DEPARTMENT OF COMPUTER SCIENCE

Cryptocurrency Pump and Dump Schemes: The Impact of Fake News and Social Media Advice

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

Wednesday 3rd May, 2023

Abstract

Pump and Dump schemes are a common occurrence in the cryptocurrency market. This paper is a study of the impact of social media advice in this type of scheme. The project aims to identify whether there's a correlation between online posts and the price of a cryptocurrency.

> My research hypothesis is that Pump and Dump Schemes are promoted through social media and that unknowing investors are being manipulated by bad actors into buying into a coin that is being artificially inflated.

As part of the project, I have achieved the following:

- I hand-picked a shortlist of 10 coins that have been associated with Pump and Dump Schemes in the past.
- I developed a Python web scraper for each of Reddit and Bitcointalk, and collected over 60,000 posts that reference the shortlisted coins.
- I developed a Python script that uses CoinGecko's API to collect historical market data for over 10,500 coins.
- Used the hand-picked coins and the historical market data to identify criteria that can be used to identify potential Pump and Dump Schemes.
- Used the criteria to find other coins that have P&D-like features.

Contents

List of Figures

List of Tables

Ethics Statement

An ethics application for this project was made with the University of Bristol's Faculty of Engineering Research Ethics Committee. The application's reference number is 13329. It received a favourable ethical opinion on 08/03/2023.

Supporting Technologies

Most of the work described in this dissertation was implemented in Python 3, therefore the technologies used were Python libraries and modules. The technologies I used are:

- Built-in libraries such as `json`, `csv`, `datetime`, `requests`, `statistics` to handle and analyse data and to make API calls.
- `Pushshift`'s API to collect Reddit posts
- `CoinGecko`'s API to collect historical market data.
- Google's Programmable Search Engine API to allow me to crawl Bitcointalk.
- `Matplotlib`'s module `pyplot` to plot graphs.
- `nltk`'s `vader` module to perform sentiment analysis on Reddit and Bitcointalk posts.
- `BeatifulSoup` to parse HTML.
- `scipy`'s `stats` module to calculate the Pearson correlation coefficient.
- The `dotenv` module for importing private keys without storing them in the code.

Notation and Acronyms

ACAR	:	Average Cumulative Abnormal Returns
API	:	Application Programming Interface
BS	:	BeautifulSoup
BTC	:	Bitcoin
CEX	:	Centralised Exchange
CSV	:	Comma Separated Values
DEX	:	Decentralised Exchange
EPS	:	Earnings Per Share
FOMO	:	Fear Of Missing Out
FTC	:	Federal Trade Commission
HTML	:	HyperText Markup Language
IP	:	Internet Protocol
IPO	:	Initial Public Offering
JSON	:	JavaScript Object Notation
MLM	:	Multi-Level Marketing
NLP	:	Natural Language Processing
NLTK	:	Natural Language Toolkit
OCR	:	Optical Character Recognition
P&D	:	Pump and Dump
SD	:	Standard Deviation
SEC	:	Securities and Exchange Commission
URL	:	Uniform Resource Locator
VADER	:	Valence Aware Dictionary and sEntiment Reasoner
env	:	Environment Variable
ftfy	:	fixes text for you

Chapter 1

Introduction

Cryptocurrencies have gained a lot of attention in recent years due to their high volatility and the potential for high returns on investment. The attention that mainstream media has given to cryptocurrencies has made them a hot new topic in the financial world. This has led a lot of people that have little to no experience in trading or investing to get involved and interested in the market, trying to make a quick profit. This is especially true for young people, who are more likely to be influenced by social media and online forums. A survey made by Blockchain Capital in 2020 says that more than 34% of Americans are likely to buy Bitcoin in the next 5 years, with a 55% for those aged 18-34 [7]. This is a significant increase from 2017, when only 19% of Americans were likely to buy Bitcoin in the next 5 years, with a 32% for those aged 18-34 [7]. This increase in interest to buy cryptocurrencies is also reflected in the number of cryptocurrency wallets, which has increased more than 4 times from 8.2 million in 2013 to nearly 35 million in 2016 [30].

However, despite the increase in mainstream media attention, cryptocurrencies are still a relatively new concept and a lot of people are not familiar with them. The combination of it being attractive for many people new in the market and the lack of regulation makes it a perfect medium for scammers to take advantage of. Fraud is a big problem in the cryptocurrency market. According to the Federal Trade Commission, reported "crypto fraud losses were $680 million last year(2021), in the last year in the roughly trillion-dollar cryptocurrency market, according to Federal Trade Commission. Losses in 2022 could double that — consumers reported $329 million in losses to the FTC in the first quarter alone. [12]. In the UK, data from Action Fraud reveals "a staggering £146,222,332 has been lost to cryptocurrency fraud since the start of this year - which is almost a third more (30 per cent) than was lost throughout the whole of 2020" [19].

Fraud has always been a problem in the financial world, through Ponzi schemes, pump and dumps, insider trading, etc. The rise of the cryptocurrency market has just created a new medium for scams that have been around longer than people know. My project will be focusing on one of the most common scams in the cryptocurrency market, pump and dumps.

One of the most known cases of P&D in the penny-stock market is the case of Jordan Belfort, the inspiration for the movie "The Wolf of Wall Street". In 1997, the Securities and Exchange Commission(SEC) charged 19 individuals and 17 firms with participating in a massive P&D scheme that defrauded investors of more than $25 million. The SEC alleged that Stratton-Oakmont Inc. and its president Jordan Belfort orchestrated the scheme, which involved artificially inflating the price of at least 35 initial public offerings (IPOs) [29]. An IPO is the process by which a private company becomes publicly traded on a stock exchange. The artificial inflation of price was done by encouraging investors to buy the stock, by making false and misleading statements about the company's prospects and finances. His brokers pushed stocks onto their unsuspecting clients, which helped inflate the stocks' prices. Then Stratton-Oakmont Inc. would sell off its own holdings in these stocks at a great profit [5, 34]. In 1999, "Belfort

pleaded guilty to fraud for the pump and dump schemes which may have cost his investors as much as $200 million. He was sentenced to four years in prison and ultimately served 22 months in prison. [25]"

The research I found on P&D in the cryptocurrency market is limited. The inspiration for this project comes from a paper that my supervisor, Dr Matthew Edwards, shared with me: "Cryptocurrency pump and dump schemes: Quantification and detection" by Victor, Friedhelm and Hagemann [36], Tanja. It analyses the coordination of malicious actors in the cryptocurrency market, with an emphasis on short-term price manipulation. This kind of P&D scheme is commonly done through Telegram/Discord groups, where the owners of the group announced the coin that they should buy. The coordinators buy the coin before announcing it to the group and then sell it when the price goes up. The paper also proposes a method to detect these schemes. It uses the properties of confirmed events of P&D schemes that have been announced in these group chats to train a binary classifier in order to detect additional suspicious trading activity, such as price, and trading volume: "A minimum volume of $50,000 and a peak gain of 5% within a time window of 30 minutes"[36].

Most of the P&D schemes that the paper analyses, and the ones that are usually the target of such Telegram/Discord groups, focus on small coins. My research will move away from what is already done by moving the focus from coordinated short-term P&D schemes towards the longer-term ones, where the hype of a coin is created through other means, with most of the participants not even knowing they are part of a P&D scheme. The paper will focus on some of the points specified in the paper I mentioned [36]: analysing the impact of fake news and coordinated social media advice. The research will therefore be conducted based on posts from Reddit and Bitcointalk. Reddit is a social news website with a wide range of communities called subreddits. Each subreddit is dedicated to a specific topic, thus allowing users to post and view content relevant to them. It is also a large hub for investment advice, with subreddits such as r/stocks, r/wallstreetbets, r/cryptocurrency, r/bitcoin, etc., so it is a great place for attempts at P&D schemes. Bitcointalk is a forum dedicated to Bitcoin and cryptocurrencies, being one of the largest forums that focus on these topics, with over 2 million registered users, who make an average of over 8,000 posts a day [17] It is also a great place to look for P&D schemes, as it is a forum dedicated to cryptocurrencies.

A final objective of the paper is to shed light on the issue of P&D schemes in the cryptocurrency market, and to raise awareness towards the fact that they do not only happen in the penny-stock and penny-cryptocurrency market, but also in the larger market, and that people should be careful when investing in cryptocurrencies, especially when it is done at the advice of strangers on the Internet because they might be lured into joining a P&D scheme that will not make them "rich quickly", but rather, make them lose the money to the benefit of those who joined the scheme before them.

In short, the goals of this project are to:

1. Build on top of what research has already been done on P&D schemes by moving towards schemes that take longer
2. Build a pipeline for detecting possible P&D schemes based on coin data and searching for online posts that reference the coins that are found
3. Raise awareness towards the issue of P&D schemes in the cryptocurrency market and the overall issue of fraud in the cryptocurrency market
4. Identifying if there is a correlation between the price of a coin and the number and sentiment of posts about it

Chapter 2

Background

This chapter will go over the background of the project, including other research that has been done on this topic and present the problem at hand and the reason why it is important to solve it.

2.1 Cryptocurrency

Cryptocurrency is a digital or virtual currency that uses cryptography for security. They operate independently from a central bank. Most of them are based on blockchain technology. A blockchain is a "consensus of replicated, shared and shared, and synchronized digital data that is geographically spread (distributed) across many sites, countries, or institutions." [14] with growing "lists of records(blocks) that are securely linked together". [18] The blocks are interconnected, every block containing the cryptographic hash of the previous block, thus giving the "chain" name. Being a distributed ledger, a blockchain is the opposite of a centralized database. The latter is usually owned and administered by a central authority, whereas the former does not require a main owner, thus not having a single and central point of failure. Consequently, any transaction that has reached the blockchain is broadcasted to all of the contributors' systems, making them impossible to revert. This is one of the many reasons cryptocurrency is prevalent in the fraud world.

The first and most well-known cryptocurrency, Bitcoin, was created in 2009 by an unknown person or a group of unknown people under the name Satoshi Nakamoto [16]. If Satoshi would be To sell all of their Bitcoin, they would be "the 15th richest person in the new world", with a net worth of up to $73 billion. [24] These large numbers that everyone is talking about when it comes to cryptocurrencies are the reason they've become so popular. Cryptocurrencies have gained popularity in recent years, particularly as an investment opportunity for getting rich quickly, and have attracted a large number of individuals who may not have previously invested in traditional financial markets. Figure 2.2 shows the estimated number of cryptocurrency users worldwide from 2016 to 2022. There is a big increase of 189% in users from 2018 to 2020, characterised by an "increased number of existing accounts as well as the heightened ability to connect the accounts to individuals" [28]. After that, the number of users is slowly but almost linear increase, with a total of 402 million users in November 2022. The attention that cryptocurrencies have received is continuingly increasing, with more and more people investing in them and more and more people being exposed to them.

One of the attractions of cryptocurrencies is their unregulated nature. Unlike traditional currencies, cryptocurrencies are not issued by a central authority, making them theoretically immune to government interference. As specified in the introduction, this is causing the cryptocurrency market to become the wild west of the financial world, with a lot of fraud and scams happening.

