ARBITRAGE OPPORTUNITIES ON DECENTRILIZED EXCHANGES: APPLICATION OF GRAPH NEURAL NETWORKS

by

Vadym Pakholchuk

A thesis submitted in partial fulfillment of the requirements for the degree of

MA in Business and Financial Economics

Kyiv School of Economics

2024

Thesis Supervisor: ______________________________ Professor Olesia Verchenko

Approved by ____
Head of the KSE Defense Committee, Professor [Type surname, name]

____

____

____

Date ______________________________

ACKNOWLEDGMENTS

The author wishes to express gratitude to my thesis advisor, professor Olesia Verchenko, who helped me on this difficult road with support, advice and inspiration. Also, I am deeply grateful to Professor Elena Besedina for her strict guidance during the business research workshops.

I would like to express my gratitude to the teachers who worked closely with us all this time and fought for our knowledge.

I would also like to express special gratitude to the KSE Foundation for their help and financial support throughout my studies at the Kyiv School of Economics, and

Also, special thanks to my family, friends, and colleagues who supported me all this time and thanks to whom I was able to reach the finish line.

TABLE OF CONTENTS

LIST OF FIGURES

Figure 18 depicts that LightGBM performs much better than Logistic Regression according to the ROC-AUC score. Both models are better than the random classifier as the area under the curve is greater than 0.5 for regression or LightGBM.

LIST OF TABLES

LIST OF ABBREVIATIONS

AMM Automated Market Makers

CEX Centralized Exchange

CPMM Constant Product Market Making

DeFi Decentralized Finance

DEX Decentralized Exchange

GNN Graph Neural Network

PoS Proof of Stake

PoW Proof of Work

CHAPTER 1. INTRODUCTION

> "…I believe the next generation for markets, the next generation for securities will be tokenization of securities…"
>
> Lary Fink CEO, BlackRock

The modern world is evolving and changing at a fascinating pace, giving rise to new ideas, technologies, and even industries. Although a few years ago, many prominent economists, investors, and financial professionals neglected the new era of modern finance, today, many of them have changed their point of view. And the quote from Larry Fink, CEO of BlackRock, where I started this proposal, is just one example of such behavior. The reason is simple. Because decentralized finance (DeFi), with its blockchain idea, is not about currency. It was about a technology that could level many processes in traditional finance. DeFi is trying to solve many problems in conventional systems: centralized control, limited access, inefficiency, lack of interoperability, and transparency.

From a very interesting perspective, it was only natural that blockchain technology would appear in the decentralized exchanges (DEXs) of automated market makers (AMMs). The existing capabilities of such open platforms allow the creation of a fully automated market based on the execution of smart contracts. This opens up new opportunities to use the same methods for any amount of money, from one to millions of dollars, while maintaining a high level of security and operational speed without any lack of transparency. Unlike traditional securities markets, market participants do not trade through the order book. DEXs allow us to eliminate unnecessary intermediaries and significantly reduce commissions and transaction costs. This is a completely new experience that is attracting a lot of attention from modern researchers, professional investors, and enthusiasts.

However, there are some open questions about arbitrage opportunities in DEXs. One might logically assume that there is some mispricing in different liquidity pools for some tokens. However, it is important to note that there are different types of arbitrage opportunities and various methods to uncover them. As a result, this research is relevant from both an academic and an industry perspective.

In summary, there are several important factors that make this topic particularly interesting:

(i) Fast-developing DeFi industry, which is challenging traditional financial systems;

(ii) The emergence of new financial instruments and infrastructures in DeFi;

(iii) As for a completely new area of modern finance, there is still room for interesting research topics;

This research aims to identify the complex cycle arbitrage opportunities, represent the market state at each timestamp as a graph network, and transform it into embeddings. They could be used in different machine learning models to predict arbitrage and improve existing approaches based on heuristics and brute force, where some market inefficiencies may exist.

One hypothesis that could be tested during the research is that the market state could be represented as a complex and complicated network. And this state could be transformed into embeddings to predict the arbitrage.

CHAPTER 2. INDUSTRY OVERVIEW AND RELATED STUDIES

2.1. Industry overview

The cryptocurrency industry is growing every year. This is evidenced by official statistics from major market players and various analytical institutions. For example, according to Statista Market Insights, the number of active users in the crypto industry will grow from 30.45 million to 670.50 million between 2017 and 2023. For example, Table 1 shows increasing revenue from new investments in the crypto industry.

Table 1. Cryptocurrencies Revenue in Billion USD (US$)

	2017	2018	2019	2020	2021	2022	2023
Revenue	1.6	2.4	1.7	7.3	35.5	20.1	40.7
Change		51.9	-29.6	336.8	386.7	-43.2	102.2

Source: Statista Market Insights

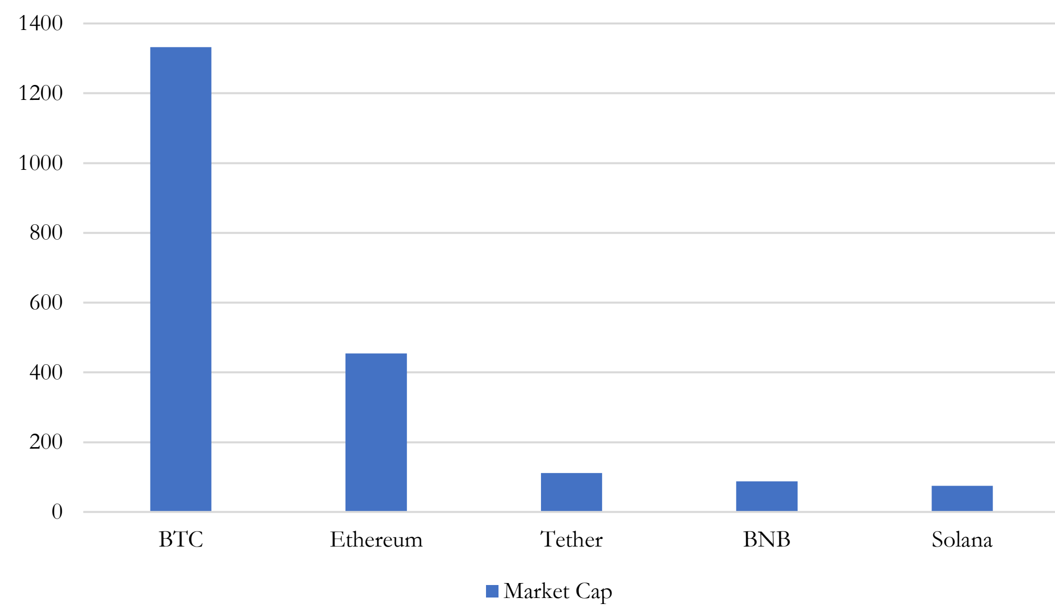
According to Statista (2024) expert estimates, the total market size currently exceeds USD 40 billion. Analysts expect the market to grow at a CAGR of at least 7.77% until 2029.

According to CoinMarketCap statistics as of June 2024, the top cryptocurrencies by market capitalization are Bitcoin, Ethereum, Tether, BNB, and Solana[1]. The top five cryptocurrencies are often used to analyze market capitalization. Market cap statistics show that despite the more advanced technology of Ethereum, Bitcoin is still in first place. This popularity can be explained by the time it has been on the market and the periodic reduction in supply. The so-called halving increases interest in Bitcoin as to the deficit good

[1] *Cryptocurrency Prices, Charts And Market Capitalizations | CoinMarketCap*. (2024). CoinMarketCap. https://coinmarketcap.com/

and way of storing value. Nowadays, a lot of investors trying to diversify their portfolio could include not only gold and other assets but also Bitcoin based on the consideration above.

Figure 1. Market Capitalization of top 5 cryptocurrencies in billion USD (US$)



Source: CoinMarketCap

The rapid development of the crypto industry is significantly influenced by its unregulated nature around the world. Although some countries are starting to introduce legislation, there are still economic zones that are free of laws. However, another important reason is technology, which has changed the world. The use of distributed computing systems to confirm transactions over the network has created a virtually self-regulating market. It enables cross-border payments with minimal fees for transfers and currency exchange.

To identify the factors that influence the development of the industry, it is worth applying the STEEP/PESTLE analysis framework. It includes political and legal, economic, environmental, and technological components. The results of the analysis are presented in Table 2.

Table 2. PESTLE analysis of the DeFi industry

Driving Force		Reason	Impact
P	Political	World instability and political concerns	Political instability can push users to use DeFi for hidden cross-border transactions.
E	Economical	Drawbacks of traditional finance and search for high returns. Some agents try to avoid sanctions.	Increasing popularity and interest in the DeFi assets
S	Social	Change in people's lifestyles and increased mobility	Strong diffusion of DeFi into modern social life
T	Technological	Recent advancements in technological progress	Improvement of security and emergence of new products and instruments
L	Legal	Unregulated in a lot of countries	Develop a more powerful regulation system and share information across countries.
E	Environmental	High energy consumption and greenhouse effect	Implementation of more energy-efficient blockchain technologies can reduce the environmental impact of DeFi.

