Building trust takes time:

Limits to arbitrage for blockchain-based assets

January 5, 2023

A blockchain replaces central counterparties with time-consuming consensus protocols to record the transfer of ownership. This settlement latency slows down cross-exchange trading which exposes arbitrageurs to price risk. Off-chain settlement, instead, exposes arbitrageurs to costly default risks. We show with Bitcoin network and order book data that cross-exchange price differences coincide with periods of high settlement latency, asset flows chase arbitrage opportunities, and that price differences across exchanges with low default risks are smaller. Blockchain-based asset trading thus faces a dilemma: Reliable consensus protocols require time-consuming settlement latency which leads to limits to arbitrage. Circumventing such arbitrage costs is possible *only* by reinstalling trusted intermediation which mitigates default risks.

JEL Codes: G00, G10, G14

Keywords: Arbitrage, Blockchain, Market Frictions

1 Introduction

Whenever two investors seek to agree on a financial contract, counterparty risk harms trading if either party may default on its contractual obligations. To guarantee the execution of the negotiated terms, traditional stock markets organize the trading process around trusted intermediaries. Typically, central clearing counterparties bear all counterparty risks between transacting parties during the time span from trade agreement until the legal transfer of ownership through security depositories. Market participants pay for the implied insurance against counterparty risks through fees and collateral deposits.

By contrast, blockchain technology promises to mitigate counterparty risks and render trusted intermediaries obsolete. In such a framework, an open network of validators establishes consensus about transaction histories. Consensus protocols serve as the regulatory framework and specify how validators reach agreement and are incentivized to collaborate. The design of consensus protocols can take different forms to ensure a reliable record of transaction histories. One widely applied consensus mechanisms is proof-of-work, where the validation process involves substantial computational effort and is hence costly to undermine (e.g., Biais et al., 2021).

The design of exchanges as facilitators of the transfer of blockchain-based assets between investors can take different forms. Decentralized exchanges (DEX) enable peer-to-peer trading and thus do not require trust into a middleman. DEXes are essentially smart-contract based algorithms, record every transaction directly on the blockchain and render exchanges pure matchmakers (e.g., Harvey et al. (2021); Lehar and Parlour (2021)). While entirely mitigating counterparty risks, this market structure is not appropriate for large trading volume, which hinders a widespread adoption of DEXes. For example, consider running a traditional limit order book: As each order submission requires blockchain validation, which involves a fee to validators, large message volume renders DEXes comparably slow, inefficient and costly.

In fact, most transactions of crypto tokens rely on fragmented, largely unregulated and so-called centralized exchanges (CEX).[1] A CEX facilitates trades, typically by running a limit order book which is maintained internally and subsequently settle transactions off-chain. Off-chain settlement means that CEXes net executed transactions internally

[1] DEX volume is negligible relative to CEX volume. According to data provider *cryptocompare*, CEXes executed 89% of digital asset volume of approximately $1.04 trillion USD in December 2021. Serious concerns regarding the limits to DEX adoption exist. Among others, Capponi and Jia (2021) and Park (2021) illustrate that DEXes render liquidity provisioning and subsequently trading on private information prohibitively costly.

such that transactions are not processed through the blockchain and validators need not be compensated for executed orderflow. To render off-chain settlement feasible, however, CEXes need to serve as custodians of their customer's funds. Moreover, market fragmentation across multiple CEXes can result in violations of the law of price. Thus, risks and marginal costs for cross-exchange arbitrageurs determine the degree to which such violations from the law of one price can persist.

The main result of this paper is that settlement latency exposes cross-exchange arbitrageurs to substantial costs. These costs arise from the so-called *settlement latency*, defined as the waiting time until validation of a blockchain transaction. Settlement latency makes arbitrage trades costly since it exposes arbitrageurs to substantial price risks: Due to the absence of a globally trusted intermediary, exploiting price differences requires the transfer of a blockchain-based asset *across* CEXes and therefore blockchain validation.

Consider the decision problem of a risk averse cross-exchange arbitrageur who monitors prices of a blockchain-based asset on two CEXes. Whenever she buys on one exchange, she has to wait until blockchain validation of the transfer of the asset before she can sell on the other exchange. Thus, the settlement latency underlying this transfer exposes the arbitrageur to the risk of adverse price movements. Consequently, risk averse arbitrageurs exploit (concurrent) price differences only if these price differences are sufficiently large to compensate for the price risk due to settlement latency.

We derive a closed-form expression for the arbitrageur's certainty equivalent and show that it increases with (*i*) price volatility on the sell-side CEX, (*ii*) the expected settlement latency, (*iii*) the variance of the settlement latency and (*iv*) the arbitrageur's risk aversion. Our characterization of the certainty equivalent also accounts for transaction costs and optimally chosen settlement fees which incentivize validators to enable faster validation (e.g., Easley et al., 2019).

Settlement latency can be effectively circumvented *only* if the arbitrageur deposits arbitrage capital as collateral of the blockchain-based asset under the custody of CEXes. While capital intensive, the arbitrageur can instantaneously acquire a marginal unit at the buy-side exchange and immediately dispose an offsetting amount of her inventory on the sell-side exchange. With such inventory, settlement latency does not affect the trading decision. However, off-chain settlement exposes traders to exchange default risks, which manifests in the risk of thefts, hacks, or exit scams (e.g., Gandal et al. (2018), Cong et al. (2021)).[2]

[2] Biais et al. (2022) document more than 50 hacks and other losses on Bitcoin exchanges and find that

As a result, the dilemma of blockchain-based settlement unfolds as follows: Settlement without trusted intermediation requires a secure validation system. However, a reliable consensus protocol relies on sufficiently high settlement latency (e.g., Hinzen et al., 2019). Marginal costs for arbitrageurs on CEXes are high when settlement latency is large. Ergo: Trustworthy blockchain-based settlement makes investments to mitigate default risks at CEXes more attractive, effectively reinforcing trusted intermediation. As long as CEXes struggle to establish themselves as trusted intermediaries, default risks render settlement latency an inevitable friction hampering arbitrage activity.[3]

Our analysis yields two major empirical predictions: (i) settlement latency is a costly friction for cross-exchange arbitrageurs and (ii) mitigating CEX default risk reduces marginal cross-exchange arbitrage costs. Our empirical analysis rests on novel, granular data to provide compelling evidence for both effects at play. We gather minute-level data from order books of 16 large CEXes that feature trading Bitcoin against US Dollar between January 2018 and October 2019. We analyze violations of the law of one price between each exchange pair. In line with Choi et al. (2018), Borri and Shakhnov (2021) and Makarov and Schoar (2020) we report substantial price differences across the 120 feasible exchange pairs through our sample. To quantify the relevance of settlement latency as a friction, we enrich our dataset by comprehensive high-frequency information about the Bitcoin network, which includes the settlement latency, i.e., the time it takes for every transaction from entering the Bitcoin network until its inclusion in the blockchain.

In line with our theoretical framework, we find that large price differences coincide with periods of high settlement latency, high latency uncertainty and high spot volatility. We find that substantial uncertainty in settlement latency due to the computational effort in Bitcoins proof-of-works mechanism contributes more than 40% of the marginal arbitrage costs. The results are robust when we control for trading costs, optimal validator fee choices and order book liquidity, in line with Roll et al. (2007) and De Long et al. (1990). Settlement latency remains relevant even beyond other well-known frictions that hamper arbitrage activity such as the presence of intermediation facilities, e.g., margin trading (e.g., Pontiff, 1996; Lamont and Thaler, 2003a,b; De Jong et al., 2009).

expected returns of cryptocurrency investors reflect these risks.

[3] Establishing trust requires costly investments in security, governance and regulatory compliance. Indeed, modern CEXes exert substantial effort for such trust-enhancing procedures, i.e., by increasing transparency of cryptocurrency holdings under custody, implementation of funds to repay damages due to security breaches or compliance with strict regulatory standards, as, e.g., the SEC BitLicense. Changpeng Zhao, the CEO of Binance, the world's largest CEX expressed this concern as follows: "By focusing on delivering a superior user experience in tandem with top-notch security, centralized exchanges essentially live or die based on their ability to "create trust" among their users."

To analyze the role of exchange default risks we use the number of Bitcoins under the custody of a CEX as a proxy for trust. This proxy relies on our theoretical analysis on arbitrage capital allocation in presence of settlement latency *and* default risks: CEXes that successfully mitigate default risks should attract more arbitrage capital. In fact, our empirical analysis reveals a strong increase of Bitcoin holdings under the custody of CEXes during our sample period. In October 2019, CEXes held more than 12.4 Billion USD of Bitcoin, an increase of 25.8% relative to the beginning of our sample period in January 2018. In line with our theoretical framework we hypothesize that this success of CEXes to mitigate default risks reduces marginal costs for arbitrageurs.

Consistent with our theoretical predictions, we find that cross-exchange price differences tend to be narrower between CEXes with more funds under custody. Simultaneously, settlement latency remains a significant driver of both the magnitudes of observed cross-exchange price differences and their variation over time, even after controlling for trust. In other words: Violations of the law of one price coincide with periods of high settlement latency even for CEXes which are perceived as highly trustworthy.

Finally, we provide evidence for cross-exchange asset flows chasing price differences but being hampered by settlement latency. We collect Bitcoin wallet IDs that are under the control of the CEXes in our sample and compile a unique and novel data set of 3.9 million cross-exchange transactions with an average daily volume of 72 million US Dollar. We use an instrument for cross-exchange price differences to tackle the inherent endogeneity arising from the simultaneity between price differences and cross-exchange asset flows. For that purpose, our instrument is the theoretical minimum price difference necessary such that the arbitrageur prefers to trade. We estimate these *arbitrage bounds* based on our granular Bitcoin order book and network data. Asset flows into an exchange significantly respond to variations in concurrent price differences, particularly those explained by variations in arbitrage bounds, while also controlling for settlement latency induced price risks and exchange-specific characteristics.

In summary, our empirical results indicate that i) violations of the law of one price coincide with periods of high settlement latency, high settlement latency uncertainty and spot volatility, ii) arbitrageurs perceive settlement latency as an economically relevant friction and act in anticipation of the related price risks and iii) trust mitigates these frictions. In this sense, our paper contributes to the literature on limits to arbitrage (e.g., De Long et al., 1990; Shleifer and Vishny, 1997; Gromb and Vayanos, 2010) by highlighting a friction that arises specifically for blockchain-based assets.

Our results contribute to a better understanding of the economic implications of blockchain technologies for trading on financial markets. In fact, the promise of fast and low-cost transaction settlement lead central banks and marketplaces to actively explore potential applications of such systems for transaction settlement (e.g., BIS, 2017; NASDAQ, 2017; ECB, 2020). The existing literature on blockchain-based asset trading mainly focuses on the incentive compatibility constraints of *validators* which limit decentralization. In particular, Abadi and Brunnermeier (2018) point out a "Blockchain dilemma" in the sense that correctness, decentralization, and cost efficiency cannot be achieved simultaneously for blockchain-based assets. Along this line, Cong et al. (2020) show that risk-averse validators have incentives to pool their mining power which can result in inefficient accumulations of mining capacities. Hinzen et al. (2019) show that network security and fast settlement are mutually exclusive and Pagnotta (2021) finds that miners' competition for block rewards amplifies price volatility and thus has severe pricing implications.

Our focus lies on the economic frictions that originate from the time-consuming effort necessary to mitigate counterparty risks for blockchain-based assets. Our results show that the consensus mechanism itself has direct implications for the marginal costs for arbitrageurs. We shed light on the important open question: to which extent can consensus protocols maintain reliable and secure validation *and* minimal third-party intermediation? Our main conclusion is that settlement latency is a severe economic friction and can be bypassed only by reinstalling core components of trusted intermediation with all associated costs or frictions.

