ECONOMIC STUDIES

DEPARTMENT OF ECONOMICS
SCHOOL OF BUSINESS, ECONOMICS AND LAW
UNIVERSITY OF GOTHENBURG
255

Decentralized Finance and Central Bank Communication

Magnus Hansson

UNIVERSITY OF
GOTHENBURG

ISBN 978-91-88199-69-0 (printed)
ISBN 978-91-88199-70-6 (pdf)
ISSN 1651-4289 (printed)
ISSN 1651-4297 (online)

Printed in Sweden,
Stema 2023

SVANENMÄRKET
Trycksak
3041 0234

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisors, Erik Hjalmarsson and Andreas Dzemski, for their guidance, expertise, and encouragement throughout my doctoral studies. Their insightful comments and constructive criticism have been invaluable in shaping my research and improving my work.

I am also grateful to my discussants, Marcin Zamojski and Isaiah Hull, for their constructive feedback and thought-provoking discussions during my final seminars. A special thanks to Randi Hjalmarsson, the head of the PhD programme, for her instrumental advice and support. I also thank Lars Nordén at Stockholm Business School for valuable comments and suggestions on my job market paper.

I would like to thank Måns Söderbom and Martin Holmén for their great work and leadership as heads of the department. I extend my appreciation to the administrators at the University of Gothenburg: Katarina Forsberg, Selma Oliviera, Ann-Christin Räätäri Nyström, Maria Siirak, Margareta Ransgård, and Lars Magnusson, as well as the administrators at the Swedish House of Finance, where I was a visiting PhD student: Jenny Wahlberg Andersson, Ulrika Dellenstedt, Anki Helmer, and Hedvig Mattsson. Your dedication and efforts have greatly facilitated my academic journey.

I am grateful for the support and companionship of my fellow PhD students, especially my cohort: David Bilén, Ville Inkinen, Vũ Tran, and Dominik Elsner. The experiences we shared have greatly enhanced my time in the programme. Additionally, I would like to thank my office roommate, Anna Lindahl, for her frendship and support.

My gratitude goes to the researchers at the Centre for Finance and Department of Economics at the University of Gothenburg, as well as those who participated in the seminar series and the PhD courses. Your diverse perspectives and expertise have enriched my understanding of the field.

I am indebted to the Ethereum community, especially Thomas Jay Rush at TrueBlocks, for their assistance with indexing the Ethereum blockchain. Your generosity in sharing your time and knowledge have greatly contributed to the success of my research.

Finally, I want to sincerely thank my parents, Inger and Göran, my sister, Ellinor, and my partner, Ingrid, for their unwavering support, encouragement, and love.

Magnus Hansson
Göteborg, April 2023

Contents

Introduction

In this dissertation, I present three distinct chapters that explore different aspects of decentralized finance (DeFi) and central bank communication. Over the past decade, both domains have undergone substantial transformations, and understanding their underlying dynamics and implications have become increasingly important for policymakers and market participants alike. The first two chapters delve into the realm of DeFi, offering insights into the market microstructure of decentralized exchanges (DEXes) that rely on blockchain technology, as well as the various ways in which market participants contribute to price discovery in these markets. The third chapter turns its attention to central bank communication and the theory of narrative economics. By harnessing advancements in natural language processing (NLP), this chapter explores themes that are present in central bank speech communication and examines their evolution over time.

Decentralized finance is an emerging financial ecosystem that employs distributed ledger technologies, like blockchains, to enable peer-to-peer financial transactions without relying on centralized intermediaries such as banks or brokers. Decentralized exchanges (DEXes) leverage blockchain platforms, such as Ethereum (Buterin, 2013; Wood, 2014), to facilitate cryptocurrency trading. In the past few years, these exchanges have experienced a dramatic surge in aggregated daily trading volume, reaching several billion dollars. Crypto markets that operate using blockchain technology display a number of distinct characteristics that set them apart from traditional financial markets. Firstly, all financial data in these markets are fully transparent and permanently stored on the blockchain's "database", granting access to the complete trade and account histories, as well as important metadata. Secondly, as maintaining an order book on a decentralized blockchain is expensive, prominent decentralized exchanges use market-making algorithms to execute trades. These automated market makers revise quotes deterministically based solely on trading. Therefore, both public and private information is incorporated into the price through trades, and a trade's price impact can be calculated precisely. Additionally, as liquidity providers do not revise quotes, there are no "stale quotes" in the traditional sense, and arbitrageurs cannot engage in *toxic arbitrage* along the lines of Foucault, Kozhan, and Tham (2017) and Aquilina, Budish, and O'Neill (2021). Nonetheless, other forms of adversarial activity may still occur. Thirdly, the underlying blockchain technology used by decentralized exchanges processes transactions in groups called *blocks*. For example, on the Ethereum blockchain, these groups of transactions are processed approximately every 14 seconds. As a result, high-frequency trading assumes a different form on the blockchain, with arbitrageurs primarily competing in transaction fees to capture profitable arbitrage opportunities instead of relying on speed. This market design bears resemblance to that proposed in Budish, Cramton, and Shim (2015), which limits the "high-frequency arms race" through a batch auction design. The first two chapters of this thesis contribute with an in-depth understanding of these crucial differences.

Central bank communication has, since the 1990s, transformed into a notably more active instrument for monetary policy. This shift has been particularly evident over the past decade when interest rates have approached the effective lower bound (ELB) of zero, constraining central banks' ability to manage the economy. In such situations, central bank communication and forward guidance play a more critical role, with the public's expectations about future policies becoming increasingly important (Blinder et al., 2008). Traditionally, the objective of central bank communication has been to convey private information to the public and influence and coordinate financial market expectations (Woodford, 2001; Amato, Morris, and Shin, 2002; Blinder et al., 2008). This implies that central bank communication should primarily focus on monetary policy. However, it remains unclear whether this reflects current practice. Narrative economics (Shiller, 2017) proposes that narratives significantly impact the economy and that trends emerge, peak, and decline in a manner similar to diseases in epidemiological models. While natural language processing has previously been used to estimate narratives (Bertsch, Hull, and Zhang, 2021) and narratives have been connected

to central bank communication (Hansen, McMahon, and Tong, 2019), the existence of pronounced trends in central bank communication is still subject to further investigation. The third chapter of this thesis establishes connections between natural language processing, more specifically Dynamic Topic Models (Blei and Lafferty, 2006), central bank communication, and narrative economics.

The first chapter, "Arbitrage in Crypto Markets: An Analysis of Primary Ethereum Blockchain Data", presents a thorough investigation of the role of arbitrageurs in the price discovery process on decentralized exchanges operating on the Ethereum blockchain. These exchanges utilize the automated market maker design and revise quotes according to trading activity. In this context, arbitrageurs are essential in returning the price to the no-arbitrage level following a shock. I classify cross-exchange and triangular arbitrages by identifying sequences of trades, executed by the same agents, that form closed loops. These types of detailed data, documenting completed arbitrages instead of merely price differences between two markets or assets, are difficult to acquire in traditional financial markets. To obtain the necessary data for this study, I run an Ethereum archive node containing the full transaction history of the Ethereum blockchain. Subsequently, I create an index of all transactions, which enables data collection from the decentralized exchanges. To investigate the extent to which decentralized markets adjust after a shock to the no-arbitrage price, I use an empirical methodology consisting of three parts. First, I construct a predictive model to estimate how far back in the trade history exchange rate changes predict arbitrage. Second, I create a counterfactual simulation in which arbitrage transactions are re-executed at different points in the trade history, to measure the length of the profitable arbitrage window. This approach is possible due to the unique features and programmability of the Ethereum blockchain, and allows for the creation of an alternative transaction history. Third, I regress arbitrage profits on previous exchange rate changes to determine the speed at which arbitrageurs profit following a shock to the no-arbitrage price. The findings reveal that the market quickly adjusts after a shock, with a small group of 20 arbitrageurs

(Ethereum accounts) capturing over 75% of all arbitrage profits in the data. This rapid adjustment underscores the importance of arbitrageurs in maintaining price efficiency in these markets, akin to how algorithmic trading in traditional markets reduces the frequency of arbitrage (Chaboud et al., 2014).

The second chapter, "Price Discovery in Constant Product Markets," presents a comprehensive analysis of the market microstructure of decentralized exchanges. The chapter focuses on how price formation occurs on decentralized exchanges that utilize the constant product rule for market making. The paper features a theoretical component, followed by an empirical section. Hasbrouck (1991) introduce the standard empirical framework for evaluating price discovery, where price revisions and order flow are estimated in a Vector Autoregressive (VAR) system. This framework cannot be applied to constant product markets employing algorithmic market making, because prices are deterministically determined, requiring only order flow to be modelled. In the theoretical portion of this paper, I derive a quadratic relationship that explicitly determines the price effect of trades, revealing their information content. In the empirical section, I follow Benos et al. (2017) and classify three types of trading: Human trading, algorithmic trading, and adversarial trading. This classification is possible due to the transparent nature of decentralized exchange data. The raw data are collected similarly to the first chapter of the thesis, by operating and indexing an Ethereum archive node. Subsequently, trade interactions are estimated in a structural VAR framework, and the impulse response functions (IRFs) – illustrating how different trade groups react to one another – are mapped to price impacts (returns) through the previously derived quadratic relationship. This study highlights the significance of large trades and sophisticated adversarial traders in driving price discovery on Uniswap, the largest decentralized exchange.

The third chapter, "Evolution of Topics in Central Bank Speech Communication", shifts the focus from the world of decentralized finance to central bank communication and narrative economics. This paper investigates the content and evolution of central bank speech

communication from 1997 through 2020, utilizing natural language processing to explore two main questions: (i) What topics do central banks commonly discuss? (ii) How have these topics evolved over time? In this study, I collect data by web-scraping speeches from all central banks associated with the Bank for International Settlements (BIS). This data collection approach is inspired by the methodology employed by Armelius et al. (2020). The data are filtered to include global and talkative institutions, resulting in a dataset of speeches from 9 central banks: Bank of Canada, Bank of England, Bank of Japan, Central Bank of Norway, ECB, Fed (including speeches from the New York Fed), Reserve Bank of Australia, Sweden's Riksbank, and Swiss National Bank. The empirical methodology combines Dynamic Topic Models (DTM) (Blei and Lafferty, 2006) and autoregressive (AR) regressions. DTM enable the estimation of central bank speech content and the evolution of these topics over time. In this setting, a topic represents a specific theme addressed by central banks. Technically, a topic is a probability distribution over words, which can be understood as the likelihood of each word being associated with that particular topic. AR modeling is employed to determine the extent to which the topics exhibit persistence over time. The findings reveal that central banks discuss a wide variety of topics, not all directly related to traditional monetary policy subjects. Moreover, the topics display strong autoregressive properties that cannot be easily attributed to standard financial control variables, suggesting that these topics might be partly driven by narratives.

References

Amato, Jeffery D., Stephen Morris, and Hyun Song Shin (Dec. 2002). "Communication and Monetary Policy". In: *Oxford Review of Economic Policy* 18.4, pp. 495–503 (cit. on p. 2).

Aquilina, Matteo, Eric Budish, and Peter O'Neill (Sept. 2021). "Quantifying the High-Frequency Trading "Arms Race"*". In: *The Quarterly Journal of Economics* 137.1, pp. 493–564 (cit. on p. 2).

Armelius, Hanna, Christoph Bertsch, Isaiah Hull, and Xin Zhang (2020). "Spread the Word: International spillovers from central bank communication". In: *Journal of International Money and Finance* 103, p. 102116 (cit. on p. 5).

Benos, Evangelos, James Brugler, Erik Hjalmarsson, and Filip Zikes (2017). "Interactions among High-Frequency Traders". In: *The Journal of Financial and Quantitative Analysis* 52.4, pp. 1375–1402 (cit. on p. 4).

Bertsch, Christoph, Isaiah Hull, and Xin Zhang (2021). "Narrative fragmentation and the business cycle". In: *Economics Letters* 201, p. 109783 (cit. on p. 2).

Blei, David and John Lafferty (2006). "Dynamic Topic Models". In: ICML '06, pp. 113–120 (cit. on pp. 3, 5).

Blinder, Alan S., Michael Ehrmann, Marcel Fratzscher, Jakob de Haan, and David-Jan Jansen (2008). "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence". In: *Journal of Economic Literature* 46.4, pp. 910–945 (cit. on p. 2).

Budish, Eric, Peter Cramton, and John Shim (July 2015). " The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response *". In: *The Quarterly Journal of Economics* 130.4, pp. 1547–1621 (cit. on p. 2).

Buterin, Vitalik (2013). *A Next-Generation Smart Contract and Decentralized Application Platform*. White paper (cit. on p. 1).

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega (2014). "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market". In: *The Journal of Finance* 69.5, pp. 2045–2084 (cit. on p. 4).

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham (2017). "Toxic Arbitrage". In: *The Review of Financial Studies* 30.4, pp. 1053–1094 (cit. on p. 2).

Hansen, Stephen, Michael McMahon, and Matthew Tong (2019). "The long-run information effect of central bank communication". In: *Journal of Monetary Economics* 108, pp. 185–202 (cit. on p. 3).

Hasbrouck, Joel (1991). "Measuring the Information Content of Stock Trades". In: *The Journal of Finance* 46.1, pp. 179–207 (cit. on p. 4).

Shiller, Robert J. (Apr. 2017). "Narrative Economics". In: *American Economic Review* 107.4, pp. 967–1004 (cit. on p. 2).

Wood, Gavin (2014). *Ethereum: A Secure Decentralised Generalised Transaction Ledger*. Yellow paper (cit. on p. 1).

Woodford, Michael (2001). "Monetary policy in the information economy". In: *Proceedings - Economic Policy Symposium - Jackson Hole*, pp. 297–370 (cit. on p. 2).

Chapter 1

Arbitrage in Crypto Markets: An Analysis of Primary Ethereum Blockchain Data

Abstract:

The Ethereum blockchain is a decentralized computing platform providing peer-to-peer financial services. Decentralized exchanges, which run on the blockchain, enable matching of buyers and sellers without any central third party, and are distinct from the centralized "off-chain" cryptocurrency markets often studied in the literature. The decentralized markets facilitate trade in cryptocurrencies and other digital assets and have daily turnovers of several billion dollars. In this paper, I study how arbitrageurs on the blockchain contribute to price discovery and price efficiency in decentralized "on-chain" markets. I collect a transaction-level dataset of primary data from the Ethereum blockchain and cleanly identify a set of completed cross-exchange and triangular arbitrages. To investigate the speed at which arbitrage opportunities are eliminated, I study how sensitive arbitrage profits are to when the trades execute. I show that most arbitrage profits are made immediately after the occurrence of price anomalies, indicating that decentralized markets adjust fast after a shock to the no-arbitrage price.

1 Introduction

The Ethereum blockchain is a decentralized computing platform, and one of its main use cases is peer-to-peer financial services. *Decentralized exchanges* run on Ethereum and accommodate trading of digital assets, including derivatives of bitcoin and ether (Ethereum's native currency), and cryptocurrencies pegged to fiat currencies (stablecoins). The aggregated daily trading volume of these exchanges is several billion dollars, and trading is settled on the blockchain without intermediaries. As it is expensive to keep distributed order books, decentralized exchanges use algorithmic market makers to determine prices. Conditional on a trade, arbitrageurs can precisely predict changes to these prices due to the transparency of the blockchain. These arbitrageurs help in the price discovery process by trading away price imbalances between exchanges (cross-exchange arbitrage) and between relative exchange rates in multiple currencies (triangular arbitrage).

My paper answers the question: How do arbitrageurs on the blockchain contribute to price efficiency in decentralized markets? Arbitrageurs on the blockchain have been shown to act similarly to high-frequency traders in traditional markets (Daian et al., 2019). However, it is unclear to what extent arbitrageurs contribute to the price discovery process. To investigate this issue, I collect a transaction-level dataset of primary data from the Ethereum blockchain, and track transactions individually. Since Ethereum data are fully transparent, I am able to identify a clean set of pure arbitrage trades, and establish how prior shocks to exchange rates predict arbitrages. To investigate the speed at which arbitrage opportunities are eliminated, I study how sensitive arbitrage profits are to when the trades execute. I show that most arbitrage profits are made almost immediately after the occurrence of price anomalies, indicating that decentralized markets adjust fast after a shock to the no-arbitrage price.

This paper introduces primary decentralized exchange data from the Ethereum blockchain to the financial literature. The main source of the data is Ethereum's largest exchange *Uniswap*. Every trade that

is executed on decentralized exchanges is stored on the blockchain. The transaction data contain detailed information about the trades, including what currencies and amounts are traded, and who made the trades. Multiple trades can be executed within one *atomic* Ethereum transaction. Arbitrageurs are expected to use this functionality and execute all trades related to an arbitrage as a bundle by coding them into one Ethereum transaction. This makes arbitrage on Ethereum as close to risk free as possible, with the caveat that transaction fees still need to be paid for failed arbitrages. Due to the unique transparency of the blockchain data, I am able to classify bundles of trades as arbitrages by identifying two or more trades that form a closed loop, so that the output amount and currency of one trade is equal to the input amount and currency of the next trade. For each arbitrage, metadata such as the number of trades, costs, and profits are gathered. The transparency of the decentralized exchanges also provide data reliability, since no manipulation of the data is possible besides what interventions stem from actual trading.[1] The previous economics literature, studying cryptocurrency arbitrage, has observed price differences on traditional trading exchanges operating outside of the blockchain ecosystem, also called *centralized exchanges*. This literature suggests that there have been large inefficiencies in the cross-border cryptocurrency markets (Makarov and Schoar, 2020), but that the price deviations have become less pronounced since 2018 (Shynkevich, 2021). My paper extends these studies by analyzing price efficiency on decentralized exchanges operating on the blockchain itself and by studying realized arbitrages. A related but distinctly different literature in computer science has studied cryptocurrency arbitrage

[1] Alexander and Dakos (2020) show that many empirical published papers studying cryptocurrency use faulty data. A study filed with the U.S. Securities and Exchange Commission by Bitwise Asset Management (SEC, 2019) found that 95% of all reported trading volume on *off-chain* cryptocurrency exchanges is non-economic "wash" trading. Similarly, Cong et al. (2019) show that many unregulated cryptocurrency exchanges engage in significant wash trading that affects exchange ranking, cryptocurrency exchange rates, and volatility. Victor and Weintraud (2021) find that 30% of all traded tokens on the order book exchanges EtherDelta and IDXE have been subject to wash trading.

with primary blockchain data. Wang et al. (2021b) give an overview of triangular arbitrages on Uniswap, and most related to my paper is the study by Berg et al. (2022), which shows that triangular arbitrages are more common in times of high market volatility and that many trades execute at unfavourable prices.

I find that arbitrages are very sensitive to the order of execution and that arbitrage windows close fast. Transactions on the Ethereum blockchain are executed sequentially in batches called *blocks*. These blocks hold on average 200 Ethereum transactions and are executed on average every 14 seconds. Arbitrageurs eliminate price anomalies fast, often as soon as they arise and often within the blocks. Exchange rate changes from trading immediately prior to the arbitrage are the strongest predictor of arbitrage profits. This implies that prices at the end of the blocks are likely to reflect market prices. Decentralized exchanges use the end-of-block prices as starting prices in the next block (Adams, Zinsmeister, and Robinson, 2020), and traders on decentralized exchanges use these to initiate their orders. Furthermore, as the blockchain is a closed system, not easily connected to the "outside world", the stored end-of-block prices are used by other decentralized finance applications as on-chain reference prices. My findings indicate that these reference prices, observed on decentralized exchanges and used by traders on the blockchain, are likely to be arbitrage-free. The automated arbitrage activity leads to improved price efficiency on the blockchain, similar to how algorithmic trading observed in traditional markets reduces the frequency of arbitrage (Chaboud et al., 2014).

To study the sensitivity of arbitrage timing, I design a counterfactual simulation where bundles of arbitrage trades are re-executed with different timings compared to their original counterparts. By creating an alternative order of the full transaction history, I can evaluate if arbitrage transactions would have been profitable in different states of the world, and accordingly measure how sensitive arbitrages are to timing. By re-executing arbitrage transactions prior to when they initially occurred, I can observe how far back arbitrage transactions are profitable and thus for how long the arbitrage opportunities existed. This gives an indication of how fast (efficient) the market is at

correcting price imbalances. The simulation is made possible because Ethereum transactions are programmatically defined and the transactions' codes can be observed on the blockchain. Therefore, it is possible to re-execute transactions with changed parameters ex-post and thus simulate an alternative reality. This is a unique feature of transparent blockchains, like Ethereum, and is used in this study to form a detailed view of price efficiency.

The counterfactual simulation shows that arbitrages are very time sensitive. Out of all realized arbitrage transactions, 69% are no longer profitable if they are re-executed as the first transaction in their own block. As it is more expensive to execute transactions early in the block, this number is adjusted to 85% if the transaction costs are also adjusted to what they would have been at the first position. In these cases, the arbitrage opportunities occur within the blocks and the arbitrage transactions capitalize on the opportunities shortly afterwards. Practically, this means that arbitrageurs monitor pending transactions and place their own trades to immediately neutralize price imbalances. If prices are observed on a block-level instead of on a transaction-level, these price anomalies are never seen. This means that the majority of the arbitrage profits are made within a window of 14 seconds (one block). As traders on decentralized exchanges observe prices from the previous block, they are therefore likely to observe and trade based on arbitrage-free prices. The analysis thus indicates that most of the price discrepancies are arbitraged away with high efficiency within the blocks, leaving the end-of-block price close to the "equilibrium" price. When re-executing the arbitrage transactions at the beginning of previous blocks, only 10% of the arbitrage opportunities exist 5 blocks back (approximately 1 minute in calendar time).

To empirically investigate how far back exchange rate changes affect arbitrages, I conduct a predictive study in which price imbalances are used to estimate whether or not an arbitrage transaction is likely to occur. In contrast to the classic market maker *agent* studied in the literature (Kyle, 1985; Hasbrouck, 1991), where the market maker can incorporate new public information into the prices, the *algorith-*

mic market maker on decentralized exchanges can only revise prices based on the latest trades. Therefore, arbitrages are solely created by trading that off-sets the no-arbitrage price. Large changes to the exchange rate should therefore signal future arbitrage transactions. To discriminate arbitrage transactions from regular trading, a random control group is constructed, consisting of randomly sampled non-arbitrage transactions. As the decentralized exchange data are fully transparent on a transaction-level basis, the exact price impacts of prior trading to these transactions are calculated. The predictors are defined by changes in the exchange rates caused by trading up to 10 blocks (approximately 2 minutes) prior to the studied transactions.

The predictive exercise shows that price changes from prior trading in the same block, and from trading up to 4 blocks back significantly predict if a transaction is a bundle of arbitrage trades or not. Thus, trading up to, approximately, one minute prior to a transaction helps to predict whether an arbitrage transaction will occur. These results hold for the full sample, as well as when running the regression with 5-month indicator variables, constructed to pick up dynamics in the data.

To evaluate how prior trading triggers arbitrages, arbitrage net profits are regressed on exchange rate changes caused by the trading up to 10 blocks prior to the arbitrage transactions. The analysis indicates to what degree arbitrageurs are able to profit from prior imbalances in the exchange rate and at what speed these opportunities are arbitraged away. This speaks to the efficiency of the decentralized markets, and to what extent these markets are able to track the no-arbitrage price.

When regressing realized arbitrage net profits on previous exchange rate changes, within-block trading is far more important for arbitrage profits than price changes from previous blocks. Trading in the same block as the arbitrage transactions and trading in the previous block are both significantly affecting net profits. However, price changes from within the block have a significantly stronger effect. Arbitrage windows therefore tend to close, on average, in 14 seconds. Furthermore, when the analysis is run using 5-month indicator

variables, the results are stronger for the later part of the sample. Between May 2021 to September 2021 and October 2021 to February 2022, only price changes from within the arbitrage block affect arbitrage net profits. The markets are more efficient and prices adjust faster to the no-arbitrage price after a shock, possibly explained by increased arbitrage competition in the later part of the sample. These results are consistent with the observations from the counterfactual simulation, and again indicate that end-of-block reference prices are likely to be arbitrage-free.

The remainder of the paper is organized as follows: Section 2 gives a background to the Ethereum ecosystem, its native currency ether, and decentralized exchanges operating on the blockchain; Section 3 describes the data collection and arbitrage classification strategy; Section 4 outline the empirical analysis and the results; and Section 5 presents concluding remarks.

2 The Ethereum Blockchain and Decentralized Exchanges

Section 2 provides the necessary background for the rest of the paper. Section 2.1 covers the Ethereum blockchain and its native currency ether, Section 2.2 introduces decentralized exchanges, and Section 2.3 demonstrates how arbitrage is conducted on decentralized exchanges. Readers familiar with the Ethereum blockchain and decentralized exchanges can comfortably skip to Section 2.3.

2.1 Ether and Ethereum

Bitcoin is a distributed ledger allowing users to transact the cryptocurrency bitcoin without any third party, and records the transactions in its *blockchain* database. Harvey et al. (2022) categorize the Ethereum blockchain, together with its main competitors (Solana, Avalanche, Cardano, and Algorand), differently than Bitcoin, since Ethereum extends Bitcoin's distributed ledger to a universal decentralized computing system. Ethereum (Buterin, 2013; Wood, 2014), building on Dwork and Naor (1992), Back (2002), and Nakamoto (2008) amongst others, is in essence a network of computer nodes sharing the same database and global state. The global state is a large data structure that describes the state of the world at a specific time, containing, for instance, account balances for all Ethereum users. Formally, Ethereum is a so-called peer-to-peer replicated state machine capable of executing user transactions without a central agency.[2] The network nodes stay in sync with each other as the global state is updated discretely by an execution model called the Ethereum Virtual Machine.[3] Nodes are operated by network participants running an Ethereum client software that follows the rules of the Ethereum protocol described in Wood (2014).

Figure 1.1 gives a general overview of the route of Ethereum transactions. Users create new transactions and send them to the Ethereum network for validation. The transactions can either be sent privately or publicly. *Miners* batch the transactions into *blocks* and spend computing resources to guarantee that the transactions are valid.[4] Once the block is validated, it is linked to the previous blocks of transactions in a *blockchain*, and the global state is updated. The rest of Section 2.1 explains each part of Figure 1.1 in more detail, and gives an institutional overview of the Ethereum blockchain.

Ethereum transactions

In traditional markets, a transaction is a *completed* agreement between a buyer and a seller for some asset. In the blockchain ecosystem the term transaction is defined more broadly, and refers to a computer instruction that is not completed until it is validated by the network. A financial "transaction" on a blockchain is therefore more similar to

[2]For an overview of cryptocurrencies and decentralized finance see Härdle, Harvey, and Reule (2020), Makarov and Schoar (2022), and Harvey et al. (2022).

[3]This should not be confused with the *Ethereum Virtual Mavericks*.

[4]On the 15th of September 2022, Ethereum changed the consensus mechanism from proof-of-work to proof-of-stake in a software upgrade called the *Merge*. In the current version of this paper only transactions from before the merge are analyzed and the term *miner* is used for this distinction and for simplicity.

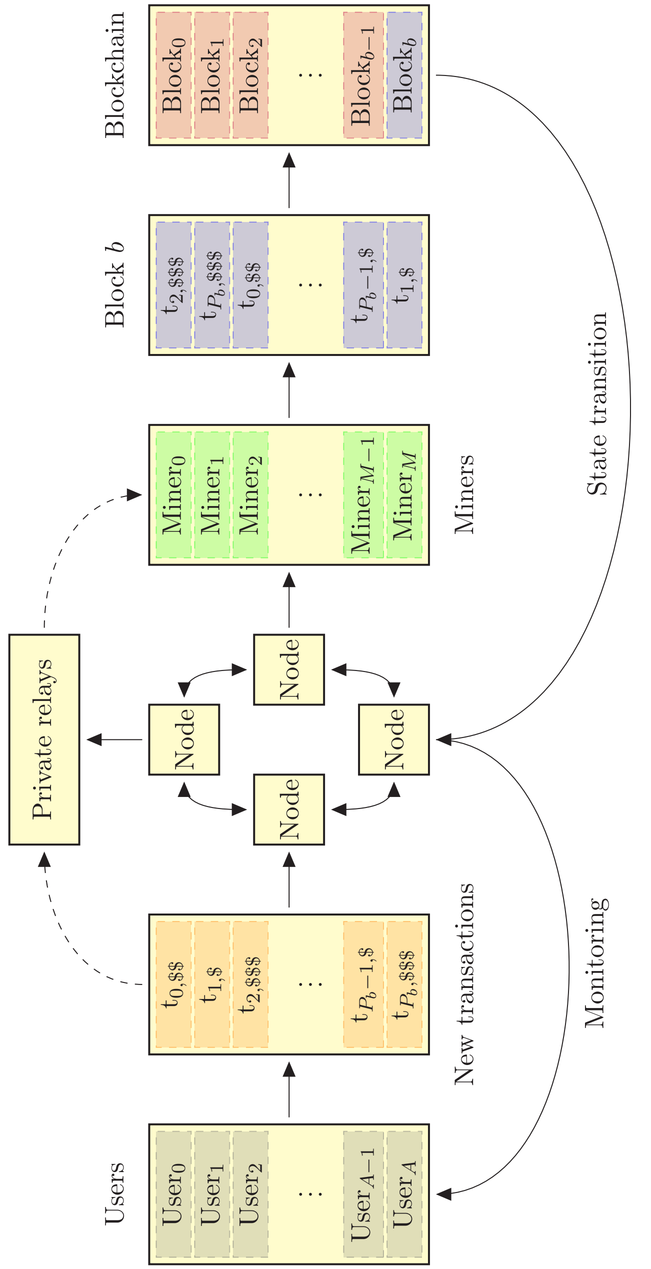


Figure 1.1: The Ethereum blockchain.

an *order* in a traditional market.

Ethereum transactions are computer code that defines a set of instructions for how the Ethereum Virtual Machine should update the global state. In the Ethereum ecosystem transactions are the sole way users interact with the Ethereum network, and thus with each other. These transactions are cryptographically signed by their users and represent one or several of the following actions: Send cryptocurrency to another user, create a *smart contract* (Szabo, 1997), and interact with an already existing smart contract. One Ethereum transaction can hold several instructions at once, for example, to interact with multiple smart contracts within a single transaction.

A regular transaction, transacting currency from one user to another, functions in a similar way to transactions in traditional banks or on the Bitcoin network, where one user can send currency to another. A *smart contract creating transaction* is an Ethereum transaction with embedded computer code that defines a software application. By sending this code to the Ethereum network, the contract is uploaded and stored in the blockchain's "database", and is accessible to every user. All *decentralized applications* on Ethereum are smart contracts. These include *decentralized finance* (DeFi) applications such as decentralized exchanges (DEXes), which facilitate trading of digital assets on the blockchain. The last type of Ethereum transactions are interactions with existing smart contracts. For example, when users trade on decentralized exchanges they send transactions to smart contracts with instructions on how to execute the trades.

Users pay a fee, denominated in Ethereum's native currency ether, to the network to execute their transactions. The size of the fee depends on the complexity of the transaction. Ethereum transactions can take up different amounts of computational work in terms of processing, memory, and network usage. This computational work is called *gas*. The ex post transaction fee is equal to the amount of gas used times the price for each gas unit denominated in ether.[5]

[5]The Ethereum Virtual Machine is a quasi-Turing-complete machine as gas costs bound the total amount of computation.

Atomicity

Ethereum transactions are *atomic*: Either every operation in the transaction is successfully executed or the whole transaction is cancelled. There are three ways an Ethereum transaction can fail: The transaction fee is set too low, a condition in the transaction code is not met, or the transaction is conflicting with the Ethereum state (Zhou et al., 2021a). Users can, therefore, condition transactions so that they only execute if certain conditions are met.

Transaction validation and blocks

Once a transaction is created and cryptographically signed by its user, it is not executed right away. The transaction is sent to a node that shares it with other nodes, and thus throughout the network. The transaction is now public and placed in a queue called the *mempool* (memory pool) to be executed. Special nodes in the network called *miners* compete with each other using processing power to validate transactions in exchange for transactions fees. During the sample period covered by the current analysis, the validation is done by a process called *proof-of-work*. Miners monitor the mempool and batch pending transactions into *blocks*. Miners are incentivized to include transactions that pay the highest transaction fees, and sort these transactions into the block based on their fees. This means that users can pay more to get their transactions executed faster (Zhou et al., 2021b). A block is considered full of transactions when the computational threshold, denominated in gas, is met. Thus, the number of transactions in a block varies depending on how complex they are. When a miner has selected a set of transactions, representing a full block, it tries to validate the block as fast as possible. Different miners can select different transactions from the mempool to try to validate. Transactions that pay a below-market transaction fee might be held in the mempool until network demand goes down and they become profitable for miners to validate. There is no guarantee that pending transactions in the mempool that pay low gas fees will ever be executed.

A new block is validated on average every 14 seconds, but block times over 1 minute are not uncommon. The expected block time is a constant and the average amount of computing power needed to mine a block is dynamically adjusted to match this. Approximately 15 transactions are processed per second, although the variation is high.[6] When a block is mined at a mining node, it is sent out to other nodes to be checked and shared. The nodes verify that the miner's proof is correct, execute the transactions in the new block, and share the block to the rest of the network. When the majority of the nodes in the network have done so, the new global state is agreed upon. The blockchain is immutable and guarantees that the transaction history cannot be changed, as any historical alteration changes the entirety of the blockchain. Every historic transaction on the blockchain must be public and accessible, such that the current state can always be verified.

It is important to note that the Ethereum global state is updated sequentially for each transaction, and each transaction operates independently on the current state. Therefore the order of transactions in a block matters.[7] The financial markets on Ethereum are combinations of continuous and discrete markets, and as it happens not too far from what Budish, Cramton, and Shim (2015) suggest.[8] The discrete update of the Ethereum state ensures that high-frequency competition relying on speed, is not as profitable as in a continuous market. Instead, arbitrageurs primarily compete with transaction fees for block inclusion. The mined transactions are executed discretely, sequentially and independently. However, which transactions are executed and in what order, is subject to the mining process. This process is economically equivalent to a continuous time auction that finishes when a block is mined, and a new auction begins (Daian et al., 2019). In this process speed is an advantage since new and more

[6]Researchers are working on scalability of blockchain systems since there is a trade-off between transaction throughput and decentralized security.

[7]For a technical overview see Appendix 5.

[8]The authors propose a batch auction where trades are executed simultaneously at the same price.

complex arbitrages can be found faster. However, it is not as important as in a continuous limit order book design, where milliseconds can decide if a trade is profitable or not. On Ethereum, to a large extent, arbitrageurs compete with payments to the miners rather than speed for inclusion in the next block.

Front-running

Since Ethereum transactions are executed sequentially and operate on the Ethereum state independently, the transaction ordering in a block is of high importance. This leads to rent-seeking behavior where users try to front-run lucrative transactions. On Ethereum all information is public and front-running is the practice of acting fast on complex public information.[9] Most front-running activity on Ethereum takes place on decentralized exchanges (Torres, Camino, and State, 2021). As miners sort transactions depending on their transaction costs, users can post transactions with higher transaction fees with the hope of getting their transactions executed before others' and capturing potential profits. Daian et al. (2019) show how transaction cost auctions take place to front-run profitable opportunities and capture what they call *miner extractable value* (MEV).[10][11] Miner extractable value is the profits a miner would be able to extract from a block by re-ordering and inserting transactions. This includes transactions fees, but also non-standard extraction methods such as arbitrages. The total amount of present miner extractable value is unknown, although estimations can be made ex post. Qin, Zhou, and

[9]For an overview of front-running on blockchains see Eskandari, Moosavi, and Clark (2020).

[10]Miner extractable value, is more often called *maximal extractable value*, and sometimes *blockchain extractable value* since not only miners participate in capturing MEV.

[11]This phenomenon has attracted attention from economists. To name one example, researchers at the Bank for International Settlements write that "Miner extractable value is an intrinsic shortcoming of pseudo-anonymous blockchains. Addressing this form of market manipulation may call for new regulatory approaches to this new class of intermediaries." (Bank for International Settlements, 2022), referring to block validators as intermediaries.

Gervais (2021) estimate that front-running profits of $541M USD were extracted during their sample period of 32 months.

Private transactions

Miners have the ability to choose which transactions to batch into blocks and attempt to validate. A transaction can be included in a block as long as it is signed by its user, regardless of whether the miner retrieved the transaction from the public mempool or in any other way. A feature of the Ethereum ecosystem is the ability to relay transactions privately to miners. This provides pre-trade privacy by bypassing the public mempool. At the time of writing, 86% of Ethereum's mining capacity is supplied by miners using private relay functionality that accepts private transactions.[12] Private relays allow the option to send transaction bundles directly to miners. Traders use this to obfuscate transactions to mitigate font-running and miners accept these transactions provided that they pay sufficient gas fees.

A transaction bundle is one or several Ethereum transactions that the miner guarantees will be executed in sequence. This is similar to the atomic functionality of the single Ethereum transaction, but extended over several transactions and not guaranteed on a protocol level. The private relay guarantees that either all transactions are executed or none at all. Private transaction bundles can target a specific block for which it is valid. Furthermore, as all nodes in the network need to execute and verify each transaction in a new block to update the global state, private transactions become public as soon as they are included in a verified block.

2.2 Decentralized exchanges

Market makers, in traditional limit order book exchanges, connect buyers and sellers to accommodate trading, and use centralized technology to keep track of the order book. In contrast, decentralized

[12] `https://docs.flashbots.net/Flashbots-auction/searchers/faq/`

exchanges are smart contracts on the blockchain, operating as autonomous and non-custodial market places without an order book. These exchanges have gained traction and constitute one of the fastest growing sectors within decentralized finance (Makarov and Schoar, 2022). Users send transactions with trading instructions to the exchange, trading is accomplished peer-to-peer and the transactions are settled on the blockchain without intermediaries.

A decentralized exchange consists of three parts: Liquidity providers that deposit liquidity to the exchange in return for trading fees, liquidity takers that trade one asset for another, and a price-discovery mechanism (Zhou et al., 2021b). The core difference distinguishing decentralized exchanges from centralized exchanges is the market making mechanism. It is expensive to keep a decentralized order book. Therefore, decentralized exchanges use market making algorithms to facilitate trading (Angeris et al., 2021).[13] These automated market maker update a price in a predetermined way, based on the latest trade.

Automated market makers

Automated market makers (Savage, 1971; Hanson, 2003; Berg and Proebsting, 2009) are a set of algorithmic markets using scoring rules for market and decision making. Uniswap (Zhang, Chen, and Park, 2018; Adams, Zinsmeister, and Robinson, 2020; Adams et al., 2021) is the largest decentralized exchange and exists as a set of smart contracts uploaded to Ethereum. Uniswap is an automated market maker that uses a *constant product formula* to decide the exchange rate between two currencies, and has been formally shown to track a reference market price (Angeris et al., 2021).

Consider two assets X and Y, and their exchange pair X/Y. The exchange rate is calculated so that the product of the respective liquidity pools is kept constant, as stipulated by,

[13] In addition to centralized and decentralized exchanges, *hybrid* exchanges exist. Hybrid exchanges keep a centralized order book off-chain, but settle trading on-chain.

$$k = x \cdot y. \tag{1.1}$$

Here k is called the *invariant* and represents the depth of the market, and x and y are the amounts of asset X and asset Y deposited by liquidity providers into the liquidity pools.[14] The value of k is initially determined by the liquidity added to the pools when the trading pair was initiated. The invariant, k, can change in three ways: Liquidity providers add or remove liquidity, trading fees after each trade are added to the pools, or by "donations". Any liquidity added to the liquidity pools, so that the ratio of x and y is not kept constant is considered a donation and changes the value of k disproportionately.

Contrary to buyers and sellers on limit order book exchanges, traders on constant product markets do not provide a price for one asset in terms of another. Instead, liquidity providers deposit liquidity to both assets' liquidity pools, and it is up to the liquidity takers, i.e., the traders, to decide which assets to trade given the current exchange rate. If a liquidity taker wants to swap δ_y of asset Y for δ_x of asset X, the liquidity pools are altered so that,

$$k = (x - \delta_x) \cdot (y + \delta_y) \tag{1.2}$$

and they have to pay,

$$\delta_y = \frac{k}{x - \delta_x} - y \tag{1.3}$$

of asset Y at the exchange rate $\frac{\delta_x}{\delta_y}$, as the invariant needs to stay constant.[15][16] The exchange rate converges to the marginal exchange rate as the liquidity pools get sufficiently large,

[14] The assets in the liquidity pools are fairly constant, as liquidity providers do not frequently move their assets across pools (Heimbach, Wang, and Wattenhofer, 2021).

[15] On Uniswap there is also a transaction fee of 0.3% to incentivize liquidity provision that is omitted in Equations 1.2 and 1.6.

[16] The constant product function in Equation 1.2, $k = (x - \delta_x) \cdot (y + \delta_y)$, can be generalized to $f(k) = f(x - \delta_x, y + \delta_y)$, where $f(\cdot)$ is an arbitrary pricing function. For an overview of automated market makers and pricing functions see Xu et al. (2021).

$$\frac{\delta_x}{\delta_y} = \frac{\delta_x}{\frac{k}{x-\delta_x} - y} = \frac{x}{y} - \frac{\delta_x}{y} \to \frac{x}{y} \quad \text{as} \quad \delta_x \to 0. \tag{1.4}$$

The difference between the exchange rate and the marginal exchange rate is the *price slippage*. If the amount traded is significant in relation to the amount in the liquidity pools, price slippage occurs, and this is how arbitrage is created. Cartea, Drissi, and Monga (2022) show that it is sub-optimal to execute large orders in one trade as it leads to price slippage and unbalancing of the exchange rate.

