323 Frankfurt am Main | Germany
Phone +49 69 798 30050 | Fax +49 69 798 30077 | info@ifk-cfs.de | www.ifk-cfs.de

The CFS Working Paper Series

presents ongoing research on selected topics in the fields of money, banking and finance. The papers are circulated to encourage discussion and comment. Any opinions expressed in CFS Working Papers are those of the author(s) and not of the CFS.

The Center for Financial Studies, located in Goethe University Frankfurt's House of Finance, conducts independent and internationally oriented research in important areas of Finance. It serves as a forum for dialogue between academia, policy-making institutions and the financial industry. It offers a platform for top-level fundamental research as well as applied research relevant for the financial sector in Europe. CFS is funded by the non-profit-organization Gesellschaft für Kapitalmarktforschung e.V. (GfK). Established in 1967 and closely affiliated with the University of Frankfurt, it provides a strong link between the financial community and academia. GfK members comprise major players in Germany's financial industry. The funding institutions do not give prior review to CFS publications, nor do they necessarily share the views expressed therein.

Center for Financial Studies | Goethe University | Theodor-W.-Adorno-Platz 3 | 60323 Frankfurt am Main | Germany
Phone +49 69 798 30050 | Fax +49 69 798 30077 | info@ifk-cfs.de | www.ifk-cfs.de

Limits to Arbitrage in Markets with Stochastic Settlement Latency*

Nikolaus Hautsch Christoph Scheuch Stefan Voigt

December 3, 2018

Distributed ledger technologies rely on consensus protocols confronting traders with random waiting times until the transfer of ownership is accomplished. This time-consuming settlement process exposes arbitrageurs to price risk and imposes limits to arbitrage. We derive theoretical arbitrage boundaries under general assumptions and show that they increase with expected latency, latency uncertainty, spot volatility, and risk aversion. Using high-frequency data from the Bitcoin network, we estimate arbitrage boundaries due to settlement latency of on average 124 basis points, covering 88% of the observed cross-exchange price differences. Settlement through decentralized systems thus induces non-trivial frictions affecting market efficiency and price formation.

JEL Codes: G00, G10, G14

Keywords: Arbitrage, Settlement Latency, Distributed Ledger, Blockchain

*Hautsch is at University of Vienna (Department of Statistics and Operations Research, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria), Research Platform "Data Science @ Uni Vienna" as well as Vienna Graduate School of Finance (VGSF) and Center for Financial Studies (CFS), Frankfurt. Scheuch and Voigt are at VGSF and WU (Vienna University of Economics and Business), Department of Finance, Accounting and Statistics, Welthandelsplatz 1, Building D4, 1020 Vienna, Austria. Corresponding author: stefan.voigt@vgsf.ac.at. We thank Sylvia Frühwirth-Schnatter, Sergey Ivliev, seminar participants at QFFE 2018, 1st International Conference on Data Science in Finance with R, 4th Konstanz-Lancaster Workshop on Finance and Econometrics, Crypto Valley Blockchain Conference 2018, HFFE 2018 conference, Pisa, University of Heidelberg, University of Vienna, University of Graz and CUNEF, Madrid, for helpful comments and suggestions. Scheuch and Voigt gratefully acknowledge financial support from the Austrian Science Fund (FWF project number DK W 1001-G16).

1 Introduction

Latency is generally defined as the time interval between the transmission of a request and the receipt of a corresponding response. In financial market trading, the order execution latency is the time it takes from sending out an market order until it enters the orderbook. By contrast, the *settlement* latency is the time until the corresponding transaction is settled, i.e., until the legal transfer of ownership is accomplished. In traditional markets, this settlement process involves several intermediaries and takes two to three business days after the execution of a trade (see, e.g., SEC (2017) or Khapko and Zoican (2017)). Due to this delay, the trading process is typically disconnected from the transfer of ownership as central securities depositories provide clearing and settlement. The implications of this separation are twofold. First, both transacting parties face counterparty risks during the settlement period. Second, the transfer of ownership does not affect the speed of trading since central clearing provides the opportunity for an immediate continuation of trading on a newly acquired position. Hence, ultimately market participants only face execution latency, which is crucial in high-frequency trading (see, e.g., Hasbrouck and Saar (2013) or Foucault et al. (2017)), but usually negligible for other market participants.

Distributed ledger technologies, such as blockchain or directed acyclic graphs, promise fast settlement without the need of such designated intermediaries. To considerably shorten the settlement process to a few minutes or seconds, distributed systems rely on protocols to establish consensus about transaction histories. Consensus protocols are still time-consuming as validators have to reach agreement on the current state of the ledger either through computationally expensive competition, decentralized voting, or simply the distribution of new information to other participants.[1] In fact, the time it takes a transaction until its inclusion in a distributed ledger is generally non-trivial and stochastic. Traders thus face uncertain waiting times until the transfer of assets between

[1]See Appendix A for more details on distributed ledgers, consensus protocols, and the underlying technological aspects.

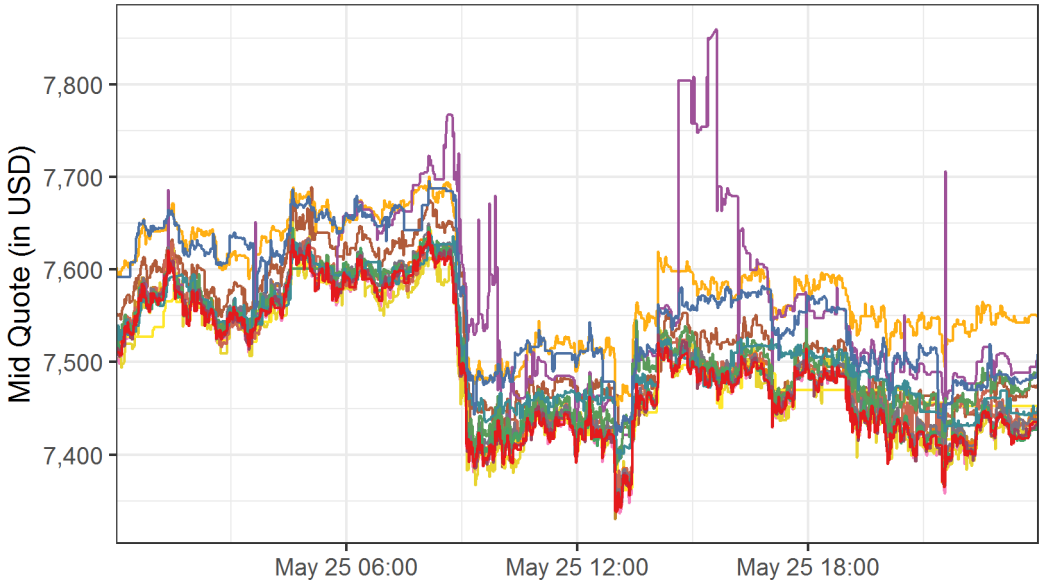
markets is settled on a distributed ledger. Unlike in traditional markets, a trader cannot dispose of her position before the transfer of ownership is confirmed by the network. This stochastic settlement latency is an inherent feature of distributed ledgers and it is several magnitudes larger than (execution) latencies in traditional markets. Even though distributed ledger technologies promise fast settlement on secure ledgers at low costs, they introduce a severe market friction which has implications for trading on and across markets.

In this paper, we show that stochastic latency introduced by distributed systems imposes limits to (statistical) arbitrage. While it is well established that transaction costs limit the possibility to exploit instantaneous differences in prices of a single asset that is traded on multiple markets (e.g., de Long et al., 1990; Pontiff, 1996), we show that stochastic settlement latency constitutes an additional friction. Settlement latency exposes arbitrageurs to the risk of adverse price movements and implies boundaries below which risk-averse arbitrageurs abstain from exploiting violations of the law of one price. Price differences below these boundaries might thus persist and are consistent with the risk-return trade-off of a rational arbitrageur. We theoretically derive these arbitrage boundaries under general conditions and thereby provide a framework to assess (statistical) arbitrage opportunities and the efficiency of markets with stochastic settlement latency.

Figure 1 illustrates the relevance of a theoretical framework to quantify (potential) arbitrage opportunities in markets with stochastic latency. It shows the midquotes of different exchanges featuring trading Bitcoin against US Dollar for a representative day in 2018. We find a commonly observed feature in cross-market Bitcoin trading[2]: the existence of substantial and persistent price differences between individual exchanges. The question to be addressed is whether these price differences are due to (neglected) arbitrage opportunities or due to market frictions, such as settlement latency.

[2]See, e.g., Makarow and Schoar (2018) or Choi et al. (2018).

Figure 1: Bitcoin-Dollar Midquotes on May 25, 2018.
This figure shows the midquotes of one Bitcoin in US Dollar on May 25, 2018, for 17 different exchanges. We gather high-frequency order book information of these exchanges by accessing their public application programming interfaces (APIs) on a minute level and calculate the midquote as the average of the best bid and best ask.



The main contribution of our paper is twofold. First, we provide a general theory on latency-induced limits to arbitrage that is applicable to any market where the processing of transactions and the transfer of assets between markets is time-consuming and where no intermediary is stepping in and guarantees immediate settlement. We consider a risk-averse arbitrageur who monitors the quotes of an asset on multiple markets. If an instantaneous transfer of the asset between markets is possible, a bid quote on one market exceeding the ask quote on another market implies a riskless profit. However, if the transfer of the asset is time-consuming, the arbitrageur faces price risk on the sell-side market. In this case, the risk averse arbitrageur only exploits price differences if the instantaneous profit compensates for price risk, which is increasing in the expected latency and latency uncertainty. Both aspects increase the variance of the arbitrageur's expected return distribution and imply boundaries below which she does not exploit price differences.

Under the assumption that the underlying price process follows a Brownian motion with locally constant drift and spot volatility, the arbitrageur's expected return follows a general class of normal variance-mean mixture distributions. These distributions are fully described through their characteristic function, which allows us to provide an exact representation of the arbitrageur's certainty equivalent of exploiting cross-market price differences. This representation holds for arbitrary concave utility functions and a general class of latency distributions. Given the certainty equivalent, the arbitrage boundary is then derived as the minimum instantaneous return required to make the arbitrageur indifferent between trading and staying idle. We also extend our framework to account for transaction costs in the form of trading fees and market impact. Moreover, we discuss the implications of so-called settlement fees that are chosen individually and allow the arbitrageur to provide incentives for a reduction of the expected latency. Latency-reducing fees typically exist in blockchain-based systems, but do not change the central insights of our framework.

To obtain expressions that convey economic intuition, we equip the arbitrageur with an explicit utility function. We then can cast the arbitrage boundaries in terms of the spot volatility, the moments of a general latency distribution, and the arbitrageur's risk aversion. While the special case of a utility function with constant *absolute* risk aversion provides analytically tractable solutions, we also derive arbitrage boundaries for the commonly used class of isoelastic utility functions with constant *relative* risk aversion.[3] For both cases, we show that the arbitrage boundary increases with (i) the arbitrageur's risk aversion, (ii) the local volatility on the sell-side market, (iii) the expected waiting time until settlement and (iv) the expected variance of this waiting time.

Our second major contribution is the estimation of arbitrage boundaries for the Bitcoin market. We collect minute-level price information from 17 exchanges that feature

[3]See, e.g., Balduzzi and Lynch (1999) and Chetty (2006) for applications of isoelastic utility in a financial context.

trading Bitcoin against US Dollar. Furthermore, we gather real-time information about the Bitcoin network, providing us with, among other things, the time it takes for every transaction from entering the Bitcoin network until its inclusion in the blockchain. This procedure yields a unique database providing a complete picture of Bitcoin transaction activities including settlement latencies and other network characteristics.

The estimation of arbitrage boundaries rests on three ingredients. First, we estimate the spot volatility using minute-level midquote data employing a kernel estimator in line with Kristensen (2010). Second, we parametrize the latency distribution according to a conditional gamma distribution, depending on network and transaction-specific characteristics that affect the settlement time. Both the spot volatility and the moments of the latency distribution are estimated based on rolling windows to rule out any look-ahead bias and to mimic the viewpoint of an arbitrageur who aims to predict her certainty equivalent based on contemporaneously available information. Third, in line with existing literature, we choose an isoelastic utility function with exogenously given coefficient of relative risk aversion.

Accounting solely for stochastic latency and assuming a relative risk aversion of 2, the average estimated arbitrage boundary amounts to 124 bp.[4] We find that 88% of all observed instantaneous price differences across markets fall into these boundaries. Adjusting additionally for transaction costs, the boundaries contain even up to 98% of the observed price differences. The few price differences exceeding the boundaries might be due to the presence of arbitrageurs with higher risk aversion or additional market frictions.[5] In fact, we show that the average implied relative risk aversion necessary to capture *all* observed price differences, amounts to 12.

In addition, we provide deeper insights into the main components determining the

[4]See Conine et al. (2017) for a review of the empirical literature on the estimation of the coefficient of relative risk aversion.

[5]These might be, for instance, information frictions (Mitchell et al., 2002), short sale constraints (Lamont and Thaler, 2003a,b; de Jong et al., 2009), or slow moving capital (Roll et al., 2007; Mitchell et al., 2007).

arbitrage boundaries. For instance, the fact that the latency is *stochastic* (and thus not perfectly predictable) is responsible for, on average, around 43% of the arbitrage boundaries. Moreover, to increase the security and to guarantee the reliability of transactions, exchanges typically consider a transaction as valid only if it is confirmed by additional subsequent blocks after it is included in the ledger for the first time. These requirements aim at protecting the system from potential double-spending attacks yielding fraudulent transactions. In our sample, exchanges require up to five additional confirmations after a transaction is included in the block for the first time. We find that this security component accounts for on average about 55% of the arbitrage boundaries. To quantify the trade-off between the number of confirmations and the resulting increase in market frictions (in terms of arbitrage boundaries), we show that the requirement of 10 confirmations increases the average arbitrage bound by more than 24%.[6]

The promise of fast and low-cost transaction settlement leads central banks and marketplaces to actively explore potential applications of distributed ledgers for transaction settlement (e.g., BIS, 2017; NASDAQ, 2017; ECB and BoJ, 2018; SIX, 2018). Our results, however, show that distributed settlement, in particular under sufficiently high security standards, is time-consuming and implies costs in form of non-trivial market frictions. These frictions impair market efficiency and substantially distort the law of one price. Our results shed some light on these costs (in terms of arbitrage boundaries) caused by the individual features underlying the settlement system and their stochastic nature.

Our theoretical results apply to all markets with stochastic settlement latency. We thus contribute to the literature on limits to arbitrage and market efficiency (e.g., Fama, 1965; de Long et al., 1990; Shleifer and Vishny, 1997) by highlighting a novel friction that impedes arbitrageurs' ability to exploit mispricing.[7] Our framework is also applicable

[6]The requirement of 10 confirmation reduces the likelihood of a successful attack to less than 5% if the attacker controls 10% of the total available computing power.

[7]We refer to Gromb and Vayanos (2010) for an extensive survey of the theoretical literature on the limits of arbitrage.

to settings with (stochastic) latency in order execution and its implications for market making (e.g., Budish et al., 2015; Menkveld and Zoican, 2017; Foucault et al., 2017).

Moreover, we contribute to the recent literature that examines price formation in Bitcoin markets. For instance, Gandal et al. (2018) demonstrate that suspicious trading activity from tradebots likely causes temporary price spikes. Similarly, Griffin and Shams (2018) find evidence for price manipulation on Bitcoin markets through the creation of tokens pegged to the Dollar. Such price movements could lead to substantial price differences across markets that remain unexploited if they coincide with periods of high stochastic latency. Our results also complement the findings of Makarow and Schoar (2018) who show that a common component explains a high share of observed price differences.

Finally, we contribute to the emerging literature on the economic impact of blockchain technology. Malinova and Park (2017) analyze the effect of various degrees of transparency on liquidity traders' strategies, while Khapko and Zoican (2017) show how flexible settlement may interact with search frictions and counterparty risk. Cong and He (2017) investigate how the informational environment for contracting might change due to decentralized consensus mechanisms. Furthermore, Biais et al. (2017) and Saleh (2018) analyze the equilibrium properties of different blockchain-based consensus protocols. Abadie and Brunnermeier (2018) discuss fundamental trade-offs of the migration to distributed ledgers with computationally expensive consensus protocols, while Chiu and Koeppl (2018) show that moving the US corporate debt market to blockchain-based settlement might create substantial welfare gains.

The remainder of the paper is organized in the following way. In Section 2, we theoretically derive arbitrage boundaries under general conditions. Section 3 presents the data, summary statistics stemming from the Bitcoin network and evidence on cross-market price differences. In Section 4, we provide estimates of arbitrage boundaries

and their main determinants, and quantify how much of the observed cross-market price differences are captured by these boundaries. Section 5 concludes.

2 Stochastic Latency and Limits to Arbitrage

2.1 Return distribution under stochastic latency

We consider an economy containing a single asset that is traded on two different markets b and s. For now, we abstract from short sales and other derivatives, but we discuss their relation to stochastic settlement latency further below. The trading activity on these markets is exogenously given and we assume that agents can continuously monitor the quotes of one unit of the asset across all markets.

Definition 1. Market $i \in \{b, s\}$ *continuously provides marginal buy quotes (asks)* A_t^i *and sell quotes (bids)* B_t^i *for a small unit of the asset, where* $B_t^i \leq A_t^i$*, at time* t*.*

We introduce transaction costs and the possibility to trade more than one unit of the asset below and show that these aspects do not affect our main insights. Our sole agent is an arbitrageur that aims at exploiting observed price differences across markets.

Definition 2. *The* arbitrageur *has unlimited capital, continuously monitors the quotes on markets* b *and* s *and considers the following strategy: if buy and sell quotes across markets imply a profit, she intends to buy one unit of the asset at the market with the lower buy quote, transfer the asset to the market with a higher sell quote and sell it as soon as the transfer is settled.*

The assumption of unlimited capital allows us to abstract from capital constraints that might impose additional limits to arbitrage (Shleifer and Vishny, 1997).

Without loss of generality, we focus on a scenario where the arbitrageur buys on market b and sells on market s. The converse case of selling on market b and buying

on market s can be handled analogously. Hence, in case of frictionless trading and no latency in settlement, the arbitrageur simply exploits observed price differences if

$$B_t^s - A_t^b > 0, \tag{1}$$

as she can buy the asset on market b at A_t^b, instantaneously transfer the asset to market s and sell it again at price B_t^s.

An instantaneous transfer is not possible, however, whenever the settlement of the transaction is time-consuming. Such a (possibly random) latency constitutes a fundamental element of distributed ledger systems that do not rely on central clearing entities.[8] It should not be confused, however, with latency in order *execution* as heavily discussed in the context of high-frequency trading (e.g., Hasbrouck and Saar, 2013; Foucault et al., 2017). Such latencies are in the order of milliseconds and thus are of several magnitudes smaller than settlement latencies. Therefore, we refrain from latency in order execution and assume that markets process orders instantaneously.

Definition 3. Latency τ *is the random waiting time until a transfer of the asset between markets is settled.*

If the buy transaction at market b takes place at time t and the transfer of the asset to market s is settled at $t + \tau$, the arbitrageur faces the sell quote $B_{t+\tau}^s$. The profit of the arbitrageur's trading decision is thus at risk, if the probability of losing money is non-zero, i.e., if

$$\mathbb{P}\left(B_{t+\tau}^s < A_t^b\right) > 0. \tag{2}$$

In this case, a risk averse arbitrageur would face limits to (statistical) arbitrage if the associated risk exceeds the expected returns (Bondarenko, 2003). To formalize the trading decision of the arbitrageur, denote the log prices as $a_t^b := \log\left(A_t^b\right)$ and $b_t^s := \log\left(B_t^s\right)$,

[8] See Appendix A for a detailed discussion of distributed ledger technologies.

respectively, to cast the payoff in log returns.

Definition 4. *The log* return *of buying on market b at time t and selling on market s at time $t + \tau$ is defined as*

$$r^{b,s}_{(t:t+\tau)} := b^s_{t+\tau} - a^b_t = \underbrace{\delta^{b,s}_t}_{\substack{\text{instantaneous} \\ \text{return}}} + \underbrace{b^s_{t+\tau} - b^s_t}_{\substack{\text{exposure to} \\ \text{price risk}}}, \tag{3}$$

where $\delta^{b,s}_t := b^s_t - a^b_t$.

The first part of the return decomposition contains the returns the arbitrageur would earn under instantaneous settlement, i.e., in the absence of any latency. The second part captures the risk of adverse price movements on the sell-side market. As the instantaneous return only depends on observed prices, the arbitrageur solely faces uncertainty about the evolution of prices on the sell-side market. The price process on the sell-side market is given as follows.