Our results are also relevant from the perspective of platform spillovers for blockchain-based assets. For instance, Capponi and Jia (2021) shows that DEX arbitrage competition can lead to elevated settlement fees and Sokolov (2021) finds that periods of elevated Ransomware activity increase the expected settlement latency. The results in our paper thus predict higher marginal arbitrage costs due to such events which are isomorphic to cross-exchange trading but nevertheless resemble negative externalities on *all* activities conducted on the blockchain, including cross-CEX trading.

The paper proceeds as follows. In Section 2 we derive arbitrage costs due to settlement latency and default risks and characterize the resulting limits to arbitrage capital. Section 3 provides a general theoretical framework to quantify latency related costs in a realistic setup with transaction costs, random settlement latency and optimal trading quantity choices. In Section 4 we introduce the Bitcoin order book and network data, in Section 5 we quantify the core components of our theoretical framework, settlement latency and spot volatility. In Section 6 we provide empirical results and show that settlement latency is an economically relevant friction for blockchain-based assets. Section 7 concludes.

2 Limits to arbitrage for blockchain-based assets

We consider an economy containing a single blockchain-based asset that is traded on two different centralized exchanges (CEX). We define an asset as blockchain-based if the securities' ownership is maintained on a shared database that can be updated without relying on trusted intermediaries or other third-party infrastructure. CEXes keep their order-matching systems off-chain, meaning they operate as escrows for their clients without recording transactions on the blockchain. Off-chain settlement has substantial benefits: First, no incentivizing payments to blockchain validators are needed and second, transaction throughput can be scaled independently of the speed with which validators append new transactions to the blockchain. However, off-chain settlement is criticized as being vulnerable to massive breaches of security and unsafe storage of information, funds, and private keys.

We assume that the trading activity on both exchanges is exogenous, and that agents can monitor the quotes of the asset across exchanges. Each exchange i continuously provides log buy quotes (asks) a_t^i and log sell quotes (bids) b_t^i (with $b_t^i \leq a_t^i$) for one marginal unit of the asset at time t.

Our sole agent is an arbitrageur who aims to exploit observed price differences across exchanges. She intends to buy a marginal unit of the asset on the exchange with the lower buy quote and sell the same amount at the exchange with a higher sell quote. Hence, in case of frictionless trading, the arbitrageur exploits observed price differences when her profits are positive, i.e., whenever

$$\delta_t := \max\left\{b_t^i - a_t^j, b_t^j - a_t^i\right\} > 0. \tag{1}$$

Cross-exchange price differences $\delta_t > 0$ can occur, for instance, as a response to private valuation shocks or asymmetric arrival of information (e.g., Foucault et al. (2017)). We assume throughout the paper that the arbitrageur cannot infer the buy-side exchange before the cross-exchange price difference occurs.

2.1 Cross-exchange arbitrage strategies

We distinguish between two major frictions for trading on price differences of blockchain-based assets across CEXes: risks related to the latency in the settlement process and risks due to defaults of CEXes. To exploit price differences δ_t across two CEXes, the arbitrageur can allocate her arbitrage capital across two possible strategies, which we term *cross-exchange arbitrage* and *inventory arbitrage.* A rational arbitrageur anticipates the implied costs of exploiting price differences and chooses the optimal arbitrage capital allocation.

Cross-exchange arbitrage. Cross-exchange arbitrage requires storing funds of the numéraire at both CEXes. We keep the model parsimonious and assume that the numéraire can be stored without default risks under CEX custody.[4] Deposits can be instantaneously exchanged for the blockchain-based asset on CEXes. If buying on one exchange and selling on the other exchange implies a profit, the arbitrageur buys a marginal unit of the asset on the exchange with the lower buy quote, transfers the asset to the exchange with a higher sell quote and sells the marginal unit as soon as the transfer is settled. Settlement in the context of blockchain-based assets is the validation of a transaction on the blockchain by validators, which indicates that the marginal unit of the asset has been deducted from a wallet under the control of the buy-side CEX and has been credited to a wallet under the control of the desired sell-side CEX. The absence of a clearing house for blockchain-based assets renders it impossible for the arbitrageur to dispose of her position before validation.

We denote the settlement latency τ as the (known) waiting time until a transfer of the asset between exchanges is settled. The simplified setup with known τ serves to illustrate the fundamental relationship between settlement latency and limits to arbitrage. In Section 3 we generalize the framework to random settlement latencies, transaction costs, optimal latency-reducing fee choice and transaction quantities beyond marginal units. However, the main insights of the theoretical framework remain the same.

Because the buy transaction takes place at time t and the transfer of the asset to the sell-side exchange is settled at $t+\tau$, the arbitrageur faces the log sell quote $b^s_{t+\tau}$, $s \in \{i, j\}$. The profit of the arbitrageur's trading decision is thus at risk if the probability of losing

[4] Introducing such an additional source of uncertainty would not affect our results qualitatively because both, cross-exchange and inventory arbitrage, require deposits of the numéraire and are thus are prone to the same additional source of risk. In our empirical analysis we mainly consider USD as safe numéraire deposits which are typically stored on CEX bank accounts outside the premises of vulnerable hacks.

money is non-zero. A risk averse arbitrageur faces limits to (statistical) arbitrage in the spirit of Bondarenko (2003) whenever the associated risk exceeds the expected return. We model the random log price change on the sell-side exchange from time t to $t+\tau$ as a Brownian motion without drift, such that the log return of the cross-exchange arbitrage transaction is

$$r_{(t:t+\tau)} := b^s_{t+\tau} - a^b_t = \underbrace{\delta^{b,s}_t}_{\substack{\text{instantaneous}\\ \text{return}}} + \underbrace{b^s_{t+\tau} - b^s_t}_{\substack{\text{exposure to}\\ \text{price risk}}} = \delta_t + \sqrt{\tau} z, \tag{2}$$

where $z \sim N(0, \sigma^2)$ is normally distributed with spot volatility σ. First, note that a risk-averse arbitrageur with mean-variance utility and risk aversion parameter γ would only exploit price differences δ_t if her certainty equivalent (CE) is positive, i.e.,

$$CE = \mathbb{E}\left(r_{(t:t+\tau)}\right) - \frac{\gamma}{2}\mathbb{V}\left(r_{(t:t+\tau)}\right) = \delta_t - \frac{\gamma}{2}\sigma^2\tau \geq 0 \Longleftrightarrow \delta_t \geq \frac{\gamma\sigma^2\tau}{2}. \tag{3}$$

Equation (3) illustrates that settlement latency implies a risk-reward trade-off for risk-averse arbitrageurs. Whenever the observed price differences δ_t are positive, but CE is negative, the arbitrageur does not trade.[5] In this case, although the trade would be profitable under the possibility of instantaneous settlement, limits to (statistical) arbitrage arise due to settlement latency. The arbitrageur requires higher expected returns if settlement latency τ, the spot volatility σ^2 or risk aversion γ increases.

Inventory arbitrage. CEXes provide effective ways to circumvent settlement latency: To conduct inventory arbitrage, the arbitrageur deposits equal fractions of the available arbitrage capital as numéraire deposits *and* collateral of the blockchain-based risky asset under the custody of each of the two exchanges. Note the important distinction to the numéraire when depositing the blockchain-based asset: such collaterals under the custody of exchanges can be subject to hacking risks. While capital intensive, inventory arbitrage does not expose the arbitrageur to settlement latency. Instead, upon spotting a price difference $\delta_t > 0$, the arbitrageur acquires a marginal unit at the buy-side CEX and immediately disposes an offsetting amount of her inventory on the sell-side CEX.

[5] The risk aversion is associated with the arbitrageur's attitude towards the risk of a single trade. Theoretically, repeatedly exploiting price differences may lead to a vanishing variance of the arbitrageurs' aggregate returns which is equivalent to a contraction of the relevant bounds. However, competition among arbitrageurs or information transmission across exchanges can imply $\mathbb{E}(z) < 0$, which induces limits to arbitrage even for risk-neutral arbitrageurs (e.g., Voigt (2020)).

As a result, her aggregate asset collateral holdings remain constant, but her aggregate numéraire deposits yield a return of δ_t. In that sense, inventory arbitrage strategies effectively render price risks due to settlement latency obsolete. However, this benefit comes at the cost of having to store the blockchain-based asset under the custody of the CEX without retaining corresponding private keys to control the holdings. As such, CEXes undermine the idea of decentralized finance without trusted intermediaries and reintroduce substantial counterparty risks. Similar to Biais et al. (2022) we model default risks as a fraction c of deposits that gets stolen while the arbitrageur waits for the arbitrage opportunity to occur. Due to default risks, profits decrease proportional to wealth and reduce the returns on allocated capital. After arrival of the arbitrage opportunity, the inventory strategy thus delivers aggregate returns of $\delta_t - \mathbb{E}(c)$. We assume that the random variable c exhibits known variance $\mathbb{V}(c) = \sigma_c^2$. Intuitively, higher perceived default risks in terms of higher expected losses due to hacking risks, $\mathbb{E}(c)$ correspond to lower trust into CEXes.

Outside option. As an outside option, the arbitrageur can decide to store her available capital into a private wallet, potentially paying interest $r_f \geq 0$ without any risks.

2.2 Optimal allocation and arbitrage capital limits

For simplicity, we assume that settlement latency is independent of default risks such that $Cov(z, c) = 0$. More specifically, the arbitrageur chooses x_τ and x_c as the fractions of wealth allocated to the cross-exchange arbitrage strategy and the inventory arbitrage strategy. The remaining fraction of wealth, $1 - x_\tau - x_c$ is invested into the outside option. Accordingly, her objective is

$$\max_{x_\tau, x_c} CE\left(x_\tau, x_c\right) = \left(x_\tau + x_c\right)\left(\delta_t - r_f\right) - x_c \mathbb{E}\left(c\right) - \frac{\gamma}{2}\left(x_\tau^2 \sigma^2 \tau + x_c^2 \sigma_c^2\right). \tag{4}$$

The solution to the optimization problem yields the optimal allocation

$$x_\tau = \frac{\delta_t - r_f}{\gamma \sigma^2 \tau} \text{ and } x_c = \frac{\delta_t - \mathbb{E}(c) - r_f}{\gamma \sigma_c^2}. \tag{5}$$

Hence, higher settlement latency (τ) decreases the allocation into the cross-exchange strategy, x_τ, and higher expected default risks $\mathbb{E}(c)$ or default risk uncertainty σ_c^2 reduce the allocation into the inventory strategy, x_c. Everything else equal, the allocation adjustments are not complements, instead, higher risks reduce the aggregate capital allocated into the arbitrage strategies. The fraction of available arbitrage capital that is *not* allocated into the risk-free outside option is

$$x_\tau + x_c = \frac{1}{\gamma \sigma_{c^2}} \left(\frac{\sigma_{c^2} + \sigma^2 \tau}{\sigma^2 \tau} \left(\delta_t - r_f \right) - \mathbb{E}\left(c \right) \right). \tag{6}$$

Hence, the arbitrageur allocates less capital for arbitrage when the risk aversion is higher, the price difference δ_t is smaller, the losses due to hacking risks are higher or the latency risk are higher. Taken together, the simple framework yields the following main insights:

1. Settlement latency implies limits to arbitrage. Higher settlement latency τ, spot volatility σ and risk aversion γ render higher required expected returns before exploiting price differences δ_t becomes incentive-compatible for the arbitrageur.