As a numerical example of how trading affects the exchange rate, consider the exchange pair X/Y, and the liquidity pools $x = 1,000$ and $y = 90$. Here the invariant, k, is equal to $1,000 \cdot 90 = 90,000$ and the marginal exchange rate is $\frac{x}{y} = \frac{1,000}{90} = 11.11$. A trader wants to buy 100 of X, and according to the constant product algorithm they have to pay,

$$\delta_y = \frac{k}{x - \delta_x} - y = \frac{90,000}{1,000 - 100} - 90 = 10 \tag{1.5}$$

of Y at the exchange rate $\frac{\delta_x}{\delta_y} = \frac{100}{10} = 10$, which is less favourable than the marginal exchange rate.[17] After the trade, the invariant, k, is still equal to $900 \cdot 100 = 90,000$, the liquidity pools have changed to $x = 900$ and $y = 100$, and the new marginal exchange rate is $\frac{x}{y} = \frac{900}{100} = 9$. If another trader, at this point, also wants to buy 100 of X, they have to pay,

$$\delta_y = \frac{k}{x - \delta_x} - y = \frac{90,000}{900 - 100} - 100 = 12.5 \tag{1.6}$$

of Y, which is more than the first trader, at the exchange rate $\frac{\delta_x}{\delta_y} = \frac{100}{12.5} = 8$. After the trade, the invariant, k, is still equal to $800 \cdot$

[17] For simplicity the example disregards trading costs.

$112.5 = 90,000$, the liquidity pools have changed to $x = 800$ and $y = 112.5$, and the new marginal exchange rate is $\frac{x}{y} = \frac{800}{112.5} = 7.11$. In the example, both traders affect the exchange rate as they alter the liquidity pools, and the second trader pay a higher price, compared to the first trader, for the same amount of currency X.

Crypto assets traded on decentralized exchanges

Crypto assets are digital assets secured by blockchain technology. The most common crypto assets are so-called *cryptocurrencies*. Ether is Ethereum's native currency, but there are thousands of other currencies within the Ethereum ecosystem. These currencies are deployed as smart contracts on the blockchain. The *ERC-20 Token Standard* (Ethereum Request for Comment 20) (Vogelsteller and Buterin, 2015) is a technical standard that allow users to create smart contracts that are fungible, i.e., interchangeable, tokens on the Ethereum blockchain. The ERC-20 standard specifies an interface for how to transfer and use the currency. The standard allows developers to create decentralized applications that can interact universally with every currency of that standard. The ERC-20 currencies can be traded on decentralized exchanges running on the Ethereum blockchain. Two examples of ERC-20 tokens are the two largest cryptocurrencies in market cap after bitcoin and ether, the stablecoins Tether (USDT) and USD Coin (USDC), which are both pegged to the US dollar.

2.3 Arbitrage on decentralized exchanges

Arbitrageurs are part of the price-discovery process of automated market makers, helping to adjust the liquidity pools to reflect the no-arbitrage price. On decentralized exchanges, prices can only change due to liquidity takers, not liquidity providers. A consequence of this is that liquidity providers cannot get sniped in the traditional sense by losing the high-frequency race to update their quotes such that stale quotes are traded on. Arbitrage opportunities arise when trading has sufficiently changed the price in one market (or asset) but not in another. This can happen through price slippage in one large trade, or

several consecutive smaller trades. Deviations from the no-arbitrage price can occur between exchanges in the same currency pair as cross-exchange arbitrage, within an exchange in different currency pairs as triangular arbitrage, or any combination of these.

There are two ways in which arbitrageurs can find profitable opportunities: (i) Observe the exchange rates in the current global state defined by the previously mined block B_b and try to cut in front of all other market participants to capture this opportunity in block B_{b+1}. (ii) Monitor the mempool to detect unconfirmed pending transactions that will affect the market price so that arbitrage opportunities occur and try to cut in front of all other market participant in capturing these opportunities (Daian et al., 2019). Arbitrageurs can predict exchange rate changes from pending transactions as all information about these transactions is public. Both the former and the latter methods are done in a process called *back-running*, where the arbitrageur adjusts the transaction fee such that their arbitrage transaction is executed directly after the transaction causing the price impact.

If the arbitrage opportunity is already manifested in the current state, arbitrageurs aim to have as high a transaction position as possible in the next block. Thus, paying up to the arbitrage profit itself in transaction fee to acquire an early execution in the coming block. If the arbitrage opportunity is pending in the mempool there are two ways in which arbitrageurs can capture the potential profits. First, the arbitrageurs can back-run the transaction by carefully choosing the transaction cost so that the probability of the arbitrage transaction being executed directly after the price-changing transaction is high. Alternatively, the arbitrageurs can "take" the transaction in the mempool, sign their own arbitrage transaction, and send these together directly to miners using private relays. The miners are incentivized to accept these bundles if the fees are sufficiently high. This process guarantees that the transactions are executed in the right order, given that the miner wins the race to confirm the block. Arbitrageurs reputedly use private relays to increase the probability of having their transactions executed directly after a large transaction off-sets the exchange rate, although the exact frequency of this practice is unknown.

Arbitrage atomicity

One significant difference between arbitrage trading on decentralized exchanges and traditional exchanges concerns the risks involved. On traditional exchanges, an arbitrage opportunity can generally only be capitalized on by submitting multiple trades. In this situation, there is a risk that the market moves in the opposite direction and the trader has to sell the position at a loss. Noise traders have been shown to create unpredictable risks that reduce the attractiveness of arbitrage (DeLong et al., 1990).

On the Ethereum blockchain, all legs of an arbitrage strategy can be included in the same transaction in which case no other transaction can interrupt the chain of trades. Furthermore, since Ethereum transactions are programmable, arbitrageurs can condition their transactions such that the transactions are cancelled if they are not profitable.[18] In principle, arbitrage on Ethereum is truly risk-free with the caveat that arbitrageurs still have to pay gas fees for cancelled transactions.

Arbitrageurs' profits and costs

The arbitrageurs have to pay two types of fees: A trading fee to the decentralized exchange, and a fee to the Ethereum network for processing the transaction. The trading fee is typically 0.3% on decentralized exchanges and goes to the liquidity providers for providing assets to the liquidity pools. Additionally, arbitrageurs have to pay gas fees to the Ethereum network to incentivize miners to process the transactions. Arbitrageurs using private relays can pay network fees to the miners either by direct payments or through regular gas payments.[19]

[18] This can be done with a simple condition asserting that the end balance need to be greater than or equal to the start balance.

[19] More about this in Section 3.1.

3 Primary transaction data

I set up and sync an Ethereum archive node (Erigon Team, 2022), which downloads the entire history of the blockchain. The node contains all transactions on the Ethereum network since its inception on the 30th of June 2015 and onwards. Trueblocks (TrueBlocks Team, 2022) is used to build an index of all transactions such that they can be filtered on interactions with decentralized exchanges.

A transaction-level data set is sourced directly from the archive node, consisting of every transaction interacting with the decentralized exchange Uniswap between 29th of July 2020 and 17th of February 2022. 37,856,529 transactions across 63,168 trading pairs, and 84 decentralized exchanges are collected.[20] Thus, each transaction in the dataset has one or more interactions with Uniswap, without restricting any further interactions with additional smart contracts, such as other decentralized exchanges. There are three different software versions of Uniswap. In this paper Uniswap version 2 is used since, during the sample period, it is by far the largest decentralized exchange on the Ethereum blockchain.

3.1 Transaction classification

Overview of the data

Uniswap allows for three main user operations: Trading, adding liquidity, and removing liquidity. From the 37,856,529 transactions interacting with the Uniswap smart contracts, 71% concern regular trading where at least one currency is exchanged for another. Of these, 43% are *simple* trades, where only one currency is exchanged for another. Furthermore, 23% of the transactions are liquidity provisions, depositing assets to liquidity pools, and 1% of the transactions are liquidity withdrawals. This is consistent with previous research, finding that liquidity providers do not frequently move their assets

[20] The transactions are collected from the Ethereum Mainnet network and the dataset includes all decentralized exchanges that use the same *application binary interface* as Uniswap version 2.

between liquidity pools (Heimbach, Wang, and Wattenhofer, 2021). Lastly, 1% of the transactions give the Uniswap smart contracts approval for spending user funds, which is an industry standard for decentralized exchanges and necessary prior to any trading. Thus, approximately 96% of the transactions in the dataset perform standard operations of a decentralized exchange. 4% of the transactions are uncategorized and are possibly combinations of the above operations, or interactions with additional smart contracts on the Ethereum blockchain.

Detecting completed arbitrage transactions

To study price efficiency, I focus on two types of pure atomic arbitrages: Cross-exchange arbitrage and triangular arbitrage. In addition, combinations of the strategies are also included in the analysis. Cross-exchange arbitrages trade on price differences between two or more exchanges. Triangular arbitrage capitalize on price deviations between three or more exchange pairs within the same exchange.[21] The cross-exchange arbitrages all include Uniswap as one of the exchanges, and the triangular arbitrages all occur on Uniswap.

From the decentralized exchange dataset 231,645 completed cross-exchange and triangular arbitrage transactions with associated metadata are extracted. The arbitrage transactions span the same time period as the full dataset, 29th of July 2020 through 17th of February 2022, trade on 82 decentralized exchanges, and across 4,663 cryptocurrency trading pairs. The transactions are publicly recorded on the Ethereum blockchain and for each arbitrage observation the following metadata are collected: Execution time, block, position in block, arbitrageurs' address, trading cost, number of trades, volume of first trade, profit, currency pairs, and decentralized exchanges used. The arbitrage transactions must satisfy the following criteria:

1. Two or more trades need to form a closed loop, such that the output amount and currency of one trade is equal to the input amount and currency of the next trade.

[21] Triangular arbitrage with more than 3 legs is sometimes called cyclic arbitrage.

2. All trades must occur within one atomic Ethereum transaction.

3. The transaction cannot perform or be connected to any other operation on the blockchain.

4. The transaction must yield a positive profit (however not a positive net profit).

5. The base currency must be Ether.

The classification criteria capture a set of clean arbitrage transactions, where the arbitrageurs' risk is bounded by the transaction cost payed to validators for trying to execute the transaction.[22] Criterion 1 assures that the transaction is an arbitrage trade. Any number of trades can take place between any number of exchanges. This covers, for example, cross-exchange arbitrage transactions with two or more trades between Uniswap and any other decentralized exchanges, as well as triangular arbitrages on Uniswap. Criterion 2 ensures that all trades in the transaction are done by the same agent and that the arbitrageurs act rationally by minimizing their risks and costs by executing all trades in one atomic transaction. Criterion 3 ensures that the transaction is a pure arbitrage transaction, meaning that it is not part of any other trading strategy. This criterion removes transactions included in sandwich bundles, flash loans, or any other blockchain operation not part of a pure arbitrage operation.[23] Criterion 4 removes arbitrage transactions with negative profits, however arbitrages with negative net profits are still included in the analysis. Since it is possible for an arbitrage transaction to be programmed such that it is cancelled if it is not profitable, transactions with negative profits are most likely due to operational mistakes. Or, the transactions are not aimed at capturing arbitrages, but have some other use case. Wang et al. (2021b) find that approximately 1% of cyclic arbitrage transactions have negative profits plausibly due to participation

[22] 31 arbitrage transactions don't pay a fee to the miner are excluded from the analysis.

[23] For a detailed description see Appendix 5.

in other decentralized finance projects, such as increasing trading volume. Lastly, Criterion 5 does not restrict the sample in any essential way, but enables straightforward comparison and profit calculations across transactions. The vast majority of arbitrage transaction fulfill this criterion as arbitrageurs pay transactions costs in ether.

Table 1.1 shows descriptive statistics for the arbitrage transactions. Approximately half of the transactions are triangular arbitrages indicated by the number of exchanges used and the number of trades per transaction. Cross-exchange arbitrages occur on 2 to 4 exchanges. The maximum number of trades for an arbitrage is 8, but few transactions have more than 3. One reason for this is likely the computational burden to find more complicated arbitrage opportunities.[24] Another reason is the larger transaction costs, which increase in two ways: A larger network fee due to a more complex Ethereum transaction, and higher exchange fees due to more trades. The volume of the arbitrage transactions, in Table 1.1, is calculated by the volume of the first trade in each arbitrage, not adding the volume of later trades in the transaction. In this way it is possible to think about the opportunity cost and required capital for the arbitrageur. Most arbitrages require capital up to \$6,000, although the maximum is over \$3 million. The distribution of the transaction costs reveal that the mean is substantial relative to the profits and that the standard deviation is high. Further, the minimum transaction cost is rounded to 0, although it is non-zero.

After Uniswap, Sushiswap is the largest decentralized exchange. Approximately 95% of the sampled arbitrages use either or both of these two exchanges (all of them use Uniswap). Further transactions, using CRO Defi Swap, Shibaswap, and Linkswap amounts to 4% of the arbitrageurs' trading. The stablecoins USDT (Tether), USDC, and DAI traded against Ether are the most traded currency pairs on

[24] In order for an arbitrageur to find an opportunity, they must identify it within the average Ethereum block time of 14 seconds. Zhou et al. (2021a) estimate that given a 3-second network delay their algorithm must detect arbitrage within a 10.5-second window. The authors show that their algorithm exceeds the time limit when trying to exploit more than 6 arbitrage cycles.

Table 1.1: Descriptive statistics of arbitrage transactions. N = 231,645. All currency units are in USD.

	Exchanges	Trades	Position	Volume	Cost	Profit	Net profit
mean	1.51	2.90	77.46	6,573.19	71.88	129.31	57.43
median	1.00	3.00	66.00	2,461.29	39.25	55.16	7.72
std	0.53	0.70	71.32	26,037.69	377.53	789.99	663.89
min	1.00	2.00	0.00	0.00	0.00	0.00	-3,936.86
25%	1.00	2.00	8.00	993.04	20.04	27.12	2.13
50%	1.00	3.00	66.00	2,461.29	39.25	55.16	7.72
75%	2.00	3.00	125.00	5,906.07	73.02	109.90	27.33
max	4.00	8.00	790.00	3,311,271.21	49,655.44	167,860.28	167,818.47
sum	358,720.00	689,714.00	18,439,829.00	1,564,878,324.27	17,113,326.74	30,784,514.82	13,671,188.08

Uniswap, and also generally the most liquid trading pairs.

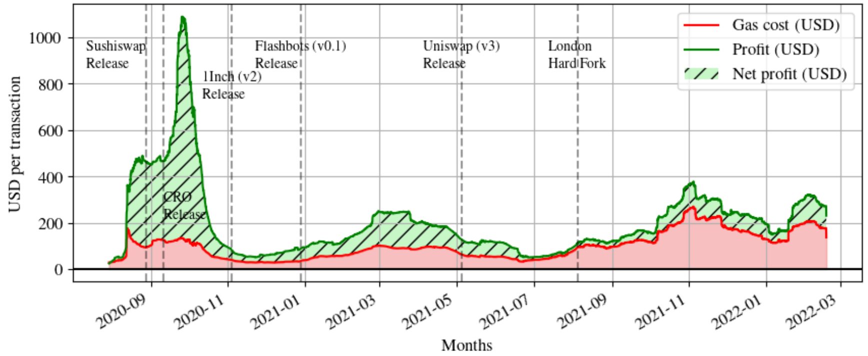


Figure 1.2: 30-day moving average of total gas costs and profits from successful arbitrages.

Unlike other financial markets, where high frequency traders compete primarily with speed, arbitrageurs on decentralized markets also compete with willingness to pay for the arbitrage opportunity, as miners prioritize high-paying transactions. Figure 1.2 shows the 30-day moving average of successful arbitrage transactions' costs and profits. The dashed area, above the cost curve and below the profits curve, pertains to average net profits. The net profits are fairly uniformly distributed over the first half of sample period, with the exception of the beginning when Sushiswap and CRO Swap were released. In the second half of the sample, competition seems to have increased as costs increase and net profits looks slightly lower. Important events such as the introduction of *Flashbots'* private relay and the decentralized exchange routing protocol *1Inch* do not, visually, seem to have any significant effect on arbitrage profits.

Approximating arbitrage transaction costs

Transactions that are sent to miners through the mempool pay transaction costs by a regular gas payment. Arbitrage transactions that use private relays have the option to pay miners directly in addition to the standard gas payment. When estimating arbitrageurs' total costs it is important to consider both alternatives. Regular gas payments are easy to observe as they are logged in the transaction data on the blockchain.[25] However, direct payments to the miner need to be considered separately. Specifically, any transfer to the miner's address needs to be considered. Miners treat direct payments and regular gas payments in the same way, and position transactions that pay the most first in the blocks.

The trading fees that the arbitrageurs pay to the decentralized exchanges do not need to be estimated as they are automatically accounted for in the trade. The total cost for each arbitrage transaction is calculated by adding the regular gas cost and any direct payment. Table 1.1 shows that the average arbitrage cost is $72, which is over 50% of the average profit. This relationship is illustrated in Figure 1.2, where the net profits are shown above the cost curve.

The total cost for successful arbitrages can be precisely calculated. However, there is a hidden cost as arbitrageurs do not always succeed in capturing arbitrages. Failed arbitrage transactions still pay a transaction fee to the miner for the attempt to execute the transaction. Assuming that failed transactions from arbitrageurs are attempted arbitrages, Wang et al. (2021b) find that most arbitrageurs have a success rate of over 90%. To estimate the total cost of the arbitrageurs the average cost should be increased by approximately 11%, and external operational costs should be added, such as hardware and electricity. Nonetheless, such an analysis is beyond the scope of this paper.

Arbitrage transactions' position in blocks

The number of transactions in an Ethereum block varies with the complexity of the transactions and current network demand. The average number of transactions per block in the arbitrage dataset is 204. The distribution of the arbitrage transactions' positions in the

[25] See Appendix 5 for a detailed description.

blocks is visualized in Figure 1.3.[26]

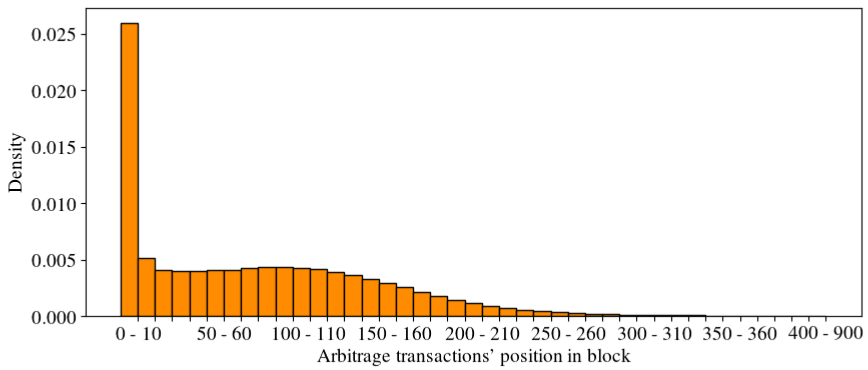


Figure 1.3: Distribution of arbitrage transactions' block position for the full sample, July 2020 to February 2022. Transactions with block position 0 are executed first in each block.

Figure 1.3 gives insight into what information the arbitrageurs use, what previous transaction might have created the arbitrage opportunity, and whether the arbitrageurs might use private relays to send their transactions to the miners. The figure shows that many arbitrage transactions are placed in the beginning of the blocks, with transaction positions 0 to 10. 4.47% of the arbitrage transactions are at block position 0, and thus the first transactions to be executed in the blocks. This indicates that these transactions capitalize on arbitrage opportunities created in previous blocks, as no other transaction is executed before the arbitrage in the current block. Other arbitrage transactions have positions in the middle or end of the blocks, suggesting that these transactions profit from arbitrage opportunities within the same block. However, arbitrageurs that use private relays do not need to wait for an exchange rate changing transaction to be executed and try to back-run it. Instead, arbitrageurs can bundle pending exchange rate changing transactions together with their arbitrage transactions and pay a high transaction fee to the miner to execute these transactions together. Thus, it is possible that privately relayed transactions tend to have a higher block position, such that the arbitrage opportunity is captured early in the block.

[26]As blocks have different number of transactions, Figure 1.9 in Appendix 5 shows the distribution of arbitrage transactions' relative block position.

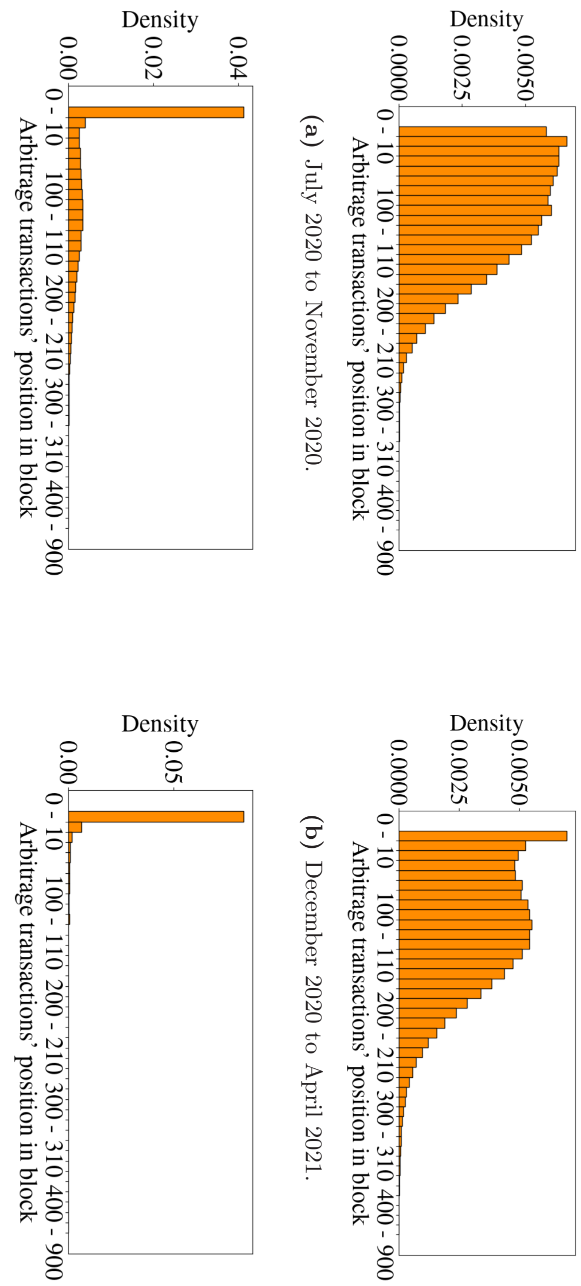
Figures 1.4a to 1.4d show the distributions of the arbitrage transactions positions in the blocks over 5-month subsamples. Interestingly, in the beginning of the sample period arbitrage transactions tend to be positioned in the middle of the blocks, and in the later part of the sample period the transactions tend to be positioned in the beginning of the blocks. One possible explanation for this is the increased amount of privately relayed arbitrage transactions in the end of the sample, which would be consistent with the seemingly increasing gas costs in the end of the sample in Figure 1.2.[27]

Figure 1.5 shows a lower bound of privately relayed arbitrage transactions over the sample period. The percentages are calculated based on the number of arbitrage transactions that use a direct payment to the miner, which is only possible through private relays. This however, creates a lower bound for the number of privately relayed transactions as the arbitrageurs do not have to pay miners directly, but can instead do so through regular gas payments. The first privately relayed arbitrage transactions can be observed around the launch of Flashbots' private relay client *MEV Geth* in December 2020, and from then on the percentage of private arbitrage transactions steadily increases. However, after July 2021, there is a declining pattern of arbitrage transactions that pay the miners directly. It is, however, unclear if the percentage of privately relayed arbitrage transactions also declines during this time period, or whether the arbitrageurs changed their primary payment method.

4 Arbitrage analysis

To investigate the occurrence and subsequent elimination of arbitrage opportunities the empirical methodology is constructed in three parts. (i) A counterfactual simulation is designed, where arbitrage transac-

[27]This observation is consistent with Figures 1.10a through 1.10d in Appendix 5, showing relative block positions over 5-month subsamples.



(a) July 2020 to November 2020.

(b) December 2020 to April 2021.

(c) May 2021 to September 2021.

(d) October 2021 to February 2022.

Figure 1.4: Distribution of arbitrage transactions' block position. Transactions with block position 0 is executed first in each block.

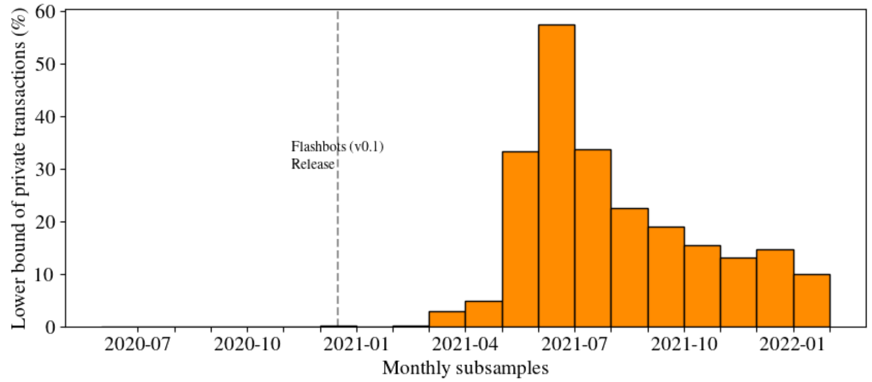


Figure 1.5: Lower bound of privately relayed arbitrage transactions for each month in the sample.

tions are re-executed in different blocks and positions. By simulating another, hypothetical, state of the blockchain it is possible to analyze how the arbitrage transactions would have behaved under different circumstances, and draw conclusions from their dependence on previous trading. (ii) A predictive model is designed to estimate how previous trading predicts arbitrages, using the realized arbitrage transactions and a randomly sampled control group of non-arbitrage transactions. (iii) Arbitrage profits are regressed on previous exchange rate changes to understand how profits differ depending on how far in the past the arbitrage-triggering price changes occurred. However, before turning to the full formal analysis, a snapshot of within-block price differences between Uniswap and Sushiswap is visualized together with some completed arbitrage transactions.

4.1 Snapshot of arbitrages

The Ethereum state is updated with each Ethereum transaction and thus multiple times within each block. Therefore, exchange rates on decentralized exchanges change within each block and prices can be measured on a transaction-level basis. Figure 1.6 shows the difference in the ETH-USDT (ether and Tether stablecoin) exchange rate between Uniswap and Sushiswap. The dashed bars show the end-of-block exchange rate differences between the two exchanges and the solid bars show the maximum differences within the blocks. The snapshot shows approximately 1 hour of trading on the 23 of May 2021. The day was chosen arbitrarily, and the time was chosen to show multiple arbitrage transactions within a short time interval. The maximum within-block price differences include the end-of-block price difference: The bars have the same height if the end-of-block price difference is the largest price difference in that block. If the bars have different heights there has been at least one occasion within that block where a larger price discrepancy has occurred and then disappeared.

At several times during the snapshot, the end-of-block price difference is lower than the within-block price difference. That is, within the block, the price difference has been greater but subsequently corrected before the block ends. In the figure, pure arbitrage transactions are marked with stars and show how within-block price anomalies are regularly arbitraged away. Unsurprisingly, there are also large price differences without any observed classified arbitrage transaction correcting the price. The classification method in this paper focus on pure arbitrage transactions, and prices may adjust due to other types of trades as well.

Table 1.2 shows some statistics for the arbitrage transactions in Figure 1.6.[28] The block positions of the arbitrage transactions reveal that they most likely have back-run pending transactions and have not captured arbitrage from previous blocks. The reason for this is that in order to compete in capturing arbitrage from the previous block

[28] One arbitrageur address, `0x...e379`, executes 5 out of the 9 arbitrages in the snapshot. In fact, this arbitrageur captures approximately 20% of the profits in the full arbitrage dataset. All arbitrageur addresses in Table 1.2, except `0x...20f2`, are present on Etherscan's list of 260 addresses that are heavily involved with miner extractable value such as arbitrage trading (`https://etherscan.io/accounts/label/mev-bot`). The competition for arbitrages is said to be increasing and solo arbitrageurs are getting out-competed by teams, both from traditional finance and the cryptocurrency industry. Rumor has it that around 20 teams are dominating the industry (Worsley, 2022). This claim is in line with the data in this paper, where 840 unique arbitrageur addresses are identified, but 76% of the net profits are captured by 20 arbitrage addresses.

Table 1.2: Descriptive statistics of arbitrage transactions between 2021-05-23 16:30:52 and 2021-05-23 17:14:57. All currency units are in USD. Volume is calcuated as the amount traded in the first leg of the arbitrage.

Time	Position	Arbitrageur	Volume	Cost	Net profit
2021-05-23 16:30:52	6	0x0000000000007f150bd6f54c40a34d7c3d5e9f56	174,988.73	149.51	476.96
2021-05-23 16:38:38	96	0x000000089341e263b85d84a0eea39f47c37a9d2	441,000.00	188.03	8,315.18
2021-05-23 16:46:02	10	0x3700006fbcde59a8b3af2c134d00e9530000e379	318,253.28	417.57	1,275.97
2021-05-23 16:46:55	83	0x3700006fbcde59a8b3af2c134d00e9530000e379	161,427.18	199.12	239.77
2021-05-23 16:57:47	94	0x0000000000007f150bd6f54c40a34d7c3d5e9f56	11,034.15	158.67	32.03
2021-05-23 16:59:59	54	0xf5b4f13bdbe12709bd3ea280ebf4b936e99b20f2	253,778.07	233.99	1,114.21
2021-05-23 17:05:39	14	0x3700006fbcde59a8b3af2c134d00e9530000e379	337,209.53	397.11	1,605.89
2021-05-23 17:11:07	153	0x3700006fbcde59a8b3af2c134d00e9530000e379	143,005.49	180.07	179.44
2021-05-23 17:14:57	148	0x3700006fbcde59a8b3af2c134d00e9530000e379	170,750.44	187.94	327.63

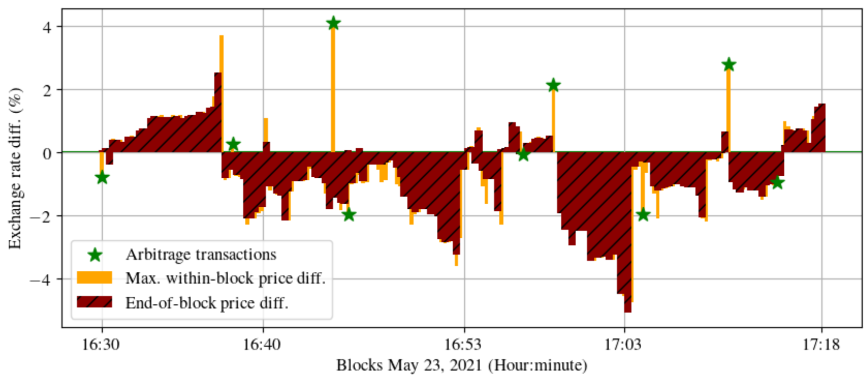


Figure 1.6: End-of-block price differences and maximum within-block price differences for the exchange pair ETH-USDT (ether and Tether stablecoin) on the decentralized exchanges Uniswap and Sushiswap. The data are taken from a snapshot of blocks from the 23 of May 2021.

the arbitrageurs would like to execute their transactions as early as possible in the block. Looking closer at the last arbitrage in Table 1.2, with block position 148, there is a previous transaction, with position 146 in the same block (now shown in Table 1.2), that trades 51 Ether for 100,000 USDT on Sushiswap, off-setting the no-arbitrage exchange rate between the exchanges. The arbitrage transaction capitalizes on this price discrepancy and brings the market back to the no-arbitrage price. This scenario showcases how an arbitrage opportunity arises and how an arbitrageur profits from the price discrepancy.

From Figure 1.6, it is clear that arbitrageurs act fast to capitalize on arbitrage opportunities and that price anomalies are often corrected within a block. This suggests that arbitrage opportunities are highly time sensitive, appearing for brief moments within blocks. Prices measured at the end of blocks, rather than at the transaction level, therefore seems more likely to be arbitrage-free.

4.2 Counterfactual simulation

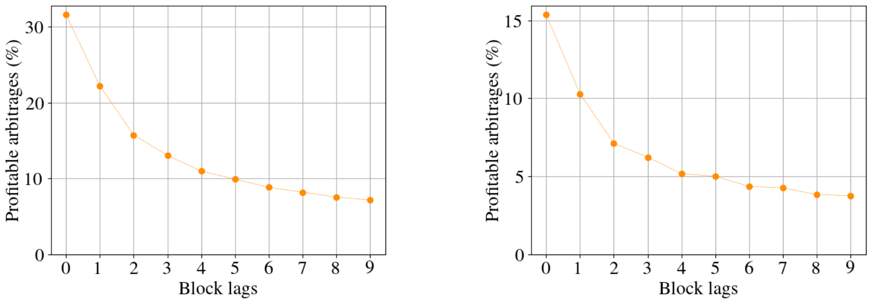
The computer code for each Ethereum transaction is fully transparent as it is necessary for node operators to be able to replay every transaction on the network up until the current state. This feature makes it possible to simulate alternative versions of the blockchain. Transactions are defined by their accompanying transaction code and any detail in the code can be changed, such as transaction cost and block position. This gives a unique opportunity to study counterfactual states of the world, in a way not possible in any traditional market. Transactions can be altered and re-ordered in any way possible and the counterfactual results can be analyzed.[29] By re-executing the arbitrage transactions in a different order in the blockchain, it is possible to analyze if the transactions would have been profitable under different circumstances. For example, if an arbitrage transaction is re-executed where the price anomaly does not exist, the transaction will either be cancelled or show negative profits.

A counterfactual simulation is designed to re-execute each arbitrage transaction as the first transaction in its' own block and as the first transaction in previous blocks, to determine at which point it is no longer profitable. The hypothesis is that if the arbitrage transactions are primarily capitalizing on exchange rate differences created in their own blocks, then the transactions would no longer be profitable if executed as the first transaction in their own blocks, i.e., the arbitrage transactions are placed before the price anomalies occur. Similarly, if an arbitrage transaction profits from exchange rate changes in the previous block, it would no longer be profitable if it were executed as the first transaction in the previous block.

Figures 1.7a and 1.7b shows the percentage of arbitrages that are still profitable when re-executed as the first transaction in their current and previous blocks.[30] Figure 1.7a keeps the original transaction costs of the arbitrage transactions, whereas Figure 1.7b uses the transaction cost the arbitrage transaction would have payed to be

[29] Appendix 5 gives an overview of the data of an Ethereum transaction.

[30] As a robustness check, the simulations are also run excluding the arbitrage transactions at position 0. See Appendix 5.



(a) Net profits with original transaction costs.

(b) Net profits with updated transaction costs.

Figure 1.7: Counter factual simulation with 9 lags.

in the simulated position. By re-executing each arbitrage transaction at the beginning of its own block, only 31% of the transactions are profitable. This suggests that a vast majority (69%) of the arbitrage transactions profit from exchange rate changes within the same block. The further away from its original position that the arbitrage transaction is re-executed, the less likely it is to be profitable. When the transactions are re-executed as the first transaction in the block that was mined 9 blocks from its original position, only 7% are profitable. Put differently, most arbitrage opportunities are eliminated within 9 blocks, and already after 4 blocks only 12% of the transactions are profitable. Since a block is mined approximately every 14 seconds, it takes around 1 minute for most of the arbitrage opportunities to disappear. However, after only an average of 14 seconds (1 block) the majority of the arbitrage profits are made.

The counterfactual simulation shows that arbitrage transactions are very sensitive to the order of executions. Arbitrageurs need to act fast and precise when back-running pending transactions from the mempool. The arbitrageurs have to identify the arbitrage opportunity within the average block window of 14 seconds, and carefully place their transaction after the exchange rate changing trade, either by adjusting the transaction cost to pay marginally less than the pending trade or by submitting the two transactions in a private transaction

bundle.

4.3 Predicting arbitrages by prior trading

On decentralized exchanges the marginal exchange rate between two assets, X and Y, is determined by the fraction of their liquidity pools, $\frac{x}{y}$. Arbitrages are solely created by agents trading against the liquidity pools so that the no-arbitrage exchange rate is off-set. Large exchange rate changes should therefore signal future arbitrage transactions. The counterfactual simulation shows that arbitrage transaction are sensitive to timing, but that some arbitrage opportunities exist as far back as 9 blocks from the original arbitrage. To empirically evaluate how far back exchange rate changes affect arbitrage transactions, I conduct a predictive study in which imbalances in the liquidity pools are used to estimate whether or not an arbitrage transaction is likely to occur. The exercise also investigates the dynamics of arbitrages in the sense that it indicates how long-lived arbitrages are by answering: How far back into the past one has to look for price imbalances to predict current arbitrage.

In the predictive exercise, $n = 231,645$ arbitrage transactions are studied. To discriminate these arbitrage transactions from regular trading, a random control group is constructed from the original dataset. Specifically, the control group consists of n randomly sampled non-arbitrage transactions. These transactions are sampled uniformly without replacement, under the criteria that they perform one trade on Uniswap. The total sample size is thus equal to $2n = 463,290$ transactions. A transaction t_{b_i,p_i} is assigned $A_i = 1$ if it is an arbitrage transaction, and $A_i = 0$ if the transaction belongs to the control group. Transaction t_{b_i,p_i}, $i = 1, \ldots, 2n$, exists in block b_i at position p_i. As the number of positions differ in each block, block b has P_b positions.

As price changes are fully deterministic on automated market makers, there is a direct mapping from trading to changes in price. This makes it possible to calculate the exact price impact of each transaction. The price impact of a transaction $t_{b,p}$ on the exchange rate $\frac{x}{y}$, is defined by the log difference in the liquidity pools x and y,

$$\Delta_{b,p}(x,y) = \log\left(\frac{x_{b,p-1}}{y_{b,p-1}}\right) - \log\left(\frac{x_{b,p}}{y_{b,p}}\right). \tag{1.7}$$

Here $\frac{x_{b,p-1}}{y_{b,p-1}}$ is the exchange rate before transaction $t_{b,p}$ is executed and $\frac{x_{b,p}}{y_{b,p}}$ it the exchange rate after transaction $t_{b,p}$ is executed. To study how a previous transaction $t_{b,p}$ affects t_{b_i,p_i}, the maximum absolute price impact of $t_{b,p}$ related to t_{b_i,p_i} is defined as,

$$\phi_{b,p}(i) = \max_{(x,y) \text{ such that } x,\, y \text{ are traded in } t_{b_i,p_i}} \left|\Delta_{b,p}(x,y)\right|. \tag{1.8}$$

Thus, Equation 1.8 describes the maximum impact on the exchange rates traded in t_{b_i,p_i} by a prior transaction $t_{b,p}$. Although, 43% of $t_{b,p}$ are only trading in one currency pair (see Section 3.1), transaction t_{b_i,p_i} can trade in several exchange rates, often across several exchanges. Here, only the maximum absolute exchange rate change is captured by $\phi_{b,p}(i)$. This is a somewhat conservative measure of how trading in transaction $t_{b,p}$ affects transaction i, and as a robustness check the *sum* is also calculated.

In order to predict if transaction t_{b_i,p_i} is an arbitrage or not, the price impacts from transactions executed prior to t_{b_i,p_i} are calculated. The maximum price impact of the transaction immediately prior to t_{b_i,p_i}, i.e., t_{b_i,p_i-1}, is defined as,

$$PrevTrans_i = \phi_{b_i,p_i-1}(i). \tag{1.9}$$

The relationship between $PrevTrans_i$ and transaction t_{b_i,p_i} indicates to what degree the transaction immediately prior to t_{b_i,p_i} helps to predict if it is an arbitrage. Arbitrageurs can profit from a large $\phi_{b_i,p_i-1}(i)$ by identifying t_{b_i,p_i-1} when it is pending in the mempool, calculate its exact price impact given the current state, and either back-run it through a private transaction bundle or a regular public transaction. If done successfully, this is the absolutely fastest way in

which price anomalies can be arbitraged away, as there are no other transaction between t_{b_i,p_i-1} and t_{b_i,p_i}. In this situation the exchange rate is back at the no-arbitrage price within the next transaction.

Furthermore, to measure the magnitude of the changes in the exchange rates in the same block as transaction t_{b_i,p_i}, the maximum price impact of all prior transactions in the block, excluding transaction t_{b_i,p_i-1}, is defined as,

$$SameBlock_i = \max\{\phi_{b_i,0}(i), \ldots, \phi_{b_i,p_i-2}(i)\}. \tag{1.10}$$

The price impact $SameBlock_i$ describes how the exchange rate changed prior to transaction t_{b_i,p_i} in block b_i. This relationship further describes to what extent prior trading in the same block predicts arbitrage. For arbitrageurs to profit from a large $SameBlock_i$, the arbitrageurs need to identify a pending transaction that will significantly affect the exchange rate, but are not able to place the arbitrage transaction immediately after it. For some reason there is at least one other transaction in between the exchange rate changing transaction and the arbitrage transaction. However, in terms of price efficiency, the exchange rate will still be arbitrage-free within the block.