Source: Author's own elaboration

First of all, it is worth noting that the development of the decentralized finance industry is heavily influenced by political factors. Political instability in many countries and uncertainty about the future push people to look for alternative solutions to minimize such risks. Changes in the political regime may result in changes in the conditions for capital outflows and, in some cases, as practice shows, in changes in the monetary policy of central banks. For example, restrictions imposed by the National Bank of Ukraine have narrowed the range of instruments available for Ukrainians to invest abroad and conduct international transactions. However, this is an example of a critical need rather than a general rule.

Another important factor in this situation is economic incentives. The disadvantages of traditional finance and the search for ways to optimize transaction costs during transfers have become a catalyst for growing interest in non-traditional finance. Modern blockchain protocols promise users minimal fees for transfers regardless of the amount. At the same time, a growing number of derivatives are emerging that offer high yields, non-standard monetization models, and attract new users.

The mobility of society is also increasing, and the diffusion of new technological progress tools into everyday life is growing. Recent innovations in the FinTech industry, such as multi-currency accounts or the integration of several banking systems, are no longer surprising. Today, people are getting used to paying for coffee or a car with cryptocurrency.

Technological advances in recent decades have driven the creation of large-scale speech models, computer vision, and general progress in the IT sector. Significant changes have also occurred in the finance industry. Distributed computing has long been used to process large amounts of data. However, blockchain has become a revolution that has changed the way we think about approaches to building modern, secure financial systems and the speed of processing transactions.

However, significant risks remain in the legal field. Different approaches to the definition of crypto assets in different countries and economic zones create prerequisites for manipulation, fraud, and the use of decentralized finance for criminal purposes.

According to a study by the European Bank (2022), not all cryptocurrencies create significant pressure on the environment, but some cryptocurrencies do. In particular, the use of the Proof of Work[2] algorithm has become a significant factor in the growth of carbon emissions by Bitcoin miners. It is worth noting, however, that concerns about

[2] Proof of Stake – "protocols are a class of consensus mechanisms for blockchains that work by selecting validators in proportion to their quantity of holdings in the associated cryptocurrency" .

negative environmental impacts have stimulated the search for new, more efficient network protocols. Proof of Stake[3] is currently one of them.

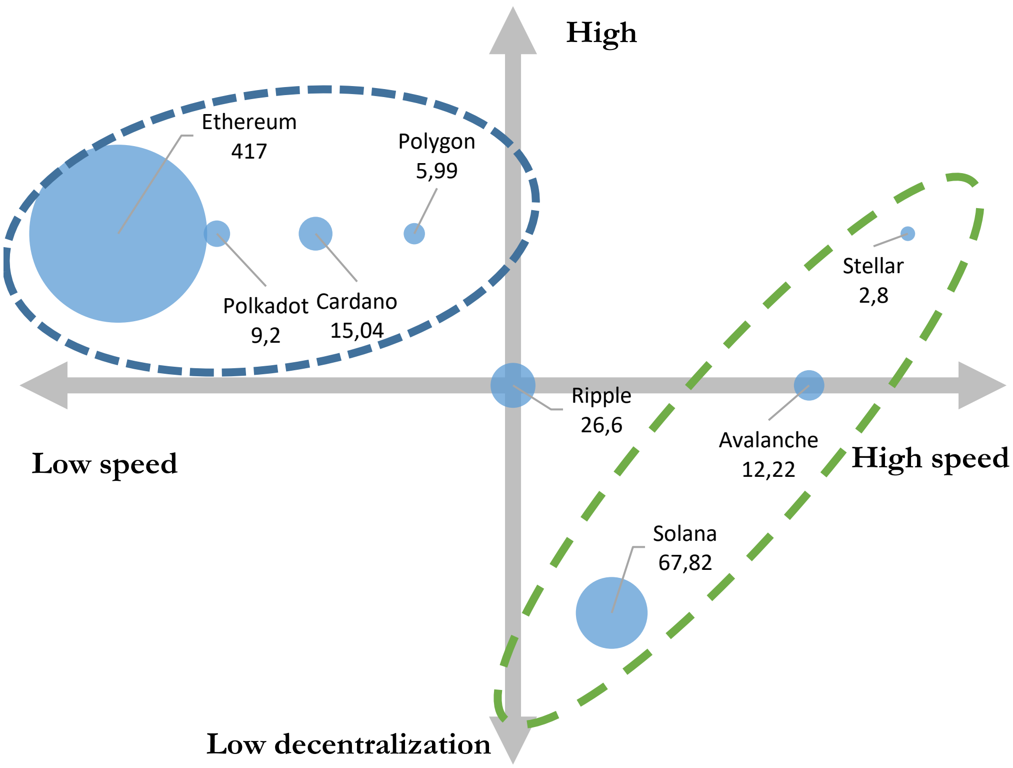
The shortcomings of Bitcoin have become the catalyst for the emergence of a significant number of cryptocurrency derivatives. Some have tried to solve problems related to payment efficiency, such as Ripple. Another class of currencies is trying to recreate traditional payment methods, such as Monero. It is worth noting that today, a huge number of new cryptocurrencies and technologies are emerging in the DeFi system.

Recently, more and more protocols in the blockchain system have been appearing every year. Each new protocol tries to solve the problems of the previous ones. Using the data on the protocols' speed, the level of decentralization of the system, and the market capitalization, we constructed the matrix shown in Figure 2.

The positional matrix shows we can identify a cluster of fairly similar protocols with high decentralization and low speed. Bitcoin is excluded from the matrix as it is not used in the development of modern DEXs and DeFi. The only purpose is storing value. The punctures in this cluster differ in speed and gas fee. New second- and third-generation networks dominate in speed but lag far behind in market capitalization. The positioning of Solana looks interesting on this matrix, as it falls out of the general pattern due to the low level of decentralization of the system. However, perhaps the most significant conclusion that can be drawn is the undisputed dominance of Ethereum in terms of market capitalization, despite all the inherent shortcomings of the system with a slow network and high gas fee.

[3] Proof of Work - "is a form of cryptographic proof in which one party (the prover) proves to others (the verifiers) that a certain amount of a specific computational effort has been expended".

Figure 2. Positioning map for popular chains

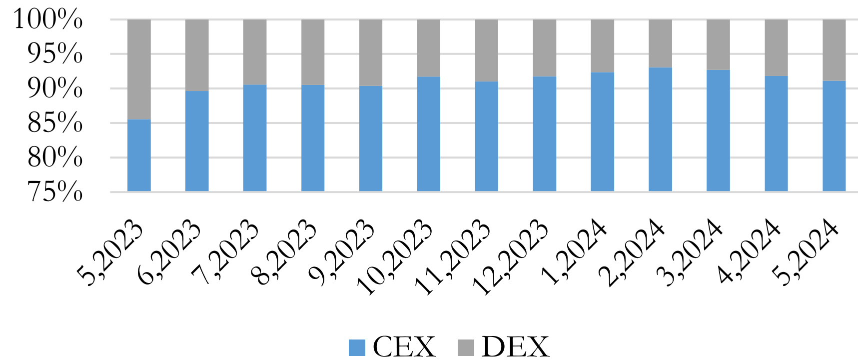


Source: CoinMarketCap

There are several exchanges on the market for crypto transactions. The largest players in the CEX market, according to CoinGecko, are Binance, Bybit, Gate.io, OKX, and Coinbase, ranked by Trust Score (Top Crypto Exchanges Ranked by Trust Score | CoinGecko, 2024). The same aggregator shows that among the top Decentralized Exchanges based on market share by trading volume are ThrustedV3, BlasterSwap, Uniswap V3, and Orca.

Another interesting feature of the modern decentralized finance industry is the popularity of decentralized exchanges. Although the volume of trading on centralized exchanges is still significant, regulatory pressure is forcing many agents to switch to trading on decentralized exchanges.

Figure 3. DEX to CEX Trade Volume (%)

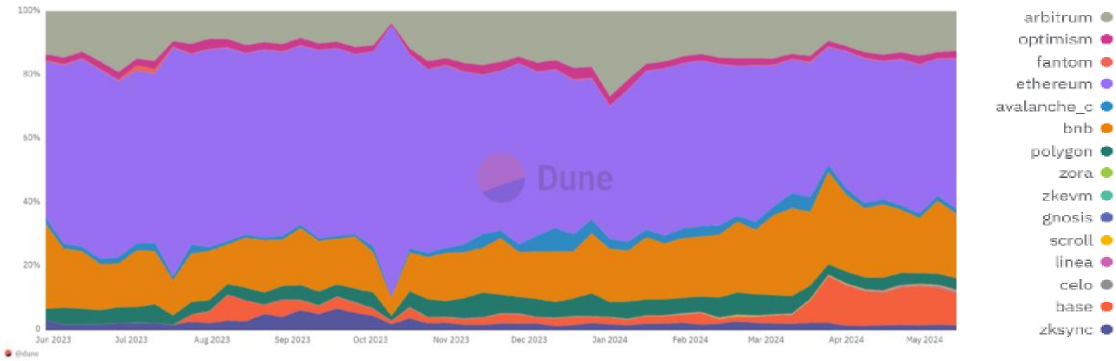


Source: The Block

An important feature of decentralized exchanges is the ability to transact without third-party intervention and to be confident that if a centralized exchange is hacked, no one can get hold of your data. Reports of even large exchanges being hacked appear in the media from time to time, adding to the interest in DEX. The DEX system is designed so that the exchange rate is set automatically by an automated market maker (AMM) system, which uses pre-defined algorithms to adjust the price based on the liquidity of a particular exchange. Today, there are many decentralized exchanges that offer their services to users and differ in their pricing algorithms.

