The Predictive Power of Social Media within Cryptocurrency Markets

Ross Christopher Phillips

A thesis submitted for the degree of

Doctor of Philosophy

Department of Computer Science

University College London

February 2019

I, Ross Christopher Phillips, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Blockchain technology has generated a great deal of interest in recent years, as has the associated area of cryptocurrency trading, not only on the part of individuals but also from traditional financial institutions and hedge funds. However, there is currently limited knowledge as to how to predict future cryptocurrency price movements. This thesis investigates whether online indicators, especially from social media, can be harnessed to predict cryptocurrency price movements – to achieve this, three experiments are conducted.

The first experiment analyses time-evolving relationships between chosen online indicators and associated cryptocurrency prices; relationships are considered over short, medium and long-term durations. The work introduces and evaluates several influential factors from the social media platform Reddit, a platform previously unexplored within cryptocurrency prediction literature. It is found that medium and longer-term relationships strengthen in bubble market regimes (compared to non-bubble regimes).

The second experiment utilises these promising new factors as inputs to a predictive model. The model used was originally designed to detect influenza epidemic outbreaks, and is repurposed here to model epidemic-like cryptocurrency price bubbles, demonstrating how social media can be used to track the epidemic spread of an investment idea. The predictive power of the model is validated through the generation of a profitable trading strategy.

Having considered quantitative count-based metrics in the previous chapters (e.g. *posts per day*, *submissions per day*, *new authors per day* etc.), the next experiment considers the content of social media submissions. More specifically, the third experiment analyses social media submission content to investigate whether certain topics of discussion precede upcoming shorter term (positive or negative) price movements. Information evidencing time-varying interest in various topics is retrieved from social media submissions, upon which hidden interactions with the associated cryptocurrency price are deciphered. It is found that certain topics precede major positive or negative price movements, and also additional analysis shows that certain discussion topics exhibit longer-term relationships with cryptocurrency market prices.

Impact statement

Cryptocurrencies and related blockchain technology have recently experienced an explosion in interest, not only from within academia and industry—the area has also become topical mainstream news. Due to this rapidly increasing interest from all angles, the work presented here has the potential to be significantly impactful in a wide range of areas.

The work provides several benefits within academia. Firstly, this work provides a foundation, in the relatively unexplored academic area of cryptocurrency price and bubble prediction, upon which further academic work could be undertaken; proposals for such extensions are provided. Secondly, the work presents a new and unexplored data source (Reddit) as a valuable source of information in cryptocurrency-related prediction. Thirdly, the work contributes to longer standing epidemic-based speculative asset bubble literature, demonstrating, in an area where data is usually hard to source, how social media can track the spread of an investment idea.

Outside of academia, there has been a flurry of activity in the cryptocurrency and blockchain area from a range of entities, including central banks, trading institutions and technology companies. The techniques for predicting price movements contained in this thesis could aid central banks to design and track their own cryptocurrencies and aid private companies in their timely purchase of cryptocurrencies, either for resale profit or for accumulation for later use of associated blockchain technology. Furthermore, the ability discovered here to detect if a speculative bubble is occurring may provide a warning to non-professional investors that any observed price rises may be the result of a (potentially unsustainable) bubble.

By design, the use of cryptocurrencies is not limited to one geographical area or jurisdiction. This inherently global nature of cryptocurrencies results in a global impact for this research. The academic impact of the research is brought about through dissemination in scholarly journals and conference proceedings. When available, a peer-reviewed open-access journal is chosen to allow the most effective and accessible dissemination. The publication of the work is also complemented by presentations in a range of contexts (for example, conference and internal presentations, and in an interview with a cryptocurrency news website).

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Denise Gorse, for her guidance, unfailing availability and scientific wisdom throughout the PhD. I appreciate her encouragement to follow my interests, especially given that cryptocurrency analytic research was a relatively new academic area when I commenced my research. I hope that Denise's mastery of English grammar has worn off on me, at least slightly, through her detailed improvements of my texts.