Assumption 1. *For a given latency τ, we model the log price changes on the sell-side $b^s_{t+\tau} - b^s_t$ as a Brownian motion with drift μ^s_t such that*

$$r^{b,s}_{(t:t+\tau)} = \delta^{b,s}_t + \tau\mu^s_t + \int_t^{t+\tau} \sigma^s_t dW^s_k, \tag{4}$$

where σ^s_t denotes the spot volatility of the bid quote process on market s, and W^s_k denotes a Wiener process. We assume that σ^s_t is constant over the interval $[t, t + \tau]$.[9]

The dynamics of the sell price thus expose the arbitrageur to uncertainty about her profits. The uncertainty is triggered by the spot volatility σ^s_t and the latency τ. We require only weak assumptions regarding the stochastic nature of the latency.

[9]Time-varying and stochastic volatility can be incorporated by means of a change of the time-scale underlying the Brownian motion. We provide the corresponding derivations in Appendix B.

Assumption 2. *The stochastic latency $\tau \in \mathbb{R}_+$ is a random variable equipped with a conditional probability distribution $\pi_t(\tau) := \pi(\tau|\mathcal{I}_t)$, where $\mathcal{I}_t$ denotes the set of available information at time t. We assume that the moment-generating function of $\pi_t(\tau)$, defined as $m_\tau(u) := \mathbb{E}_t(e^{u\tau})$ for $u \in \mathbb{R}$, is finite on an interval around zero.*

Assumptions 1 and 2 allow us to fully characterize the return distribution $\pi_t\left(r^{b,s}_{(t:t+\tau)}\right)$ through the interval from t to $t+\tau$ for a wide range of latency distributions.

Lemma 1. *Under Assumptions 1 and 2, the returns follow a normal variance-mean mixture with probability distribution*

$$\pi_t\left(r^{b,s}_{(t:t+\tau)}\right) = \int_{\mathbb{R}_+} \pi_t\left(r^{b,s}_{(t:t+\tau)}|\tau\right) \pi_t(\tau)\, d\tau, \tag{5}$$

and corresponding characteristic function[10]

$$\varphi_{r^{b,s}_{(t:t+\tau)}}(u) = e^{iu\delta_t^{b,s}} m_\tau\left(iu\mu_t^s - \frac{1}{2}u^2(\sigma_t^s)^2\right). \tag{6}$$

Proof. See Equation (2.2) in Barndorff-Nielsen et al. (1982) and thereafter. □

Lemma 1 characterizes the impact of stochastic latency on the return distribution. To provide an illustrative example, we parametrize the probability distribution of the stochastic latency as an exponential distribution with locally-constant scale parameter $\lambda_t := \lambda(\mathcal{I}_t)$. The probability density function of the latency is then given by

$$\pi_t(\tau) = \lambda_t e^{-\lambda_t \tau}, \tag{7}$$

with conditional mean $\mathbb{E}_t(\tau) = \lambda_t^{-1}$ and conditional variance $\mathbb{V}_t(\tau) = \lambda_t^{-2}$. The moment

[10]The characteristic function of a random variable provides an alternative way to fully describe the behavior and properties of the probability distribution of the random variable. More specifically, the characteristic function $\varphi_X(u)$ of a random variable X is defined as $\varphi_X(u) = \mathbb{E}(e^{iuX})$, where i is the imaginary unit and $u \in \mathbb{R}$ is the argument of the characteristic function.

generating function of the exponential distribution is $m_\tau(u) = \left(1 - \lambda_t^{-1} u\right)^{-1}$. Thus, Lemma 1 yields

$$\varphi_{r^{b,s}_{(t:t+\tau)}}(u) = \frac{e^{iu\delta_t^{b,s}}}{1 - i\frac{\mu_t^s}{\lambda_t}u + \frac{(\sigma_t^s)^2}{2\lambda_t}u^2}, \tag{8}$$

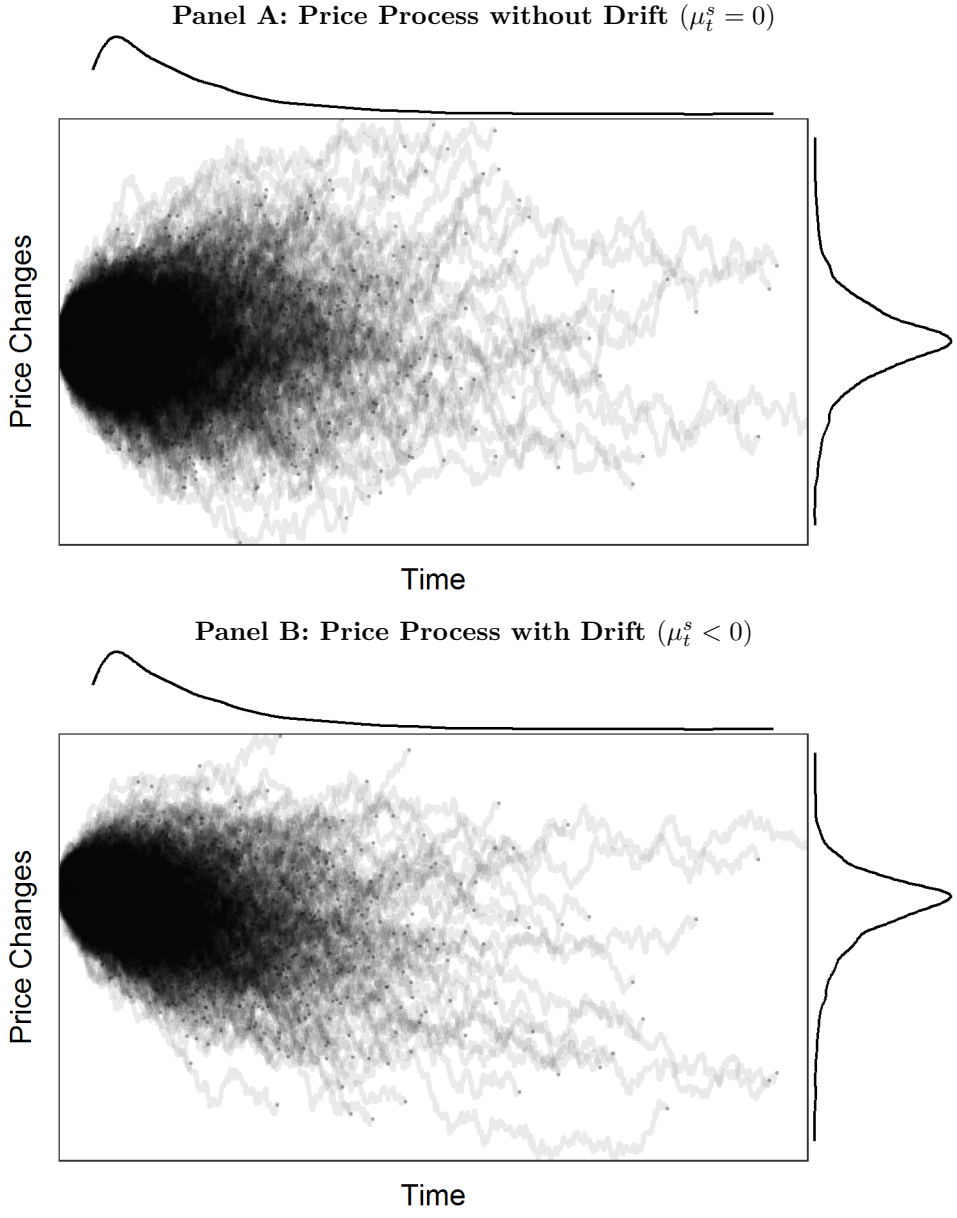
which corresponds to the characteristic function of an asymmetric Laplace distribution with $\mathbb{E}_t\left(r^{b,s}_{(t:t+\tau)}\right) = \delta_t^{b,s} + \frac{\mu_t^s}{\lambda_t}$ and $\mathbb{V}_t\left(r^{b,s}_{(t:t+\tau)}\right) = \frac{1}{\lambda_t}\left((\mu_t^s)^2 + (\sigma_t^s)^2\right)$ (e.g., Kotz et al., 2012). Without a drift ($\mu_t^s = 0$), the distribution collapses to a symmetric Laplace distribution with location parameter $\delta_t^{b,s}$, scale parameter $\frac{\sigma_t^s}{\sqrt{2\lambda_t}}$, and corresponding probability density function

$$\pi_t\left(r^{b,s}_{(t:t+\tau)}\right) = \frac{\sqrt{2\lambda_t}}{2\sigma_t^s}\exp\left(-\frac{\sqrt{2\lambda_t}}{\sigma_t^s}\left|r^{b,s}_{(t:t+\tau)} - \delta_t^{b,s}\right|\right), \tag{9}$$

with $\mathbb{E}_t\left(r^{b,s}_{(t:t+\tau)}\right) = \delta_t^{b,s}$ and $\mathbb{V}_t\left(r^{b,s}_{(t:t+\tau)}\right) = (\sigma_t^s)^2\,\mathbb{E}_t(\tau)$. Hence, not surprisingly, in the absence of a drift in the underlying Brownian motion, the (conditionally) expected return implied by the arbitrage strategy is equal to the instantaneous return $\delta_t^{b,s} = b_t^s - a_t^b$. The (conditional) variance equals the (locally constant) spot variance on market s, $(\sigma_t^s)^2$, scaled by the (conditional) expected waiting time until the settlement of the transaction, λ_t^{-1}. Hence, the higher the volatility on the sell-side market or the longer the expected waiting time until the transfer is settled, the higher is the risk of extreme adverse price movements.

Figure 2 provides a graphical illustration of the resulting distributions. The plots show simulated draws from a Brownian motion stopped at randomly sampled waiting times. The marginal distribution at the top of each figure illustrates the exponential distribution of the waiting times. The marginal distribution on the right-hand side shows the resulting sampling distribution of the price process which converges in the limit to a Laplace distribution. Panel A shows the resulting distribution for a price process without drift, while the price process in Panel B includes a negative drift.

Figure 2: Distribution of Returns under Stochastic Latency.
This figure illustrates the impact of stochastic latency (horizontal axis) on the distribution of returns (vertical axis) if log prices follow a Brownian motion and if latencies are exponentially distributed. The individual paths correspond to sample draws of the price process and the dots correspond to the terminal value of the stopped Wiener process. The marginal distribution on the top corresponds to the sampled latencies. The marginal distribution on the right-hand side corresponds to the sampled distribution of returns which converges in the limit to a Laplace distribution. Panel A shows the resulting distribution for a price process without drift, while the price process in Panel B includes a negative drift.



2.2 Arbitrage boundaries for risk averse arbitrageurs

To quantify the arbitrageur's assessment of risk, we have to to equip her with a corresponding utility function.

Assumption 3. *The arbitrageur has an utility function $U_\gamma(r)$ with risk aversion parameter γ, where r are the log returns implied by her trading decision. Furthermore, assume $U'_\gamma(r) > 0$ and $U''_\gamma(r) < 0$.*

The arbitrageur maximizes the expected utility $\mathbb{E}_t\left(U_\gamma(r)\right)$, which we express in terms of the certainty equivalent (CE). The CE is the riskless return which yields the same expected utility as the risky payoff r, i.e., $U_\gamma(CE) = \mathbb{E}_t\left(U_\gamma(r)\right)$. We derive the CE of exploiting instantaneous cross-market price differences in the following theorem.

Theorem 1. *Under Assumptions 1 - 3, the certainty equivalent (CE) of trading is given by*

$$CE = \delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s + \sum_{k=2}^{\infty} \frac{U_\gamma^{(k)}\left(\delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s\right)}{k! U'_\gamma\left(\delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s\right)} \mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - \delta_t^{b,s} - \mathbb{E}_t(\tau)\mu_t^s\right)^k\right), \quad (10)$$

where $U_\gamma^{(k)}\left(\mu_r\right) := \frac{\partial^k}{\partial \mu_r^k} U_\gamma\left(\mu_r\right)$.

Proof. See Appendix C. □

Theorem 1 allows us to compare the utility of trading versus staying idle (which yields a riskless return of zero). Therefore, the arbitrageur is willing to exploit cross-market price differences if and only if the CE of trading given by Equation (10) is positive. A positive CE corresponds to a statistical arbitrage opportunity in the sense of positive *expected* risk-adjusted profits. Whenever the observed price differences $\delta_t^{b,s}$ are positive, but CE is negative, the arbitrageur does not exploit the arbitrage opportunity. In this case, although the trade would be profitable under instantaneous settlement, limits to

(statistical) arbitrage arise due to stochastic latency. Hence, the arbitrageur is indifferent between trading and staying idle if price differences $\delta_t^{b,s}$ imply $CE = 0$.

Definition 5. *We define the arbitrage boundary d_t^s as the minimum price difference necessary such that the arbitrageur prefers to trade if $\delta_t^{b,s} > d_t^s$. Formally, d_t^s is the maximum of zero and the (unique) root of*

$$F(d) = d + \mathbb{E}_t(\tau)\mu_t^s + \sum_{k=2}^{\infty} \frac{U_\gamma^{(k)}\left(d + \mathbb{E}_t(\tau)\mu_t^s\right)}{k! U_\gamma'\left(d + \mathbb{E}_t(\tau)\mu_t^s\right)} \mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - d - \mathbb{E}_t(\tau)\mu_t^s\right)^k\right).^{11} \quad (11)$$

Price differences below the arbitrage boundary might persist as the arbitrageur prefers not to trade in such a scenario. To provide a more intuitive representation of the arbitrage boundary, we now consider a special case by assuming that the arbitrageur is equipped with constant absolute risk aversion (CARA). In line with Schneider (2015), we ignore the impact of higher order moments above the fourth degree of the Taylor representation in Equation (10). These assumptions yield an analytically tractable formulation of the arbitrage bound.

Lemma 2. *If, in addition to Assumptions 1 and 2, the arbitrageur has an exponential utility function $U_\gamma(r) := \frac{1-e^{-\gamma r}}{\gamma}$ with risk aversion $\gamma > 0$, then the arbitrage boundary is*

$$\begin{aligned} d_t^s = &- \mathbb{E}_t(\tau)\,\mu_t^s + \frac{\gamma}{2}\left(\mathbb{V}_t(\tau)\left(\mu_t^s\right)^2 + \left(\sigma_t^s\right)^2 \mathbb{E}_t(\tau)\right) \\ &- \frac{\gamma^2}{6}\left(3\mu_t^s\left(\sigma_t^s\right)^2 \mathbb{V}_t(\tau) + \left(\mu_t^s\right)^3 \mathbb{E}_t\left(\left(\tau - \mathbb{E}_t(\tau)\right)^3\right)\right) \\ &+ \frac{\gamma^3}{24}\left(\left(\mu_t^s\right)^4 \mathbb{E}_t\left(\left(\tau - \mathbb{E}_t(\tau)\right)^4\right) + 6\left(\sigma_t^s\right)^2\left(\mu_t^s\right)^2\left(\mathbb{E}(\tau)^3 + \mathbb{E}_t\left(\tau^3\right) - 2\mathbb{E}_t(\tau)^2\right) + 3\mathbb{E}_t\left(\tau^2\right)\left(\sigma_t^s\right)^4\right). \end{aligned} \quad (12)$$

Proof. See Appendix C. □

[11] By definition of CE, $F(d) = U_\gamma^{-1}\left(\mathbb{E}_t\left(U_\gamma\left(d + \mu_t^s\tau + \int_t^{t+\tau} \sigma_t^s W_k^s\right)\right)\right)$. Since $U_\gamma'(r) > 0$, the expectation is increasing in d. Moreover, since $U_\gamma''(r) < 0$, the inverse $U_\gamma^{-1}(r) > 0$ is also strictly concave. Thus, $F(d)$ is strictly increasing and has a unique root.

In the absence of a drift ($\mu_t^s = 0$), the arbitrage boundary of Lemma 2 further simplifies to

$$d_t^s = \frac{\gamma}{2}(\sigma_t^s)^2 \mathbb{E}_t(\tau) + \frac{\gamma^3}{8}(\sigma_t^s)^4 \left(\mathbb{V}_t(\tau) + \mathbb{E}_t(\tau)^2\right). \tag{13}$$

Hence, the arbitrage boundary d_t^s positively depends on (i) the arbitrageur's risk aversion, (ii) the local volatility on the sell-side market s, (iii) the expected waiting time until settlement, and (iv) the variance of the waiting time, $\mathbb{V}_t(\tau)$. Moreover, note that the boundary does not depend on characteristics of the buy-side market.

The boundary crucially depends on the arbitrageur's risk aversion γ. In our setting, this risk aversion corresponds to the arbitrageur's attitude towards the risk of a single trade. Theoretically, an arbitrageur can reduce her risk by repeatedly exploiting cross-market price differences as a law of large numbers may lead to a vanishing variance of the arbitrageur's aggregate returns. While such a repeated strategy can make the arbitrageur becoming risk-neutral in the theoretical limit (by means of a law of large numbers), in practice, this case is equivalent to the assumption of a lower risk aversion. Then, Equation (13) still applies, but implies a contraction of the relevant boundaries.

Accordingly, for a risk neutral arbitrageur, we have $d_t^s = 0$ and she would exploit any positive price difference $\delta_t^{b,s} > 0$. In this case, any price differences between the two markets should be absorbed immediately. The empirical evidence for the Bitcoin market according to Figure 1, however, suggests that this is not the case. According to our model, the existence of persistent price differences between the two markets (which are obviously not traded away) indicates that the markets are populated by risk averse arbitrageurs who do not exploit price differences below the threshold d_t^s. We thus denote the interval $[0, d_t^s]$ as the *no-trade region* for selling on market s implied by stochastic latency and risk aversion.

The lower bound d_t^s is a fundamental pillar of markets with stochastic latency, as the implied costs of stochastic latency affect the entire action and contracting space of

market participants. The only possibility for the arbitrageur to circumvent the exposure to stochastic latency would be the ability to sell *instantaneously* at the more expensive market to lock in the price difference. This is, however, only possible if the arbitrageur already has an inventory of the asset on the expensive (sell) market, or if she can borrow the asset on that market. Then, she can instantaneously buy on the cheap market and sell on the expensive market without the need of waiting for the verification of the transaction.

The first alternative, however, bears considerable price risk. To be able to exploit instantaneous price differences whenever they arise, it is necessary to keep inventory on the sell-side market over longer periods. Such a strategy continuously exposes the arbitrageur to price risk until an arbitrage opportunity arises (without any guarantee that this event will ever occur), which makes this strategy riskier than the strategy described in Definition 2. Second, if the arbitrageur can hedge the price risk (e.g., through borrowing), she substitutes the price risk with the implied hedging costs. In a competitive market, however, potential lenders of the asset demand a compensation for the opportunity costs of exploiting the price difference themselves. The boundaries d_t^s should thus enter the price of the hedging instrument and diminish the attractiveness of the strategy.

The choice of CARA (exponential utility) provides convenient analytical results, however, the case of constant relative risk aversion (CRRA) is more realistic as it implies that decision-making is independent of the arbitrageur's initial wealth (Harvey et al., 2010). The following lemma yields the arbitrageur's optimal decision for a CRRA utility function.

Lemma 3. *If, in addition to Assumptions 1 and 2, the arbitrageur has an isoelastic utility function* $U_\gamma(r) := \frac{r^{1-\gamma}}{1-\gamma}$ *with risk aversion parameter* $\gamma > 1$*, then the arbitrage boundary for* $\mu_t^s = 0$ *is given by*

$$d_t^s = \frac{1}{2}\sigma_t^s \sqrt{\gamma \mathbb{E}_t(\tau) + \sqrt{\gamma^2 \mathbb{E}_t(\tau)^2 + 2\gamma(\gamma+1)(\gamma+2)\mathbb{E}_t(\tau^2)}}. \tag{14}$$

Proof. See Appendix C. □

The main insights of the simple CARA model remain unchanged: the higher the expected latency, the variance of the latency, or the risk aversion, the less likely is the arbitrageur to trade.

2.3 The effect of transaction costs

Most markets feature trading fees that agents pay upon the execution of a trade. For instance, traders frequently pay fees as a percentage of the trading volume when they execute trades on centralized exchanges. Similarly, broker-dealers usually charge markups for the execution of trades in over-the-counter markets. Moreover, markets typically exhibit limited supply in the form of price-quantity schedules that agents are willing to trade, possibly leading to substantial price impacts for large trading quantities. To incorporate trading fees and liquidity effects into our framework, we make the following assumption.

Assumption 4. *Trading the quantity $q \geq 0$ on market i exhibits proportional transaction costs such that the average per unit sell and buy quotes are*

$$B_t^i(q) = B_t^i \left(1 - \rho^{i,B}(q)\right) \tag{15}$$

$$A_t^i(q) = A_t^i \left(1 + \rho^{i,A}(q)\right), \tag{16}$$

with $\rho^{i,B}(q) \geq 0$ and $\rho^{i,A}(q) \geq 0$.