Although much of the arbitrage action happens within the block, the counterfactual simulation (Section 4.2) shows that some arbitrages live across blocks. By analyzing how far back transactions affect arbitrages, it is possible to get a complete measure of how fast arbitrageurs are able to correct exchange rate deviations. To investigate how far back exchange rate changes affect t_{b_i,p_i}, the maximum change in the exchange rates from transactions in previous blocks $b_i - s$, $s = 1, \ldots, 10$ is measured as,

$$PrevBlock_{i,s} = \max\{\phi_{b_i-s,0}(i), \ldots, \phi_{b_i-s,P_{b_s}}(i)\}. \tag{1.11}$$

The relation between $PrevBlock_{i,s}$ and transaction t_{b_i,p_i}, describes how far back exchange rate changes affect arbitrages. One reason for why arbitrageurs are not always able to back-run arbitrage creating

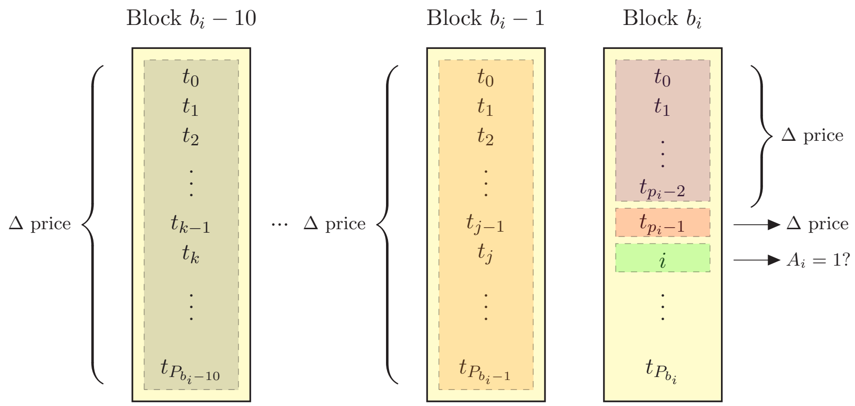


Figure 1.8: For each transaction in the sample data, the maximum price impact of previous transactions is calculated in groups. The notation in the figure is simplified as the blocks are labeled in the headings instead of as subscripts.

transactions is that some blocks are mined faster than average. Although, arbitrageurs have on average 14 seconds to observe pending transactions, calculate their price impacts, and take action, this is not always the case as some blocks are mined as fast as within 1 second.

Figure 1.8 visually illustrates how $PrevTrans_i$, $SameBlock_i$, and $PrevBlock_{i,1}, \ldots, PrevBlock_{i,10}$ are calculated. Tables 1.3 and 1.4 show the distributions of $PrevTrans_i$, $SameBlock_i$, and $PrevBlock_{i,1}, \ldots, PrevBlock_{i,4}$ for the arbitrage transactions, $A_i = 1$, and the non-arbitrage control group, $A_i = 0$, respectively. The transactions just prior to the arbitrage transactions affect the exchange rate the most on average, with a 3% impact. 27% of the $PrevTrans_i$ observations are non-zero for the arbitrages. These statistics are significantly lower for the control group, where $PrevTrans_i$ is 0.22% on average, and barely 5% are non-zero. Similarly, $SameBlock_i$, and $PrevBlock_{i,1}, \ldots, PrevBlock_{i,10}$ are all significantly higher for the arbitrages compared to the control group.

To predict whether a transaction is an arbitrage, i.e., $A_i = 1$, or a regular transaction from the non-arbitrage control group, i.e., $A_i = 0$,

Table 1.3: Descriptive statistics of the maximum price changes by trading prior to the arbitrage transactions. All units are in log differences multiplied by 100 to be interpreted as percentages.

	$PrevTrans_i$	$SameBlock_i$	$PrevBlock_{i,1}$	$PrevBlock_{i,2}$	$PrevBlock_{i,3}$	$PrevBlock_{i,4}$
mean	3.16	0.57	2.11	1.46	0.91	0.66
median	0.00	0.00	0.00	0.00	0.00	0.00
std	56.24	7.90	37.87	37.85	30.82	20.13
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00	0.00	0.00
75%	0.45	0.00	0.49	0.05	0.01	0.01
max	6,914.57	1,927.32	7,525.22	6,936.49	7,828.48	7,201.81
sum	735,546.11	132,226.65	491,477.84	340,514.61	211,374.69	153,207.66
% nonzero	27.01	20.37	48.38	42.26	37.35	35.43

Table 1.4: Descriptive statistics of the maximum price changes by trading prior to the control transactions. All units are in log differences multiplied by 100 to be interpreted as percentages.

	$PrevTrans_i$	$SameBlock_i$	$PrevBlock_{i,1}$	$PrevBlock_{i,2}$	$PrevBlock_{i,3}$	$PrevBlock_{i,4}$
mean	0.22	0.38	0.50	0.48	0.45	0.43
median	0.00	0.00	0.00	0.00	0.00	0.00
std	5.42	8.70	11.84	11.00	10.37	6.25
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00	0.00	0.00
75%	0.00	0.00	0.00	0.00	0.00	0.00
max	1,749.97	4,143.30	3,425.05	2,901.43	4,715.84	2,462.47
sum	68,607.19	116,404.43	151,211.56	146,227.55	138,145.99	131,055.41
% nonzero	4.91	18.57	26.40	25.35	24.86	24.78

I fit a probit regression with predictors $PrevTrans_i$, $SameBlock_i$, and $PrevBlock_{i,1}, \ldots, PrevBlock_{i,10}$ on the full sample and with 5-month indicator variables to investigate any dynamics in the data.

Table 1.5: Probit: Arbitrage trades and non-arbitrage trades regressed on the maximum of previous price changes. The first column shows a pooled regression. Columns 2 through 5 show the results from one estimation using dummy variables for each time period. The interactions of the time dummy variables and the previous price changes are shown in the table. 2N=463,290 for both regressions. The numbers of observations for each time period are: 114,701; 154,408; 133,478; and 47,987.

	(1) All	(2) July 2020 - Nov 2020	(3) Dec 2020 - April 2021	(4) May 2021 - Sep 2021	(5) Oct 2021 - Feb 2022
$PrevTrans_i$	0.0468***	1.623***	0.603***	0.0837***	0.0305***
	(0.00188)	(0.0271)	(0.0163)	(0.00475)	(0.00200)
$SameBlock_i$	0.0526***	0.300***	0.503***	0.113***	-0.0114
	(0.00890)	(0.0374)	(0.0364)	(0.0193)	(0.0103)
$PrevBlock_{i,1}$	0.0489***	0.665***	0.102***	0.0833***	0.0260***
	(0.00263)	(0.0293)	(0.00660)	(0.00622)	(0.00312)
$PrevBlock_{i,2}$	0.0288***	0.0162**	0.225***	0.0800***	0.0171***
	(0.00275)	(0.00494)	(0.0141)	(0.00831)	(0.00351)
$PrevBlock_{i,3}$	0.0222***	0.0133*	0.0223***	0.120***	0.0181***
	(0.00336)	(0.00614)	(0.00605)	(0.0150)	(0.00529)
$PrevBlock_{i,4}$	0.0214***	0.00929	0.175***	0.0161*	0.0744***
	(0.00489)	(0.00657)	(0.0309)	(0.00759)	(0.0181)
$PrevBlock_{i,5}$	0.00751	-0.101**	0.0404	0.00100	0.0232
	(0.00766)	(0.0330)	(0.0323)	(0.00874)	(0.0191)
$PrevBlock_{i,6}$	0.0185*	-0.148***	0.00448	0.0512	0.0592***
	(0.00823)	(0.0335)	(0.00992)	(0.0273)	(0.0174)
$PrevBlock_{i,7}$	-0.00395	-0.0137	0.0238	-0.00271	-0.00411
	(0.00411)	(0.0251)	(0.0291)	(0.0102)	(0.00441)
$PrevBlock_{i,8}$	-0.00940***	-0.0523**	-0.0951***	-0.00783	-0.00582*
	(0.00272)	(0.0161)	(0.0227)	(0.0220)	(0.00268)
$PrevBlock_{i,9}$	-0.00592	-0.223***	0.00288	0.00488	0.000773
	(0.00761)	(0.0394)	(0.0282)	(0.0234)	(0.00843)
$PrevBlock_{i,10}$	0.00332	0.00236	-0.0881**	0.0362	0.0137
	(0.00637)	(0.00697)	(0.0312)	(0.0220)	(0.0192)
N	463290	463290	463290	463290	463290

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results are presented in Table 1.5 and show that exchange rate changes in transactions up to 4 blocks prior to transaction t_{b_i,p_i},

significantly predict if the transaction is an arbitrage.[31] Thus, trading up to 1 minute before a transaction helps to predict if an arbitrage transaction will occur. One explanation for these results are that price imbalances build up. The price could be different between two exchanges without arbitrage being profitable. Arbitrageurs would arbitrage away price imbalances only when it is profitable to do so. However, the probability of a profitable price imbalance increases as trading is pushing the exchange rate in the off-setting direction. These findings are consistent with those of the counterfactual simulation in Section 4.2, in which some arbitrages are profitable over blocks. Approximately 12% (5% with updated transaction costs) of the arbitrage transactions are profitable 4 blocks back, after which there is clear levelling off in the fraction of profitable arbitrages.

4.4 The effect of prior trading on arbitrage profits

Due to the unique features of the Ethereum data, costs and profits can be precisely calculated for the successful arbitrage transactions. As a further step in understanding arbitrage trading on decentralized exchanges, net profits are regressed on $PrevTrans_i$, $SameBlock_i$, and $PrevBlock_{i,1}, \ldots, PrevBlock_{i,10}$.[32] In the regression, the block times of the blocks b_i, b_{i-1}, and b_{i-2} are used as control variables. The reason is that if the block b_{i-1} is mined fast, arbitrageurs might not have enough time to capture the arbitrage opportunity in the same block. If 2 consecutive blocks are mined fast, arbitrageurs might not be able to capture the profits for 2 blocks. The estimation is conducted on the sample of $n = 231,645$ arbitrage transactions, using one pooled estimation and one with 5-month dummy variables.

Table 1.6 presents the results from the regressions. The first column displays the results from the pooled regression. It shows that previous exchange rate changes up to $PrevBlock_{i,1}$ significantly af-

[31] A probit model using the summation of previous price changes without taking the absolute value is used as a robustness check and shows similar results, see Table 1.7 in Section 5.

[32] As a robustness check, a regression is run using the summation of the previous price changes without taking the absolute value, see Table 1.9 in Appendix 5.

Table 1.6: OLS estimation: Arbitrage net profits regressed on the maximum of previous price changes. The first column shows a pooled regression. Columns 2 through 5 show the results from one estimation using dummy variables for each time period. The interactions of the time dummy variables and the previous price changes are shown in the table. N=231,639 for both regressions. The numbers of observations for each time period are: 42,211; 86,830; 81,754; and 22,278.

	(1) All	(2) July 2020 - Nov 2020	(3) Dec 2020 - April 2021	(4) May 2021 - Sep 2021	(5) Oct 2021 - Feb 2022
$PrevTrans_i$	19.94***	254.8***	1058.2***	17.38**	6.162*
	(2.409)	(38.97)	(22.81)	(6.359)	(2.614)
$SameBlock_i$	390.7***	533.1***	589.6***	445.9***	139.9***
	(20.56)	(82.55)	(66.50)	(26.61)	(40.78)
$PrevBlock_{i,1}$	10.01**	504.7***	20.99*	14.53	2.923
	(3.583)	(59.37)	(8.860)	(8.336)	(4.418)
$PrevBlock_{i,2}$	4.265	-1.536	92.63***	15.28	-0.926
	(3.576)	(6.722)	(20.49)	(10.53)	(4.690)
$PrevBlock_{i,3}$	0.475	1.029	3.045	-3.545	-1.318
	(4.504)	(8.392)	(8.095)	(20.64)	(7.496)
$PrevBlock_{i,4}$	5.217	2.468	10.97	4.199	10.08
	(6.758)	(8.972)	(49.82)	(11.14)	(28.75)
$PrevBlock_{i,5}$	105.1***	16.66	753.8***	-24.07	42.79
	(20.29)	(84.37)	(61.53)	(30.60)	(34.24)
$PrevBlock_{i,6}$	13.75	13.37	-35.31	7.801	23.95
	(15.19)	(66.59)	(22.08)	(42.61)	(27.96)
$PrevBlock_{i,7}$	8.867	-25.75	86.93	-7.196	-6.202
	(15.69)	(50.60)	(53.98)	(30.25)	(22.28)
$PrevBlock_{i,8}$	53.28**	47.64	264.0***	22.86	11.81
	(18.21)	(104.1)	(66.12)	(39.03)	(22.41)
$PrevBlock_{i,9}$	10.69	108.6	-80.16	62.00	1.829
	(10.82)	(101.6)	(68.48)	(39.86)	(11.93)
$PrevBlock_{i,10}$	5.085	-0.124	-86.59	-0.280	4.505
	(8.725)	(9.463)	(65.54)	(30.20)	(36.96)
$SameBlock_i$ Block time	0.196	0.126	0.126	0.124	0.125
	(0.108)	(0.107)	(0.107)	(0.107)	(0.107)
$PrevBlock_{i,1}$ Block time	0.234*	0.200	0.201	0.200	0.201
	(0.109)	(0.108)	(0.108)	(0.108)	(0.108)
$PrevBlock_{i,2}$ Block time	0.0544	0.0239	0.0241	0.0244	0.0248
	(0.109)	(0.108)	(0.108)	(0.108)	(0.108)
N	231639	231639	231639	231639	231639

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

fect net arbitrage profits. These results are different from the predictive study, and indicate that arbitrageurs' profits are generated by trading much closer to the arbitrage than the trading that predicts the arbitrage. Furthermore, the results show that changes in the exchange rates from transactions within the arbitrage transaction's block, $PrevTrans_i$ and $SameBlock_i$, strongly affect arbitrage net profits. This holds true across both estimations, but changes somewhat over time.[33] Looking at the later part of the sample, May 2021 to September 2021 and October 2021 to February 2022, only price changes in the current block, $PrevTrans_i$ and $SameBlock_i$, affect net profits. This indicates that the decentralized exchanges have become more efficient over time and that arbitrage competition seems to have increased. In the later part of the sample it is, on average, no longer possible to use information from the last state (block) of the blockchain and previous states (blocks) for profitable arbitrage trading. Arbitrageurs need to observe pending transactions in order to profit. One explanation for these results is the increased usage of private relays in the later part of the sample, which allows arbitrageurs to capture arbitrage opportunities with a higher precision.

The atomicity of the Ethereum transaction ensures that the arbitrage risk is reduced. On the blockchain, compared to traditional markets, arbitrageurs have the advantage to be able to calculate the exact price changes of pending transactions, and are thus able to precisely forecast arbitrage opportunities. Although speed is of importance, arbitrageurs on the blockchain have an average of 14 seconds to do their calculations before the next block is mined. These features are reflected in the empirical results that suggest that in most cases, these 14 seconds are sufficient for the arbitrageurs to act and thus eliminate the arbitrage before the end of the block.

[33] The results are consistent with estimating the same model using only the 10 most traded arbitrage pairs (n=27,697). However, using only the most traded pairs the market seems more efficient, and $PrevTrans_i$ and $SameBlock_i$ have stronger effect on net profits. One plausible explanation for this is that only cross-exchange arbitrage between one currency pair occur in this subsample. These arbitrage opportunities are easier to find, and one would expect price anomalies to be corrected faster.

Furthermore, the results have some implications for the use of exchange rates from decentralized exchanges as reference prices on the blockchain. Since Ethereum is an isolated system, it is unable to receive external data from the "outside" world. Therefore, decentralized exchanges are used for reference pricing, and smart contracts can query market information from the exchanges on-chain. As deviations from the no-arbitrage price are prone to be arbitraged away within the block, the end-of-block prices are likely to be arbitrage-free and suitable as reference prices.

5 Conclusion

In this paper, I show that arbitrageurs contribute to price efficiency on decentralized exchanges by neutralizing price anomalies. This happens very fast, and most of the arbitrage opportunities are created and capitalized on within the Ethereum block. These effects are stonger in the later part of the sample, where only trading in the same block as the arbitrage transaction affects its profits. Arbitrageurs in the later part of the sample have to monitor pending transactions in order to profit. The results speaks to an increased arbitrage competition over time. The speed at which price anomalies are arbitraged away implies that end-of-block prices are likely to be arbitrage-free. This is important as traders place orders based on the price from the previous block, and other on-chain applications use these prices as reference prices.

The results show that arbitrages are created by trading that offsets the no-arbitrage price. A natural question arises: Do most arbitrage opportunities need to occur in the first place? One way around a large price impact is to split an order into multiple smaller orders. The trades could be routed over several exchanges and exchange pairs such that no arbitrage opportunity arises. The trade-off for this kind of order routing would be between the expected price slippage and the increased transaction costs for splitting the trades. This would lead to less arbitrage opportunities occurring and, by design, an even more efficient market.

References

Adams, Hayden, Noah Zinsmeister, and Dan Robinson (2020). *Uniswap v2 core*. Tech. rep. Uniswap (cit. on pp. 12, 23).

Adams, Hayden, Noah Zinsmeister, Moody Salem, River Keefer, and Dan Robinson (2021). *Uniswap v3 core*. Tech. rep. Uniswap (cit. on p. 23).

Alexander, C. and M. Dakos (2020). "A critical investigation of cryptocurrency data and analysis". In: *Quantitative Finance* 20.2, pp. 173–188 (cit. on p. 11).

Angeris, Guillermo, Hsien-Tang Kao, Rei Chiang, Charlie Noyes, and Tarun Chitra (2021). *An analysis of Uniswap markets* (cit. on p. 23).

Back, Adam (2002). *Hashcash - A Denial of Service Counter-Measure*. White paper (cit. on p. 16).

Bank for International Settlements (June 2022). *Miners as intermediaries: extractable value and market manipulation in crypto and DeFi*. Tech. rep. (cit. on p. 21).

Berg, Henry and Todd A Proebsting (2009). "Hanson's automated market maker". In: *The Journal of Prediction Markets* 3.1, pp. 45–59 (cit. on p. 23).

Berg, Jan Arvid, Robin Fritsch, Lioba Heimbach, and Roger Wattenhofer (2022). *An Empirical Study of Market Inefficiencies in Uniswap and SushiSwap* (cit. on p. 12).

Budish, Eric, Peter Cramton, and John Shim (July 2015). " The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response *". In: *The Quarterly Journal of Economics* 130.4, pp. 1547–1621 (cit. on p. 20).

Buterin, Vitalik (2013). *A Next-Generation Smart Contract and Decentralized Application Platform*. White paper (cit. on p. 15).

Cartea, Álvaro, Fayçal Drissi, and Marcello Monga (June 2022). "Decentralised Finance and Automated Market Making: Execution and Speculation". In: *SSRN Electronic Journal* (cit. on p. 25).

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega (2014). "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market". In: *The Journal of Finance* 69.5, pp. 2045–2084 (cit. on p. 12).

Cong, Lin, Xi Li, Ke Tang, and Yang Yang (2019). "Crypto Wash Trading". In: *SSRN Electronic Journal* (cit. on p. 11).

Daian, Philip, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels (2019). *Flash Boys 2.0: Frontrunning, Transaction Reordering, and Consensus Instability in Decentralized Exchanges* (cit. on pp. 10, 20, 21, 27).

DeLong, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann (1990). "Noise Trader Risk in Financial Markets". In: *Journal of Political Economy* 98.4, pp. 703–738 (cit. on p. 28).

Dwork, Cynthia and Moni Naor (1992). "Pricing via Processing or Combatting Junk Mail". In: *Proceedings of the 12th Annual International Cryptology Conference on Advances in Cryptology*, pp. 139–147 (cit. on p. 16).

Erigon Team (Feb. 2022). *Erigon* (cit. on pp. 29, 63).

Eskandari, Shayan, Seyedehmahsa Moosavi, and Jeremy Clark (2020). "SoK: Transparent Dishonesty: Front-Running Attacks on Blockchain". In: *Financial Cryptography and Data Security*, pp. 170–189 (cit. on p. 21).

Hanson, Robin (2003). "Combinatorial information market design". In: *Information Systems Frontiers* 5.1, pp. 107–119 (cit. on p. 23).

Härdle, Wolfgang Karl, Campbell R Harvey, and Raphael C G Reule (Feb. 2020). "Understanding Cryptocurrencies*". In: *Journal of Financial Econometrics* 18.2, pp. 181–208 (cit. on p. 16).

Harvey, Campbell R., Tarek Abou Zeid, Teun Draaisma, Martin Luk, Henry Neville, Andre Rzym, and Otto van Hemert (May 2022). "An Investor's Guide to Crypto". In: *SSRN Electronic Journal* (cit. on pp. 15, 16).

Hasbrouck, Joel (1991). "Measuring the Information Content of Stock Trades". In: *The Journal of Finance* 46.1, pp. 179–207 (cit. on p. 13).

Heimbach, Lioba, Ye Wang, and Roger Wattenhofer (2021). *Behavior of Liquidity Providers in Decentralized Exchanges* (cit. on pp. 24, 30).

Kyle, Albert S. (1985). "Continuous Auctions and Insider Trading". In: *Econometrica* 53.6, pp. 1315–1335 (cit. on p. 13).

Makarov, Igor and Antoinette Schoar (2020). "Trading and arbitrage in cryptocurrency markets". In: *Journal of Financial Economics* 135.2, pp. 293–319 (cit. on p. 11).

— (Apr. 2022). *Cryptocurrencies and Decentralized Finance (DeFi)*. Working Paper 30006. National Bureau of Economic Research (cit. on pp. 16, 23).

Nakamoto, Satoshi (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. White paper (cit. on p. 16).

Qin, Kaihua, Liyi Zhou, and Arthur Gervais (2021). *Quantifying Blockchain Extractable Value: How dark is the forest?* (Cit. on pp. 21, 80).

Savage, Leonard J. (1971). "Elicitation of Personal Probabilities and Expectations". In: *Journal of the American Statistical Association* 66.336, pp. 783–801 (cit. on p. 23).

SEC (Mar. 2019). *Memorandum: Meeting with Bitwise Asset Management, Inc., NYSE Arca, Inc., and Vedder Price P.C.* Tech. rep. (cit. on p. 11).

Shynkevich, Andrei (2021). "Bitcoin arbitrage". In: *Finance Research Letters* 40, p. 101698 (cit. on p. 11).

Szabo, Nick (Sept. 1997). "Formalizing and Securing Relationships on Public Networks". In: *First Monday* 2.9 (cit. on p. 18).

Torres, Christof Ferreira, Ramiro Camino, and Radu State (Aug. 2021). "Frontrunner Jones and the Raiders of the Dark Forest: An Empirical Study of Frontrunning on the Ethereum Blockchain".

In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, pp. 1343–1359 (cit. on p. 21).

TrueBlocks Team (Feb. 2022). *TrueBlocks: Lightweight indexing for any EVM-based blockchain* (cit. on p. 29).

Victor, Friedhelm and Andrea Marie Weintraud (Apr. 2021). “Detecting and Quantifying Wash Trading on Decentralized Cryptocurrency Exchanges”. In: *Proceedings of the Web Conference 2021* (cit. on p. 11).

Vogelsteller, Fabian and Vitalik Buterin (Nov. 2015). *EIP-20: Token Standard*. Ethereum Improvement Proposal (cit. on p. 26).

Wang, Dabao, Siwei Wu, Ziling Lin, Lei Wu, Xingliang Yuan, Yajin Zhou, Haoyu Wang, and Kui Ren (May 2021a). “Towards A First Step to Understand Flash Loan and Its Applications in DeFi Ecosystem”. In: *Proceedings of the Ninth International Workshop on Security in Blockchain and Cloud Computing*. ACM (cit. on p. 79).

Wang, Ye, Yan Chen, Shuiguang Deng, and Roger Wattenhofer (2021b). *Cyclic Arbitrage in Decentralized Exchange Markets* (cit. on pp. 12, 31, 35).

Wood, Gavin (2014). *Ethereum: A Secure Decentralised Generalised Transaction Ledger*. Yellow paper (cit. on pp. 15, 16, 60, 63).

Worsley, Nathan (2022). “MEV as an inner experience”. MEV.Day at Devconnect Amsterdam (cit. on p. 40).

Xu, Jiahua, Krzysztof Paruch, Simon Cousaert, and Yebo Feng (2021). *SoK: Decentralized Exchanges (DEX) with Automated Market Maker (AMM) Protocols* (cit. on p. 25).

Zhang, Yi, Xiaohong Chen, and Daejun Park (2018). *Formal specification of constant product (xy=k) market maker model and implementation* (cit. on p. 23).

Zhou, Liyi, Kaihua Qin, Antoine Cully, Benjamin Livshits, and Arthur Gervais (2021a). “On the Just-In-Time Discovery of Profit-Generating Transactions in DeFi Protocols”. eng. In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 919–936 (cit. on pp. 19, 32).

Zhou, Liyi, Kaihua Qin, Christof Ferreira Torres, Duc V Le, and Arthur Gervais (May 2021b). “High-Frequency Trading on Decentralized On-Chain Exchanges”. In: *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 428–445 (cit. on pp. 19, 23).

Appendices

Ethereum state transition

Formally, the transition to the next global state can be described with the following set of equations (Wood, 2014),

$$\sigma_{t+1} \equiv \Upsilon(\sigma_t, T_i) \tag{1.12}$$
$$\Sigma_{b+1} \equiv \Pi(\Sigma_b, B_{b+1}) \tag{1.13}$$
$$B_{b+1} \equiv (B_H, B_T, B_U) \tag{1.14}$$
$$B_T \equiv (T_0, T_1, \ldots, T_I) \tag{1.15}$$

where Υ is the Ethereum state transition function operating on a transaction-level basis. T_i, $i = 0, \ldots, I$, is a valid transaction and σ_{t+1} is the state at transaction time $t+1$. Π is the block level state transition function, Σ_{b+1} is the global block state and B_{b+1} is a block in block time $b+1$. B_{b+1} contains valid transactions B_T and the block information B_H and B_U, called headers, containing important metadata about the current and previous blocks.[34] The global state is a mapping between addresses and account states and is updated each time a new block is added to the blockchain. Importantly, Equation 1.12 shows that each valid transaction affects the Ethereum Virtual Machine state sequentially, indicating that the blockchain state changes several times within each block.

Descriptives

[34]See Appendix 5 for a full description of the block data.

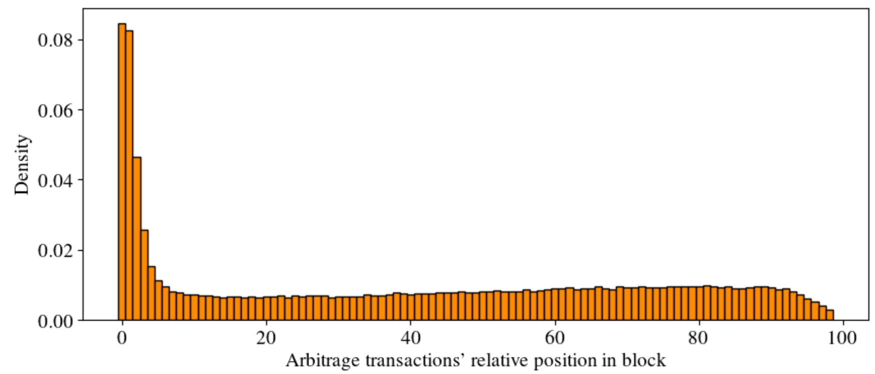
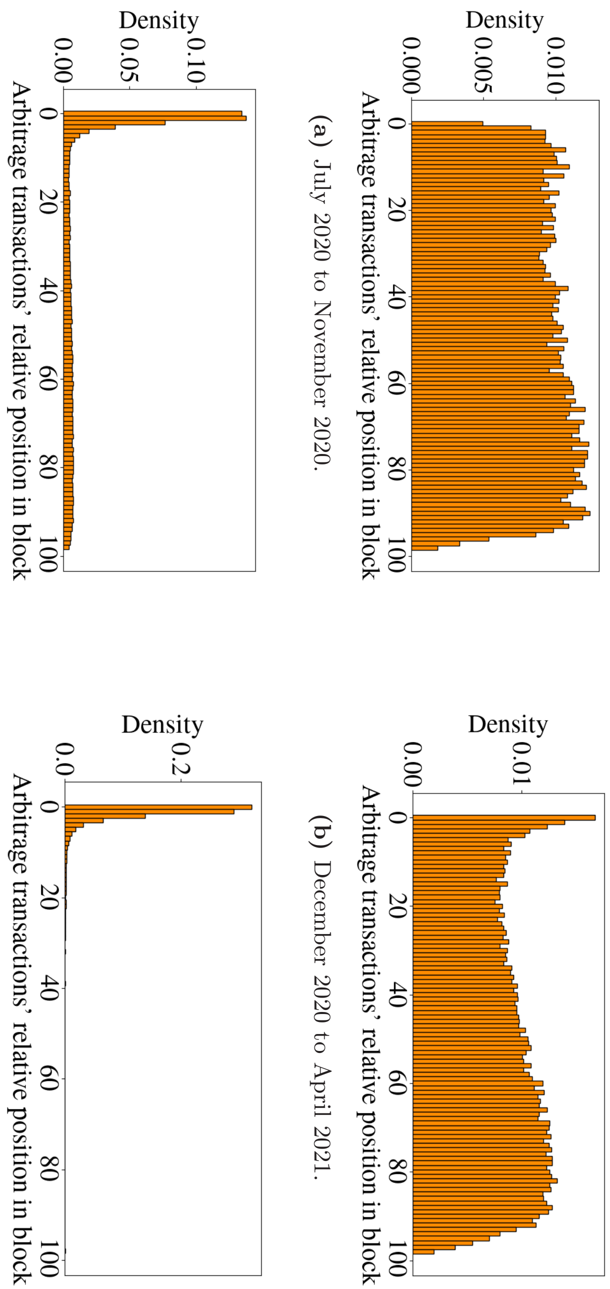


Figure 1.9: Distribution of arbitrage transactions' relative block position for the full sample, July 2020 to February 2022. Transactions with block position 0 is executed first in each block. The relative position is calculated as the percentage.

Robustness analysis

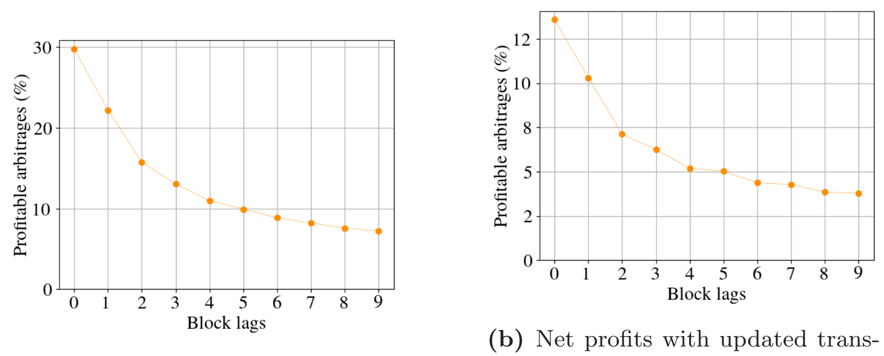
4.47% of the arbitrage transactions are executed at block position 0. These transactions are removed in the simulations presented in Figures 1.11a and 1.11b.



(c) May 2021 to September 2021.

(d) October 2021 to February 2022.

Figure 1.10: Distribution of arbitrage transactions' relative block position. Transactions with block position 0 is executed first in each block. The relative position is calculated as the percentage.



(a) Net profits with original transaction costs, where the arbitrage transactions at position 0 are removed.

(b) Net profits with updated transaction costs, where the arbitrage transaction at position 0 are removed.

Figure 1.11: Counter factual simulation with 9 lags.

Data

This appendix describes the on-chain Ethereum data used in this paper in detail. Ethereum on-chain data is structured into four tables `blocks`, `transactions`, `receipts` and `traces`, and can be accessed through a JSON RPC API. The Ethereum protocol specify a number of necessary fields for blocks, transactions and receipts (Wood, 2014). In addition, trace logs are outputs from the Ethereum Virtual Machine that consist of additional information about the transactions. Depending on what Ethereum client software is used to query the data, the output can vary slightly. The sections 5, 5, 5 and 5 show block, transaction, receipt and trace call responses from an Erigon (Erigon Team, 2022) archive node.[35][36] The output descriptions have been compiled from the official Ethereum documentation, OpenEthereum's documentation, and Wood (2014). Unnecessary verbose outputs are, at places, replaced with ....

[35]The archive node is running on a Debian 11.2 machine with AMD Ryzen 5 1600 (6-core, 3.2GHz), 64GB of RAM and 4TB SSD.

[36]Example transaction hash: 0x0e5e386a2e3a80f1843f6520ebe2f0f118fd1939b36d8a3c00e2e90d2c88df8e.

Table 1.7: Probit: Arbitrage trades and non-arbitrage trades regressed on the sum of previous price changes. The first column shows a pooled regression. Columns 2 through 5 show the results from one estimation using dummy variables for each time period. The interactions of the time dummy variables and the previous price changes are shown in the table. 2N=463,290 for both regressions. The numbers of observations for each time period are: 114,701; 154,408; 133,478; and 47,987.

	(1) All	(2) July 2020 - Nov 2020	(3) Dec 2020 - April 2021	(4) May 2021 - Sep 2021	(5) Oct 2021 - Feb 2022
$PrevTrans_i$	0.0470***	1.624***	0.598***	0.0838***	0.0307***
	(0.00188)	(0.0271)	(0.0162)	(0.00475)	(0.00199)
$SameBlock_i$	0.000283	0.160***	0.506***	0.193***	-0.0433***
	(0.0100)	(0.0373)	(0.0402)	(0.0370)	(0.0108)
$PrevBlock_{i,1}$	0.0699***	0.640***	0.353***	0.0865***	0.0322***
	(0.00352)	(0.0301)	(0.0129)	(0.00702)	(0.00412)
$PrevBlock_{i,2}$	0.0259***	0.0158**	0.195***	0.0794***	0.0153***
	(0.00288)	(0.00532)	(0.0143)	(0.0100)	(0.00353)
$PrevBlock_{i,3}$	0.0250***	0.0128*	0.0544***	0.166***	0.0202***
	(0.00381)	(0.00615)	(0.0111)	(0.0211)	(0.00524)
$PrevBlock_{i,4}$	0.0115**	0.0106	0.109***	0.00909	0.0153
	(0.00436)	(0.00659)	(0.0299)	(0.00588)	(0.0204)
$PrevBlock_{i,5}$	-0.0124	-0.0534	-0.00303	-0.00318	-0.0739**
	(0.00800)	(0.0375)	(0.0280)	(0.00896)	(0.0232)
$PrevBlock_{i,6}$	-0.00993	-0.168***	-0.00335	0.00995	-0.0307
	(0.0105)	(0.0333)	(0.0133)	(0.0250)	(0.0218)
$PrevBlock_{i,7}$	-0.00938*	-0.0132	0.0303	-0.00992	-0.00654
	(0.00429)	(0.0238)	(0.0318)	(0.0115)	(0.00458)
$PrevBlock_{i,8}$	-0.0112***	-0.121***	-0.0381	-0.0402	-0.00691*
	(0.00276)	(0.0251)	(0.0309)	(0.0247)	(0.00270)
$PrevBlock_{i,9}$	-0.0722***	-0.282***	0.0609	-0.0430	-0.0543***
	(0.0131)	(0.0402)	(0.0338)	(0.0298)	(0.0164)
$PrevBlock_{i,10}$	-0.00510	0.00150	-0.0740*	0.00816	-0.0235
	(0.00786)	(0.00850)	(0.0334)	(0.0360)	(0.0231)
N	463290	463290	463290	463290	463290

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.8: OLS estimation: Arbitrage net profits regressed on the maximum of previous price changes without control variables. The first column shows a pooled regression. Columns 2 through 5 show the results from one estimation using dummy variables for each time period. The interactions of the time dummy variables and the previous price changes are shown in the table. N=231,639 for both regressions. The numbers of observations for each time period are: 42,211; 86,830; 81,754; and 22,278.

	(1) All	(2) July 2020 - Nov 2020	(3) Dec 2020 - April 2021	(4) May 2021 - Sep 2021	(5) Oct 2021 - Feb 2022
$PrevTrans_i$	19.94***	255.8***	1058.4***	17.46**	6.157*
	(2.409)	(38.97)	(22.81)	(6.359)	(2.614)
$SameBlock_i$	391.0***	534.0***	590.4***	446.1***	139.9***
	(20.56)	(82.55)	(66.50)	(26.61)	(40.78)
$PrevBlock_{i,1}$	10.12**	506.7***	21.05*	14.68	2.982
	(3.583)	(59.35)	(8.860)	(8.335)	(4.418)
$PrevBlock_{i,2}$	4.323	-1.492	92.85***	15.42	-0.881
	(3.576)	(6.722)	(20.49)	(10.53)	(4.690)
$PrevBlock_{i,3}$	0.475	1.024	3.051	-3.426	-1.347
	(4.503)	(8.391)	(8.095)	(20.64)	(7.496)
$PrevBlock_{i,4}$	5.257	2.502	11.35	4.232	9.946
	(6.758)	(8.972)	(49.82)	(11.14)	(28.75)
$PrevBlock_{i,5}$	105.0***	16.69	753.9***	-24.27	42.63
	(20.29)	(84.37)	(61.53)	(30.60)	(34.24)
$PrevBlock_{i,6}$	13.78	12.87	-35.31	7.896	23.93
	(15.19)	(66.59)	(22.08)	(42.61)	(27.96)
$PrevBlock_{i,7}$	9.073	-25.60	86.77	-7.050	-6.054
	(15.69)	(50.60)	(53.98)	(30.25)	(22.28)
$PrevBlock_{i,8}$	53.37**	47.92	264.1***	23.02	11.77
	(18.21)	(104.1)	(66.12)	(39.02)	(22.41)
$PrevBlock_{i,9}$	10.70	108.3	-79.80	62.10	1.846
	(10.82)	(101.6)	(68.47)	(39.86)	(11.93)
$PrevBlock_{i,10}$	5.112	-0.0813	-86.35	-0.440	4.468
	(8.725)	(9.463)	(65.54)	(30.20)	(36.96)
N	231639	231639	231639	231639	231639

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.9: OLS estimation: Arbitrage net profits regressed on the sum of previous price changes with control variables. The first column shows a pooled regression. Columns 2 through 5 show the results from one estimation using dummy variables for each time period. The interactions of the time dummy variables and the previous price changes are shown in the table. N=231,639 for both regressions. The numbers of observations for each time period are: 42,211; 86,830; 81,754; and 22,278.

	(1) All	(2) July 2020 - Nov 2020	(3) Dec 2020 - April 2021	(4) May 2021 - Sep 2021	(5) Oct 2021 - Feb 2022
$PrevTrans_i$	19.95***	258.3***	1057.6***	17.38**	6.197*
	(2.407)	(38.91)	(22.73)	(6.362)	(2.610)
$SameBlock_i$	811.8***	756.3***	741.6***	1610.8***	171.8**
	(32.37)	(93.89)	(75.36)	(56.73)	(52.67)
$PrevBlock_{i,1}$	13.61**	559.0***	27.50	19.01*	4.480
	(4.727)	(62.61)	(17.58)	(9.407)	(5.730)
$PrevBlock_{i,2}$	10.96**	-0.312	258.2***	20.99	-0.764
	(3.909)	(7.234)	(20.81)	(12.49)	(5.102)
$PrevBlock_{i,3}$	-4.082	1.118	3.815	-25.78	-3.060
	(5.134)	(8.389)	(14.94)	(29.70)	(7.481)
$PrevBlock_{i,4}$	1.575	2.196	53.09	1.412	-3.399
	(5.948)	(8.953)	(46.22)	(8.225)	(34.40)
$PrevBlock_{i,5}$	129.0***	37.91	413.4***	-50.76	64.74
	(25.03)	(81.84)	(47.96)	(37.26)	(64.78)
$PrevBlock_{i,6}$	21.10	38.80	-43.79	-53.03	72.51
	(25.00)	(73.71)	(72.12)	(39.37)	(46.44)
$PrevBlock_{i,7}$	1.598	-11.90	24.12	-10.38	-46.28
	(21.59)	(39.60)	(50.79)	(54.26)	(42.29)
$PrevBlock_{i,8}$	24.43	8.829	156.0*	-82.63	12.15
	(19.80)	(118.6)	(64.61)	(42.95)	(24.39)
$PrevBlock_{i,9}$	46.49*	107.7	31.48	47.67	5.783
	(22.37)	(105.9)	(56.21)	(57.82)	(28.93)
$PrevBlock_{i,10}$	5.383	-0.345	-35.73	-8.733	7.001
	(10.64)	(11.56)	(60.69)	(51.37)	(35.65)
$SameBlock_i$ Block time	0.187	0.118	0.118	0.117	0.118
	(0.108)	(0.107)	(0.107)	(0.107)	(0.107)
$PrevBlock_{i,1}$ Block time	0.229*	0.192	0.194	0.192	0.193
	(0.109)	(0.108)	(0.108)	(0.108)	(0.108)
$PrevBlock_{i,2}$ Block time	0.0558	0.0276	0.0277	0.0284	0.0285
	(0.109)	(0.108)	(0.108)	(0.108)	(0.108)
N	231639	231639	231639	231639	231639

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.10: OLS estimation: Arbitrage net profits regressed on the sum of previous price changes without control variables. The first column shows a pooled regression. Columns 2 through 5 show the results from one estimation using dummy variables for each time period. The interactions of the time dummy variables and the previous price changes are shown in the table. N=231,639 for both regressions. The numbers of observations for each time period are: 42,211; 86,830; 81,754; and 22,278.