Another attraction of the cryptocurrency market is the anonymity that it offers. Even

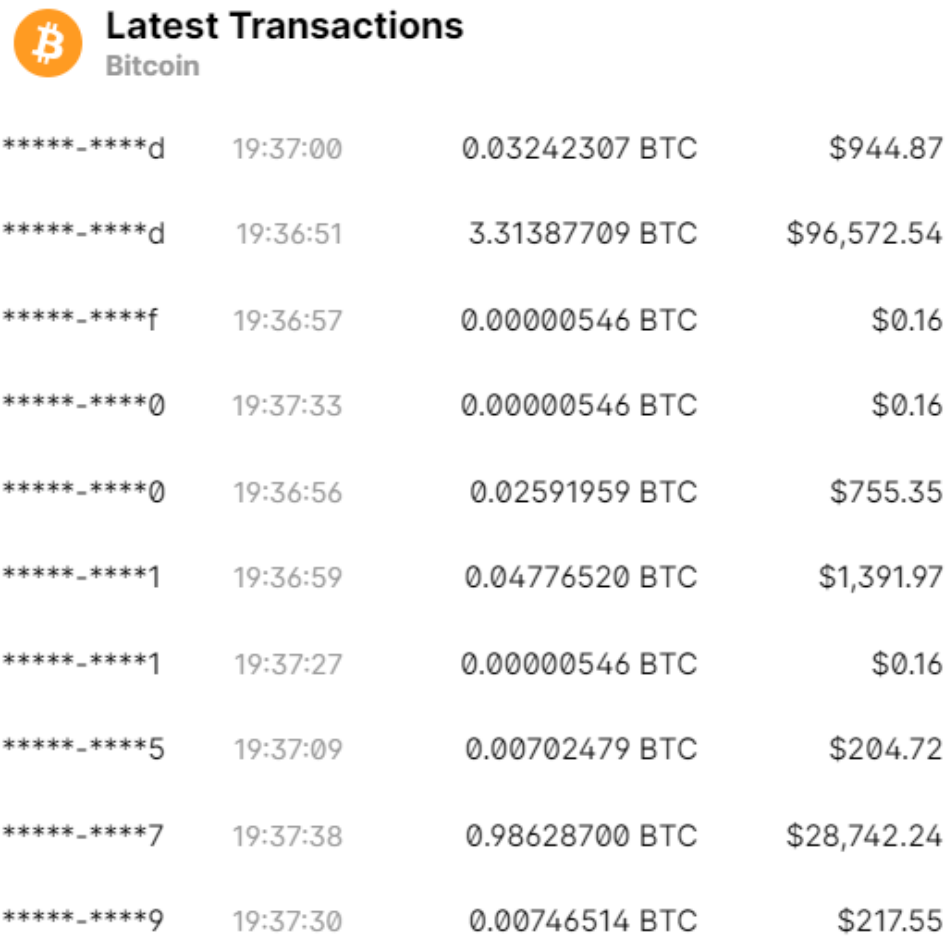

Figure 2.1: A snapshot of recent Bitcoin transactions available on blockchain.com [6]

though all of the transactions are recorded publicly and are available for people to find on the blockchain, the identity of the people making the transactions is not known. Figure All that a person that wants to see more can do is track the transactions that a wallet does, but they can not find out the identity of the owner from the blockchain itself. This is different from wallets created on websites that offer this service, as those are directly tied to an identity. The fact that cryptocurrencies offer anonymity makes them a good medium for illegal activities. Users can easily make transactions involving illegal goods such as drugs and guns, and they can also use them to do other illegal activities, such as money laundering, tax evasion, pump and dump schemes, Ponzi schemes, etc. As seen in Figure 2.1, where there's a list of the most recent Bitcoin transactions at the time of writing this, all transactions are free for anyone to see and easily accessible online: there's no need to have a wallet or to be part of the blockchain to have access to this data. Note I hid the wallet ids, they are actually shown on the website.

2.2 Cryptocurrency Fraud

There are many types of fraud that exist in the cryptocurrency space, most of them already existing in other markets. The rise of mainstream attention and the sheer amount of new people joining the space makes it a great target for fraudsters. When comparing the cryptocurrency space to the already existing situations, the downside of the former is that, being unregulated, with no authority to protect users, scams are easy to conduct and there's no way for people that have been scammed to get their money back. In this section, I will go present some of the most

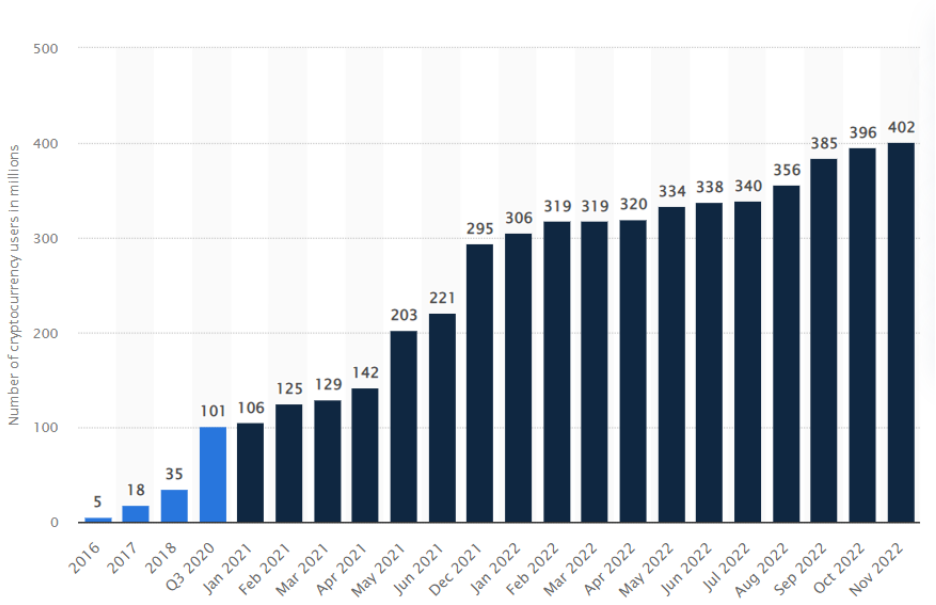

Figure 2.2: Estimate of the monthly number of cryptocurrency users worldwide 2016-2022 [33]

prevalent types of cryptocurrency scams:

1. Phishing emails and websites. Phishing is the practice of sending people emails or websites that appear to be from reputable sources, with the scope of making the target individuals reveal personal information, such as their identity, passwords, and bank accounts. This is especially dangerous in the cryptocurrency space because if, for example, the person being targeted by a fraudster shares their wallet password, all of their funds might instantly leave their possession. This is less possible in the context of a bank account, as they have checks in place that flag activity that is out of the ordinary, such as sending all of their funds to an unknown account.

2. Fake exchanges. Websites that pretend to be exchanges, but are just a way for fraudsters to steal money. They usually offer a service that is too good to be true, such as offering Bitcoin or other cryptocurrencies at prices well below the market price.

3. Advance-fee scam. This is a type of scam that has been around for a very long time and has been adopted into the cryptocurrency space, too. It consists of the scammer promising the victim a large sum of cryptocurrencies. However, for cashing out the money the victim would have to deposit or send the fraudster some cryptocurrency, most commonly Bitcoin. These are effective in the cryptocurrency space for the same reason others work: there's no one to be held accountable for the fraud, and there's no way for the victim to get their money back.

4. Ponzi schemes. They "get their name from notorious 1920s swindler Charles Ponzi (although he may have gotten the idea from an earlier scammer, William Miller, who was nicknamed "520 Percent" for the exaggerated returns he promised investors in a late-19th-century con). The basic premise has not changed in more than a century: A crooked broker touts a surefire investment, guaranteeing lavish returns." [1] The victims are persuaded to invest in an asset or give their money directly to the fraudster, with the promise that they will increase their value, often doubling or tripling it. The scheme works as long as there are new investors that join the scheme. The fraudster then uses the new investors' funds to pay out the investors that came before. The scheme falls when there are not enough new people joining in to account for the payouts of previous investors. In

the cryptocurrency space, this scheme can be done more easily with none of the victims being able to cash out at all. These schemes are usually run in cryptocurrencies created by the authors themselves, rather than using an already existing coin. Their balance looks like is increasing in their wallet, and they get paid in said cryptocurrency, but they are not allowed to sell the cryptocurrency at all or the cryptocurrency loses all of its value or liquidity overnight.

5. Pyramid schemes. They work similarly to the Ponzi schemes, but now the investors get paid for recruiting new investors. "Many pyramid schemes resemble multilevel marketing (MLM) businesses, which also involve a chain of adding new people to the operation. The key difference is that while a legitimate MLM focuses on bringing in people to sell the product, pyramid promoters emphasize the recruitment itself." [1]. Most of these are strongly marketed through social media influencers, who are incentivised to spread the word about the scheme because they'll get paid and have succeeded in getting people to invest because of how many people's attention they can absorb.

6. Honeypots. These happen when coin creators launch a new coin that is rigged: only a certain list of wallets, set by the creators, can withdraw from the coin, unbeknownst to the investors. All of what victims can see is that the price keeps increasing. This happens because, since no one but the creators can sell, there is no one that sells, essentially creating a monopoly. In a monopoly, there are many buyers, but only one or a small number of sellers (which is forced by the fact that the coin is rigged), thus creating an unjust market where the price of the coin keeps increasing. The increase in price makes other users invest too, hoping that the price will keep increasing and they will be able to sell when they feel like it, only to find out that they can not actually sell at all.

7. Rug-pulls. This is also usually done by the developers of a coin. It happens with coins that have small liquidity, and there are individuals that own a high amount of said coin. "A rug-pull is when developers run away with funds from a cryptocurrency - in essence, they remove all the liquidity from the pool and drive the price of the coin to zero, preventing trading at all. As the phrase implies, it happens suddenly and without warning - usually near a peak trading value so the scammers can take as much money as possible." [35]

8. Pump & Dump Schemes. This is similar to a rug-pull and happens when the coin at hand loses all or most of its value due to malicious actors selling the coin after getting people to invest in it, leaving the users with a loss of money. They're usually harder to spot and it is difficult to prove that the price increase and the sudden increase are done intentionally, rather than being just a product of how the market works. P%D schemes will be the focus of my project.

With so many types of scams that users need to avoid and with no one to protect them, it is not a surprise that scams are as prevalent in the cryptocurrency space as they are, especially given how many inexperienced investors join the market every day. Figure 2.3 posted by the Federal Trade Commission shows how alarmingly fast the total losses due to cryptocurrency investment scams increase. They report that "Since October 2020 (until Q1 2021), reports have skyrocketed, with nearly 7,000 people reporting losses of more than $80 million on these scams.", with a reported median loss of $1,900, with 12 times more reports than in the same period in the prior year, with 1,000% more in reported losses [13]. They raise awareness of the fact that online many people appear friendly and apparently want to share their "tips" with the victims, when in reality they only want to influence people to invest too, thus recruiting new people into their schemes. As seen in the explanations about the most prevalent scam types, there's a high tendency and incentive for scammers to recruit as many people and to share their "tips" on all social media mediums they have a following on. Even though younger people would likely be the

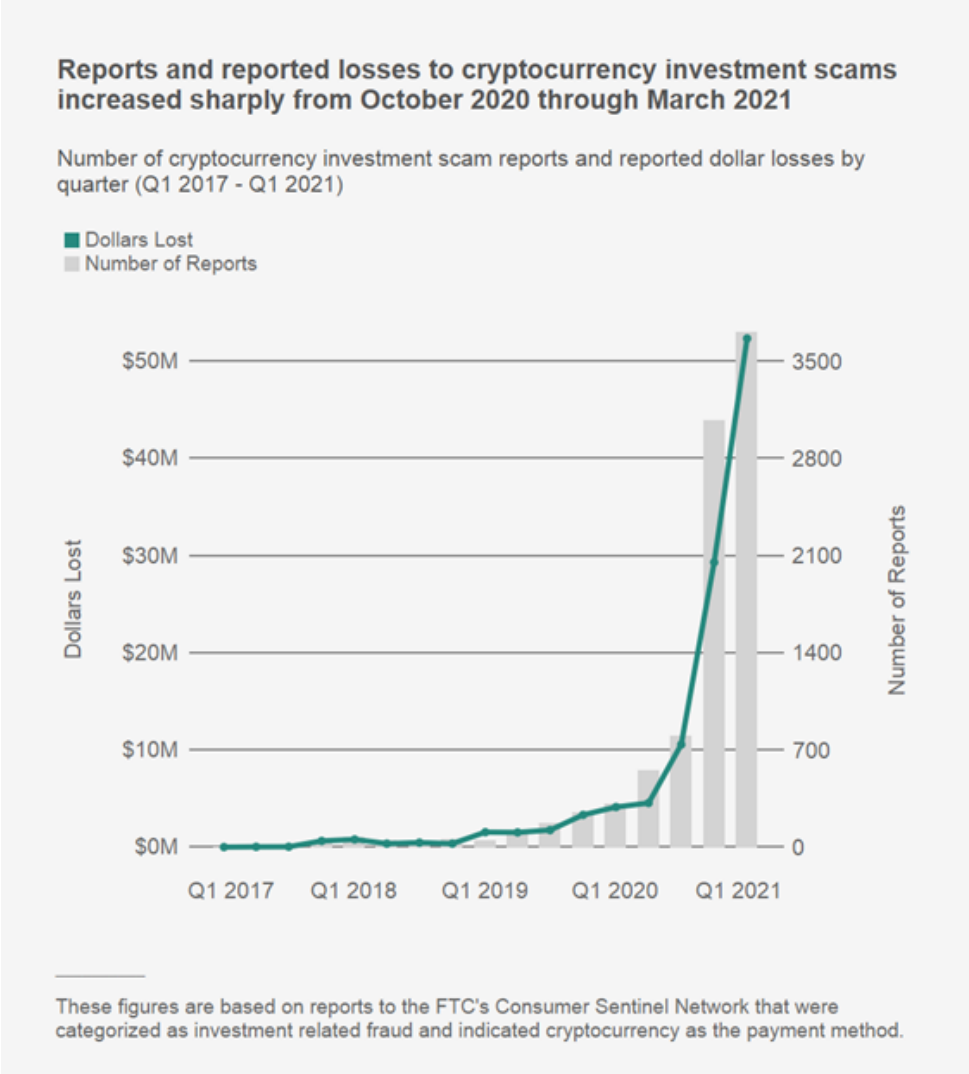

Figure 2.3: Scam losses in the cryptocurrency space [13]

target of such scams due to inexperience in the field and the "hustling" trend, cryptocurrency scams affect all age groups. FTC's findings show that even though people aged between 20 and 49 were five times more likely to report losing money on cryptocurrency investment scams, people over 50 that reported their losses have a much higher median loss of $3,250 [13]. This shows that cryptocurrency scams affect anyone that is curious and inexperienced enough. Since the field is not regulated and there's a general stigma with people admitting they've been the victim of fraud, the real number of scams might be bigger than what's reported.