According to Dune Analytics, DeFi lama conducts most of its operations based on Ethereum. At the same time, according to the block aggregator, the most popular exchanges have long been Uniswap, SushiSwap, Pancakeswap, Curve, and Balancer. The architecture of the exchanges does not involve depositing assets in the exchanges' wallets. Instead, users use a smart contract system to interact with the DEX infrastructure to place orders.

Figure 4. Weekly DEX volume by chain (%)



Source: The Dune DEX Volume by Chain

DEX pricing mechanisms use automated market makers (AMMs). This is based on organizing the interaction between several participants. Firstly, these are liquidity providers who typically stack assets in order to receive remuneration. This allows traders to exchange cryptocurrencies for a fee they pay to the liquidity providers.

As most decentralized exchanges use Constant-function market makers (CFMM) [4], there will always be deviations from the fair market price within the exchange. However, the price is quickly brought back to normal by arbitrageurs, who trade to make money but ensure the market remains efficient.

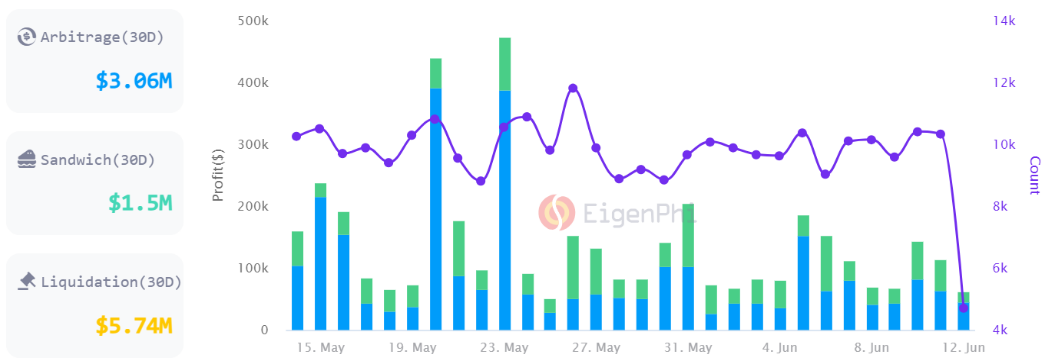
At the same time, the issue that concerns all market participants and still needs to be addressed is the security of transactions and protection against maximum extractable value (MEV)[5] Arbitrage is specific to the decentralized financial system and reduces the remuneration of transaction participants (Harvey et al., 2020).

---

[4] Constant-function market makers (CFMM) – "are a paradigm in the design of trading venues where a trading function and a set of rules determine how liquidity takers (LTs) and liquidity providers (LPs) interact, and how markets are cleared. The trading function is deterministic and known to all market participants".

[5] maximum extractable value (MEV) – "is the maximum amount of value miners or network validators can extract by rearranging and reordering transactions waiting for confirmation"

Figure 5. Performance of MEV Types excluding Liquidations (Mil. Dollars)



Source: EigenPhi

Market data indicates that even now, arbitrage opportunities exist (MEV Data | EigenPhi | Wisdom of DeFi, May 2024). According to EigenPhi statistics, as of June 2024, the last 30 days on the Ethereum chain brought traders 3.06 mil. dollars of profit by arbitrage, identified as from trading cycles on Etherscan.io. On the other hand, sandwiching profit[6] was about 1.5 m. dollars. An even larger amount corresponds to liquidation[7]. However, during an analysis of the previous periods, it was determined that this is more like an outlier.

Summarizing the review of the industry statistics, we can conclude that there is still a prospect for arbitrage operations in the decentralized finance system based on the relevant frequency of their existence. It should be noted that the role of arbitrageurs is crucial for maintaining the fair value of crypto assets. It is thanks to their activities that the system manages to remain fully self-regulated according to the foundations of DEXes functioning and attract new users. Leading investment funds are now joining the DeFi trend in various forms: VC for new projects and big investment funds like BlackRock

[6] Sandwiching – "is a form of market manipulation on Decentralized Exchanges (DEXs). It involves a malicious actor who identifies a large pending transaction and then carefully places two transactions around it: one before and the other after the targeted transaction"

[7] Liquidation – "is the process of forcibly closing a trader's positions in the cryptocurrency market".

founding new specialized ETH. The year 2024 is a defining year for the industry as a whole, marking the launch of specialized ETFs. However, we can expect the trend to continue in the coming years, largely due to DeFi's technological advantages over traditional finance.

2.2. Related studies

The development of DeFi attracts the attention of numerous researchers in this area. Starting from the historical publication of Nakamoto (2008) about introducing Bitcoin concepts, the literature describing the industry expanded enormously. A lot of good research questions have been discussed in recent years. Some of them were about blockchain technology, the development of payment systems, security concerns, specifically the block mining process, etc. But still, there are many questions we could find in the field. We have to investigate the related studies to find insights into what was done and what could be improved, bringing something new to the literature.

For example, some interesting and non-trivial research was done by Avarikioti et al. (2020). The authors describe the perspectives of new protocols in blockchain systems from the game theory perspective. They prove the positive impact of such technologies on achieving the social optimum.

The idea of applying the game theory methodology is not new. Eyal and Sirer (2014) tried to see the miner's dilemma from the same perspective. According to the research, in order to operate efficiently, the blockchain system requires a significant amount of computational resources. In such a situation, the logical consequence is the emergence of large pools where participants can share resources and benefit from them. However, this situation could also be described as a normal form of game in which participants could cooperate or defect. The article describes several combinations of players and pools and, as a result, suggests that emerging pools are natural and beneficial for the industry. They also emphasized the disadvantages of the defecting strategy, especially in the long run.

The logical development of this article is the work of Kiayias et al. (2016), in which researchers described two forms of the preview game but narrowed it down to small players and perfect vs imperfect information about the block release and next starting point of the pool. The conclusion of the research proves the previous results from the game theory perspective, as the more computational power you have, the more incentives for participants to deviate from the cooperative strategy.

The more recent research by Koutsoupias et al. (2019) investigates the same idea about small players in blockchain networks and how they interact with big pools. However, the overall result remains the same for even small miners; the best response strategy to any other strategy of the big players is to play cooperatively.

Economic perspectives on the problem were described by Huberman et al. (2020) and Easley et al. (2019). The first paper compares the traditional and DeFi financial systems in terms of fees, speed, and monopolization. In conclusion, they derive the theoretical foundations of how to improve the fee efficiency in the Blockchain system. The same fee problem was investigated by Easley et al. (2019) from the game theory perspective. They showed how externalities affected the fees and derived an empirical model explaining the size of the fee depending on the market state.

As we suggested in the industry overview section, one of the driving forces of the DeFi is technological advancements. Nayak et al. (2016) reveal some concerns about selfish-mining. This idea describes how miners could apply attacks in blockchain networks for their gains. As a result, they showed how agents could manipulate the network, revealing the lack of security. Liu et al. (2018) further developed this area of the research. They investigate the behaviors of miners in the context of coordinated attacks by multiple actors.

The growing emphasis on the security question brings a lot of attention to this topic in the research areas. For example, Kwon et al. (2017) and Daian et al. (2020) showed the importance of security threats for blockchain, especially for decentralized exchanges. The last was a groundbreaking research that changed the world of DeFi forever. This research brings light to specific instruments traders are using to exploit vulnerabilities of the chain protocols and extract maximum value from trade operations. Today, these instruments are known as frontrunning and sandwiching attacks. The other very important result of this research is the comparison of some trading techniques from traditional finance.

A very comprehensive literature overview was conducted by Fangl et al. (2022). The authors organized 146 research papers on cryptocurrency trading, building a foundation for future researchers to understand the current state and platform.

One of the foundation papers in the area of arbitration is definitely the work of Makarov and Schoar (2020). They investigate the price deviation of Bitcoin among several crypto exchanges in different countries. The emphasis of the paper is on the significant impact of capital regulation on arbitrage effectiveness. The important result of the research was the decomposition of the price movements into common and idiosyncratic components to explain price deviations. The same issue was investigated by Kristoufek and Bouri (2022) later, who tried to identify the sources of arbitrage opportunities for Bitcoin on centralized exchanges. Interestingly, they used the Grey correlation in a simple autoregressive process of the first order to measure the arbitrage opportunities. To identify the most important factors, they used economic and blockchain features.

Grimberg et al. (2020) explore the triangular arbitrage[8] on centralized exchanges. They used the Kaiko order book dataset as a basis and enriched it with Binance information. The research narrowed to several anchor tokens such as BTC, BNB, ALTS, etc. The results indicate that almost 3% of the total transactions on Binance were performed by bots and correspond to triangular arbitrage.

[8] "Triangular arbitrage is a strategy where you find price discrepancies between three currencies and buy and sell them in a specific order to make a profit".

Zhou et al. (2021) conducted an analysis of high-frequency trading on decentralized exchanges. They explored the Uniswap data and analyzed the probability of sandwiching attacks depending on the external event. The value of the research is also in a very detailed explanation and investigation of DEXes microstructure, how AMM operates, and why the slippage occurs. Other interesting evidence about the impact of CEX security system breaches was introduced by Aspris et al. (2021).