The presence of transaction costs changes the objective function of the arbitrageur who focuses on maximizing returns net of transaction costs defined as

$$\tilde{r}^{b,s}_{(t:t+\tau)} = b^s_{t+\tau} - b^s_t + \delta^{b,s}_t - \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right) = r^{b,s}_{(t:t+\tau)} - \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right). \tag{17}$$

Due to the concavity of the utility function, transaction costs decrease the expected utility of the arbitrageur. A different interpretation of Equation (17) is that transaction costs only increase the instantaneous return required to make the arbitrageur indifferent between trading and staying idle. The following lemma summarizes the arbitrageurs' decision problem in the presence of transaction costs.

Lemma 4. *Under assumptions 1 - 4, the arbitrageur prefers to trade a quantity $q > 0$ over staying idle if*

$$\delta_t^{b,s} - \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right) > d_t^s. \tag{18}$$

Proof. See Appendix C. □

2.4 The effect of latency-reducing settlement fees

In distributed ledger systems, validators typically receive a reward for confirming transactions. This reward (at least partly) comprises of fees that originators of transactions offer to potential validators. Since the information that can be added to the ledger at any point in time is usually limited, such fees aim to provide validators with incentives to prioritize the settlement of transactions that include a higher fee (Easley et al., 2017). In particular, by offering a higher fee, arbitrageurs can thus decrease the settlement latency they face. We extend our framework to incorporate such latency-reducing *settlement fees* as follows.

Assumption 5. *A settlement fee $f > 0$ implies a latency distribution $\pi_t(\tau|f)$ that can be ordered in the sense that for $\tilde{f} > f$, $\pi_t(\tau|f)$ first-order stochastically dominates $\pi_t\left(\tau|\tilde{f}\right)$, i.e., $\mathbb{P}\left(\tau \leq x|\tilde{f}\right) > \mathbb{P}(\tau \leq x|f)$ for all $x \in \mathbb{R}_+$.*

The ordering of latency distributions in Assumption 5 implies a lower CE of trading for $\tilde{f} > f$ (e.g., Hadar and Russell, 1969; Levy, 1992). Denote by $d_t^s(f)$ the arbitrage boundary associated with the latency distribution $\pi_t(\tau|f)$. Theorem 1 then implies that

$d_t^s(f) > d_t^s(\tilde{f})$, i.e., by paying a higher settlement fee, the arbitrageur can reduce the risk associated with stochastic latency and becomes more likely to trade. For simplicity, we assume that $d_t^s(f)$ is differentiable such that Assumption 5 implies $\frac{\partial}{\partial f} d_t^s(f) < 0$.

While settlement fees reduce the latency, they are costly for the arbitrageur. Since the arbitrageur does not hold inventory of the asset on the buy-side market, she has to acquire the additional quantity f to spend it in the settlement process. We thereby assume that the arbitrageur has to pay the settlement fee in terms of the underlying asset. Given the transaction costs from the previous section, the choice of f thus also affects the trading quantity q. The following lemma summarizes the arbitrageur's decision problem in the presence of transaction costs and settlement fees.

Lemma 5. *Under assumptions 1 - 5, the arbitrageur prefers to trade a quantity $q > 0$ and pay a settlement fee $f > 0$ over staying idle if*

$$\delta_t^{b,s} - \log\left(\frac{1+\rho^{b,A}(q+f)}{1-\rho^{s,B}(q)}\right) > d_t^s(f). \tag{19}$$

Proof. See Appendix C. □

2.5 Optimal choice of trading quantities and settlement fees

While trading a larger quantity might deliver higher total returns, it comes at the cost of higher transaction costs on both the buy-side and sell-side market. Moreover, paying higher settlement fees leads to lower arbitrage boundaries, but at the cost of additional transaction costs on the sell-side market. The arbitrageur's trading decision thus features a trade-off between q and f with endogenous arbitrage boundaries. Formally, the arbitrageur aims to maximize total returns

$$\max_{\{q,f\}\in\mathbb{R}_+^2} B_t^s\left(1-\rho^{s,B}(q)\right)q - A_t^b(1+\rho^{b,A}(q+f))(q+f) \tag{20}$$

subject to the constraint

$$\delta_t^{b,s} - \log\left(\frac{1+\rho^{b,A}(q+f)}{1-\rho^{s,B}(q)}\right) \geq d_t^s(f). \tag{21}$$

We characterize the arbitrageur's optimal choice of trading quantities and settlement fees in the following lemma.

Lemma 6. *A total return maximizing arbitrageur chooses trading quantities $q^* > 0$ and settlement fees $f^* \geq 0$ such that*

$$\delta_t^{b,s} - \log\left(\frac{1+\rho^{b,A}(q^*+f^*)}{1-\rho^{s,B}(q^*)}\right) = d_t^s(f^*). \tag{22}$$

Moreover, a total return maximizing arbitrageur only pays a settlement fee $f^* > 0$ to trade a quantity $q^* > 0$ if the following necessary conditions are met:

$$\frac{1-\rho^{s,B}(q^*)}{q^*} > \rho^{s,B'}(q^*) \tag{23}$$

$$-\frac{\partial}{\partial f} d_t^s(f^*) > \frac{\rho^{s,B'}(q^*)}{1+\rho^{s,B}(q^*)}. \tag{24}$$

Otherwise, the arbitrageur optimally sets $f^* = 0$.

Proof. See Appendix C. □

The first part of the lemma states that the arbitrageur always chooses trading quantities and settlement fees such that the constraint in Equation (21) is binding. If the constraint would not be binding, then the arbitrageur could either trade a larger quantity or pay a lower settlement fee to increase her objective.

The second part of the lemma provides conditions for the choice of the settlement fee. According to Equation (23) the arbitrageur increases the settlement fee as long as the marginal price impact is below the average price impact. However, Equation (24)

shows that the reduction of the arbitrage bound through a higher settlement fee must exceed the implied opportunity costs, i.e., the possible gain in selling a higher quantity. Consequently, the arbitrageur tends to pay a higher settlement fee if the sell-side market is very liquid (keeping the marginal price impact low) and the settlement fee has a high impact on the boundary (i.e., reducing the latency and thus risk). If these conditions are violated, the arbitrageur optimally chooses not to pay any settlement fee, but might still decide to trade.

3 Bitcoin Network and Orderbook Data

We utilize data from the Bitcoin network, which constitutes the most popular decentralized protocol since Nakamoto (2008) published the concept and the underlying code. As of April 2018, Bitcoin (BTC) can be traded against fiat currencies and other cryptocurrencies on more than 400 markets that differ substantially in terms of location, regulation, security, technology, and fee structure. Bitcoin is continuously traded with an average daily trading volume of more than $4.5B as of April 2018.

To estimate the arbitrage boundaries according to Lemma 3, we need to estimate the local volatility of the price process as well as the moments of the latency distribution. For the former, we collect minute-level order book information, i.e., the list of all open buy and sell quotes, as well as exchange-specific trading fees from several trading venues that feature trading BTC against USD. To characterize the distribution of latencies, we collect real-time information from the Bitcoin network.

3.1 Bitcoin orderbook data

We gather orderbook information from the public application programming interfaces (APIs) of the 17 largest cryptocurrency exchanges that feature BTC versus USD trad-

ing.[12] We retrieve all open buy and sell orders for the first 25 levels on a minute level since April 2018. The granularity of our data yields detailed information on orderbook depth.[13]

Table 1 lists the corresponding exchanges and provides summary statistics of the underlying orderbook data based on our sample ranging from April 1, 2018 until September 30, 2018. The number of observations varies across exchanges, as the APIs can be unstable, i.e., they temporarily do not send responses to requests. During such periods, trading was not possible anyway and thus the missing data does not affect our analysis.

We observe a strong heterogeneity with respect to exchange-specific liquidity. For instance, whereas investors could have traded BTC versus USD at *Coinbase Pro* with an average spread of 0.05 USD, the average quoted spread at *Gatecoin* was about 150 USD since April 2018. For most exchanges, however, the relative bid-ask spreads are comparable to those from equity markets such as NASDAQ or NYSE, where relative spreads range from 5 basis points (bp) for large firms to 38 bp for small firms (Brogaard et al., 2014). To illustrate the importance of the orderbook depth, we also compute the change in marginal prices for a given trading quantity. For instance, buying one BTC at *Lykke* increases the marginal price on average by 37 bp. Selling 10 BTC at *Bitfinex* decreases the marginal price on average by 4 bp.

The exchanges also exhibit substantial heterogeneity in terms of trading-related characteristics. We therefore collect information on trading fees (i.e., taker fees) that directly feed into the price differences. Taker fees range from 0% on *Lykke* to 1% on *Gemini*. Furthermore, exchanges have different requirements with respect to the number of block confirmations before they proceed to process BTC deposits. For instance, *Kraken* requires 6 confirmations, i.e., incoming transactions must be included in at least 6 blocks.

[12]Some exchanges do not feature fiat currencies. However, they allow trading BTC against Tether, a token that is backed by one USD for each token and trading close to par with USD.

[13]To the best of our knowledge, none of these exchanges offers the opportunity to place hidden orders such that our data set indeed reflects a real-time image of the available liquidity at the distinct exchanges.

Table 1: Descriptive Statistics of Orderbook Data.

This table reports descriptive statistics of the orderbook data we employ in our empirical analysis. We gather high-frequency orderbook information of 17 exchanges by accessing their public application programming interfaces (APIs) on a minute level. *Orderbooks* denotes the number of successfully retrieved orderbooks between April 1, 2018 until September 30, 2018. *Spread* is the average quoted spread in USD, *Spread (in bp)* is the average spread relative to the quoted ask price. *Ask Depth* and *Bid Depth* correspond to the average percentage price change (in bp) relative to the best buy/sell price when buying 1 (10) BTC. We compute the latter as $\text{Ask Depth}^i(k) := 10,000/T \sum_{t=1}^{T}(A_t^i(k)/A_t^i - 1)$, where $A_t^i(k)$ corresponds to the price (in USD) of buying k BTC on market i at time t (analogously for $\text{Bid Depth}^i(k)$). *Taker Fee* are the associated trading fees in percentage points relative to the trading volume. *Conf.* refers to the number of blocks that the exchange requires to consider incoming transactions as being valid. Empty cells indicate missing values.

	Orderbooks	Spread (USD)	Spread (bp)	Ask Depth (1)	Ask Depth (10)	Bid Depth (1)	Bid Depth (10)	Taker Fee	Conf.	Company Location
Binance	253,690	2.44	3.37	1.99	9.69	2.01	10.03	0.10	2	Tokyo, Japan
Bitfinex	253,122	0.24	0.32	0.52	4.16	0.49	3.92	0.20	3	Central, Hong Kong
bitFlyer	252,983	14.91	20.28	14.04	99.68	26.89	139.75	0.15		Tokyo, Japan
Bitstamp	252,151	4.37	6.00	2.29	10.04	2.13	9.82	0.25	3	London, UK
Bittrex	253,563	12.29	17.04	13.20	43.40	12.63	43.15	0.25	2	Las Vegas NV, USA
BTCC	108,871	110.87	146.26	131.46	482.29	148.13	1,422.00	0.10	2	Shanghai, China
CEX.IO	251,951	12.03	16.52	7.11	42.89	6.74	41.92	0.25	3	London, UK
Gate	253,811	53.68	71.02	18.29	85.16	7.93	63.27	0.20	2	Sparta NJ, USA
Gatecoin	246,099	149.27	189.51	334.22	17,180.13	187.02	1,645.68	0.35	6	Wanchai, Hong Kong
Coinbase Pro	252,915	0.05	0.07	0.92	4.24	0.94	5.61	0.30	3	San Francisco CA, USA
Gemini	251,223	2.01	2.82	2.16	9.64	2.52	11.24	1.00	3	New York NY, USA
HitBTC	253,205	3.00	3.98	2.63	15.77	2.43	15.61	0.10	2	Hong Kong
Kraken	253,571	2.36	3.23	2.16	10.15	2.16	10.41	0.26	6	San Francisco CA, USA
Liqui	252,209	33.78	47.01	83.90	171.89	55.95	102.06	0.25		Kiev, Ukraine
Lykke	254,793	29.20	37.61	36.58	151.91	20.51	51.86	0.00	3	Zug, Switzerland
Poloniex	254,000	6.43	8.70	4.24	18.44	4.84	19.84	0.20	1	Wilmington DE, USA
xBTCe	238,057	4.54	6.17	1.01	5.84	0.90	5.58	0.25	3	Charlestown, Nevis

The objective of these requirements is to reduce the possibility of an attack that aims at revoking previous transactions (i.e., a so-called 'double-spending attack'). In such a case, a potential attacker has to alter all blocks containing the corresponding transaction. The probability that an attacker catches up with the honest chain decreases exponentially with the number of blocks the attacker has to alter. For instance, in the case of a confirmation requirement of 10 blocks, the probability of a successful attack is less than 0.01% (5%), if the attacker has a share of 30% (10%) of the total available computing power (Nakamoto, 2008). As we discuss below, these requirements confront arbitrageurs with an increase in the settlement latency.

3.2 Bitcoin network data

By running a full node[14] and gathering transaction-specific information in real time, we collect data on the status of the Bitcoin network on a minute level. The full node allows us to retrieve all transactions entering the Bitcoin network and waiting for verification. Each transaction contains a unique identifier, a timestamp of the initial announcement to the network, and, among other details, the fee (per byte) the initiator of the transaction offers validators to verify the transaction.[15]

All transactions enter the so-called *mempool* (short for memory pool) which is essentially a collection of all unconfirmed transactions. These transactions wait in the mempool until they are picked up by validators and get verified. The size of the mempool thus reflects the number of transactions that wait for confirmation. By design, however, the Bitcoin protocol restricts the number of transactions that can enter a single block. This

[14]Any computer that connects to the Bitcoin network is referred to as a *node*. *Full nodes* download every transaction and block and check them against specific consensus rules. If a transaction complies with the rules, it is distributed to validators. If a newly validated block complies to the rules, it is broadcast to other nodes.

[15]The fee per byte is more relevant than the total fee associated with a transaction as block sizes are limited in terms of bytes. In principle, a transaction can have multiple inputs and outputs, i.e., several addresses that are involved as senders or recipients of a transaction, which increases the number of bytes.

restriction induces competition among the originators of transactions who can offer higher *settlement fees*, as discussed in Section 2.4, to make it attractive for validators to include transactions in the next block. Consequently, transactions with no or very low settlement fees may not attract validators and thus stay in the mempool until they become verified eventually.

Validators bundle transactions that wait for verification and try to solve a computationally expensive problem which involves numerous trials until the first validator finds the solution. By design of the Bitcoin protocol, validators successfully find a solution and append a block on average every 10 minutes.[16] The system's protocol, however, also limits the number of transactions that can be included in a single block. Even though the expected block validation time is 10 minutes, it is uncertain when a transaction is included in a block for the first time. For any transaction this induces stochastic settlement latency. The probability of being included in the next block decreases, however, with the number of transactions that wait for settlement and increases with the settlement fee the investor is willing to pay.

Any transaction in the Bitcoin network has to go through the mempool, irrespective of its origin. The entire cross-section of transactions thus determines the latencies for the transfer of BTC across exchanges. Table 2 provides summary statistics of the recorded transactions. The average settlement fee is about 0.73 USD. The distribution of fees is highly skewed with a median of 0.15 USD. The average waiting time until the verification of a transaction is about 18 minutes, while the median is about 8.2 minutes. The time until verification, however, should not be confused with the time it takes until a new block is mined. Whereas, during our sample period, validators announce a new block to the network on average every 9.5 minutes, transactions may not be included in the subsequent block but instead have to wait longer until they get verified.

The Bitcoin network is prone to high fluctuation both in terms of usage and validators'

[16] We provide a more detailed discussion of the underlying consensus protocol in Appendix A.

Table 2: Descriptive Statistics of the Bitcoin Network.
This table reports descriptive statistics for Bitcoin transaction data from April 1, 2018, until September 30, 2018. The sample contains all transactions announced to the Bitcoin network. Transactions leave the mempool either because they are included in a block (validation) or after a certain waiting time (no validation). Our sample comprises of 22,572,426 transactions that are verified in 25,762 blocks. *Fee per Transaction* is the total settlement fee per transaction. *Fee per Byte* is the total fee per transaction divided by the size of the transaction in bytes where 100,000,000 Satoshi are 1 Bitcoin. We approximate the USD price by the average minute-level midquote across all exchanges in our sample. *Latency* is the time until the transaction is either validated or leaves the mempool without verification (in minutes). *Mempool Size* is the number of other transactions in the mempool at the time a transaction of our sample enters the mempool. *Block Validation Time* is the time (in minutes) between two consecutive block announcements to the Bitcoin network.

	Mean	SD	5 %	25 %	Median	75 %	95 %
Fee per Byte (Satoshi)	24.16	106.00	2.31	5.02	8.31	19.25	102.64
Fee per Transaction (USD)	0.73	7.26	0.04	0.08	0.15	0.38	2.78
Latency (Minutes)	18.13	41.97	0.88	3.55	8.28	17.57	58.30
Transaction Size (Byte)	474.93	2087.21	142.00	191.00	225.00	333.00	933.00
Mempool Size	2938.35	3474.47	193.00	779.00	1765.00	3685.00	9878.00
Block Validation Time (Minutes)	9.53	9.33	0.55	2.87	6.70	13.18	28.13

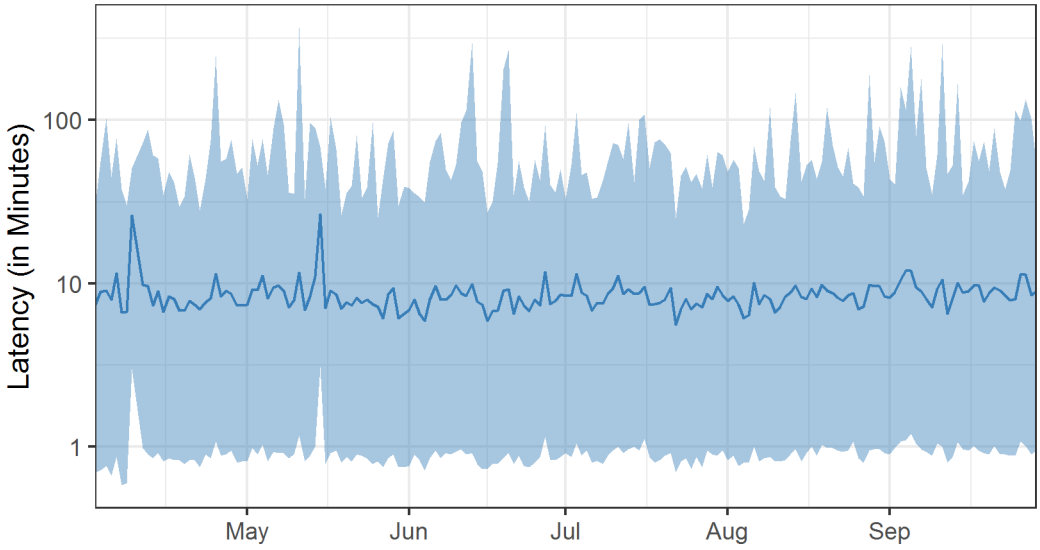
activity. Figure 3 illustrates the high variation of observed latencies. The daily median of all transactions verified on a particular day in our sample ranges from 5.6 to 26.4 minutes indicating considerable changes over time. The shaded area in Figure 3 shows the 5% and 95% daily quantiles, indicating large cross-sectional dispersion of latencies.

3.3 Price differences across markets

To provide systematic empirical evidence on the extent of (potential) arbitrage opportunities and thus violations of the law of one price, we compute the observed instantaneous cross-market price differences of all $N(N-1)/2$ exchange pairs, with the total number of exchanges $N = 17$, defined as

$$\Delta_t := \begin{pmatrix} 0 & \cdots & \delta_t^{N,1} \\ \vdots & \ddots & \vdots \\ \delta_t^{1,N} & \cdots & 0 \end{pmatrix} = \begin{pmatrix} 0 & \cdots & b_t^1 - a_t^N \\ \vdots & \ddots & \vdots \\ b_t^N - a_t^1 & \cdots & 0 \end{pmatrix}, \tag{25}$$

Figure 3: Distributions of Observed Latencies in the Bitcoin Network.
The figure illustrates the distribution of transaction settlement latency over time. The solid line corresponds to the median latency of all transactions verified on the particular day. The shaded area illustrates the corresponding 5% and 95% quantiles (in minutes). For illustrative purposes we use the log scale for the horizontal axis.



where b_t^i and a_t^i correspond to the (log) best bid and ask prices at exchange i and time t, respectively. These price differences are akin to a setting where arbitrageurs trade marginal quantities on each market. The effective price difference, however, depends on taker fees and the trading quantities which might create price impact by passing through higher orderbook levels. We thus define the transaction cost adjusted price differences as

$$\tilde{\Delta}_t := \begin{pmatrix} 0 & \cdots & \tilde{\delta}_t^{N,1} \\ \vdots & \ddots & \vdots \\ \tilde{\delta}_t^{1,N} & \cdots & 0 \end{pmatrix} = \begin{pmatrix} 0 & \cdots & \tilde{b}_t^1\left(q_t^{1,N}\right) - \tilde{a}_t^N\left(q_t^{1,N}\right) \\ \vdots & \ddots & \vdots \\ \tilde{b}_t^N\left(q_t^{N,1}\right) - \tilde{a}_t^1\left(q_t^{N,1}\right) & \cdots & 0 \end{pmatrix}, \tag{26}$$

where $\tilde{b}_t^i(q_t^{i,j})$ is the transaction cost adjusted (log) sell price of $q_t^{i,j}$ units of the asset on exchange i at time t and $\tilde{a}_t^i(q_t^{i,j})$ is the transaction cost adjusted (log) buy price of $q_t^{i,j}$ units of the asset.