2. Trust into CEXes in the sense of low perceived default risks $\mathbb{E}\left(c\right)$ attracts more arbitrage capital.

3. Trading volume chases arbitrage opportunities until the resulting price pressure renders marginal trading costs too large.

3 Settlement latency and limits to arbitrage

As shown in a simplified framework in Section 2, settlement latency costs increase with volatility, latency and risk aversion. In this section, we substantially relax the framework to reflect uncertainty in settlement latency, and to allow for general utility functions, transaction costs, order book liquidity and the resulting opportunity to trade optimal quantities. We thus provide a general characterization of the decision framework of a utility maximizing arbitrageurs facing settlement latency. The resulting theoretical framework is sufficiently general to empirically quantify the costs due to latency in order to analyze the economic magnitudes and implications of settlement latency. First, we generalize the return dynamics compared to the simplistic framework from Section 2.

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

$$r^{b,s}_{(t:t+\tau)} = \delta^{b,s}_t + \tau \mu^s_t + \int\limits_t^{t+\tau} \sigma^s_t dW^s_k, \tag{7}$$

where σ_t^s denotes the spot volatility of the bid quote process on the exchange s, and W_k^s denotes a Wiener process. We assume that σ_t^s is constant over the interval $[t, t+\tau]$.[6]

We now let the latency τ denote the *random* waiting time until a transfer of the asset between CEXes is settled. Only weak assumptions regarding the stochastic nature of the settlement latency τ are required.

Assumption 2. *The settlement 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 fully characterize the return distribution $\pi_t\left(r_{(t:t+\tau)}^{b,s}\right)$ through the interval of random length from t to $t+\tau$.

Lemma 1. *Under Assumptions 1 and 2, $r_{(t:t+\tau)}^{b,s}$ exhibits the probability distribution*

$$\pi_t\left(r_{(t:t+\tau)}^{b,s}\right) = \int_{\mathbb{R}_+} \pi_t\left(r_{(t:t+\tau)}^{b,s}|\tau\right) \pi_t(\tau)\, d\tau, \tag{8}$$

and corresponding characteristic function[7]

$$\varphi_{r_{(t:t+\tau)}^{b,s}}(u) = e^{iu\delta_t^{b,s}} m_\tau\left(iu\mu_t^s - \frac{1}{2}u^2(\sigma_t^s)^2\right). \tag{9}$$

Proof. See Appendix A. □

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

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

[6] Time-varying and stochastic volatility can be incorporated by means of a change of the timescale of the underlying Brownian motion. We provide the corresponding derivations in Appendix B. Both the time-variability of σ_t^s and the presence of jumps would further increase the price risk the arbitrageur is facing. In that sense, the bounds derived in this paper are conservative.

[7] The characteristic function fully describes the behavior and properties of a probability distribution. For a random variable X, $\varphi_X(u)$ 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.

The arbitrageur maximizes the expected utility $\mathbb{E}_t\left(U_\gamma(r)\right)$, which we express in terms of the CE. We derive the CE of exploiting concurrent cross-exchange price differences in the following theorem.

Theorem 1. *Under Assumptions 1 - 3, the certainty equivalent (CE) resulting from the cross-exchange arbitrage trade 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 $U_\gamma^{(k)}(r) := \frac{\partial^k}{\partial r^k} U_\gamma(r)$.

Proof. See Appendix A. □

Theorem 1 allows us to compare the expected utility of executing the arbitrage trade versus staying idle (which yields a riskless return of zero). The arbitrageur is willing to exploit cross-exchange price differences if and only if the CE given by Equation (10) is positive.

Definition 1. *We define the arbitrage bound d_t^s as the minimum price difference necessary such that the arbitrageur prefers to trade. Formally, d_t^s is the maximum of zero and the unique root*[8] *of*

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

Definition 1 is a generalization of the arbitrage bounds derived in Equation (3). Below we follow Schneider (2015) and ignore the impact of higher order moments above the fourth degree of the Taylor representation in Equation (11). Under the additional assumption of a power utility function, Lemma 2 provides an analytical closed-form expression for d_t^s.

Lemma 2. *If, in addition to Assumptions 1 and 2, the arbitrageur has an isoelastic utility function $U_\gamma(r) := \frac{(1+r)^{1-\gamma}}{1-\gamma}$ with risk aversion parameter $\gamma > 1$, the arbitrage bound*

[8] By definition of the CE, we have $F(d) = U_\gamma^{-1}\left(\mathbb{E}_t\left(U_\gamma\left(d + \mu_t^s\tau + \int_t^{t+\tau} \sigma_t^s W_k^s\right)\right)\right)$. 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.

for $\mu_t^s = 0$ is given by

$$d_t^s = \frac{1}{2}\sigma_t^s \sqrt{\gamma \mathbb{E}_t\left(\tau\right) + \sqrt{\gamma^2 \mathbb{E}_t\left(\tau\right)^2 + 2\gamma(\gamma+1)(\gamma+2)\left(\mathbb{V}_t(\tau) + \mathbb{E}_t(\tau)^2\right)}}. \tag{12}$$

Proof. See Appendix A. □

Hence, d_t^s positively depends on (i) the arbitrageur's risk aversion, γ, (ii) the local volatility on the sell-side exchange, σ_t^s, (iii) the (conditionally) expected waiting time until settlement, $\mathbb{E}_t\left(\tau\right)$, and (iv) the conditional variance of the waiting time, $\mathbb{V}_t\left(\tau\right)$.

3.1 Transaction costs

Most CEXes demand trading fees that agents pay upon the execution of an off-chain transaction. Market participants typically pay fees as a percentage of the trading volume. Similarly, broker-dealers usually charge markups for the execution of trades in over-the-counter exchanges. Moreover, exchanges 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 exchange i exhibits proportional transaction costs such that the average per unit sell and buy quotes are*

$$B_t^i(q) = B_t^i\left(1 - \rho^{i,B}(q)\right) \tag{13}$$

$$A_t^i(q) = A_t^i\left(1 + \rho^{i,A}(q)\right), \tag{14}$$

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

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

$$\begin{aligned} \tilde{r}_{(t:t+\tau)}^{b,s} &= b_{t+\tau}^s - b_t^s + \delta_t^{b,s} - \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right) \\ &= r_{(t:t+\tau)}^{b,s} - \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right). \end{aligned} \tag{15}$$

Intuitively, Equation (15) shows that transaction costs increase the instantaneous return required to make the arbitrageur indifferent between trading and staying idle. The following lemma summarizes the arbitrageur's decision problem in the presence of transaction costs.

Lemma 3. *Under assumptions 1 - 4, the arbitrageur prefers to trade a quantity $q > 0$ over staying idle if*

$$\delta_t^{b,s} - \log\left(\frac{1 + \rho^{b,A}(q)}{1 - \rho^{s,B}(q)}\right) > d_t^s. \tag{16}$$

Proof. See Appendix A. □

In the context of blockchain-based assets, settlement fees play a pivotal role in the architecture of many consensus protocols. Validators typically receive a reward for confirming transactions which (at least partly) comprises fees that originators of transactions offer to provide validators incentives to prioritize the settlement of transactions that include a higher fee (e.g., Easley et al., 2019).

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 $\tilde{f} > f$.[9] Denote by $d_t^s(f)$ the arbitrage bound 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 settlement 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 exchange, she has to acquire the additional quantity f to spend it in the settlement process. In line with practical implementation in most systems, where cryptocurrencies are transferred, we assume that the arbitrageur has to pay the settlement fee in terms of the underlying asset. Given the transaction costs from above, the choice of f thus also affects the trading quantity q. The following lemma characterizes the arbitrageur's decision problem in the presence of transaction costs and settlement fees.

[9] We refer to Hadar and Russell (1969) and Levy (1992) for an explicit analysis of the relation between stochastic dominance and expected utility.

Lemma 4. *Under assumptions 1 - 5, the arbitrageur prefers to trade a quantity $q > 0$ and pay a settlement fee $f > 0$ over staying idle if*

$$\delta_t^{b,s} - \log\left(\frac{1 + \rho^{b,A}(q + f)}{1 - \rho^{s,B}(q)}\right) > d_t^s(f). \tag{17}$$

Proof. See Appendix A. □

Lemma 2 shows that blockchain congestion directly affects trading decisions of cross-exchange arbitrageurs. Consider a scenario where a blockchain-based platform generates excess demand for validation services which manifests in a substantial increase of settlement fees. Marginal costs for cross-CEX arbitrageurs would increase and thus violations from the law of one price may persist. Such *spillovers* can be induced, for instance, by DEX arbitrage competition (e.g., Capponi and Jia, 2021) or Ransomware activity (e.g., Sokolov, 2021).

3.2 Optimal trading quantity

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

$$\max_{\{q,f\}\in\mathbb{R}_+^2} B_t^s\left(1 - \rho^{s,B}(q)\right) q - A_t^b(1 + \rho^{b,A}(q + f))(q + f) \tag{18}$$

subject to the constraint

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

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

Lemma 5. *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:*

$$\frac{1-\rho^{s,B}(q^*)}{q^*} > \frac{\partial}{\partial q}\rho^{s,B}(q^*) \tag{20}$$

$$-\frac{\partial}{\partial f}d_t^s(f^*) > \frac{\frac{\partial}{\partial q}\rho^{s,B}(q^*)}{1+\rho^{s,B}(q^*)}. \tag{21}$$

Otherwise, the arbitrageur optimally sets $f^* = 0$. Moreover, a total return maximizing arbitrageur chooses trading quantities $q^* > 0$ and settlement fees $f^* \geq 0$ such that

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

Proof. See Appendix A. □

The first part of the lemma provides conditions for the choice of the settlement fee. According to Equation (20), the arbitrageur chooses a positive settlement fee as long as the marginal price impact for the trading quantity is below the average price impact. However, Equation (21) 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. As a consequence, the arbitrageur tends to pay a higher settlement fee if the sell-side exchange is very liquid (keeping the marginal price impact low) and the settlement fee has a high impact on the arbitrage bound (i.e., reducing the latency and thus risk). If any of these two conditions is violated, the arbitrageur optimally chooses not to pay any settlement fee, but might still decide to trade.

The second part of the lemma states that the arbitrageur always chooses trading quantities and settlement fees such that the constraint in Equation (19) binds. If the constraint would not be binding, the arbitrageur could trade a larger quantity to increase her total returns at the expense of higher transaction costs.

4 Bitcoin order book and network data

We gather novel and granular data on Bitcoin, the largest blockchain-based asset in terms of market valuation to assess the economic relevance of settlement latency as a friction for cross-exchange arbitrage.

Testable implications of the theoretical framework are at least threefold: We show in Section 3 that marginal costs for cross-exchange arbitrage increase with settlement latency, latency uncertainty and the volatility of the blockchain-based asset. In Section 2 we find that lower perceived expected losses due to hacking risks render inventory strategies more attractive such that arbitrageurs decide to store funds under custody of the CEX. As a result, order flow should chase cross-exchange price differences as long as these are below marginal costs for arbitrageurs.

We first collect Bitcoin order book data to investigate price differences across a large sample of CEXes at high frequencies and to estimate spot volatility of Bitcoin. Second, we enrich our data with Bitcoin blockchain network data in order to quantify settlement latency. We use the parametrization to provide empirical evidence for each of the main implications of settlement latency for blockchain-based asset trading highlighted above.

4.1 Bitcoin order book data

We gather order book information from the application programming interfaces (APIs) of the 16 largest CEXes in terms of trading volume in January 2018 that feature BTC versus USD trading.[10] We retrieve all open buy and sell orders for the first 25 order book levels on a minute interval from January 1, 2018, to October 31, 2019.