	(1) All	(2) July 2020 - Nov 2020	(3) Dec 2020 - April 2021	(4) May 2021 - Sep 2021	(5) Oct 2021 - Feb 2022
$PrevTrans_i$	19.95***	259.2***	1057.9***	17.45**	6.191*
	(2.407)	(38.90)	(22.73)	(6.362)	(2.610)
$SameBlock_i$	812.5***	757.5***	742.5***	1611.1***	172.0**
	(32.37)	(93.88)	(75.36)	(56.73)	(52.67)
$PrevBlock_{i,1}$	13.76**	561.0***	27.73	19.14*	4.558
	(4.727)	(62.60)	(17.58)	(9.407)	(5.729)
$PrevBlock_{i,2}$	11.01**	-0.258	258.5***	21.09	-0.725
	(3.909)	(7.234)	(20.81)	(12.49)	(5.102)
$PrevBlock_{i,3}$	-4.089	1.116	3.868	-25.85	-3.081
	(5.134)	(8.389)	(14.94)	(29.70)	(7.481)
$PrevBlock_{i,4}$	1.613	2.224	53.43	1.438	-3.434
	(5.948)	(8.953)	(46.22)	(8.225)	(34.40)
$PrevBlock_{i,5}$	128.7***	38.21	413.3***	-51.11	64.42
	(25.03)	(81.84)	(47.96)	(37.25)	(64.78)
$PrevBlock_{i,6}$	21.23	38.19	-43.68	-52.92	72.61
	(25.00)	(73.71)	(72.12)	(39.37)	(46.44)
$PrevBlock_{i,7}$	1.839	-11.78	23.95	-10.22	-46.08
	(21.59)	(39.60)	(50.79)	(54.26)	(42.29)
$PrevBlock_{i,8}$	24.54	9.306	156.5*	-82.48	12.13
	(19.80)	(118.6)	(64.61)	(42.95)	(24.39)
$PrevBlock_{i,9}$	46.60*	107.5	31.57	47.85	5.843
	(22.37)	(105.9)	(56.21)	(57.82)	(28.93)
$PrevBlock_{i,10}$	5.439	-0.299	-35.32	-8.716	6.906
	(10.64)	(11.56)	(60.69)	(51.37)	(35.65)
N	231639	231639	231639	231639	231639

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Block data

Listing 1.1 shows the response from the archive node when calling the function `eth_getBlockByHash`. Here I provide a description of the output data,

- `baseFeePerGas`: A scalar value equal to the minimum fee per gas required to be included in the block.[37]
- `difficulty`: A scalar value corresponding to the difficulty level of this block. This can be calculated from the previous block's difficulty level and the timestamp.
- `extraData`: An arbitrary byte array containing data relevant to this block.
- `gasLimit`: A scalar value equal to the current limit of gas expenditure per block.
- `gasUsed`: A scalar value equal to the total gas used in transactions in this block.
- `hash`: The Keccak 256-bit hash of this block's header.
- `logsBloom`: The Bloom filter composed from indexable information (logger address and log topics) contained in each log entry from the receipt of each transaction in the transaction list.
- `miner`: The 160-bit address to which the fees collected from the successful mining of this block be transferred.
- `mixHash`: A 256-bit hash which proves combined with the nonce that a sufficient amount of computation has been carried out on this block.

[37]This is only present in `type 2` transactions after the implementation of EIP-1559 in the London Hard Fork 2021-08-05.

- `nonce`: A 64-bit hash which proves combined with the mix-hash that a sufficient amount of computation has been carried out on this block.
- `number`: A scalar value equal to the number of ancestor blocks. The genesis block has a number of zero.
- `parentHash`: The Keccak 256-bit hash of the parent block's header.
- `receiptsRoot`: The Keccak 256-bit hash of the root node of the Merkle Patricia tree structure populated with the receipts of each transaction in the transaction list portion of the block.
- `sha3Uncles`: The Keccak 256-bit hash of the uncles list portion of this block.
- `size`: A scalar value equal to the size of the block.
- `stateRoot`: The Keccak 256-bit hash of the root node of the state Merkle Patricia tree, after all transactions are executed and finalized.
- `timestamp`: A scalar value equal to the reasonable output of Unix's time() at this block's inception.
- `totalDifficulty`: A scalar value equal to the total difficulty of the chain until this block.
- `transactions`: A list of Keccak 256-bit hashes of the transactions included in this block.
- `transactionsRoot`: The Keccak 256-bit hash of the root node of the Merkle Patricia tree structure populated with each transaction in the transaction list portion of this block.
- `uncles`: A list of Keccak 256-bit hashes of the uncle blocks.

```
{
  "difficulty": "0xc4bbf8674df01",
  "extraData": "0x30783438363937363635366636653230353036
      636663663",
  "gasLimit": "0xbe150c",
  "gasUsed": "0xbbd420",
  "hash": "0xd85f9b3690a8aca172d096a408024c12da45eb4621e0
      8982eaf886f1d12f5d49",
  "logsBloom": "0xdfe041d47520195087193 3f0a87d5da05a28b29
      80014c3ec829dd10a7aa24a1454803c5d660542a6c2213366390
      6cdd546d934080aa698f2ab981a70db2a4dad11131500c7ce630
      3f82c04bd18214ce15ad2095f23480d5458cdd4ea9175d101761
      408849a0ec5b88031830c02268e3dcfe414221e648dc6032d5c9
      2f6a8fc627e04b31787792426df52f560a8a38e0bc003d4816
      ffd9cfbf911f5ef065dc8d7831e1640707c61da0df797ac0528b
      183d3a100018ac06a61a1170c009c2ad28140d8e86ae1b406303
      e846a688f6d85dc04088ec1c0fa443009327343a606b00da0983
      59ca22185403525678cd5911a9d66758715b0da2954193d2707
      ba360a84",
  "miner": "0x1ad91ee08f21be3de0ba2ba6918e714da6b45836",
  "mixHash": "0xf58be2dacfb26108447da3d7809e44829fe35d9ac
      0bdba9115d3a41364bfa29c",
  "nonce": "0x21f0257c209b32dd",
  "number": "0xacee03",
  "parentHash": "0xe424fb2b560b5c7d405dacf2b92cee2dfc8972
      6b365e10b034d776d7b1a16365",
  "receiptsRoot": "0xcf2a24c67500957ccf0faeff4dc0d3b26806
      2d898e084d1f062867442cb887e5",
  "sha3Uncles": "0x1dcc4de8dec75d7aab85b567b6ccd41ad31245
      1b948a7413f0a142fd40d49347",
  "size": "0xa839",
  "stateRoot": "0x13649aca35de0a40e1b1f53c50eab6b27660138
      77c8474a7bfd8b1b81350f53b",
  "timestamp": "0x5fbf71b6",
  "totalDifficulty": "0x4045661d0d677d859d1",
  "transactions": [
    "0xa988d729ecb71c6402fbb893cb696e35f32b9a257eba0fc4be
        77adad443832bd",
    "0x0af408473617105f24ed80117a267315eeadc65048fce857cf
        52529419629e3d",
    "0x578d168a72a9a054f89155a5c38d64401a517f5fb46a64fc77
        ee796873205541",
    ...
    "0xc4713088a14c4be8954d03083bae9a28280ef55b4001005b72
        df0eaf22ffc87c",
    "0x4eb4d31fdf54adab612ed64c0ab836bf0b9ab2f6d87a7e3b0f
        736e125f947133",
    "0x4c2531e6dccb65e3d64f9f41ebb1dd4f86bf9b6aceadeecef6
        69df417899ada2"
  ],
  "transactionsRoot": "0xdbc7459cf1eb23471bdedf2cb02a9140
      d4e2e1956b4366316b1b83b82aeb4a8c",
  "uncles": []
}
```


Listing 1.1: Erigon archive node block response.

Transaction data

Listing 1.2 shows the response from the archive node when calling the function `eth_getTransactionByHash`. Here I provide a description of the output data,

- `blockHash`: The Keccak 256-bit hash of the block's header which this transaction is included in.
- `blockNumber`: A scalar value equal to the block's number which this transaction is included in.
- `from`: The 160-bit address of the sender.
- `gas`: A scalar value equal to the maximum amount of gas units that can be consumed by the transaction.
- `gasPrice`: A scalar value specifying the gas price provided by the sender in wei.
- `maxPriorityFeePerGas`: A scalar value equal to the maximum amount of gas to be included as a tip to the miner.[38]

[38] See footnote 37.

- `maxFeePerGas`: A scalar value equal to the maximum amount of gas to be paid.[39]

- `hash`: The Keccak 256-bit hash of this transaction.

- `input`: An unlimited size byte array specifying the EVM-code for the account initialisation procedure.

- `nonce`: A scalar value equal to the number of transactions sent by the sender.

- `to`: The 160-bit address of the message call's recipient.

- `transactionIndex`: A scalar value equal to this transactions' position in the block.

- `value`: A scalar value equal to the number of Wei to be transferred to the message call's recipient or, in the case of a contract creation, as an endowment to the newly created account.

- `type`: A scalar value indicating transaction type (0 for legacy transaction and 2 for transaction type after EIP-1559).

- `accessList`: Optional list of addresses and storage keys that the transaction plans to access.

- `chainId`: A scalar value indicating which chain this transaction is on (1 for Ethereum Mainnet).

- `v`, `r` and `s`: Values corresponding to the signature of the transaction and used to determine the sender of the transaction.


```
{
  "blockHash": "0xd85f9b3690a8aca172d096a408024c12da45eb4
     621e08982eaf886f1d12f5d49",
  "blockNumber": "0xacee03",
  "from": "0x000000007cb2bd00ae5eb839930bb7847ae5b039",
  "gas": "0x4c959",
```


[39] See footnote 37.


```
  "gasPrice": "0x1ddc4aadade",
  "hash": "0x0e5e386a2e3a80f1843f6520ebe2f0f118fd1939b36d
     8a3c00e2e90d2c88df8e",
  "input":  "0x0000000000000000000000000000000000000000
     0000000000000000002800000000000000000000000000000000
     000000000000000000000000000d4c2000000000000000000
     0000000000000000b3f879cb30fe243b4dfee438691c...882f0
     000000000000000000000000000000000000000001662e93021
     bfb0ca856e1000000000000000000000000000000000000000
     00000c5beefbbfa5688",
  "nonce": "0x37b8",
  "to": "0x0000000000080c886232e9b7ebbfb942b5987aa",
  "transactionIndex": "0xd",
  "value": "0x0",
  "type": "0x0",
  "v": "0x26",
  "r": "0x83de603b9714fbd2b5446b9061bedd5bf8ba4868567595d
     82022378ae700054f",
  "s": "0x708e255f2a98120084267dbce82c739c6c1275a03c10f4c
     6cd1d7e4c5d361730"
}
```


Listing 1.2: Erigon archive node transaction response.

Receipt data

Listing 1.3 shows the response from the archive node when calling the function `eth_getTransactionReceipt`. Here I provide a description of the output data,

- `blockHash`: The Keccak 256-bit hash of the block's header which this transaction is included in.

- `blockNumber`: A scalar value equal to the block's number which this transaction is included in.

- `contractAddress`: The Keccak 256-bit hash of the address if a contract was created, otherwise `null`.

- `cumulativeGasUsed`: A scalar value equal to the total amount of gas used when this transaction was executed in the block.

- `effectiveGasPrice`: A scalar value equal to the gas price used by the transaction.
- `from`: The 160-bit address of the sender.
- `gasUsed`: A scalar value equal to the gas used by the transaction.
- `logs`: A list of log objects.
  - `address`: The 160-bit address to the contract emitting the event.
  - `topics`: An array of Keccak 256-bit hashes of contract functions including arguments.
  - `data`: Byte array specifying the arguments for the contract function called.
  - `blockNumber`: A scalar value equal to the block's number which this transaction is included in.
  - `transactionIndex`: A scalar value equal to this transactions' position in the block.
  - `blockHash`: The Keccak 256-bit hash of the block's header which this transaction is included in.
  - `logIndex`: A scalar value equal to this logs' position in logs.
  - `removed`: Boolean indicating if the log was removed in a reorg.
- `logsBloom`: The Bloom filter composed from indexing information.
- `status`: A scalar value indicating if the transaction was successfully mined.
- `to`: The 160-bit address of the message call's recipient.
- `transactionHash`: The Keccak 256-bit hash of this transaction.

- `transactionIndex`: A scalar value equal to this transactions' position in the block.
- `type`: A scalar value indicating transaction type (0 for legacy transaction and 2 for transaction type after EIP-1559).


```
{
  "blockHash": "0xd85f9b3690a8aca172d096a408024c12da45eb4
      621e08982eaf886f1d12f5d49",
  "blockNumber": "0xacee03",
  "contractAddress": null,
  "cumulativeGasUsed": "0x172b16",
  "effectiveGasPrice": "0x1ddc4aadade",
  "from": "0x000000007cb2bd00ae5eb839930bb7847ae5b039",
  "gasUsed": "0x197a4",
  "logs": [
    {
      "address": "0xc02aaa39b223fe8d0a0e5c4f27ead9083c756
          cc2",
      "topics": [
        "0xddf252ad1be2c89b69c2b068fc378daa952ba7f163c4a1
            1628f55a4df523b3ef",
        "0x000000000000000000000000a478c2975ab1ea89e81968
            11f51a7b7ade33eb11",
        "0x000000000000000000000000fbc312fa3b5be4e7631db2
            901ae7e0e79a764c9b"
      ],
      "data": "0x0000000000000000000000000000000000000000
          000000d68fba3c3c11e53108",
      "blockNumber": "0xacee03",
      "transactionHash": "0x0e5e386a2e3a80f1843f6520ebe2f
          0f118fd1939b36d8a3c00e2e90d2c88df8e",
      "transactionIndex": "0xd",
      "blockHash": "0xd85f9b3690a8aca172d096a408024c12da4
          5eb4621e08982eaf886f1d12f5d49",
      "logIndex": "0x39",
      "removed": false
    },
    ...
    }
  ],
  "logsBloom": "0x002000004020000000000000108001000000000000
      0000000000000000000000000000000000000000000000002000
```

```
    000000200000008000000000000000000000000000000200000
    1000000080000002000000000000000000000000000000000100
    0000000008000000000000000000000000000000000000000001
    0000000000000000000000020000000000000000000000000001
    0000800000041000000000000000000000000000000000000000
    0000000000000000000020000000000000000004200040000000
    0000000400000000002000000001000000000000000040000200
    0000000000000000400000000000000000008000000000000000
    00000",
  "status": "0x1",
  "to": "0x00000000000080c886232e9b7ebbfb942b5987aa",
  "transactionHash": "0x0e5e386a2e3a80f1843f6520ebe2f0f11
    8fd1939b36d8a3c00e2e90d2c88df8e",
  "transactionIndex": "0xd",
  "type": "0x0"
}
```

Listing 1.3: Erigon archive node receipt response.

Trace call data

Listing 1.4 shows the response from the archive node when calling the function `trace_replayTransaction`. Here I provide a description of the output data,

- `output`: String.
- `stateDiff`: Array.
- `trace`:
  - `action`:
    - `from`: The 160-bit address of the trace initiator.
    - `callType`: String of the type of call.
    - `gas`: Maximum amount of gas allowed for this trace.
    - `input`: Byte array of data specifying the EVM-code for the action.
    - `to`: The 160-bit address to the trace recipient.
    - `value`: A scalar value equal to the number of Wei to be transferred to the recipient.
  - `result`:
    - `gasUsed`: Gas used by the trace.
    - `output`: Byte array of the result of the call for this trace.
  - `subtraces`: The number of children traces.
  - `traceAddress`: Array of a particular trace address in the trace tree.
  - `type`: String of the type of trace.
- `vmTrace`: VmTrace object.

```
{
  "output": "0x000000000000000000000000000000000000000000
      0000000000000000000004",
  "stateDiff": null,
  "trace": [
      ...,
      {
      "action": {
        "from": "0x00000000000080c886232e9b7ebbfb942b5987
            aa",
        "callType": "call",
        "gas": "0x35b60",
        "input": "0x022c0d9f00000000000000000000000000000
            0000000000000000000000000000000000000000000000
            00000000000000000000000000000000000d68fba3c3c1
            1e531080000000000000000000000000fbc312fa3b5be4e
            7631db2901ae7e0e79a764c9b000000000000000000000
            0000000000000000000000000000000000000000080000
            0000000000000000000000000000000000000000000000
            0000000000000800000000000000000000000000c02aaa3
            9b223fe8d0a0e5c4f27ead9083c756cc20000000000000
            00000000000000000000000000000000010d0ca8041bb5
            d3108000000000000000000000000c3d03e4f041fd4cd3
            88c549ee2a29a9e5075882f00000000000000000000000
            000000000000000000001662e93021bfb0ca856e1",
```

```
        "to": "0xa478c2975ab1ea89e8196811f51a7b7ade33eb11",
        "value": "0x0"
      },
      "result": {
        "gasUsed": "0x209dc",
        "output": "0x"
      },
      "subtraces": 4,
      "traceAddress": [
        2
      ],
      "type": "call"
    }
    ...,
  ],
  "vmTrace": null
}
```


Listing 1.4: Erigon archive node trace response.

Empirical classification

The arbitrage classification consists of three parts: Detecting arbitrage transactions, detecting sandwich arbitrage bundles, and detecting transactions using flash swaps. This appendix describes the empirical classification strategy with more technical details.

Arbitrage detection

To detect arbitrage transactions the following process is used,

1. Necessary swap actions:
   (a) At least two `Swap` events are emitted.
   (b) All `Swap` events must form a loop, the input asset and amount of any swap action must be the output asset and amount of the previous action.
   (c) The input asset of the first swap action and the output asset of the last swap action must be the same, closing the loop.
2. Atomic transaction: All swap actions must be included in a single transaction.
3. Pure arbitrage:
   (a) The transaction receipt log should only contain `Transfer`, `Sync` and `Swap` events, ensuring that nothing other than DEX trading takes place in the transaction.
   (b) The transaction needs to be profitable.
   (c) The transaction needs to pay a non-zero fee to the miner.
   (d) Flash swap transactions were classified and then removed using the following conditions (Wang et al., 2021a):
      i. The length of the parameter `data` in the transaction's `trace` is greater than zero.

ii. The internal transaction triggered by `uniswapV2Call` must include the invocation of `transfer` or `transferFrom` function.

iii. The receiver address of `transfer` or `transferFrom` function must be the pair contract.

(e) Sandwich bundles were classified and then removed using the following criterion (Qin, Zhou, and Gervais, 2021):

i. The transactions must be executed by the same address.

ii. The transactions must be in the same block and their transaction positions must be within one step from each other.

iii. The transactions' `swap` events must include the same tokens and trade in opposite directions.

iv. There must be one other transaction in between the transactions trading at least one currency pair of the transactions.

4. Simple arbitrage: A token pair should at most occur in two `Swap` events, ensuring that only one arbitrage trade is executed per transaction.

Chapter 2

Price Discovery in Constant Product Markets

Abstract:

Constant product markets are the most common type of automated market maker designs, which has become increasingly popular in the advent of blockchain technology. This paper develops a price discovery framework for constant product markets that comprises three parts. Firstly, I derive a quadratic relationship which expresses trades in terms of price changes. Secondly, I model trade interactions in a structural VAR system. Thirdly, I translate the impulse responses from the VAR to returns through the quadratic equation. The empirical analysis reveals how large trades carry important market information and how a small but sophisticated group of adversarial traders, much like high-frequency traders in traditional markets, play a key role in price discovery on the largest decentralized exchange Uniswap.

1 Introduction

Decentralized exchanges that employ distributed ledger technology primarily use automated market makers to facilitate cryptocurrency trading. These exchanges have gained widespread popularity due to the rapid adoption of blockchain technology, resulting in daily trading volumes of several billion dollars for these trading platforms. The most frequently used automated market making mechanism is the constant product market, which determines prices by holding the relative value of the market maker's inventory constant according to a fixed rule. Despite its simplicity, this market design offers numerous advantages and has become the primary market design for these trading venues.

The focus of this paper is to study price discovery in constant product markets. I develop a framework for price discovery by deriving a deterministic formula for price revision in this market. The derived equation reveals that price changes are predominantly influenced by liquidity takers, while liquidity providers have only an indirect impact on prices. The automated market maker deterministically revise prices based solely on trading. Therefore, unlike what is the case on traditional markets, both public and private information is incorporated into the price by trades alone. This paper's empirical analysis involves categorizing three types of trades on the decentralized exchange Uniswap, which are analogous to human trading, algorithmic trading, and adversarial high-frequency trading in traditional financial markets. The findings indicate that large trades from all trade groups, have a significant and persistent economic impact on price discovery for one of the largest currency pairs on the Uniswap exchange. Furthermore, a small but sophisticated subset of adversarial trades carry important market information and execute in the same direction as future permanent price changes.

The theoretical part of this paper presents a framework for evaluating trades' informativeness and their contribution to price discovery on constant product markets. The framework consists of three parts. Firstly, I derive a formula for transforming trades to price impacts.

Secondly, I estimate trade interactions. Thirdly, I transform trade responses into price changes. The central idea in the methodology comes from recognizing that price revisions on constant product markets are deterministically determined by trades according to a fixed rule. Accordingly, a quadratic equation is derived that mechanically maps trades to price changes. Specifically, the price change from a trade is equivalent to the relative size of a trade plus the square of the relative size of the trade, where the relative size is defined as the signed volume divided by the depth of the market. In this equation, the depth of the market (liquidity provision) only indirectly influences prices by determining the magnitude of price changes resulting from a given trade. In the end, trading activity can be modelled in a structural VAR system and the results can precisely be transformed to price impacts. This framework follows the structure in Hasbrouck (1991), with the crucial difference that price revisions are never estimated but mechanically derived due to the market making design of the constant product market. An advantage of this approach is that no assumptions need to be made regarding how trades affect prices. However, a limitation is that it is challenging to disentangle the effects that private and public information have on prices.

In the first empirical step, I study trade interactions among certain subgroups of trades on the decentralized exchange Uniswap (Adams, Zinsmeister, and Robinson, 2020). Uniswap is a 24-hour market operating on the Ethereum blockchain, allowing cryptocurrency trading to be settled peer-to-peer without any centralized intermediary such as banks or brokers. A trade-level dataset is collected from the Ethereum blockchain, constituting the full trade history of the ether-dollar exchange pair (where ether is Ethereum's native cryptocurrency) from November 2020 through May 2021. During this time, Uniswap operated using the standard constant product market making mechanism, and was the primary trading venue for professional and retail traders. The comprehensive and transparent nature of the data allows for the classification of various trade subgroups, including regular manual (human) trading, algorithmic trading routed through different decentralized finance trading applications, and adversarial trading, involving sophisticated trading strategies like front-running and arbitrage. The trades are further divided by size into small (less than $\$1,000$), medium (between $\$1,000$ and $\$5,000$), and large (more than $\$5,000$) sized trades. The interactions between the trade groups are modeled using a structural vector autoregression (SVAR) system, following an approach similar to the one outlined in Benos et al. (2017).

The results from the structural VAR show that all trade groups exhibit positive autocorrelation, indicating a tendency for purchases to follow previous purchases and sales to follow previous sales. Moreover, large informed adversarial and algorithmic trades execute in the opposite direction of large uninformed trades from the regular trade group and the algorithmic trade group. Among the trade groups, regular (human) trades demonstrate the strongest trade reversal effects and exhibit negligible responses to the other groups. Conversely, the trades within the algorithmic trade group display a diverse range of informed and uninformed characteristics, actively responding to other trade groups, while also experiencing adverse selection from the arbitrage trade group. Small and medium sized trades appear to have no meaningful trade interactions with the other groups and lack any significant influence on their trading.

The second step of the empirical methodology involves converting the impulse responses obtained from the structural VAR estimation to returns by utilizing a quadratic relationship that maps trades to price changes. The ultimate price impact of a trade is commonly regarded as its informativeness. To assess this impact, I calculate the total price effect of all trading activity (responses) following a shock to each trade group (impulses). In practice, this methodology is comparable to other studies that estimate the informativeness (price changes) of trades from different market participants (e.g., Hendershott and Riordan (2013) and Chaboud, Hjalmarsson, and Zikes (2021)).

The return-transformation of the IRFs shows that large trades carry the highest information value in the ether-dollar market on Uniswap. Despite some partial trade reversal effects, large trades overall have an economic significant and permanent effect on price. The adversarial trades execute in the same direction as future price

changes and contribute to a faster price discovery, these results are consistent with previous studies in the high-frequency trading literature (e.g., Brogaard, Hendershott, and Riordan (2014), Chaboud et al. (2014), and Benos et al. (2017)). Small and medium sized trades also align with future price changes, but their impact on price is too insignificant to have any meaningful economic effect.

The remainder of this paper is organized as follows: Section 2 outlines the institutional details of automated market makers and their trading activity. Section 3 presents a framework for price discovery analysis on constant product markets, which provides the foundation for the empirical analysis. Section 4 provides a description of the Uniswap trade data collected from the Ethereum blockchain. In Section 5, the interactions among the trade groups are modelled using a structural VAR model. In Section 6, the IRFs obtained from the SVAR model are transformed into precise price impacts. Finally, Section 7 offers some concluding remarks.

2 Automated market makers and decentralized exchanges

Decentralized markets enable investors to trade directly with each other without the need for any centralized authority. Traditionally, foreign exchange markets and real estate markets are considered decentralized, as traders do not participate through a centralized intermediary. For example, traders on foreign exchange markets often get bid and ask quotes from a broker-dealer network.[1]

Decentralized markets can be facilitated by various technologies, and more recently the advent of decentralized blockchains has created new financial markets commonly referred to as decentralized exchanges (DEXes). These markets are distinct from traditional decentralized and centralized markets in several ways. Firstly, most of these markets use an automated market maker algorithm to determine

[1] As one example, Hagströmer and Menkveld (2019) study information revelation in a network of decentralized foreign exchange markets by collecting trading from nine different trading venues.

prices, resulting in a transparent and deterministic market making process, wherein exact price impacts from trading can be calculated ex ante and ex post. Secondly, the underlying blockchain technology is technically equivalent to a state machine that treats transactions in discrete groups called *blocks*, this implies that trades on decentralized exchanges are also executed in discrete blocks. Thirdly, cryptocurrencies are traded in these markets. Fourthly, transactions are settled on the blockchain (database), with the Ethereum blockchain being the most prominent, without any third-party such as brokers or banks. Fifthly, the decentralized exchanges are accessible to any participant with an internet connection, and they accommodate various highly heterogeneous categories of informed and uninformed traders. Sixthly, the data from decentralized exchanges are fully transparent, enabling precise tracking of individual accounts and trades. Lastly, there are no spreads in the traditional sense in these markets, instead liquidity providers earn fees based on trading volume. In this paper, I delve into various significant economic implications arising from these differences. Among them, the discrepancy in the price discovery process stands out as one of the most noteworthy.

2.1 Automated market makers

An automated market maker (AMM) is a market design where bid and ask quotes are set by an algorithmic scoring rule. Historically, AMMs have been utilized in prediction markets (Hanson, 2003). However, AMMs have more recently gained significance in decentralized finance (DeFi) as the primary market making mechanism in decentralized exchanges (Adams, Zinsmeister, and Robinson, 2020; Adams et al., 2021). AMM markets are categorically different from traditional market making designs (and limit order book markets), in which designated market maker agents post bid and ask quotes. In AMMs, prices are mechanically determined by a deterministic pricing function, and the market making rule only revises prices based on past trading activity. Therefore, information can only be incorporated into the price through market participants affecting the price by trading.

An AMM market for a single exchange pair comprises three components: A liquidity pool for the first asset, a liquidity pool for the second asset, and the market making mechanism. Two types of agents participate in this market: Liquidity providers and liquidity takers (traders). No fixed obligations or formal barriers to entry exist in this market, and any agent may decide to be a liquidity provider, liquidity taker, or both. Liquidity providers contribute liquidity to the liquidity pools, while liquidity takers can purchase either currency by depositing a fixed amount of the other currency into its respective liquidity pool.

To understand this properly, consider two assets X and Y, and their exchange rate X/Y. In this setting, Y can be viewed as a cryptocurrency (or a financial security), while X can be regarded as the US dollar. An exchange pair of X and Y is initially established on the decentralized exchange by a liquidity provider depositing a positive amount of asset X, $\Delta x_t^{LP} > 0$, and a positive amount of asset y, $\Delta y_t^{LP} > 0$, into two separate liquidity pools associated with each asset. These amounts, Δx_t^{LP} and Δy_t^{LP}, can initially differ and their ratio determines the initial price. In this setting, liquidity providers play a role similar to that in traditional markets. However, the only actions the liquidity providers can take are to deposit or to remove inventory from the liquidity pools at the current market price.

The price for which one can trade an infinitesimal amount of X for Y (at time t) is the marginal price, which equals the ratio of the assets in the liquidity pools of the exchange pair, $p_t^M = \frac{x_t}{y_t}$. This price is analogous to the mid-price in a traditional market, which is the average of the bid and ask quotes. The marginal price is different from the less favourable execution price at which Δx_t can be exchanged for Δy_t (at time t), defined by $p_t^E \equiv \frac{\Delta x_t}{\Delta y_t}$. The execution price is similar to the average price on a limit order book, where a trader has to "walk the book" to fill their order. The difference between these two prices is the *price slippage*, and can be regarded as a hidden cost for the trader.

After the liquidity pool has been created, a liquidity taker (trader) can exchange Δx_t for Δy_t by adding Δx_t to the liquidity pool of asset X and removing Δy_t from the liquidity pool of asset Y, or vice versa.

The execution price for Δy_t in terms of Δx_t is determined by the market making mechanism. The most commonly used type of AMM is the constant function market maker (CFMM), where the exchange rate between two assets, X and Y, is determined so that a function of their respective liquidity pools is kept constant before and after each trade,

$$f(x_t + \Delta x_t, y_t - \Delta y_t) = f(x_t, y_t) = k_t. \tag{2.1}$$

Here the invariant, k_t, represents the depth of the market (inventory in the liquidity pools) at time t, which reflects the market's ability to absorb quantities without significant price impact. Although k_t is constant across trades, it is increasing in both x_t and y_t, and changes as liquidity providers deposit or remove liquidity; it therefore inherits the time subscript. $f(\cdot)\colon \mathbb{R}^2_+ \mapsto \mathbb{R}$ is the deterministic pricing function, x_t and y_t are the amounts of assets X and Y in their respective liquidity pools. Δx_t and Δy_t are the changes to these liquidity pools by the liquidity taker and represents the amounts traded.

There are various constant function market makers, including the constant sum market maker and the constant product market maker, with pricing functions $f(x_t, y_t) = x_t + y_t$ and $f(x_t, y_t) = x_t \cdot y_t$ respectively. The constant product market maker is the most prevalent type of AMM and is employed by leading decentralized exchanges on blockchain platforms. Despite its simplicity, this market making mechanism displays attractive features, which are explored in greater detail in Section 3.

Liquidity providers have the flexibility, at any time, to add or remove liquidity of an exchange pair at the current marginal price, p_t^M. This is done by depositing or removing Δx_t^{LP} and Δy_t^{LP} to the respective liquidity pools of assets X and Y, such that the marginal price stays constant,

$$p_t^M = \frac{x_t}{y_t} = \frac{x_t + \Delta x_t^{LP}}{y_t + \Delta y_t^{LP}} = p_{t+1}^M. \tag{2.2}$$

Δx_t^{LP} and Δy_t^{LP} share the same sign, negative when liquidity is removed, and positive when liquidity is added. Liquidity providers hold a specific portion of the total liquidity provided and receive trading fees based on this fraction. Typically, the total trading fee shared among the liquidity providers amount to 0.3% of the trading volume and is paid by the liquidity taker with each trade. However, providing liquidity comes with exchange rate and inventory holding risks. *Impermanent loss* or *divergence loss* refers to the difference between the value of the liquidity provision and the counterfactual value of the assets if they were not deposited in the first place (Cartea, Drissi, and Monga, 2022). This loss occurs if the exchange rate changes in either direction from the time the liquidity was deposited. Therefore, liquidity providers lose to informed traders who change the price in any direction away from the initial exchange rate. Conversely, liquidity providers benefit from uninformed noise traders who do not affect the long-run price.

2.2 Decentralized exchanges on the Ethereum blockchain

The Ethereum blockchain (Buterin, 2013) is often referred to as a "world computer" due to its ability to enable any user to deploy immutable software applications on its network. These applications are referred to as *smart contracts* (Szabo, 1997), which are computer code written by developers and uploaded onto Ethereum's blockchain (database). Smart contracts contain state variables and functions that can modify these state variables. The computing power of Ethereum is relatively low, but the network's innovation lies in its distributed database, which is updated without any trusted third party by network participants following the rules outlined in Wood (2014). In this sense, Ethereum can be considered a decentralized contracts platform, rather than any form of currency.

Every interaction on the Ethereum blockchain requires an Ethereum transaction initiated by an *externally owned account* (EOA), similar to a user account, which specifies the user's Ethereum address. Transactions are sent continuously to the Ethereum network and the network groups transactions together in discrete *blocks*, under the condition that they pay a sufficiently high network fee. Transactions (in the blocks) are sorted and executed based on their transaction fees paid to the network. This process is economically similar to a continuous-time auction. The transactions in each block are executed discretely and sequentially, and once executed, the block is added to the full history of previous transactions (blocks) in a blockchain.

Ether is Ethereum's native cryptocurrency, which is utilized to pay network transaction fees for deploying applications to and interacting with the network. Hence, when using decentralized exchanges, traders are required to pay two types of fees: A transaction fee to the Ethereum network and a trading fee to the liquidity providers on the exchange, equivalent to the bid-ask spread seen in traditional markets. Furthermore, the price slippage between the marginal price and the execution price constitutes a third operational cost for the trader.

Decentralized exchanges on Ethereum are discrete markets subjects to a continuous time auction. The underlying blockchain technology operates as a discrete state machine, but transactions are submitted continuously, resulting in a seemingly similar design as advocated in Budish, Cramton, and Shim (2015), where the "high-frequency arms race" is limited through a batch auction design. Hence, on decentralized exchanges on Ethereum, arbitrageurs compete primarily with transaction fees to capture profitable arbitrage opportunities, rather than with speed as in traditional markets. However, an adversarial race for profits still exists as Ethereum transactions are executed sequentially and independently, allowing for front-running.

Ethereum supports thousands of cryptocurrencies that operate as smart contracts following the ERC-20 standard for cryptocurrencies (Vogelsteller and Buterin, 2015). The adoption of this standard is a significant development for the Ethereum ecosystem, as it establishes a unified currency standard for DeFi applications. *Stablecoins* are an example of these cryptocurrencies, which are designed to be backed by

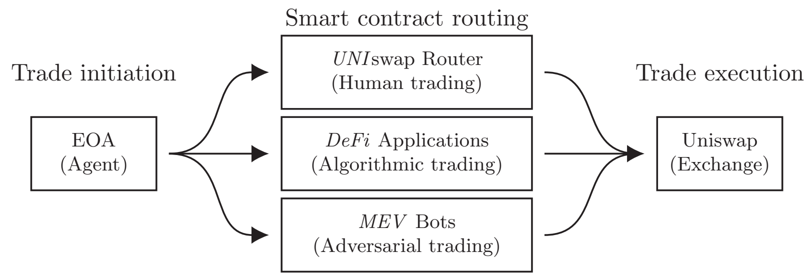
fiat currency. USD Coin (USDC) and Tether (USDT) are examples of stablecoins that are pegged to the US dollar. These currencies are programmed as smart contracts and uploaded to the blockchain, where they can be traded on decentralized exchanges. The most prominent decentralized exchange on Ethereum, Uniswap (Adams, Zinsmeister, and Robinson, 2020; Adams et al., 2021), uses the constant product market maker design, which allows users to trade these crypto assets.

2.3 Various trading activity on decentralized exchanges

The impact of trades on price discovery in markets with asymmetrically informed agents varies with the speed at which they reveal public information and the private information they disclose. Trading is generally motivated by the agent's liquidity needs or private information, which may include advanced knowledge of public information, especially in crypto markets where all trade data are public. Trades carrying significant information have a lasting effect on price and exhibit a pronounced information effect. Although all trades initially impact price depending on their volume, the persistent impact on price is determined by the information content of the trade. Uninformed participants' trades, driven by liquidity needs, are expected to have only a temporary impact on price. On the other hand, informed participants' trades that contain valuable market information impact the price through both their initial price impact and their lasting effect.

Trades on decentralized exchanges that operate on the Ethereum blockchain can be categorized into three types depicted in Figure 2.1: Regular trades that have been sent directly to the exchange (Uniswap in this paper), labeled UNI; trades that have been processed through decentralized finance applications before arriving at the exchange, labeled DeFi; and adversarial trades (including arbitrages) executed by traders known as "searchers", labeled MEV. These categories are analogous to the trading types in traditional financial markets studied in Chaboud, Hjalmarsson, and Zikes (2021), namely "human trading" (manual trading), algorithmic trading (AT), and high-frequency trading (HFT).

Figure 2.1: Routing of Uniswap trades. Every trade on Uniswap is initiated by an *externally owned account* (EOA). The trade can then be sent directly to the Uniswap exchange through the Uniswap router smart contract, or it can first be sent to an intermediary smart contract before arriving at the exchange. The intermediary smart contracts are categorized as either MEV bots used by searchers or decentralized finance (DeFi) smart contracts. In the end, all trades are executed on Uniswap.



UNI trades (human trading)

Most trades on decentralized exchanges can be categorized as "regular trades" (UNI). These trades are directly submitted to the exchange, through an Ethereum transaction to the decentralized exchange's smart contract. These trades can easily be identified by searching for transactions where the sender is an *Externally Owed Account* (EOA) and the recipient is the decentralized exchange. Many of these transactions are likely initiated by individuals who visit the decentralized exchange's web page, connecting their Ethereum wallet applications, and submitting their trades. As such, these trades closely resemble human trading on decentralized exchanges and are expected to have a moderate impact on the price discovery process.

DeFi trades (algorithmic trading)

Before reaching the decentralized exchange, orders can be routed through various smart contracts. This includes routing trades through decentralized finance application, such as *liquidity aggregators* that consolidate liquidity from multiple exchanges to offer better execution prices for users. These applications operate similarly to the collection of algorithms provided by broker-dealers to traders in traditional markets.[2] Algorithmic trading smart contracts can be employed to determine the venue for trade execution, split large orders into smaller trades, and protect traders from front-running, thereby reducing adversarial costs. There is free competition to create these AT smart contracts, and any trader can generally use any decentralized finance application. Hence, the categorization of trades as "regular" or "DeFi" is determined naturally by the choices made by the traders themselves.

A DeFi trade can be identified by observing if its' Ethereum transaction was sent to an intermediary smart contract before arriving at the exchange. These trades are most likely initiated by a mixture of informed and uninformed agents, as these protocols provide more advanced trading features than a single decentralized exchange, but are also accessible for any retail trader.

MEV trades (adversarial trading)

In the event that the law of one price (LOP) breaks down, arbitrageurs intervene by selling overpriced assets and buying under-priced assets to re-establish the LOP. These arbitrage opportunities may occur either within a single exchange as triangular arbitrages or across exchanges as cross-exchange arbitrages. In decentralized exchanges, arbitrage activity is automated, and searchers process pending orders to identify deviations from the efficient price and trade accordingly. Since liquidity providers do not revise the quotes in response to changing market conditions but instead rely on the automated market maker rule to update prices after trades, there are no stale quotes from the liquidity providers that arbitrageurs can snipe. Therefore, *toxic arbitrage*, along the lines of Foucault, Kozhan, and Tham (2017) and Aquilina, Budish, and O'Neill (2021) do not exist in these markets.

[2] Algorithmic trading is generally defined as using computer automation to submit and manage orders.

Arbitrageurs have a dual impact on liquidity providers, one adversarial and one favourable. On one hand, they contribute to the adversarial selection effect that liquidity providers have to endure, not by sniping stale quotes but by modifying the price which results in impermanent loss. On the other hand, arbitrageurs can aid liquidity providers by neutralizing price anomalies that cause a temporary high impermanent loss. Furthermore, by helping decentralized exchanges track the market price, arbitrageurs provide fair pricing to the market, increased trading volume, and higher trading fees for liquidity providers. Therefore, arbitrageurs on decentralized exchanges are part of the price discovery process, as they rebalance liquidity pools, benefiting all investors as prices are brought closer to fundamentals, similar to the arbitrageurs studied in Gromb and Vayanos (2002).

However, this does not imply that the automated market maker design is immune to negative externalities arising from informed agents and unhealthy market competition. According to Park (2021), the constant product pricing function used to determine the execution price, always provides an opportunity for profitable front-running due to the deterministic price impact of trades. As a result, other traders may be "sniped". For instance, sandwich attacks are a common phenomenon in which an adversarial agent places a trade just before and after a victim's trade. The first trade is executed in the same direction as the victim's trade, causing the execution price for the victim to change so that they pay the maximum allowable price slippage, up to their willingness-to-pay for the trade. The adversarial agent then profits from the last trade by trading in the opposite direction to the victim's trade.

Maximal extractable value (MEV) refers to the various ways of extracting profits from (mainly) decentralized exchanges, such as healthy arbitrage and adversarial sandwich attacks (Daian et al., 2019).[3] Much of the MEV trading carried out by searchers is similar to high-frequency trading (HFT) in traditional markets. Similar to HFT, there is free entry to become a MEV searcher, as the blockchain is an open protocol and all data are public, leading to a high competition environment. The searchers have a positive effect on price discovery, as they trade against transitory price movements, and in the direction towards the efficient price. This effect has been studied in traditional markets, for example, Chaboud et al. (2014) show that arbitrageurs improve price efficiency. However, the activities of searchers also result in adversarial selection costs for other traders. MEV trades can be empirically identified by observing if the trades are routed through smart contracts operated by MEV searchers. These trades are executed by sophisticated agents, considered informed traders, and are expected to have a significant impact on price discovery.