I am also thankful to the examiners of my PhD transfer viva, Prof. Tomaso Aste and Dr. Paolo Tasca. Their insightful recommendations, given their areas of expertise, provided important finishing touches and contributed towards a more robust final thesis. In addition I would like to acknowledge the academic, financial and industrial support provided by the Financial Computing and Analytics Centre for Doctoral Training and, more broadly, University College London.

Finally, but by no means least, I would like to express my appreciation to Julia, Paul and Rachael. Their unequivocal support and encouragement, not only during the PhD, will always be appreciated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation for this research

Cryptocurrencies, of which Bitcoin [1] is the most well-known, have recently experienced a new wave of interest. It has become commonplace to see TV coverage, news articles, blog posts and discussion on social media platforms about cryptocurrencies and related (blockchain) technology. Since the introduction of Bitcoin in 2008, thousands of other cryptocurrencies have come into existence; the cryptocurrencies often aim to serve a specific purpose or provide some particular functionality. As well as triggering the launch of more cryptocurrencies and creating mainstream excitement, there has been a flurry of activity from a range of interested parties.

Central banks have investigated the possibility of launching their own cryptocurrencies. For example, the Bank of England worked with researchers to design their own cryptocurrency called RSCoin [2] which has some decentralised properties while some aspects remain centralised (monetary supply and issuance). Also, Russia has recently announced the intention to issue their own cryptocurrency, referred to by the press as the 'CryptoRuble'.

With the emergence of blockchain-based smart contracts, many existing companies have publically announced their interest in researching and integrating blockchain technologies into existing business processes. Several consortiums have been set up to amalgamate the research efforts of otherwise competing companies. One example is the R3 consortium (https://www.r3.com/) which combines the research and development efforts of financial institutions while investigating the integration of blockchain technology in financial use cases. Another example, the Ethereum Enterprise Alliance (https://entethalliance.org/), is a partnership boasting members from a range of industries including investment banks (e.g. Credit Suisse and JP Morgan), technology companies (e.g. Microsoft and Intel), energy companies (e.g. BP) and a plethora of start-ups.

Traditional financial institutions and hedge funds, new and existing, have allocated funds for investment and trading within cryptocurrency markets, attracted to cryptocurrencies by the substantial returns and volatility seen historically [3]. The strategies engaged range from fundamental-based investment in new projects (through initial coin offerings), fundamental-based investment in existing cryptocurrencies (through holding a basket of the usually well-established cryptocurrencies), and more active trading strategies attempting to generate buy and sell signals and shorter-term holding periods. Growing interest in trading cryptocurrencies is not limited to trading firms; the number of individuals trading cryptocurrencies has grown significantly over recent years. It was reported in late 2017 that a leading US-based retail cryptocurrency exchange (Coinbase) was receiving up to 100,000 new user registrations per day.[1]

While some of the above groups are not directly interested in profiting from cryptocurrency price changes, knowledge of price dynamics is vital for each group for their own reasons. As a first example, consider the case of a national central bank issued cryptocurrency. Here, stability (lack of substantial price movements/volatility) would be vital and would be expected to be a consideration of the cryptocurrency design (RSCoin attempted one such mechanism to maintain price stability). However, stability is not a luxury available with existing cryptocurrencies. Having an understanding of what drives the price of current cryptocurrencies would aid the development of future more stable ones, potentially ones to be issued by central banks. As a second example, companies deploying smart contracts to a public blockchain usually have to pay to deploy and interact with them; payment is made in the cryptocurrency associated with that blockchain. Therefore companies running these applications (who otherwise may not

[1] https://www.bloomberg.com/news/articles/2017-11-02/bitcoin-exchange-added-100-000-users-in-a-day-as-price-exploded

have an interest in trading cryptocurrencies) have to make decisions on maintaining their cryptocurrency funds, including when is the best time to buy their cryptocurrency. As a third and final example, it is very obvious that entities trading cryptocurrencies for profit would have an interest in predicting future price movements.