Table 1 gives the corresponding exchanges and provides summary statistics of the underlying order book data of our sample period. We observe a strong heterogeneity of exchange-specific liquidity. For instance, whereas investors could have traded BTC versus USD at *Coinbase Pro* with an average spread of 0.45 USD, the average quoted spread at *Gatecoin* has been about 337 USD. For most exchanges, however, the relative bid-ask spreads are comparable to those from equity exchanges such as NASDAQ or NYSE, where relative spreads range from 5 basis points (bp) for large firms to 38 bp for small firms (e.g., Brogaard et al., 2014).

The exchanges also exhibit substantial heterogeneity in terms of trading-related characteristics. Taker fees range from 0% on *Lykke* to 1% on *Gemini.* Other potential transaction costs are withdrawal fees that have to be paid upon the transfer of BTC from the exchange to any other exchange or private wallet address. Exchanges charge up to 0.003 BTC for withdrawal requests, which corresponds to roughly 30 USD in prices as of January 2018, irrespective of the withdrawn amount. Furthermore, exchanges have

[10] 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. In response to the results documented in Griffin and Shams (2020) substantial doubts on the backing of Tether by USD arose. Our empirical results remain qualitatively unaffected if we adjust for temporaneous price deviations between the USD and USDT (Tether).

Table 1: Descriptive Statistics of the order book Sample

	Order books	Spread (USD)	Spread (bp)	Taker Fee	With. Fee	Conf.	Margin	Business
Binance	941,399	2.61	3.29	0.10	0.00100	2	✓	✗
Bitfinex	938,703	0.62	0.74	0.20	0.00080	3	✓	✓
bitFlyer	919,182	15.13	20.52	0.15	0.00080		✓	✓
Bitstamp	938,483	5.11	6.33	0.25	0.00000	3	✗	✓
Bittrex	940,523	9.07	13.20	0.25	0.00000	2	✗	✓
CEX.IO	936,378	11.73	15.07	0.25	0.00100	3	✓	✓
Gate	907,874	81.24	90.92	0.20	0.00200	2	✗	✗
Gatecoin	560,111	336.52	515.87	0.35	0.00060	6	✗	✓
Coinbase Pro	941,539	0.45	0.54	0.30	0.00000	3	✓	✓
Gemini	912,944	2.57	3.40	1.00	0.00200	3	✗	✓
HitBTC	919,686	2.96	3.68	0.10	0.00085	2	✗	✗
Kraken	936,970	2.63	3.24	0.26	0.00100	6	✓	✓
Liqui	491,516	30.15	45.13	0.25			✓	✗
Lykke	918,768	44.04	57.95	0.00	0.00050	3	✗	✗
Poloniex	916,876	5.38	7.51	0.20		1	✓	✗
xBTCe	887,289	13.34	17.87	0.25	0.00300	3	✓	✗

Notes: This table reports descriptive statistics of orderbook data used in our study. We gather high-frequency orderbook information of 16 exchanges by accessing the public application programming interfaces (APIs) every minute. *Orderbooks* denotes the number of successfully retrieved orderbook snapshots between January 1, 2018 and October 31, 2019. *Spread (USD)* is the average quoted spread in USD, *Spread (bp)* is the average spread relative to the quoted best ask price (in basis points). *Taker Fee* are the associated trading fees in percentage points relative to the trading volume. *With. Fee* are the withdrawal fees in BTC. *Conf.* refers to the number of blocks that the exchange requires to consider incoming transactions as being valid. Empty cells indicate missing values. *Margin* refers to the existence of BTC shorting instruments at the exchange. *Business* indicates whether the exchange allows business accounts and hence access for institutional investors.

different requirements with respect to the number of block confirmations before they proceed to process BTC deposits. For instance, *Kraken* requires that incoming transactions must be included in at least 6 blocks. 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 scenario, 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 (Nakamoto, 2008). As we discuss below, these requirements confront arbitrageurs with a mechanical increase in the settlement latency.

Finally, we collect information about two exchange characteristics that might reduce marginal arbitrage costs. On the one hand, some exchanges offer margin trading instruments which allow traders to take short positions on BTC. However, such margin trading always comes at the cost of substantial collateral deposits which the exchanges control. On the other hand, some exchanges allow businesses to open an account which provides institutional investors, who might have lower risk aversion, with the opportunity to hold inventories and exploit price differences. Holding inventories at exchanges is costly though, since it is associated with continuous exposure to exchange-specific default or hacking risks. We demonstrate in Section 6, that the mere presence of margin trading instruments or access for institutional investors is not a sufficient condition to offset the impact of settlement latency.

4.2 Bitcoin network data

To quantify the settlement latency for Bitcoin, we gather transaction-specific information from blockchain.com, a popular provider of Bitcoin network data. We download all blocks verified between January 1, 2018, and October 31, 2019, and extract information about all verified blockchain transactions in this period. 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.[11]

Any transaction in the Bitcoin network, irrespective of its origin, has to go through the so-called *mempool* which is a collection of all unconfirmed transactions. These transactions wait 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, the Bitcoin protocol restricts the number of transactions that can enter a single block. This restriction induces competition among the originators of transactions who can offer higher settlement fees 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. Relaxing this artificial supply constraint might reduce issues pertaining to settlement latency but at the cost of reduced network security (see, e.g., Hinzen et al., 2019).

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 (during our sample period, new blocks are announced to the network on average every 9.7 minutes). Settlement latency, however,

[11] 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.

Table 2: **Descriptive Statistics of Transactions in the Bitcoin Network**

	Mean	SD	5 %	25 %	Median	75 %	95 %
Fee per Byte (in Satoshi)	47.41	183.08	1.21	5.00	14.06	45.52	200.25
Fee per Transaction (in USD)	1.98	24.19	0.02	0.09	0.28	1.12	7.54
Latency (in Min)	41.03	289.26	0.73	3.55	8.82	20.75	109.52
Mempool Size (in Number)	10,018.74	14,876.52	432.00	1,812.00	4,503.50	11,057.50	41,884.50
Transaction Size (in Bytes)	507.28	2174.13	192.00	225.00	248.00	372.00	958.00

Notes: This table reports descriptive statistics of the Bitcoin transaction data used in our study. The sample contains all transactions settled in the Bitcoin network from January 1, 2018, to October 31, 2019. Our sample comprises 139,704,737 transactions that are verified in 99,129 blocks. *Fee per Byte* is the total fee per transaction divided by the size of the transaction in bytes in Satoshi where 100,000,000 Satoshi are 1 Bitcoin. *Fee per Transaction* is the total settlement fee per transaction (in USD). 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). *Transaction Size* denotes the size of a transaction in bytes. *Mempool Size* is the number of other transactions in the mempool at the time a transaction of our sample enters the mempool.

should not be confused with the time it takes until a new block is mined. Even though the expected block validation time is 10 minutes, it is ex-ante uncertain when a transaction is included in a block for the first time. The number of outstanding transactions serves as a proxy for fluctuations in congestion of the Bitcoin network. Whereas on average 1,644 transactions have been included per block in our sample period, the average number of transactions in the mempool is above 10,000 with temporarily more than 41,000 transactions waiting for verification. For any transaction this induces uncertainty in the settlement latency. The probability of being included in the next block decreases with the number of transactions that wait for settlement and increases with the settlement fee the investor is willing to pay.

Table 2 provides summary statistics of the recorded transactions. The average settlement fee per transaction is about 2 USD. The distribution of fees exhibits a strong positive skewness with a median of 0.28 USD. The average waiting time until the verification of a transaction is about 41 minutes, while the median is about 8.8 minutes.

4.3 Price differences across exchanges

To provide systematic empirical evidence on the extent of violations of the law of one price, we compute the observed instantaneous cross-exchange price differences, adjusted for transaction costs, of all 120 exchange pairs (with the total number of exchanges

$N = 16$), defined as

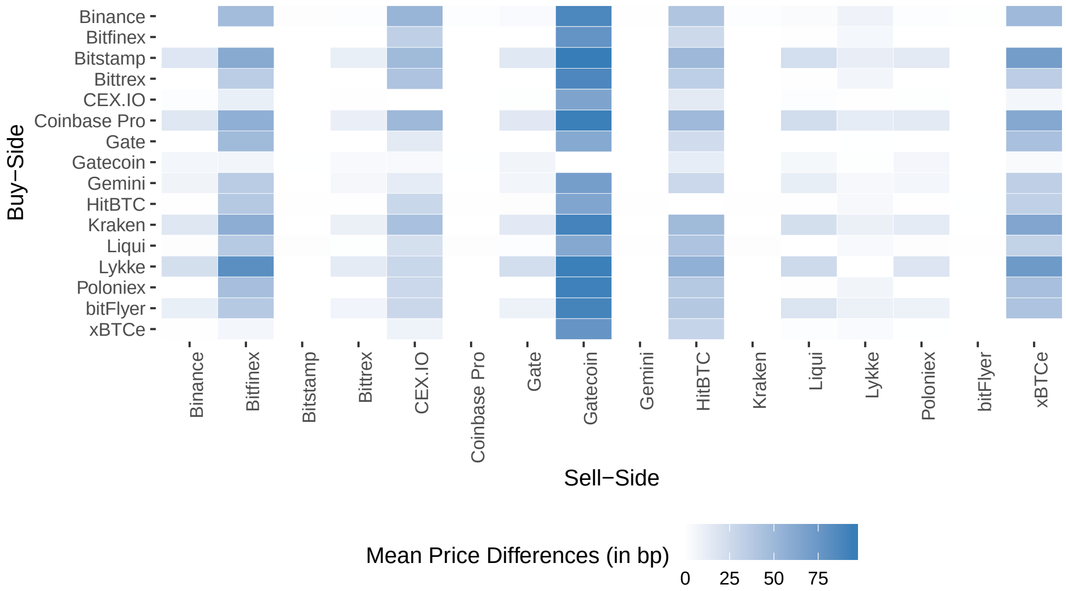
$$\tilde{\Delta}_t := \begin{pmatrix} 0 & \cdots & \tilde{\delta}_t^{N,1} \\ \vdots & \ddots & \vdots \\ \tilde{\delta}_t^{1,N} & \cdots & 0 \end{pmatrix} = \begin{pmatrix} 0 & \cdots & \tilde{b}_t^1\left(q_t^{N,1}\right) - \tilde{a}_t^N\left(q_t^{N,1}\right) \\ \vdots & \ddots & \vdots \\ \tilde{b}_t^N\left(q_t^{1,N}\right) - \tilde{a}_t^1\left(q_t^{1,N}\right) & \cdots & 0 \end{pmatrix}, \quad (23)$$

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.

In line with our definition in Section 3.1, transaction costs are proportional to the trading quantity. We choose $q_t^{i,j}$ as the quantity that maximizes the resulting return for the exchange pair i and j given the prevailing order books at the time t, the taker fees of exchanges i and j and withdrawal fees of exchange j. 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 transaction cost adjusted order book queues and apply a grid search algorithm to identify the trading quantity that maximizes the total return for each exchange pair. As a last step, we check if the resulting trading quantity exceeds the withdrawal fee that the buy-side exchange charges for outgoing transactions (see Table 1). If the optimal trading quantity is below the withdrawal fee, we set the trading quantity to zero. This data-driven approach thus mimics the strategy of an arbitrageur who aims to maximize profits by optimally accounting for the prevailing order book depth and other trading-related fees. As price differences obviously can only be positive in one trading direction, we set negative price differences to zero as such scenarios (even without latency) do not correspond to arbitrage opportunities. The resulting matrix of price differences thus contains only non-negative values.