Transaction costs are proportional to the trading quantity and correspond to our definition in Assumption 4. We choose $q_t^{i,j}$ as the quantity which maximizes the resulting return for the exchange pair i and j given the prevailing orderbooks at time t and the taker fees of exchanges i and j. If the limit orderbook would not be available, we would have to rely on a specific transaction cost function. In our setting, however, we avoid such assumptions and use the fully observed orderbook on a minute level. Accordingly, we account for proportional exchange-specific taker fees (as reported in Table 1), which increase the average buy price and decrease the average sell price. We then use the resulting shifted orderbook queues and apply a grid search algorithm to identify the trading quantity that maximizes the total return for each exchange pair. This data-driven approach thus just mimics the choice of an arbitrageur who aims at maximizing arbitrage profits by optimally accounting for the prevailing orderbook depth. As price differences obviously can only be positive for one direction, we set negative price differences to zero as (even without latency) such scenarios do not correspond to arbitrage opportunities. The resulting matrix of price differences thus contains only positive values.

Figure 4 visualizes the resulting average price differences for each exchange-pair. Panel A shows the price differences based on best bid and best ask according to Equation (25). The heatmap shows that some exchanges exhibit quotes that tend to deviate quite systematically from (nearly) all other exchanges. For instance, *CEX.IO*, *Gatecoin* and *HitBTC* quote on average higher bids than most other exchanges and thus exhibit large price differences when used as a sell-side market. Conversely, other exchange pairs do not feature large average price differences. For instance, there are hardly any price differences between *Gate* and other markets. Panel B of Figure 4 visualizes the average price differences adjusted for transaction costs, corresponding to elements of $\tilde{\Delta}_t$. The magnitudes of price differences decrease on average by 50 bp, indicating the relevance of transaction costs for the Bitcoin markets. However, price differences adjusted for transaction costs

are still of relevant magnitudes. Bid prices on *Gatecoin*, for instance, exceed most other exchanges on average by roughly 70 bp.

Figure 5 depicts the 25%, 50% and 75% quantiles of the minute-level price differences across all exchange pairs. We observe a substantial variation over time. The median price difference across all exchange pairs is on average 24 bp. The 25% (75%) quantile is on average 10 bp (67 bp), indicating a large dispersion of price differences in our sample period. The average price difference over all exchange pairs and over the whole sample period is 55 bp, indicating outliers that raise the mean considerably above the median. Panel B of Figure 5 depicts the corresponding price differences adjusted for transaction costs as of Equation (26). Interesting to note is the apparent trend towards quoted price differences which do not exhibit arbitrage opportunities when adjusting for transaction costs.

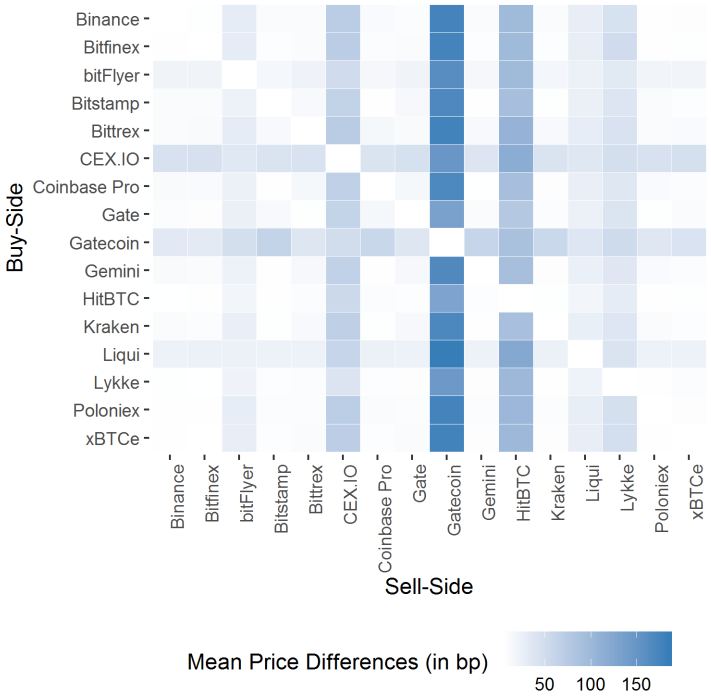
Figure 6 visualizes the dispersion of observed price differences adjusted for transaction costs for each month in our sample. The kernel estimates indicate that during our sample period, price differences clustered more around zero whereas the occurrence of extreme price differences decreased. These dynamics could be due to either increased competition among arbitrageurs or liquidity providers, or decreased arbitrage boundaries (e.g., because of lower risk aversion) of market participants.

4 Quantifying Arbitrage Boundaries in Bitcoin Markets

According to Lemma 3, the estimation of arbitrage boundaries at time t requires to predict the spot volatility $(\sigma_t^s)^2$, which we assume to be locally constant through the interval until the settlement of a transaction and the conditional mean and variance of the latency, $\mathbb{E}_t(\tau)$ and $\mathbb{V}_t(\tau)$.

Figure 4: Price Differences between Exchanges.
This heat map shows the mean price differences (in basis points) across time for each exchange pair in our sample. Price differences are based on minute-level best bid and best ask for each exchange. The darker the color, the higher the average price difference through our sample period in the specific exchange pair. White or very light colors indicate that there are on average no or few price differences for a specific exchange pair.

Panel A: Price Differences Based on Best Bid and Best Ask



Panel B: Price Differences Adjusted for Transaction Costs

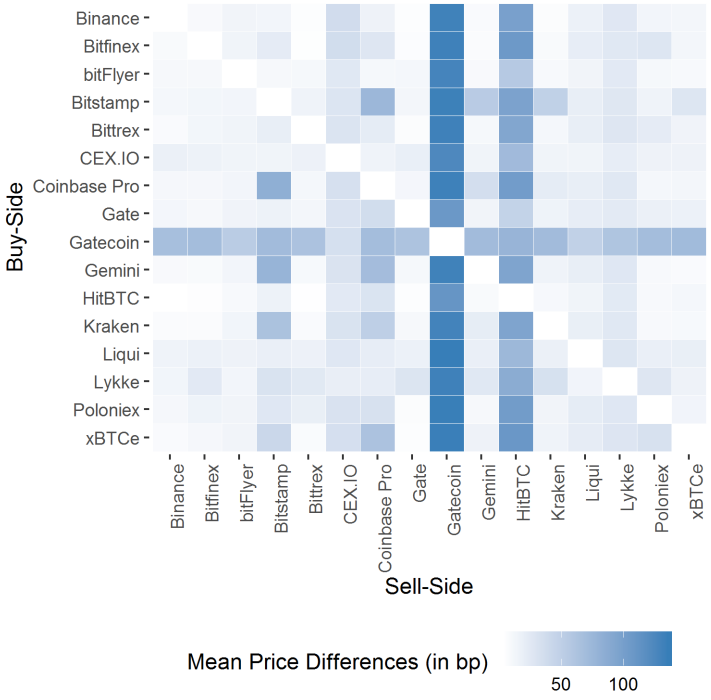
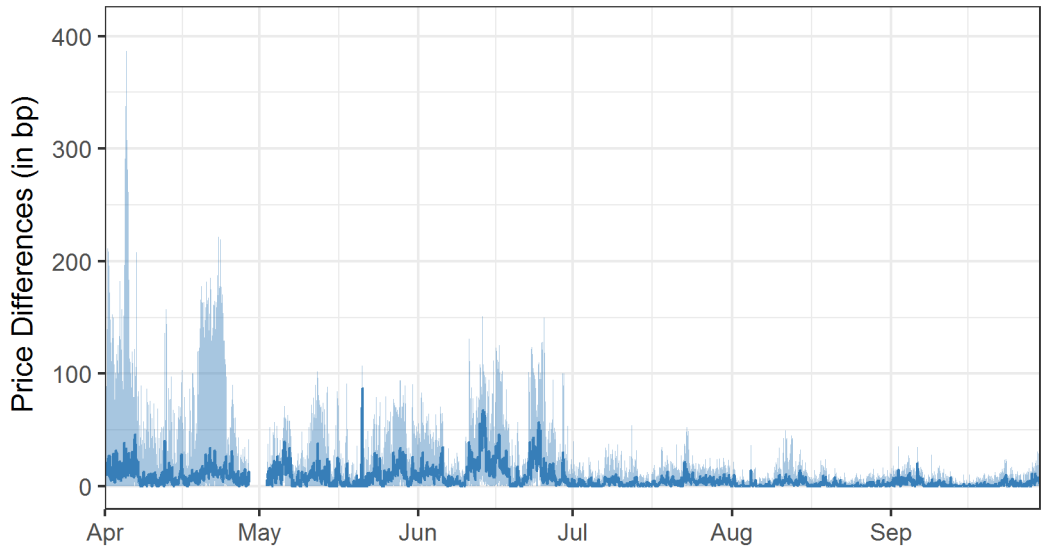


Figure 5: Price Differences over Time.

This figure shows the 25%, 50% and 75% quantiles of the price difference $\delta_t^{b,s}$ (in bp) across exchange pairs on hourly basis. Price differences are based on minute-level best bid and best ask for each exchange. Price differences adjusted transaction costs, $\tilde{\delta}_t^{b,s}$, are computed based on the transaction cost adjusted prices of trading $q_t^{b,s}$ units of BTC where $q_t^{b,s}$ is chosen as the quantity which maximizes the resulting return. We then aggregate minute-level price differences to hourly exchange-pair specific price differences and plot the 25%, 50% and 75% quantiles across all exchange pairs.

Panel A: Price Differences Based on Best Bid and Best Ask



Panel B: Price Differences Adjusted for Transaction Costs

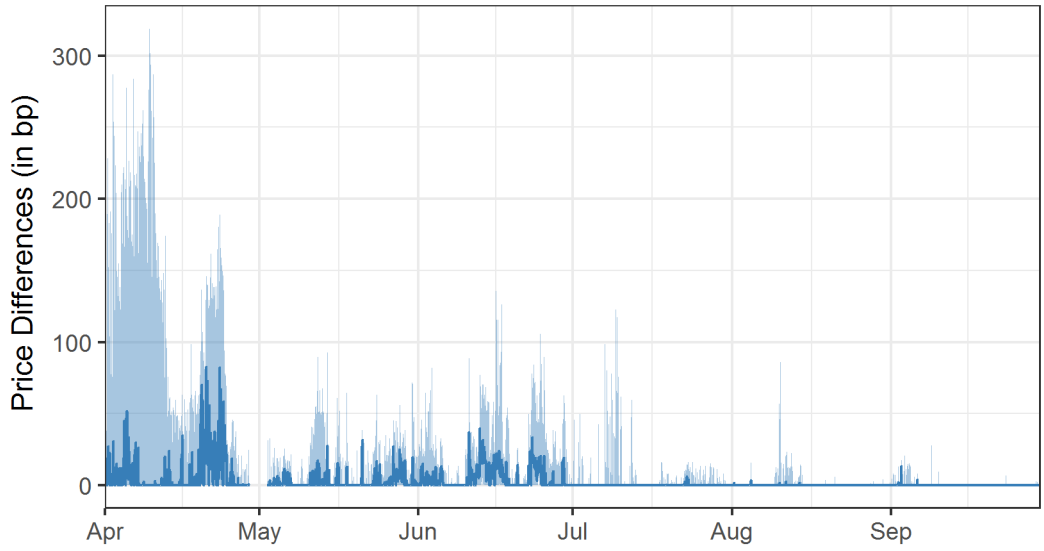
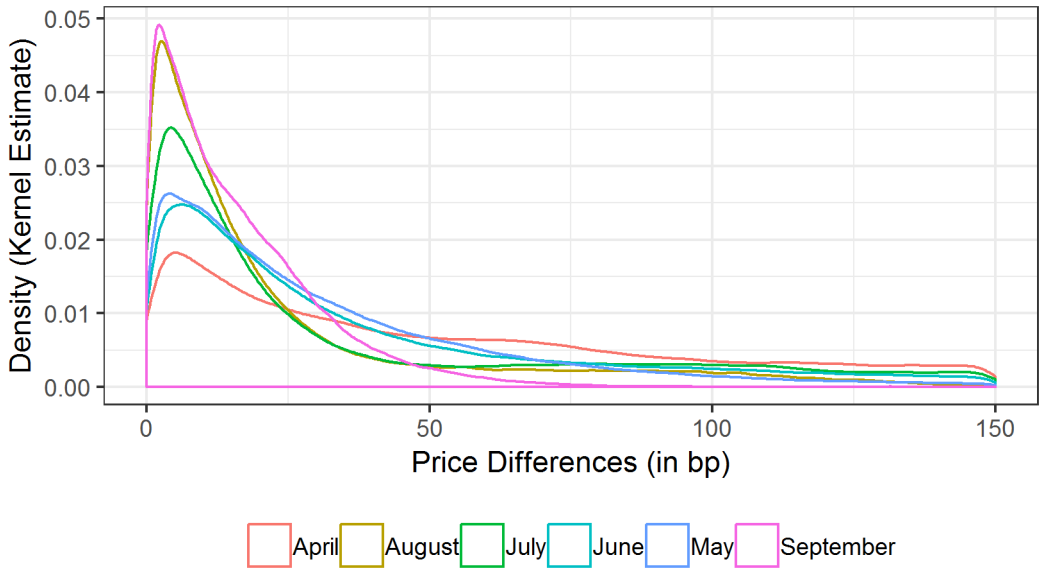


Figure 6: Price Differences over Time.
This figure shows the dispersion of observed price differences for each month in our sample *after* adjusting for transaction costs. The lines correspond to density estimates of all observed price differences (across all exchange-pairs) for the corresponding month.



4.1 Spot volatility prediction

To compute an estimate of the spot volatility $\widehat{(\sigma_t^s)}^2$, we follow the approach of Kristensen (2010). For each market s and time t, we estimate $(\sigma_t^s)^2$ by

$$\widehat{(\sigma_t^s)}^2(h) = \sum_{l=1}^{t} K(l-t,h)\left(b_l^s - b_{l-1}^s\right)^2, \tag{27}$$

where $K(l-t,h)$ is a one-sided Gaussian kernel smoother with bandwidth h. The choice of a one-sided kernel rules out any forward-looking bias, in the sense that only information up to t is utilized. The choice of the bandwidth h involves a trade-off between the variance and the bias of the estimator. Considering too many observations introduces a bias if the volatility is time-varying, whereas shrinking the estimation window through a lower bandwidth results in a higher variance of the estimator. Kristensen (2010) thus proposes

to choose h such that the Integrated Squared Error (ISE)

$$\widehat{\mathrm{ISE}}_T(h) = \sum_{i \in \{I_T\}} \left[\left(b_i^s - b_{i-1}^s\right)^2 - \widehat{\left(\sigma_{T,i}^s\right)}^2(h) \right]^2 \tag{28}$$

is minimized. Here, I_T refers to all observations on day T and $\widehat{\left(\sigma_{T,i}^s\right)}^2(h)$ is the spot variance estimator for timestamp i on day T based on bandwidth h. The optimal bandwidth on day $T+1$ is thus chosen as $h = \arg\min_{h>0} \widehat{\mathrm{ISE}}_T(h)$. The bandwidth choice implies that information on day T is used for the estimation on day $T+1$.

Table 3 summarizes the distributions of the exchange-specific time series of spot volatility estimates. We trim the distribution of all estimates at 1% on both tails to eliminate outliers (e.g., due to flickering quotes). Since the underlying is identical, as expected, the resulting estimates do not differ substantially across exchanges. The average minute-level volatility across exchanges is about 0.08%, which translates to a daily volatility of about 3%.[17] The average daily volatility of the S&P 500 index during the same period yields roughly 0.65%, thus Bitcoin represents a rather volatile asset. Given that the cross-sectional correlation between exchange-specific spot volatilities is high, Figure 7 displays the corresponding cross-market average on an hourly basis.

4.2 Latency prediction

Figure 8 plots the empirical distribution of the observed latencies at four randomly chosen dates. For each day, the measured latencies are clustered around 8 minutes, but reveal a strong skewness, indicating extreme outliers with substantially longer waiting times. Based on the shape of the empirical distribution, we parametrize the latency as being conditionally gamma distributed.

Accordingly, the conditional probability density function of transaction i with latency

[17] We convert minute-level estimates to the daily level by multiplying it with the square root of the number of minutes on any given trading day, i.e., $\sqrt{1440}$.

Table 3: Summary of Exchange-Specific Spot Volatility Estimates.
This table summarizes the estimation results of exchange-specific volatilities following Kristensen (2010). The estimates correspond to time-weighted average of squared price changes with a one-sided Gaussian kernel. We report all estimates in percent and on a minute-level.

	Mean	SD	5 %	25 %	Median	75 %	95 %
Binance	0.08	0.04	0.03	0.05	0.07	0.09	0.15
Bitfinex	0.07	0.04	0.02	0.04	0.06	0.09	0.15
bitFlyer	0.08	0.04	0.04	0.06	0.07	0.10	0.16
Bitstamp	0.08	0.04	0.03	0.05	0.07	0.09	0.15
Bittrex	0.11	0.05	0.05	0.09	0.11	0.14	0.20
BTCC	0.06	0.07	0.00	0.00	0.05	0.10	0.22
CEX.IO	0.08	0.04	0.03	0.05	0.07	0.10	0.15
Gate	0.08	0.04	0.03	0.05	0.07	0.10	0.15
Gatecoin	0.09	0.07	0.00	0.03	0.06	0.12	0.24
Coinbase Pro	0.07	0.04	0.02	0.04	0.06	0.08	0.14
Gemini	0.07	0.04	0.03	0.04	0.06	0.09	0.15
HitBTC	0.07	0.04	0.02	0.04	0.06	0.09	0.14
Kraken	0.07	0.04	0.02	0.04	0.06	0.09	0.14
Liqui	0.08	0.04	0.02	0.05	0.08	0.10	0.15
Lykke	0.07	0.05	0.02	0.04	0.06	0.10	0.16
Poloniex	0.08	0.04	0.03	0.05	0.07	0.09	0.15
xBTCe	0.07	0.05	0.00	0.04	0.06	0.09	0.15

Figure 7: Time Series of Spot Volatility Estimates.
This figure summarizes the cross-market average of minute-level volatilities following Kristensen (2010). The shaded areas correspond to the range across exchanges. The estimates correspond to time-weighted average of squared price changes with a one-sided Gaussian kernel. For each hour, we compute the average volatility across all exchanges. The zeros at the end of April and beginning of May result from missing data.