One starting point in understanding what drives cryptocurrency prices is to understand what asset class cryptocurrencies behave like, as different asset classes are likely to have different inherent drivers of their price. The original intention was for Bitcoin to be an alternative to traditional money; however due to a confluence of reasons it arguably does not currently exhibit the attributes required to be classed as money (a more comprehensive explanation of the attributes of money and Bitcoin's lack of them will be provided in Chapter 2). Whether Bitcoin acts like money may change as the technology is updated. Due to the relative newness of the entire cryptocurrency ecosystem many cryptocurrencies, including Bitcoin, are unfinished projects and although they may have working versions, they have broader aspirations. Investors often buy cryptocurrencies based on what could be achieved in the future rather than what is presently possible. In fact much research attempting to define what cryptocurrencies are—or at least why cryptocurrencies are owned—concludes that cryptocurrencies have many of the traits of speculative investment vehicles [4] [5]. Given the current speculative nature of cryptocurrency trading and investment, price changes, at least to some extent, could be determined by investors' changing interest in different cryptocurrency projects; online indicators and social media usage could offer one such way to track interest. The ability of online indicators, especially social media indicators, to predict more traditional financial asset prices (e.g. company stocks) has been covered thoroughly within the literature (a survey of current techniques and research avenues is provided in [6]). It seems intuitive that such indicators will provide at least as good predictive power within cryptocurrency markets as in more traditional financial markets, for a number of reasons described below.

First, the majority of trading volume in traditional financial markets is generated by large companies (e.g. proprietary trading shops, high-frequency trading firms, investment banks and pension funds). Employees of such large companies would turn to colleagues to discuss investment hypotheses rather than publically on online social media platforms. Conversely cryptocurrency markets have historically been traded more by individuals. Individual home traders who—lacking both proximity to colleagues with whom they can discuss ideas and contractual obligations preventing them from doing so publically—are more likely to turn to online platforms to discuss and research investment ideas.

Second, a lack of professional certification and consistent regulation, combined with the anonymity of trader identities, can cultivate an environment where people are more likely to create social media activity around cryptocurrencies they own (informally termed in finance as “talking your book”). This activity ranges from creating genuine discussion relating to a particular project to promising future price rises based on false information (extreme cases are referred to as ‘pump and dump’ schemes [7] and are reprimanded in more traditional markets). The volume of messages on social media relating to a particular cryptocurrency (and how this number changes over time) could indicate the number of people expecting a future price rise.

Third, cryptocurrencies are often described as internet currencies [8] and are traded online through private exchange websites. Information about pertinent events (e.g. development progress, hacks and partnerships) disseminates through blog posts, social media content and chat rooms. Some of the trading exchanges even have integrated public chat rooms that facilitate discussion between traders. By contrast most stocks are traded through electronic submission of orders from desktop applications (some of which have built-in market news feeds), thus reducing (though not completely) a reliance on the internet to disseminate information.

Fourth, the lack of barriers to entry into the cryptocurrency sphere (free data streams, free exchange access), combined with aspirational stories of novices making fortunes, attract many with limited trading or financial background to cryptocurrency trading. The demand of such individuals for learning material and advice on cryptocurrencies is likely to be higher than an experienced stock trader trading a new stock as part of their portfolio, and whereas stock markets have private analyst written reports distributed internally and to selected clients, cryptocurrency commentary and knowledge is predominantly documented online. The accessing of such online material and the new arrival of market participants may produce a visible digital footprint which can provide valuable information.