Figure 1 shows the average price differences for each exchange pair. The heatmap shows that some exchanges exhibit quotes that tend to deviate quite systematically from (nearly) all other exchanges. For instance, *Bitfinex*, *CEX.IO*, *Gatecoin* and *HitBTC* quote on average higher bid prices than most other exchanges and thus exhibit large price differences when used as a sell-side exchange. Conversely, other exchange pairs do not feature large average price differences. For instance, there are hardly any price differences whenever *Coinbase Pro* or *Kraken* serve as sell-side exchanges.

Figure 1: Price Differences between Exchanges



Notes: The heatmap shows the average price differences, adjusted for transaction costs, $\tilde{\delta}_t^{b,s}$, across time for each exchange pair in our sample. Price differences are based on minute-level transaction cost adjusted bids and asks for each exchange according to Equation (23). We account for exchange-specific taker fees according to Table 1 and compute the quantity which maximizes the return for each exchange pair using a grid search algorithm. 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.

5 Expected costs due to settlement latency

Spot volatility σ_t, the expected settlement latency and the variance of settlement latency are the central parameters of our theoretical framework in Section 3. We estimate these values based on our Bitcoin order book and network data.

5.1 Spot volatility

To estimate the spot volatility, we follow the approach of Kristensen (2010). For each exchange s and minute t, we estimate $(\sigma_t^s)^2$ by

$$\widehat{(\sigma_t^s)}^2(h_T) = \sum_{l=1}^{\infty} K(l-t,h_T)\left(b_l^s - b_{l-1}^s\right)^2, \tag{24}$$

where $K(l-t,h_T)$ is a one-sided Gaussian kernel smoother with bandwidth h_T and b_l^s corresponds to the quoted bid price on the exchange s at minute l. The choice of the bandwidth h_T 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_T such that information on day $T-1$ is used for the estimation on day T. Formally, the bandwidth on any day of our sample is the result of minimizing the Integrated Squared Error of estimates on the previous day, i.e.,

$$h_T = \arg\min_{h>0} \sum_{l=1}^{1440} \left[\left(b_l^s - b_{l-1}^s\right)^2 - \widehat{\left(\sigma_l^s\right)}^2(h) \right]^2, \tag{25}$$

where l refers to the minutes on day $T-1$ and $\widehat{\left(\sigma_l^s\right)}^2(h)$ is the spot variance estimator for minute l on day $T-1$ based on bandwidth h.

For each exchange, we trim the distribution of all estimates at 1% on both tails to eliminate outliers (e.g., due to flickering quotes). Since the underlying asset is identical, the resulting estimates—as expected—do not differ substantially across exchanges. The average minute-level volatility across exchanges is about 0.09%, which translates into a daily volatility of about 3.4%, significantly higher than the average daily volatility of the S&P 500 index during the same period, which yields roughly 0.65%.[12]

5.2 Latency prediction

We use all verified transactions to quantify settlement latency of the Bitcoin blockchain. In line with Chiu and Koeppl (2019) and Easley et al. (2019) we expect that transaction fees and mempool congestion play an important role in the determination of the expected time until verification. Accordingly, we employ a Gamma regression, where the conditional probability density function of latency τ_i with rate parameter β_i and shape parameter α_T is given by

$$\pi(\tau_i|\theta_T) = \frac{\beta_i^{\alpha_T}}{\Gamma\left(\alpha_T\right)} \tau_i^{\alpha_T - 1} e^{-\beta_i \tau_i}, \tag{26}$$

where

$$\theta_T := (\theta_T^\beta, \alpha_T)' \in \mathbb{R}^k \text{ and } \beta_i = \exp(-x_i' \theta_T^\beta), \alpha_T > 0. \tag{27}$$

[12] 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., $\sqrt{1440}$.

Here, $x_i \in \mathbb{R}^K$ includes an intercept and denotes (pre-determined) covariates driving τ_i, $\theta_T^\beta \in \mathbb{R}^K$ denotes the corresponding vector of parameters and $\Gamma(x) := \int_{\mathbb{R}_+} z^{x-1}e^{-z}dz$ is the Gamma function. The Gamma distribution collapses to an exponential distribution for $\alpha_T = 1$. We estimate the parameter vector θ_T using all verified transactions on day $T-1$ via maximum likelihood, both with and without covariates. In addition, we estimate an exponential model by fixing $\alpha_T = 1$. As covariates x_i we include settlement fees and the (log) size of the mempool. 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. Blockchain congestion, i.e., the number of transactions waiting for verification at the time when a transaction is announced serves as a proxy for competition among transactions.

In Table 3, we provide summary statistics of the estimated parameters. The numbers in the brackets denote the 5% and 95% quantiles of the time series of estimated parameters. The marginal effect of settlement fees is statistically significant and has the expected sign for nearly all days, i.e., higher fees predict a lower latency. The mempool size exhibits a positive impact on latencies through our sample period, i.e., congestion of the mempool decreases the probability of inclusion of an transaction in the next block (see, e.g., Huberman et al., 2021; Easley et al., 2019). A likelihood ratio test against a model without covariates indicates that the regressors are jointly highly significant. We therefore find clear 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 in nearly 93% of all days.

To predict the (conditional) moments of the latency distribution, while avoiding any look-ahead bias, we use the estimated parameter $\hat{\theta}_T$ based on transactions from day $T-1$ to parameterize the latency distribution for every minute t of day T. We provide further evidence for the predictability of settlement latency by computing the in-sample as well as out-of-sample root mean square prediction errors (MSPEs). In particular, for the in-sample MSPE, we use all transactions that feed into the estimation of $\hat{\theta}_T$ (i.e., all transactions verified on day $T-1$). The out-of-sample MSPE is based on predictions for all transactions verified on day T using the estimated parameter vector $\hat{\theta}_T$. We find that the in-sample MSPE is on average smaller for the unrestricted model specifications and that the unrestricted models exhibit on average a lower out-of-sample MSPE compared to their restricted counterparts. As a consequence, we predict the latency using the unrestricted Gamma model.

Accordingly, the conditional mean and variance of settlement latency τ, induced by

Table 3: Parameter Estimates for the Duration Models

	Exponential		Gamma	
	W/o Covariates	W/ Covariates	W/o Covariates	W/ Covariates
Intercept	3.31	1.41	3.86	1.19
	[2.510, 4.246]	[-0.070, 3.675]	[2.626, 5.250]	[0.013, 2.596]
α			0.62	0.63
			[0.358, 0.902]	[0.365, 0.900]
Fee per Byte		-0.22		-0.22
		[-0.486, -0.031]		[-0.501, -0.031]
Mempool Size		0.23		0.31
		[-0.043, 0.452]		[0.059, 0.530]
LR (Covariates)	91.33		74.59	
LR (Gamma vs. Exponential)	92.68			
MSPE (In-Sample)	65.67	65.74	65.67	66.02
MSPE (Out-of-Sample)	70.97	70.81	70.97	70.55

Notes: This table reports summary statistics for the estimated parameters of the Gamma duration model given by Equation (26). *Fee* denotes fee per byte and *Mempool Size* refers to the number of unconfirmed transactions in the mempool. We estimate each model for each day in our sample, where we consider all transactions confirmed on a particular day. We report the time series averages of the estimated parameters. Values in brackets correspond to the 5% and 95% percent quantiles of the estimated parameters. *LR (Covariates)* summarizes likelihood ratio tests of the corresponding unrestricted duration model with covariates against the restricted model *without* covariates. *LR (Gamma vs. Exponential)* summarizes likelihood ratio tests of the Gamma duration model against the exponential specification. The reported values denote the percentage of days where the null hypothesis that the likelihood of the more general model equals the likelihood of the restricted model is rejected at the 95% significance level. *MSPE* refers to the mean squared prediction error for out-of-sample and in-sample tests, respectively.

a transaction at minute t on day T with characteristics x_t, is given by

$$\widehat{\mathbb{E}}_t(\tau) = \hat{\alpha}_T \exp(x_t'\hat{\theta}_T^{\beta}), \qquad \text{and} \qquad \widehat{\mathbb{V}}_t(\tau) = \hat{\alpha}_T \exp(2x_t'\hat{\theta}_T^{\beta}), \tag{28}$$

where x_t consists of the mempool size and the fee an arbitrageur is willing to pay at time t. While the mempool size is observable at any point in time, we use the optimal fee as derived in Lemma 5 as a proxy for the individually chosen settlement fees.

We cannot reject the null hypothesis that the correlation between volatility and expected latency is significantly different from zero, which suggests that settlement latency constitutes a source of risk which is not captured by price fluctuations. In other words, periods of high spot volatility $(\sigma_t^s)^2$ are not driven by high cross-exchange asset flows. Instead, settlement latency seems to be primarily driven by intraday fluctuation patterns of high network activity which are not necessarily related to cross-exchange arbitrage activity. For instance, Sokolov (2021) finds that network activity spikes during periods of ransomware attacks which increases network fees and thus, according to our theoretical framework, renders cross-exchange arbitrage activity more expensive.

5.3 Estimation of arbitrage bounds

Having estimated spot volatilities $(\sigma_t^s)^2$ as well as the first two moments of the settlement latency distribution, $\widehat{\mathbb{E}}_t(\tau)$ and $\widehat{\mathbb{V}}_t(\tau)$, we analyze the contribution of these components to the arbitrage bounds d_t^s. For that purpose, we estimate the arbitrage bounds according to our theoretical framework in Section 3 and delineate $\hat{d}_t^s$ into individual components.

Based on the empirically relevant CRRA case of Lemma 2, the estimated arbitrage bounds $\hat{d}_t^s$ at minute t are given by

$$\hat{d}_t^s = \frac{1}{2}\hat{\sigma}_t^s\sqrt{\gamma m_1 + \sqrt{\gamma^2 m_1^2 + 2\gamma(\gamma+1)(\gamma+2)m_2}}, \tag{29}$$

with

$$m_1 = \widehat{\mathbb{E}}_t(\tau) + \widehat{\mathbb{E}}_t(\tau_B)\cdot(B^s-1), \tag{30}$$

$$m_2 = \widehat{\mathbb{V}}_t(\tau) + \widehat{\mathbb{V}}_t(\tau_B)\cdot(B^s-1)^2 + \left(\widehat{\mathbb{E}}_t(\tau_B)\cdot(B^s-1) + \widehat{\mathbb{E}}_t(\tau)\right)^2, \tag{31}$$

where $\hat{\sigma}_t^s$ denotes the square-root of the estimated spot volatility on the sell-side exchange, and $\widehat{\mathbb{E}}_t(\tau)$ and $\widehat{\mathbb{V}}_t(\tau)$ denote the estimated conditional mean and variance of the latency distribution, respectively. Moreover, B^s refers to the number of blocks that the sell-side exchange s requires considering incoming transactions as valid (see Table 1). This exchange-specific security requirement thus further increases the settlement latency beyond the waiting time until a transaction's validation in the first block.[13]

We thus decompose the settlement latency into two components: the time it takes until a transaction is included in the blockchain (i.e., the first block), τ, and the additional time τ_B until exchanges accept the transaction as de facto being immutable. While τ is partially under the control of the arbitrageur, the validation time of subsequent blocks is exogenous. 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 partially under control of the Bitcoin

[13] *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.