2.3 Pump and Dump Schemes in the Stock Market

A Pump and Dump scheme is a type of fraud that involves inflating the price of an asset through false and misleading positive statements, in order to sell a low-value asset at a higher price. After the fraudster sells "dumps" their overvalued asset, its price usually falls, causing the investors to lose money. [32]. This is most common in small-cap cryptocurrencies and penny stocks. While in the past operators of such a scheme relied on cold calling, there now are much better and cheaper ways to spread hype about an asset: social media, investment research websites, spam, etc. The operators of P&D schemes prefer to target people that are inexperienced in the financial market. Unlike other types of fraud, P&D schemes are harder to spot and pinpoint,

and the victims only realise they are being taken advantage of when they are left with an asset worth almost nothing. Unless the target of the fraud does their due diligence and researches the asset that is being sold to them, they will likely get stuck with an asset that sounds too good to be true. When talking about stocks, it is sometimes difficult to get proper research done, and people may be easy to convince, especially as the malicious actors keep learning about how to persuade people better. When talking about cryptocurrencies the high volatility of price and the idea of getting rich fast plays a strong role in influencing people to buy. People who are not familiar with the concept can think of Cryptocurrency as a black box that generates money, and the fraudsters use this against the victims. Operators of the scheme hype the cryptocurrency project in such a way that the fear of missing out outweighs the risk that victims might know that these cryptocurrencies imply.

P&D schemes have been around for long before the first cryptocurrency was made. Investopedia's article "The Pioneers of Financial Fraud" mentions an occurrence of a P&D scheme in the late 1800s. "Daniel Drew used techniques known as a corner, poop and scoop, and pump and dump to defraud stock market investors." [4] Another example of a P&D before the age of cryptocurrencies is the case of a 15-year-old student named Jonathan Lebed. During the dot-com bubble, he bought stock and "posted hundreds of messages on Yahoo Finance message boards recommending that stock to others." [26] The article also presents a point that this project wants to show: the correlation between posts on assets that are part of Pump and Dump schemes and the price/volume of said assets. "The average daily trading volume of the small companies he dealt in was about 60,000 shares; on the days he posted his messages, volume soared to more than a million shares. More to the point, he had made money. Between September 1999 and February 2000, his smallest one-day gain was $12,000. His biggest was $74,000." [26] When Jonathan made posts on stocks that he had, the trading volume would increase to almost 17 times what the initial volume was. This shows that P&D schemes can have a huge impact on the trading volume, and thus the price of stocks.

Allen and Gale [3] show that the manipulators are large the large investors, who hold a big percentage of a stock. They "pump the stock with a series of buy orders then dump the stock, taking advantage of the disposition effect of investors to make a profit. The manipulator's strategic action not only profits the manipulator but also brings about higher volatility, higher trading volume, short-term price continuation, and then price reversal." [23] Yu Chuan Huang and Yao Jen Cheng's research "Stock manipulation and its effects: pump and dump versus stabilization" [23] examine the manipulation of stock prices in Taiwan stock markets. They use a set of hand-collected data to examine the characteristics and patterns of manipulated stocks. Figure 2.4 shows the usual shape of the price movements surrounding stock price manipulation. In the figure, "the ACARs (Average Cumulative Abnormal Returns) for the whole sample, as well as for the positive-EPS (Earnings Per Share) and negative-EPS subsamples, are shown from 100 days before the manipulation period to 100 days after the manipulation period. The two vertical grey lines represent the beginning and end of the manipulation period. The ACARs of the manipulation period are presented over a grid showing different stages of the manipulation. m1 corresponds to the beginning of the manipulation period, m5 corresponds to the middle, and m10 corresponds to the end of the manipulation period." [23]

The shape of the price movements surrounding stock P&Ds is similar to the ones that appear in cryptocurrencies. The only difference is that, in the short-term P&Ds that are caused by coordinated Telegram/Discord communities the price spikes and drops very rapidly, sometimes within the span of minutes, whereas in long-term P&D schemes that appear in the stock market, the variation of the price takes a longer time, with the effects of a scheme spanning months, as seen in Figure 2.4.

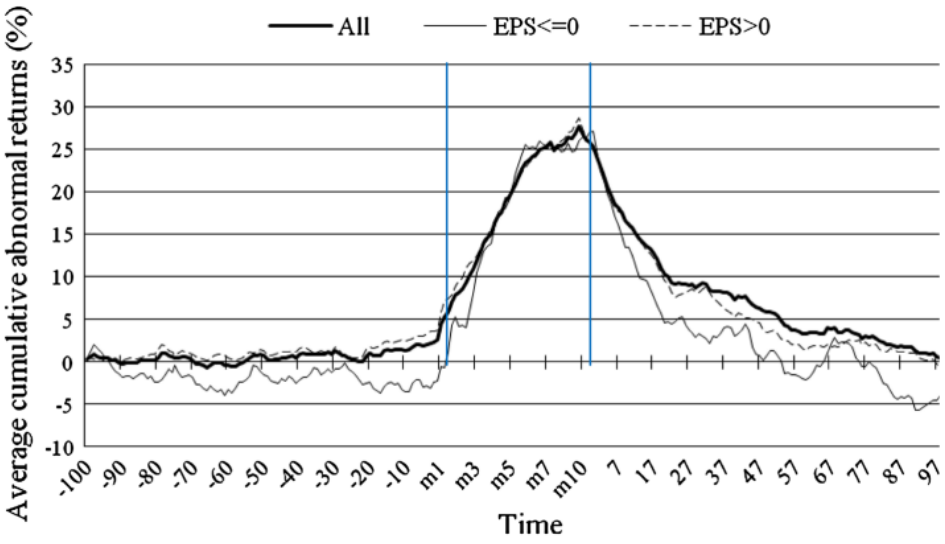

Figure 2.4: Price movements surrounding stock manipulation [23]

2.4 Short-term Pump and Dump Schemes in the Cryptocurrency Market

As presented before, P&D schemes take a slightly different approach in the cryptocurrency market than they do in the stock one. In this section, I will present the research others have done on the topic of P&D schemes. My project differs from others' findings as it focuses on longer schemes, the same ones that are prevalent in stocks, rather than the short-term ones. The world of short P&D schemes is most prevalent in Telegram and Discord group chats. The system is complex, with multiple layers of depth to it and multiple groups of actors that get involved. They form a coordinated network, that works together to recruit as many people to join the scheme.

2.4.1 Actors

There are 4 groups of actors that willingly take part in a P&D scheme. Most of the participants are recruited through social media or find advertisements of pumps on cryptocurrency exchange websites.

- Pump organisers: these are the owners of the groups and they are the ones calling the shots. They decide which coins will be pumped and when. The organisers also give the participants promises about how much money pumps will make them. They also take pride in previous pumps by broadcasting how big and sustained they were. They are the final winners of the P&D scheme, profiting off of participants.

- Pump outsiders: these are the general audience, participants that join the pumps when they are announced and they are the ones that usually lose money in the scheme. Many of them end up buying the coin after the price has already been artificially inflated and end up selling at a loss, thus being the final victims. However, they still join the scheme in the hopes that they will be able to buy and sell fast enough, and either profit off of other outsiders or participants that do not even know that they're joining a scheme and just invest because of the fear of missing out, thinking the coin will increase in value and keep its value for longer.

- Pump insiders: these are the outsiders that have an advantage over the others. Li, Shin,

and Wang [27] have read through P&D group chats and have found that "some pump-group organizers offer premium memberships to allow certain investors to receive pump signals before others do". In a scheme where being the first matters, the premiums they pay towards the organisers pay off as they end up profiting off of outsiders. This can be looked at as insider trading, with the participants having to pay for the insider information.

- Cryptocurrency exchanges. There are some exchanges that organise the pumps themselves or partner up with the organisers to have the pumps run on their exchange. The exchanges profit off of this through being insiders of the scheme by knowing the coins that will be pumped in advance, earning money through fees from the surge in the number of transactions as a byproduct of the P&D scheme. There is also the third benefit of marketing: they give the participants a reason to join their platform and might end up with user retention, which is always a plus.

2.4.2 The process

The whole scheme takes the form of a pyramid, where the organisers get the biggest slice of the pie, then the participants that pay to get the pump signal early, then the outsiders that either get lucky enough to sell fast or end up selling at a loss or being left with a coin with no value that they are unable to sell because of a lack of demand.

A general P&D process can be split into 5 major parts: [38]

- Setup: the pump organisers gather a large number of recruits into their Discord or Telegram channels. As said before, these channels are not singular, they form a big network that works together and announce the same pump. No matter whether the channels are owned by the same group of people or not, it is always feasible for them to work together as they can attract more recruits that they can profit off of. The groups are configured so that only the administrators are able to post, and basic users only being allowed to read the messages. This is likely done to avoid interference. Users might want to trick participants into buying the wrong coin, or they can use the group to accuse the owners of not holding up the promise that "they are all going to win money" after they lose money on a scheme. Silencing the victims is an easy way for avoiding other users realise they are most likely going to be taken advantage of and make them not partake in the scheme.

- Pre-pump announcement: once groups gather enough users, "typically above 1,000" [38], the owners decide what coin is going to be pumped, but do not announce it yet. They only announce the date and time at which the pump will start. They also announce the exchange that the pump will take place on and the pairing coin, that is the coin that will be used to buy the one that will be announced. This is usually well-established, such as Bitcoin. The coins the organisers usually choose are obscure coins with low trading and low market capitalisation [21].

  The organisers also instruct the participants on how to behave during the scheme: they should buy as soon as possible, promote the pumped coin to others to attract other investors (therefore being recruiters to the fraud themselves) and hold. They also encourage the participants by advertising an expected price increase, influencing the victims to hold the coin for longer, until the target is reached. Figure 2.5 shows a graph posted by an organiser on the Telegram group chat "Big Pumps Binance" with the caption "$NAS (NAS/BTC) is the coin we've been talking about all this time... it's probably the highest potential spot gem on Binance and across all other exchanges", showing how they expect the coin they chose to behave in the future, all with the scope of influencing people to join. The figure shows a potential of "3,500%" profit. In reality, the price evolution of the advertised coin is miles away from their prediction and can be seen in Figure 2.5. It can be seen that the price takes a whole different path than the predicted one, and keeps

Figure 2.5: Graph of the prediction of price evolution of a coin advertised on a P&D group

Figure 2.6: Graph showing the real price evolution of the Nebulas coin paired with BTC [9]

decreasing steadily rather than increasing at the advertised rate. The degree of how wrong the prediction is goes to show that the organisers of P&D schemes have no background that they base their predictions on, being just speculations: they use big numbers that recruited people like, hoping to convince. Giving unlicensed investment advice is illegal in the UK under Financial Services and Markets Act 2000. This is all to the benefit of the organisers: the more people hold, the more the coin's price increases. There is evidence to show that this is also the period when the organisers buy the coin too: well before the insiders and outsiders do, not risking buying too late. This can be easily seen in coin data given that most of the coins usually show little to no trading prior to the scheme: there's a small spike in volume before the scheme starts. A general visual representation of this can be seen in Figure 2.7 even starts. [27] Insiders are given the coin name before the others.