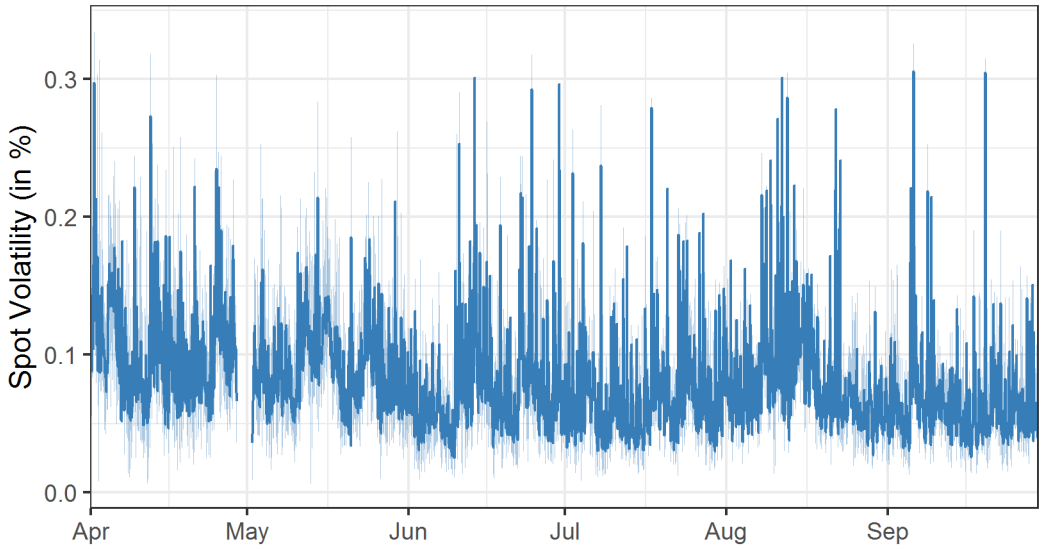
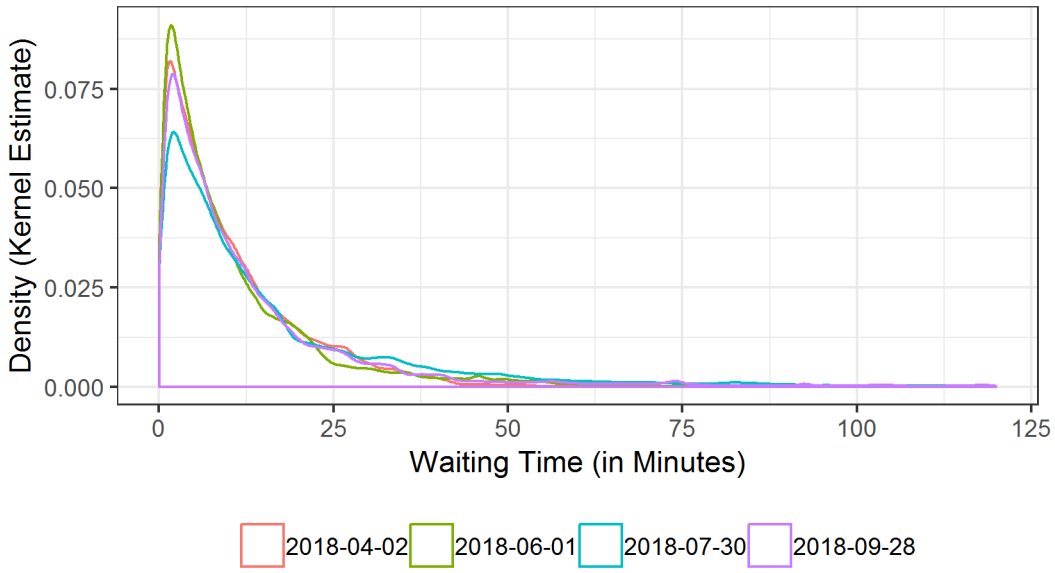


Figure 8: Distribution of Observed Latencies.
This figure illustrates the dispersion of the verification latency at four randomly chosen days. The lines correspond to density estimates of the latency of all transactions verified on the particular day.



τ_i with parameterized rate parameter β_i and shape parameter α_T is given by

$$\pi(\tau_i|\theta_T) = \frac{\beta_i^{\alpha_T}}{\Gamma(\alpha_T)} \tau_i^{\alpha_T - 1} e^{-\beta_i \tau_i}, \tag{29}$$

where

$$\beta_i = \exp(-x_i' \theta_T^{\beta}), \alpha_T > 0, \tag{30}$$

with x_i including an intercept and denoting (pre-determined) covariates driving τ_i, $\theta_T^{\beta} \in \mathbb{R}^K$ denoting the corresponding vector of parameters and $\Gamma(x) := \int_{\mathbb{R}_+} z^{x-1} e^{-z} dz$ being the Gamma function. The gamma distribution collapses to an exponential distribution for $\alpha_T = 1$. We estimate the parameter vector $\theta_T := (\theta_T^{\beta}, \alpha_T)'$ on a *daily* basis using the sample of all verified transactions $\{\tau_1, \ldots, \tau_m\}$ on a particular day T.

As covariates x_i, we include settlement fees as described in Section 3.2 and the size of the mempool. The number of transactions waiting for verification serves as a proxy for

competition among transactions and thus indicates the length of the queue of transactions waiting for verification. The settlement fees enter as *fees per byte* as the relevant metric for validators who face a restriction in terms of the maximum size of a block (in bytes). Both characteristics are available at the time of the trading decision.

We estimate the parameters $\hat{\theta}_T$ by maximum likelihood for each day in our sample, both with and without covariates. In addition, we estimate an exponential model by fixing $\alpha_T = 1$. In Table 4, we provide summary statistics of the estimated parameters. The numbers in the brackets denote the 5% and 95% of the time-series of estimated parameters. The covariates are statistically significant and have the expected sign for nearly all days, i.e., higher fees and lower mempool activity predict a lower latency. Likelihood ratio tests indicate that the covariates have joint explanatory power. We therefore find evidence that the waiting time until a transaction enters the next block of the blockchain is predictable. We moreover find that the exponential distribution is rejected in favor of the more general gamma distribution.

To avoid any look-ahead bias, we use the day-T parameter estimates, $\hat{\theta}_T$, to parametrize the conditional moments of the latency distribution on the subsequent day $T+1$. Accordingly, the (conditional) mean and variance of the latency τ induced by a transaction at time t on day $T+1$ is given by

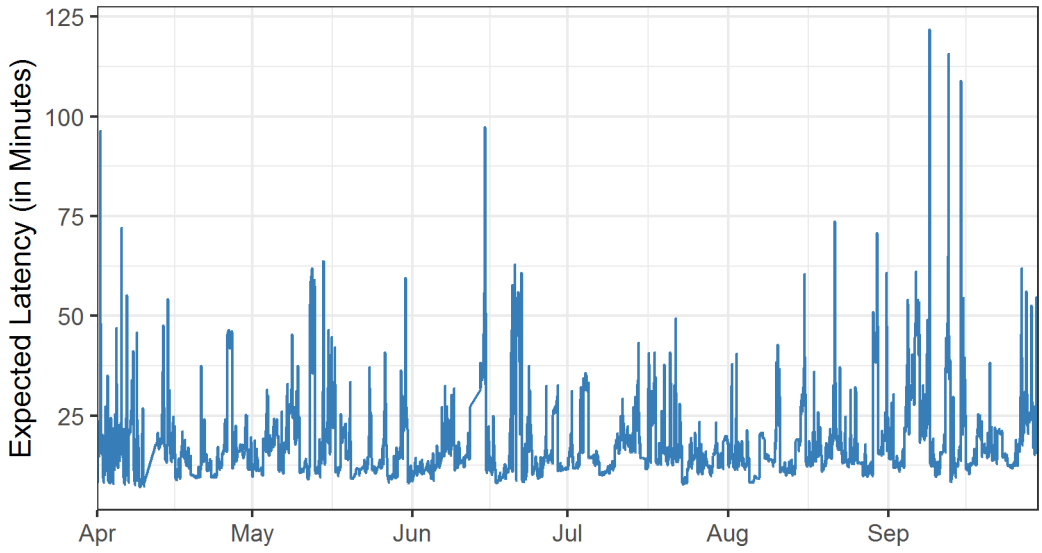
$$\widehat{\mathbb{E}}_t(\tau) = \hat{\alpha}_T \exp(x_t' \hat{\theta}_T^{\beta}), \qquad \text{and} \qquad \widehat{\mathbb{V}}_t(\tau) = \hat{\alpha}_T \exp(2x_t' \hat{\theta}_T^{\beta}). \tag{31}$$

As a proxy for the (individually chosen) settlement fees we use the fees recommended by the Bitcoin network. Intuitively, these recommendations are based on an algorithm that provides the lowest fee that during the recent history resulted in a high fraction of transactions that got validated in the next possible block. This fee, however, is higher than the (unobservable) 'optimal' fee an arbitrageur would choose if she would not only aim at minimizing the latency but also at maximizing her expected returns (net of transaction

Table 4: Parameter Estimates for the Duration Models.
This table reports summary statistics for the estimated parameters of the gamma duration model given by Equation (29). *Fee* denotes fee per byte and *Mempool Size* is the number of transactions in the mempool waiting for verification. The model is estimated on a daily basis, where the reported values denote the time series average of the estimated parameters. Values in brackets correspond to the 5% and 95% percent quantiles of the estimated parameters. *F-Test* denotes the fraction of days (in percent) at which the hypothesis $\theta_T = 0$ is rejected at the 95% percent level. *LR (Covariates)* rests on a likelihood ratio test against a model *without* covariates. *LR (Gamma vs. Exp.)* rests on a likelihood ratio test against the exponential specification. The reported value denotes the fraction of days (in percent) where the hypothesis that *the likelihood of the more general model equals the likelihood of the restricted model* is rejected at the 95% significance level.

	Exponential		Gamma	
	W Covariates	W/o Covariates	W Covariates	W/o Covariates
Intercept	2.75	2.79	2.9	2.98
	[2.26 , 3.45]	[2.29 , 3.53]	[2.08 , 4.04]	[2.21 , 4.19]
α	-	-	0.86	0.82
	-	-	[0.55, 1.16]	[0.51 , 1.12]
Fee per Byte	-0.04	-	-0.04	-
	[-0.1 , -0.01]	-	[-0.1 , -0.01]	-
Mempool Size	0.19	-	0.19	-
	[-0.03 , 0.49]	-	[-0.03 , 0.49]	-
F-test	100	-	99.52	-
LR (Covariates)	100	-	99.43	-
LR (Gamma vs.Exp.)	100	-	-	-

Figure 9: Expected Latency in the Bitcoin Network.
This figure summarizes the conditional expected latency. The estimates on day $T+1$ are based on the fitted parameter $\hat{\theta}_T^{\mathcal{G}}$ of the gamma model. We plot the average expected latency on an hourly basis through our sample period.



costs). Therefore, the expected arbitrage boundaries resulting from these recommended fees are smaller than in the case of lower ('optimal') settlement fees. Consequently, our estimates are conservative in the sense that they represent the smallest interval in which price differences might persist.

Figure 9 provides the corresponding predictions of latencies. The predictions vary considerably, mainly due to the dynamics of the mempool size. The average expected latency for the entire sample is 13.4 minutes. During certain periods, however, the expected latency exceeds 90 minutes, thus exposing the arbitrageur to substantial price risk if the spot volatility during these periods is high as well.

4.3 Estimates of arbitrage boundaries

In the empirically relevant CRRA case based on Lemma 3, the estimated arbitrage boundaries $\hat{d}_t^s$ are given by

$$\hat{d}_t^s = \frac{1}{2}\hat{\sigma}_t^s \sqrt{\gamma c_1 + \sqrt{\gamma^2 c_1^2 + 2\gamma(\gamma+1)(\gamma+2)c_2}}, \tag{32}$$

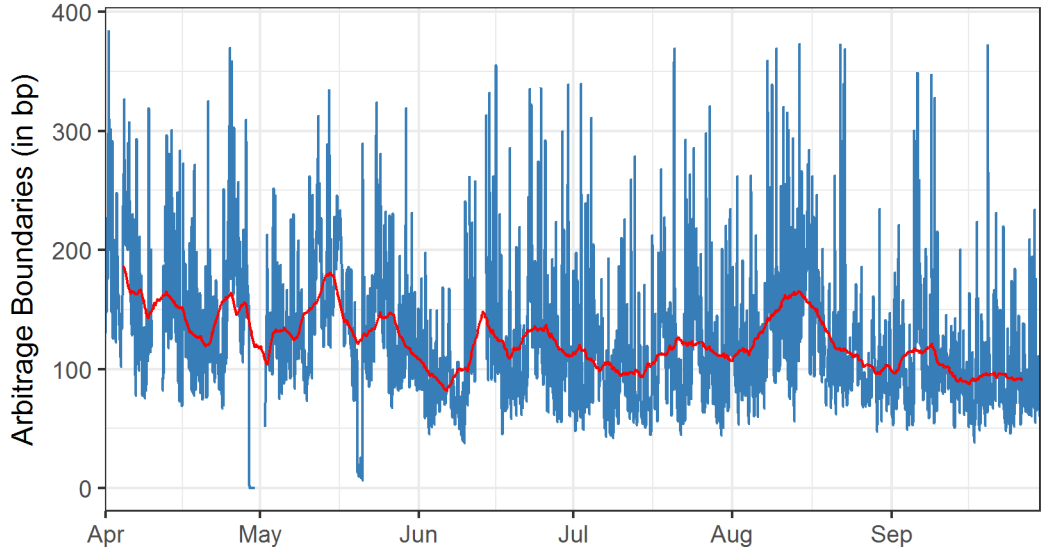
$$c_1 = \hat{\mathbb{E}}_t(\tau) + \hat{\mathbb{E}}_t(\tau_B) \cdot (B^s - 1), \tag{33}$$

$$c_2 = \hat{\mathbb{V}}_t(\tau) + \hat{\mathbb{V}}_t(\tau_B) \cdot (B^s - 1)^2 + \left(\hat{\mathbb{E}}_t(\tau_B) \cdot (B^s - 1) + \hat{\mathbb{E}}_t(\tau)\right)^2, \tag{34}$$

where $\hat{\sigma}_t^s$ denotes the square-root of the estimated spot volatility on the sell-side market, and $\hat{\mathbb{E}}_t(\tau)$ and $\hat{\mathbb{V}}_t(\tau)$ denote the estimated conditional mean and variance, respectively, of the latency distribution. Moreover, B^s refers to the number of blocks that the sell-side exchange s requires to consider incoming transactions as being valid. This exchange-specific security requirement thus further increases the settlement latency beyond the waiting time until a transaction's validation in the first block.[18] We thus decompose the latency into two components: the time it takes until a transaction is included in the blockchain, τ, and the additional time until exchanges accept the transaction as de facto being immutable. While τ is predictable using information on the current state of the Bitcoin network (see Section 4.2), the validation time of subsequent blocks is actually unpredictable. In fact, we do not find evidence against non-zero autocorrelation in waiting times and constant volatility in the block validation time. This evidence supports the notion that the validation times of blocks are partly under control of the Bitcoin network and are internally impaired by the computational complexity of the underlying cryptographic problem. As a result, we can safely assume that the waiting times between subsequent blocks after the first one, which includes the current transaction, are indepen-

[18] *bitFlyer* and *Liqui* do not report a minimum number of confirmations. They rather use a discretionary system depending on the individual transaction and the state of the network. In this case, we assume the number of confirmations to be equal to the median across all exchanges that provide such information, which is 3.

Figure 10: Estimated Arbitrage Boundaries over Time.
This figure shows the average estimated arbitrage boundary based on CRRA utility with risk aversion $\gamma = 2$. We estimate the boundaries using spot volatility estimates following Kristensen (2010) and conditional moments of the latency based on a gamma distribution. The solid blue line shows the hourly averages across all exchanges. The solid red line corresponds to the weekly moving average over the hourly averages.



dently and identically distributed. As validators append a new block on average every 9.5 minutes in our sample (see Table 2), we use this magnitude as the best-possible prediction of the time between two subsequent blocks, $\hat{\mathbb{E}}_t(\tau_B)$. Accordingly, $\hat{\mathbb{V}}_t(\tau_B)$ denotes the (sample) variance of the time between two consecutive blocks.

In line with Conine et al. (2017), we fix the coefficient of risk aversion to $\gamma = 2$ and estimate $\hat{d}_t^s$ for each exchange on a minute level. Figure 10 illustrates the time variation of the average arbitrage boundaries across all exchanges. We observe a substantial variation of these boundaries over time. The correlation between arbitrage boundaries and spot volatilities (expected latency) is 95% (16%) which indicates that volatility is the main driver of the variation in arbitrage boundaries.

Table 5 gives summary statistics of the resulting time series of arbitrage boundaries, which, on average vary from 54 bp to 250 bp. Our theoretical framework allows us to directly analyze the relevance of the latency *uncertainty*. As the variance of the arbi-

Table 5: Summary of Arbitrage Boundaries.
This table summarizes the time series of estimated arbitrage boundaries for the individual sell-side markets. We compute arbitrage boundaries for risk aversion parameter $\gamma = 2$ for the case with isoelastic utility. We estimate the boundaries using the spot volatility estimator of Kristensen (2010) and conditional moments of the latency based on a gamma distribution. We report all values in basis points (except otherwise noted). *Uncertainty* corresponds to the (percentage) contribution of the uncertainty in latency to the median arbitrage boundary. *Security* gives the (percentage) contribution of the required number of confirmations to the median arbitrage boundary.

	Mean	SD	5 %	25 %	Median	75 %	95 %	Uncertainty	Security
Binance	96.64	53.94	40.61	63.17	84.79	117.75	194.70	38.09	45.07
Bitfinex	116.55	70.11	39.50	71.10	100.92	146.32	244.49	44.34	57.63
bitFlyer	136.25	74.14	58.76	92.60	119.14	162.36	271.20	44.64	58.15
Bitstamp	125.72	72.29	52.68	82.16	109.69	151.34	248.40	44.42	57.62
Bittrex	143.91	59.43	68.99	106.63	136.07	173.09	251.26	37.97	45.08
BTCC	100.92	124.09	0.00	0.11	65.97	136.32	353.93	37.78	43.88
CEX.IO	130.55	65.62	47.52	89.26	122.08	160.68	239.71	44.36	58.21
Gate	101.94	52.80	41.96	68.76	91.92	124.28	194.80	38.48	45.42
Gatecoin	249.62	246.77	9.02	83.76	168.65	326.00	808.93	50.35	71.98
Coinbase Pro	110.31	69.94	36.09	66.40	93.45	137.81	236.91	44.04	57.77
Gemini	117.30	69.86	41.45	73.15	101.42	146.74	239.14	44.37	57.80
HitBTC	87.75	53.01	30.79	55.02	76.58	107.75	177.76	38.14	44.83
Kraken	170.14	101.48	61.60	104.25	148.79	210.70	350.80	50.23	72.15
Liqui	128.17	64.41	36.11	86.59	122.77	160.59	235.52	44.29	58.21
Lykke	120.86	82.12	32.95	66.61	100.31	155.64	274.89	43.64	57.49
Poloniex	54.18	34.28	21.11	32.60	45.28	65.33	117.48	12.39	0.00
xBTCe	112.35	79.92	0.00	65.57	95.03	145.76	251.42	44.40	58.21

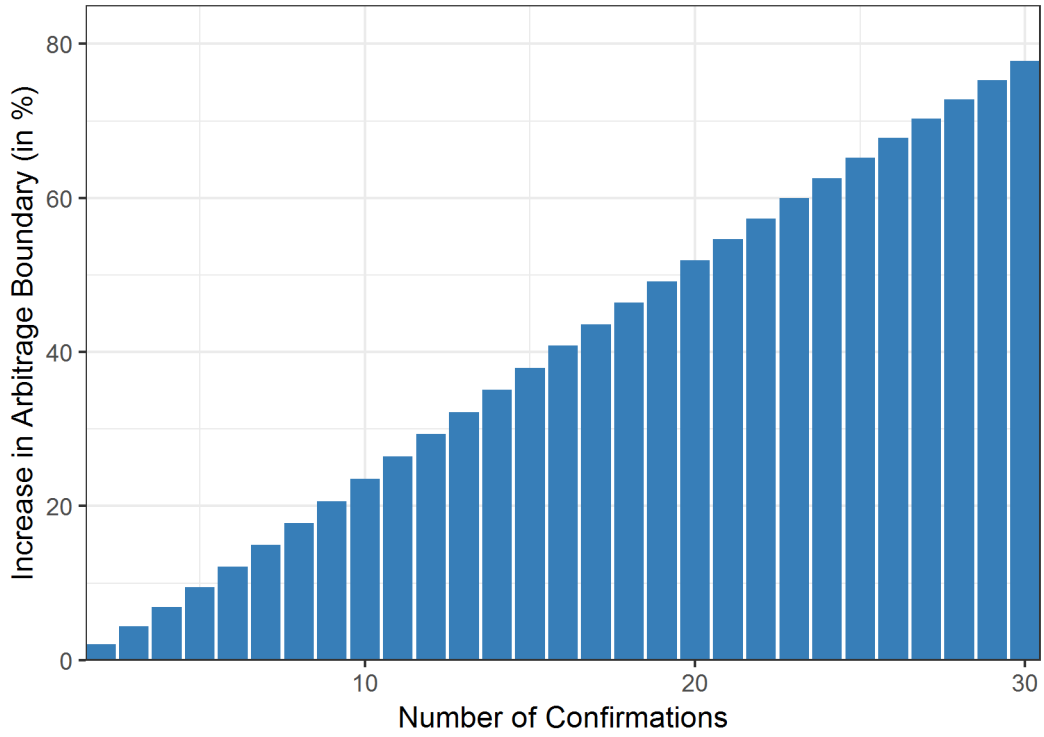
trageurs' returns increases with the (conditional) variance of the settlement latency, we can compare the estimated arbitrage boundaries to the (hypothetical) case of a deterministic latency. The second to last column in Table 5 gives the increase in arbitrage boundaries when adjusting for the randomness in latency. The values correspond to the (percentage) difference between the median arbitrage boundary and boundaries based on the assumption $\mathbb{V}_t(\tau) = \mathbb{V}_t(\tau_B) = 0$. We find that the impact of the randomness in latency is considerable and accounts on average for approximately 43% of the arbitrage boundaries.