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 independently and identically distributed. As validators append a new block on average every 9.7 minutes in our sample, we use this magnitude as the best-possible prediction of the time between two subsequent blocks, $\widehat{\mathbb{E}}_t\left(\tau_B\right)$. Accordingly, $\widehat{\mathbb{V}}_t\left(\tau_B\right)$ denotes the (sample) variance of the time between two consecutive blocks.

We fix the coefficient of risk aversion to $\gamma = 2$ and estimate $\hat{d}_t^s$ for each exchange on a minute level.[14] Table 4 gives summary statistics of the resulting time series of arbitrage bounds due to settlement latency. We observe that the estimated bounds range, on average, between 91 bp and 197 bp. We acknowledge that the choice of the risk aversion coefficient determines the level of the estimated arbitrage bound $\hat{d}_t^s$. In that sense, interpreting the magnitude of the arbitrage bound itself is challenging. However, instead of only interpreting $\hat{d}_t^s$ in relation to observed price differences, we can focus on the relative importance of the core components that determine $\hat{d}_t^s$ in Equation (29).

While the conditional moments of the latency distribution affect the time series variation of the bounds, the cross-sectional variation is driven by the exchange-specific spot volatilities and the required number of confirmations, B^s. For instance, *Gatecoin* and *Kraken* require $B^s = 6$ confirmations and produce on average the highest bounds, while *Poloniex* requires only $B^s = 1$ confirmation yielding the smallest median bound. To quantify the effect of the exchange-specific security component B^s, we decompose the arbitrage bounds into the component resulting from the latency until a transaction is included in a block for the first time, τ, and the component resulting from the waiting time until a transaction fulfills exchange-specific security requirements, $(B^s - 1)\tau_B$. The second to last column in Table 4 gives the increase in the median arbitrage bound when we take the exchange-specific number of confirmations into account. The values correspond to the (percentage) difference between the median arbitrage bound as of Equation (29) and the respective bounds based on the assumption $B^s = 1$ for all exchanges. We observe that the impact of exchange-specific security components on arbitrage bounds is substantial and accounts on average for 23% of the bounds.

Moreover, our theoretical framework allows us to directly analyze the relevance of the latency *variance*. As the uncertainty of the arbitrageurs' returns increases with the variance of the settlement latency, we can compare the estimated arbitrage bounds to

[14] Our estimation follows Conine et al. (2017), who estimate an average coefficient of relative risk aversion of about 2 over an extensive sample period.

Table 4: **Summary of Exchange-Specific Arbitrage Bounds**

	Mean	SD	5%	25%	Median	75%	95%	Security	Uncertainty
Binance	114.75	318.76	24.35	42.10	68.92	125.59	320.28	13.54	41.53
Bitfinex	117.22	299.25	18.89	42.47	73.26	136.19	324.19	23.98	40.85
bitFlyer	130.85	317.68	33.02	57.07	86.62	145.18	333.88	24.09	40.72
Bitstamp	126.34	294.72	28.45	50.53	80.46	145.61	341.72	23.69	40.79
Bittrex	129.03	277.80	30.94	57.25	89.41	143.51	333.37	14.32	41.63
CEX.IO	120.84	286.39	29.46	52.72	81.69	136.05	305.50	24.44	40.60
Gate	101.50	277.20	24.12	43.81	68.78	117.27	260.03	14.04	41.48
Gatecoin	196.89	219.90	2.62	46.70	118.29	274.82	638.77	45.95	40.26
Coinbase Pro	114.84	305.25	17.89	40.75	71.77	132.79	318.48	24.44	40.68
Gemini	115.36	343.30	21.07	43.27	72.42	130.54	309.53	24.44	40.77
HitBTC	101.22	287.97	19.10	37.64	62.72	112.79	273.14	14.14	41.36
Kraken	135.07	271.66	25.37	54.09	91.53	164.15	357.11	41.86	40.50
Liqui	90.79	60.20	23.51	49.96	77.40	115.62	201.88	28.97	39.98
Lykke	133.43	379.31	18.58	44.51	80.57	150.73	381.17	25.21	40.61
Poloniex	94.69	264.09	18.49	33.32	55.53	104.34	260.68	0.00	45.13
xBTCe	106.16	246.56	19.90	40.74	70.58	131.44	281.96	24.15	40.78

Notes: This table provides descriptive statistics of estimated arbitrage bounds for each sell-side exchange. We compute arbitrage bounds for a CRRA utility function with risk aversion parameter $\gamma = 2$. We estimate the bounds using the spot volatility estimator of Kristensen (2010) and out-of-sample predictions of the conditional moments of the latency based on a Gamma duration model. We report all values in basis points (except otherwise noted). *Security* gives the (percentage) contribution of the required number of confirmations to the median arbitrage boundary. *Uncertainty* corresponds to the (percentage) contribution of the uncertainty in latency to the median arbitrage boundary.

the (hypothetical) case of a *deterministic* latency. The last column in Table 4 reports the percentage increase in arbitrage bounds when adjusting for the randomness in latency. The values correspond to the percentage difference between the median arbitrage bound and bounds based on the assumption $\mathbb{V}_t(\tau) = \mathbb{V}_t(\tau_B) = 0$. We find that the impact of the latency volatility is substantial and accounts on average for 41% of the arbitrage bounds.

6 Settlement latency and cross-exchange activity

The preceding analysis demonstrates that the estimated arbitrage bounds are of sizeable magnitudes compared to the cross-exchange price differences in our sample. In this section we investigate the two major empirical predictions of our analysis in depth: Based on our findings in Section 2 we hypothesize, that (i) settlement latency is a costly friction for cross-exchange arbitrageurs and (i) that mitigating CEX default risk reduces marginal cross-exchange arbitrage costs.

6.1 Settlement latency and price differences

First, we investigate the relationship between the observed price difference and expected settlement latency, latency volatility as well as spot volatility. The theoretical analysis in Section 3 yields that marginal costs for arbitrageurs increase with settlement latency. In turn, periods of high price risks for arbitrageurs due to settlement latency are consistent with larger observed price differences.

Table 5 gives the estimation results of linear regressions of hourly averages of cross-exchange price differences of sell-side exchanges on exchange-specific fixed effects and various regressors. In columns (1) and (2), we include the average estimated exchange-specific arbitrage bound or, alternatively, its underlying components, i.e., the average hourly sell-side spot volatility, the hourly median and the variance of realized waiting times of transactions entering the mempool until being included in a block for the first time (where we rescale the variance to have a mean of zero and a standard deviation of one). Consistent with our theoretical framework, we find a statistically significant positive relationship between cross-exchange price differences and the components of the arbitrage bounds. The marginal effect is statistically and economically significant: a 1 bp increase in arbitrage bounds is on average associated with a 0.3 bp increase of price differences. Substituting the estimated arbitrage bounds by their components confirms that large price differences are consistent with periods of high price risk due to settlement latency.

In columns (3) and (4), we interact the estimated arbitrage bounds with sell-side exchange-specific dummy variables indicating whether the exchange offers margin trading instruments (*Margin*) and access for institutional traders (*Business Accounts*). We find that exchanges with margin trading are less sensitive to settlement latency, but still yield a significant relationship between price differences and settlement latency. This means that the costs of margin trading for investors tend to exceed the risk-adjusted latency-implied price risk, presumably due to substantial margin requirements by CEXes, which, in turn, implies higher default risks. Similarly, exchanges, which feature access for institutional traders are less sensitive to arbitrage bounds, consistent with the notion that large institutions are more likely to exhibit a lower risk aversion than individual arbitrageurs.

In the last two columns of Table 5, we control for the number of Bitcoins under the custody of a CEX as a proxy for trust. We extract the number of Bitcoins under the control of wallets that the data provider *glassnode* associates with each sell-side exchange directly from the Bitcoin blockchain. In Section 2 we show, that if investors perceive the risks associated with exchanges as low, they should be willing to store more of their holdings directly under the custody of CEXes. Thus, the observed number of Bitcoins serve as a proxy for the perceived default risks of the CEX.

The time series of inventories at CEXes exhibits an annualized aggregate growth of 13.4%. At of the end of our sample, 12.4 Billion USD worth of Bitcoin were under the custody of CEXes. The data thus indicates that perceived default risks of CEXes decreased during our sample period. It should be noted, however, that whereas some exchanges increased their inventory by large amounts (e.g., 400% at Coinbase, 141% at Binance, and 32% at Bitstamp), some other exchanges face net inventory outflows in our sample period. For cross-exchange trading in such cases, settlement latency thus potentially became more relevant due to higher implied costs of inventory arbitrage strategies. Consistent with our hypothesis, we find in Table 5 that cross-exchange price differences tend to be narrower between CEXes with more funds under custody.

Finally, in line with Roll et al. (2007), we find that trading costs, measured by the magnitude of bid-ask spreads, are an additional significant market friction that is related with higher cross-exchange price differences.

6.2 Settlement latency and cross-exchange flows

In the last step of our analysis, we exploit the estimated arbitrage bounds in order to shed light on the relationship between cross-exchange price differences and transfers of assets between CEXes. Our theoretical analysis implies that during periods of large price differences, i.e., in periods, where price differences likely exceed the arbitrage bounds, transfers of funds between CEXes should increase.

We therefore extend our data by cross-exchange asset flows. Since exchanges are reluctant to provide the identity of their customers, it is virtually impossible to identify actual transactions by arbitrageurs. However, we take the overall transfer of assets between two different exchanges as a measure for the trading activity of cross-exchange arbitrageurs. For each exchange, we thus collect a list of addresses that are likely under the control of the exchanges in our sample.[15] Bitcoin transactions are pseudonymous in the sense that each transaction publicly reveals all addresses associated with the transaction, but it is hard to map these addresses to their respective physical or legal owners. Exchanges typically control a large number of addresses to keep track of individual users'

[15] We thank Sergey Ivliev for his tremendous support on this front.

Table 5: Price Differences and Sources of Price Risk

Dependent Variable:	Price Differences					
	(1)	(2)	(3)	(4)	(5)	(6)
Arbitrage Bound (in %)	0.307***		0.440***	0.442***	0.333***	
	(15.98)		(18.62)	(12.84)	(17.61)	
Spot Volatility (in %)		5.416***				5.659***
		(16.99)				(18.14)
Latency Median (in Min)		0.003***				0.002***
		(3.92)				(3.02)
Latency Variance		0.078***				0.105***
		(3.53)				(4.77)
Arbitrage Bound × Margin			-0.258***			
			(-7.07)			
Arbitrage Bound × Business				-0.220***		
				(-5.38)		
Inventory					-1.349***	-1.349***
					(-60.42)	(-60.54)
Spread (in %)	0.111***	0.075*	0.093**	0.101***	0.099***	0.062
	(2.91)	(1.95)	(2.42)	(2.65)	(2.59)	(1.63)
Exchange Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.162	0.163	0.162	0.162	0.212	0.213
Exchange-Hour Observations	213,984	213,984	213,984	213,984	213,622	213,622

Notes: This table provides OLS estimates based on a regression of hourly average sell-side exchange-specific price differences and the main components of price risk due to stochastic settlement latency. *Price Differences* is the sell-side exchange-specific average hourly price difference from all other exchanges (in percent). *Spot Volatility* is the average hourly sell-side spot volatility estimate based on one-sided Gaussian kernel estimates (Kristensen, 2010). *Latency* denotes the hourly median (variance) of the waiting time of transactions entering the Bitcoin mempool, where we rescale the variance to have a mean of zero and a standard deviation of one. *Arbitrage Bound* corresponds to the average hourly sell-side exchange estimated arbitrage bound.*Margin* is a dummy variable that indicates the availability of margin trading instruments and *Business Accounts* indicates whether exchanges offer access for institutional investors. Inventory is the number of Bitcoins controlled by all wallets associated with the sell-side exchange at hour $t-1$. We compute *Spread* as the hourly sell-side exchange-specific average percentage spread. We report t-statistics based on heteroskedasticity-robust standard errors in parentheses. ***,**, and * indicate statistical significance on the $1\%, 5\%$ and 10% levels (two-tailed), respectively.