- Pump: this is when the coin is announced to the whole group and people start buying. The organisers start with a countdown and then the coin accouncemented. The method by which is done is usually through sending an OCR-proof image. Figure 2.8 shows the countdown and the OCR-proof image. Only 12,000 group members saw the announcement out of the total of 62,307 members [36] The choice of format for broadcasting the coin name is done to prevent bots from getting the name of the coin from the group chat and investing in it. A potential bot can then set a limit order to automatically sell the coin when it rises, thus making a profit. The key element in P&D schemes is the speed with which the

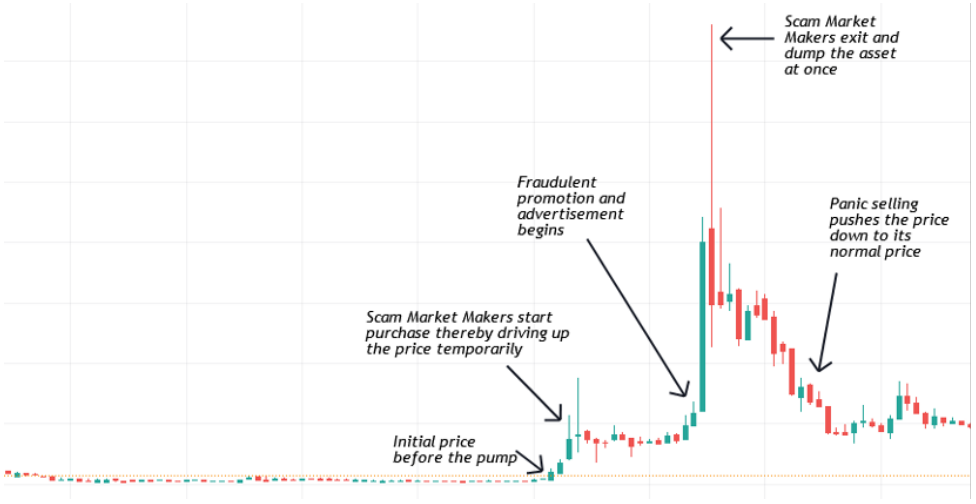

Figure 2.7: Graph showing the main parts of a Pump and Dump scheme [10]

participants react to the price change of the coin. For example, if a participant sees that their coin has reached a satisfying value and decides to sell, in the time it takes for them to send a sell request through the price might drop abruptly, possibly below the price at which they bought. A potential outsider running a bot can give them an advantage over the insiders, yielding the premium fees they pay to become a worse deal.

- Dump: after the pump starts, for up to a few minutes, the price keeps increasing while the organisers keep telling people to buy. When the price starts to drop, investors usually get panicked and start selling, making the price fall even faster. "The Anatomy of a Cryptocurrency Pump-and-Dump Scheme" found that "the price might be re-boosted by the second wave of purchasers who buy the dips (as encouraged by channel admins), chances are the price will rapidly bounce back to the start price, sometimes even lower." Attempts of the organisers to coordinate the users to hold and buy the dip can be seen in Figure 2.9. This particular pump had a higher exposure than the one done on POLY: 59,200 users saw the pump announcement while the current number of members is 56,788. When the coin reaches a price close to the one prior to the pump, it usually marks the end of the scheme. [38] The buyers that realise they bought the coin at a higher price than what it's worth now usually just hold the coin, rather than sell at a loss. Unless the coin is popular or is hit by P&Ds often, they are essentially stuck with a coin of low value with little to no trading volume.

- Post-pump review: after the scheme is done, the administrators do an analysis of the pump. This usually involves advertising the price at the peak and the percentual increase and hyping the next pump, promising an even higher peak price. If a pump goes bad, the organisers do not mention it at all. As only the owners of the groups can write messages in them, no one can publicly complain.

2.4.3 Effects of Pump and Dump Schemes

In this section, I will present the results of Pump and Dump schemes that I found during my research. Given that most of the coins selected for price manipulation are small, mostly obscure coins, it is not a surprise that they are the ones being affected the most. Studies have shown that P&D schemes can have a significant effect on the price and liquidity of cryptocurrencies, while the biggest impact is on small coins. A research paper made in 2021 shows that the small cryptocurrencies listed on PancakeSwap exhibit higher price and volume increases as a result of a P&D. The study also shows that the effect of a P&D on the price and volume of a

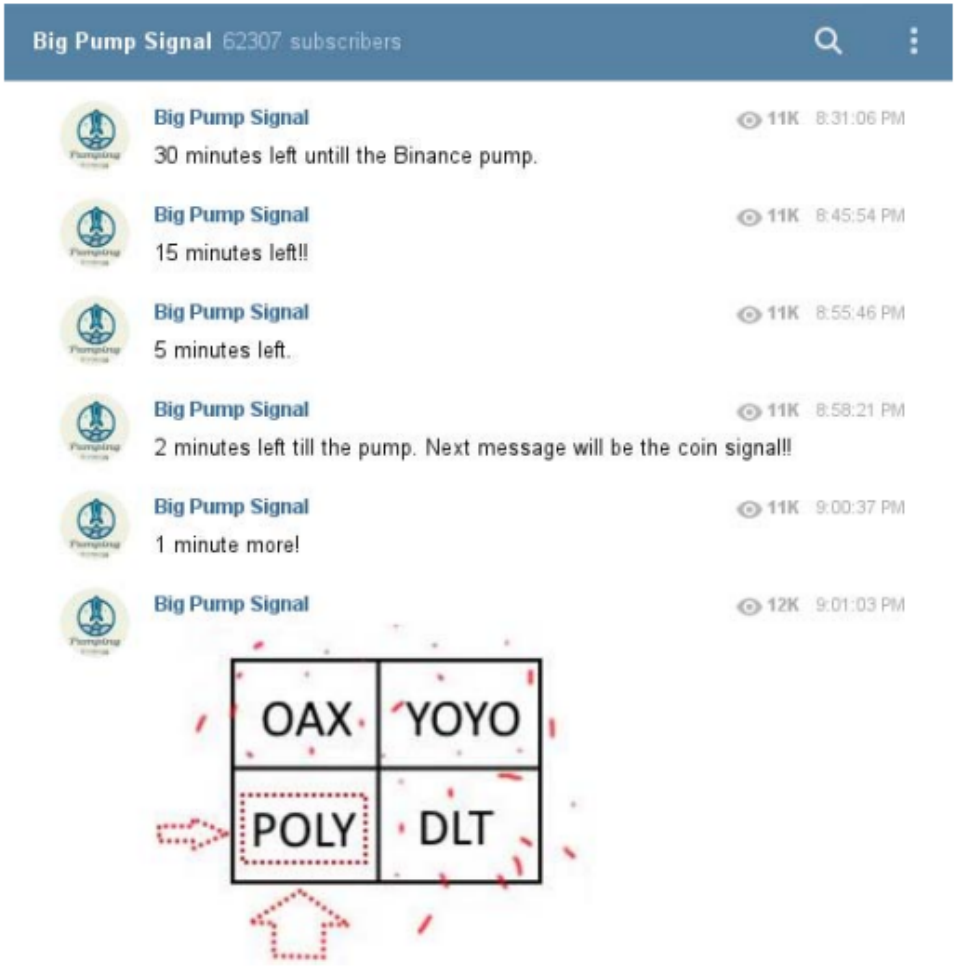

Figure 2.8: Screenshot from a Telegram group giving a countdown before announcing that the coin being pumped is POLY using an OCR-proof image [36]

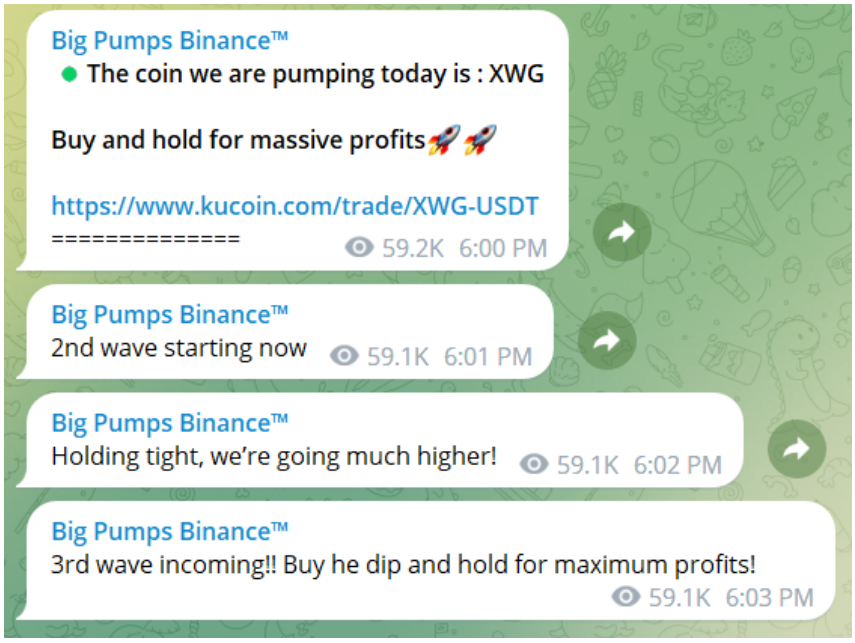

Figure 2.9: Pump organiser announcing a coin and coordinating the participants during the scheme

cryptocurrency is more significant when the coin is listed on a decentralised exchange, such as PancakeSwap than when it is listed on a centralised exchange(CEX). [27]

A centralised exchange is run by a central authority that monitors and facilitates transactions. They provide the necessary platform for transactions to take place. The transactions are settled off-chain in a central database, and offer the users the possibility of trading using fiat money. Decentralised exchanges are not run by a third party. They run on a system of smart contracts and run without any centralised oversight. They are on-chain, and trades are matched by an automated market maker. [15]

When looking at who benefits the most from a P&D, it is obvious that the pump organisers are the ones making the most profit. They make money both off of the premiums the insiders pay, and thanks to the advantage they have by buying the coins earlier than anyone else. Studies also show that the profits P&Ds make are higher in decentralised exchanges than in centralised ones. Researchers estimate that organisers and insiders realize "large returns, raging from 18% on CEXs to 152% on PancakeSwap". During the sample period, the insiders make "one Bitcoin (about $10,000) and 3.3 Binance Coin (about $1,200) for an average P&D on CEXs and PancakeSwap respectively." [27]. The return that the outsiders make are more difficult to quantify due to the much larger number of individuals and the trading volume. The insiders' buy prices are also easier to identify due to the small surge in transactions right before the pump is announced.

When looking at the impact on a coin, the data shows that the increase of price from the start of the pump to the peak shows a median of 14.4% and a median generated trading volume of $374,593, with some of the coins reaching over $1 million trading volume. 100 days after a pump, a coin that has been the target of a scheme performs on average 10% better than coins that haven't.[36]

2.4.4 Pump Types

As mentioned before, not all short-term pump attempts end up being successful, and pump organisers do not mention these. One way in which pumps fail is when the organisers see suspicious activity prior to the announcement. Investors can use previously pumped coins to establish a pattern and guess what the next coin will be. Another way in which investors might find out what coin will be pumped before they're supposed to is if an organiser leaks it to them. It is also possible to find the pumped coin before the insiders do and before the pump is announced through automatic detection. As the pump organisers buy the coins before everyone else, it is possible to analyse this pattern to predict future pumps. The paper "The Anatomy of a Cryptocurrency Pump-and-Dump Scheme" trains a random forest model that manages to predict pumps with this data. In the end, no matter how others might find out about the coin that will be pumped, if the organisers notice suspicious activity they will cancel the pump. [38]

The pumps that end up going through can be split into three categories and can be seen in Figure 2.10 [36]:

- Sustained pumps: these happen when the price does not immediately drop and manages to hold for longer than the usual minute-long pumps. This kind of pump ties well into the type of long P&D schemes that my research looks for. The pump is sudden, as in short-term pumps, but, in the end, it drops abruptly, but holds for longer. They are a byproduct of the participants not panic-selling and are the ones desired by the organisers. If a pump is long and sustained enough it has the possibility of having its price increased even further, giving an even better return on investment.

- Short-term pump and dumps: these are the most common, and are the ones that are analysed by most papers. They are characterised by a sudden spike in price followed by a sudden drop in spike, shortly after, as seen in Figure 2.10. Usually caused by panic selling.