However, when talking about how DEX CMPP operates, it is necessary to mention the work by Mohan (2022) exploring the pricing model for different models of AMM. The author gives important insights from the geometric perspective about how the AMM DEX exchanges operate and why the arbitrage could occur. This is a result of the liquidity slippage on DEX when traders add or remove tokens from the pools.

Wang et al. (2022) suggest ideas on how to create new features to identify arbitrage opportunities on DEXes. They give insights on how to identify and measure profitable arbitrage on exchanges. The work is based on the ETH network protocol and is complemented by a lot of quantitative conclusions. The main result of their research is that DEXes are highly inefficient and show a lot of arbitrage opportunities. But a lot of them are realized by a small number of participants.

Zhang & Wang (2023) tried to model arbitrage in decentralized exchanges with deep reinforcement learning. They combined assumptions about the market microstructure from the game's theoretical perspective and described market agents and their set of strategies how they influence the liquidity and exchange rates of liquidity pools. In this fundamental work, researchers provide valuable insights into the application of zero intelligence, moving averages, and two-point arbitrage for different levels of time aggregation. The experiment was conducted for the eight most popular currencies. The interesting conclusion is that even on a daily level, some agents could achieve a positive profit from arbitrage transactions.

This literature review is not completely exhaustive but gives at least a short overview of the current trends and directions of the recent research in blockchain, DeFi, and arbitrage. The majority of the research is represented from the game theoretic perspective. There is a very small amount of modern papers about arbitrage. It is necessary to emphasize the increasing interest in decentralized exchanges. The investigation of related studies gave us significant insights into methodologies and data sources but highlighted the lack of research in this area. This work brings new perspectives to academic discussions trying to extent work of Zhang & Wang (2023), and implement new methods with graph neural networks and graph embeddings to predict arbitrage opportunities. The results could be used by either market professionals or academic researchers.

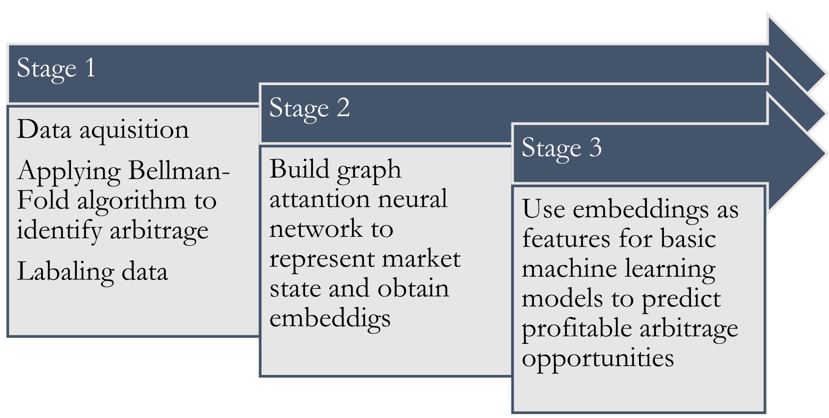
CHAPTER 3. METHODOLOGY

Recent studies focusing on arbitrage opportunities in the DEX market have primarily focused on simple statistical arbitrage across different exchanges. This could be explained as atomic arbitrage. This study will focus on atomic arbitration, as it can be carried out with almost no risk if the terms of smart contracts are properly organized. The advantage of atomic arbitration is that if a transaction fails to execute, it simply does not execute, so there is no profit but no loss.

The increasing popularity of arbitrage could be explained by the emergence of flash loans. Flash loans, which replicate traditional margin trading, appear to be a new window of opportunity in this context. However, they can make trading safer. When you submit a transaction to smart contact, all conditions are checked and if any condition is not met, the loan will not be granted. This reduces the risks for both the trader and the lender.

However, even taking into account the potential opportunities offered by the decentralized financial infrastructure, finding arbitrage opportunities remains a non-trivial task. In our study, we propose to test machine learning algorithms for finding signal-based arbitrage opportunities.

Figure 6. General view of research three stage methodological framework



Source: Author’s own elaboration

3.1. Arbitrage identification

To start with, it is necessary to explain the methodological foundations of using a constant market pricing model (CPMM) by automated market makers on decentralized exchanges. The overall CPMM logic will be explained above. This will give us an idea of why the price discrepancies occur and how we are going to apply the constant pricing model in triangular arbitrage. The CPMM is very popular among different forks of the Uniswap V2[9] because of its simplicity and specifically based on arbitrage opportunities. Centralized exchanges are dealing with almost infinite liquidity. However, the DEX operates using liquidity pools. To maintain the exchange rate, the following equation is used:

$$x \times y = k \tag{1}$$

In equation (1) x represents the amount of token USDT and y of token ETH in the specified liquidity pool. k is a constant invariant according to DEX's functioning rules. Consider the possible situation of selling token USDT with the amount dy and buying token ETH with the amount dx. This operation will result in a new combination of quantities. They could be computed using Equation (2) and in Figure 7.

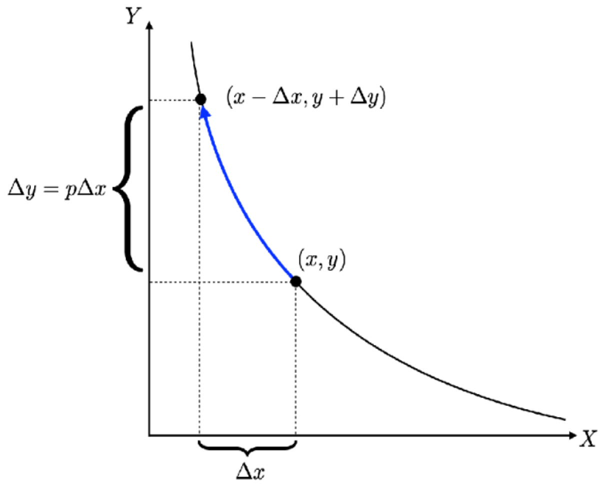
$$(x - dx) \times (y + dy) = k \tag{2}$$

Now we could introduce the next step and sell received tokens of ETH on the other exchange at a better rate. In this case, the arbitrage profit will be defined as the following function:

$$F(dy_A) = dy_B - dy_A \tag{3}$$

[9] Uniswap V2 is a decentralized cryptocurrency exchange

Figure 7. Constant Product Market Makers on DEX



Source: Aoyagi and Ito (2021).

To explain the overall process logic, let's consider two decentralized exchanges: DEX A and DEX B. Where we define dy_B as the number of tokens sold on DEX B and dy_A as the amount we bought on DEX A. The optimal solution to this problem could be found by solving the first-order condition and is represented in equation (4):

$$dy_A^* = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{4}$$

Where we define variables a, b, c as below and derivation is presented in the Appendix:

$$a = k^2 \tag{5}$$

$$b = 2kY_AX_B \tag{6}$$

$$c = (Y_AX_B) - (1 - f)^2X_AY_AX_BY_B \tag{7}$$

$$k = (1-f)X_B + (1-f)^2 X_A \tag{8}$$

The notation represents the following:

$f \in [0; 1]$ – swap fee on different DEXes;

X_A – reserve out on DEX A (removing token liquidity from the pool);

Y_A – reserve in on DEX A (adding new token liquidity to the pool);

X_B – reserve out on DEX A;

Y_B – reserve in on DEX A;

So, we found out how to calculate the optimal quantity to execute the identified arbitrage opportunity, but we still have to establish the algorithm for finding such opportunities.

The most well-known solution for this problem is using the Bellman-Ford algorithm. The following method gives an opportunity to identify arbitrage opportunities without implying any optimization solutions about the size of the buy/sell orders, etc. The final cycle price is calculated as the multiplication of all token prices. It should be greater than 1. The reason for that is that we need the output amount to be greater than the input. Let's consider that we have at least three exchange rates for several tokens: USDT/ETH, ETH/USDC, and USDC/USDT. In case of multiplication of these token prices for $1, the result will be greater than 1; we could assume the existence of an arbitrage opportunity.