assets. However, algorithms are available which link addresses to certain exchanges (e.g., Meiklejohn et al., 2013; Foley et al., 2019). Usually, the matching procedure is based on either having observed an address being advertised to belong to an exchange or by actively sending small amounts of Bitcoin to exchanges. We gather 62.6 million unique exchange addresses which allow us to identify 3.9 million cross-exchange transactions

Table 6: Cross-Exchange Flows and Arbitrage Opportunities

Dependent Variable:	Exchange Inflows (in 100k USD)		Log(Exchange Inflows)	
	(1)	(2)	(3)	(4)
Price Differences (in %)	2.407***	2.525***	0.468***	0.462***
	(17.03)	(15.09)	(17.63)	(15.53)
Spread (in %)	-0.355***	-0.376***	-0.068***	-0.066***
	(-3.74)	(-3.73)	(-3.67)	(-3.61)
Exchange Fixed Effects	Yes	Yes	Yes	Yes
Exchange-Hour Observations	213,984	213,984	213,984	213,984

Notes: This table provides the estimated marginal effects based on a two-stage least square regression of cross-exchange asset flows on price differences and bid-ask spreads. *Inflows* are the average hourly inflows (in BTC) to market s from all other markets in our sample. *Price Differences* denote the price differences on sell-side market s and are the fitted values of the regression outlined in Table 5 and denote price differences on sell-side market s. In columns (1) and (3), we instrument price differences with all components of arbitrage bounds. Columns (2) and (4) correspond to the estimation results where we directly use the estimated arbitrage bounds as an instrument. We compute *Spread* as the hourly sell-side exchange-specific average percentage spread. We report t-statistics based on heteroskedasticity-robust standard errors in parentheses. ***,**, and * indicate statistical significance at the 1%, 5% and 10% levels (two-tailed), respectively.

with an average daily volume of 72 million USD in our sample period.[16]

Table 6 gives the estimates of a two-stage least squares regression of hourly cross-exchange flows on hourly averaged cross-exchange price differences as well as exchange-specific fixed effects and the bid-ask spread as a proxy for trading costs. The dependent variable is the sum of cross-exchange flows into a given exchange per hour, which we use in both absolute numbers and in logarithmic terms to reduce the impact of outliers.[17]

We have to take into account that cross-exchange asset flows and price differences are jointly determined, giving rise to a simultaneity problem. On the one hand, arbitrage activity is expected to increase with higher price differences (in excess of arbitrage bounds). On the other hand, price differences should decrease in response to arbitrage trades as arbitrageurs enforce adjustments towards the law of one price. We therefore instrument the price differences by the estimated arbitrage bounds (columns (2) and (4)) and, alternatively, by their respective components, i.e., the spot volatility, median settlement

[16] We compute the average daily volume by extracting the hourly sum of net flows (inflows to an exchange minus the outflows in BTC) and multiplying it by the hourly average midquote across all exchanges.

[17] Note that for any given hour and exchange, arbitrage opportunities involving a particular exchange might arise using the exchange as a sell-side market in some trades and as a buy-side market in some other trades. Therefore, to distinctly quantify the direction of flows, we compute the aggregate sum of cross-exchange flows *into* a given exchange.

latency and variance of realized latencies (columns (1) and (3)). These variables satisfy the two necessary conditions for the validity of an instrument. First, we find a positive correlation between price differences and arbitrage bounds after controlling for other exogenous variables (see Table 5). Second, the only role arbitrage bounds play in influencing cross-exchange flows is through their effect on the endogenous price differences.

Throughout all specifications, we find a significant positive relationship between cross-exchange flows into an exchange and (instrumented) price differences: a one percentage point increase in price differences is on average associated with a 0.5% increase in asset flows into an exchange in a given hour. These results are robust when we control for bid-ask spreads, which are negatively related to inflows coming from other exchanges. The negative marginal effect of the bid-ask spread is consistent with the notion that higher transaction costs deter arbitrageurs' activity. Hence, the regression results indicate that cross-exchange flows increase in response to larger price differences triggered by larger arbitrage bounds. This provides evidence for arbitrageurs chasing profitable arbitrage opportunities by actively transferring assets across markets.

7 Conclusions

Many market participants believe that blockchain technologies have the potential to radically transform the transfer of assets. Replacing trusted intermediaries and central clearing parties by a blockchain may increase efficiency and security, and may lower transaction costs. However, a new friction emerges for blockchain-based assets as the potential merits come at the cost of latency in the settlement process. The inability to perform quick cross-exchange transactions implies limits to arbitrage as market participants cannot react sufficiently fast to potential violations of the law of one price.

We show that settlement latency implies limits to arbitrage as it is not worth for risk-averse arbitrageurs to exploit cross-exchange price differences in periods with high volatility and long validation times. We formally derive the resulting no-trade price bounds for concave utility functions and a general class of latency distributions.

Using data from the Bitcoin market in 2018 and 2019, we show that price differences remain large during periods of high spot volatility, settlement latency and settlement latency uncertainty. Cross-exchange asset flows chase price differences which indicates that market participants perceive these restrictions. Deviations from the law of one price are hence particularly large during times of high latency-implied price risk.

These results shed some new light on the inherent trade-off between costs and benefits of central clearing versus blockchain-based settlement. While central clearing counterparties take on counterparty risk to guarantee instantaneous trading on non-settled positions, blockchains renders trusted intermediation obsolete. The degree of trustworthiness for blockchain-based assets, however, depends on the complexity of the validation process, which ultimately causes settlement latency. We document that the economic costs of latency-related trading frictions for blockchain-based assets are substantial.

To put trading funds under the custody of an exchange and to rely on an exchange's protection against counterparty risk by providing collateral requires trustworthiness of the exchange. Though we observe an increase in trust in exchanges, measured by the increase in funds under the custody of exchanges, our results indicate that intermediation services are still insufficiently utilized to exploit cross-exchange price differences. In fact, we demonstrate that settlement latency remain a statistically and economically significant driving force of time-varying price differences, even when we control for exchange-specific inventory holdings and margin trading possibilities. These results indicate that circumventing settlement latency via alternative strategies is not sufficiently pervasive to completely offset the impact of arbitrage bounds. A possible reason is a lack of trust in the capabilities of CEXes to serve as central counterparties.

This paper thus contributes to an ongoing debate on the organization of clearing on financial markets and the role of third-party intermediation for reliable settlement systems. Our analysis demonstrates that a decentralized system cannot easily replace central clearing. Removing the frictions (and costs) induced by third-party intermediation cause novel trading frictions with non-trivial implications for pricing. First, limits to arbitrage implied by settlement latency may harm 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 assets, as risk neutral probabilities are not uniquely defined. Third, the implied costs of settlement latency depend on the design of the blockchain and should influence the decision whether to migrate to a decentralized settlement system.

References

Abadi, J. and M. Brunnermeier (2018). Blockchain Economics. Working Paper.

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

Barndorff-Nielsen, O. E., 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 (2021). The Blockchain Folk Theorem. *Review of Financial Studies 32*(5), 1662–1715.

Biais, B., C. Bisiere, M. Bouvard, C. Casamatta, and A. J. Menkveld (2022). Equilibrium Bitcoin Pricing. *The Journal of Finance (forthcoming).*

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. *Review of Financial Studies 16*(3), 875–919.

Borri, N. and K. Shakhnov (2021). The Cross-section of Cryptocurrency Returns. *Review of Asset Pricing Studies (forthcoming).*

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

Capponi, A. and R. Jia (2021). The Adoption of Blockchain-based Decentralized Exchanges. Working paper.

Chiu, J. and T. V. Koeppl (2019). Blockchain-Based Settlement for Asset Trading. *Review of Financial Studies 32*(5), 1716–1753.

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

Cong, L. W., Z. He, and J. Li (2020). Decentralized Mining in Centralized Pools. *The Review of Financial Studies 34*(3), 1191–1235.

Cong, L. W., X. Li, K. Tang, and Y. Yang (2021). Crypto Wash Trading. 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 (2019). From Mining to Markets: The Evolution of Bitcoin Transaction Fees. *Journal of Financial Economics 134*(1), 91–109.

ECB (2020). Report on a digital euro. European Central Bank - Eurosystem.

Foley, S., J. Karlsen, and T. J. Putniņš (2019). Sex, Drugs, and Bitcoin: How Much Illegal Activity is Financed Through Cryptocurrencies? *Review of Financial Studies 32*(5), 1798–1853.

Foucault, T., R. Kozhan, and W. W. Tham (2017). Toxic Arbitrage. *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 95*, 86 – 96.

Griffin, J. M. and A. Shams (2020). Is Bitcoin Really Untethered? *The Journal of Finance 75*(4), 1913–1964.

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. *American Economic Review 59*(1), 25–34.

Harvey, C. R., A. Ramachandran, and J. Santoro (2021). *DeFi and the Future of Finance*. John Wiley & Sons.

Hinzen, F. J., K. John, and F. Saleh (2019). Bitcoin's Limited Adoption Problem. Working Paper.

Huberman, G., J. D. Leshno, and C. Moallemi (2021, 03). Monopoly without a Monopolist: An Economic Analysis of the Bitcoin Payment System. *The Review of Economic Studies 88*(6), 3011–3040.

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.

Lehar, A. and C. A. Parlour (2021). Decentralized Exchanges. Working paper.

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

Makarov, I. and A. Schoar (2020). Trading and Arbitrage in Cryptocurrency Markets. *Journal of Financial Economics 135*(2), 293–319.

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

Meiklejohn, S., M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage (2013). A Fistful of Bitcoins: Characterizing Payments among Men with no Names. In *Proceedings of the 2013 conference on Internet measurement conference*, pp. 127–140. ACM.

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.citigroup.com/citi/news/2017/170522a.htm`.

Pagnotta, E. S. (2021, 01). Decentralizing Money: Bitcoin Prices and Blockchain Security. *The Review of Financial Studies 35*(2), 866–907.

Park, A. (2021). The Conceptual Flaws of Constant Product Automated Market Making. Working Paper.

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.

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.

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

Sokolov, K. (2021). Ransomware Activity and Blockchain Congestion. *Journal of Financial Economics 141*(2), 771–782.

Voigt, S. (2020). Liquidity and Price Informativeness in Blockchain-Based Markets. Working Paper.