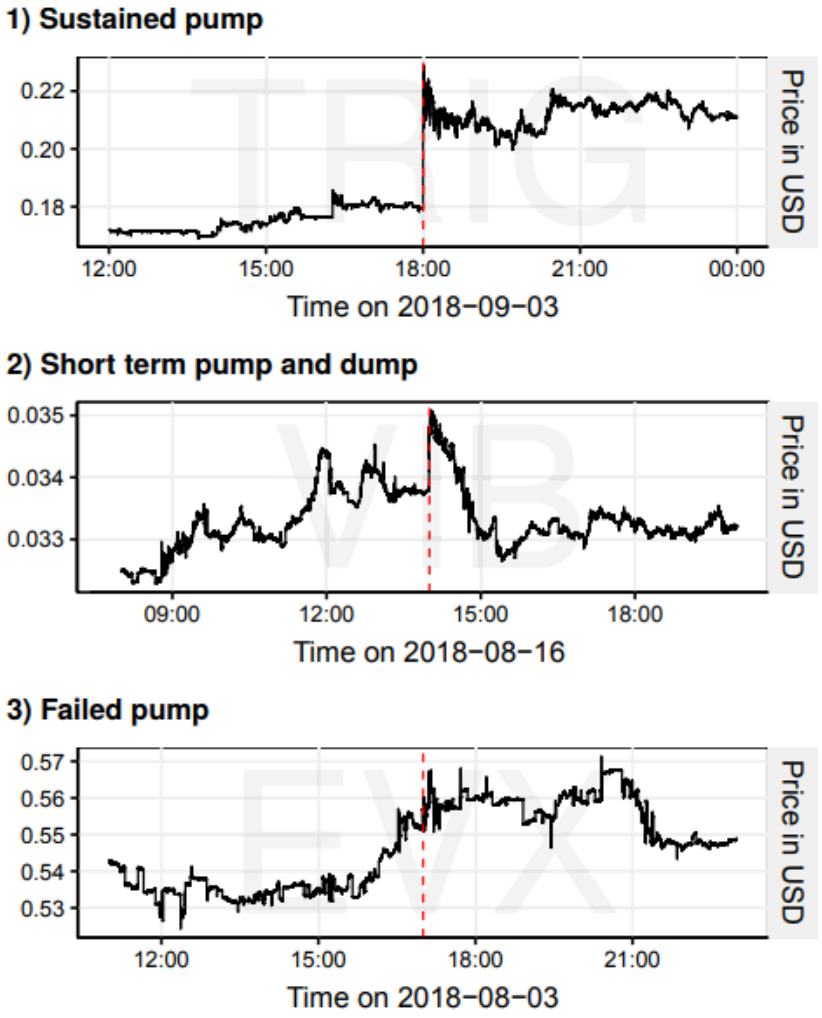

Figure 2.10: Types of pump events [36]

- Failed pumps: characterised by a pump that is too small with a peak that can not be differentiated from how the price is affected by the usual market, without interference. They are caused by not enough people buying / people buying too little.

Chapter 3

Project Execution

My project consists of 6 main stages, each of them having a different degree of complexity. The stages are:

1. Researching previous findings on the topic of pump and dump schemes in both the stock and the cryptocurrency ones. The reason for researching both is that my analysis moves away from Telegram/Discord coordinated short-term pumps towards longer pumps that are prevalent in the stock market. This stage was presented in the background chapter.

2. Scraping historical coin data off of an exchange. The exchange of choice ended up being CoinGecko.

3. Selecting a shortlist of cryptocurrencies that have been associated with pump and dump schemes in the past.

4. Scraping Reddit and Bitcointalk for mentions of the selected coins.

5. Applying natural language processing on the scraped posts to make sentiment analysis on them.

6. Using the historical coin data of the shortlist of selected coins to search the scraped coin data for other coins that show pump and dump-like behaviour.

All the code and the dissertation were written in Visual Studio Code due to the lightweight feel and ease of use. To manage the code, I chose to use Git as a version control software and GitHub as a method of backing up my progress in case of a hardware failure. Despite my code and collected data leaving the machine I am working on, they are stored in a private repository, being available only to me.

The code is solely written in Python 3. I chose this programming language due to the ease of use which allowed me to focus on the problem at hand, rather than have me get stuck on the gimmicks that lower-level languages come with. Python also had the most resources and the most packages that helped me speed up the execution process, such as making API requests, parsing JSON, CSV and HTML, plotting graphs, and applying analysis techniques to my data. Performance was not an issue since most of the time was usually spent on non-resource-heavy tasks or waiting for API responses and waiting between API calls.

I chose to store the data in the project's directories themselves, using both JSON and CSV as formats. I used CSV for the coin price data that I scraped off CoinGecko, due to the simple format, having just 4 fields for each entry. JSON was used for more complex data, such as posts, as I needed it to be structured better, e.g. store the sentiment of a post within the post's JSON bounds or have different fields based on which website the posts were scraped off of. Since the data was included in the Git commits, the collected data was always backed up in case it got lost or corrupted.

I wanted to follow good programming practices so all of my API keys were stored in an .env file and imported into my code using the "dotenv" package. The file is ignored by the source control software and, in conclusion, my private keys do not leave my machine.

3.1 Coin Data Collection

The collection of coin data was the first practical step and was a necessity for graphs and further analysis. Despite having an initial shortlist of coins that I could collect, I wanted to get data for as many coins as possible. This is because one of the scopes of the project was creating a pipeline of going from coin data to the posts made about said coin.

When choosing the source of my data, my first thoughts were to try Binance or CoinMarketCap. I did not use the latter because the free tier did not have access to historical data. Binance wanted me to deposit a sum of money prior to getting an API, so I ended up not using it either. My final choice of data source was CoinGecko. At the time of the collection, an API call I made gave me a list of 12314 coins that I could have access to. The list of available coins was stored in a CSV file, with the id, symbol and name being saved for every coin.

Getting the historical coin data was done using two Python scripts. One of them was a helper script, that was used to send the requests and write the response to a CSV file. The other script handled the logic of making the calls. It parses the list of all the coins available on the website, which I previously made, and makes requests for each. Being on a free API plan, the server only allowed 10-50 requests a minute, timing the API key out after too many calls were made. I went around this by making the program check whether the server sent the expected response and, if it did not, I made the program wait for one minute before trying again. This abides by ethical considerations regarding using the API, my program does not overload the server with requests that it can not serve and waits until it's ready to serve them, by pooling the server every minute to see if it is ready.

The data contains the prices as early as April 2013 in the case of Bitcoin, up to the time when it was collected. Each coin is stored in a separate file, with one multiple entries for each, based on how long they've been stored in CoinGecko's database and how old the cryptocurrencies are. Each entry has a timestamp, followed by the price, market cap and total traded volume of the coin at that timestamp. One limit of the collected data is that the historical data has low resolution, with only one entry per day. While it is not within the scope of the project to analyse short-term pumps that need minute-by-minute data, coin movements within a day are harder to see with only one aggregated value per day.

Another limitation of the data is that it might be subject to bias. There's selection bias involved, the coins that I collected the data for do not represent the whole coin population. This is due to the fact that CoinGecko does not aggregate all of the existing coins, focusing on only more than half of the total of over 22,900 as reported by CoinMarketCap. [22] As a result of this, the data does not include obscure coins, which are more likely the target of P&Ds. This is not a problem in my case as I am looking for longer pumps, that are advertised on social media and are the target of mainstream attraction, being more popular. This will also indicate that my findings regarding P&Ds represent the lower bound because the coins that are missing are the usual target of P&Ds anyway. Another bias that the coins might be subject to is survivorship bias. This happens because the sample does not include all coins that failed or have been delisted. It is provable in my dataset because the last entry for any coin is on December 2019. This means that any coins that were delisted before December 2019 are not included in the sample. This is not too big of an issue since the coin data covers the period before the end of 2019 anyway.

In the end, the coin data is good for the study and has a big enough population to satisfy the requirements for good statistical evidence. Even though some of the population is missing, it does not necessarily represent the target of the study.

3.2 Coin Selection

Due to not finding prior analysis of long-term pump and dumps, one of the aspects that needed to be addressed was finding a shortlist of coins that have been associated with P&D schemes in the past. The historical data found on these coins will then be used as a pattern for searching for long-term pump and dumps later. Due to the nature of the subject and the long time that these schemes take, there isn't a way to prove that the selected coins are P&Ds beyond reasonable doubt. There have been a small number of arrests made for P&D schemes in the cryptocurrency world, such as the case of a 39-year-old man that has been charged with fraud and embezzlement by the Dutch police after he attempted to defraud investors by luring them to invest in a cryptocurrency he made, only to sell all of his coins afterwards, profiting off of them. [31] However, being a new topic, and sometimes the fraudster's identity is unknown due to the anonymity cryptocurrencies provide, it's difficult for law enforcement officers to pinpoint the identities of the bad actors.

In conclusion, the list contains 10 coins that are selected based on allegations of being connected to P&D schemes in the past. The list is not exhaustive as it is only a starting point for analysis. The time ranges selected as the start and end of a P&D scheme on the selected coins use both what the online mentions say and the P&D-like patterns that surge in coin data based on the studies presented in the background section.

Some of the most known cases of pump and dump among the selected are:

3.2.1 Dogecoin (DOGE)

Dogecoin is a so-called meme coin, a coin that originates from an Internet joke or has an overall humouristic characteristic. Dogecoin itself resembles the Doge online humouristic reference that started going viral around 2010. It is one of the most popular cases of social media manipulating the price of a cryptocurrency. An article on CoinCodex [37] documents how Elon Musk has influenced the price of Dogecoin over the years. Prior to December 2020, the coin had a value of under a penny. After Musk started tweeting about it, the price started increasing, as shown in Figure 3.1. Subfigure 3.1a shows an 18% price increase from $0.039 to $0.046 shortly after Elon Musk's tweet "One word: Doge". Subfigure 3.1b shows another instance of price manipulation after Musk tweeted "Doge", increasing the price of the coin by 40%, from $0.0408 to $0.057 within a couple of hours.[37] These price spikes are directly related to his tweets and this kind of market manipulation, while illegal in the case of stocks in most countries, is difficult for authorities to detect and prove, especially in the case of cryptocurrencies. This kind of market manipulation is done, while at a smaller scale, by other malicious actors through social media in an attempt to increase the price of the coins they own. While not as successful as Elon Musk, it is possible if done by a big enough number of people.

3.2.2 Verge (XVG)

These two coins have also been subject to a pump and dump scheme. "The Commodity Futures Trading Commission today filed a complaint in the U.S. District Court for the Southern District of New York charging businessman and computer programmer John McAfee, and his former employee Jimmy Gale Watson" [11] for conducting a pump and dump scheme. They secretly bought Verge and Reddcoin coins and then touted them on Twitter between December 2017 to October 2018. The malicious advertisement managed to make a spike in price with people joining the hype, letting the two fraudsters sell at a profit. this effect is confirmed by the price movement in that timeframe. Before the start of the pump, Verge had a pre-pump value of $0.0063 and a peak price of $0.2619, having its price increase of over 4000%. The coin then went on to lose almost 90% of its value over the next 3 months. Reddcoin had an even bigger effect, with an almost 11500% price increase from the start of the pump to the peak, and then

a loss of around 84% of its peak value. Reddcoin was not included in the analysis due to a lack of mentions on the websites that I scraped. When looking at numbers, there can be made the case that the Verge pump is not done solely by the two mentioned above, but also by others that jump the hype train and advertise the coin to others so they can sell at a profit, too.

3.2.3 Bitcion Cash (BCH)

Bitcoin Cash is another popular coin that has been associated with a pump and dump.[20] For the timestamps that I selected for this coin the price increases by 227% from the start of the pump and ends up losing around %83 of it towards the end of the dump.

3.3 Scraping Reddit and Bitcointalk

When it comes to scraping posts I ended up choosing these two websites. I chose Bitcointalk because is one of the largest forums that focus on cryptocurrencies, the choice for Reddit was made due to the large number of users that use it daily. It also has a big investing scene, with the subreddits like r/CryptoCurrency and r/Bitcoin which have 6.3 and 4.9 million members respectively. I managed to get the data I expected from these websites, but I ended up using a different method for each.

3.3.1 Bitcointalk

Scraping Bitcointalk could have been done in one of two ways. One of the ways was to do a wget-approach: start from the home page and follow all possible links and store every webpage found on the way. This would have proven to be too time-consuming as Bitcointalk times users out pretty fast and does not support crawling. There were more than 16 million posts and at most 20 posts/page I was looking at over 800,000 pages to scrape. With a half a second wait time between requests, to not upset the server and get my IP address banned, this all adds up to over 18 days of continuous scraping, which is not doable.