3 Price discovery in constant product markets

This section presents an empirical framework to study the process of price discovery in the constant product market. Several aspects of the market microstructure in the constant product market bear similarities to those of traditional markets. However, differences in the market's operations lead to disparities in the way price discovery occurs.

The market conditions of the constant product maker are reminiscent of the classic market maker setting explored in Hasbrouck (1991). However, instead of a market maker posting bid and ask quotes, prices in constant product markets are governed by the constant product rule. Liquidity providers in this market are willing to buy or sell any asset provided that the product of the liquidity pools, known as the invariant, remains constant. This results in significant implications for how information is integrated into the price, as com-

[3]Technically, MEV is defined as the maximal value that can be extracted from a block by including, excluding, or changing the order of transactions. This could, for example, be to font-run a high value trade on a decentralized exchange. However, it could also relate to other blockchain activity outside of decentralized exchanges, such as "liquidations" in on-chain debt markets.

pared to the conventional agent market making setting.

In conventional settings, market makers have the ability to include public information (such as earnings and merger announcements released after the most recent trade) in their quote revisions. As a result, prices can change without actual trading. The designated market maker updates prices with both private information from recent trades and potential new public information. In contrast, in the constant product setting with the deterministic market maker algorithm, prices can only be adjusted based on previous trades. This implies that both public and private information are integrated into the price through trading on decentralized exchanges using the constant product rule.

In order to gain a deeper understanding of the impact of trading on prices, consider the two assets X and Y and their exchange rate X/Y. As previously, the exchange pair comprises two liquidity pools with $x_t > 0$ units of asset X and $y_t > 0$ units of asset Y, at time t. In this setting, time is referred to as "trade time" and is incremented discretely with each trade (buy or sell) or liquidity provision (deposit or removal).

A sell is defined as selling asset Y for asset X, while a buy is defined as buying asset Y for asset X. Assets X and Y can refer to any assets, however in the empirical section of this paper, X denotes dollars, and Y denotes Ethereum's native cryptocurrency, *ether*. A trade at time t is characterized by its signed (dollar) volume Δx_t. If the trade buys asset Y (removes Δy_t of asset Y from its liquidity pool), Δx_t is positive (inputs Δx_t of asset X into its liquidity pool). On the other hand, if the trade sells Y (inputs Δy_t of asset Y into its liquidity pool), Δx_t is negative (removes Δx_t of asset X from its liquidity pool). In this model, the sequence of trades and subsequent marginal price (mid-price) revisions occur in the following order: First, the marginal price, p_t^M at time t, is determined by the trade in the previous time period, Δx_{t-1}. Then, the trade Δx_t takes place at this price. Afterwards the automated market maker updates the marginal price for the subsequent time period, p_{t+1}^M.

In effect, all orders placed in this market are market orders. To

purchase Δy_t units of asset Y, a trader has to pay Δx_t units of asset X. In the constant product market, the execution price at time t is determined so that the product of the liquidity pools is constant. This means that,

$$f(x_t, y_t) = f(x_t + \Delta x_t, y_t - \Delta y_t) = k_t, \tag{2.3}$$

where the pricing function, $f(x_t, y_t)$, is simply defined as the product of the assets in the liquidity pools, yielding the constant product pricing rule,

$$x_t \cdot y_t = (x_t + \Delta x_t) \cdot (y_t - \Delta y_t) = k_t. \tag{2.4}$$

The amount of asset X, Δx_t, a trader has to pay for a certain amount of asset Y, Δy_t, is equivalent to the well-known expression of price multiplied by quantity, represented as,

$$\Delta x_t = p_t^E \cdot \Delta y_t. \tag{2.5}$$

It is worth noting that in the constant product market, the amount of asset X, Δx_t, a trader has to pay for Δy_t amount of asset Y, remains the same regardless of whether the order is executed in one large trade or multiple consecutive trades. Therefore, the order in which a sequence of trades is executed does not matter for the final inventory of the liquditiy pools. Furthermore, the execution price, p_t^E, (dollars in terms of ether) the trader has to pay is defined by,

$$p_t^E \equiv \frac{\Delta x_t}{\Delta y_t} = \frac{x_t}{y_t - \Delta y_t}. \tag{2.6}$$

Since Δx_t is a function of Δy_t, the right-hand side of the equality in Equation 2.6 is derived by solving for Δx_t in Equation 2.4 and substituting it into the definition for the execution price. The execution price function exhibits several desirable properties. In particular, it increases as the trade size increases, due to $\frac{\partial p_t^E}{\partial \Delta y_t} > 0$, and this increase occurs at an increasing rate because $\frac{\partial^2 p_t^E}{\partial \Delta (y_t)^2} > 0$. As a trader's desired quantity approaches the inventory of the liquidity pool, the execution price approaches infinity,

$$p_t^E \to \infty \quad \text{as} \quad \Delta y_t \to y_t. \tag{2.7}$$

Conversely, as the trade volume becomes relatively small compared to the size of the liquidity pools, the execution price approaches the marginal price and the price slippage goes to zero,

$$p_t^E \to p_t^M \quad \text{as} \quad \Delta y_t \to 0. \tag{2.8}$$

By solving for the trade Δx_t in Equation 2.4 and substituting it into the expression for the execution price (Equation 2.6), one can decompose the execution price into the marginal price and a multiplier that depends on the trade's size and the liquidity pool's inventory, leading to a more comprehensive understanding of this relationship,

$$p_t^E = \frac{\Delta x_t}{\Delta y_t} = p_t^M \cdot \left(\frac{1}{1 - \frac{\Delta y_t}{y_t}} \right) \to p_t^M = \frac{x_t}{y_t} \quad \text{as} \quad \frac{\Delta y_t}{y_t} \to 0. \tag{2.9}$$

The execution price can converge to the marginal price either by $\Delta y_t \to 0$ or by $y_t \to \infty$. This means that the execution price approaches the marginal price, either as the trade size goes to zero or as the inventory of the liquidity pool approaches infinity (or both). This relationship can be thought of as the mechanism behind a trade's price impact, where larger trades or illiquid markets will result in larger price impacts.

In a traditional market making setting, the mid-price is typically the primary variable used for price discovery analysis. Likewise, in the constant product setting, the marginal price is a suitable choice for studying price discovery. As noted earlier, the marginal price

is analogous to the mid-price. The *efficient price* is defined as the "end-of-trading price", given all information available at time $t-1$. Therefore, the marginal price variable can be considered equivalent to the efficient price since no other price changes occur besides those resulting from the trade itself. The revised price fulfills a zero-profit condition for the market maker, as no revenue is taken by the algorithmic market maker. However, a trading fee of typically 0.3%, which is disregarded in this model, is shared among the liquidity providers. This fee serves the same purpose as the spread for liquidity providers in conventional markets.

The immediate price impact of a trade represents all new information provided by that trade, including both private and public information. The relative change in price resulting from a trade is,

$$r_{t+1} = \frac{p_{t+1}^M}{p_t^M} = \frac{y_t(x_t + \Delta x_t)}{x_t(y_t - \Delta y_t)}. \tag{2.10}$$

It is possible to rewrite Equation 2.10 to demonstrate that it is explicitly dependent on the trade and its liquidity pool. By solving for the change in the liquidity pool of asset Y, i.e., $y_t - \Delta y_t$, in Equation 2.4, and substituting it into Equation 2.10, it is possible to represent the price impact solely in terms of the (dollar) trade and the liquidity pool of asset X,

$$r_{t+1} = 1 + 2\frac{\Delta x_t}{x_t} + \frac{(\Delta x_t)^2}{x_t^2}. \tag{2.11}$$

From this representation it is clear that the price impact can be deterministically expressed as a function of the relative change in one asset and that it is well-defined for both a buy and a sell.[4] Price revisions are also completely neutral, in the sense that the algorithmic market maker cannot discriminate between informed and uninformed traders. This representation is distinct from the standard model with a designated market maker agent, where price impacts cannot be modelled

[4]For a sell, the sign of Δx_t is negative, leading to $r_{t+1} = 1 - 2\frac{\Delta x_t}{x_t} + \frac{(\Delta x_t)^2}{x_t^2}$.

individually, but only in an average sense as the revisions could be due to non-trade public information. This is advantageous when studying price discovery, as the price impacts of trades does not need to be estimated, but can be precisely calculated.

It is important to note that Equation 2.11 implies that liquidity providers who deposit and withdraw liquidity do not directly affect the marginal price. However, they do impact the depth of the market, k_t, which in turn affects the magnitude of price impacts. As the depth of the liquidity pools increases (decreases), the price impact of trading decreases (increases).[5] Strategic deposits and withdraws by liquidity providers can have an uneven effect on price impacts if they target different types of trades in the market. However, empirical studies indicate that liquidity provision is conducted rather passively on decentralized exchanges.[6]

At this point it is appropriate to define the relative trade variable which is normalized by the current inventory of the liquidity pool,

$$\phi_t \equiv \frac{\Delta x_t}{x_t}. \tag{2.12}$$

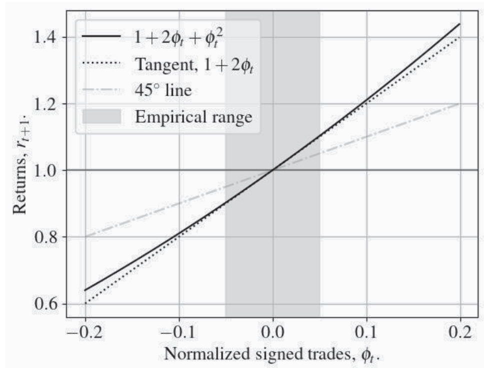
This definition offers several advantages from an empirical standpoint. Firstly, it accounts for any changes in the liquidity pool that may occur during the sample period. This is crucial, since the average trade's price impact can vary throughout the sample period due to fluctuations in the liquidity pool size. Secondly, this definition simplifies

[5]Hasbrouck, Rivera, and Saleh (2022) explores the relationship between offering higher fees to liquidity providers, which is analogous to widening the spread in traditional markets, and lowering fees to encourage higher trading volumes. Their findings indicate that raising trading fees on decentralized exchanges can lead to a rise in the equilibrium trading volume. O'Neill (2022) indicates that for a particular exchange pair with a specified trading fee and trading volume, liquidity providers converge to an "equilibrium amount" of liquidity provided for that currency pair.

[6]Heimbach, Wang, and Wattenhofer (2021) show that approximately 70% of liquidity providers only provide liquidity to one trading pair and do not frequently move their funds. Along the same lines, Cartea, Drissi, and Monga (2022) show that liquidity providers have deposited liquidity non-strategically at a significant loss in the ether-dollar exchange pair on Uniswap.

empirical modeling, as it allows for straightforward modeling of the relative trades' evolution, which can then be transformed into price changes using Equation 2.11.

Figure 2.2: Deterministic Price Revision Formula. This figure illustrates Equation 2.11 for the interval $[-0.2, 0.2]$. All data used in the empirical section of this paper falls within the empirical range of $[-0.05, 0.05]$, which is represented by the grey area in the figure.



The parabolic curve represented by Equation 2.11 is depicted in Figure 2.2, spanning the relevant range of $[-0.2, 0.2]$. This curve is convex in this range, with the tangent line positioned below the curve. The derivative, $\frac{\partial r_{t+1}}{\partial \phi_t} = 2 + 2\phi_t$, indicates that the price change exceeds one unit, and thus the slope is steeper than the 45° line. Furthermore, the second derivative, $\frac{\partial^2 r_{t+1}}{\partial \phi_t^2} = 2 > 0$, demonstrates that as the relative trade size increases, the price change becomes increasingly significant. As a result, the price impact of a trade makes a consecutive trade in the same direction more expensive, as the new marginal price is higher. This process acts as a discrete and automated method of adjusting the price to meet the demand of order flow, similar to how a traditional market maker revises their quotes in response to new information (order flow) in the market. In the empirical range of $[-0.05, 0.05]$, the relationship between the variables is relatively linear, while in a narrower range around $\phi_t = 0$, where most trades take place, the change in price is approximately linear.

Hasbrouck (1991) introduces the conventional method for price discovery, which incorporates the impact of a trade's information through its price impact and persistence effect on price. This approach employs a VAR (vector autoregression) model that includes two autoregressive equations. The first equation represents the price revision (return) as a function of previous price revisions and trading activities, while the second equation represents the trade as a function of previous trading and price revisions. While no structural assumptions are made regarding the nature of information, it is commonly assumed in this context that trading is driven by private information and that public information is incorporated into the price by the market maker.

In contrast to the standard market, where designated market makers can incorporate public information into price revisions, the constant product market utilizes the deterministic rule to revise the price solely based on the direction and the volume of the most recent trade. Consequently, the automated market maker loses a degree of freedom in its ability to factor in additional public information when revising prices. As a result, it becomes difficult to disentangle the price impact of a trade into public and private information. Nonetheless, it is worth noting that traditional markets also incorporate public information into prices through trading. For instance, market makers may be required to smooth prices, which limits the speed and size of price adjustments, and traders may utilize public information, such as earnings announcements or forecasts, to inform their decisions. For example, Andersen et al. (2003) demonstrate that the foreign exchange market's price discovery is driven by public information in the form of macroeconomic news announcements.

In terms of modelling price discovery in the constant product market, the persistent effect of a trade on price, is solely based on the trade's relationship to future trading, as future trading is the only factor that will affect the future price. All trades have an initial price impact determined by their volume relative to the current inventory of the liquidity pool. Informed trades that carry significant market information have the same initial price impact as uninformed trades of the same size and marginal price (mid-price). However, informed trades have a persistent impact on price as they influence future trading behavior. As a result, this paper's price discovery analysis consists of two parts. Firstly, I model the evolution of relative trades, ϕ_t, using a VAR framework similar to that in Benos et al. (2017), where different types of trades are categorized. Secondly, I map the relationship of the relative trades to returns through Equation 2.11.

4 Data and trade classification

4.1 The Uniswap decentralized exchange

This paper's empirical findings are based on transaction-level data from the ether-dollar exchange rate from Uniswap version 2. Uniswap version 2 is a decentralized exchange operating on the Ethereum blockchain, and this paper refers to it as simply "Uniswap". The main advantage of utilizing data from this decentralized exchange is that Uniswap operates as a constant product market in its simplest form and precisely reflects the modelling in Section 3. As a result of this market design, liquidity provision is likely to be non-strategic.[7] It is a 24-hour market where ERC-20 cryptocurrencies are traded every day of the week. However, neither ether (Ethereum's native cryptocurrency) nor the US dollar can be traded directly on Uniswap because they do not conform to the ERC-20 cryptocurrency standard. Thus, the exchange pair studied in this paper involves Wrapped Ether (wETH) and the "stablecoin" USD Coin (USDC). These cryptocurrencies are pegged to ether and the US dollar, respectively, and allow them to be traded on Uniswap. The wETH-USDC currency pair is one of the most liquid pairs on Uniswap and was the first trading pair introduced on the decentralized exchange.

The primary data are obtained by running and indexing an

[7]For instance, Uniswap version 3 employs a feature known as *concentrated liquidity*, whereby liquidity is concentrated within a defined price range for specific trading fees. This allows liquidity providers to more easily capitalize on strategic opportunities.

Ethereum archive node containing the complete transaction history of the blockchain (Erigon Team, 2022; TrueBlocks Team, 2022). The Ethereum transaction data are transformed into a final trade-level dataset, which include every trade and liquidity provision made in the wETH-USDC exchange pair between the 17th of November 2020 and the 5th of May 2021. During this time period, Uniswap was the primary exchange for both retail and professional traders on Ethereum, making it an ideal representation of trading activity on a constant product market.[8] While all trades are collected, the frequency is much lower than in traditional markets, where trading occurs on a millisecond basis. On Uniswap, typically 1.25 trades take place per block, equivalent to approximately one trade every 11 seconds.

In addition to the trade data, metadata are gathered to describe the route each trade followed before execution at Uniswap. This information is used to classify trades into three categories: UNI trades (human trading), DeFi trades (algorithmic trading), and MEV trades (adversarial trading).

4.2 Various trading activity on Uniswap

Decentralized exchanges attract a diverse range of liquidity takers, each with unique trading behaviors and private information. Trades represent private information and demand for liquidity, and different market participants exhibiting distinct trade characteristics. The routing of trades to Uniswap occurs through three primary channels. The first and most common route involves trades sent directly to

[8]The raw data covers trading activity from July 29th, 2020 up to the present day. However, the dataset only includes transactions that took place between November 17th, 2020 and May 5th, 2021. There are two reasons for this restriction. Firstly, prior to November 17th, 2020, Uniswap had introduced an incentives program designed to attract liquidity to the exchange, which rewarded liquidity providers with extra commissions and resulted in a significant increase in liquidity in the wETH-USDC pool. Secondly, on May 5th, 2021, Uniswap version 3 was launched, causing much of the trading activity to shift to the new version. After the release of the new version, Uniswap's web interface routes through version 2 only if it is specifically selected or has a better price, leading to a significant amount of retail trading activity being concentrated on version 3.

Table 2.1: Trade volume distribution. This table presents descriptive statistics of dollar trade volume for the trade groups: UNI, DeFi, and MEV. Within the UNI and DeFi groups, there are three distinct categories based on the size of trades, namely small, medium, and large.

	std	mean	25%	50%	75%	max	N
$\text{UNI}_t^{\text{Small}}$	307.32	462.88	199.73	435.94	700.00	1,010.07	369,317
$\text{UNI}_t^{\text{Medium}}$	1,191.06	2,577.34	1,550.00	2,280.43	3,435.91	5,000.00	376,031
$\text{UNI}_t^{\text{Large}}$	58,471.71	27,476.41	7,514.64	11,384.91	24,671.53	4,000,000.00	304,569
$\text{DeFi}_t^{\text{Small}}$	320.09	403.42	107.66	340.00	651.65	1,010.06	42,389
$\text{DeFi}_t^{\text{Medium}}$	1,190.26	2,645.12	1,606.51	2,408.85	3,572.43	5,000.00	46,429
$\text{DeFi}_t^{\text{Large}}$	97,450.48	52,065.37	9,230.00	19,825.00	59,481.79	4,902,855.46	66,445
$\text{MEV}_t^{\text{Small}}$	275.36	596.76	399.37	634.49	830.05	1,009.76	2,009
$\text{MEV}_t^{\text{Medium}}$	1,136.91	2,866.38	1,887.72	2,810.15	3,832.93	5,000.00	8,825
$\text{MEV}_t^{\text{Large}}$	110,594.23	80,692.72	13,803.42	50,609.97	109,010.96	3,992,969.49	37,670

Uniswap through the Uniswap router smart contract. The second route involves trades that pass through various DeFi applications before reaching Uniswap. The third route involves trades facilitated by adversarial trading MEV bots, which execute complex trading strategies such as arbitrage and sandwich attacks. Analogous to human, algorithmic, and high-frequency trading in traditional markets, these three trade groups represent different forms of trading on decentralized exchanges.

In this paper, I categorize human trading as trades sent directly to Uniswap, labelled *UNI*; trades routed through decentralized finance applications, labelled *DeFi*; and trades routed through searchers' MEV smart contracts, labelled *MEV*. As trades on Uniswap are executed in a sequential manner, these categories are mutually exclusive (disjoint), therefore each trade is assigned one label (indicator) equal to 1 and the others are set to 0. This means that there is no contemporaneous trading in the distinct groups.

Informed traders typically trade larger volumes at any given price point (Easley and O'Hara, 1987), and their price impact is assumed to be more persistent than that of non-informed traders. To further analyze these trade groups, I have divided them into small, medium, and large trades based on cutoff points of 1,010 and 5,000 dollars, representing the 33rd and 66th percentiles of trade volumes.

Table 2.1 illustrates the distribution of trade volumes by dollar amount across the different trade groups. The small, medium, and large trade categories are evenly represented in the UNI and DeFi groups. However, the MEV group is primarily composed of large trades. Small trades with volumes under 1,010 dollars are likely to have minimal immediate price impact and play a minor role in price discovery. Similarly, medium-sized trades, with average volumes of approximately 3,000 dollars, are most likely driven by retail users' liquidity demands. On the other hand, the large trade groups have significantly higher volumes, with mean trade volumes of 27,000, 52,000, and 81,000 dollars. These trades are expected to have a substantial immediate price impact due to their larger volume.

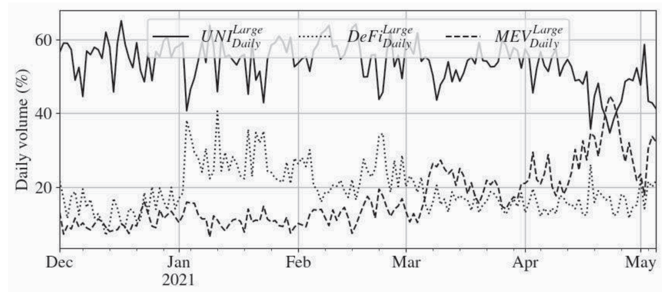
Figure 2.3 shows the daily relative trading volume for each trade group, and Panel 2.3a highlights the dominance of large trades in all groups. The relative volume of MEV trades increases throughout the sample period, starting at 10% and reaching 30% towards the end. The small and medium regular (UNI) trading volume experiences a decline over the sample period, which could be attributed to industry growth and price appreciation during this period. Based on the figure, it appears that large trades play a crucial role in price discovery since they constitute the majority of total trading volume.

UNI trades (human trading)

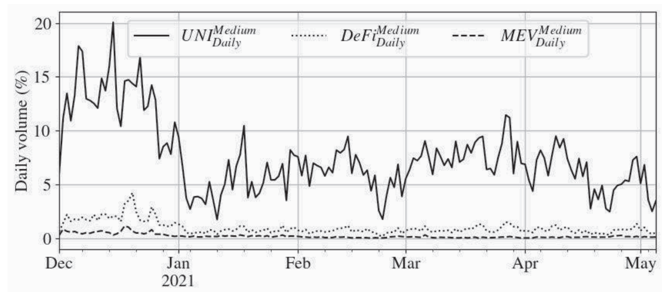
Of all trades, about one million (84%) are sent directly from a user account to the Uniswap exchange through the Uniswap router, and are classified as "human" trades since no other smart contract is involved in the transaction. These trades are executed by 201,990 unique Ethereum addresses, averaging about 5 trades per address. While an agent may use multiple Ethereum addresses, Victor and Weintraud (2021) estimate that 82% of Ethereum accounts belong to unique users, suggesting that many retail traders use these transactions.

Figure 2.3: Relative trade volume. This figure shows the daily relative dollar trading volume in the wETH-USDC exchange pair on Uniswap, divided into the following categories: 1) Trades that are directly sent to Uniswap, labeled *UNI*. 2) Trades that are routed through decentralized finance smart contracts, labeled *DeFi*. 3) Trades that are routed through smart contracts related to maximal extractable value, labeled *MEV*. Panel 2.3a shows trades with a volume of over 5,000 dollars, Panel 2.3b shows trades with a volume between 1,000 and 5,000 dollars, and Panel 2.3c shows trades with a volume of less than 1,000 dollars.

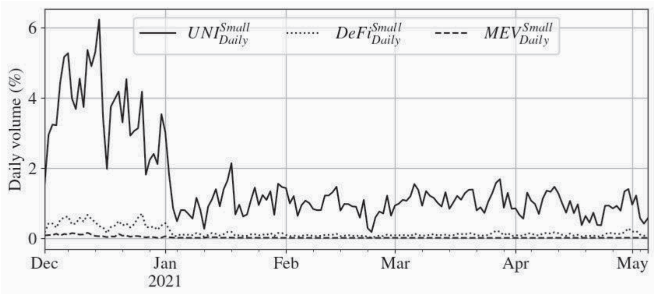
(a) Large trades. Daily relative volume of large trades ($>$ $5,000$ dollars).



(b) Medium trades. Daily relative volume of medium trades ($> 1,000$ and $< 5,000$ dollars).



(c) Small trades. Daily relative volume of small trades ($<$ $1,000$ dollars).



DeFi trades (algorithmic trading)

The dataset contains 155,265 trades categorized as DeFi trades, accounting for around 12% of all trades. These trades route through at least one of 2,088 smart contracts before reaching the Uniswap exchange. They are executed by 43,243 unique Ethereum addresses, which averages to about 4 trades per account.

As Ethereum is an open protocol with low barriers to creating new smart contracts, there are numerous smart contracts on the platform, including widely used commercial applications and those designed by individuals. Among the DeFi trades in the dataset, 35% are routed through the three most common decentralized finance smart contracts: *0x Protocol*, *Metamask*, and *1inch*. The 0x Protocol is an infrastructure for decentralized exchanges that aggregates liquidity from various sources, including on-chain decentralized exchanges and off-chain centralized markets. Metamask is a "wallet" software application for Ethereum that simplifies interaction with the blockchain. 1inch is a decentralized exchange aggregator that routes trades to obtain the best execution price across multiple decentralized exchanges.

MEV trades (adversarial trading)

The trades categorized as MEV trades are identified through Etherscan, a leading Ethereum blockchain explorer and analytics platform, which identifies smart contracts with significant MEV activity. Out of all trades in the dataset, only 4% or 48,504, are MEV trades. These trades are initiated by 605 externally owned accounts and routed through just 18 MEV smart contracts. This indicates that searchers may utilize multiple addresses to initiate transactions. One possible explanation is that a group of searchers share the same deployed smart contract, but different traders' bots initiate the trades. This aligns with industry rumors that suggest around 20 teams are responsible for the majority of MEV extraction on Ethereum. It is also consistent with prior research that indicates 20 arbitrageurs capture 76% of the profits on Uniswap (Hansson, 2022).

Notably, the three most active MEV bots in the data account for 89% of the total MEV trades, equivalent to 43,406 trades. This suggests that the MEV trades are executed by a highly active and informed group of traders, which is unsurprising given the sophisticated trading strategies required for such trades.

4.3 Is liquidity provision strategic?

In the dataset, only 1% of the total observations (15,208) are liquidity provisions, with 9,347 adding liquidity and 5,861 removing liquidity. If liquidity provision is a strategic action, it may be targeted towards specific trade types and market participants, making it difficult to compare the price impact of different groups of trades in a price discovery analysis. The liquidity is provided by 6,618 unique Ethereum addresses. The majority of these addresses provide liquidity infrequently, with 98% (6,514 addresses) adding or removing liquidity less than 10 times, and 43% of addresses making only one liquidity provision. This is consistent with previous empirical findings suggesting that liquidity provision is non-strategic (Heimbach, Wang, and Wattenhofer, 2021; Cartea, Drissi, and Monga, 2022).

A strategic liquidity provision approach is *just-in-time* (JIT) liquidity, where a liquidity provider deposits liquidity just before a trade, and removes it just after the trade. This is feasible because pending trades on decentralized exchanges on the Ethereum network are publicly visible. The purpose of this strategy is to benefit from short-term profits generated by large volume trades. JIT liquidity benefits the trader by offering a more favorable execution price (lower price slippage), but it decreases the fees received by passive liquidity providers. This can affect the price discovery process by decreasing the initial price impact of targeted trades.

JIT liquidity happens on some decentralized exchanges, however, it is unclear if it happens on Uniswap (version 2), which is studied in this paper. To confirm that liquidity provision is non-strategic, I identify JIT liquidity by classifying a pair of liquidity deposit and removal as a JIT liquidity observation if it meets the following criteria: 1) They are in the same block. 2) They are exactly two positions apart in the block. 3) There is a trade in between the deposit and the removal of liquidity. The analysis concludes that no strategic JIT liquidity has occurred historically in the wETH-USDC exchange pair on Uniswap (version 2). This finding is essential as it enables price discovery analysis to be conducted without specific consideration for strategic liquidity provision.

5 Trade group interactions

5.1 Structural VAR specification

Hasbrouck (1991) put forward a price discovery framework consisting of two autoregressive equations. The first equation represents trades as a function of past trading and price revisions, while the second equation expresses price revisions as a function of previous trading and price revisions. The focus of this section is on estimating the first equation, which models trade interactions. The price revision equation is not estimated since price revisions are deterministically determined by trading in the constant product market.

To study the information content of trades and their interactions,

the following relative trade variables are defined,

$$\phi_{t,\text{Small}}^{\text{UNI}} \equiv \frac{\text{UNI}_t^{\text{Small}}}{x_t}, \quad \phi_{t,\text{Medium}}^{\text{UNI}} \equiv \frac{\text{UNI}_t^{\text{Medium}}}{x_t}, \quad \phi_{t,\text{Large}}^{\text{UNI}} \equiv \frac{\text{UNI}_t^{\text{Large}}}{x_t}. \tag{2.13}$$

The normalized trade variables for the DeFi and MEV trade groups, $\left[\phi_{t,\text{Small}}^{\text{DeFi}}, \phi_{t,\text{Medium}}^{\text{UNI}}, ..., \phi_{t,\text{Large}}^{\text{MEV}}\right]$, are defined analogously. Dividing each dollar trade at time t by the inventory of the dollar liquidity pool at that time provides two significant advantages. Firstly, the price impact of the average trade can vary during the sample period due to changes in the size of the liquidity pool. Normalizing each trade by the current size of the liquidity pool controls for this and ensures that trades are comparable over the sample. Secondly, this definition simplifies the transformation of the results from the modelling into price changes using the deterministic return formula, Equation 2.11, as derived in Section 3.

Consequently, the structured VAR model is specified as follows,

$$\boldsymbol{A}\boldsymbol{y}_t = \boldsymbol{c} + \sum_{i=1}^{p} \boldsymbol{A}_i^* \boldsymbol{y}_{t-i} + \boldsymbol{B}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim iid(\boldsymbol{0}, \boldsymbol{I}), \tag{2.14}$$

$$\boldsymbol{y}_t = \left[\phi_{t,\text{Small}}^{\text{UNI}}, \quad \ldots, \quad \phi_{t,\text{Large}}^{\text{MEV}}\right]', \quad \boldsymbol{A} = \boldsymbol{I}, \quad \boldsymbol{B} = \boldsymbol{\Sigma}\boldsymbol{I}. \tag{2.15}$$

Here the left-hand side includes, $\boldsymbol{y}_t$, the normalized trades on group level, and the structural $\boldsymbol{A}$ matrix, which equals the identity matrix. This results in the SVAR model commonly referred to as the "B model". The right-hand side consists of a 9×1 vector, $\boldsymbol{c}$, representing the group-specific intercepts, p number of 9×9 coefficient matrices, $\boldsymbol{A}_i^*$, the lagged normalized trades on group level, $\boldsymbol{y}_{t-i}$, a 9×9 diagonal variance-covariance matrix, $\boldsymbol{B}$, and the error term, $\boldsymbol{\epsilon}_t$, a 9×1 vector representing the unanticipated components of the trades. The model is estimated in trade-time, which is favorable as the precise order of trades is important, allowing for detailed front-running analysis. This is particularly crucial when analyzing MEV trades, which are believed to be triggered by other trades.

The structural diagonal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ imply that the covariances are zero among the trade groups. This assumption is made to ensure that the model closely resembles reality. Because of the construction of how trades are executed on the Ethereum blockchain, only one trade can occur at time t, hence no contemporaneous effects exist between trades. In practice, the covariance matrix is close to diagonal when estimating a reduced form VAR due to the construction of the trade data. However, a correctly specified SVAR model, with no contemporaneous effects, is advantageous when analyzing the impulse responses, which otherwise often impose the Cholesky decomposition on the residual covariance matrix, which in this case would lead to misspecified contemporaneous effects.

In theory, the VAR system (Equation 2.14) could be of any order, however in practice it is estimated up to some lag s. In this paper, the model is estimated using 30 lags, which correspond to approximately 6 minutes in calendar time. The choice of this lag length is based on both estimations with aggregated data and computational limitations.[9]

5.2 Empirical results

Table 2.2 provides a summary of the results obtained from the SVAR estimation. The table shows the sum of all coefficients associated with each trade group for each estimated equation in the SVAR. The rows represent the equations in the SVAR model, while the columns show the summed coefficients for each trade group (in the same order as the columns). Technically, the table is equivalent to the sum of the $\boldsymbol{A}_i^*$ matrices in Equation 2.14, which can be thought of as a one-sided long-run covariance matrix (it is not symmetric). The p-values for the coefficient restrictions, which are shown in parentheses, correspond to the F-test of the null hypothesis that the sum of the coefficients is equal to zero. For the off-diagonal elements, these tests correspond to Granger causality tests, which determines if one variable is useful

[9] Estimations using aggregated data yield similar results across different lag lengths (10, 20, 30, 50, and 75 lags).

in predicting another.

The diagonal of Table 2.2 shows positive and significant autocorrelations for each trade group. This suggests that purchases tend to follow purchases and sales tend to follow sales. This is a well-established pattern in the finance literature, particularly for short lags. Despite the constant product market's inherent trade reversal mechanism, which results in price increases (decreases) following consecutive purchases (sales), it is unsurprising that this pattern is also present on decentralized exchanges. The autocorrelation is especially strong for large MEV trades and small UNI trades, but surprisingly low for large DeFi trades.

Table 2.2: SVAR Results. This table shows the aggregate results for the full SVAR model. This is equivalent to the sum of the $\boldsymbol{A}_i^*$ matrices in Equation 2.14. The rows represent each equation in the estimated model, and the columns represent the summed coefficients over all lags for each trade group. Thus, the diagonal shows the autocorrelation for each trade group. P-values are presented in parentheses, and constitute the F-test of the null hypothesis that the sum of coefficients is equal to 0.

SVAR equations	Sum of SVAR Coefficients								
	$\phi_{t,\mathrm{Small}}^{\mathrm{UNI}}$	$\phi_{t,\mathrm{Medium}}^{\mathrm{UNI}}$	$\phi_{t,\mathrm{Large}}^{\mathrm{UNI}}$	$\phi_{t,\mathrm{Small}}^{\mathrm{DeFi}}$	$\phi_{t,\mathrm{Medium}}^{\mathrm{DeFi}}$	$\phi_{t,\mathrm{Large}}^{\mathrm{DeFi}}$	$\phi_{t,\mathrm{Small}}^{\mathrm{MEV}}$	$\phi_{t,\mathrm{Medium}}^{\mathrm{MEV}}$	$\phi_{t,\mathrm{Large}}^{\mathrm{MEV}}$
$\phi_{t,\mathrm{Small}}^{\mathrm{UNI}}$	0.2783	0.5482	1.1414	0.008	0.0359	0.4448	-0.0008	0.0067	0.2387
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.1)	(0.08)	(0.06)	(0.28)
$\phi_{t,\mathrm{Medium}}^{\mathrm{UNI}}$	0.0245	0.1958	0.5514	-0.0004	-0.0067	0.1085	-0.0009	-0.0082	-0.058
	(0.00)	(0.00)	(0.00)	(0.09)	(0.00)	(0.01)	(0.00)	(0.00)	(0.99)
$\phi_{t,\mathrm{Large}}^{\mathrm{UNI}}$	-0.001	-0.0019	0.2022	-0.0002	-0.0041	-0.1632	-0.0001	-0.0024	-0.228
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\phi_{t,\mathrm{Small}}^{\mathrm{DeFi}}$	0.0563	0.1888	-0.7727	0.1225	0.1638	0.9978	0.0264	0.0436	0.14
	(0.00)	(0.00)	(0.51)	(0.00)	(0.00)	(0.29)	(0.00)	(0.00)	(0.86)
$\phi_{t,\mathrm{Medium}}^{\mathrm{DeFi}}$	0.0075	-0.0044	0.1862	0.009	0.1746	0.628	0.0041	0.047	0.249
	(0.00)	(0.71)	(0.36)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.07)
$\phi_{t,\mathrm{Large}}^{\mathrm{DeFi}}$	-0.0007	-0.002	0.0465	-0.0001	-0.0025	0.0519	-0.0001	-0.0016	-0.1123
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\phi_{t,\mathrm{Small}}^{\mathrm{MEV}}$	0.1081	-0.255	-1.0653	0.1688	0.9272	5.2956	0.1765	0.5282	2.2862
	(0.00)	(0.14)	(0.72)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.26)
$\phi_{t,\mathrm{Medium}}^{\mathrm{MEV}}$	0.0166	-0.0199	0.0206	0.0122	0.1617	1.0536	0.0096	0.1971	0.8639
	(0.00)	(0.38)	(0.96)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\phi_{t,\mathrm{Large}}^{\mathrm{MEV}}$	-0.0004	-0.0007	0.0967	-0.0001	-0.0024	0.0062	-0.0	-0.0014	0.2678
	(0.00)	(0.04)	(0.00)	(0.02)	(0.00)	(0.2)	(0.00)	(0.00)	(0.00)

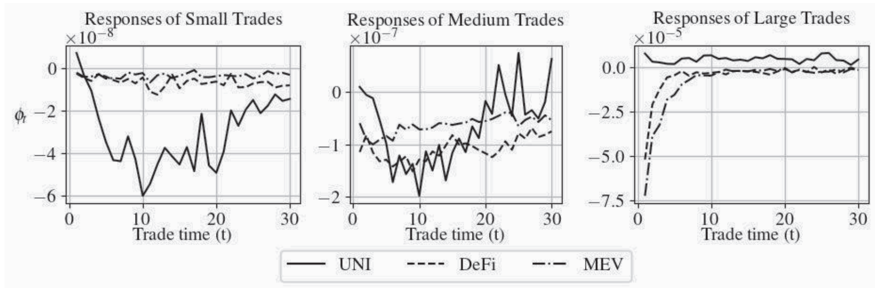
The off-diagonal elements in Table 2.2 illustrate the cross-categorical trade interactions, indicating how each trade group affects the others in aggregate. For example, large UNI trades (third row) have a negative correlation with all other groups, indicating that these groups typically trade in the opposite direction of large UNI trades. This negative relationship is particularly strong for large MEV and DeFi trades, which trade heavily in the opposite direction. One possible explanation for this pattern is that agents are asymmetrically informed and the other groups profits from trade reversal resulting from large (uninformed) UNI trades. Another likely explanation is that large UNI trades are driven by traders' liquidity needs, and that they are willing to pay the price slippage to execute their orders. It is worth noting that large MEV trades, which are believed to have a significant impact on the market due to their informational value, are only found to be statistically significant trading against large UNI and DeFi trades (as well as other MEV trades). In both cases, the effects are strongly negative, indicating that MEV trades act in a corrective and adversarial manner, trading in the opposite direction to large trades that have a significant immediate impact on prices.

Turning to the impulse response analysis, Figure 2.4 shows the responses of all variables to shocks in the large trade groups. Moreover, Figures 2.5 and 2.6 present similar responses for the shocks to medium and small sized trade groups, respectively. Analyzing the responses to the shocks of large UNI, DeFI, and MEV trades illustrated in Figure 2.4, three relationships stand out.

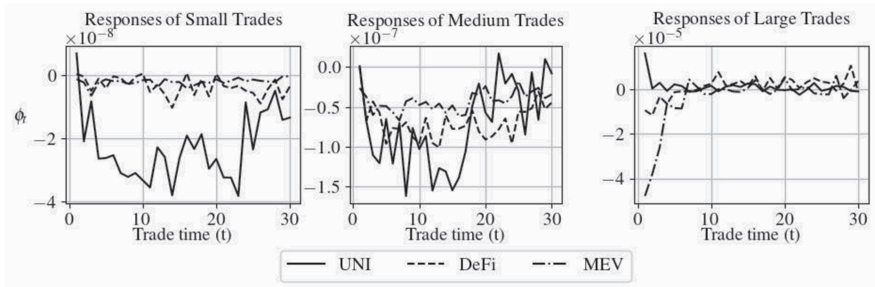
Firstly, the last figure in Panel 2.4a depicts the responses of large UNI trades, large DeFi trades and large MEV trades to a shock in large UNI trades. Both the DeFi and MEV groups demonstrate large negative reactions to a positive UNI trade, suggesting that they trade in the opposite direction. The trade reversal effect is immediate and most pronounced for the first 5 time periods (equivalent to 1 minute in calendar time), and gradually declining afterward. Automated MEV searchers are continuously monitoring incoming orders to Uniswap, and they compete to capture arbitrage profits created by significant price impacts. These agents perform a similar function as high-frequency traders in traditional markets, which have been shown to contribute to faster price discovery (Brogaard, Hendershott, and Riordan, 2014; Chaboud et al., 2014).

Figure 2.4: Large Trade Impulses. Panels 2.4a to 2.4c display all responses, across all trade categories (sizes and trade types), from a one standard deviation shock to the trade variables $\phi^{\text{UNI}}_{t,\text{Large}}$, $\phi^{\text{DeFi}}_{t,\text{Large}}$, and $\phi^{\text{MEV}}_{t,\text{Large}}$, at time t. The initial shocks at time 0 are not displayed in the figures due to their sizes. The index t refers to trades. All graphs use the same units for the y-axis, which refer to the normalized signed trades, ϕ_t. Additionally, the plotted lines on all graphs share the same names. The responses displayed in the panels exhibit a direct connection to the correlations reported in Table 2.2.

(a) IRF UNI: Responses from a one standard deviation shock in $\phi^{\text{UNI}}_{t,\text{Large}}$.

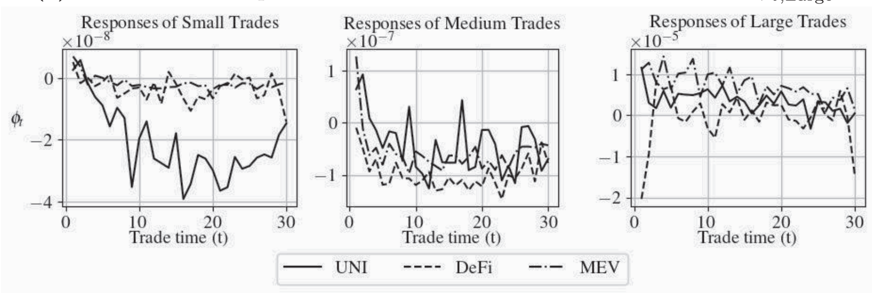


(b) IRF DeFi: Response from a one standard deviation shock in $\phi^{\text{DeFi}}_{t,\text{Large}}$.