Appendix

A Proofs

Proof of Lemma 1. The proof of the lemma is an application of Equation (2.2) in Barndorff-Nielsen et al. (1982). □

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

$$\begin{aligned}\mathbb{E}_t\left(r^{b,s}_{(t:t+\tau)}\right) &= (-i)\frac{\partial}{\partial u}\varphi_{r^{b,s}_{(t:t+\tau)}}(u)\bigg|_{u=0}\\ &= \delta^{b,s}_t e^{iu\delta^{b,s}_t} m_\tau\left(iu\mu^s_t - \frac{1}{2}u^2(\sigma^s_t)^2\right)\\ &\qquad + e^{iu\delta^{b,s}_t} m'_\tau\left(iu\mu^s_t - \frac{1}{2}u^2(\sigma^s_t)^2\right)\left(\mu^s_t + iu(\sigma^s_t)^2\right)\bigg|_{u=0}\\ &= \delta^{b,s}_t + \mathbb{E}_t(\tau)\mu^s_t, \end{aligned} \tag{A1}$$

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

In the 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

$$U_\gamma\left(r^{b,s}_{(t:t+\tau)}\right) = \sum_{k=0}^{\infty} \frac{U^{(k)}_\gamma\left(\mu_r\right)}{k!}\left(r^{b,s}_{(t:t+\tau)} - \mu_r\right)^k, \tag{A2}$$

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

$$\mathbb{E}_t\left(U_\gamma\left(r^{b,s}_{(t:t+\tau)}\right)\right) = U_\gamma\left(\mu_r\right) + \sum_{k=2}^{\infty} \frac{U^{(k)}_\gamma\left(\mu_r\right)}{k!}\mathbb{E}_t\left(\left(r^{b,s}_{(t:t+\tau)} - \mu_r\right)^k\right). \tag{A3}$$

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:

$$\mathbb{E}_t\left(U_\gamma\left(r^{b,s}_{(t:t+\tau)}\right)\right) = U_\gamma\left(CE\right) = U_\gamma\left(\mu_r\right) + U'_\gamma\left(\mu_r\right)\left(CE - \mu_r\right). \tag{A4}$$

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 (A3) and (A4), plugging in (A1), and solving for CE. □

Proof of Lemma 2. The proof follows directly from applying Theorem 1 together with the derivatives of the isoelastic utility function which yields

$$d^s_t - \frac{1}{2}\frac{\gamma}{d^s_t}(\sigma^s_t)^2\mathbb{E}_t\left(\tau\right) - \frac{1}{8}\frac{\gamma(\gamma+1)(\gamma+2)}{(d^s_t)^3}(\sigma^s_t)^4\mathbb{E}_t\left(\tau^2\right) = 0. \tag{A5}$$

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

$$(d^s_t)^4 - \frac{1}{2}\gamma(\sigma^s_t)^2\mathbb{E}_t\left(\tau\right)(d^s_t)^2 - \frac{1}{8}\gamma(\gamma+1)(\gamma+2)(\sigma^s_t)^4\mathbb{E}_t\left(\tau^2\right) = 0. \tag{A6}$$

All four solutions of the quartic polynomial are given by

$$d_t^s = \pm \frac{1}{\sqrt{2}} \sqrt{\frac{\gamma}{2}(\sigma_t^s)^2 \mathbb{E}_t\left(\tau\right) \pm \sqrt{\frac{\gamma^2}{4}(\sigma_t^s)^4 \mathbb{E}_t\left(\tau\right)^2 + \frac{\gamma(\gamma+1)(\gamma+2)}{2}(\sigma_t^s)^4 \mathbb{E}_t\left(\tau^2\right)}}. \tag{A7}$$

However, since

$$\frac{\gamma}{2}(\sigma_t^s)^2 \mathbb{E}_t\left(\tau\right) < \sqrt{\frac{\gamma^2}{4}(\sigma_t^s)^4 \mathbb{E}_t\left(\tau\right)^2 + \frac{\gamma(\gamma+1)(\gamma+2)}{2}(\sigma_t^s)^4 \mathbb{E}_t\left(\tau^2\right)} \tag{A8}$$

holds for all $\gamma > 0$, $\sigma_t^s > 0$ and $\mathbb{E}_t\left(\tau^2\right) > 0$, the expression in the lemma gives the unique positive real root. □

Proof of Lemma 3. The Taylor representation of $U_\gamma(\tilde{r})$ yields for $\rho^* := \log\left(\frac{1+\rho^{b,A}(q)}{1-\rho^{s,B}(q)}\right)$:

$$\begin{aligned}\mathbb{E}_t\left(U_\gamma(\tilde{r})\right) = {} & \delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s - \rho^* \\ & + \sum_{k=2}^{\infty} \frac{U_\gamma^{(k)}\left(\delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)}{k! U_\gamma'\left(\delta_t^{b,s} + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)} \mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - \rho^* - \delta_t^{b,s} - \mathbb{E}_t(\tau)\mu_t^s\right)^k\right).\end{aligned} \tag{A9}$$

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

$$\begin{aligned}\tilde{F}(d) := {} & d + \mathbb{E}_t(\tau)\mu_t^s - \rho^* \\ & + \sum_{k=2}^{\infty} \frac{U_\gamma^{(k)}\left(d + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)}{k! U_\gamma'\left(d + \mathbb{E}_t(\tau)\mu_t^s - \rho^*\right)} \mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - \rho^* - d - \mathbb{E}_t(\tau)\mu_t^s\right)^k\right).\end{aligned} \tag{A10}$$

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

$$\delta_t^{b,s} > d_t^s + \ln\left(\frac{1+\rho_t^{b,A}(q)}{1-\rho_t^{s,B}(q)}\right). \tag{A11}$$

□

Proof of Lemma 4. The proof directly follows from Lemma 3 and Theorem 1. □

Proof of Lemma 5. We cast the arbitrageur's optimization problem in terms of the Lagrangian

$$\begin{aligned}\mathcal{L}(q,f;\xi) =& B_t^s(1-\rho^{s,B}(q))q + A_t^b(1+\rho^{b,A}(q+f))(q+f) \\ & - \xi\left(d_t^s(f) - \delta_t^{b,s} + \log\left(1+\rho^{b,A}(q)\right) - \log\left(1-\rho^{s,B}(q)\right)\right) \end{aligned} \tag{A12}$$

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

$$\begin{aligned} q = 0 \quad \vee \quad & B_t^s\left((1-\rho^{s,B}(q)) - \rho^{s,B'}(q)q\right) \\ & - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right) \\ & - \xi\left(\frac{\rho^{b,A'}(q+f)}{1+\rho^{b,A}(q+f)} - \frac{\rho^{s,B'}(q)}{1+\rho^{s,B}(q)}\right) = 0 \end{aligned} \tag{A13}$$

$$\begin{aligned} f = 0 \quad \vee \quad & - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right) \\ & - \xi\left(\frac{d}{df}d_t^s(f) + \frac{\rho^{b,A'}(q+f)}{1+\rho^{b,A}(q+f)}\right) = 0 \end{aligned} \tag{A14}$$

$$\begin{aligned} \xi = 0 \quad \vee \quad & d_t^s(f) - \delta_t^{b,s} \\ & + \log\left(1+\rho^{b,A}(q+f)\right) - \log\left(1-\rho^{s,B}(q)\right) = 0, \end{aligned} \tag{A15}$$

We first consider the case of $\xi = 0$. Conditions (A13) and (A14) now become

$$\begin{aligned} q = 0 \quad \vee \quad & B_t^s\left((1-\rho^{s,B}(q)) - \rho^{s,B'}(q)q\right) \\ & - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right) = 0 \end{aligned} \tag{A16}$$

$$f = 0 \quad \vee \quad - A_t^b\left((1+\rho^{b,A}(q+f)) + \rho^{b,A'}(q+f)(q+f)\right) = 0 \tag{A17}$$

which only holds if

$$1+\rho^{b,A}(q+f) = -\rho^{b,A'}(q+f)(q+f). \tag{A18}$$

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 (19) 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

$$\begin{aligned}
B^s_t\left((1-\rho^{s,B}(q_1)) - \rho^{s,B'}(q_1)q_1\right) - A^b_t\left((1+\rho^{b,A}(q_1)) + \rho^{b,A'}(q_1)(q_1)\right)& \\
-\xi_1\left(\frac{\rho^{b,A'}(q_1)}{1+\rho^{b,A}(q_1)} - \frac{\rho^{s,B'}(q_1)}{1+\rho^{s,B}(q_1)}\right) &= 0 \qquad \text{(A19)}\\
d^s_t(f_1) - \delta^{b,s}_t + \log\left(1+\rho^{b,A}(q_1)\right) - \log\left(1-\rho^{s,B}(q_1)\right) &= 0 \qquad \text{(A20)}\\
f_1 &= 0. \qquad \text{(A21)}
\end{aligned}$$

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

$$\begin{aligned}
B^s_t\left((1-\rho^{s,B}(q_2)) - \rho^{s,B'}(q_2)q_2\right)& \\
-A^b_t\left((1+\rho^{b,A}(q_2+f_2)) + \rho^{b,A'}(q_2+f)(q_2+f_2)\right)& \\
-\xi\left(\frac{\rho^{b,A'}(q_2+f_2)}{1+\rho^{b,A}(q_2+f_2)} - \frac{\rho^{s,B'}(q_2)}{1+\rho^{s,B}(q_2)}\right) &= 0 \qquad \text{(A22)}\\
-A^b_t\left((1+\rho^{b,A}(q_2+f_2)) + \rho^{b,A'}(q_2+f_2)(q_2+f_2)\right)& \\
-\xi\left(\frac{d}{df}d^s_t(f_2) + \frac{\rho^{b,A'}(q_2+f_2)}{1+\rho^{b,A}(q_2+f_2)}\right) &= 0 \qquad \text{(A23)}\\
d^s_t(f_2) - \delta^{b,s}_t + \log\left(1+\rho^{b,A}(q_2+f_2)\right) - \log\left(1-\rho^{s,B}(q_2)\right) &= 0. \qquad \text{(A24)}
\end{aligned}$$

However, combining (A22) and (A22) shows that the solutions are only admissible if

$$\xi = \frac{B^s_t\left((1-\rho^{s,B}(q_2)) - \rho^{s,B'}(q_2)q_2\right)}{\frac{d}{df}d^s_t(f_2) - \frac{\rho^{s,B'}(q_2)}{1+\rho^{s,B}(q_2)}} > 0. \qquad \text{(A25)}$$

Equation (A25) 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) $-\frac{d}{df}d^s_t(f_2) > \frac{\rho^{s,B'}(q_2)}{1-\rho^{s,B}(q_2)}$ and $1-\rho^{s,B}(q_2) > \rho^{s,B'}(q_2)q_2$

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

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 $\theta(\tau) := \int\limits_t^{t+\tau} (\sigma_k^s)^2 \, dk < \infty \quad \forall \tau$, 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

$$r_{(t:t+\tau)}^{b,s} = \delta_t^{b,s} + \int\limits_t^{t+\tau} \sigma_k^s dW_k^s. \tag{B1}$$

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

$$\mathbb{E}_t\left(\left(r_{(t:t+\tau)}^{b,s} - \delta_t^{b,s}\right)^2\right) = \theta(\tau) - \theta(0) = \mathbb{E}_t\left(W_{\theta(\tau)}^s - W_{\theta(0)}^s\right). \tag{B2}$$

In other words, the time-changed Brownian motion $W_{\theta(t)}^s$ has the same distribution as the log returns given in Equation (B1) (e.g., Durrett, 1984; Barndorff-Nielsen et al., 2002). We can thus rewrite the return process as

$$r_{(t:t+\tau)}^{b,s} = \delta_t^{b,s} + \int\limits_t^{t+\theta(\tau)} dW_k^s, \tag{B3}$$

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 $\tau(\theta)$,

$$\mathbb{P}_t\left(\theta(\tau) = y\right) = \mathbb{P}_t\left(\tau = \theta^{-1}\left(y\right)\right) \quad \forall y > 0. \tag{B4}$$

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

$$\varphi_{\theta(\tau)}\left(u\right) = \mathbb{E}_t\left(e^{i\theta(\tau)u}\right) = \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^\infty \varphi_\tau\left(s\right)e^{-is\tau}dse^{i\theta(\tau)u}d\tau. \tag{B5}$$

Lévy's characterization allows extending these ideas to more general non-deterministic integrands and to stochastic time-changes. Although Equation (B5) allows deriving theoretical arbitrage bounds 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.