Since Bitcointalk's search service also had a 1.5-minute timeout between calls, I ended up using Google's Programmable Search Engine API to search for coin mentions on Bitcointalk. Due to the free tier having limitations on the number of calls, I made a crawler that scraped the links retrieved in the first 5 pages of results. Most of the links returned by the first 5 pages were from the same topic, meaning there was no reason to go beyond the 5.

Another crawler then takes the response and crawls all the pages available on a topic. Since the links had an easily predictable pattern, I could go through all the possible URLs that a topic has. After the whole scraping process, I ended up with over 21,100 posts out of which 4,715 were between the timestamps selected as the pump start and pump end of each coin. Due to the smaller number of total users, some coins ended up having no mentions for the specified timeframe. This does not necessarily mean that there was not an attempt to pump the coin, it just means that the coin received attention on other social media platforms. The users that work together to advertise coins come from all types of backgrounds and each uses their platform of choice. The fact that there are two different platforms being used for post collection helps with any biases that could impact the analysis.

3.3.2 Reddit

For scraping Reddit I used Pushshift's API. It offered the most freedom and allowed scraping for free. Since this website has a wider range of topics, I ended up creating a list of 70 investing-related subreddits (subgroups of the website) that I would use. The API allowed me to choose which subreddit to search and also allowed me to select the timestamps between which I want the results of the request to be. This allowed me to only scrape the posts that I needed, rather

(a) Dogecoin price increase on 20 Dec

(b) Dogecoin price increase on 4 Feb

Figure 3.1: Dogecoin price spikes shortly after Elon Musk's tweets about the coin [37]

than scraping everything like I did on Bitcointalk. Since the API send a maximum number of 500 posts, this limited the number of posts that I could get for a subreddit. I ended up going around this limitation by using an algorithm design paradigm called Divide et Impera (Divide and Conquer), which is splitting a problem into smaller sub-problems. Algorithm 3.1 shows how I implemented this algorithmic paradigm. Whenever the API responded with 500 elements, I would split the time range into two halves and search them, until I was left with responses that were shorter than the maximum limit. This assured me that no posts were left out. The actual function uses some extra checks to see e.g. if the response was successful but they were excluded from this section for simplicity. The function uses recursive calls split the time range in two and to count the total number of posts that have been found in a subreddit for a specific coin in the end.

Reddit ended up returning almost 42,000 posts, all within the specified time range. Table 3.1 shows the number of posts retrieved for each coin, for each of the two websites. With a total of 46690 collected posts, the sample size is big enough for satisfactory analysis. Some coins, such as Verge and Ripple are more popular on Reddit, while Dragonchain and Electroneum seem to have more mentions on Bitcointalk. Combined, the two websites give a good number of posts for each of the selected coins, with a minimum sample size of 594 for Dragonchain and a maximum sample size of 18633 for Dogecoin.

```
Input: coin_name, subreddit, after, before
Output: found_in_query
posts ← get_reddit_posts(coin_name, subreddit, size=500, after=after, before=before)
found_number ← len(posts)
found_in_query ← 0
if total_found == 0 then
    return 0;
end
if total_found < 500 then
    foreach post in posts do
        if is_not_empty_or_duplicate(post) then
            found_in_query += 1
            write_post_to_file(post)
        end
    end
    return found_in_query
end
else
    middle ← after + (before - after) / 2
    return multi_query(coin_name, subreddit, after, middle) + multi_query(coin_name,
      subreddit, middle, before);
end
return 0;
```

Algorithm 3.1: Divide et Impera implementation for getting all the posts in a subreddit for a particular coin

3.4 Natural Language Processing

Natural Language Processing is the application of computational techniques to analyse and synthesise natural language. The reason for using NLP was to do sentiment analysis on the scraped posts so that the overall sentiment that these posts reflect could be analysed in comparison to the price.

For this task, I used NLTK ("Natural Language Toolkit"), which is a suite of libraries that

Coin	Bitcointalk	Reddit	Total
BitShares (BTS)	723	462	1185
Bitcoin Cash (BCH)	1058	6164	7222
Dogecoin (DOGE)	20	18613	18633
Dragonchain (DRGN)	519	75	594
EOS (EOS)	36	3258	3294
Electroneum (ETN)	1163	85	1248
Ripple (XRP)	0	8764	8764
Stratis (STRAX)	1168	788	1956
TRON (TRX)	20	1739	1759
Verge (XVG)	8	2027	2035
Total	**4715**	**41975**	**46690**

Table 3.1: Number of posts scraped for each coin

has an implementation of a sentiment analysis that I needed. I made a Python program that went through all of the scraped posts, applied preprocessing such as switching all letters to lowercase for consistency or using the "ftfy" ("fixes text for you") module which fixes text encoding errors, and then applied the VADER ("Valence Aware Dictionary and sEntiment Reasoner") sentiment analysis tool from NLTK. The tool uses a dictionary that maps lexical features to sentiment scores. The tool returns a negative, positive and neutral sentiment score for each input, all bounded between 0 and 1. A compound score is also returned, which is an average of the previous three and is normalised between -1 and +1. The higher the score, the more positive the sentiment transmitted by the text is, and the lower the score, the more negative it is.

An example of a negative text is "Verge is dead. Im just terrified with the release of the wraith and all the problems and bugs waiting to wraith on us. Better close your position now or idk just gamble ur money.Its your money not mine." The post shows the author's concern with the coin Verge on the "vergecurrency" subreddit. The post was made on the 1st of January 2018, shortly after Verge reached its peak and started losing value fast. The post has a compound score of -0.9509, which means very negative, and the score is confirmable by just reading the text.

An example of a positive post was about Dogecoin, on the 4th of May 2021, four days before the coin reached its all-time high on the 8th of May. It has a compound score of 0.99, which is the far opposite of what the previously cited post was. It is a long post, but some snippets read as follows: "Dogecoin is virtually unstoppable now. With these kinds of gains, The Fear of missing out will be so great that people will feel like they have no choice but to put money in.", and "Everyone is starting to wake up to crypto making real money.". Both of these snippets show how strongly the author wants people to invest so that his coins will increase in value. They put a high emphasis on the fear of missing out (FOMO), which commonly makes people make decisions without being sure of the choice: "I'll say it again, this is the mother of all FOMOS. [...] This coin is changing people's lives[...]". Even with all of this artificially inflated hype, the coin went on to lose 75% of its value over the next 45 days, giving huge losses to people that surrendered to the FOMO and jumped in on the Dogecoin hype.

Chapter 4

Data Analysis

4.1 Sentiment Analysis

When looking at the results of sentiment analysis, the compound value is used to threshold the overall sentiment into three classes: negative, neutral and positive. Following the advice of the developers of VADER, based on the compound score we have:

- Negative for score ≤ -0.05
- Neutral for score > -0.05 and score < 0.05
- Positive for score ≥ 0.05

For the visualisation of this data, I decided to split the posts into two big groups: one for posts made before the peak of each coin and one for posts made after the peak. The coin-by-coin data can be seen in Table 4.1. Roughly 85% of all of the analysed posts show a neutral sentiment. The percentages at the bottom of the table show that either the data or the sentiment analyser have a bias towards positive and neutral sentiment. The number of negative posts after the peak seems to be higher, on average, than the number of negative posts before the peak. Since the data looks like it's biased, although the increase in negative posts is only 0.44%, increase in neutral posts relative to the total number of posts in that time period shows that indeed, people's opinion about the coins decreases slightly after the peak, when it starts to lose value.

A more important thing to look at is the shift in the number of neutral and positive posts from before the peak to after the peak. The percentage of positive posts relative to the total number of posts in the same period increases by almost 5%, while the percentage of neutral posts increases by 43.56%. This shift of sentiment from positive towards neutral, although small, shows that there is a small correlation between the sentiment of posts and when they're made in the scope of the lifespan of a supposed P%D. This shift could be caused by possible bad actors that try to influence others to buy during the pump period, but stop talking about it after they sell.

4.2 Using selected coins to find pump and dump-like events

In this section, I will present the methods I used for searching for pump and dump-like events in the scraped coin data using the list selected in Section 3.2. To be noted that finding these coins does not prove that they have been subject to such schemes, only that they show patterns that also appear in other coins that have been associated with P&D schemes in the past.

4.2.1 The parameters

By going through the coin data of the pre-selected list of coins and applying statistical analysis, I managed to get a statistical pattern. Table 4.2 shows the mean price, volume, and market cap

Coin	Before Peak			After Peak		
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
BitShares (BTS)	102	273	253	80	237	240
Bitcoin Cash (BCH)	292	668	513	899	2567	2283
Dogecoin (DOGE)	1299	5319	5269	793	3076	2877
Dragonchain (DRGN)	26	75	113	51	83	246
EOS (EOS)	80	311	188	389	1555	771
Electroneum (ETN)	226	223	493	78	82	146
Ripple (XRP)	551	1812	1517	573	2943	1368
Stratis (STRAX)	133	294	423	104	398	604
TRON (TRX)	130	331	391	166	386	355
Verge (XVG)	63	170	213	306	668	615
Total	2902	9476	9373	3439	11995	9505
Percentage	13.34%	43.56%	43.09%	13.78%	48.09%	38.11%

Table 4.1: Cryptocurrency sentiment analysis results

Value	**Pump Start to Peak (Peak Value) / (Start Value)**	**Peak to Dump End (Peak Value) / (End Value)**
Median Price Ratio	13.80	6.46
Price Ratio SD	37.95	1.91
Median Volume Ratio	38.53	12.97
Volume Ratio SD	289.08	22.42
Median Market Cap Ratio	14.23	0.16
Market Cap Ratio SD	38.95	0.06
Median Number of Days	39	84

Table 4.2: Median values extracted from the pre-selected coins

relative increase/decrease for each of the two periods in the lifespan of a P&D (pump start to peak and peak to dump end), and the median number of days that each of these takes. They also show the standard deviation (SD) for each value. The SD will be used as a helper for finding the coins that abide by the found patterns while making sure the pattern is not too generalistic and does not apply to all coins. The values, as indicated in the table header, represent the peak values divided to pump start values and dump end values respectively. Note that due to the market cap value being missing for some days or completely for some coins, the value was not used for searching other coins so that coins with incomplete data could still be searched.

The median price ratios are quite big considering the data is looking at long-term pumps. The prices increase almost 14 times from the start of the pump to a peak, and then they get divided by almost a factor of 7. These price movements are big and it will be easy to identify them in the coin data. Given that the P&D schemes we're looking for are long, the median numbers of days reflect this, with a median of 39 days from start to peak and an even longer 84 days from peak to end, the time period that a scheme runs for justifies how big the ratios are.

4.2.2 Searching for matching coins

Searching for matching coins was done by using another Python script. The program parses each coin's data and for each day, it checks a range of days after it (19 to 59 days after) to see if there's any day that has coin data that passes a threshold. If there is more than one day that passes the threshold, the day on which the coin had the biggest price is selected. This is the first part, which connects pump start days to peak days. For each of these pump start -¿ peak pairs,

the program then checks for days that are 45 to 125 days after the day of the peak, applying the same thresholding method as before.

The range of days the program checks for each starting day is determined by the median number of days for its respective period and half of the same number as the range (start from half of the median and check until 1.5 * median days after the starting day). This has the scope of accommodating for P%D schemes that are shorter or longer than the median.

The thresholds that the program checks for are based on the median values and the standard deviation for each. I've made 6 runs, adjusting the threshold by adding a multiple of the standard deviations to the median values (0 to 5 standard deviations). If any day's price/volume passes this threshold, it gets selected as a coin that matches the pattern. Assuming that price and volume ratios are normally distributed for these kinds of day ranges, the number of standard deviations being used for the threshold shows what percentage of the population the coins are in. Given that the 0 standard deviations run only returned only 12% of the coins that I had data for, the actual irregularity factor of the found coins is higher. In other words, the percentage of the population that the found coins lie in is smaller than what a usual normal distribution would imply. Table 4.3 shows how many coins each run found, which of them pass a minimum of $1000 trading volume, and what percentage of the population it lies in, assuming that the population is normally distributed.

When looking at the number of coins returned, after 2 standard deviations above the means there isn't a large drop in the number of found coins. This effect can have one or more of two causes:

1. The coins are less popular and much more vulnerable to price spikes. If a coin has little to no prior trading, a moderate transaction can drive the price up, even if there was no intention to pump it.

2. This ties a bit into the first point, but if the coin has a very low price, a big price spike doesn't imply a pump attempt. In the case of the coin Elite (1337), the 360x price increase is marked by a peak trading volume of just $83, with a volume ratio increase of almost 20000.