Secondly, after a large DeFi trade, large MEV trades off-set the transient price impact by trading in the opposite direction. This effect is depicted in the last figure in Panel 2.4b, and resembles the

(c) IRF MEV: Response from a one standard deviation shock in $\phi_{t,\text{Large}}^{\text{MEV}}$.



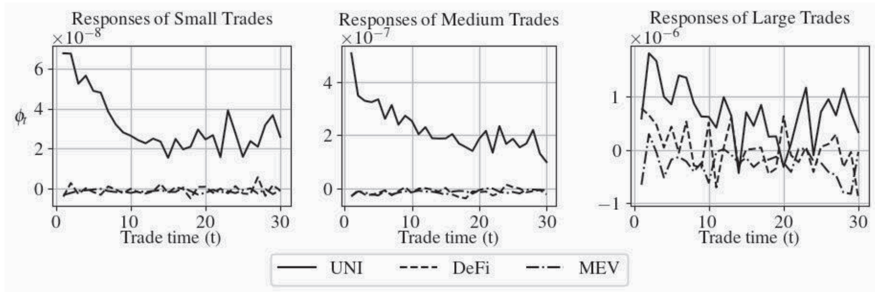
responses to a large UNI trade.

Thirdly, in Panel 2.4c, the last figure shows a large MEV trade followed by an initial trade reversal by the large DeFi trade group. This response can be attributed to the informed trades within the DeFi trade group. DeFi trades are particularly susceptible to immediate reversal effects, but may also impact other trade groups in the same way. This could be due to the diverse set of decentralized finance applications utilized by DeFi trades. For example, the wallet application Metamask is heavily used by retail agents, and thus, many of the trades it processes may have low information value and be driven by liquidity needs. In contrast, 1Inch is a decentralized exchange aggregator that routes trades across multiple decentralized exchanges to achieve the best execution price, and trades routed through this algorithmic trading venue may carry valuable information that can help correct price imbalances.

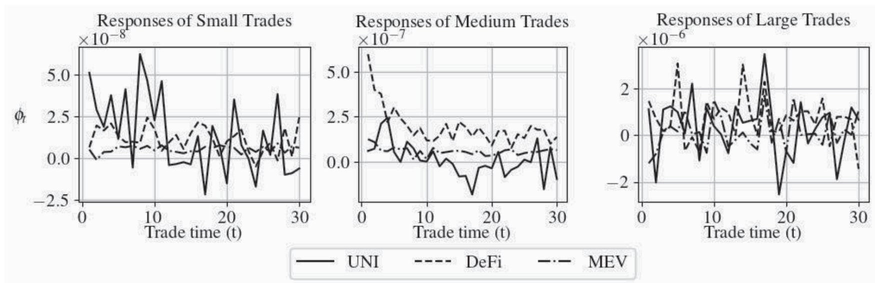
The responses of medium and small sized trades to large trade shocks appear to lack a definitive and significant interpretation. However, when examining Figures 2.5 and 2.6 illustrating medium and small sized trade shocks, clear patterns of autocorrelation emerge (part of the diagonal effect observed in Table 2.2). It is uncertain whether these effects arise from independent trades or sequences of trades following the Kyle and Obizhaeva (2016) model. Nonetheless, it is more likely that consecutive trades are independent trades as

Figure 2.5: Medium Trade Impulses. Panels 2.4a to 2.4c display all responses, across all trade categories (sizes and trade types), from a one standard deviation shock to the trade variables $\phi_{t,\text{Medium}}^{\text{UNI}}, \phi_{t,\text{Medium}}^{\text{DeFi}}$, and $\phi_{t,\text{Medium}}^{\text{MEV}}$, at time t. The initial shocks at time 0 are not displayed in the figures due to their sizes. The index t refers to trades. All graphs use the same units for the y-axis, which refer to the normalized signed trades, ϕ_t. Additionally, the plotted lines on all graphs share the same names. The responses displayed in the panels exhibit a direct connection to the correlations reported in Table 2.2.

(a) IRF UNI: Impulse response of a one standard deviation positive shock in $\phi_{t,\text{Medium}}^{\text{UNI}}$, at time 0.

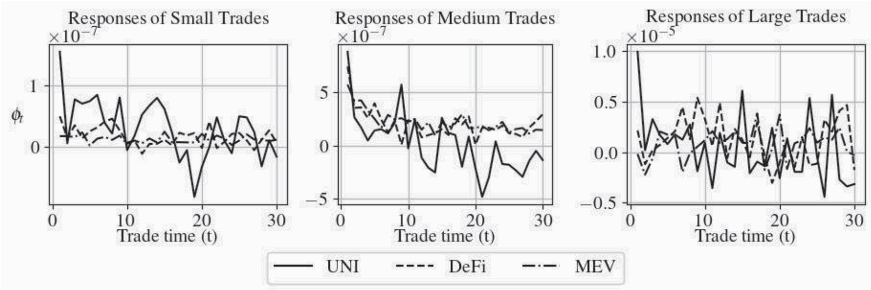


(b) IRF DeFi: Impulse response of a one standard deviation positive shock in $\phi_{t,\text{Medium}}^{\text{DeFi}}$, at time 0.



combining them would result in the same price slippage as executing them separately, albeit at a higher cost.

(c) IRF MEV: Impulse response of a one standard deviation positive shock in $\phi^{\text{MEV}}_{t,\text{Medium}}$, at time 0.



6 Price discovery analysis

6.1 Return transformation

In Section 5.2, the relationship between the trade groups is estimated. This section turns to the second equation in the Hasbrouck (1991) model, which models price revisions. In contrast to the traditional market setting, in a constant product market, the trade interaction estimates can be directly transformed into price changes. One significant advantage of this approach is that it does not require any assumptions about how trades impact prices. By aggregating all responses (trading) following a shock (trade) from one of the trade groups, it is possible to map out the ultimate price effect for each trade group.

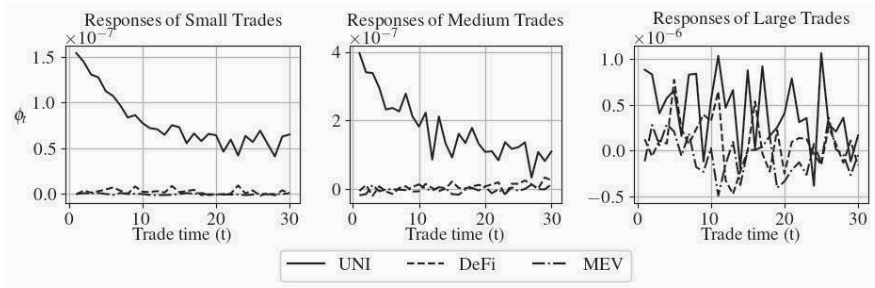
A trade's information effect is typically measured by its price impact and its persistence on price. In the constant product market, the deterministic change in mid-price from a relative trade, is defined for both a buy (ϕ_t) and a sell ($-\phi_t$) as,

$$r_{t+1} = 1 + 2\phi_t + \phi_t^2. \tag{2.16}$$

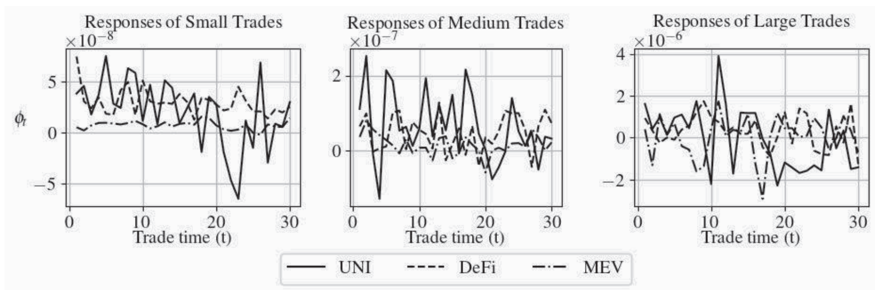
Here, r_{t+1} is the change in price from time t to time $t+1$, and ϕ_t is the normalized trade (signed volume divided by the inventory of the

Figure 2.6: Small Trade Impulses. Panels 2.4a to 2.4c display all responses, across all trade categories (sizes and trade types), from a one standard deviation shock to the trade variables $\phi^{\text{UNI}}_{t,\text{Small}}$, $\phi^{\text{DeFi}}_{t,\text{Small}}$, and $\phi^{\text{MEV}}_{t,\text{Small}}$, at time t. The initial shocks at time 0 are not displayed in the figures due to their sizes. The index t refers to trades. All graphs use the same units for the y-axis, which refer to the normalized signed trades, ϕ_t. Additionally, the plotted lines on all graphs share the same names. The responses displayed in the panels exhibit a direct connection to the correlations reported in Table 2.2.

(a) IRF UNI: Impulse response of a one standard deviation positive shock in $\phi^{\text{UNI}}_{t,\text{Small}}$, at time 0.

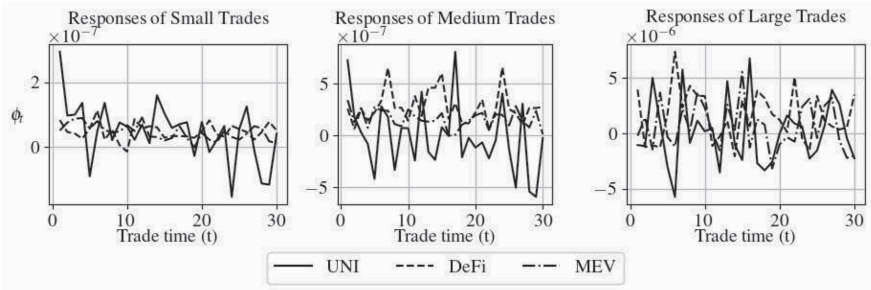


(b) IRF DeFi: Impulse response of a one standard deviation positive shock in $\phi^{\text{DeFi}}_{t,\text{Small}}$, at time 0.



liquidity pool) at time t. Price discovery analyses in traditional markets often assume that trades impact prices linearly. In the constant

(c) **IRF MEV**: Impulse response of a one standard deviation positive shock in $\phi^{\mathrm{MEV}}_{t,\mathrm{Small}}$, at time 0.



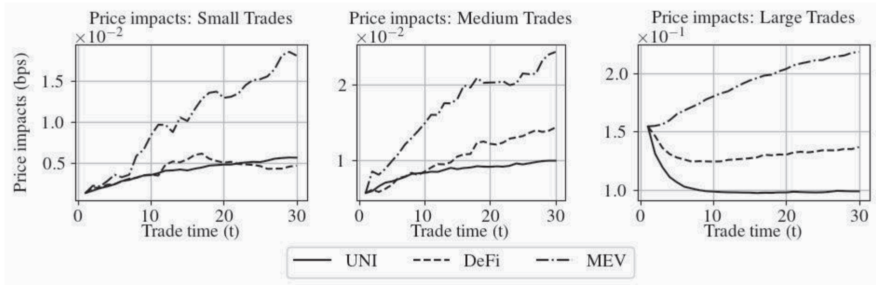
product setting, no assumption is needed regarding the relationship between trades and prices since it is explicitly determined. However, returns are approximately linear in the normalized trades because the squared term is expected to be very small for sufficiently liquid trading pairs.

6.2 Empirical results

Figure 2.7 displays the cumulative price change (return) after a one standard deviation shock (trade) in each trade group, illustrating both the immediate and ultimate effect on price. To measure and aggregate the cumulative returns for each trade group, the following steps are taken: Firstly, the trade responses for all trade groups are captured after each positive shock. Secondly, the nine responses are aggregated for each shock over the next 30 trades. Lastly, the aggregated cumulative trading response is transformed to cumulative returns by Equation 2.16, showing the price evolution after each shock.

Unsurprisingly, the large trades have the highest initial price impact, as they are subject to the greatest initial shocks. Persistent and economically significant price effects can be seen for all trade groups (UNI, DeFi, and MEV) when considering the large trades. This means that these trades carry important market information and have a permanent effect on price.

Figure 2.7: Cumulative Price Impacts. This figure shows the cumulative price impacts from all trading responses after a one standard deviation shock to each trade group. The initial shocks are equivalent to 0.13, 0.56, and 15.43 basis points for the small, medium, and large trade groups respectively. It displays the "total effect" on price after a shock to each trade group. The figure is constructed by transforming the impulse response functions to cumulative returns by applying Equation 2.16 and aggregating them over each trade group. The index t refers to trades.



There is a clear distinction between the large MEV trades and the other large trade groups, as the effect of MEV trades on prices is noticeably more persistent. Following large MEV trades, the initial price impacts are reinforced and these trades execute in the same direction as future price changes. These results are consistent with previous studies in the high-frequency trading literature (e.g., Brogaard, Hendershott, and Riordan (2014) and Benos et al. (2017)), which suggests that the behavior of searchers is comparable to that of high-frequency traders in traditional markets.

Large trades from the UNI and DeFi groups display some partial trade reversal effects, as subsequent trades following the shock are executed in the opposite direction. This phenomenon is particularly evident in the aftermath of a substantial UNI trade, where the initial price impact is counteracted by roughly 30%. One possible explanation for this pattern is that large UNI and DeFi trades are partially driven by liquidity needs and partially guided by informed trading decisions. On the other hand, medium and small trades generate price

Table 2.3: Cumulative Price Impacts. This table shows the cumulative price impacts for each trade group from a one standard deviation shock, 30 trades into the future. The initial shocks are equivalent to 0.13, 0.56, and 15.43 basis points for the small, medium, and large trade groups respectively. The cumulative changes are equal to the first observation subtracted from the last observation in Figure 2.7.

	Small			Medium			Large		
	UNI	DeFi	MEV	UNI	DeFi	MEV	UNI	DeFi	MEV
Initial Shock (bps)	0.13	0.13	0.13	0.56	0.56	0.56	15.43	15.43	15.43
Change (bps)	0.44	0.27	1.68	0.42	0.87	1.77	-5.53	-1.74	6.13
Total Impact (bps)	0.57	0.40	1.81	0.98	1.44	2.33	9.90	13.68	21.56
Volume (M$)	170	17	1	969	122	25	8,368	3,459	3,039
Change × Volume ($)	7,468	463	201	40,343	10,735	4,469	-4,625,155	-603,497	1,862,435
Impact × Volume (M$)	96	6	2	949	176	58	82,857	47,338	65,521

impacts in the same direction as future price changes, albeit with a substantially lower magnitude.

Table 2.3 presents the price impacts of each trade group following a one standard deviation shock. The first three rows of the table are also illustrated in Figure 2.7, where the first row shows the initial shock, the second row shows the subsequent price changes after the shock, and the third row shows the ultimate price impact after 30 time periods (equivalent to approximately 6 minutes in calendar time) for each trade group. Additionally, the last three rows of the table display, for each respective trade group, the trade volume, the price changes after the initial shock multiplied by the volume, and the total price impacts multiplied by the volume. These final metrics reveal to what extent the trade groups ultimately influence prices throughout the sample period.

The price impacts of the large trade groups are significantly higher, even after accounting for the total trading volume, compared to the small and medium sized trade groups. Despite some partial trade reversal effects due to adverse selection, particularly in the case of large UNI trades, the large UNI and DeFi trade groups exhibit persistent and overall economically significant price impacts. Moreover, the ultimate price impacts of all large trade groups have a high level of economic significance. In contrast, small and medium sized trades

have minimal short-term and long-term price impacts. These trades contribute little to the price discovery process, even when controlling for total volume.

7 Conclusion

This paper demonstrates that liquidity-takers drive price revisions in constant product markets, while liquidity provision has only an indirect impact. Both public and private information are integrated into prices through trades. To measure the price impact of trades on a constant product market, I propose a three-step empirical methodology. First, I develop a deterministic formula that describes how price revisions occur after trading. Second, I model trade interactions in a structural VAR system. Third, I use the deterministic returns formula to translate impulse responses from trade interactions into price changes.

The empirical section of this paper involves categorizing three trade groups on Uniswap, a decentralized exchange operating on Ethereum. These groups are similar to human trades, algorithmic trades, and adversarial (high-frequency) trades in traditional markets. I investigate how these groups interact and contribute to the price discovery process using trade-level data from the ether-dollar exchange pair from November 2020 to May 2021.

The results show that large trades from all trade groups carry significant market information and have a persistent impact on price. Additionally, a small group of sophisticated adversarial automated traders execute large trades in the opposite direction of large uninformed trades and in the direction of future permanent price changes.

References

Adams, Hayden, Noah Zinsmeister, and Dan Robinson (2020). *Uniswap v2 core*. Tech. rep. Uniswap (cit. on pp. 83, 86, 91).

Adams, Hayden, Noah Zinsmeister, Moody Salem, River Keefer, and Dan Robinson (2021). *Uniswap v3 core*. Tech. rep. Uniswap (cit. on pp. 86, 91).

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Clara Vega (2003). "Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange". In: *The American Economic Review* 93.1, pp. 38–62 (cit. on p. 102).

Aquilina, Matteo, Eric Budish, and Peter O'Neill (Sept. 2021). "Quantifying the High-Frequency Trading "Arms Race"*". In: *The Quarterly Journal of Economics* 137.1, pp. 493–564 (cit. on p. 93).

Benos, Evangelos, James Brugler, Erik Hjalmarsson, and Filip Zikes (2017). "Interactions among High-Frequency Traders". In: *The Journal of Financial and Quantitative Analysis* 52.4, pp. 1375–1402 (cit. on pp. 84, 85, 103, 121).

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan (2014). "High-Frequency Trading and Price Discovery". In: *The Review of Financial Studies* 27.8, pp. 2267–2306 (cit. on pp. 85, 114, 121).

Budish, Eric, Peter Cramton, and John Shim (July 2015). " The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response *". In: *The Quarterly Journal of Economics* 130.4, pp. 1547–1621 (cit. on p. 90).

Buterin, Vitalik (2013). *A Next-Generation Smart Contract and Decentralized Application Platform*. White paper (cit. on p. 89).

Cartea, Álvaro, Fayçal Drissi, and Marcello Monga (Nov. 2022). "Decentralised Finance and Automated Market Making: Predictable Loss and Optimal Liquidity Provision". In: *SSRN Electronic Journal* (cit. on pp. 89, 100, 109).

Chaboud, Alain, Erik Hjalmarsson, and Filip Zikes (2021). "The evolution of price discovery in an electronic market". In: *Journal of Banking & Finance* 130, p. 106171 (cit. on pp. 84, 91).

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega (2014). "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market". In: *The Journal of Finance* 69.5, pp. 2045–2084 (cit. on pp. 85, 95, 114).

Daian, Philip, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels (2019). *Flash Boys 2.0: Frontrunning, Transaction Reordering, and Consensus Instability in Decentralized Exchanges* (cit. on p. 94).

Easley, David and Maureen O'Hara (1987). "Price, trade size, and information in securities markets". In: *Journal of Financial Economics* 19.1, pp. 69–90 (cit. on p. 105).

Erigon Team (Feb. 2022). *Erigon* (cit. on p. 104).

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham (2017). "Toxic Arbitrage". In: *The Review of Financial Studies* 30.4, pp. 1053–1094 (cit. on p. 93).

Gromb, Denis and Dimitri Vayanos (2002). "Equilibrium and welfare in markets with financially constrained arbitrageurs". In: *Journal of Financial Economics* 66.2. Limits on Arbitrage, pp. 361–407 (cit. on p. 94).

Hagströmer, Björn and Albert J. Menkveld (2019). "Information Revelation in Decentralized Markets". In: *The Journal of Finance* 74.6, pp. 2751–2787 (cit. on p. 85).

Hanson, Robin (2003). "Combinatorial information market design". In: *Information Systems Frontiers* 5.1, pp. 107–119 (cit. on p. 86).

Hansson, Magnus (Nov. 2022). "Arbitrage in Crypto Markets: An Analysis of Primary Ethereum Blockchain Data". In: *SSRN Electronic Journal* (cit. on p. 109).

Hasbrouck, Joel (1991). "Measuring the Information Content of Stock Trades". In: *The Journal of Finance* 46.1, pp. 179–207 (cit. on pp. 83, 95, 102, 110, 118).

Hasbrouck, Joel, Thomas J Rivera, and Fahad Saleh (Aug. 2022). "The Need for Fees at a DEX: How Increases in Fees Can Increase DEX Trading Volume". In: *SSRN Electronic Journal* (cit. on p. 100).

Heimbach, Lioba, Ye Wang, and Roger Wattenhofer (2021). *Behavior of Liquidity Providers in Decentralized Exchanges* (cit. on pp. 100, 109).

Hendershott, Terrence and Ryan Riordan (2013). "Algorithmic Trading and the Market for Liquidity". In: *The Journal of Financial and Quantitative Analysis* 48.4, pp. 1001–1024 (cit. on p. 84).

Kyle, Albert S. and Anna A. Obizhaeva (2016). "Market Microstructure Invariance: Empirical Hypotheses". In: *Econometrica* 84.4, pp. 1345–1404 (cit. on p. 116).

O'Neill, Peter (Jan. 2022). *Can Markets be Fully Automated? Evidence From an "Automated Market Maker"*. Working Paper (cit. on p. 100).

Park, Andreas (Apr. 2021). "Conceptual Flaws of Decentralized Automated Market Making". In: *SSRN Electronic Journal* (cit. on p. 94).

Szabo, Nick (Sept. 1997). "Formalizing and Securing Relationships on Public Networks". In: *First Monday* 2.9 (cit. on p. 89).

TrueBlocks Team (Feb. 2022). *TrueBlocks: Lightweight indexing for any EVM-based blockchain* (cit. on p. 104).

Victor, Friedhelm and Andrea Marie Weintraud (Apr. 2021). "Detecting and Quantifying Wash Trading on Decentralized Cryptocurrency Exchanges". In: *Proceedings of the Web Conference 2021* (cit. on p. 106).

Vogelsteller, Fabian and Vitalik Buterin (Nov. 2015). *EIP-20: Token Standard*. Ethereum Improvement Proposal (cit. on p. 90).

Wood, Gavin (2014). *Ethereum: A Secure Decentralised Generalised Transaction Ledger*. Yellow paper (cit. on p. 89).

Chapter 3

Evolution of Topics in Central Bank Speech Communication

Abstract:

This paper studies the content of central bank speech communication from 1997 through 2020 and asks the following questions: (i) What global topics do central banks commonly discuss? (ii) How have these topics evolved over time? I turn to natural language processing, and more specifically Dynamic Topic Models, to answer these questions. The analysis consists of an aggregate study of nine major central banks and a case study of the Federal Reserve, which allows for region specific control variables. I show that: (i) Central banks address a broad range of topics. (ii) The topics are well captured by Dynamic Topic Models. (iii) The global topics exhibit strong and significant autoregressive properties not easily explained by financial control variables, suggesting that the topics could exhibit a "narrative effect".

1 Introduction

> I'm a weatherman, I'm a showman, and I'm an economist. I'm expected to be, and I am, a storyteller. I tell stories about the future.[1] (Stefan Ingves, governor of the Swedish Riksbank)

Central bank communication affects financial markets (Cook and Hahn, 1989; Cochrane and Piazzesi, 2002; Nakamura and Steinsson, 2018) and makes monetary policy more predictable. The primary objectives of central bank communication have traditionally been two-fold: First, to disclose private information to the public, and second, to influence and coordinate the expectations of financial markets (Woodford, 2001; Amato, Morris, and Shin, 2002; Blinder et al., 2008). This suggests that central bank communication should focus on topics that are closely tied to monetary policy. However, is this always the case?

In this paper, I give a descriptive analysis of speeches delivered by the major central banks associated with the Bank for International Settlements (BIS). The study aims to address two primary questions: (i) What global topics do central banks commonly discuss? (ii) How have these topics evolved over time? To answer these questions, I turn to natural language processing (NLP) to analyze the content of speeches from 9 major central banks using Dynamic Topic Models (DTM) (Blei and Lafferty, 2006).

The analysis shows that the central banks talk about a broad range of topics, not all related to the classic theory of central bank communication. The estimated topics are persistent and exhibit a large significant autoregressive effect which is robust to different model specifications. To control for underlying variables that may drive this persistence, a case study is conducted using data solely from the Federal Reserve. The results of the case study suggest that topic persistence cannot be easily explained by controlling for regional

[1] https://www.bloomberg.com/news/articles/2019-09-16/wall-street-used-to-crunch-numbers-they-ve-moved-on-to-stories

financial variables that are associated with the topics. This suggests that the topics are subject to a "narrative effect", where economic narratives, as described by Shiller (2017), may drive the persistence.[2]

The body of research relating narratives to central bank communication is growing and there is evidence that narratives affect the economy (Nyman et al., 2018; Hansen, McMahon, and Tong, 2019; Ellen, Larsen, and Thorsrud, 2019). Furthermore, textual analysis has been employed in previous research to estimate economic narratives. For instance, Bertsch, Hull, and Zhang (2021) use the Dynamic Embedded Topic Model to study business cycle narratives. The narratives present in central bank communication could either originate from the global economy or be created by the central banks themselves, which would align with the concept of *gradualism*. Gradualism is commonly used by central banks to make small adjustments to interest rates over time, rather than implementing abrupt changes. For instance, between 2001 and 2003, the Fed employed gradualism to reduce the interest rate by 550 basis points, with thirteen cuts in total.[3] Gradualism is founded on the theory of uncertainty in policy making, in which policymakers are inclined to gradually introduce a policy when its effect on the economy is ambiguous (Brainard, 1967). Similarly, central bank communication can introduce narratives that gradually prepare the public for future policy changes, such as changes in financial market regulation or an introduction of central bank digital currency (CBDC). At present, many of the central banks are investigating a potential introduction of digital money and many of them are actively communicating on the matter through speeches and reports

[2] While the concept of narrative is a relatively recent addition to economic discourse, it has been explored in other fields. For example, Narrative Psychology was pioneered by Sarbin (1986). In the humanities, the theory of narrative has been studied for several decades, with early works such as Barthes and Duisit (1975) and Bruner (1991) defining a narrative as "an account of events occurring over time". For a comprehensive overview of the subject, see Mitchell (1981).

[3] This was discussed by Mr Ben Bernanke, member of the Board of Governors of the US Federal Reserve System, at his speech *Gradualism* at an economics luncheon co-sponsored by the Federal Reserve Bank of San Francisco (Seattle Branch) and the University of Washington, Seattle, 20 May 2004.

(Bank for International Settlements, 2020).[4] Central banks have the option to employ verbal communication to inform and prepare the public for future structural changes, ensuring that monetary policy and regulatory changes are achieved as expected. Nevertheless, the extent to which central banks utilize this approach proactively remains undocumented.

Central banks primarily rely on two types of communication: (i) Written disclosures of meeting minutes and reports. (ii) Speeches and verbal communication. Since the 1990s, central bank communication strategies have undergone a transformation, going from opaque secrecy, to greater transparency, to active use of communication as a tool for monetary policy (Woodford, 2005; Blinder et al., 2008; Blinder, 2018). One notable change in central bank communication strategies is the significant increase in the number of speeches delivered. Consequently, there has been a substantial growth in text data pertaining to central banks, which can now be analyzed using machine learning techniques. Text data are multi-dimensional and contain a wealth of information. With the help of advances in computational linguistics, it is possible to reduce the dimensionality of these data and use it for economic analysis. In 1997, the number of speeches from central banks affiliated with the BIS was 119, but by 2019, this number had risen to 423, indicating a substantial surge in verbal communication. Compared to announcements, central bank speeches offer a richer and more extensive range of information, as they are significantly longer and greater in number.

Previous research has analyzed central bank speeches, with Jansen and Haan (2005) finding that speeches from the ECB affect the volatility of the euro-dollar exchange rate. Andersson, Dillén, and Sellin (2006) study speeches from the Swedish Riksbank and find that they affect market prices, particularly those delivered by the head of the Riksbank. Born, Ehrmann, and Fratzscher (2014) demonstrate that central bank speeches about financial stability can significantly affect market returns and volatility. Two studies similar to this one that use topic modeling are Hansen, McMahon, and Prat (2017) and Armelius et al. (2020). Hansen, McMahon, and Prat (2017) examine central bank transparency using topic modelling in an event study around 1993, the year the Fed began to release the FOMC meeting transcripts. Armelius et al. (2020) study spillover effects in sentiment from central bank speeches and show that cross-country effects affect both central bank communication and macroeconomic variables, with the Fed having a unique impact on creating sentiment spillover effects.

[4]The ECB, Bank of Japan, Sweden's Riksbank, Swiss National Bank, Bank of England and the Fed, are actively investigating and reporting on CBDCs.

Furthermore, when the interest rate is close to the efficient lower bound (ELB), central bank communication is of increased importance, and forward guidance may be the main policy tool (Blinder et al., 2008). At these times, the public's expectations of the central bank's future policy are crucial, which indicates that central bank communication might be weighted towards forward guidance. Yet, the results of this paper suggest that the content of central bank communication is broad, also at times when the interest rates are close to the ELB. This is consistent with previous research indicating that central bank communication may not sufficiently target or affect the general public (Kumar et al., 2015; Lamla and Vinogradov, 2019; Coibion et al., 2020), and the fact that trust in central banks is relatively low (Hayo and Neuenkirch, 2014; European Commission, 2019).[5] Blinder (2018) predicts that "central banks will keep trying to communicate with the general public, as they should, but for the most part, they will fail".

This study offers several contributions to the existing literature. Firstly, it provides a comprehensive and dynamic analysis of the evolution of central bank speech communication. Secondly, it explores the use of Dynamic Topic Models (Blei and Lafferty, 2006) in analyzing central bank communication. Thirdly, the study examines the autoregressive properties of the content of central bank speeches, which demonstrates strong persistence, even after controlling for underlying

[5]According to the Eurobarometer survey, public trust in the ECB is low (European Commission, 2019), but increases as communication from the ECB increases (Hayo and Neuenkirch, 2014).

financial variables. Finally, the paper draws connections between central bank communication, topic modelling, and narrative economics, suggesting ideas for applications of topic modelling in narrative economics research.

The rest of the paper is organized as follows: Section 2 provides a description of the data; Section 3 outlines the methodology used in this study; Section 4 presents the main findings, which includes a case study investigating the persistence of topics in speeches from the Federal Reserve while controlling for regional financial variables; and finally, Section 5 concludes the paper.

2 Data

The central bank speech data are downloaded from the BIS website.[6] 14,423 central bank speeches are collected from 113 institutions, over the time period 1997 through 2020. The speeches are in text format, meaning that they have been transcribed into English sentences (translated when necessary). Armelius et al. (2020) are the first to use the BIS data source, which has not been extensively studied in the literature despite its rich information.

To ensure a homogenous and sufficiently talkative global subsample, the dataset is limited to include only global institutions that have given more than 200 speeches over the sample period. This leads to exclusion of local central bank branches, for example the Bank of Spain, and less talkative central banks, such as the Central Bank of Brazil. Ultimately, the analysis is conducted on a dataset consisting of 7,379 speeches from 9 different central banks, including the Bank of Canada, the Bank of England, the Bank of Japan, the Central Bank of Norway, the ECB, the Fed (including speeches from the New York Fed), the Reserve Bank of Australia, Sweden's Riksbank, and the Swiss National Bank.

The preprocessing of the text data follows standard methodology (Gentzkow, Kelly, and Taddy, 2019). The data are first transformed from pdf format to text format.[7] Each document is split into lower case tokens (words), removing punctuation, numbers and web links. Headers and footers of the documents are removed, together with reference lists. A common list of stop words is applied to filter out words of little importance to the topic modelling. Through "lemmatization" the tokens are replaced by their dictionary form, for example *banks* becoming *bank*. Bigrams (sequences of two adjacent tokens) and trigrams (sequences of three adjacent tokens) are created to replace commonly followed tokens, such as *central bank* and *real interest rate*. To simplify interpretation and reduce dimensionality of the data, extreme tokens that appear in the corpus less than 20 times or in more than 50% of the documents are filtered out. [8] Table 3.1 shows the data dimensionality reduction at each step in the preprocessing. After preprocessing, the data consist of 4,280,706 tokens and the vocabulary (alias dictionary) of 20,697 unique tokens.[9]

Table 3.1: This table shows the data dimensionality reduction from each step in the preprocessing. The columns show how many total words and how many unique words are in the corpus after employing each step of the preprocessing. The final corpus consists of 4,280,706 total words and a vocabulary of 20,697 unique words.

	Raw text	Remove stopwords	Lemmatization	Bigrams and trigrams	Filter extremes
Total words	22,762,644	11,903,054	11,737,316	9,908,456	4,280,706
Unique words	66,735	58,478	53,116	65,282	20,697

For the Fed case study, control variables are obtained from the Wharton Research Data Service (WRDS), which comprise 1-year US treasury bond yields, US inflation, S&P 500 Index returns, and the CBOE Volatility Index (VIX). To down-sample the S&P 500 Index

[6]The data are scraped using the Request and Beautiful Soup Python libraries.

[7]The Textract Python library is used for this task.

[8]The degree of filtering is in this paper determined by a grid search over topic coherence (Newman et al., 2010), using the full sample and an LDA model. The evaluation is conducted using the topic coherence measures described in Röder, Both, and Hinneburg (2015).

[9]A kernel density estimation shows that the number of total words and number of unique words used in the speeches are smoothly distributed (not clustered) in the data.

data and the VIX data, the maximum values in each quarter are chosen to retain as much of the variance as possible.

3 Methodology

To analyze the central bank speech data, I utilize a combination of Dynamic Topic Models (DTM) (Blei and Lafferty, 2006) and autoregressive (AR) regressions. While DTM has been infrequently used in the field of finance and economics, it offers significant advantages by allowing topics to change dynamically over time. This means that the word distributions defining the topics are not fixed, which allows the researcher to study the time evolution of the latent topics discovered by the model. Additionally, using DTM enhances transparency of the topics, as one can determine whether a given topic pertains to the same subject throughout the sample period. This feature is especially important when using the estimated topics in time series modeling.

In the applied literature, Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) has emerged as the standard topic model, drawing from the foundations of Latent Semantic Indexing (LSI) first introduced by Deerwester et al. (1990) and later extended by Papadimitriou et al. (1997) and Hofmann (1999).[10] LDA has gained significant traction in the fields of finance and economics, with Gentzkow, Kelly, and Taddy (2019) providing an overview of its use in natural language processing for economic analysis.

Similar to the LDA model, DTM are a set of generative probabilistic models used for discrete data, particularly for text analysis. DTM incorporate a time component to the LDA framework, allowing topics to evolve over time. DTM are in the category of unsupervised ma-

[10] Further advancements have been made to the LDA algorithm. The Hierarchical Dirichlet Process (HDP) considers a hierarchical topic structure to determine the number of topics (Teh et al., 2004), while Correlated Topic Models (CTM) (Blei and Lafferty, 2007) assume a correlation between topics. Additionally, computer software has been developed to simplify natural language processing and topic modeling for applied researchers (McCallum, 2002; Bird, Loper, and Ewan, 2009; Řehůřek and Sojka, 2010).

chine learning models and utilize a "bag-of-words" structure, meaning that the order of the words does not matter. However, in contrast to the static LDA model, the order of the documents does matter. The dynamic model addresses the static assumption of the LDA model by updating the parameters of the distributions for each time slice, which in this paper is done annually. This is done by introducing a state space model using a logistic normal distribution. In this framework, each document is generated from a mixture of topics and each topic is generated from a mixture of words from the vocabulary.

Given a model with K topics, D documents, and a vocabulary with V terms, let $\beta_{t,k}$ be a V-dimensional vector representing topic k at time t, where $t = 1, \ldots, T$ and $k = 1, \ldots, K$. $\beta_{t,k}$ evolves with a Gaussian random walk, $\beta_{k,t}|\beta_{k,t-1} \sim \mathcal{N}(\beta_{k,t-1}, \sigma^2 I)$, meaning that the word distribution over topics change over time. Furthermore, let α_t be a D-dimensional mean parameter vector of the logistic normal distribution for the topic proportions, following a Gaussian random walk, $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$. The generative process for a sequential corpus at time slice t involves chaining together a set of topic models in the following manner:

1. Draw topics distributions over dictionary $\beta_{k,t}|\beta_{k,t-1} \sim \mathcal{N}(\beta_{k,t-1}, \sigma^2 I)$.
2. Draw mean parameters of document distributions over topics $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$.
   (b) For each word position $n \in N_d$:
      (i) Draw topic $Z \sim Mult(\pi(\eta))$.
      (ii) Draw word $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

Here $\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum \exp(\beta_{k,t,w})}$ maps the multinomial natural parameters to the mean parameters.[11] From a practical point of view, one

[11] Note that the process is similar to that of LDA (Blei, Ng, and Jordan, 2003).

does not generate the corpus, but rather backs out the underlying latent distributions, given a corpus, with variational Bayesian inference.[12] The initialization of the parameters is done by first estimating a static LDA model. The variance of the Gaussian random walks is set to a fixed value of 0.005 in the implementation used in this paper, as described in the computer code associated with Blei and Lafferty (2006) and Řehůřek and Sojka (2010). It is common in the literature to not estimate these hyperparameters, but this approach is considered a limitation of the methodology, as it restricts the extent to which topics can change over time.

In a practical sense, the model can be understood to yield two kinds of results. Firstly, the model generates a set of K V-dimensional topic distributions for each time slice t, which are functions of the vocabulary and define the K estimated topics in the model. Each word in the vocabulary at each time slice is assigned a probability in a topic's probability distribution, indicating how likely that word is to be drawn from that topic at time t. Consequently, the most probable words in a topic's distribution represent the theme of that topic, and the topic names are manually assigned based on these themes. By monitoring the changes in a topic's distribution across time it is possible to study how the topic evolves.

Secondly, using a trained (estimated) model each document (central bank speech) in the corpus can be assigned a static K-dimensional distribution over topics. For each topic, a probability is assigned, resulting in a set of probabilities that describe how likely the document is to have been generated from each of the topics. By classifying the documents and averaging the distributions on a monthly or quarterly basis, it is possible to track the evolution of topics discussed in the

However in LDA, the topics and the topic proportions would be sampled from the static Dirichlet distribution, which is the conjugate prior to the Categorical distribution. This distribution is a generalization of the Bernoulli distribution or a special case of the Multinomial distribution (one draw instead of many), which simplifies the estimation process and allows for efficient use of Gibbs sampling. For a practical overview of LDA, see Griffiths and Steyvers (2004).

[12] Blei and Lafferty (2006) also discuss Variational Kalman Filtering and Variational Wavelet Regression.

documents over time. Thus, the model's output enables the study of both the evolution of topics discussed throughout the sample period, and the terminology-evolution of the topics themselves.

The progression of topics discussed and the within topic changes are both functions of the underlying central bank speech data (corpus). When significant global events occur in financial markets, central bank speeches are expected to discuss these events and to include contemporary relevant language. Therefore economic events drive topics and current topics have a higher probability of being addressed. However, the emerging field of narrative economics (Shiller, 2017) suggests that there are other factors (narratives) that affect how topics develop and spread. Narrative economics provide an additional explanatory theory for any unexplained topic persistence in the model. Narratives within the global economy, as well as narratives created by central banks, can contribute to the probability of certain words appearing in topic distributions at each time interval, and the probability of certain topics being discussed at each time interval.

Furthermore, selecting the appropriate number of topics, K, is a non-trivial and extensively researched area. In the DTM framework, the number of topics is assumed to be known and must therefore be specified before estimation. A common approach for determining the number of topics is to estimate several models with varying numbers of topics and evaluate them based on a particular metric. A typical method is to evaluate the model according to its *topic coherence*, introduced by Newman et al. (2010). To be comprehensible to humans, the words that contribute the most to a topic's distribution must be semantically related, and this can be quantified using various coherence metrics. For this study, the number of topics was chosen via a grid search that assessed the coherence metrics implemented by Röder, Both, and Hinneburg (2015) using an LDA model applied to the entire corpus, combined with manual analysis.

Common problems in unsupervised topic modelling are residual topics without a clear meaning and topics that are indistinguishable from each other in a meaningful way. Some researchers address these issues by opting for a large number of topics, such as 100, and disregarding irrelevant ones. Alternatively, similar topics can be clustered and manually classified as one topic, but these approaches may result in overfitting and ambiguity. In this paper, however, I demonstrate that by preprocessing the central bank speech data appropriately, meaningful and distinctive topics can be achieved without the need for a significant increase in the number of clusters or human intervention.