Moreover, the arbitrage boundaries differ across exchanges due to an additional source of variation: we can decompose arbitrage boundaries into the latency until a transaction is included in a block for the first time, τ, and the latency until a transaction fulfills exchange-specific security requirements, τ_B. While the conditional moments of the la-

tency distribution affect the time series variation of the boundaries, the cross-sectional variation is only driven by the exchange-specific spot volatilities and the required number of confirmations, B^s. For instance, even though the spot volatility estimates are highly correlated across exchanges, the arbitrage boundaries of *Gatecoin* are large particularly due to the occurrence of periods of high volatility. *Gatecoin* and *Kraken* require $B^s = 6$ confirmations. Thus, these exchanges show on average the highest boundaries, while *Poloniex* requires only $B^s = 1$ confirmation yielding the smallest boundary on average. The last column in Table 5 gives the increase in the median arbitrage boundary when taking exchange-specific number of confirmations into account. The values correspond to the (percentage) difference between the median arbitrage boundary as of Equation (32) and corresponding boundaries based on the assumption $B^s = 1$ for all exchanges. We observe that the impact of exchange-specific security components on arbitrage boundaries is substantial and accounts for 55% of the arbitrage boundaries, on average.

This analysis provides deeper insights into the implied costs of distributed settlement under sufficiently high security standards. These costs materialize in terms of considerable limits of arbitrage that constitute a significant market friction. Such no-arbitrage regions allow for violations of the law of one price and thus have implications for pricing and quoting of market makers. To shed more light on these costs imposed by confirmation requirements, we quantify the relationship between the level of security and the resulting latency. For each exchange, we compute arbitrage boundaries for a given hypothetical number of confirmations and compare it to the baseline case of no additional security requirements (i.e., whenever the inclusion in the first upcoming block is sufficient). Figure 11 shows the increase in the average (across time and the cross-section) arbitrage boundary for varying numbers of confirmations. We observe that (on average) arbitrage boundaries increase by 2% if the security requirements are increased by one block. For instance, requiring 10 confirmations increases the average arbitrage boundary by more

Figure 11: Number of Confirmations and Arbitrage Boundaries.
This figure visualizes the trade-off between a higher number of confirmations and the resulting increase in arbitrage boundaries. For each exchange, we compute arbitrage boundaries for a hypothetical number of confirmations (horizontal axis). We then compare the resulting arbitrage boundaries to the baseline case of no additional confirmations, i.e., whenever an inclusion in the first block is sufficient. For each number of confirmation, we compute the time-series and cross-sectional average increase in arbitrage boundaries relative to the baseline case (vertical axis).



than 20%.

4.4 Quantifying violations of arbitrage limits

To quantify to which extent observed cross-market price differences exceed the estimated arbitrage boundaries, we define the price differences in excess of arbitrage boundaries as

$$\mathcal{E}_t := \left(\Delta_t - \begin{pmatrix} \hat{d}_t^1 \\ \vdots \\ \hat{d}_t^N \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} \right) \odot \Psi_t \quad , \tag{35}$$

where the (i,j)-th element of Ψ_t is defined as

$$\Psi_{t,i,j} = \mathbb{1}\left\{b_t^i - a_t^j > \hat{d}_t^i\right\}, \tag{36}$$

and $\mathbb{1}\{.\}$ is the indicator function, and $\odot$ corresponds to the element-wise multiplication operator. Similarly, we define the price differences *adjusted for transaction costs* in excess of arbitrage boundaries as

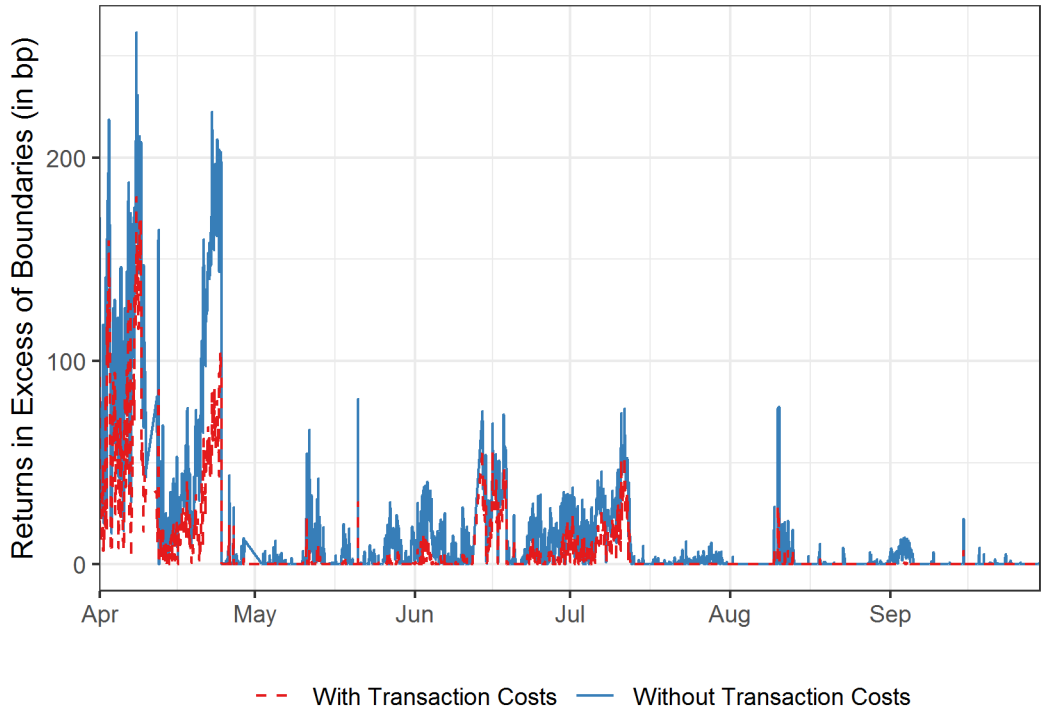
$$\tilde{\mathcal{E}}_t := \left(\tilde{\Delta}_t - \begin{pmatrix} \hat{d}_t^1 \\ \vdots \\ \hat{d}_t^N \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} \right) \odot \tilde{\Psi}_t, \quad \text{with } \tilde{\Psi}_{t,i,j} = \mathbb{1}\left\{\tilde{b}_t^i(q_t^{i,j}) - \tilde{a}_t^j(q_t^{i,j}) > \hat{d}_t^i\right\}. \tag{37}$$

Figure 12 plots the time series of cross-sectional average returns in excess of the arbitrage boundaries. The blue line corresponds to price differences at the best bid and best ask *not* adjusted for transaction costs according to Equation (35). The red dashed line shows the corresponding excess price differences after adjusting for transaction costs according to Equation (37). Taking transaction costs into account lowers the returns in excess of arbitrage boundaries on average by 15% in our sample.

To quantify which proportion of observed differences are within the arbitrage boundaries, we compute the fractions $\min\{1, \hat{d}_t^s/\delta_t^{b,s}\}$, where $\delta_t^{b,s}$ corresponds to the differences of the (log) best bid and ask prices at exchanges b and s at time t according to Equation (25), and $\hat{d}_t^s$ is the arbitrage boundary for exchange s at time t.

Panel A of Figure 13 visualizes the time series of the average across all exchange-pair-based price differences within arbitrage boundaries for different values of the risk aversion parameter γ. We find that even for very low values of risk aversion, roughly 88% of observed price differences fall within the estimated arbitrage boundaries. By construction, a higher risk aversion widens the no-trade regions of the arbitrageur and

Figure 12: Average Returns in Excess of Arbitrage Boundaries.
This figure shows the time series of the average minute-level returns in excess of the estimated arbitrage boundaries. The solid blue line corresponds to price differences based on the best bid and best ask of the individual exchange pairs, $\mathcal{E}_t$. The red line displays the corresponding excess price differences after adjusting for transaction costs, $\tilde{\mathcal{E}}_t$.



thus implies higher limits to arbitrage.

Panel B of Figure 13 provides the corresponding proportions $\min\{1, \hat{d}_t^{s,b}/\tilde{\delta}_t^{b,s}\}$, when we adjust the boundaries for the impact of transaction costs utilizing optimal trading quantities, $q_t^{b,s}$, according to Equation (26). On average, 98% of all observed price differences are within arbitrage boundaries due to stochastic latency when transaction costs are taken into account. Therefore, the vast majority of substantial price differences observed at the Bitcoin market are within no-trade regions of rational arbitrageurs, taking into account costs of trading and the risks associated with stochastic latency.

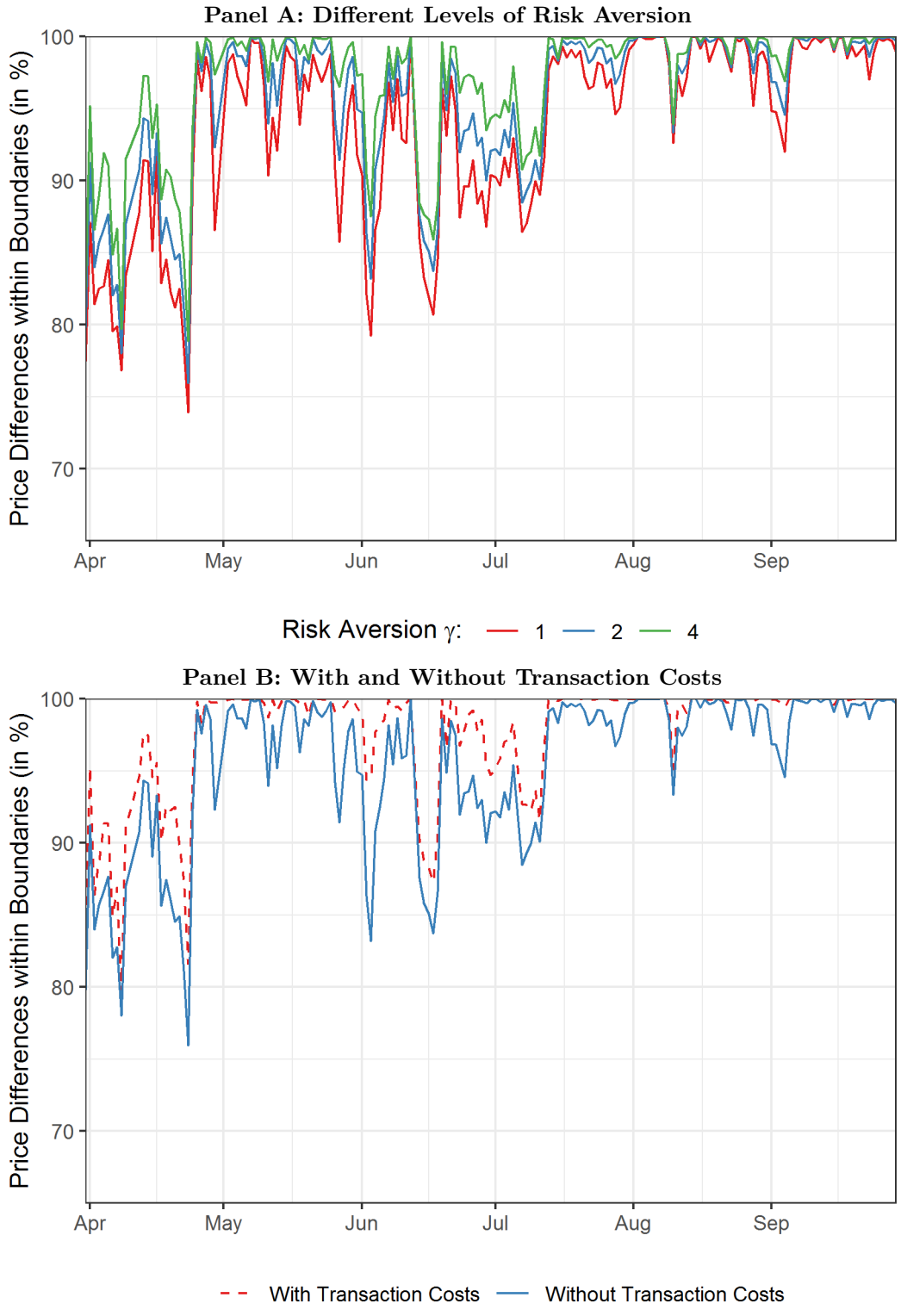
Obviously, observations outside the arbitrage boundaries might arise due to additional market frictions, which are not captured by our theory, e.g., capital controls (Choi et al., 2018), or can be explained by higher risk aversion of the arbitrageurs. Instead of displaying arbitrage limits for different values of (relative) risk aversion, one may alternatively compute the *implied* relative risk aversion, which is necessary to encompass *all* price differences observed. In particular, we compute the risk aversion parameter $\hat{\gamma}_t^{b,s}$, for which the observed price difference of exchange pair $\{b, s\}$ at time t are located within the implied limits to arbitrage. The interpretation of $\hat{\gamma}_t^{b,s}$ is straightforward: if the risk aversion of an arbitrageur is below $\hat{\gamma}_t^{b,s}$, it would be rational to trade. We compute $\hat{\gamma}_t^{b,s}$ according to the following lemma.

Lemma 7. *Define* $\hat{\gamma}_t^{b,s}$ *as the root of the cubic polynomial*

$$\left(\tilde{\delta}_t^{b,s}\right)^4 - \frac{1}{8}\left(\hat{\sigma}_t^s\right)^4 c_2 \left(\hat{\gamma}_t^{b,s}\right)^3 - \frac{3}{8}\left(\hat{\sigma}_t^s\right)^4 c_2 \left(\hat{\gamma}_t^{b,s}\right)^2 - \frac{1}{2}\left(\hat{\sigma}_t^s\right)^2 \left(c_1 \left(\tilde{\delta}_t^{b,s}\right)^2 + \frac{1}{2}\left(\hat{\sigma}_t^s\right)^2 c_2\right) \hat{\gamma}_t^{b,s} = 0, \tag{38}$$

where, analogously to Equations (33) *and* (34), $c_1 = \hat{\mathbb{E}}_t(\tau) + \hat{\mathbb{E}}(\tau_B) \cdot (B^s - 1)$ *and* $c_2 = \hat{\mathbb{V}}_t(\tau) + \hat{\mathbb{V}}(\tau_B) \cdot (B^s - 1)^2 + \left(\hat{\mathbb{E}}(\tau_B) \cdot (B^s - 1) + \hat{\mathbb{E}}_t(\tau)\right)^2$. *Then, price differences (adjusted for transaction costs)* $\tilde{\delta}_t^{b,s}$ *constitute a (statistical) arbitrage opportunity for an arbitrageur with risk aversion* γ *only if* $\gamma < \hat{\gamma}_t^{b,s}$.

Figure 13: Explanatory Power of Limits to Arbitrage.
These figures show the share of price differences explained by arbitrage boundaries. For each exchange pair, we compute $\hat{d}_t^s/\Delta_t^{b,s}$ and then take the hourly average. In Panel A, we plot the results for different levels of risk aversion. In Panel B, we plot the results with and without taking transaction costs into account for risk aversion $\gamma = 2$.



Proof. See Appendix C. □

The exchange pair specific implied risk aversion $\hat{\gamma}_t^{b,s}$ is defined in a way such that the observed price differences $\tilde{\delta}_t^{b,s}$, adjusted for transaction costs, coincide with the arbitrage boundaries for an isoelastic utility function with risk aversion parameter $\hat{\gamma}_t^{b,s}$. As the arbitrage boundaries monotonically increase with risk aversion, any value of γ below $\hat{\gamma}_t^{b,s}$ constitutes a trading opportunity for the arbitrageur. Conversely, $\gamma > \hat{\gamma}_t^{b,s}$ reflects that the observed price differences do not justify (unconstrained) trading because an arbitrageur with a higher risk aversion obtains higher (expected) utility by trading less or not at all. As the asset is traded on N markets, we define $\gamma_t^{\max}$ as the minimum risk aversion parameter for which all observed price differences fall within the implied boundaries to arbitrage, i.e.,

$$\hat{\gamma}_t^{\max} := \max_{i,j \in \{1,\dots,N\}} \hat{\gamma}_t^{i,j}. \tag{39}$$

Figure 14 shows the time series of implied risk aversion parameters $\hat{\gamma}_t^{\max}$. On average across our sample, the implied minimum risk aversion equals 12. During more recent periods, the decrease in observed price differences reduces $\hat{\gamma}_t^{\max}$ considerably and suggests that arbitrage opportunities vanish for reasonable values of relative risk aversion. Estimates of γ in the recent literature range from as little as 0.35 to as much as 9.0 (e.g., Hansen and Singleton, 1982; Aivazian et al., 1986; Chetty, 2006). Our data suggests that particularly since August 2018 the risk aversion of active arbitrageurs required to exploit remaining price differences (after adjusting for transaction costs) would be close to zero. An alternative interpretation of this finding is that arbitrage opportunities at Bitcoin markets vanish because (i) observed price differences decrease and (ii) the remaining instantaneous returns do not justify the substantial latency risk anymore even for arbitrageurs that are nearly risk neutral.

Finally, to assess the overall efficiency of the market, we compute the proportion of observations (on a minute level and per exchange pair), where the price differences exceed

Figure 14: Implied Risk Aversion.
This figure shows the time series of the implied risk aversion parameter $\hat{\gamma}_t^{\max}$, which yields the smallest relative risk aversion such that all observed price differences (adjusted for transaction costs) fall within the implied limits to arbitrage. The solid blue line shows the hourly averages. The solid red line corresponds to the weekly moving average over the hourly averages.

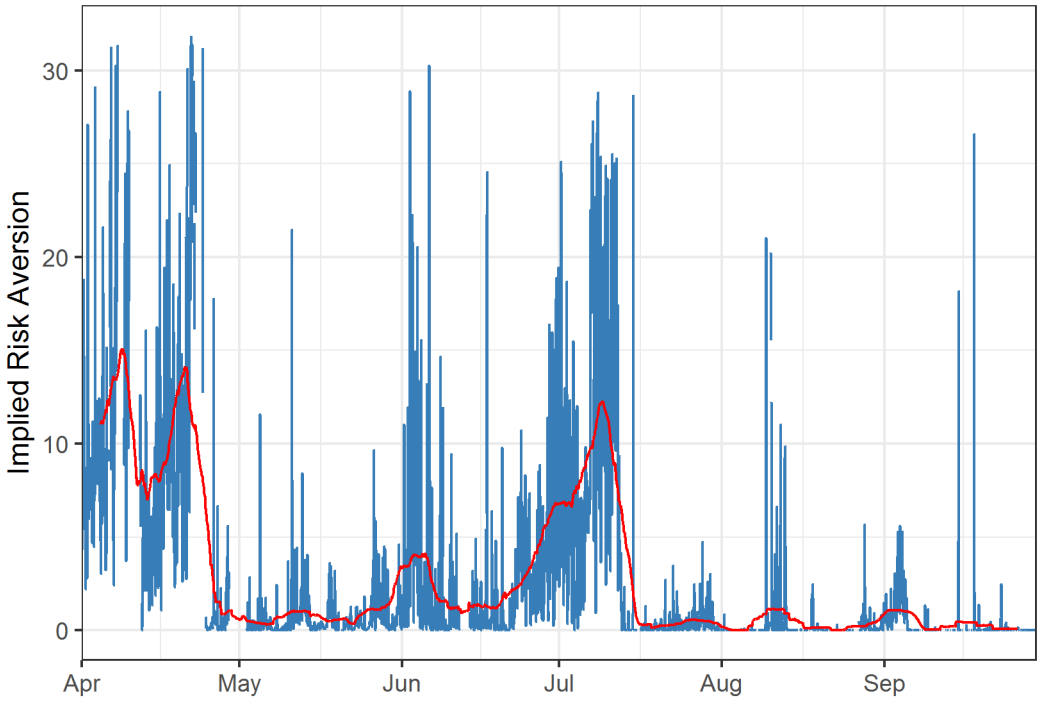


Table 6: Frequency of Arbitrage Opportunities.
This table shows the fraction of all exchange pair observations of a given month that exceed a certain threshold. The first column corresponds to the fraction of minute-level price differences greater than zero, the second column corresponds the fraction of price differences in excess of arbitrage boundaries implied by stochastic latency. The third column gives the fraction of price differences in excess of boundaries taking transaction costs into account.

Month	$\delta_t^{b,s} > 0$	$\delta_t^{b,s} > \hat{d}_t^s$	$\tilde{\delta}_t^{b,s} > \hat{d}_t^s$
2018-April	66.11	29.10	21.75
2018-May	63.05	15.31	2.91
2018-June	68.59	24.54	8.42
2018-July	60.66	16.53	7.36
2018-August	55.16	3.72	1.20
2018-September	55.16	6.69	0.89

the arbitrage boundaries. Table 6 shows the resulting summary statistics on a monthly basis. We observe arbitrage opportunities, $\delta_t^{b,s} > 0$, in 63% of all observations. However, adjusting for stochastic latency, we only observe arbitrage opportunities, $\delta_t^{b,s} > \hat{d}_t^s$, in 17% of all cases. The number decreases to 8% when we additionally take transaction costs into account ($\tilde{\delta}_t^{b,s} > \hat{d}_t^s$). Interestingly, the occurrence of arbitrage opportunities as implied by our theoretical framework dropped significantly since August 2018.