To get around this, a filter was applied to the end list of coins so that only coins with a trading volume bigger than $1000 at the start of the pump and the end of the dump were left. This helps trim off some of the low trading volume coins. Note that by how the algorithm works it checking that the volume at the peak is higher than $1000 is not necessary as it is always bigger than the volumes at the start and the end. The number of coins that pass this filter is shown on the 3rd column of Table 4.2. An example of a coin returned by this is shown in Figure 4.1. The visual pattern the graph shows is similar to how a P&D looks like: a 16-fold price increase followed by a decrease to one-16th of the peak price. The start and end of the scheme are on the left and right extremities of the X-axis. Anyway, due to how volatile the cryptocurrency market is and how difficult it is to pinpoint market manipulation just by data alone, the coins returned by the analysis have not necessarily been subject to P&D schemes, they just show the same patterns found in coins that have been associated with such schemes in the past.

4.3 Measuring correlation between market data and online activity

In this section, I will show what method of analysis I used to check for correlation and causation between the market data of the preselected coins and the online activity related to them (the scraped posts). The correlation metric I decided to go for was the Pearson correlation coefficient, which measures the strength of the linear relationship between two variables, ranging from -1 to

Standard Deviations	Number of coins	Nuber of coins with $> \$1000$ trading volume	Percentage of population
0	1331	417	50%
1	457	45	16%
2	292	20	2.5%
3	223	11	0.15%
4	171	7	$< 0.01\%$
5	140	4	$< 0.01\%$

Table 4.3: Number of coins that pass the thresholds when checked against the pre-selected coin medians and standard deviations

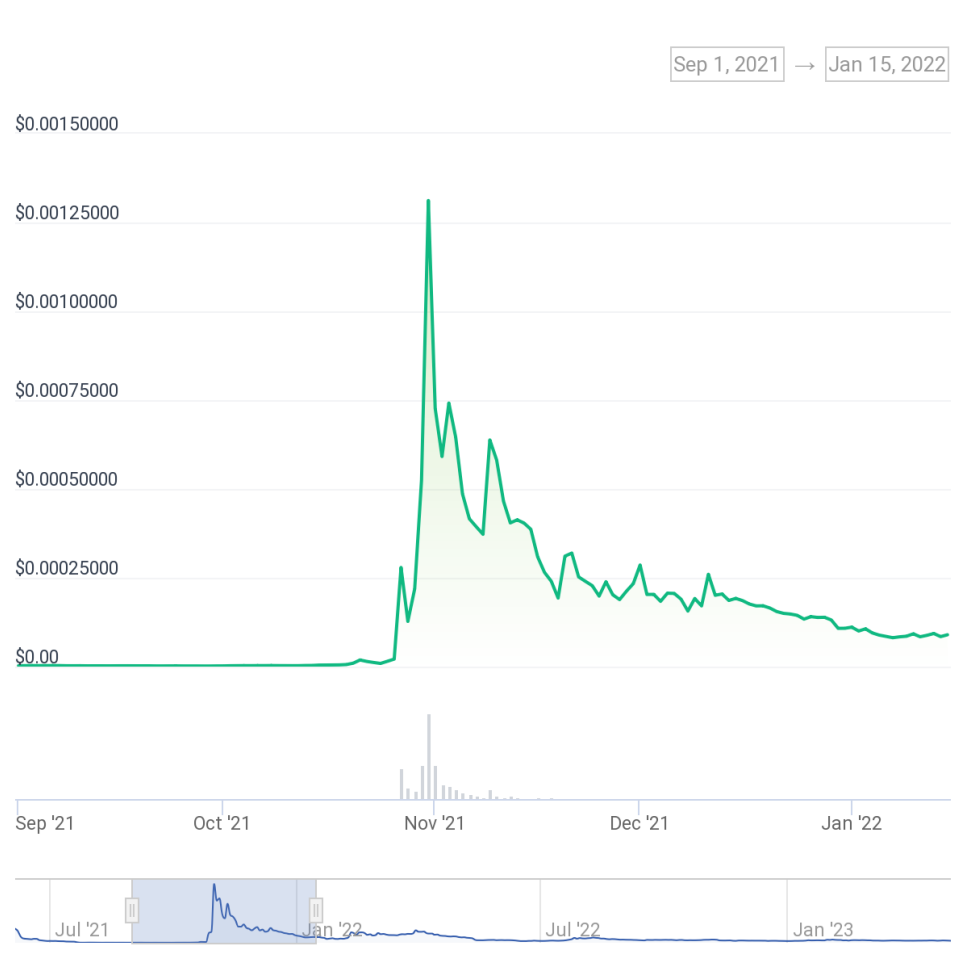

Figure 4.1: The price graph of CorgiCoin, a coin returned by the P&D pattern matching algorithm. Extracted from coingecko.com [8]

Coin Name	Pearson coefficient for no lag	Best Pearson	Lag days
BitShares (BTS)	0.037	0.075	5
Bitcoin Cash (BCH)	0.588	0.588	0
Dogecoin (DOGE)	0.361	0.389	1
Dragonchain (DRGN)	0.195	0.212	-1
EOS (EOS)	0	0	0
Electroneum (ETN)	0	0	0
Ripple (XRP)	0.561	0.595	2
Stratis (STRAX)	0.19	0.221	1
TRON (TRX)	0.874	0.874	0
Verge (XVG)	0.705	0.775	5

Table 4.4: Pearson coefficients for correlation between the number of mentions and the price of a cryptocurrency

+1. A negative value shows a negative linear correlation, meaning that an increase in one value correlates to a decrease in the other, while a positive value shows a positive linear correlation. The farther the value is from 0, the stronger the correlation is.

Correlation does not imply causation. In an attempt to check if there's causation, the method that I used is called "Non-directional lagged interactions" as presented by Shay Palachy Affek. [2] The method of doing this is by "lagging" the set of data that shows the number of daily posts each day, and then using the lagged data to check for correlation with the set of data containing the prices. For example, if the set containing posts lagged by 3 days has a high correlation with the original set of prices, it could imply that a change in the number of posts today causes a change in the price 3 days later.

Table 4.4 shows the correlation between the original sets of data. It also shows the best possible correlation value found by lagging the posts set by a range of days (from -5 to 5). A positive "Lag days" value shows that a change in the number of mentions is reflected after some days in the price, and a negative value says that the change in price is reflected in a change in the number of mentions.

Some of the coins show zero correlation, such as EOS and Electroneum. This is likely caused by there being no relation between the values whatsoever. The coin with the highest correlation is the coin TRON with a Pearson coefficient of 0.874. It shows a very strong correlation. Since the value is found for a lag day of 0. This means that the number of posts doesn't necessarily imply the price, they grow and shrink together. When looking at the coin Verge we can see that it is the coin with the highest Pearson coefficient that has a positive number of lag days. This means that a change in the number of posts is reflected strongly in the price around 5 days later. However, due to the Pearson coefficients being so close to each other, the causation can not be proven beyond reasonable doubt. Figure 4.2 shows a visualisation of the two data sets with no lag for the coin Verge. It shows how closely the two lines follow each other. Note that a possible causation factor is not that visible in the graph due to the long time the data covers. This long period means that it is unfeasible for the line reflecting the price to be constantly behind the line reflecting the number of posts.

Another coin to look at is Dragonchain, which has the best Pearson coefficient as a negative lag number of days. This goes to show that it is the price that is the one that causes changes in mentions. This causality is not strong though, as the Pearson coefficient is rather low.

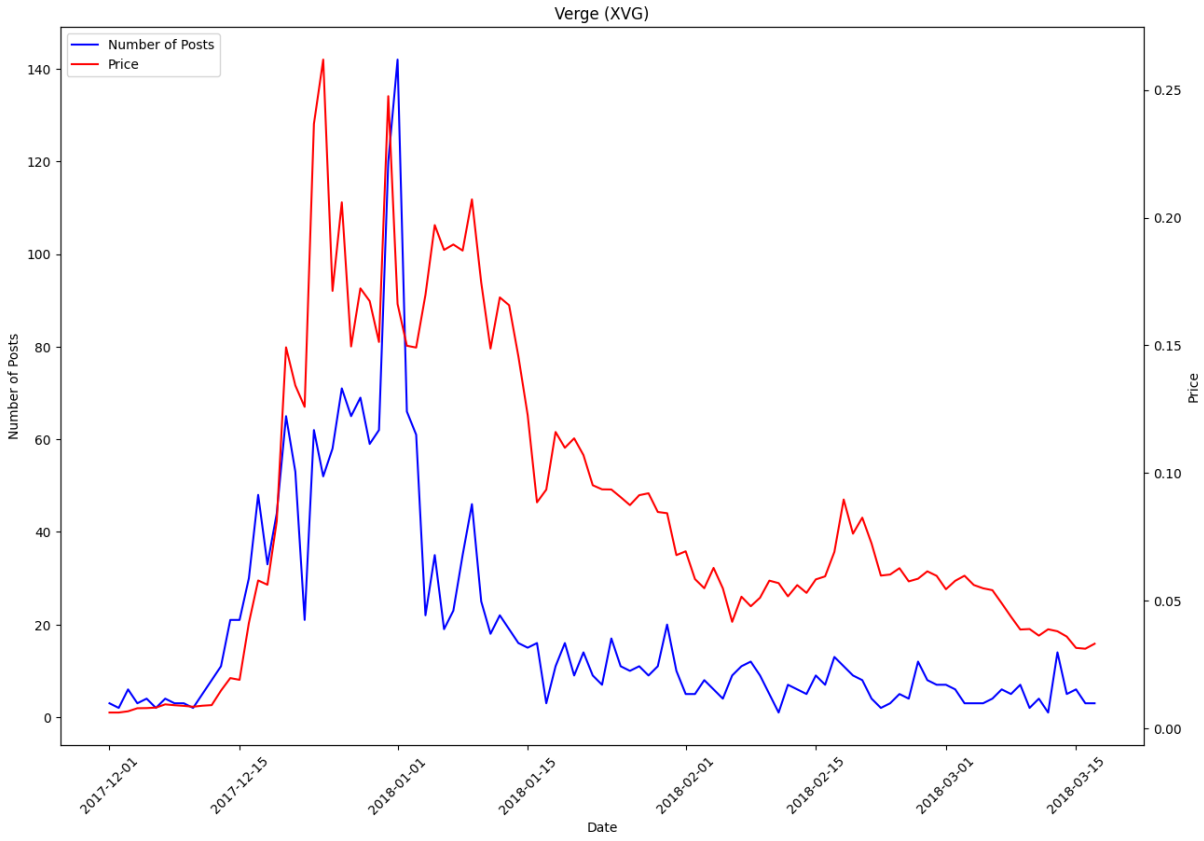

Figure 4.2: Visualisation of the correlation between the number of mentions and the price of Verge (XVG)

Chapter 5

Critical Evaluation

5.1 Coin Data Collection

This part went really well, and I am happy with the results. I managed to get a big enough list of coins from a reputable source. As said before, one limitation was the lack of granularity that CoinGecko's API provided but it was not a problem in the scope of the project. The sample size also showed all kinds of coins: from ones that aren't traded that much to the very popular ones. The selection and survivorship biases that the data could have shown also do not pose as a threat to the analysis that this project conducts.

5.2 Coin Selection

This is one section that I am not very happy with. The sample size of the 10 pre-selected coins is rather small. Had I had more time or focused on this section more I would have liked to get a sample size of at least 20-25 coins. One of the reasons behind the small sample size is the fact that there aren't many arrests being made on P&D schemes, especially the long-term ones, most mentions online of a coin being a P&D scheme being just allegations. I ended up including both coins that are subject to allegations, and ones that show clear P&D factors or legal action related to them. The whole topic is a bit in the grey area and the fact that cryptocurrencies offer so much anonymity does not help either. Due to this, the pattern that these coins form is rather restrictive and the results of searching for other P&Ds using it are affected too.

5.3 Scraping

When it comes to scraping, there are both good and bad outcomes. On one hand, the scrapers managed to get lots of mentions on the provided coins, with a total of almost 47,000 scraped posts, which is a good sample size. The whole execution is satisfactory, the actual coin collection time was low and the implementation of the scraping algorithms didn't take long.

On the other side, my data handling was sluggish, and it took a while to get it right. The JSON pattern that I chose to store the posts in was difficult to manage and took a lot of reorganising whenever I wanted to process it in some way. The fact that I've done the sentiment analysis after the scraping instead of while doing it also proved to be a bad choice as I had to insert the sentiment of each post inside the post's JSON bounds, and it proved to be quite buggy during development, but I managed to get it right in the end. Had I planned better I could've organised my data better and it would've been easier to manage it later.