4 Results

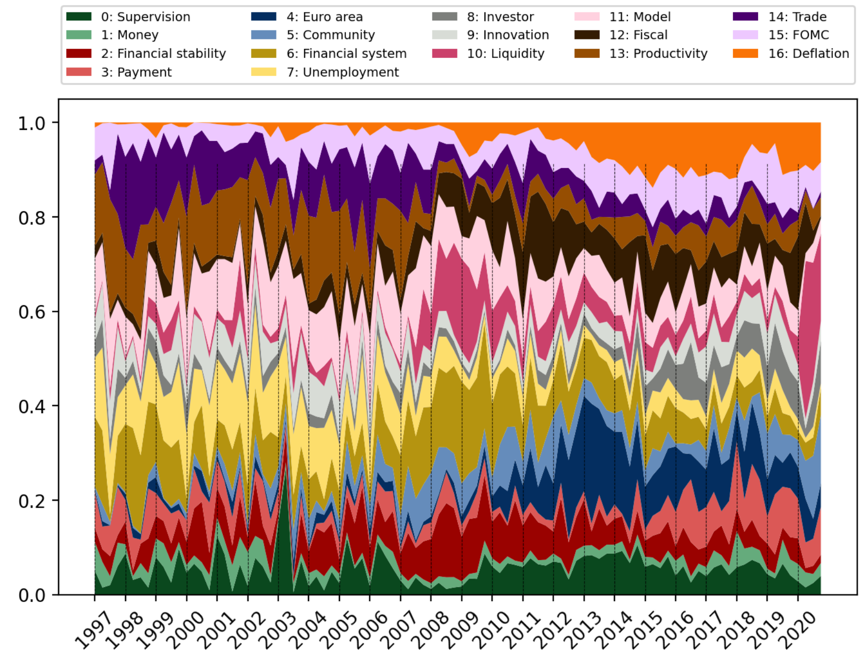
4.1 Central bank topics

Optimizing the number of topics with respect to coherence, yield a total of 29 topics. The resulting topics are then categorized into 12 local and 17 global topics, based on the dominant themes of the most probable words in the estimated probability distributions. The local topics are region specific, as indicated by the high probability of words being associated with the corresponding regions. For example, the local topic related to the Swedish economy include top words such as *Sweden*, *Riksbank*, and *Swedish*, and words relevant to the Swedish economy, such as *Krona* and *Repo rate*. Moreover, local topics have a low probability of being discussed in international settings. For instance, the Swedish topic is primarily mentioned by the Swedish Riksbank.

Figure 3.1 presents the average normalized probability distributions for the global topics during each quarter in the sample period. These probabilities are derived by classifying each central bank speech in the corpus by the estimated topic model, which reflects the probability of the speech addressing each topic. The quarterly averages are obtained by averaging the topic distributions across speech time. It is important to note that the standard deviation of the Gaussian random walk, which models within-topic persistence, does not impact which topics the central banks address over time.

Figure 3.1 illustrates that the model successfully identifies 17 global topics that persist throughout the sample period, providing

Figure 3.1: This figure shows the average normalized global topic probabilities for each quarter. The mass of each line, represents the probability that the topic is addressed in that time period. These probabilities are obtained by classifying each central bank speech in the corpus by the estimated topics, and calculating the average quarterly distribution. Figures 3.4a through 3.4q, in Appendix 5, display each topic in this figure as a line graph.



insight into the subjects that central banks commonly discuss. While the topics are predominantly related to central bank matters, such as monetary policy, the analysis reveals that they encompass a diverse range of themes, indicating that central bank communication does not solely revolve around coordinating financial market expectations or disclosing private information to the public. This finding challenges the conventional belief that central bank communication is primarily aimed at coordinating financial market expectations.

Throughout the sample period, central bank communication has exhibited a notable degree of diversity, covering a wide range of global topics, such as payment systems, small business communities, innovation and technology, economic modelling, and productivity. This pattern is consistent even during periods where global interest rates are near the effective lower bound (ELB). It could be argued that these topics, in the long run, help central banks to communicate their monetary policy objectives effectively. By informing the public about matters such as payment systems or innovation and technology, central banks can construct narratives and prepare the economy for future changes in monetary policy paradigms. Consequently, central bank communication can be regarded as indirectly connected to the conventional definition of central bank communication.

Certain topics exhibit seemingly constant probability mass over the sample period, such as the first topic related to supervision and regulation (topic 0). In contrast, other topics demonstrate distinct trends. Following the global financial crisis in 2008, there is a clustering of topics relating to the financial system, financial stability, and liquidity. Conversely, some topics display steady upward trends. For instance, the topic of payment systems exhibits a recent positive trend, reflecting the contemporary technological advancements in this area. Many countries are witnessing significant growth in digital transactions, and several central banks are exploring the potential of central bank digital currencies.

4.2 Within topic word-distribution

Compared to static topic modeling techniques like LDA, Dynamic Topic Models offer the advantage of being able to track topics over time. Table 3.2 presents an example of this capability by displaying the evolution of the topic related to supervision and regulation on a yearly basis. The table shows the words with the highest probability of belonging to the topic throughout the sample period. It can be observed that the topic remains relatively consistent over time, suggesting that it pertains exclusively to supervision and regulation both in the beginning and end of the sample period. This is important since it allows for more precise conclusions to be drawn from

subsequent econometric analysis that employs the estimated topics. If one determines that central banks consistently discuss a topic over time, it is crucial to determine if the topic's content varies throughout time.

Although a topic may remain consistent in terms of its theme throughout the sample, the vocabulary associated with it can change over time, resulting in an increase or decrease in the probability of certain words. For instance, in Table 3.2, the term *Regulation* has a higher probability in the end of the sample period, while, the term *Capital requirement* remains a top word both in 1997 and later years. It should be noted that the model incorporates a degree of persistence in the word-probability distribution of topics. The standard deviation in the model regulates the rate at which topics can change annually. The lower the standard deviation, the less the variation in topics over time. A standard deviation of zero renders the model static.

Table 3.2: This table presents the year-by-year evolution of the within-topic probability distribution associated with the topic about supervision and regulation. The distributions for each year are arranged in descending order, listing the terms with the highest probabilities first (probabilities enclosed in parentheses).

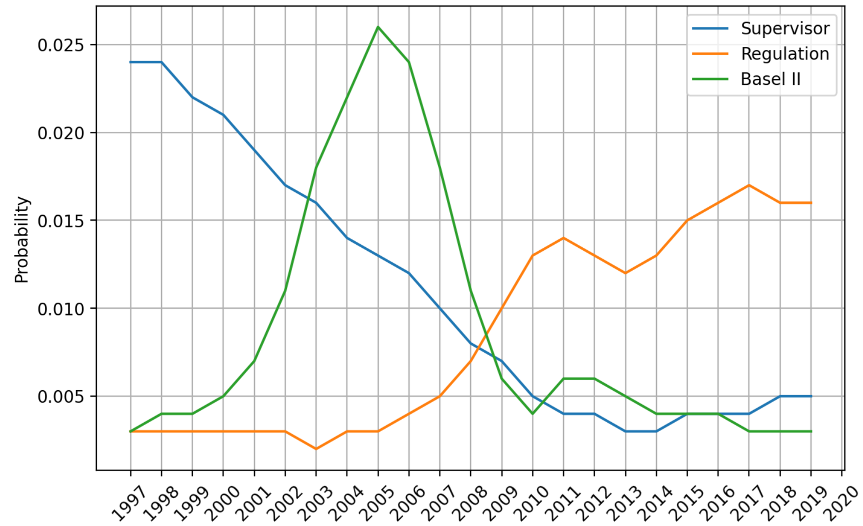
1997	1998	1999
Supervisor (0.024)	Supervisor (0.024)	Supervisor (0.022)
Standard (0.021)	Standard (0.02)	Standard (0.017)
Approach (0.02)	Approach (0.018)	Approach (0.016)
Supervisory (0.009)	Supervisory (0.009)	Supervisory (0.008)
Internal (0.008)	Internal (0.008)	Market Discipline (0.008)
Institution (0.008)	Institution (0.008)	Institution (0.007)
Risk Management (0.007)	Market Discipline (0.008)	Internal (0.007)
Market Discipline (0.007)	Risk Management (0.007)	Risk Management (0.007)
Capital Requirement (0.006)	Exposure (0.006)	Exposure (0.007)
Exposure (0.006)	Proposal (0.006)	Proposal (0.006)

2018	2019	2020
Regulation (0.016)	Regulation (0.016)	Regulation (0.016)
Requirement (0.009)	Stress Test (0.01)	Stress Test (0.01)
Stress Test (0.009)	Approach (0.009)	Approach (0.01)
Approach (0.008)	Requirement (0.009)	Requirement (0.009)
Capital Requirement (0.008)	Capital Requirement (0.008)	Rule (0.008)
Rule (0.007)	Rule (0.007)	Capital Requirement (0.008)
Regulatory (0.007)	Regulatory (0.007)	Regulatory (0.008)
Regime (0.006)	Framework (0.007)	Framework (0.007)
Standard (0.006)	Regime (0.006)	Regime (0.006)
Framework (0.005)	Stress Testing (0.006)	Stress Testing (0.006)

In Figure 3.2, the probability of the tokens *Basel II*, *Regulation*, and *Supervisor* from the topic about supervision and regulation is illustrated across the sample period. The graph indicated a shift in language from using the term *Supervisor* to *Regulation*, with the curves intersecting around 2008. This is consistent with the implementation of new regulations in the global financial industry due to the global financial crisis. Additionally, the token *Basel II* exhibits a bell-shaped curve in the figure, with its highest probability occurring shortly after the publication of the Basel II Accord in June 2004. This bell-shaped pattern aligns with the epidemiology of narratives theory discussed in Shiller (2017), where a narrative begins, develops, peaks, and eventually declines.

Gentzkow, Kelly, and Taddy (2019) stress the significance of human cross-checking in cases where the outcomes of natural language processing are utilized for descriptive or statistical analyses rather than solely for prediction purposes. Auditing a subsample of 20-30 documents can provide insight into whether the model accurately captures the relevant information in the corpus. In topic modeling, it is crucial to ensure that the topics proficiently explain the documents they generate. While documents are produced from a mixture of topics, some documents are likely to be generated solely from one topic, making them ideal candidates for manual examination. The speech “Implementing Basel II – choices and challenges” by Ms Susan Schmidt Bies at the Fed has the highest probability (99%) of being generated from the topic about supervision and regulation. This can be verified by reading the speech,

Figure 3.2: This figure displays the average probabilities of the tokens *Basel II*, *Supervisor*, and *Regulation*, associated with the topic about supervision and regulation.



In my remarks, I will focus primarily on the choices and challenges associated with Basel II implementation. In particular, I want to reaffirm the Federal Reserve’s commitment to Basel II and the need for continual evolution in risk measurement and management at our largest banks and then discuss a few key aspects of Basel II implementation in the United States. Given the international audience here today, I also plan to offer some thoughts on cross-border implementation issues associated with Basel II, including so-called home-host issues. (Ms Susan Schmidt Bies, Member of the Board of Governors of the US Federal Reserve System, at the Global Association of Risk Professionals’ Basel II and Banking Regulation Reform, Barcelona, 16 May 2006.)

Upon manual validation of a small subset of the central bank speeches in the corpus, it can be concluded that the reduction of dimensionality to topic space accurately captures the content of the documents.

4.3 Persistence in topics

Table 3.3: Estimated coefficients from the AR(1) models, using quarterly data. Each equation is represented as a column. The t-statistics (reported in parenthesis) are based on Newey and West (1987) standard errors with 1 lag.

	(1) Regulation	(2) Money	(3) Financial stability	(4) Payment	(5) Euro area	(6) Community	(7) Financial system	(8) Unemployment	(9) Investor
1 Lag	0.228	0.385***	0.468***	0.587***	0.892***	0.644***	0.767***	0.719***	0.693***
	(1.85)	(4.77)	(4.21)	(7.33)	(15.51)	(5.47)	(10.25)	(9.47)	(6.58)
Constants	0.0230***	0.00756***	0.0188***	0.0109***	0.00387*	0.00985***	0.0113**	0.00964**	0.00626***
	(5.85)	(6.06)	(5.23)	(4.99)	(2.59)	(3.56)	(3.30)	(3.40)	(3.44)
N	95	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(10) Innovation	(11) Liquidity	(12) Model	(13) Fiscal	(14) Productivity	(15) Trade	(16) FOMC	(17) Deflation
1 Lag	0.525***	0.770***	0.180	0.640***	0.579***	0.638***	0.380**	0.926***
	(6.25)	(15.56)	(1.76)	(7.03)	(6.28)	(7.16)	(3.33)	(23.67)
Constants	0.0125***	0.00789***	0.0387***	0.0126***	0.0175***	0.0133***	0.0206***	0.00234*
	(5.47)	(3.54)	(7.07)	(3.76)	(4.48)	(3.85)	(5.59)	(2.50)
N	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Next, I estimate a set of autoregressive (AR) regressions to examine the degree of persistence in the central bank speeches classified by the global topics. This indicates if the central banks tend to discuss particular topics persistently. The analysis relies on quarterly and monthly data, consistent with the approaches taken by Armelius et al. (2020) and Nyman et al. (2018). Specifically, the following AR(1) model is estimated for each global topic,

$$\theta_{k,t} = \alpha_k + \varphi_k \theta_{k,t-1} + \boldsymbol{\beta}_k \boldsymbol{X}_t + \boldsymbol{\gamma}_k \boldsymbol{X}_{t-1} + \epsilon_{k,t}. \tag{3.1}$$

Table 3.4: Estimated coefficients from the AR(1) models, using monthly data. Each equation is represented as a column. The t-statistics (reported in parenthesis) are based on Newey and West (1987) standard errors with 1 lag.

	(1) Regulation	(2) Money	(3) Financial stability	(4) Payment	(5) Euro area	(6) Community	(7) Financial system	(8) Unemployment	(9) Investor
1 Lag	0.118	0.222**	0.286***	0.363***	0.716***	0.258***	0.473***	0.426***	0.274***
	(1.82)	(2.69)	(3.78)	(5.27)	(12.62)	(3.88)	(7.65)	(7.75)	(3.57)
Constants	0.0249***	0.00962***	0.0247***	0.0165***	0.00923***	0.0191***	0.0257***	0.0209***	0.0139***
	(10.13)	(9.06)	(9.46)	(8.51)	(5.06)	(8.08)	(8.16)	(8.59)	(8.29)
N	287	287	287	287	287	287	287	287	287

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(10) Innovation	(11) Liquidity	(12) Model	(13) Fiscal	(14) Productivity	(15) Trade	(16) FOMC	(17) Deflation
1 Lag	0.181**	0.731***	0.140*	0.430***	0.380***	0.417***	0.131	0.669***
	(3.26)	(14.22)	(2.09)	(6.41)	(6.16)	(5.94)	(1.90)	(12.26)
Constants	0.0210***	0.00842***	0.0397***	0.0201***	0.0265***	0.0214***	0.0288***	0.00835***
	(12.00)	(5.93)	(11.88)	(7.51)	(8.64)	(8.20)	(10.44)	(5.59)
N	287	287	287	287	287	287	287	287

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Here, $\theta_{k,t}$ is the average probability (from the classified speeches) for the global topic k at time t, where $k = 1, \ldots, K$ and $t = 1, \ldots, T$. $\boldsymbol{X}_t$ is a vector of region specific control variables at time t, which is employed in the Fed case study detailed in Section 4.4.

Tables 3.3 and 3.4 report the estimated coefficients from the AR(1) models represented by Equation 3.1 (without regional control variables), using aggregated quarterly and monthly observations for each topic respectively. The tables show strong significant autoregressive effects in the majority of the topics.[13] The findings suggest that there is substantial persistence in the global topics, both on a quarterly and monthly basis, indicating that central banks tend to continue talking about a particular topic once they have begun to do so. One possible explanation for this is that the underlying macroeconomic variables reflected in the topics are themselves persistent, which will be con-

[13] The results are robust to estimation in a VAR system, where the autoregressive effects dominate the cross-sectional effects.

trolled for in the second part of the analysis in Section 4.4. Another explanation is that narratives may drive the persistence in the topics, either as part of the global economy or of narratives set by the central banks themselves. However, a few topics do not exhibit significant autoregressive effects. For instance, the topic related to supervision and regulation does not show any significant persistence, implying that this topic is discussed sporadically throughout the sample without any particular trends.

4.4 FED case study

In this section, I estimate Equation 3.1 (as presented in Section 4.3) using only the Fed speech data, first without any control variables and then with regional financial control variables. Central banks tend to address the current economic environment in their communication, and thus underlying financial and macroeconomic variables are important determinants of central bank communication topics. Central banks globally display heterogeneity in the topics they address, as well as what economic variables may affect their communication. The Fed is more likely to discuss topics related to US inflation and the US stock market than topics related to European macroeconomic and financial conditions. By focusing on speeches from one central bank (the Fed in this case), I can control for the regional variables associated with that bank's speeches.

Figure 3.3 illustrates the relationship between the probability that the Fed addresses the topic about the financial system, and three financial control variables: the CBOE Volatility Index (VIX), the S&P 500 Index returns, and 1-year US treasury returns. The figure suggests potential co-movements between these variables. Therefore, it is possible that by controlling for the financial variables discussed in the topic about the financial system, the autoregressive feature of the topic might be fully explained. If the persistence is entirely accounted for by the control variables, it would suggest that other factors, such as global or local central bank narratives, are not driving the observed persistence. However, if the topic continues to exhibit persistence, it is conceivable that a "narrative effect" is one driving factor.

Figure 3.3: This figure shows the average quarterly probability (normalized) that the Fed addresses the topic about the *Financial system*. The topic is visualized together with the normalized control variables: CBOE Volatility Index (VIX), S&P 500 Index returns, 1-year US treasury bond returns.

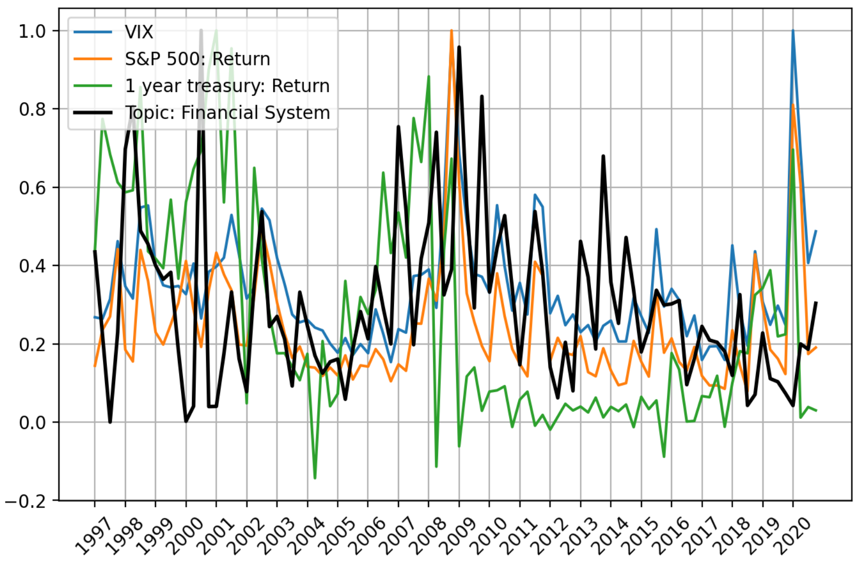


Table 3.5 reports the autoregressive coefficients of the average topic probabilities, based on 1,841 speeches from the Fed and the New York Fed, covering the period from 1997 to 2020. In this table, no control variables are included in the analysis, and only quarterly data are used, as some months in the sample period have no Fed speeches. The speeches are classified using the DTM model in Section 4. The model is not re-estimated but trained on the full corpus, including speeches from the other 8 central banks analyzed in this paper. As a result, the model may be less descriptive of the Fed data, which could be a possible reason for the slightly weaker results compared to Table

3.3, in Section 4.[14]

Table 3.6 presents the autoregressive coefficients of the global topics discussed in the Fed speeches, with the inclusion of regional control variables. The addition of control variables weakens the autoregressive properties in the global topics, but does not entirely account for them. The persistent topics exhibit smaller coefficients and lower significance with the inclusion of controls, as compared to the estimation without controls. However, most of the topics that show persistence without controls also show persistence with the controls. Only two exceptions are observed, the persistence in the topics related to the financial system and trade appears to be entirely explained by the control variables.

These findings indicate that there are additional factors beside the underlying financial variables contributing to the persistence of topics. Although further investigation is needed, these results align with the theory of narrative economics, suggesting that communication at the topic level is *story-based*. Narrative-based communication is more likely to spread in various forms such as conversations, news, and social media (Shiller, 2017), and being communicated by a central bank, it becomes more accessible for the general public (Haldane and McMahon, 2018).

Table 3.5: Estimated coefficients from the AR(1) models, using quarterly data from the Fed. Each equation is represented as a column. The t-statistics (reported in parenthesis) are based on Newey and West (1987) standard errors with 1 lag.

	(1) Regulation	(2) Money	(3) Financial stability	(4) Payment	(5) Euro area	(6) Community	(7) Financial system	(8) Unemployment	(9) Investor
1 Lag	0.121	0.163	0.236	0.374**	0.685***	0.368*	0.280*	0.430***	0.400***
	(1.04)	(1.44)	(1.91)	(3.07)	(7.68)	(2.26)	(2.25)	(5.71)	(4.65)
Constants	0.0267***	0.0100***	0.0297***	0.0174***	0.0104***	0.0212***	0.0353***	0.0188***	0.0120***
	(6.35)	(6.36)	(5.96)	(4.61)	(3.71)	(4.34)	(4.97)	(6.07)	(4.60)
N	95	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[14]Table 3.11 in Appendix 5 shows that the autoregressive results are more robust when using topic probabilities from a DTM model that is trained solely on the Fed corpus.

	(10) Innovation	(11) Liquidity	(12) Model	(13) Fiscal	(14) Productivity	(15) Trade	(16) FOMC	(17) Deflation
1 Lag	0.151	0.690***	0.0870	0.468***	0.478***	0.342***	0.139	0.564***
	(1.40)	(6.88)	(0.73)	(4.11)	(4.40)	(4.38)	(1.73)	(5.18)
Constants	0.0246***	0.0105***	0.0435***	0.0192***	0.0216***	0.0255***	0.0315***	0.0117***
	(6.31)	(3.90)	(6.92)	(3.95)	(4.92)	(6.14)	(7.09)	(4.09)
N	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.6: Estimated coefficients from the AR(1) models with control variables, using quarterly data from the Fed. Each equation is represented as a column. The t-statistics (reported in parenthesis) are based on Newey and West (1987) standard errors with 1 lag.

	(1) Regulation	(2) Money	(3) Financial stability	(4) Payment	(5) Euro area	(6) Community	(7) Financial system	(8) Unemployment	(9) Investor
1 Lag	0.0211	0.170	0.174	0.381**	0.583***	0.309	0.262	0.270*	0.367***
	(0.18)	(1.37)	(1.60)	(2.81)	(5.48)	(1.93)	(1.88)	(2.15)	(3.79)
Constants	0.0497***	0.0112**	0.0258	0.0191*	0.0332*	0.0122	0.0208	0.0192**	0.0201*
	(4.49)	(2.77)	(1.58)	(2.01)	(2.35)	(0.83)	(1.82)	(2.85)	(2.46)
N	95	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(10) Innovation	(11) Liquidity	(12) Model	(13) Fiscal	(14) Productivity	(15) Trade	(16) FOMC	(17) Deflation
1 Lag	0.0928	0.542***	-0.0121	0.334*	0.421***	0.127	0.114	0.479***
	(0.79)	(4.51)	(-0.10)	(2.25)	(4.06)	(0.98)	(1.31)	(3.94)
Constants	0.0401***	-0.0118	0.0572***	0.0183	0.0405**	0.0310**	0.0472***	0.0206
	(3.95)	(-0.74)	(4.37)	(1.43)	(3.36)	(3.03)	(3.60)	(1.84)
N	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5 Conclusion

The empirical findings show that central banks talk about a wide range of global topics, not all immediately related to the traditional theory of central bank communication. However, with a broader set of topics, central banks can reveal private information and prepare society for long term monetary policy shifts and structural changes. Topic trends occur and vocabulary changes over time, but most topics have significant probability mass throughout the sample period, even at times when the interest rate is close to the ELB.

Furthermore, the topics are well captured by Dynamic Topic Models. Both in terms of quantitative measures, such as coherence scores, as well as manual investigation linking the topics to the representative documents. Thus, the dimension reduction of the corpus to topic space is able to, in a meaningful way, capture the relevant central bank communication. This encourages the use of topic modelling, and more specifically DTM, in other social science applications with similar data.

Topic modelling has an interesting application in estimating narratives. The observed topic persistence is consistent with the theory of narrative economics and proposes that the central bank communication on the topic level might be story-based. The evolution of word-probabilities within the topics is also consistent with the epidemiology models of narrative economics.

References

Amato, Jeffery D., Stephen Morris, and Hyun Song Shin (Dec. 2002). "Communication and Monetary Policy". In: *Oxford Review of Economic Policy* 18.4, pp. 495–503 (cit. on p. 130).

Andersson, Malin, Hans Dillén, and Peter Sellin (2006). "Monetary policy signaling and movements in the term structure of interest rates". In: *Journal of Monetary Economics* 53.8, pp. 1815–1855 (cit. on p. 132).

Armelius, Hanna, Christoph Bertsch, Isaiah Hull, and Xin Zhang (2020). "Spread the Word: International spillovers from central bank communication". In: *Journal of International Money and Finance* 103, p. 102116 (cit. on pp. 133, 134, 146).

Bank for International Settlements (Oct. 2020). *Central bank digital currencies: foundational principles and core features*. Tech. rep. (cit. on p. 132).

Barthes, Roland and Lionel Duisit (1975). "An Introduction to the Structural Analysis of Narrative". In: *New Literary History* 6.2, pp. 237–272 (cit. on p. 131).

Bertsch, Christoph, Isaiah Hull, and Xin Zhang (2021). "Narrative fragmentation and the business cycle". In: *Economics Letters* 201, p. 109783 (cit. on p. 131).

Bird, Steven, Edward Loper, and Klein Ewan (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. (cit. on p. 136).

Blei, David and John Lafferty (2006). "Dynamic Topic Models". In: ICML '06, pp. 113–120 (cit. on pp. 130, 133, 136, 138).

— (Aug. 2007). "A correlated topic model of Science". In: *The Annals of Applied Statistics* 1 (cit. on p. 136).

Blei, David, Andrew Ng, and Michael Jordan (Mar. 2003). "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022 (cit. on pp. 136, 137).

Blinder, Alan S. (2018). "Through a Crystal Ball Darkly: The Future of Monetary Policy Communication". In: *AEA Papers and Proceedings* 108, pp. 567–571 (cit. on pp. 132, 133).

Blinder, Alan S., Michael Ehrmann, Marcel Fratzscher, Jakob de Haan, and David-Jan Jansen (2008). "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence". In: *Journal of Economic Literature* 46.4, pp. 910–945 (cit. on pp. 130, 132, 133).

Born, Benjamin, Michael Ehrmann, and Marcel Fratzscher (2014). "Central Bank Communication on Financial Stability". In: *The Economic Journal* 124.577, pp. 701–734 (cit. on p. 132).

Brainard, William C. (1967). "Uncertainty and the Effectiveness of Policy". In: *The American Economic Review* 57.2, pp. 411–425 (cit. on p. 131).

Bruner, Jerome (1991). "The Narrative Construction of Reality". In: *Critical Inquiry* 18.1, pp. 1–21 (cit. on p. 131).

Cochrane, John H. and Monika Piazzesi (2002). "The Fed and Interest Rates: A High-Frequency Identification". In: *The American Economic Review* 92.2, pp. 90–95 (cit. on p. 130).

Coibion, Olivier, Yuriy Gorodnichenko, Saten Kumar, and Mathieu Pedemonte (2020). "Inflation expectations as a policy tool?" In: *Journal of International Economics* 124. NBER International Seminar on Macroeconomics 2019, p. 103297 (cit. on p. 133).

Cook, Timothy and Thomas Hahn (1989). "The effect of changes in the federal funds rate target on market interest rates in the 1970s". In: *Journal of Monetary Economics* 24.3, pp. 331–351 (cit. on p. 130).

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (Sept. 1990). "Indexing by Latent Semantic Analysis". In: *Journal of the American Society for Information Science (1986-1998)* 41.6, p. 391 (cit. on p. 136).

Ellen, Saskia ter, Vegard H. Larsen, and Leif Anders Thorsrud (2019). *Narrative monetary policy surprises and the media*. Working Paper (cit. on p. 131).

European Commission (Nov. 2019). *Standard Eurobarometer (EB 92) – Public opinions in the European Union*. Tech. rep. (cit. on p. 133).

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". eng. In: *Journal of Economic Literature* 57.3, pp. 535–574 (cit. on pp. 134, 136, 144).

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.1, pp. 5228–5235 (cit. on p. 138).

Haldane, Andrew and Michael McMahon (May 2018). "Central Bank Communications and the General Public". In: *AEA Papers and Proceedings* 108, pp. 578–83 (cit. on p. 150).

Hansen, Stephen, Michael McMahon, and Andrea Prat (Oct. 2017). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*". In: *The Quarterly Journal of Economics* 133.2, pp. 801–870 (cit. on p. 133).

Hansen, Stephen, Michael McMahon, and Matthew Tong (2019). "The long-run information effect of central bank communication". In: *Journal of Monetary Economics* 108, pp. 185–202 (cit. on p. 131).

Hayo, Bernd and Edith Neuenkirch (2014). "The German public and its trust in the ECB: The role of knowledge and information search". In: *Journal of International Money and Finance* 47, pp. 286–303 (cit. on p. 133).

Hofmann, Thomas (1999). "Probabilistic latent semantic indexing". In: *SIGIR '99* (cit. on p. 136).

Jansen, David-Jan and Jakob De Haan (2005). "Talking heads: the effects of ECB statements on the euro–dollar exchange rate". In: *Journal of International Money and Finance* 24.2. Exchange Rate Economics, pp. 343–361 (cit. on p. 132).

Kumar, Saten, Olivier Coibion, Hassan Afrouzi, and Yuriy Gorodnichenko (2015). "Inflation Targeting Does Not Anchor Inflation

Expectations: Evidence from Firms in New Zealand". In: *Brookings Papers on Economic Activity*, pp. 151–208 (cit. on p. 133).

Lamla, Michael J. and Dmitri V. Vinogradov (2019). "Central bank announcements: Big news for little people?" In: *Journal of Monetary Economics* 108, pp. 21–38 (cit. on p. 133).

McCallum, Andrew Kachites (2002). "MALLET: A Machine Learning for Language Toolkit" (cit. on p. 136).

Mitchell, W. J. T. (1981). *On narrative*. Phoenix book. University of Chicago Press (cit. on p. 131).

Nakamura, Emi and Jón Steinsson (Jan. 2018). "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect*". In: *The Quarterly Journal of Economics* 133.3, pp. 1283–1330 (cit. on p. 130).

Newey, Whitney K. and Kenneth D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix". In: *Econometrica* 55, p. 703 (cit. on pp. 146, 147, 150, 151, 161, 163).

Newman, David, Jey Lau, Karl Grieser, and Timothy Baldwin (Jan. 2010). "Automatic Evaluation of Topic Coherence." In: pp. 100–108 (cit. on pp. 135, 139).

Nyman, Rickard, Sujit Kapadia, David Tuckett, David Gregory, Paul Ormerod, and Robert Smith (Jan. 2018). "News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment". In: *SSRN Electronic Journal* (cit. on pp. 131, 146).

Papadimitriou, Christos H., Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala (1997). *Latent Semantic Indexing: A Probabilistic Analysis* (cit. on p. 136).

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50 (cit. on pp. 136, 138).

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. Shanghai, China, pp. 399–408 (cit. on pp. 135, 139).

Sarbin, Theodore R (1986). *Narrative psychology: The storied nature of human conduct.* Praeger Publishers/Greenwood Publishing Group (cit. on p. 131).

Shiller, Robert J. (Apr. 2017). "Narrative Economics". In: *American Economic Review* 107.4, pp. 967–1004 (cit. on pp. 131, 139, 144, 150).

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei (2004). "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes". In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. NIPS'04. Vancouver, British Columbia, Canada: MIT Press, pp. 1385–1392 (cit. on p. 136).

Woodford, Michael (2001). "Monetary policy in the information economy". In: *Proceedings - Economic Policy Symposium - Jackson Hole*, pp. 297–370 (cit. on p. 130).

— (Dec. 2005). *Central Bank Communication and Policy Effectiveness*. Working Paper 11898. National Bureau of Economic Research (cit. on p. 132).

Appendices

Results

Table 3.7: This table shows the full set of estimated topics from the DTM. The topics are divided into local topics and global topics. The probabilities that each topic is addressed in the corpus are shown in the second column.

Topics	Average P (%) in corpus
0: Supervision	3.17
1: Money	1.61
2: Financial stability	7.25
3: Payment	2.77
4: Euro area	2.63
5: Community	3.3
6: Financial system	2.56
7: Unemployment	1.61
8: Investor	5.33
9: Innovation	2.63
10: Liquidity	1.93
11: Model	4.37
12: Fiscal	5.51
13: Productivity	3.88
14: Trade	4.37
15: FOMC	2.14
16: Deflation	2.55
Sum global:	57.61
17: Australia	2.97
18: Canada	7.02
19: ECB	3.01
20: Euro area	4.54
21: Federal reserve	4.67
22: Japan economy	1.76
23: Japan	5.29
24: Norway	2.12
25: Riksbank	2.89
26: Residual	4.08
27: SNB	2.54
28: UK	1.53
Sum local:	42.39
Mean	3.45
Sum	100.0

Table 3.8: This table shows the number of speeches held by the central banks each year.

	1997	1998	1999	2000	2001	2002	2003	2004	2005
CAN	7	6	9	9	12	18	18	18	24
ENG	13	15	12	22	21	14	7	10	12
JPN	21	16	15	22	16	21	28	15	20
FED	52	43	42	39	48	55	55	97	86
NOR	0	0	6	7	10	19	14	13	20
ECB	0	8	46	45	36	40	31	56	51
NYC	5	7	5	7	2	4	3	8	10
AUS	9	12	9	12	8	12	11	9	13
SWE	11	13	28	31	29	29	21	24	23
CHE	1	1	3	2	4	11	11	28	27
Sum	119	121	175	196	186	223	199	278	286

	2006	2007	2008	2009	2010	2011	2012	2013	2014
CAN	25	29	24	24	23	31	25	28	22
ENG	17	17	21	28	38	34	25	36	37
JPN	14	19	18	32	32	45	50	42	44
FED	76	85	88	70	73	62	51	59	46
NOR	20	15	16	17	16	11	10	10	8
ECB	58	88	149	130	121	139	103	152	130
NYC	12	9	4	11	23	22	18	29	26
AUS	9	10	17	17	29	32	27	26	26
SWE	36	22	27	23	25	28	15	17	13
CHE	27	26	29	25	19	16	21	15	14
Sum	294	320	393	377	399	420	345	414	366

	2015	2016	2017	2018	2019	2020	Sum
CAN	23	27	26	29	19	23	499
ENG	47	38	37	44	48	32	625
JPN	35	33	39	27	40	24	668
FED	57	42	54	49	79	57	1465
NOR	8	8	10	9	10	5	262
ECB	136	120	169	135	143	99	2185
NYC	34	32	29	26	28	22	376
AUS	34	25	33	38	34	23	475
SWE	9	11	5	12	8	6	466
CHE	16	15	10	16	14	7	358
Sum	399	351	412	385	423	298	7379

Table 3.9: This table shows the average topic distribution for each central bank. The average distributions are calculated by taking the mean over the topic distribution for all classified speeches for each central bank.

	0	1	2	3	4	5	6	7	8	9
CAN	0.009	0.012	0.501	0.018	0.011	0.011	0.002	0.01	0.001	0.002
ENG	0.04	0.016	0.042	0.035	0.03	0.015	0.231	0.013	0.001	0.0
JPN	0.004	0.01	0.004	0.015	0.02	0.012	0.002	0.003	0.001	0.255
FED	0.086	0.008	0.039	0.01	0.029	0.109	0.001	0.003	0.0	0.001
NOR	0.009	0.022	0.008	0.018	0.033	0.011	0.004	0.009	0.522	0.001
ECB	0.01	0.016	0.005	0.082	0.026	0.006	0.002	0.087	0.001	0.0
NYC	0.103	0.006	0.04	0.021	0.023	0.129	0.001	0.012	0.001	0.001
AUS	0.009	0.013	0.071	0.014	0.031	0.011	0.007	0.006	0.001	0.001
SWE	0.023	0.027	0.003	0.033	0.024	0.004	0.002	0.008	0.003	0.001
CHE	0.025	0.034	0.013	0.031	0.037	0.022	0.003	0.01	0.002	0.001
Mean	0.032	0.016	0.073	0.028	0.026	0.033	0.026	0.016	0.053	0.026

	10	11	12	13	14	15	16	17	18	19
CAN	0.003	0.015	0.046	0.001	0.019	0.016	0.039	0.021	0.067	0.004
ENG	0.003	0.027	0.082	0.001	0.055	0.052	0.035	0.026	0.153	0.02
JPN	0.003	0.006	0.022	0.001	0.013	0.011	0.022	0.015	0.029	0.004
FED	0.109	0.027	0.085	0.0	0.017	0.017	0.031	0.04	0.055	0.002
NOR	0.001	0.017	0.029	0.002	0.084	0.012	0.026	0.017	0.029	0.006
ECB	0.001	0.009	0.023	0.002	0.014	0.009	0.016	0.026	0.042	0.192
NYC	0.001	0.064	0.018	0.16	0.001	0.01	0.026	0.019	0.064	0.072
AUS	0.002	0.274	0.039	0.001	0.068	0.043	0.021	0.037	0.171	0.001
SWE	0.002	0.03	0.026	0.001	0.114	0.007	0.019	0.018	0.039	0.02
CHE	0.002	0.015	0.039	0.378	0.042	0.022	0.027	0.033	0.046	0.046
Mean	0.019	0.044	0.055	0.039	0.044	0.022	0.025	0.03	0.07	0.03

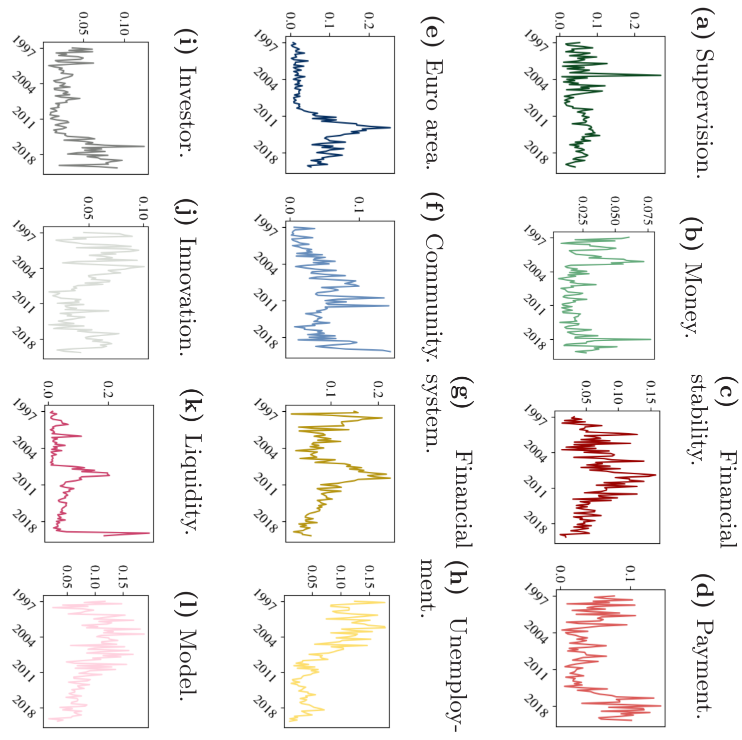
	20	21	22	23	24	25	26	27	28
CAN	0.002	0.064	0.007	0.002	0.004	0.035	0.064	0.007	0.008
ENG	0.001	0.037	0.006	0.001	0.002	0.017	0.04	0.005	0.016
JPN	0.434	0.038	0.005	0.002	0.002	0.02	0.033	0.005	0.013
FED	0.002	0.046	0.004	0.001	0.001	0.099	0.036	0.132	0.008
NOR	0.001	0.061	0.01	0.018	0.004	0.01	0.018	0.004	0.011
ECB	0.002	0.057	0.095	0.002	0.179	0.018	0.023	0.002	0.055
NYC	0.004	0.036	0.004	0.001	0.001	0.034	0.052	0.086	0.007
AUS	0.002	0.035	0.008	0.003	0.001	0.025	0.089	0.007	0.009
SWE	0.002	0.042	0.014	0.489	0.002	0.017	0.018	0.003	0.007
CHE	0.004	0.051	0.023	0.01	0.016	0.012	0.035	0.003	0.017
Mean	0.045	0.047	0.018	0.053	0.021	0.029	0.041	0.025	0.015

Table 3.10: This table shows the estimated coefficients from the AR(1) models with control variables, using quarterly data from the Fed. Each equation is represented as a column. The t-statistics (reported in parenthesis) are based on Newey and West (1987) standard errors with 1 lag.