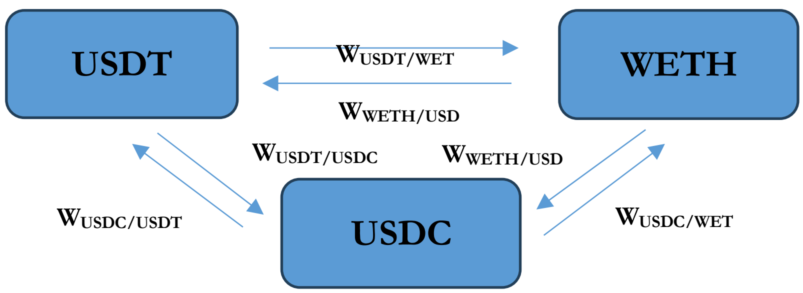
$$P_{\frac{USDT}{ETH}} \times P_{\frac{ETH}{USDC}} \times P_{\frac{USDC}{USDT}} > 1 \tag{9}$$

$$\ln\left(P_{\frac{USDT}{WETH}}\right) + \ln\left(P_{\frac{WETH}{USDC}}\right) + \ln\left(P_{\frac{USDC}{USDT}}\right) < 0 \tag{10}$$

The whole cycle could be modeled as a graph, where graph nodes are tokens, and different exchange rates will be represented as edges. The quantitative measure of the edges should be defined as weights (components) from the equation (9). We could even modify it to include fees. As a result, the edge weights will be defined as (11):

$$Weight_{\frac{token1}{token2}} = \ln\left(P_{\frac{token1}{token2}} - fee\right) \tag{11}$$

Figure 8. Example of Bellman-Ford cycle



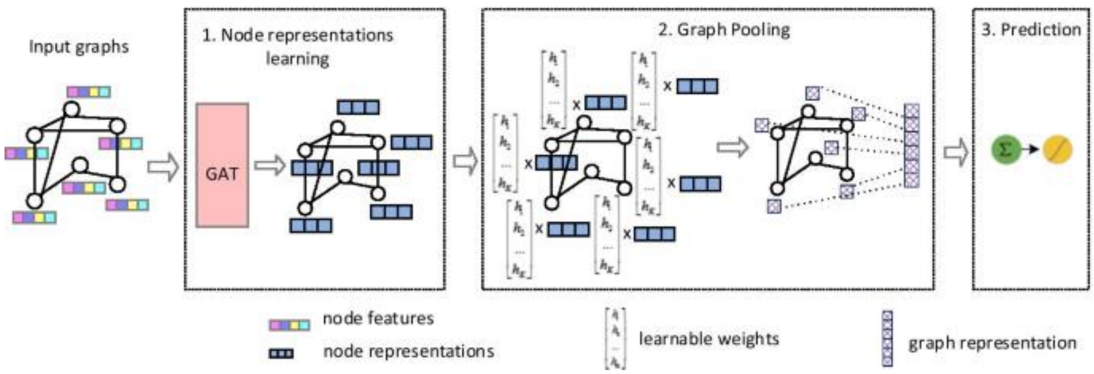
Source: Author's own elaboration.

An example state of the market represented in the graph is shown in Figure 8. The Bellman-Ford method is usually used to identify arbitrage. In this method, the algorithm searches the most negative cycle in the graph as the sum of the logarithm of its weights. This step should be applied to build a graph for each DEX order book on each timestamp.

3.2. Building graph Neural Network

To represent each state of the market, an embedding Graph Neural Network (GNN) could be used. GNN is one of the deep learning methods used to analyze big unstructured data and represent some entities and their relations. The biggest problem in standard GNN is that the basic feature map is constructed from node characteristics. But today, researchers could try to experiment with Graph Attention Networks (GAT). GAT, despite ordinary features, includes edge attributes as additional features. For the goal of market state representation, where almost all information is represented as exchange rates and trade volumes, GAT could become the best solution. Figure 8 shows how typical GAT looks. The network from input graph node and edge features learns graph representation. The learned weight for nodes and edges is fed into the pooling layer and represented as one graph embedding. For prediction, simple fully connected layers with a softmax activation function could be used.

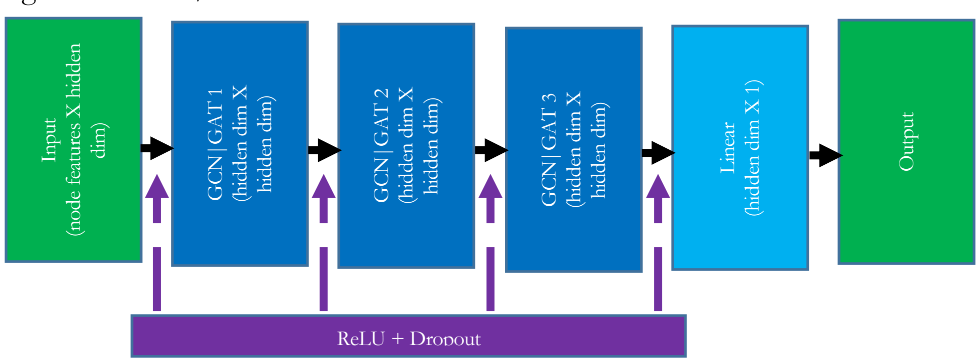
Figure 9. Graph Neural Network (GAT) architecture example



Source: Hu et al. (2021)

Based on the theoretical foundation described above, the proposed model was built using a Graph Convolutional Network (GCN) as the backbone (Figure 9).

Figure 10. GCN/GAT network architecture



Source: Author's own elaboration

Figure 10 depicts that three Graph Convolutional layers with ReLU activation function were used. The output was built as a linear multilayer perceptron with a sigmoid activation function. As a regularization method, dropout was applied with a probability of 30%. Mean pooling was used before pushing results to output.

3.3. Prediction of the Arbitrage

The output embeddings were used in the logistic regression classifier and gradient boosting classifier. One of the most important questions in the experiment is to set up key metrics. For this task, such metrics as accuracy, precision, recall, F1 (mean harmonized for precision and recall), and ROC-AUC curve. True and false positive/negative values from the classifier are used to calculate this metric.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

The ROC-AUC curve is widely used in the assessment of binary classifiers. In general, it represents how much better our model is in comparison to any random model.

CHAPTER 4. DATA

Nowadays, it is quite easy to become totally lost with a large number of new projects in DeFi. Even the Trading Strategy site contains data about more than five thousand DEXes. To find the total quantity of trading pairs, we have to multiply this number by hundreds. These statistics make us narrow our research to just several trading pairs.

Exchanges that carry the highest daily trading volume are very interesting for our research as they correspond to high liquidity. The other important thing we must consider is the data availability for different DEXes for different timestamps. Could we retrieve it using any API requests?

The simplicity of technical implementation is also an important factor. Finally, some exchanges offer a unique number of trading pairs, depending on the protocol used. During the analytical derivation, we have emphasized the role of fees, but here it remains crucial again, as we should choose exchanges with the lowest ones.

Almost all modern DEXes use the Ethereum protocol. As a result, all data about transactions is public and transparent. But from an implementation point of view, there are three main sources to acquire such data:

(i) Directly from the protocol network. This method requires knowledge of Solidity and creating a local node of the network;

(ii) Data distributors. This is a more preferable and convenient method, but usually prepaid for the quantity requested;

(iii) Block parsing. This is an open source with a lot of flexibility, but requires a huge amount of time and is very complex and complicated;

For this research, we chose the second solution as the most convenient and affordable. Python SDK from Trading Strategies allows you to get access to historical DEX trading data with a free API key. It uses open-high-low-close-volume data on different time levels: 1m, 5m, 15m, 1h, 4h, daily, weekly, and monthly. Also, the site gives access to data describing liquidity pools with CPMM for the same time levels. Additionally, we could use information about AAVE supply and borrow rates as external data.

Almost 50 GB of trading data and 30 GB were retrieved from the Trading Strategies back-testing data API. The mentioned API gives access to 237464 trading pairs and 5851 DEXs in total. Due to the lack of computational resources, several limitations were applied to narrow the research. First of all, trading pairs were filtered to include tokens with transaction fees less than 30 basic percent points (BPS) and identified 50090. Only 50072 pairs from them were volatile enough. But only 1620 pairs were quoted in stable coin and narrowed to just 196 pairs with enough volume. The top ten best-suited trading pairs are presented in Table 3.

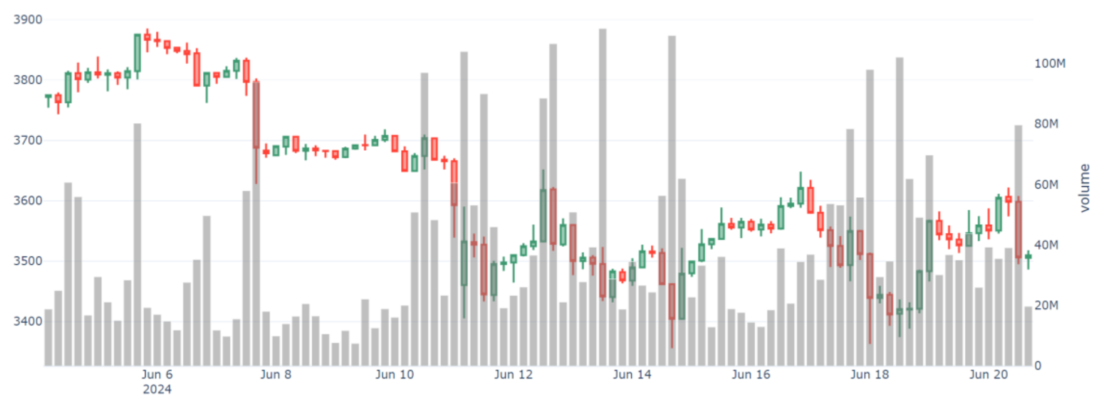
Table 3. Top 10 of the best-suited trading pairs for analysis

Number	DEX	Base Token Symbol	Quote Token Symbol	fee	Buy volume 30 days
1	uniswap-v3	WMATIC	USDT	5	3.977633e+07
2	quickswap	WMATIC	USDC	30	2.071649e+06
3	uniswap-v3	WMATIC	USDC	5	3.805003e+07
4	uniswap-v3	WMATIC	DAI	30	1.743066e+06
5	sushi	WETH	UST	30	1.513046e+06
6	uniswap-v2	WETH	USDT	30	2.158244e+08
7	uniswap-v3	WETH	USDT	5	1.226157e+09
8	uniswap-v2	WETH	USDC	30	2.120530e+08
9	sushi	WETH	USDC	30	1.790754e+07
10	quickswap	WETH	USDC	30	4.798150e+06

Source: Author's own elaboration

The resulting token list included several tokens: SOL, ADA, WBNP, WBTC, WETH, USD, DAI, USDC, BUSD, USDT, TRX, WMANA, and WMATIC. The data format for these tokens was represented as trading candles. So, they included open, close, high, and low prices and trading volume. In Figure 11, only one of the trading pairs is shown. The chart indicates a huge volume of transactions during the period. Also, there is high volatility in the assets. This particular aspect is initially the reason why potential arbitrage opportunities appear in the market.