5 Conclusions

Many market participants believe that distributed ledger technology has the potential to radically transform the transfer of assets. The promises of greater efficiency and higher security, however, come at the cost of stochastic latency in the settlement process without the option to dispose positions during this waiting time. We theoretically show that stochastic latency implies limits to arbitrage as it exposes arbitrageurs to price risk. We develop a framework which allows us to derive arbitrage boundaries for arbitrary concave utility functions and a general class of latency distributions. The arbitrage boundaries increase with spot volatility, risk aversion, expected latency and uncertainty in latency. Furthermore, we show how to estimate these arbitrage boundaries for the realistic case

of a utility function with constant relative risk aversion.

To quantify the limits to arbitrage in the Bitcoin market, we utilize more than 22 million transactions from the Bitcoin blockchain and high-frequency orderbook data from several exchanges. We find that stochastic latency is a quantitatively important market friction and imposes arbitrage boundaries of 124 bp on average. Approximately 43% of these boundaries are due to the randomness in waiting times until settlement.

The quantification of latency-based limits to arbitrage is essential to assess the efficiency of a market relying on distributed ledger technology. In fact, we show that on average, 88% of observed Bitcoin price differences are within the corresponding arbitrage boundaries. Additionally adjusting for transaction costs, this proportion increases to 98%.

Stochastic latency constitutes a novel market friction that might have far reaching implications. First, limits to arbitrage implied by stochastic latency reduce price efficiency, as the lower activity of arbitrageurs reduces the information flow across markets. Second, deviations from the law of one price affect the pricing of securities, as risk neutral probabilities are not uniquely defined. Third, the implied costs of stochastic latency depend on the design of the distributed ledgers and should influence the decision whether to migrate to a distributed settlement system. Fourth, the magnitude of these boundaries provides market makers flexibility in setting their quotes without being exploited by arbitrageurs. Overall, our paper provides a first step to understand the impact of stochastic settlement latency on financial markets.

References

Abadie, J. and M. Brunnermeier (2018). Blockchain Economics. Working paper.

Aivazian, V. A., J. L. Callen, I. Krinsky, and C. C. Kwan (1986). An Empirical Portfolio Analysis of Financial Asset Substitutability - The Case of the United States Household Sector. *Quarterly Review of Economics and Business 26*(2), 47–65.

Arditti, F. D. (1967). Risk and the Required Return on Equity. *The Journal of Finance 22*(1), 19–36.

Balduzzi, P. and A. W. Lynch (1999). Transaction Costs and Predictability: Some Utility Cost Calculations. *Journal of Financial Economics 52*(1), 47 – 78.

Barndorff-Nielsen, O., J. Kent, and M. Sørensen (1982). Normal Variance-Mean Mixtures and z Distributions. *International Statistical Review / Revue Internationale de Statistique 50*(2), 145–159.

Barndorff-Nielsen, O. E., E. Nicolato, and N. Shephard (2002). Some recent developments in stochastic volatility modelling. *Quantitative Finance 2*(1), 11–23.

Biais, B., C. Bisiere, M. Bouvard, and C. Casamatta (2017). The Blockchain Folk Theorem. Working paper.

BIS (2017). Distributed Ledger Technology in Payment, Clearing and Settlement: An Analytical Framework. Bank for International Settlements, Committee on Payments and Market Infrastructures.

Bondarenko, O. (2003). Statistical Arbitrage and Securities Prices. *The Review of Financial Studies 16*(3), 875–919.

Brogaard, J., T. Hendershott, and R. Riordan (2014). High-Frequency Trading and Price Discovery. *The Review of Financial Studies 27*(8), 2267–2306.

Budish, E., P. Cramton, and J. Shim (2015). The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. *The Quarterly Journal of Economics 130*(4), 1547–1621.

Chetty, R. (2006). A New Method of Estimating Risk Aversion. *American Economic Review 96*(5), 1821–1834.

Chiu, J. and T. V. Koeppl (2018). Blockchain-based Settlement for Asset Trading. Working paper.

Choi, K. J., A. Lehar, and R. Stauffer (2018). Bitcoin Microstructure and the Kimchi Premium. Working paper.

Cong, L. W. and Z. He (2017). Blockchain Disruption and Smart Contracts. Working paper.

Conine, T. E., M. B. McDonald, and M. Tamarkin (2017). Estimation of Relative Risk Aversion Across Time. *Applied Economics 49*(21), 2117–2124.

de Jong, A., L. Rosenthal, and M. A. van Dijk (2009). The Risk and Return of Arbitrage in Dual-Listed Companies. *Review of Finance 13*(3), 495–520.

de Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1990). Noise Trader Risk in Financial Markets. *Journal of Political Economy 98*(4), 703–738.

Durrett, R. (1984). *Brownian Motion and Martingales in Analysis*. Wadsworth Advanced Books & Software.

Easley, D., M. O'Hara, and S. Basu (2017). From Mining to Markets: The Evolution of Bitcoin Transaction Fees. Working paper.

ECB and BoJ (2018). Securities Settlement Systems: Delivery-versus-Payment in a Distributed Ledger Environment. STELLA - A Joint Research Project of the European Central Bank and the Bank of Japan.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business 38*(1), 34–105.

Foucault, T., R. Kozhan, and W. W. Tham (2017). Toxic Arbitrage. *The Review of Financial Studies 30*(4), 1053–1094.

Gandal, N., J. Hamrick, T. Moore, and T. Oberman (2018). Price Manipulation in the Bitcoin Ecosystem. *Journal of Monetary Economics, Forthcoming 95*, 86 – 96.

Griffin, J. M. and A. Shams (2018). Is Bitcoin Really Un-Tethered? Working paper.

Gromb, D. and D. Vayanos (2010). Limits of Arbitrage. *Annual Reviews of Financial Economics 2*(1), 251–275.

Hadar, J. and W. R. Russell (1969). Rules for Ordering Uncertain Prospects. *The American Economic Review 59*(1), 25–34.

Hansen, L. P. and K. J. Singleton (1982). Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica 50*(5), 1269–1286.

Harvey, C. R., J. C. Liechty, M. W. Liechty, and P. Müller (2010). Portfolio Selection with Higher Moments. *Quantitative Finance 10*(5), 469–485.

Hasbrouck, J. and G. Saar (2013). Low-Latency Trading. *Journal of Financial Markets 16*(4), 646–679.

Khapko, M. and M. Zoican (2017). Smart Settlement. Working paper.

Kotz, S., T. Kozubowski, and K. Podgorski (2012). *The Laplace Distribution and Generalizations: a Revisit with Applications to Communications, Economics, Engineering, and Finance*. Springer Science & Business Media.

Kristensen, D. (2010). Nonparametric Filtering of the Realized Spot Volatility: A Kernel-Based Approach. *Econometric Theory 26*, 60–93.

Lamont, O. A. and R. H. Thaler (2003a). Anomalies: The Law of one Price in Financial Markets. *Review of Finance 17*(4), 191–202.

Lamont, O. A. and R. H. Thaler (2003b). Can the Market Add and Subtract? Mispricing in Tech Stock Carve-Outs. *Journal of Political Economy 111*(2), 227–268.

Levy, H. (1992). Stochastic Dominance and Expected Utility: Survey and Analysis. *Management Science 38*(4), 555–593.

Makarow, I. and A. Schoar (2018). Trading and Arbitrage in Cryptocurrency Markets. Working paper.

Malinova, K. and A. Park (2017). Market Design with Blockchain Technology. Working paper.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance 7*(1), 77–91.

Menkveld, A. J. and M. A. Zoican (2017). Need for Speed? Exchange Latency and Liquidity. *The Review of Financial Studies 30*(4), 1188–1228.

Mitchell, M., L. H. Pedersen, and T. Pulvino (2007). Slow Moving Capital. *American Economic Review 97*(2), 215–220.

Mitchell, M., T. Pulvino, and E. Stafford (2002). Limited Arbitrage in Equity Markets. *The Journal of Finance 57*(2), 551–584.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. Working Paper.

NASDAQ (2017). Nasdaq and Citi Announce Pioneering Blockchain and Global Banking Integration. National Association of Securities Dealers Automated Quotations, URL: `https://www.nasdaq.com/article/nasdaq-and-citi-announce-pioneering-blockchain-and-global-banking-integration-cm792544`.

Pontiff, J. (1996). Costly Arbitrage: Evidence from Closed-End Funds. *Quarterly Journal of Economics 111*(4), 1135–1152.

Roll, R., E. Schwartz, and A. Subrahmanyam (2007). Liquidity and the Law of One Price: The Case of the Futures–Cash Basis. *The Journal of Finance 62*(5), 2201–2234.

Saleh, F. (2018). Blockchain Without Waste: Proof-of-Stake. Working paper.

Schneider, P. (2015). Generalized Risk Premia. *Journal of Financial Economics 116*(3), 487–504.

Scott, R. C. and P. A. Horvath (1980). On the Direction of Preference for Moments of Higher Order than the Variance. *The Journal of Finance 35*(4), 915–919.

SEC (2017). Amendment to Securities Transaction Settlement Cycle - A Small Entity Compliance Guide. U.S. Securities and Exchange Commission, Amendment to Rule 15c6-1(a) under the Exchange Act, URL: `https://www.sec.gov/tm/t2-sbrefa`.

Shleifer, A. and R. W. Vishny (1997). The Limits of Arbitrage. *The Journal of Finance 52*(1), 35–55.

SIX (2018). SIX to Launch Full End-to-End and Fully Integrated Digital Asset Trading, Settlement and Custody Service. Swiss Infrastructure and Exchange, URL: `https://www.six-group.com/en/home/media/releases/2018/20180706-six-digitalexchange.html`.

Appendix

A Distributed ledger technology and stochastic latency

In its essence, the distributed ledger technology (DLT) is a digital record-keeping system that allows for the verification, updating and storage of transfers of ownership without the need for a designated third party. It relies on a single ledger that is distributed among many different parties who are incentivized to ensure a truthful representation of the transaction history. Nakamoto (2008) first popularized the idea of DLTs in a financial context with the Bitcoin protocol and the underlying concept of blockchain. Nowadays, however, Bitcoin is just one of several hundred applications that use the blockchain technology, while other forms of DLTs, in particular directed acyclic graphs (DAG), are actively explored as well. In the following, we first describe the building blocks of DLT before we turn to a more detailed discussion of blockchains and DAGs.

A.1 Fundamentals of distributed ledgers

DLT solves the fundamental problems that arise in the context of digital transfer of ownership. Transactions are pieces of information that agents authorize to be sent to other agents. A record-keeping system has to ensure that transactions are signed and recorded in the correct order. In principle, a single authority could verify signatures and consistency of transactions, but it would be prone to failures. As a result, it might be desirable to distribute this type of information to a system with multiple machines that can sustain the failure of single units. A fault-tolerant design would enable a system to continue its operation, even when one or several units stop working.

To achieve this goal, DLT essentially combines two fundamental concepts. First, a distributed ledger is based on *asymmetrical cryptography* that enables digital signatures of transactions. On the one hand, the sender of a transaction wants to be the sole owner

of the signature that allows the transfer of assets from her private wealth. On the other hand, a record-keeping system requires information about the identities of the parties involved. Cryptographic algorithms ensure that any private keys are only known to their owners, while public keys may be disseminated widely. That is, everybody can check whether a private key is valid, but nobody can back out the private key from public information.[19]

Second, a distributed ledger is conceptually a *distributed system* which checks whether transactions should be in the system and in which order. In a distributed system, many machines are connected through a network and ensure that the system keeps operating even when some machines fail or try to mess up the system. For instance, if the sender of a transaction is also a potential validator, then she has an incentive for dishonest behavior, such as double-spending or revoking transactions. Individual machines have to reach some form of consensus about actual transaction histories. This consensus can be achieved through different network structures such as blockchain or DAG.

A.2 Blockchain

In the context of blockchain, the typical solution to the consensus problem involves competition among potential validators for the right to append information to the ledger.[20] The most common consensus protocol, Proof-of-Work (PoW), involves solving a computationally expensive problem where the winner gets the right to update the ledger and typically receives a reward. This particular form of DLT is called blockchain since transactions are not verified individually, but rather appended to the ledger in blocks. Validators bundle transactions that wait for verification and try to solve the problem.

[19]The most simple illustrative example for asymmetric cryptography is the multiplication of prime numbers. One can easily multiply two prime numbers (private key) to get a large number (public key), but it can be difficult to infer the initial set of numbers from the product.

[20]The problem is more severe in a *permissionless* blockchain where anybody can access and potentially update the blockchain. Other variants, where only few institutions or individuals are entitled to direct access to the blockchain, so-called *permissioned* blockchains, limit the problem to few players.

However, the system's protocol limits the number of transactions that can be included in a single block. This limit leads to a queue of unconfirmed transactions and validators are free to choose the transactions they try to append to the blockchain. Average verification times thus not only depend on the number of unconfirmed transactions, but also on the fee associated with a transaction, as validators find it more attractive to include transactions with high fees in their blocks.[21]

The computationally difficult problem typically relies on cryptographic hash functions, which map an input of arbitrary size to output of fixed size and cannot be inverted. In the Bitcoin network, validators bundle the information of several transactions and a reference to the current state of the blockchain and plug the data into a hash function. The hash function converts this input into a sequence of characters and numbers of certain length. The system's protocol then requires that the output starts with a certain number of zeros. The probability of calculating a hash that starts with many zeros is very low and to generate a new hash, validators include a random number called *nonce* that can lead to a very different output. The difficulty of the problem is then determined by the number of leading zeros validators have to find. Depending on total available computational power, the system regularly adjusts the target to achieve an average of 10 minutes between two consecutive blocks.

While validators in a PoW system utilize substantial computational resources to win the competition for block generation, validators might also be randomly chosen based on their wealth. In the so called Proof-of-Stake (PoS) protocol, validators stake their tokens to be able to create blocks. The higher a validator's stake, the higher are the chances of creating the next block. After successfully appending a new block, the validator receives transaction fees just as in the case of PoW. If the validator submits an incorrect block or is offline during a staking period, then she is penalized and (at least partly) loses her stake.

[21] Biais et al. (2017) provide an extensive discussion of the equilibrium properties of the PoW game, while Easley et al. (2017) analyze the role of transaction fees.

The penalty might either arise explicitly through a deduction of funds from the stake or implicitly as dishonest behavior creates a feedback on the value of the stake. In particular, if a validator appends to the blockchain in a way that perpetuates disagreement, then she imposes a cost upon all users of the particular blockchain. Such behavior lowers the value of the whole network and is also reflected in a lower valuation of the misbehaving validator's stake. The endogenous feedback between validators' behavior and the value of their stakes incentives them to eventually reach consensus.[22]

Other consensus protocols combine features of PoW and PoS. For instance, delegated Proof-of-Stake (DPoS) relies both on stakeholders, who elect validators and have voting rights proportional to their stake, and validators, who exert effort to append information to the ledger. The reputation of validators determines their chance for reelection, while stakeholders have incentives to select truthful validators.

A.3 Directed acyclic graphs

While blockchains record transactions in blocks, DAGs store information in single transactions. More specifically, any transaction represents a node in a graph (i.e., a set of vertices connected by edges) and each new transaction confirms at least one previous transaction, depending on the configuration of the underlying protocol. The longer the branch on which a transaction is based, the more certain is its validity. Intuitively, only once a transaction is broadcasted sufficiently throughout the network, it is verified. DAGs thus hinge on a steady flow of new transactions that enter the network to verify and reference old transactions. The connections between transactions are directed (i.e., the edges in the form of confirmations are one way) and the whole graph is acyclic (i.e., it is impossible to traverse the entire graph starting from a single edge). Given a high number of transactions, DAG ledgers scale better and can achieve consensus faster than blockchains which rely on fixed block sizes and limited verification rates.

[22]See Saleh (2018) for a game-theoretical foundation of the stability of the PoS protocol.

A.4 Stochastic latency in distributed systems

A distributed system features stochastic latency in transaction verification, as it is ex-ante unclear how long it takes until validators achieve consensus or to broadcast that consensus through the network. For PoW, latency depends on the time it takes for validators to find a solution to the computationally expensive problem. In the Bitcoin protocol, for instance, validators append a new block on average every 10 minutes, while the process takes about 20 seconds in the Ethereum protocol. For PoS, latency depends on how long disagreement on the correct order of transactions persists. For both blockchain-based protocols, the information still needs to be distributed to all other nodes in the network, possibly facing technological limitations that prevent instant percolation. Technological limits are also particularly relevant for DAGs which rely on a large number of nodes that verify transactions and distribute information through the network. Overall, any distributed system that refrains from using designated third-parties bearing the counterparty risk associated with transactions thus features stochastic latency.

B Latency distribution under stochastic volatility

We can relax the assumption that σ_t^s is constant over the interval $[t, t+\tau]$ by allowing σ_t^s to vary over time. More specifically, let $\sigma_t^s : \mathbb{R}_+ \to \mathbb{R}_+$ with $\theta(\tau) := \int_t^{t+\tau} (\sigma_k^s)^2 \, dk < \infty \quad \forall \tau$, i.e., the volatility of the sell-side market follows a (deterministic) path with bounded integrated variance. Assuming $\mu_t^s = 0$, we can then rewrite the log returns of the arbitrageur for given latency τ as

$$r_{(t:t+\tau)}^{b,s} = \delta_t^{b,s} + \int_t^{t+\tau} \sigma_k^s dW_k^s. \tag{A.1}$$

The integral above corresponds to a Gaussian process with independent increments. More specifically, we get

$$\mathbb{E}_t\left(\left(r^{b,s}_{(t:t+\tau)} - \delta^{b,s}_t\right)^2\right) = \theta(\tau) - \theta(0) = \mathbb{E}_t\left(W^s_{\theta(\tau)} - W^s_{\theta(0)}\right). \tag{A.2}$$

In other words, the time changed Brownian motion $W^s_{\theta(t)}$ has the same distribution as the log returns given in Equation (A.1) (see, e.g. Durrett, 1984; Barndorff-Nielsen et al., 2002). We can thus rewrite the return process as

$$r^{b,s}_{(t:t+\tau)} = \delta^{b,s}_t + \int_t^{t+\theta(\tau)} dW^s_k, \tag{A.3}$$

The implications of Lemma 1 still hold, but we need to compute the moment generating function of the transformed latency $m_{\theta(\tau)}(u)$, which depends on the latency distribution and the dynamics of the volatility process. First, note that, as $\theta(\tau)$ is strictly increasing, the probability integral transformation yields the distribution of $\tau(\theta)$,

$$\mathbb{P}_t\left(\theta(\tau) = y\right) = \mathbb{P}_t\left(\tau = \theta^{-1}(y)\right) \quad \forall y > 0. \tag{A.4}$$

Finally, the distribution of $\theta(\tau)$ is fully described via its characteristic function which is of the form

$$\varphi_{\theta(\tau)}(u) = \mathbb{E}_t\left(e^{i\theta(\tau)u}\right) = \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^\infty \varphi_\tau(s)\, e^{-is\tau} ds e^{i\theta(\tau)u} d\tau. \tag{A.5}$$

Lévy's characterization allows to extend these ideas to more general non-deterministic integrands and to stochastic time-changes. Although Equation (A.5) allows to derive theoretical arbitrage boundaries based on Theorem 1 for every continuous local martingale, we restrict our analysis to analytically more tractable and intuitive dynamics of the

price process and the associated settlement latency.

C Proofs

Proof of Theorem 1. First, note that the characteristic function in Lemma 1 yields the first moment μ_r of the returns as given by

$$\begin{aligned}\mathbb{E}_t\left(r^{b,s}_{(t:t+\tau)}\right) &= (-i)\frac{\partial}{\partial u}\varphi_{r^{b,s}_{(t:t+\tau)}}(u)\Big|_{u=0} \\ &= \delta_t^{b,s} e^{iu\delta_t^{b,s}} m_\tau\left(iu\mu_t^s - \frac{1}{2}u^2(\sigma_t^s)^2\right) \\ &\quad + e^{iu\delta_t^{b,s}} m'_\tau\left(iu\mu_t^s - \frac{1}{2}u^2(\sigma_t^s)^2\right)\left(\mu_t^s + iu(\sigma_t^s)^2\right)\Big|_{u=0} \\ &= \delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s, \end{aligned} \tag{A.6}$$

since $m_\tau(0) = 1$ and $m'_\tau(0) = \mathbb{E}_t(\tau)$ by definition of the moment generating function.