	(1) Regulation	(2) Money	(3) Financial stability	(4) Payment	(5) Euro area	(6) Community	(7) Financial system	(8) Unemployment	(9) Investor	(10) Innovation	(11) Liquidity	(12) Model	(13) Fiscal	(14) Productivity	(15) Trade	(16) FOMC	(17) Deflation
1 Lag	0.0211	0.170	0.174	0.381**	0.583***	0.309	0.262	0.270*	0.367***	0.0928	0.542***	-0.0121	0.334*	0.421***	0.127	0.114	0.479***
	(0.18)	(1.37)	(1.60)	(2.81)	(5.48)	(1.93)	(1.88)	(2.15)	(3.79)	(0.79)	(4.51)	(-0.10)	(2.25)	(4.06)	(0.98)	(1.31)	(3.94)
Inflation	-0.217	0.0326	-0.0803	-0.396	-0.305	-0.141	0.225	-0.145	-0.108	0.0291	-0.416	0.373	0.518	0.645	0.165	0.383	0.114
	(-0.57)	(0.24)	(-0.13)	(-1.21)	(-1.08)	(-0.39)	(0.53)	(-0.46)	(-0.71)	(0.13)	(-1.20)	(1.31)	(1.15)	(1.41)	(0.44)	(1.37)	(0.35)
Lag Inflation	-0.165	-0.366**	-0.125	0.375	0.0950	-0.0224	-0.381	0.350	-0.157	0.204	-0.0432	0.212	0.0725	-0.175	-0.0773	-0.586	-0.525
	(-0.49)	(-3.33)	(-0.30)	(1.09)	(0.24)	(-0.06)	(-0.84)	(1.32)	(-0.88)	(0.76)	(-0.11)	(0.69)	(0.17)	(-0.42)	(-0.26)	(-1.56)	(-1.84)
1 year bond	-0.793	-0.0410	-0.778	-0.251	-0.301	-0.925	-0.596	1.587**	0.237	-0.223	-1.578*	-0.384	-0.684	-0.0599	2.037**	-0.259	0.0466
	(-1.66)	(-0.14)	(-1.10)	(-0.39)	(-0.63)	(-1.59)	(-0.70)	(2.90)	(0.84)	(-0.49)	(-2.16)	(-0.70)	(-1.24)	(-0.06)	(3.12)	(-0.42)	(0.11)
Lag 1 year bond	-0.00155	0.142	-0.323	0.493	-1.148*	-0.0752	1.010	-0.161	-0.388	0.384	0.943	0.549	-0.953	0.611	0.151	-0.336	-1.151*
	(-0.00)	(0.41)	(-0.45)	(0.69)	(-2.15)	(-0.11)	(1.30)	(-0.28)	(-1.44)	(0.86)	(1.32)	(1.17)	(-1.37)	(0.82)	(0.25)	(-0.57)	(-2.33)
SP500 returns	-0.244	-0.394*	1.201*	-0.275	0.424	0.0697	-0.299	0.0832	-0.635*	0.166	0.313	1.319**	-0.0479	0.773	-0.0949	0.158	-0.213
	(-0.47)	(-2.04)	(1.99)	(-0.47)	(1.13)	(0.11)	(-0.48)	(0.19)	(-2.49)	(0.50)	(0.63)	(2.93)	(-0.10)	(1.35)	(-0.22)	(0.38)	(-0.59)
Lag SP500 returns	0.331	-0.184	-0.176	-0.0587	0.258	-0.00507	0.157	-0.0580	0.273	0.450	-0.0878	-0.685	-0.327	-0.223	-0.258	-0.159	0.186
	(0.73)	(-0.92)	(-0.37)	(-0.11)	(0.50)	(-0.01)	(0.27)	(-0.18)	(1.01)	(0.91)	(-0.17)	(-1.53)	(-0.70)	(-0.41)	(-0.62)	(-0.33)	(0.46)
VIX	0.0000797	0.000707*	-0.000968	0.000413	-0.000873	0.000115	0.000524	-0.000402	0.000834*	-0.000293	0.000172	-0.00175**	0.000750	-0.000922	-0.000430	-0.000179	0.000363
	(0.13)	(2.40)	(-1.40)	(0.54)	(-1.57)	(0.14)	(0.68)	(-0.73)	(2.03)	(-0.64)	(0.28)	(-2.97)	(1.11)	(-1.15)	(-0.68)	(-0.30)	(0.47)
Lag VIX	-0.000614	-0.000143	0.000469	-0.000207	-0.000109	0.000461	0.0000854	0.000188	-0.000656	-0.000881	0.000762	0.000661	0.0000888	-0.000413	0.000365	-0.000162	-0.000240
	(-1.00)	(-0.47)	(0.52)	(-0.27)	(-0.14)	(0.66)	(0.10)	(0.39)	(-1.94)	(-1.24)	(0.82)	(0.95)	(0.15)	(-0.56)	(0.59)	(-0.21)	(-0.46)
Constants	0.0497***	0.0112**	0.0258	0.0191*	0.0332*	0.0122	0.0208	0.0192**	0.0201*	0.0401***	-0.0118	0.0572***	0.0183	0.0405**	0.0310**	0.0472***	0.0206
	(4.49)	(2.77)	(1.58)	(2.01)	(2.35)	(0.83)	(1.82)	(2.85)	(2.46)	(3.95)	(-0.74)	(4.37)	(1.43)	(3.36)	(3.03)	(3.60)	(1.84)
N	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 3.4: Figures 3.4a through 3.4q display each topic in Figure 3.1 in Section 4.1 as line plots. Each figure displays the normalized probability for that topic. The probabilities are calculated by classifying the central bank speeches in the corpus by the estimated topics. The figures show the probability that the topics are addressed in each quarter during the sample period.



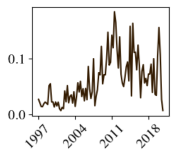
Fed robustness

Table 3.11: This table shows the estimated coefficients from the AR(1) models, using quarterly data from the Fed. The documents are classified with a DTM model trained on the Fed speeches alone. Each equation is represented as a column. The t-statistics (reported in parenthesis) are based on Newey and West (1987) standard errors with 1 lag.

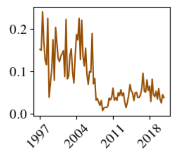
	(1) Labor	(2) Forecast	(3) Supervision	(4) Housing	(5) Payment	(6) Small business	(7) Innovation	(8) Regulation	(9) Spending	(10) Financial markets	(11) FED	(12) Community	(13) Compliance	(14) Trade	(15) Treasury
1 Lag	0.672***	0.171	0.241	0.797***	0.0876	0.132	0.640***	0.489***	0.464***	0.402***	0.579***	0.277*	0.567***	0.226*	0.519***
	(7.68)	(1.60)	(1.87)	(7.95)	(0.70)	(1.13)	(6.63)	(4.26)	(3.91)	(3.71)	(5.49)	(2.31)	(5.90)	(2.55)	(5.15)
Inflation	0.882	1.024*	-0.115	-0.588	-0.224	0.0300	0.398	-0.266	0.946	-0.134	-0.707	0.468	-0.107	-0.141	-1.342**
	(1.54)	(2.03)	(-0.15)	(-1.00)	(-0.45)	(0.08)	(0.67)	(-0.52)	(1.14)	(-0.23)	(-1.35)	(1.00)	(-0.28)	(-0.31)	(-2.90)
Lag Inflation	-1.284*	-1.517**	1.844*	-0.609	0.0924	-0.342	0.872*	-0.0738	-0.545	1.087	0.298	0.753	-0.0731	-0.104	-0.292
	(-2.45)	(-2.85)	(2.16)	(-1.61)	(0.25)	(-0.72)	(2.28)	(-0.10)	(-0.57)	(1.68)	(0.35)	(1.54)	(-0.17)	(-0.27)	(-0.71)
1 year bond	0.404	0.915	1.204	1.032	2.253	1.403	1.162	0.288	2.605	-1.660	-3.243*	-2.436***	-0.0541	-0.0578	-1.456
	(0.45)	(0.89)	(0.72)	(1.38)	(1.76)	(1.51)	(0.88)	(0.33)	(1.24)	(-1.80)	(-2.08)	(-3.69)	(-0.08)	(-0.09)	(-1.84)
Lag 1 year bond	-2.348**	0.341	2.363	-0.295	2.467	-1.662	1.556	-2.346*	-0.589	0.891	1.720	-0.474	-0.527	1.327	-0.119
	(-2.87)	(0.38)	(1.32)	(-0.51)	(1.79)	(-1.98)	(1.51)	(-2.04)	(-0.21)	(0.91)	(1.33)	(-0.55)	(-0.84)	(1.90)	(-0.14)
SP500 returns	-1.454*	0.569	0.862	0.330	-2.589**	-0.235	-0.246	-0.794	0.0402	0.303	1.481	0.693	0.182	-0.0307	0.499
	(-2.34)	(0.61)	(0.76)	(0.55)	(-2.88)	(-0.27)	(-0.36)	(-1.10)	(0.04)	(0.49)	(1.25)	(0.99)	(0.39)	(-0.06)	(0.77)
Lag SP500 returns	0.112	0.135	-0.125	-0.338	0.284	-0.141	0.210	0.779	0.130	-0.410	-0.438	-0.188	-0.337	-0.312	0.395
	(0.15)	(0.17)	(-0.12)	(-0.60)	(0.31)	(-0.18)	(0.27)	(0.92)	(0.08)	(-0.73)	(-0.46)	(-0.30)	(-0.61)	(-0.54)	(0.59)
VIX	0.00232*	-0.00102	-0.00154	-0.000343	0.00362**	0.000210	-0.000281	0.000446	-0.000613	0.000131	-0.00170	-0.000469	-0.000563	-0.000418	-0.000140
	(2.57)	(-0.82)	(-0.97)	(-0.43)	(3.04)	(0.15)	(-0.31)	(0.45)	(-0.40)	(0.17)	(-1.28)	(-0.52)	(-0.80)	(-0.62)	(-0.18)
Lag VIX	-0.000877	-0.00135	0.000293	0.000483	-0.000948	0.000859	-0.000203	-0.000150	-0.000193	0.000230	0.00127	-0.000238	0.000594	-0.000314	0.000576
	(-0.70)	(-1.20)	(0.18)	(0.55)	(-0.74)	(0.75)	(-0.18)	(-0.12)	(-0.10)	(0.26)	(0.67)	(-0.27)	(0.60)	(-0.39)	(0.56)
Constants	0.0481	0.114***	0.0385	0.00583	0.00329	0.0230	0.0153	0.0410*	0.0528*	0.0373*	0.0263	0.0689**	0.0277	0.0594***	0.00835
	(1.97)	(5.30)	(1.57)	(0.48)	(0.18)	(1.24)	(1.06)	(2.44)	(2.24)	(2.08)	(0.86)	(3.30)	(1.86)	(4.00)	(0.53)
N	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

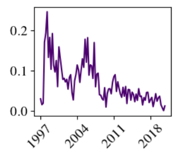
(m) Fiscal.



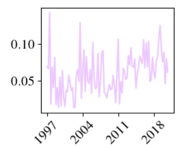
(n) Productivity.



(o) Trade.



(p) FOMC.



(q) Deflation.

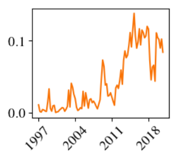
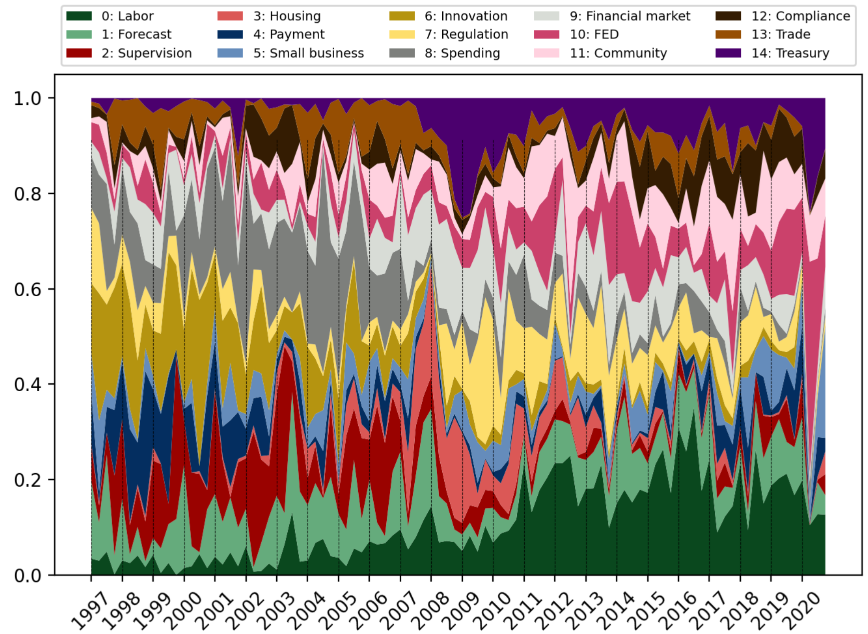


Figure 3.5: This figure illustrates the average normalized probability distributions for the topics for each quarter, given by the classified documents using DTM trained on the data from the Fed and the New York Fed alone. In each quarter all speeches are classified, their probability distributions averaged, and plotted.



Previous doctoral theses in the Department of Economics, Gothenburg

Avhandlingar publicerade innan serien Ekonomiska Studier startades
(Theses published before the series Ekonomiska Studier was started):

Östman, Hugo (1911), Norrlands ekonomiska utveckling
Moritz, Marcus (1911), Den svenska tobaksindustrien
Sundbom, I. (1933), Prisbildning och ändamålsenlighet
Gerhard, I. (1948), Problem rörande Sveriges utrikeshandel 1936/38
Hegeland, Hugo (1951), The Quantity Theory of Money
Mattsson, Bengt (1970), Cost-Benefit analys
Rosengren, Björn (1975), Valutareglering och nationell ekonomisk politik
Hjalmarsson, Lennart (1975), Studies in a Dynamic Theory of Production and its Applications
Örtendahl, Per-Anders (1975), Substitutionsaspekter på produktionsprocessen vid massaframställning
Anderson, Arne M. (1976), Produktion, kapacitet och kostnader vid ett helautomatiskt emballageglasbruk
Ohlsson, Olle (1976), Substitution och odelbarheter i produktionsprocessen vid massaframställning
Gunnarsson, Jan (1976), Produktionssystem och tätortshierarki – om sambandet mellan rumslig och ekonomisk struktur
Köstner, Evert (1976), Optimal allokering av tid mellan utbildning och arbete
Wigren, Rune (1976), Analys av regionala effektivitetsskillnader inom industribranscher
Wästlund, Jan (1976), Skattning och analys av regionala effektivitetsskillnader inom industribranscher
Flöjstad, Gunnar (1976), Studies in Distortions, Trade and Allocation Problems
Sandelin, Bo (1977), Prisutveckling och kapitalvinster på bostadsfastigheter
Dahlberg, Lars (1977), Empirical Studies in Public Planning
Lönnroth, Johan (1977), Marxism som matematisk ekonomi
Johansson, Börje (1978), Contributions to Sequential Analysis of Oligopolistic Competition

<u>Ekonomiska Studier</u>, utgivna av Nationalekonomiska institutionen vid Göteborgs Universitet. Nr 1 och 4 var inte doktorsavhandlingar. (The contributions to the department series 'Ekonomiska Studier' where no. 1 and 4 were no doctoral theses):

2. **Ambjörn, Erik** (1959), Svenskt importberoende 1926-1956: en ekonomisk-statistisk kartläggning med kommentarer
3. **Landgren, K-G.** (1960), Den "Nya ekonomien" i Sverige: J.M. Keynes, E. Wigfors och utecklingen 1927-39
5. **Bigsten, Arne** (1979), Regional Inequality and Development: A Case Study of Kenya
6. **Andersson, Lars** (1979), Statens styrning av de kommunala budgetarnas struktur (Central Government Influence on the Structure of the Municipal Budget)
7. **Gustafsson, Björn** (1979), Inkomst- och uppväxtförhållanden (Income and Family Background)
8. **Granholm, Arne** (1981), Interregional Planning Models for the Allocation of Private and Public Investments
9. **Lundborg, Per** (1982), Trade Policy and Development: Income Distributional Effects in the Less Developed Countries of the US and EEC Policies for Agricultural Commodities
10. **Juås, Birgitta** (1982), Värdering av risken för personskador. En jämförande studie av implicita och explicita värden. (Valuation of Personal Injuries. A comparison of Explicit and Implicit Values)
11. **Bergendahl, Per-Anders** (1982), Energi och ekonomi - tillämpningar av input-output analys (Energy and the Economy - Applications of Input-Output Analysis)
12. **Blomström, Magnus** (1983), Foreign Investment, Technical Efficiency and Structural Change - Evidence from the Mexican Manufacturing Industry
13. **Larsson, Lars-Göran** (1983), Comparative Statics on the Basis of Optimization Methods
14. **Persson, Håkan** (1983), Theory and Applications of Multisectoral Growth Models
15. **Sterner, Thomas** (1986), Energy Use in Mexican Industry.
16. **Flood, Lennart** (1986), On the Application of Time Use and Expenditure Allocation Models.
17. **Schuller, Bernd-Joachim** (1986), Ekonomi och kriminalitet - en empirisk undersökning av brottsligheten i Sverige (Economics of crime - an empirical analysis of crime in Sweden)
18. **Walfridson, Bo** (1987), Dynamic Models of Factor Demand. An Application to Swedish Industry.
19. **Stålhammar, Nils-Olov** (1987), Strukturomvandling, företagsbeteende och förväntningsbildning inom den svenska tillverkningsindustrin (Structural Change, Firm Behaviour and Expectation Formation in Swedish Manufactury)
20. **Anxo, Dominique** (1988), Sysselsättningseffekter av en allmän arbetstidsförkortning (Employment effects of a general shortage of the working time)
21. **Mbelle, Ammon** (1988), Foreign Exchange and Industrial Development: A Study of Tanzania.
22. **Ongaro, Wilfred** (1988), Adoption of New Farming Technology: A Case Study of Maize Production in Western Kenya.
23. **Zejan, Mario** (1988), Studies in the Behavior of Swedish Multinationals.
24. **Görling, Anders** (1988), Ekonomisk tillväxt och miljö. Förorenings-struktur och ekonomiska effekter av olika miljövårdsprogram. (Economic Growth and Environment. Pollution Structure and Economic Effects of Some Environmental Programs).
25. **Aguilar, Renato** (1988), Efficiency in Production: Theory and an Application on Kenyan Smallholders.
26. **Kayizzi-Mugerwa, Steve** (1988), External Shocks and Adjustment in Zambia.
27. **Bornmalm-Jardelöw, Gunilla** (1988), Högre utbildning och arbetsmarknad (Higher Education and the Labour Market)
28. **Tansini, Ruben** (1989), Technology Transfer: Dairy Industries in Sweden and Uruguay.
29. **Andersson, Irene** (1989), Familjebeskattning, konsumtion och arbetsutbud - En ekonometrisk analys av löne- och inkomstelasticiteter samt policysimuleringar för svenska hushåll (Family Taxation, Consumption and Labour Supply - An Econometric Analysis of Wage and Income Elasticities and Policy Simulations for Swedish Households)
30. **Henrekson, Magnus** (1990), An Economic Analysis of Swedish Government Expenditure
31. **Sjöö, Boo** (1990), Monetary Policy in a Continuous Time Dynamic Model for Sweden

32. **Rosén, Åsa** (1991), Contributions to the Theory of Labour Contracts.
33. **Loureiro, Joao M. de Matos** (1992), Foreign Exchange Intervention, Sterilization and Credibility in the EMS: An Empirical Study
34. **Irandoust, Manuchehr** (1993), Essays on the Behavior and Performance of the Car Industry
35. **Tasiran, Ali Cevat** (1993), Wage and Income Effects on the Timing and Spacing of Births in Sweden and the United States
36. **Milopoulos, Christos** (1993), Investment Behaviour under Uncertainty: An Econometric Analysis of Swedish Panel Data
37. **Andersson, Per-Åke** (1993), Labour Market Structure in a Controlled Economy: The Case of Zambia
38. **Storrie, Donald W.** (1993), The Anatomy of a Large Swedish Plant Closure
39. **Semboja, Haji Hatibu Haji** (1993), Energy and Development in Kenya
40. **Makonnen, Negatu** (1993), Labor Supply and the Distribution of Economic Well-Being: A Case Study of Lesotho
41. **Julin, Eva** (1993), Structural Change in Rural Kenya
42. **Durevall, Dick** (1993), Essays on Chronic Inflation: The Brazilian Experience
43. **Veiderpass, Ann** (1993), Swedish Retail Electricity Distribution: A Non-Parametric Approach to Efficiency and Productivity Change
44. **Odeck, James** (1993), Measuring Productivity Growth and Efficiency with Data Envelopment Analysis: An Application on the Norwegian Road Sector
45. **Mwenda, Abraham** (1993), Credit Rationing and Investment Behaviour under Market Imperfections: Evidence from Commercial Agriculture in Zambia
46. **Mlambo, Kupukile** (1993), Total Factor Productivity Growth: An Empirical Analysis of Zimbabwe's Manufacturing Sector Based on Factor Demand Modelling
47. **Ndung'u, Njuguna** (1993), Dynamics of the Inflationary Process in Kenya
48. **Modén, Karl-Markus** (1993), Tax Incentives of Corporate Mergers and Foreign Direct Investments
49. **Franzén, Mikael** (1994), Gasoline Demand - A Comparison of Models
50. **Heshmati, Almas** (1994), Estimating Technical Efficiency, Productivity Growth And Selectivity Bias Using Rotating Panel Data: An Application to Swedish Agriculture
51. **Salas, Osvaldo** (1994), Efficiency and Productivity Change: A Micro Data Case Study of the Colombian Cement Industry
52. **Bjurek, Hans** (1994), Essays on Efficiency and Productivity Change with Applications to Public Service Production
53. **Cabezas Vega, Luis** (1994), Factor Substitution, Capacity Utilization and Total Factor Productivity Growth in the Peruvian Manufacturing Industry
54. **Katz, Katarina** (1994), Gender Differentiation and Discrimination. A Study of Soviet Wages
55. **Asal, Maher** (1995), Real Exchange Rate Determination and the Adjustment Process: An Empirical Study in the Cases of Sweden and Egypt
56. **Kjulin, Urban** (1995), Economic Perspectives on Child Care
57. **Andersson, Göran** (1995), Volatility Forecasting and Efficiency of the Swedish Call Options Market
58. **Forteza, Alvaro** (1996), Credibility, Inflation and Incentive Distortions in the Welfare State
59. **Locking, Håkan** (1996), Essays on Swedish Wage Formation
60. **Välilä, Timo** (1996), Essays on the Credibility of Central Bank Independence
61. **Yilma, Mulugeta** (1996), Measuring Smallholder Efficiency: Ugandan Coffee and Food-Crop Production
62. **Mabugu, Ramos E.** (1996), Tax Policy Analysis in Zimbabwe Applying General Equilibrium Models
63. **Johansson, Olof** (1996), Welfare, Externalities, and Taxation; Theory and Some Road Transport Applications.
64. **Chitiga, Margaret** (1996), Computable General Equilibrium Analysis of Income Distribution Policies in Zimbabwe
65. **Leander, Per** (1996), Foreign Exchange Market Behavior Expectations and Chaos
66. **Hansen, Jörgen** (1997), Essays on Earnings and Labor Supply
67. **Cotfas, Mihai** (1997), Essays on Productivity and Efficiency in the Romanian Cement Industry
68. **Horgby, Per-Johan** (1997), Essays on Sharing, Management and Evaluation of Health Risks
69. **Nafar, Nosratollah** (1997), Efficiency and Productivity in Iranian Manufacturing Industries
70. **Zheng, Jinghai** (1997), Essays on Industrial Structure, Technical Change, Employment Adjustment, and Technical Efficiency
71. **Isaksson, Anders** (1997), Essays on Financial Liberalisation in Developing Countries: Capital mobility, price stability, and savings
72. **Gerdin, Anders** (1997), On Productivity and Growth in Kenya, 1964-94
73. **Sharifi, Alimorad** (1998), The Electricity Supply Industry in Iran: Organization, performance and future development
74. **Zamanian, Max** (1997), Methods for Mutual Fund Portfolio Evaluation: An application to the Swedish market
75. **Manda, Damiano Kulundu** (1997), Labour Supply, Returns to Education, and the Effect of Firm Size on Wages: The case of Kenya
76. **Holmén, Martin** (1998), Essays on Corporate Acquisitions and Stock Market Introductions
77. **Pan, Kelvin** (1998), Essays on Enforcement in Money and Banking
78. **Rogat, Jorge** (1998), The Value of Improved Air Quality in Santiago de Chile
79. **Peterson, Stefan** (1998), Essays on Large Shareholders and Corporate Control
80. **Belhaj, Mohammed** (1998), Energy, Transportation and Urban Environment in Africa: The Case of Rabat-Salé, Morocco
81. **Mekonnen, Alemu** (1998), Rural Energy and Afforestation: Case Studies from Ethiopia
82. **Johansson, Anders** (1998), Empirical Essays on Financial and Real Investment Behavior
83. **Köhlin, Gunnar** (1998), The Value of Social Forestry in Orissa, India
84. **Levin, Jörgen** (1998), Structural Adjustment and Poverty: The Case of Kenya
85. **Ncube, Mkhululi** (1998), Analysis of Employment Behaviour in Zimbabwe
86. **Mwansa, Ladslous** (1998), Determinants of Inflation in Zambia
87. **Agnarsson, Sveinn** (1998), Of Men and Machines: Essays in Applied Labour and Production Economics
88. **Kadenge, Phineas** (1998), Essays on Macroeconomic Adjustment in Zimbabwe: Inflation, Money Demand, and the Real Exchange Rate
89. **Nyman, Håkan** (1998), An Economic Analysis of Lone Motherhood in Sweden
90. **Carlsson, Fredrik** (1999), Essays on Externalities and Transport
91. **Johansson, Mats** (1999), Empirical Studies of Income Distribution

92. **Alemu, Tekie** (1999), Land Tenure and Soil Conservation: Evidence from Ethiopia
93. **Lundvall, Karl** (1999), Essays on Manufacturing Production in a Developing Economy: Kenya 1992-94
94. **Zhang, Jianhua** (1999), Essays on Emerging Market Finance
95. **Mlima, Aziz Ponary** (1999), Four Essays on Efficiency and Productivity in Swedish Banking
96. **Davidsen, Björn-Ivar** (2000), Bidrag til den økonomisk-metodologiske tenkningen (Contributions to the Economic Methodological Thinking)
97. **Ericson, Peter** (2000), Essays on Labor Supply
98. **Söderbom, Måns** (2000), Investment in African Manufacturing: A Microeconomic Analysis
99. **Höglund, Lena** (2000), Essays on Environmental Regulation with Applications to Sweden
100. **Olsson, Ola** (2000), Perspectives on Knowledge and Growth
101. **Meuller, Lars** (2000), Essays on Money and Credit
102. **Österberg, Torun** (2000), Economic Perspectives on Immigrants and Intergenerational Transmissions
103. **Kalinda Mkenda, Beatrice** (2001), Essays on Purchasing Power Parity, RealExchange Rate, and Optimum Currency Areas
104. **Nerhagen, Lena** (2001), Travel Demand and Value of Time - Towards an Understanding of Individuals Choice Behavior
105. **Mkenda, Adolf** (2001), Fishery Resources and Welfare in Rural Zanzibar
106. **Eggert, Håkan** (2001), Essays on Fisheries Economics
107. **Andrén, Daniela** (2001), Work, Sickness, Earnings, and Early Exits from the Labor Market. An Empirical Analysis Using Swedish Longitudinal Data
108. **Nivorozhkin, Eugene** (2001), Essays on Capital Structure
109. **Hammar, Henrik** (2001), Essays on Policy Instruments: Applications to Smoking and the Environment
110. **Nannyonjo, Justine** (2002), Financial Sector Reforms in Uganda (1990-2000): Interest Rate Spreads, Market Structure, Bank Performance and Monetary Policy
111. **Wu, Hong** (2002), Essays on Insurance Economics
112. **Linde-Rahr, Martin** (2002), Household Economics of Agriculture and Forestry in Rural Vienam
113. **Maneschiöld, Per-Ola** (2002), Essays on Exchange Rates and Central Bank Credibility
114. **Andrén, Thomas** (2002), Essays on Training, Welfare and Labor Supply
115. **Granér, Mats** (2002), Essays on Trade and Productivity: Case Studies of Manufacturing in Chile and Kenya
116. **Jaldell, Henrik** (2002), Essays on the Performance of Fire and Rescue Services
117. **Alpizar, Francisco, R.** (2002), Essays on Environmental Policy-Making in Developing Countries: Applications to Costa Rica
118. **Wahlberg, Roger** (2002), Essays on Discrimination, Welfare and Labor Supply
119. **Piculescu, Violeta** (2002), Studies on the Post-Communist Transition
120. **Pylkkänen, Elina** (2003), Studies on Household Labor Supply and Home Production
121. **Löfgren, Åsa** (2003), Environmental Taxation – Empirical and Theoretical Applications
122. **Ivaschenko, Oleksiy** (2003), Essays on Poverty, Income Inequality and Health in Transition Economies
123. **Lundström, Susanna** (2003), On Institutions, Economic Growth and the Environment
124. **Wambugu, Anthony** (2003), Essays on Earnings and Human Capital in Kenya
125. **Adler, Johan** (2003), Aspects of Macroeconomic Saving
126. **Erlandsson, Mattias** (2003), On Monetary Integration and Macroeconomic Policy
127. **Brink, Anna** (2003), On the Political Economy of Municipality Break-Ups
128. **Ljungwall, Christer** (2003), Essays on China's Economic Performance During the Reform Period
129. **Chifamba, Ronald** (2003), Analysis of Mining Investments in Zimbabwe
130. **Muchapondwa, Edwin** (2003), The Economics of Community-Based Wildlife Conservation in Zimbabwe
131. **Hammes, Klaus** (2003), Essays on Capital Structure and Trade Financing
132. **Abou-Ali, Hala** (2003), Water and Health in Egypt: An Empirical Analysis
133. **Simatele, Munacinga** (2004), Financial Sector Reforms and Monetary Policy in Zambia
134. **Tezic, Kerem** (2004), Essays on Immigrants' Economic Integration
135. INSTÄLLD
136. **Gjirja, Matilda** (2004), Efficiency and Productivity in Swedish Banking
137. **Andersson, Jessica** (2004), Welfare Environment and Tourism in Developing Countries
138. **Chen, Yinghong** (2004), Essays on Voting Power, Corporate Governance and Capital Structure
139. **Yesuf, Mahmud** (2004), Risk, Time and Land Management under Market Imperfections: Applications to Ethiopia
140. **Kateregga, Eseza** (2005), Essays on the Infestation of Lake Victoria by the Water Hyacinth
141. **Edvardsen, Dag Fjeld** (2004), Four Essays on the Measurement of Productive Efficiency
142. **Lidén, Erik** (2005), Essays on Information and Conflicts of Interest in Stock Recommendations
143. **Dieden, Sten** (2005), Income Generation in the African and Coloured Population – Three Essays on the Origins of Household Incomes in South Africa
144. **Eliasson, Marcus** (2005), Individual and Family Consequences of Involuntary Job Loss
145. **Mahmud, Minhaj** (2005), Measuring Trust and the Value of Statistical Lives: Evidence from Bangladesh
146. **Lokina, Razack Bakari** (2005), Efficiency, Risk and Regulation Compliance: Applications to Lake Victoria Fisheries in Tanzania
147. **Jussila Hammes, Johanna** (2005), Essays on the Political Economy of Land Use Change
148. **Nyangena, Wilfred** (2006), Essays on Soil Conservation, Social Capital and Technology Adoption
149. **Nivorozhkin, Anton** (2006), Essays on Unemployment Duration and Programme Evaluation
150. **Sandén, Klas** (2006), Essays on the Skill Premium
151. **Deng, Daniel** (2006), Three Essays on Electricity Spot and Financial Derivative Prices at the Nordic Power Exchange
152. **Gebreeyesus, Mulu** (2006), Essays on Firm Turnover, Growth, and Investment Behavior in Ethiopian Manufacturing

153. **Islam, Nizamul Md.** (2006), Essays on Labor Supply and Poverty: A Microeconometric Application
154. **Kjaer, Mats** (2006), Pricing of Some Path-Dependent Options on Equities and Commodities
155. **Shimeles, Abebe** (2006), Essays on Poverty, Risk and Consumption Dynamics in Ethiopia
156. **Larsson, Jan** (2006), Four Essays on Technology, Productivity and Environment
157. **Congdon Fors, Heather** (2006), Essays in Institutional and Development Economics
158. **Akpalu, Wisdom** (2006), Essays on Economics of Natural Resource Management and Experiments
159. **Daruvala, Dinky** (2006), Experimental Studies on Risk, Inequality and Relative Standing
160. **García, Jorge** (2007), Essays on Asymmetric Information and Environmental Regulation through Disclosure
161. **Bezabih, Mintewab** (2007), Essays on Land Lease Markets, Productivity, Biodiversity, and Environmental Variability
162. **Visser, Martine** (2007), Fairness, Reciprocity and Inequality: Experimental Evidence from South Africa
163. **Holm, Louise** (2007), A Non-Stationary Perspective on the European and Swedish Business Cycle
164. **Herbertsson, Alexander** (2007), Pricing Portfolio Credit Derivatives
165. **Johansson, Anders C.** (2007), Essays in Empirical Finance: Volatility, Interdependencies, and Risk in Emerging Markets
166. **Ibáñez Díaz, Marcela** (2007), Social Dilemmas: The Role of Incentives, Norms and Institutions
167. **Ekbom, Anders** (2007), Economic Analysis of Soil Capital, Land Use and Agricultural Production in Kenya
168. **Sjöberg, Pål** (2007), Essays on Performance and Growth in Swedish Banking
169. **Palma Aguirre, Grisha Alexis** (2008), Explaining Earnings and Income Inequality in Chile
170. **Akay, Alpaslan** (2008), Essays on Microeconometrics and Immigrant Assimilation
171. **Carlsson, Evert** (2008), After Work – Investing for Retirement
172. **Munshi, Farzana** (2008), Essays on Globalization and Occupational Wages
173. **Tsakas, Elias** (2008), Essays on Epistemology and Evolutionary Game Theory
174. **Erlandzon, Karl** (2008), Retirement Planning: Portfolio Choice for Long-Term Investors
175. **Lampi, Elina** (2008), Individual Preferences, Choices, and Risk Perceptions – Survey Based Evidence
176. **Mitrut, Andreea** (2008), Four Essays on Interhousehold Transfers and Institutions in Post-Communist Romania
177. **Hansson, Gustav** (2008), Essays on Social Distance, Institutions, and Economic Growth
178. **Zikhali, Precious** (2008), Land Reform, Trust and Natural Resource Management in Africa
179. **Tengstam, Sven** (2008), Essays on Smallholder Diversification, Industry Location, Debt Relief, and Disability and Utility
180. **Boman, Anders** (2009), Geographic Labour Mobility – Causes and Consequences
181. **Qin, Ping** (2009), Risk, Relative Standing and Property Rights: Rural Household Decision-Making in China
182. **Wei, Jiegen** (2009), Essays in Climate Change and Forest Management
183. **Belu, Constantin** (2009), Essays on Efficiency Measurement and Corporate Social Responsibility
184. **Ahlerup, Pelle** (2009), Essays on Conflict, Institutions, and Ethnic Diversity
185. **Quiroga, Miguel** (2009), Microeconomic Policy for Development: Essays on Trade and Environment, Poverty and Education
186. **Zerfu, Daniel** (2010), Essays on Institutions and Economic Outcomes
187. **Wollbrant, Conny** (2010), Self-Control and Altruism
188. **Villegas Palacio, Clara** (2010), Formal and Informal Regulations: Enforcement and Compliance
189. **Maican, Florin** (2010), Essays in Industry Dynamics on Imperfectly Competitive Markets
190. **Jakobsson, Niklas** (2010), Laws, Attitudes and Public Policy
191. **Manescu, Cristiana** (2010), Economic Implications of Corporate Social Responsibility and Responsible Investments
192. **He, Haoran** (2010), Environmental and Behavioral Economics – Applications to China
193. **Andersson, Fredrik W.** (2011), Essays on Social Comparison
194. **Isaksson, Ann-Sofie** (2011), Essays on Institutions, Inequality and Development
195. **Pham, Khanh Nam** (2011), Prosocial Behavior, Social Interaction and Development: Experimental Evidence from Vietnam
196. **Lindskog, Annika** (2011), Essays on Economic Behaviour: HIV/AIDS, Schooling, and Inequality
197. **Kotsadam, Andreas** (2011), Gender, Work, and Attitudes
198. **Alem, Yonas** (2011), Essays on Shocks, Welfare, and Poverty Dynamics: Microeconometric Evidence from Ethiopia
199. **Köksal-Ayhan, Miyase Yesim** (2011), Parallel Trade, Reference Pricing and Competition in the Pharmaceutical Market: Theory and Evidence
200. **Vondolia, Godwin Kofi** (2011), Essays on Natural Resource Economics
201. **Widerberg, Anna** (2011), Essays on Energy and Climate Policy – Green Certificates, Emissions Trading and Electricity Prices
202. **Siba, Eyerusalem** (2011), Essays on Industrial Development and Political Economy of Africa
203. **Orth, Matilda** (2012), Entry, Competition and Productivity in Retail
204. **Nerman, Måns** (2012), Essays on Development: Household Income, Education, and Female Participation and Representation
205. **Wicks, Rick** (2012), The Place of Conventional Economics in a World with Communities and Social Goods
206. **Sato, Yoshihiro** (2012), Dynamic Investment Models, Employment Generation and Productivity – Evidence from Swedish Data
207. **Valsecchi, Michele** (2012), Essays in Political Economy of Development
208. **Teklewold Belayneh, Hailemariam** (2012), Essays on the Economics of Sustainable Agricultural Technologies in Ethiopia
209. **Wagura Ndiritu, Simon** (2013), Essays on Gender Issues, Food Security, and Technology Adoption in East Africa
210. **Ruist, Joakim** (2013), Immigration, Work, and Welfare
211. **Nordén, Anna** (2013), Essays on Behavioral Economics and Policies for Provision of Ecosystem Services
212. **Yang, Xiaojun** (2013), Household Decision Making, Time Preferences, and

Positional Concern: Experimental Evidence from Rural China

213. **Bonilla Londoño, Jorge Alexander** (2013), Essays on the Economics of Air Quality Control
214. **Mohlin, Kristina** (2013), Essays on Environmental Taxation and Climate Policy
215. **Medhin, Haileselassie** (2013), The Poor and Their Neighbors: Essays on Behavioral and Experimental Economics
216. **Andersson, Lisa** (2013), Essays on Development and Experimental Economics: Migration, Discrimination and Positional Concerns
217. **Weng, Qian** (2014), Essays on Team Cooperation and Firm Performance
218. **Zhang, Xiao-Bing** (2015), Cooperation and Paradoxes in Climate Economics
219. **Jaime Torres, Monica Marcela** (2015) Essays on Behavioral Economics and Policy Design
220. **Bejenariu, Simona** (2015) Determinants of Health Capital at Birth: Evidence from Policy Interventions
221. **Nguyen, Van Diem** (2015) Essays on Takeovers and Executive Compensation
222. **Tolonen, Anja** (2015) Mining Booms in Africa and Local Welfare Effects: Labor Markets, Women's Empowerment and Criminality
223. **Hassen, Sied** (2015) On the Adoption and Dis-adoption of Household Energy and Farm Technologies
224. **Moursli, Mohamed-Reda** (2015) Corporate Governance and the Design of Board of Directors
225. **Borcan, Oana** (2015) Economic Determinants and Consequences of Political Institutions
226. **Ruhinduka, Remidius Denis** (2015) Essays on Field Experiments and Impact Evaluation
227. **Persson, Emil** (2016) Essays on Behavioral and Experimental Economics, Cooperation, Emotions and Health.
228. **Martinangeli, Andrea** (2017) Bitter divisions: inequality, identity and cooperation
229. **Björk, Lisa** (2017) Essays on Behavioral Economics and Fisheries: Coordination and Cooperation
230. **Chegere, Martin Julius** (2017) Post-Harvest Losses, Intimate Partner Violence and Food Security in Tanzania
231. **Villalobos-Fiatt, Laura** (2017) Essays on forest conservation policies, weather and school attendance
232. **Yi, Yuanyuan** (2017) Incentives and Forest Reform: Evidence from China
233. **Kurz, Verena** (2017) Essays on behavioral economics: Nudges, food consumption and procedural fairness
234. **Yashodha, Yashodha** (2017) Contract Choice and Trust in Informal Groundwater Markets
235. **Meles, Tensay Hadus** (2017) Power Outages, Increasing Block Tariffs and Billing Knowledge
236. **Mühlrad, Hanna** (2018) The Impact of Reproductive and Birth Technologies on Health, Fertility and Labor Outcomes
237. **Gakii Gatua, Josephine** (2018) Primary Health Care Interventions and Social Ties in Kenya
238. **Felgendreher, Simon** (2018) Essays on behavioral economics and the effects of the colonial rule on Java
239. **Demeke, Eyoual** (2019) Essays on Environmental and Behavioral Economics
240. **Kiss, Tamas** (2019) Predictability in Equity Markets: Estimation and Inference
241. **Khomenko, Maksym** (2019) Essays on the Design of Public Policies and . Regulations
242. **Mukanjari, Samson** (2019) Climate Policy and Financial Markets
243. **Schürz, Simon** (2020) Economic and Intergenerational Decision-Making in Families
244. **Lau, Debbie** (2020) Empirical Essays on Education and Health Policy Evaluation
245. **Ho, Hoang-Anh** (2020) Essays on Culture, Institutions, and Economic Development
246. **Rubio Ramos, Melissa** (2020) The Economics of Coercive Institutions, Conflict, and Development
247. **Lindahl, Anna** (2021) Empirical tests of exchange rate and stock return models
248. **Sjöholm, Carolin** (2021) Essays on public policy in the informal sector context
249. **Tian, Ruijie** (2021) Impacts of Climate Policy and Natural Disasters: Evidence from China
250. **Larsson, Sebastian** (2021) Providers and Profiteers: Essays on Profits and Competition in the Provision of Public Services
251. **Norrgren, Lisa** (2022) Carpe Diem or Seize your Health? The Economics of Time Preferences, Health, and Education
252. **Andersson, Lina** (2022) Emotions in Game Theory: Fear, friendliness and hostility
253. **Enlund, Jakob** (2022) Three essays in applied economics
254. **Berggren, Andrea** (2023) Choices among Doctors, Students and Primary Care Providers: Empirical Evidence from Sweden
255. **Hansson, Magnus** (2023) Decentralized Finance and Central Bank Communication