Figure 11. ETH/USDT price candles with trading volume



Source: Author's own elaboration

Table 4 shows that some trading pairs have a history starting from 2020, while others started trading only in 2021 or 2022. On the other hand, despite the wide range of maximum and minimum timestamps, some tokens have low liquidity and just a few updates in the order books. However, some trading pairs have a lot of transactions on a 1-minute aggregation level. For example, pair ETH-USDC has 4,217,565 records in the dataset, which is a lot.

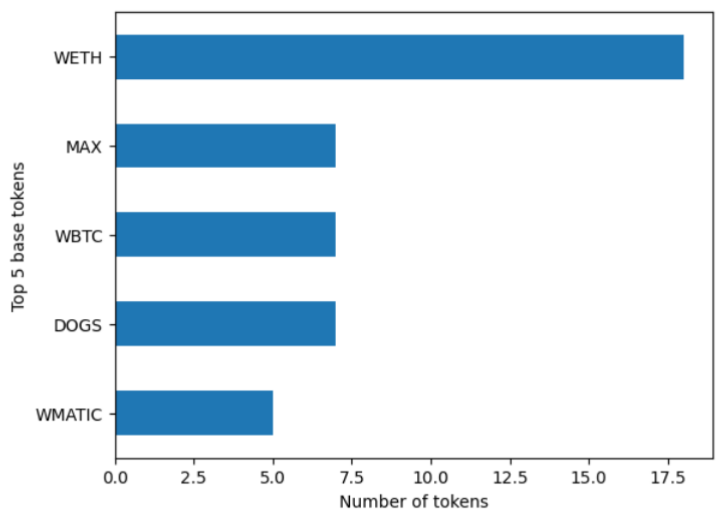
According to the performed exploration data analysis, the most popular base tokens are WEATH, MAX, WBTC, DOGS, and WMATIC, as shown in Figure 12. It is worth mentioning that the most popular exchanges in the dataset are pancakeswap-v2, uniswap-v3, uniswap-v2, sushi, quickswap, and pancakeswap, as shown in Figure 12.

Table 4. Trading pair liquidity and history analysis

Pair	Timestamp			Trading volume, USD		
	min	count	max	min	mean	max
eth-dai	14.05.2020 0:08	1088392	12.08.2024 21:08	0.0005	29482.326	96929600
eth-dai-fee-30	23.09.2021 10:45	149969	13.08.2024 17:52	0.0000	934.722	218741.1
eth-usdc	05.05.2020 21:09	4217565	12.08.2024 21:08	0.0005	21193.898	27747550
eth-usdc-fee-5	31.08.2021 14:15	1284074	13.08.2024 18:01	0.0000	42739.335	10503090
eth-usdt	19.05.2020 9:06	2516946	12.08.2024 21:08	0.0010	24189.975	61941810
eth-usdt-fee-5	06.05.2021 4:55	1397817	13.08.2024 18:00	0.0000	71141.923	31226770
eth-ust	25.12.2020 9:21	60211	11.08.2024 23:26	1.5536	12977.217	5219099
matic-dai-fee-30	21.12.2021 16:20	198673	13.08.2024 17:58	0.0000	105.268	39903.07
matic-usdc	09.10.2020 9:11	1277302	12.08.2024 21:08	0.0005	4785.401	8042628

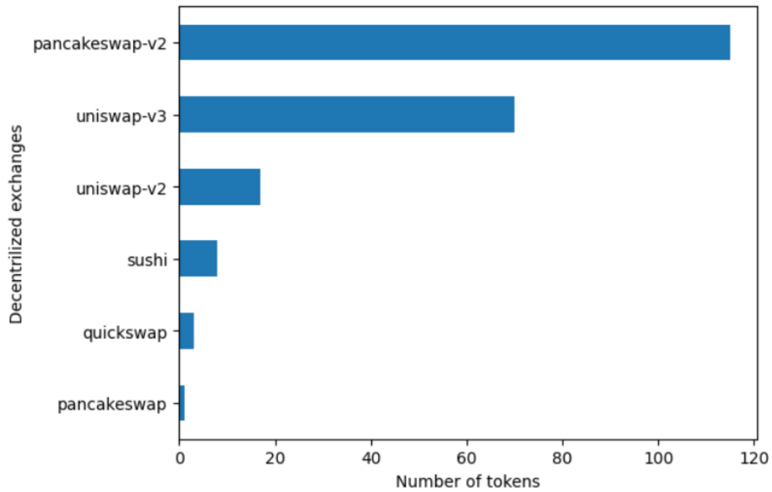
Source: Author's own elaboration

Figure 12. Top 5 tokens in the dataset



Source: Author's own elaboration

Figure 13. Top 5 decentralized exchanges in the dataset



Source: Author's own elaboration

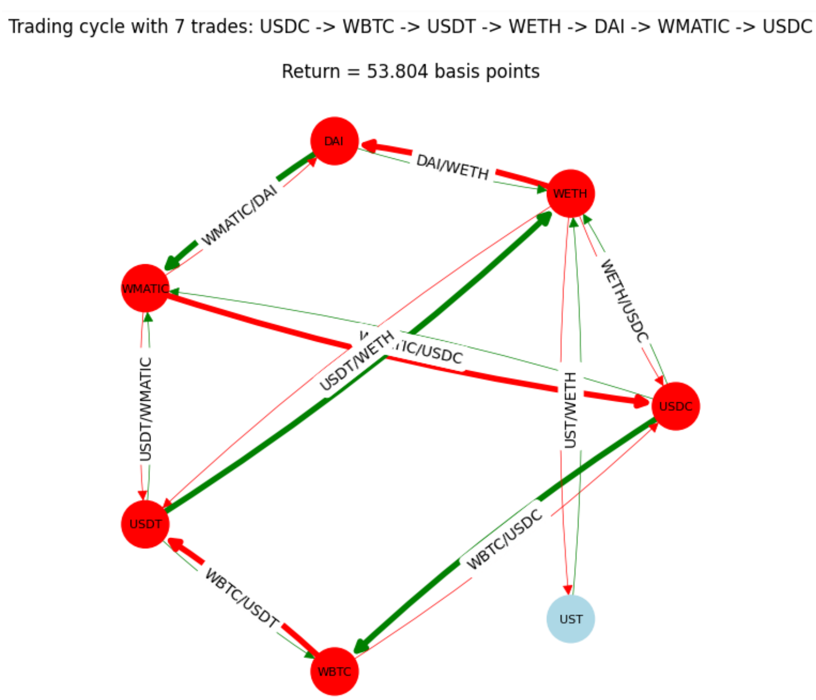
Due to the lack of computational resources and the big size of the dataset, there is a time limit for the modeling part. For arbitrage identification, 29165402 records were used. Even this small amount of data was computed for several hours. Because of that, during modeling, the overall dataset size was decreased to 319304 records with 18361 unique timestamps on a 1m level of aggregation over a two-week period.

The results of the modeling are presented in the next chapter.

5.1. Graph Arbitrage Identification

Bellman-Fold algorithm helps to identify the best arbitrage cycle sequence and return in BPS, as shown in Figure 14. The green arrows represent buying operations, while the red depicts the selling of the tokens. The overall profit of the cycle is 53.8104 BPS, which could be multiplied by 0.01% to transform into percent.

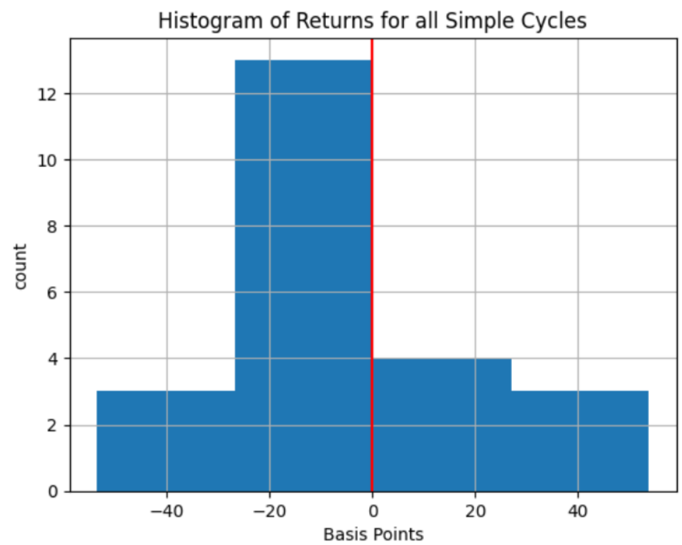
Figure 14. Trading cycle example



Source: Author’s own elaboration

It is necessary to emphasize that not all cycles are profitable and the size of the return is not very significant. As Figure 15 depicts, the mean and median of the distribution are below zero, and the overall distribution is highly right-skewed.