In spirit of Arditti (1967) and Scott and Horvath (1980), we express the expected utility of the arbitrageur by a Taylor expansion which results in a function of the higher-order moments of the return distribution. A Taylor expansion of a general utility function $U_\gamma(r)$ around the mean μ_r yields

$$U_\gamma\left(r^{b,s}_{(t:t+\tau)}\right) = \sum_{k=0}^{\infty}\frac{U_\gamma^{(k)}(\mu_r)}{k!}\left(r^{b,s}_{(t:t+\tau)} - \mu_r\right)^k, \tag{A.7}$$

where $U_\gamma^{(k)}(\mu_r) := \frac{\partial^k}{\partial \mu_r^k} U_\gamma(\mu_r)$. Then, taking expectations yields

$$\mathbb{E}_t\left(U_\gamma\left(r^{b,s}_{(t:t+\tau)}\right)\right) = U_\gamma(\mu_r) + \sum_{k=2}^{\infty}\frac{U_\gamma^{(k)}(\mu_r)}{k!}\mathbb{E}_t\left(\left(r^{b,s}_{(t:t+\tau)} - \mu_r\right)^k\right). \tag{A.8}$$

Following Markowitz (1952), we next consider a first-order Taylor expansion for the CE. We thus implicitly assume that the risk premium, $\mu_r - CE$, is small and that higher-order

moments vanish:

$$\mathbb{E}_t\left(U_\gamma\left(r^{b,s}_{(t:t+\tau)}\right)\right) = U_\gamma\left(CE\right) = U_\gamma\left(\mu_r\right) + U'_\gamma\left(\mu_r\right)\left(CE - \mu_r\right). \tag{A.9}$$

Moreover, the first-order Taylor expansion provides a convenient closed-form approximation of the certainty equivalent which is linear in the moments of the return distribution. We obtain the equation in the theorem by equating (A.8) and (A.9), plugging in (A.6), and solving for CE:

$$CE = \delta^{b,s}_t + \mathbb{E}_t(\tau)\mu^s_t + \sum_{k=2}^{\infty} \frac{U^{(k)}_\gamma\left(\delta^{b,s}_t + \mathbb{E}_t(\tau)\mu^s_t\right)}{k!U'_\gamma\left(\delta^{b,s}_t + \mathbb{E}_t(\tau)\mu^s_t\right)} \mathbb{E}_t\left(\left(r^{b,s}_{(t:t+\tau)} - \delta^{b,s}_t - \mathbb{E}_t(\tau)\mu^s_t\right)^k\right). \tag{A.10}$$

□

Proof of Lemma 2. For the exponential utility, we have $U^{(k)}\left(r\right)/U'\left(r\right) = \left(-\gamma\right)^{k-1}$ for $k \geq 1$. Therefore, from Theorem 1 we have

$$CE = \delta^{b,s}_t + \mathbb{E}_t\left(\tau\right)\mu^s_t - \frac{\gamma}{2}\mu_{r^{b,s}_{(t:t+\tau)}}(2) + \frac{\gamma^2}{6}\mu_{r^{b,s}_{(t:t+\tau)}}(3) - \frac{\gamma^3}{24}\mu_{r^{b,s}_{(t:t+\tau)}}(4) + \mathcal{O}\left(r\right), \tag{A.11}$$

where $\mu_{r^{b,s}_{(t:t+\tau)}}(k) := \mathbb{E}_t\left(\left(r^{b,s}_{(t:t+\tau)} - \delta^{b,s}_t - \mathbb{E}_t\left(\tau\right)\mu^s_t\right)^k\right)$ is the k-th order central moment of the returns and $\mathcal{O}\left(r\right)$ corresponds to the Taylor approximation error which we neglect subsequently. Recognizing that by definition $m_{r^{b,s}_{(t:t+\tau)}}(iu) = \varphi_{r^{b,s}_{(t:t+\tau)}}(u)$, we can derive the moment generating function of the returns given by

$$m_{r^{b,s}_{(t:t+\tau)}}(u) = e^{u\delta^{b,s}_t} m_\tau\left(u\mu^s_t + \frac{1}{2}u^2(\sigma^s_t)^2\right). \tag{A.12}$$

The central moment generating function is defined as

$$C_{r^{b,s}_{(t:t+\tau)}}(u) = \mathbb{E}_t\left(\exp\left(u\left(r_{(t:t+\tau)} - \mathbb{E}_t(r^{b,s}_{(t:t+\tau)})\right)\right)\right) = \exp\left(-u\mathbb{E}_t(r^{b,s}_{(t:t+\tau)})\right) m_{r^{b,s}_{(t:t+\tau)}}(u)\,. \tag{A.13}$$

Thus, we have

$$\mu_{r^{b,s}_{(t:t+\tau)}}(k) = \frac{\partial^k}{\partial u^k} C_{r^{b,s}_{(t:t+\tau)}}(u)\Big|_{u=0} = \frac{\partial^k}{\partial u^k}\exp\left(-\mathbb{E}_t(\tau)\,\mu^s_t u\right) m_\tau\left(u\mu^s_t + \frac{1}{2}u^2\left(\sigma^s_t\right)^2\right)\Big|_{u=0}. \tag{A.14}$$

Basic calculus then yields

$$\mu_{r^{b,s}_{(t:t+\tau)}}(2) = \mathbb{V}_t(\tau)\left(\mu^s_t\right)^2 + \left(\sigma^s_t\right)^2\mathbb{E}_t(\tau) \tag{A.15}$$

$$\mu_{r^{b,s}_{(t:t+\tau)}}(3) = 3\mu^s_t\left(\sigma^s_t\right)^2\mathbb{V}_t(\tau) + \left(\mu^s_t\right)^3\mathbb{E}_t\left(\left(\tau - \mathbb{E}_t(\tau)\right)^3\right) \tag{A.16}$$

$$\begin{aligned}\mu_{r^{b,s}_{(t:t+\tau)}}(4) = {} & \left(\mu^s_t\right)^4\mathbb{E}_t\left(\left(\tau - \mathbb{E}_t(\tau)\right)^4\right) + 3\mathbb{E}_t\left(\tau^2\right)\left(\sigma^s_t\right)^4 \\ & + 6\left(\sigma^s_t\right)^2\left(\mu^s_t\right)^2\left(\mathbb{E}_t(\tau)^3 + \mathbb{E}_t\left(\tau^3\right) - 2\mathbb{E}_t(\tau)\,\mathbb{E}_t\left(\tau^2\right)\right).\end{aligned} \tag{A.17}$$

Then, we plug in equations (A.15)-(A.17) into (A.11). Finally, recognizing that the arbitrageur exploits price differences if and only if $CE > 0$, we can solve for the minimum instantaneous price differences $\delta^{b,s}_t$ which completes the proof. □

Proof of Lemma 3. The proof follows directly from applying Theorem 1 together with the derivatives of the utility function which yields

$$d^s_t - \frac{1}{2}\frac{\gamma}{d^s_t}(\sigma^s_t)^2\mathbb{E}_t(\tau) - \frac{1}{8}\frac{\gamma(\gamma+1)(\gamma+2)}{(d^s_t)^3}(\sigma^s_t)^4\mathbb{E}_t\left(\tau^2\right) = 0. \tag{A.18}$$

Then, by Descartes' rule of signs there is exactly one positive real root to the polynomial

$$(d^s_t)^4 - \frac{1}{2}\gamma(\sigma^s_t)^2\mathbb{E}_t(\tau)\,(d^s_t)^2 - \frac{1}{8}\gamma(\gamma+1)(\gamma+2)(\sigma^s_t)^4\mathbb{E}_t\left(\tau^2\right) = 0. \tag{A.19}$$

All four solutions of the quartic polynomial are given by

$$d_t^s = \pm\frac{1}{\sqrt{2}}\sqrt{\frac{\gamma}{2}(\sigma_t^s)^2\mathbb{E}_t\left(\tau\right) \pm \sqrt{\frac{\gamma^2}{4}(\sigma_t^s)^4\mathbb{E}_t\left(\tau\right)^2 + \frac{\gamma(\gamma+1)(\gamma+2)}{2}(\sigma_t^s)^4\mathbb{E}_t\left(\tau^2\right)}}. \quad \text{(A.20)}$$

However, since

$$\frac{\gamma}{2}(\sigma_t^s)^2\mathbb{E}_t\left(\tau\right) < \sqrt{\frac{\gamma^2}{4}(\sigma_t^s)^4\mathbb{E}_t\left(\tau\right)^2 + \frac{\gamma(\gamma+1)(\gamma+2)}{2}(\sigma_t^s)^4\mathbb{E}_t\left(\tau^2\right)} \quad \text{(A.21)}$$

holds for all $\gamma > 0$, $\sigma_t^s > 0$ and $\mathbb{E}_t\left(\tau^2\right) > 0$, the unique positive real root is given by

$$d_t^s = \frac{1}{\sqrt{2}}\sqrt{\frac{\gamma}{2}(\sigma_t^s)^2\mathbb{E}_t\left(\tau\right) + \sqrt{\frac{\gamma^2}{4}(\sigma_t^s)^4\mathbb{E}_t\left(\tau\right)^2 + \frac{\gamma(\gamma+1)(\gamma+2)}{2}(\sigma_t^s)^4\mathbb{E}_t\left(\tau^2\right)}}. \quad \text{(A.22)}$$

□

Proof of Lemma 4. The Taylor representation of $U_\gamma(\tilde{r})$ yields for $\rho^* := \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right)$:

$$\begin{aligned}\mathbb{E}_t\left(U_\gamma(\tilde{r})\right) = {} & \delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s - \rho^* \\ & + \sum_{k=2}^{\infty} \frac{U_\gamma^{(k)}\left(\delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)}{k!U_\gamma'\left(\delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)}\mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - \rho^* - \delta_t^{b,s} - \mathbb{E}_t(\tau)\mu_t^s\right)^k\right).\end{aligned} \quad \text{(A.23)}$$

Let d_t^s be the arbitrage boundary (in absence of transaction costs) as defined in Equation (11). Then, $d_t^s + \ln\left(\frac{1+\rho_t^{b,A}(q)}{1-\rho_t^{s,B}(q)}\right)$ is a root of the function

$$\begin{aligned}\tilde{F}(d) := {} & d + \mathbb{E}_t(\tau)\mu_t^s - \rho^* \\ & + \sum_{k=2}^{\infty} \frac{U_\gamma^{(k)}\left(d + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)}{k!U_\gamma'\left(d + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)}\mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - \rho^* - d - \mathbb{E}_t(\tau)\mu_t^s\right)^k\right).\end{aligned} \quad \text{(A.24)}$$

Therefore, $\mathbb{E}_t\left(U_\gamma(\tilde{r})\right)$ is positive if and only if

$$\delta_t^{b,s} > d_t^s + \ln\left(\frac{1+\rho_t^{b,A}(q)}{1-\rho_t^{s,B}(q)}\right). \tag{A.25}$$

□

Proof of Lemma 5. The proof directly follows from Lemma 4 and Theorem 1. □

Proof of Lemma 6. We cast the arbitrageur's optimization problem in terms of the Lagrangian

$$\begin{aligned}\mathcal{L}(q,f;\xi) =& B_t^s(1-\rho^{s,B}(q))q + A_t^b(1+\rho^{b,A}(q+f))(q+f)\\ &- \xi\left(d_t^s(f) - \delta_t^{b,s} + \log\left(1+\rho^{b,A}(q)\right) - \log\left(1-\rho^{s,B}(q)\right)\right)\end{aligned} \tag{A.26}$$

and observe that the corresponding Karush-Kuhn-Tucker (KKT) conditions imply

$$\begin{aligned} q=0 \quad \vee \quad & B_t^s\left((1-\rho^{s,B}(q)) - \rho^{s,B'}(q)q\right) - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right)\\ & - \xi\left(\frac{\rho^{b,A'}(q+f)}{1+\rho^{b,A}(q+f)} - \frac{\rho^{s,B'}(q)}{1+\rho^{s,B}(q)}\right) = 0 \end{aligned} \tag{A.27}$$

$$\begin{aligned} f=0 \quad \vee \quad & - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right)\\ & - \xi\left(\frac{d}{df}d_t^s(f) + \frac{\rho^{b,A'}(q+f)}{1+\rho^{b,A}(q+f)}\right) = 0 \end{aligned} \tag{A.28}$$

$$\xi = 0 \quad \vee \quad d_t^s(f) - \delta_t^{b,s} + \log\left(1+\rho^{b,A}(q+f)\right) - \log\left(1-\rho^{s,B}(q)\right) = 0, \tag{A.29}$$

We first consider the case of $\xi = 0$. Conditions (A.27) and (A.28) now become

$$\begin{aligned} q=0 \quad \vee \quad & B_t^s\left((1-\rho^{s,B}(q)) - \rho^{s,B'}(q)q\right)\\ & - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right) = 0 \end{aligned} \tag{A.30}$$

$$f=0 \quad \vee \quad - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right) = 0 \tag{A.31}$$

which only holds if

$$1 + \rho^{b,A}(q+f) = -\rho^{b,A'}(q+f)(q+f). \tag{A.32}$$

Since $\rho^{b,A'}(q+f) > 0$ by Assumption 4, this cannot be the case for any $q > 0$ or $f > 0$. Also note that $\xi = q = f = 0$ implies a contradiction. Therefore, the constraint (21) cannot be slack at the optimum and there does not exist a candidate solution for $\xi = 0$.

Next, we turn to the analysis of $\xi > 0$. The simple case of $q = 0$ does not deliver any positive returns and it does not make sense for the arbitrageur to pay any fee $f > 0$. If anything, the arbitrageur would prefer not to trade at all, i.e. $q = f = 0$. We are left with the two interesting cases of $q > 0$.

For $f = 0$, the KKT conditions give the candidate solution $\{q_1, f_1, \xi_1\}$ as solutions to the system of equations

$$B_t^s \left((1 - \rho^{s,B}(q_1)) - \rho^{s,B'}(q_1) q_1 \right) - A_t^b \left((1 + \rho^{b,A}(q_1)) + \rho^{b,A'}(q_1)(q_1) \right) - \xi_1 \left(\frac{\rho^{b,A'}(q_1)}{1 + \rho^{b,A}(q_1)} - \frac{\rho^{s,B'}(q_1)}{1 + \rho^{s,B}(q_1)} \right) = 0 \tag{A.33}$$

$$d_t^s(f_1) - \delta_t^{b,s} + \log\left(1 + \rho^{b,A}(q_1)\right) - \log\left(1 - \rho^{s,B}(q_1)\right) = 0 \tag{A.34}$$

$$f_1 = 0. \tag{A.35}$$

For $f > 0$, we can get the candidate solution $\{q_2, f_2, \xi_2\}$ as solutions to

$$\begin{aligned} & B_t^s \Big((1 - \rho^{s,B}(q_2)) - \rho^{s,B'}(q_2) q_2 \Big) \\ -A_t^b & \Big((1 + \rho^{b,A}(q_2 + f_2)) + \rho^{b,A'}(q_2 + f)(q_2 + f_2) \Big) \\ & -\xi \left(\frac{\rho^{b,A'}(q_2 + f_2)}{1 + \rho^{b,A}(q_2 + f_2)} - \frac{\rho^{s,B'}(q_2)}{1 + \rho^{s,B}(q_2)} \right) = 0 \end{aligned} \tag{A.36}$$

$$\begin{aligned} -A_t^b & \Big((1 + \rho^{b,A}(q_2 + f_2)) + \rho^{b,A'}(q_2 + f_2)(q_2 + f_2) \Big) \\ & -\xi \left(\frac{d}{df} d_t^s(f_2) + \frac{\rho^{b,A'}(q_2 + f_2)}{1 + \rho^{b,A}(q_2 + f_2)} \right) = 0 \end{aligned} \tag{A.37}$$

$$d_t^s(f_2) - \delta_t^{b,s} + \log\big(1 + \rho^{b,A}(q_2 + f_2)\big) - \log\big(1 - \rho^{s,B}(q_2)\big) = 0. \tag{A.38}$$

However, combining (A.36) and (A.36) shows that the solutions are only admissible if

$$\xi = \frac{B_t^s \big((1 - \rho^{s,B}(q_2)) - \rho^{s,B'}(q_2) q_2 \big)}{\frac{d}{df} d_t^s(f_2) - \frac{\rho^{s,B'}(q_2)}{1 + \rho^{s,B}(q_2)}} > 0. \tag{A.39}$$

Equation (A.39) now provides us with necessary conditions for a solution to the problem that entails a strictly positive settlement fee. Namely, $q_2 > 0$, $f_2 > 0$ $\xi_2 > 0$ can only be solution if one of the following two conditions holds

(i) $-\frac{d}{df} d_t^s(f_2) > \frac{\rho^{s,B'}(q_2)}{1 - \rho^{s,B}(q_2)}$ and $1 - \rho^{s,B}(q_2) > \rho^{s,B'}(q_2) q_2$

(ii) $-\frac{d}{df} d_t^s(f_2) < \frac{\rho^{s,B'}(q_2)}{1 - \rho^{s,B}(q_2)}$ and $1 - \rho^{s,B}(q_2) < \rho^{s,B'}(q_2) q_2$.

However, condition (ii) cannot hold at the maximum since $1 - \rho^{s,B}(q_2) < \rho^{s,B'}(q_2) q_2$ means that the trading quantity is such that the marginal price impact exceeds the average price impact. In this case, the arbitrageur would reduce the trading quantity to raise her total return. Consequently, (i) remains as the necessary condition for a candidate solution with a positive settlement fee which completes the proof. □

Proof of Lemma 7. The proof follows directly from applying Theorem 1 together with

the derivatives of the utility function which yields

$$d_t^s - \frac{1}{2}\frac{\gamma}{d_t^s}(\sigma_t^s)^2\mathbb{E}_t(\tau) - \frac{1}{8}\frac{\gamma(\gamma+1)(\gamma+2)}{(d_t^s)^3}(\sigma_t^s)^4\mathbb{E}_t\left(\tau^2\right) = 0. \tag{A.40}$$

Then, by Descartes' rule of signs there is exactly one positive real root to the polynomial

$$(d_t^s)^4 - \frac{1}{2}\gamma(\sigma_t^s)^2\mathbb{E}_t(\tau)(d_t^s)^2 - \frac{1}{8}\gamma(\gamma+1)(\gamma+2)(\sigma_t^s)^4\mathbb{E}_t\left(\tau^2\right) = 0. \tag{A.41}$$

By definition, d_t^s corresponds to the arbitrage boundary for a given risk aversion γ. The arbitrageur prefers to trade if observed price differences $\tilde{\delta}_t^s$ exceed the boundary. Therefore, rewriting Equation (A.41) in terms of γ and replacing d_t^s with $\tilde{\delta}_t^s$ yields a cubic polynomial in γ :

$$\begin{aligned}\left(\tilde{\delta}_t^{b,s}\right)^4 - \frac{1}{8}(\hat{\sigma}_t^s)^4\mathbb{E}_t\left(\tau^2\right)\left(\hat{\gamma}_t^{b,s}\right)^3 - \frac{3}{8}(\hat{\sigma}_t^s)^4\mathbb{E}_t\left(\tau^2\right)\left(\hat{\gamma}_t^{b,s}\right)^2 - \\ \frac{1}{2}(\hat{\sigma}_t^s)^2\left(\mathbb{E}_t(\tau)\left(\tilde{\delta}_t^{b,s}\right)^2 + \frac{1}{2}(\hat{\sigma}_t^s)^2\mathbb{E}_t\left(\tau^2\right)\right)\hat{\gamma}_t^{b,s} = 0\end{aligned} \tag{A.42}$$

Replacing the (conditional) expected latencies with the values given by Equations (33) and (34) completes the proof. □

CENTER FOR FINANCIAL STUDIES

CFS Working Paper Series

Recent Issues

All CFS Working Papers are available at *www.ifk-cfs.de*.

No.	Authors	Title
615	Winfried Koeniger and Marc-Antoine Ramelet	*Home Ownership and Monetary Policy Transmission*
614	Christos Koulovatianos and Dimitris Mavridis	*Increasing Taxes After a Financial Crisis: Not a Bad Idea After All...*
613	John Donaldson, Christos Koulovatianos, Jian Li and Rajnish Mehra	*Demographics and FDI: Lessons from China's One-Child Policy*
612	Hans Gersbach	*Contingent Contracts in Banking: Insurance or Risk Magnification?*
611	Christian Leuz	*Evidence-Based Policymaking: Promise, Challenges and Opportunities for Accounting and Financial Markets Research*
610	Christian Leuz and João Granja	*The Death of a Regulator: Strict Supervision, Bank Lending and Business Activity*
609	Christian Leuz, Steffen Meyer, Maximilian Muhn, Eugene Soltes, and Andreas Hackethal	*Who Falls Prey to the Wolf of Wall Street? Investor Participation in Market Manipulation*
608	Brandon Gipper, Luzi Hail, and Christian Leuz	*On the Economics of Audit Partner Tenure and Rotation: Evidence from PCAOB Data*
607	Vanya Horneff, Raimond Maurer and Olivia S. Mitchell	*Putting the Pension Back in 401(k) Retirement Plans: Optimal versus Default Longevity Income Annuities*

Center for Financial Studies | Goethe University | Theodor-W.-Adorno-Platz 3 | 60323 Frankfurt am Main | Germany
Phone +49 69 798 30050 | Fax +49 69 798 30077 | info@ifk-cfs.de | www.ifk-cfs.de