Another limitation is that scraping Bitcointalk can not be reproduced continuously due to how I implemented it. As mentioned in Section 3.3, I used Google's Programmable Search Engine API to find mentions of the coins on Bitcointalk. This is an issue as the API's free tier is fairly restricted in terms of the number of available calls. In conclusion, the scraping of a larger

list of coins would need to either have the user pay or spread the data collection over multiple days. Had I planned better I could have tried to scrape the whole website in advance and only store the posts (stripping all HTML clutter). This would have allowed me to search for other coins in the already scraped data, rather than having to search again.

Another limitation that is present in my project and is related to this part is the small number of sources. An initial plan was to scrape news websites too, but it ended up not being worth the time as I would have either needed to use a news aggregator to get all the data from the same source, or create a web scraper for each website that I would have wanted to scrape, which would prove to be too much work for too little effort: think how many mentions of a particular coin can one news website have?

An issue that I should've planned better for was scraping Twitter. As a lot of the research on P&Ds in both the cryptocurrency and the stocks market points towards Twitter, it would have been sensible to scrape it too. Unfortunately, Twitter's API began to be very restrictive lately and I was also unable to get an API key, due to me having to pay for one. I also tried looking for other databases that store Twitter data and broadcast it to users, the same way Pushshift does with Reddit's data. This proved to be an issue, as I am certain that scraping Twitter for data would have improved my findings greatly.

5.4 Natural Language Processing

This part was fairly straightforward and didn't prove to be too difficult. The only issue I encountered was, as specified in Section 5.3, inserting the sentiment of a post into the posts file. A thing that I didn't know how to address and I'm not sure if is a problem is the fact that most of the posts that were long had a positive sentiment. I'm not sure whether this is because I was not using the right tool for the job or if it is just a normal thing that is happening in my data (longer posts being mostly made when having a positive thought about a coin).

5.5 Analysis

Due to the nature of the topic, the findings given by the analysis are mostly supposed ones, without being able to give a straight answer to all types of analysis I tried to make. I find the sentiment analysis part to be a good contribution, as there is an obvious decrease in the number of positive posts about a coin after the peak of a supposed P&D scheme. However, it is not possible to say whether the price affects the sentiment or the sentiment affects the price.

When looking at the coins matched by the parameters of the pre-selected coins, there was a good contribution in finding coins that look like they have been subjected to P&D schemes, and the found coins can be used for further analysis of possible schemes that might have happened before. Although the basis on which these coins were found is not foolproof, the extreme values shown by some of the found coins still go to show that something is happening to them. Again, we can not tell if they have been subjected to P&Ds but further research can improve on this.

Analysing the correlation between the number of posts and price gave a range of results. For some coins, the correlation was high, while for others the correlation was 0. Also, analysing causation didn't give great results, only a couple of the coins having a more-than-weak correlation between the lagged posts dataset and the coins dataset. One reason for this could be that the dataset of coins is too small. Maybe for a bigger dataset, there could be made a better analysis of whether online advertisements of coins affect the price or not.

Chapter 6

Conclusion

In this paper, I analysed previous research on cryptocurrency fraud, with an accent on Pump and Dump schemes. The paper shows information about P&D schemes in the stock market and analyses how they behave in the cryptocurrency one. It aims to fill a gap in the world of long-term P&D schemes, the ones which are usually done by either a rich investor with many assets that tries to turn a profit on their coin, or by people jumping on a hype train based on a cryptocurrency, who have been influenced to buy a coin and want to influence others to buy as well, by creating hype around said coin. The paper focuses on looking if online posts about a coin have a correlation with its price. It uses a list of pre-selected coins that have been associated with P&D schemes in the past and uses them as a basis for looking for other coins, while also searching for their mentions online for analysing correlation.

The analysis involves multiple techniques, such as using the Pearson correlation coefficient to check for correlation and using NLP for analysing how the sentiment around a coin changes as the coin progresses through a possible P&D scheme. Due to the grey area the topic of P&D schemes in the cryptocurrency market is currently in, with little to no evidence on coins that have truly been subjected to this kind of long-term scheme, it is difficult to say if the findings show that P&D schemes are conducted in this way. The correlations that have been found could just be the subject of normal reactions of people when seeing their coin increase or decrease in value.

6.1 Further work

Further work on this topic could be:

1. Including Twitter mentions in the research
2. Finding a way of using mainstream news websites in the research and seeing if they have a bigger impact on how a coin behaves
3. Researching for more cryptocurrencies that have been subjected to long-term P&D schemes and using them to form a larger dataset of confirmed fraudulent activity. It could maybe include only coins that have strong evidence and arrests around them, but that is difficult as of now.

Bibliography

[1] AARP. Ponzi and pyramid schemes, 2021. Accessed: 2023-04-28. URL: https://www.aarp.org/money/scams-fraud/info-2021/ponzi-vs-pyramid-schemes.html.

[2] Shay Palachy Affek. "inferring causality in time series data". 2019. URL: https://towardsdatascience.com/inferring-causality-in-time-series-data-b8b75fe52c46.

[3] Franklin Allen and Douglas Gale. Stock-price manipulation. *The Review of Financial Studies*, 1992.

[4] Andrew Beattie. The pioneers of financial fraud, 2022. Accessed: 2023-04-28. URL: https://www.investopedia.com/articles/financial-theory/09/history-of-fraud.asp.

[5] Biography. Jordan belfort - wife, children & facts. Accessed: 2023-04-27. URL: https://www.biography.com/business-leaders/jordan-belfort.

[6] Blockchain.com. Blockchain explorer, 2022. Accessed: 2023-04-28. URL: https://www.blockchain.com/explorer.

[7] Spencer Bogart. Bitcoin is (still) a demographic mega-trend: Data update, Dec 2020. Accessed: 2023-04-28. URL: https://blockchain.capital/bitcoin-is-still-a-demographic-mega-trend-data-update/.

[8] coingecko.com. Corgicoin price chart (corgi). URL: https://www.coingecko.com/en/coins/corgicoin.

[9] coinmarketcap.com. Price graph of Nebulas coin on CoinMarketCap . Accessed: 2023-04-29. URL: https://coinmarketcap.com/currencies/nebulas-token/nas/btc/.

[10] colibritrader. Pump and Dump Trading Strategy (The Ultimate Guide), 2021. URL: https://www.colibritrader.com/pump-and-dump-guide/.

[11] Commodity Futures Trading Commission. Cftc charges two individuals with multi-million dollar digital asset pump-and-dump scheme, 2021. URL: https://www.cftc.gov/PressRoom/PressReleases/8366-21.

[12] Federal Trade Commission. Fighting crypto scams with crypto intelligence, blockchain analytics ... Accessed: 2023-04-27. URL: https://newsroom.mastercard.com/news/perspectives/2022/ciphertrace-crypto-fraud-tracking/.

[13] Federal Trade Commission. Cryptocurrency buzz drives record investment scam losses, 2021. Accessed: 2023-04-28. URL: https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2021/05/cryptocurrency-buzz-drives-record-investment-scam-losses.

[14] Michael Crosby, Pradan Pattanayak, Sanjeev Verma, Vignesh Kalyanaraman, et al. Blockchain technology: Beyond bitcoin. *Applied Innovation*, 2(6-10):71, 2016.

[15] Ivan Cryptoslav. Centralized vs Decentralized Exchanges. Accessed: 2023-04-29. URL: https://coinmarketcap.com/alexandria/article/centralized-vs-decentralized-exchanges.

[16] The Economist. Who is satoshi nakamoto?, 2015. Accessed: 2023-04-28. URL: https://www.economist.com/the-economist-explains/2015/11/02/who-is-satoshi-nakamoto.

[17] Entrepreneur. How to build an audience on bitcointalk, the online forum. Accessed: 2023-04-27. URL: https://www.entrepreneur.com/money-finance/how-to-build-an-audience-on-bitcointalk-the-online-forum/316613.

[18] Bahar Farahani, Farshad Firouzi, and Markus Luecking. The convergence of iot and distributed ledger technologies (dlt): Opportunities, challenges, and solutions. *Journal of Network and Computer Applications*, 177:102936, 2021.

[19] Action Fraud. Cryptocurrency fraud leads to millions in losses so far this year. Accessed: 2023-04-27. URL: https://www.actionfraud.police.uk/news/cryptocurrency-fraud-leads-to-millions-in-losses-so-far-this-year.

[20] Simon Golstein. Coinbase suspends bitcoin cash, investigates insider trading, 2017. URL: https://www.financemagnates.com/cryptocurrency/news/coinbase-suspends-bitcoin-cash-investigate-insider-trading/.

[21] JT Hamrick, Farhang Rouhi, Arghya Mukherjee, Amir Feder, Neil Gandal, Tyler Moore, and Marie Vasek. An examination of the cryptocurrency pump-and-dump ecosystem. *Information Processing & Management*, 2021.

[22] Coryanne Hicks. Coryanne hicks, 2023. URL: https://www.forbes.com/advisor/investing/cryptocurrency/different-types-of-cryptocurrencies/.

[23] Yu Chuan Huang and Yao Jen Cheng. Stock manipulation and its effects: pump and dump versus stabilization. *Review of Quantitative Finance and Accounting*, 2015.

[24] The Independent. Bitcoin: How much is satoshi nakamoto worth?, 11 2022. Accessed: 2023-04-28. URL: https://www.independent.co.uk/life-style/gadgets-and-tech/bitcoin-satoshi-nakamoto-wealth-net-worth-b1957878.html.

[25] Investopedia. Who is jordan belfort, the wolf of wall street? Accessed: 2023-04-27. URL: https://www.investopedia.com/investing/who-is-jordan-belfort/.

[26] Michael Lewis. Jonathan lebed's extracurricular activities, 2001. URL: https://www.nytimes.com/2001/02/25/magazine/jonathan-lebed-s-extracurricular-activities.html.

[27] Tao Li, Donghwa Shin, and Baolian Wang. Cryptocurrency pump-and-dump schemes. *Available at SSRN 3267041*, 2021.

[28] University of Cambridge. 3rd global cryptoasset benchmarking study, 2020. Accessed: 2023-04-28. URL: https://www.jbs.cam.ac.uk/wp-content/uploads/2020/12/2020-ccaf-3rd-global-cryptoasset-benchmarking-study.pdf.

[29] National Museum of Crime and Punishment. Jordan belfort. Accessed: 2023-04-27. URL: https://www.crimemuseum.org/crime-library/white-collar-crime/jordan-belfort/.

[30] Michel Rauchs, Garrick Hileman, et al. Global cryptocurrency benchmarking study. *Cambridge Centre for Alternative Finance Reports*, 2017.

[31] James Rubin. Dutch police arrest three in pump-and-dump scheme involving self-made cryptocurrency, 2021. URL: https://www.coindesk.com/markets/2021/06/05/dutch-police-arrest-three-in-pump-and-dump-scheme-involving-self-made-cryptocurrency/.

[32] U.S. Securities and Exchange Commission. Pump-and-dump schemes, 2021. Accessed: 2023-04-28. URL: https://www.investor.gov/introduction-investing/investing-basics/glossary/pump-and-dump-schemes.

[33] Statista. Estimate of the monthly number of cryptocurrency users worldwide 2016-2022, 2023. Accessed: 2023-04-28. URL: https://www.statista.com/statistics/1202503/global-cryptocurrency-user-base/.

[34] TheStreet. How jordan belfort became the wolf of wall street. Accessed: 2023-04-27. URL: https://www.thestreet.com/video/how-did-wolf-wall-street-make-his-money.

[35] u/reagsters. Crypto scams : Rug-pulls, pump n' dump, and pre-launch coin sales, 2021. Accessed: 2023-04-28. URL: https://www.reddit.com/r/CryptoCurrency/comments/q1kqz2/crypto_scams_rugpulls_pump_n_dump_and_prelaunch/.

[36] Friedhelm Victor and Tanja Hagemann. Cryptocurrency pump and dump schemes: Quantification and detection. In *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2019.

[37] Peter Wind. Elon musk & dogecoin — how did the world's richest man get involved with a meme cryptocurrency?, 2023. URL: https://coincodex.com/article/21927/elon-musk-dogecoin/.

[38] Jiahua Xu, Benjamin Livshits, et al. The anatomy of a cryptocurrency pump-and-dump scheme. In *USENIX Security Symposium*, pages 1609–1625, 2019.