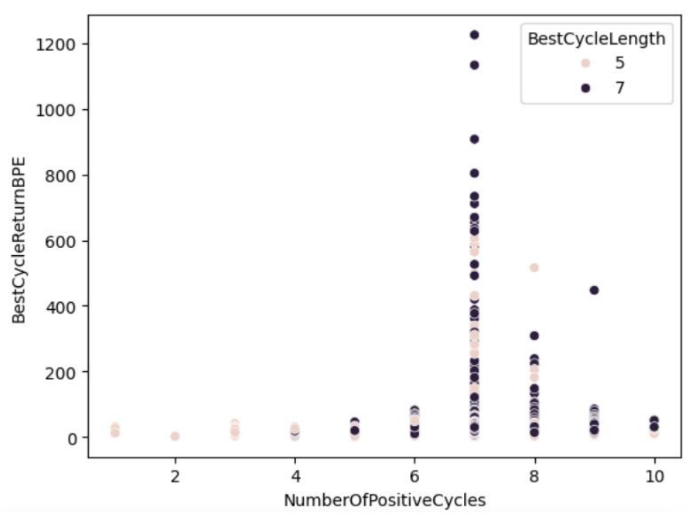
Figure 15. Return Distribution for all Simple Cycles



Source: Author's own elaboration

One interesting thing is that the biggest part of profitable cycles is devoted to graphs with a large number of positive cycles. Maybe for that reason, the best cycle length varies from 5 to 7. The longer the cycle, the greater the profitability of arbitrage.

Figure 16. Relation between cycle length and return

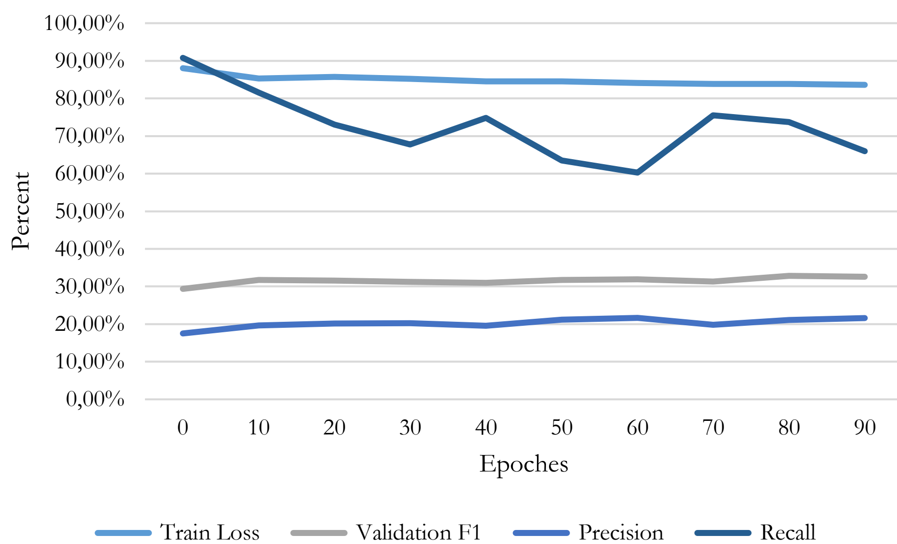


Source: Author's own elaboration

5.2. Modeling results

To test the hypothesis, all data was divided into the train (80%), validation (10%), and test (10%) sets. Figure 17 shows the evolution of training loss during different epochs in the range from 0 to 100 for GAT. The total modeling time was approximately 17min 24s (the total time of the computation cycle). The best model parameters were saved and pickled. The output embeddings from the best model were used in the classic machine learning models as logistic regression and gradient boosting.

Figure 17. Training loss evolution for GCN/GAT

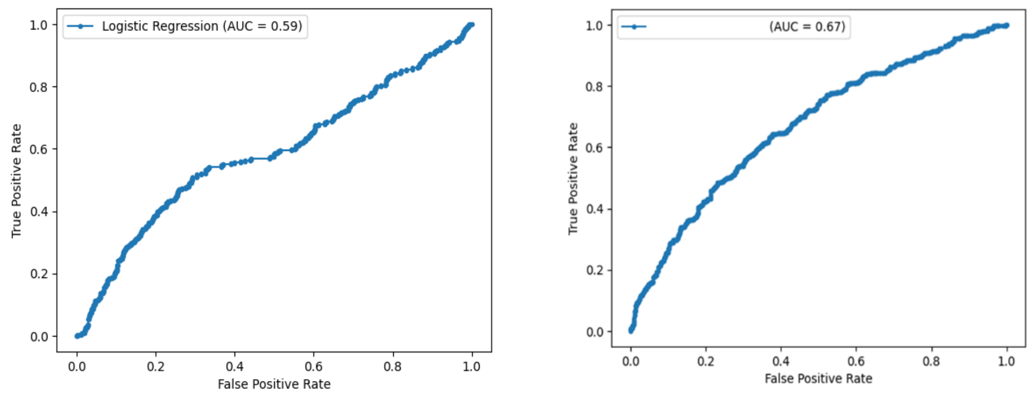


Source: Author's own elaboration

All metrics were computed using equations (12)-(15). The validation accuracy for Logistic Regression is 71.9%, while for LightGBM, it is 70.53%. The same difference was on the test set. Test accuracy for Logistic Regression was 71.97%, and for LightGMB, it was 71.42%. In Figure 18, the ROC-AUC curves illustrate that both models are better than the random classifiers.

Figure 18 depicts that LightGBM performs much better than Logistic Regression according to the ROC-AUC score. Both models are better than the random classifier as the area under the curve is greater than 0.5 for regression or LightGBM.

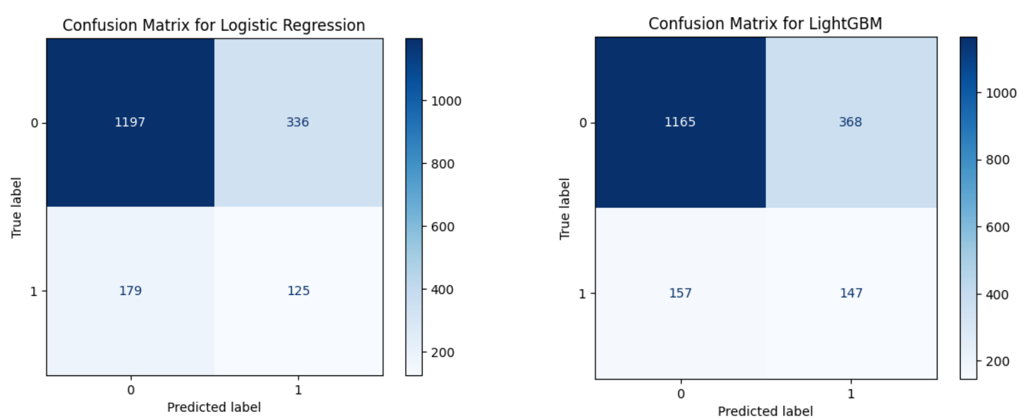
Figure 18. ROC-AUC curves for Logistic Regression and LightGBM



(a) Logistitic Regression (b) LightGBM

Source: Author's own elaboration

Figure 19. Confusion Matrixes



(a) Logistitic Regression (b) LightGBM

Source: Author's own elaboration

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

The traditional finance system is being criticized. People are also getting interested in new ways of financing, like cryptocurrencies and blockchain technology. The world's largest investment funds are interested in products that digitize real assets, make transactions more transparent, avoid intermediaries, and reduce fees.

Many new blockchain projects are emerging, but the biggest change is in finance. Today, we see a shift in how financial markets function for different assets. There is more reliability and fewer intermediaries, including through decentralized exchanges.

Investors have always been interested in the prospect of risk-free earnings in the financial market. It is for this reason that arbitrage has remained a subject of particular interest to financial experts. Decentralized blockchain exchanges are a case in point. Investors are particularly interested in them due to the specific pricing features of the liquidity pool system. Indeed, it is this liquidity system that creates the conditions for price discrepancies between different exchanges, which attracts arbitrageurs. They play a crucial role in the market by conducting opposite transactions, eliminating inefficiencies, and maintaining a fair price.

This study analyses arbitrage opportunities on various decentralized exchanges at the minute level of data aggregation. This equates to over 50 gigabytes of data, which, due to limited computing capacity, necessitated the narrowing of the study period to two weeks. This resulted in the analysis of several tens of thousands of records for the most liquid tokens. The investigation focused on the potential for cyclical arbitrage between different tokens on different exchanges, driven by sequential exchanges.

Our data substantiates the existence of arbitrage opportunities in cyclical transactions. It is worth noting that the majority of cycles are not profitable, as commissions actually reduce the profit of market participants. The most profitable cycles include at least five transactions. Our calculations show that the majority (304 in our dataset) of these potential arbitrage opportunities yield relatively small profits and are highly competitive. However, it can be assumed that, given capitalization and high-frequency trading, market participants can secure substantial income.

Our hypothesis stated that the specific arbitrage state of the market can be represented as a graph with certain tokens as nodes and exchange rates as edges. Such a graph can be used to predict potential arbitrage, which can then help build more optimized models for finding optimal arbitrage cycles. This hypothesis was tested using graph neural networks, which allowed us to obtain a vector (embedding) representation of the exchange rate graph and use it in predictive models. The logistic regression and boosting models showed better results than the random classifier, which supports the hypothesis that such factors are effective.

The decentralized finance industry is growing rapidly and is constantly evolving. The methods and algorithms used today are unlikely to be effective in the future. Therefore, this study can serve as a methodological basis for further research in exploiting arbitrage opportunities.

REFERENCES

Aoyagi, J., & Ito, Y. (2021). Liquidity implication of constant product market makers. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3808755

Aspris, A., Foley, S., Svec, J., & Wang, L. (2021). Decentralized exchanges: The 'wild west' of cryptocurrency trading. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3717330.

Avarikioti, Z., Heimbach, L., Wang, Y., & Wattenhofer, R. (2020). Ride the lightning: The game theory of payment channels. In *International Conference on Financial Cryptography and Data Security* (pp. 264–283)

Chains and DEXs volumes. (2024). Dune. https://dune.com/marchissio94/chains-and-dexs-volumes

Cryptocurrencies - CIS | Statista Market Forecast. (2024). Statista. https://www.statista.com/outlook/dmo/fintech/digital-assets/cryptocurrencies/cis#methodology

Cryptocurrency Prices, Charts And Market Capitalizations | CoinMarketCap. (2024). CoinMarketCap. https://coinmarketcap.com/

Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2019). Flash Boys 2.0: Frontrunning, Transaction Reordering, and Consensus Instability in Decentralized Exchanges. *ArXiv, abs/1904.05234*.

Easley, D., O'Hara, M., & Basu, S. (2019). From mining to markets: The evolution of bitcoin transaction fees. *Journal of Financial Economics*, *134*(1), 91–109. https://doi.org/10.1016/j.jfineco.2019.03.004

Eyal, I. (2015). The miner's dilemma. In *2015 IEEE Symposium on Security and Privacy* (pp. 89–103). IEEE.

Eyal, I., & Sirer, E. G. (2014). Majority is not enough: Bitcoin mining is vulnerable. In *International Conference on Financial Cryptography and Data Security* (pp. 436–445).

Fang, F., Ventre, C., Basios, M., Kanthan, L., Martinez-Rego, D., Wu, F., & Li, L. (2022). Cryptocurrency trading: A comprehensive survey. *Financial Innovation*, *8*(1). https://doi.org/10.1186/s40854-021-00321-6

Grimberg, P., Lauinger, T., & McCoy, D. (2020). Empirical analysis of indirect internal conversions in cryptocurrency exchanges. *arXiv preprint arXiv:2002.12274*

Harvey, C. R., Ramachandran, A., & Santoro, J. (2020). DeFi and the future of finance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3711777.

Huberman, G., Leshno, J. D., & Moallemi, C. C. (2017). Monopoly without a monopolist: An economic analysis of the bitcoin payment system. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3025604.

Hu, J., Cao, L., Li, T., Dong, S., & Li, P. (2021). GAT-LI: A graph attention network based learning and interpreting method for functional brain network classification. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04295-1

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

Nayak, K., Kumar, S., Miller, A., & Shi, E. (2016). Stubborn mining: Generalizing selfish mining and combining with an eclipse attack. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 305–320)

Kiayias, A., Koutsoupias, E., Kyropoulou, M., & Tselekounis, Y. (2016). Blockchain mining games. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 365–382).

Koutsoupias, E., Lazos, P., Ogunlana, F., & Serafino, P. (2019). Blockchain mining games with pay forward. In *The World Wide Web Conference* (pp. 917–927)

Kristoufek, L., & Bouri, E. (2022). Exploring sources of statistical arbitrage opportunities among Bitcoin exchanges. *Finance Research Letters*, 103332. https://doi.org/10.1016/j.frl.2022.103332

Kwon, Y., Kim, D., Son, Y., Vasserman, E., & Kim, Y. (2017). Be selfish and avoid dilemmas: Fork after withholding (FAW) attacks on Bitcoin. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 195–210).

Liu, H., Ruan, N., Du, R., & Jia, W. (2018). On the strategy and behavior of bitcoin mining with n-attackers. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (pp. 357–368)

Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293-319. https://doi.org/10.1016/j.jfineco.2019.07.001

MEV Data | EigenPhi / Wisdom of DeFi. (2024). EigenPhi. https://eigenphi.io/

Mining the environment – is climate risk priced into crypto-assets? (2022). European Central Bank. https://www.ecb.europa.eu/press/financial-stability-publications/macroprudential-bulletin/html/ecb.mpbu202207_3~d9614ea8e6.en.html

Mohan, V. (2022). Automated market makers and decentralized exchanges: A DeFi primer. *Financial Innovation*, *8*(1). https://doi.org/10.1186/s40854-021-00314-5

Open defi API - trading strategy. (2024). Trading Strategy - Algorithmic Trading Protocol. https://tradingstrategy.ai/api/explorer/#/Exchange/web_exchanges

Top Crypto Exchanges Ranked by Trust Score | CoinGecko. (2024). CoinGecko. https://www.coingecko.com/en/exchanges

Wang, Y., Chen, Y., Wu, H., Zhou, L., Deng, S., & Wattenhofer, R. (2022, April). Cyclic arbitrage in decentralized exchanges. In *Companion Proceedings of the Web Conference 2022* (pp. 12-19).

Zhang, J., & Wang, H. (2023). Deep reinforcement learning for arbitrage in decentralized exchanges. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4666504.

Zhou, L., Qin, K., Torres, C. F., Le, D. V., & Gervais, A. (2021, May). High-frequency trading on decentralized on-chain exchanges. In 2021 *IEEE Symposium on Security and Privacy (SP)* (pp. 428-445).

APPENDIX

Derivation of optimal profit quantity for executing arbitrage opportunity on CPMM DEX.

The amount of swapped token quantity is defined by the equation, including the fees:

$$dx = \frac{dy_A(1-f)X_A}{Y_A + dy_A(1-f)} \tag{A1}$$

$$dy = \frac{dx(1-f)Y_B}{X_B + dx(1-f)} \tag{A2}$$

Now we could substitute (A1) for dx into the (A2):

$$\begin{aligned} dy = \frac{dx(1-f)Y_B}{X_B + dx(1-f)} &= \frac{\frac{dy_A(1-f)X_A}{Y_A + dy_A(1-f)}(1-f)Y_B}{X_B + \frac{dy_A(1-f)X_A}{Y_A + dy_A(1-f)}(1-f)} \\ &= \frac{dy_A(1-f)^2 X_A Y_B}{Y_A X_B + dy_A((1-f)X_B + (1-f)^2 X_A)} \end{aligned} \tag{A3}$$

Now let's consider an equation (3) and find its derivative:

$$F(dy_A) = dy_B - dy_A \tag{3}$$

$$F'(dy_A) = d'y_B - 1 = 0 \tag{A4}$$

As dy_B is defined by (A3) and is a complex function we could find it derivative considering the nominator and denominator as separate functions:

$$dy_B = \frac{dy_A(1-f)^2 X_A Y_B}{Y_A X_B + dy_A((1-f)X_B + (1-f)^2 X_A)} = \frac{f}{g} \tag{A5}$$

$$d'y_B = \frac{f'g - gf'}{g^2} \tag{A6}$$

$$f' = (1-f)^2 X_A Y_B \tag{A7}$$

$$g' = (1-f)^2 X_B + (1-f)^2 X_A \tag{A8}$$

Now substitute (A7) and (A8) into the (A6):

$$\begin{aligned} f'g - fg' &= (1-f)^2 X_A Y_B \left(Y_A X_B + dy_A\big((1-f)X_B + (1-f)^2 X_A\big)\right) \\ &\quad - dy_A(1-f)^2 X_A Y_B((1-f)^2 X_B + (1-f)^2 X_A) \\ &= (1-f)^2 X_A Y_B X_B Y_A \end{aligned} \tag{A9}$$

$$\begin{aligned} g^2 &= (Y_A X_B + dy_A((1-f)X_B + (1-f)^2 X_A))^2 \\ &= k^2 d^2 y_A + 2kY_A X_B dy_A + (Y_A X_B)^2 \end{aligned} \tag{A10}$$

Where $k = (1-f)X_B + (1-f)^2 X_A$

Now we can find the optimal $d'y_B$. From (A6) we could obtain:

$$f'g - fg' = g^2$$

$$(1-f)^2 X_A Y_B X_B Y_A = k^2 d^2 y_A + 2kY_A X_B dy_A + (Y_A X_B)^2$$

$$k^2 d^2 y_A + 2kY_A X_B dy_A + (Y_A X_B)^2 - (1-f)^2 X_A Y_B X_B Y_A = 0$$

Substituting some parts for convenience and solving as a quadratic equation results in the following optimal:

$$a = k^2, b = 2kY_AX_B, c = (Y_AX_B) - (1-f)^2X_AY_AX_BY_B$$

$$k = (1-f)X_B + (1-f)^2X_A$$

$$dy_A^* = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{4}$$