Previous research has found a relationship between social media activity and Bitcoin price [9] whereby price and social media activity appear to reinforce one another in a positive feedback loop. It was found that as the price of Bitcoin rises, so does the volume of social media activity relating to Bitcoin, further increasing the price. This dynamic creates the possibility of an asset price bubble[2] as seen in other financial markets (notable examples include the internet bubble [10] and housing price bubbles [11]). This feedback loop may be one of the factors that have caused a number of past Bitcoin price bubbles [12]. Bubbles have not only been noted by academics; it is common in (possibly sensationalist) news articles to claim—rightly or wrongly—that a particular cryptocurrency is experiencing a bubble.

Price bubbles provide a number of characteristics important to the above interested parties including positive price dynamics (significant returns being attractive to traders) and negative price dynamics (volatility and lack of stability being something central banks and companies using blockchain technology would ideally want to avoid). As a result having more knowledge of what causes cryptocurrency price bubbles, or at least how to predict them, can go a long way towards understanding and predicting cryptocurrency price movements.

1.2 Research objectives

Given the current speculative nature of cryptocurrencies and the link between cryptocurrency markets and social media (both of these having been mentioned above and to be covered thoroughly in Chapter 2), the hypothesis explored here is that social media can be harnessed, in various ways, to gain information predictive of future cryptocurrency price movements. Three experiments are constructed to investigate this hypothesis:

1. **Do informative relationships exist between online indicators and cryptocurrency price movements?**

   The first experiment analyses the correlations between chosen online indicators and cryptocurrency prices, to discover whether relationships of value exist. Where relationships are found, it is checked whether they are leading relationships (i.e. whether the considered online indicator is leading price movements). The relationships are evaluated over different

[2] It should be noted there is little academic agreement on what exactly constitutes a financial bubble. As a result different sources are likely to be using different definitions and therefore the term ‘bubble’ is used loosely here. This disagreement and different bubble detection mechanisms are covered in more detail in Section 4.1.1.

time durations—short, medium and long term—to determine within which intervals relationships exist. Particular attention is given to whether relationships change dependent on the current trading market regime. Due to the prevalence and implications of price bubbles in cryptocurrency markets, two regimes are considered, bubble and non-bubble.

2. **Can the discovered indicators be used to predict cryptocurrency price bubbles?**

   The second experiment utilises the promising social media indicators identified in the first experiment as inputs into a predictive model. Given the relationships discovered in the first experiment between social media use and cryptocurrency price bubbles, the model attempts to harness social media use to predict the transition of the cryptocurrency market into and out of price bubble-like regimes. The value of the indicators and the model are further validated through the generation of a trading strategy that is evaluated using historical data.

3. **Are specific discussion topics, occurring on social media, indicative of intraday price movements?**

   The third experiment analyses the textual content of social media submissions. Having seen in the previous experiments that count-based indicators are more predictive over the medium and longer term (experiment 1), especially in the successful prediction of price bubbles (experiment 2), the focus here moves to content-based indicators to investigate whether shorter-term price movements can in this way be predicted. To achieve this aim, topics of discussion within social media submissions are analysed, identifying which discussion topics lead and lag both positive and negative price movements. Overall this experiment gives an understanding of which topics are indicative of upcoming price movements. The experiment also contains additional analysis which identifies that longer-term relationships between particular discussion topics and certain cryptocurrency price changes exist.

Each of the above tasks (and the overall aim of the research) involves the analysis of data sourced from online social media. As discussed more comprehensively in Chapter 3, several online data sources are used in this work, including a relatively novel and unexplored social media data source, Reddit. Thus, this work highlights the value of a little-used data source and provides evidence that it is a valuable source of information in cryptocurrency markets.

1.3 Thesis structure

The structure of this thesis is as follows. Chapter 2 reviews the background and literature relevant to this research, including cryptocurrencies, cryptocurrency trading and the use of social media to predict cryptocurrency prices. Chapter 3 outlines the data sources used including a review of the social media data sources available and justification of the final social media data sources chosen. Chapter 4 investigates whether relationships between cryptocurrency prices and online indicator based factors exist, and investigates whether such relationships change over time or with the market regime (in particular, in bubble regimes). Chapter 5 produces a model able to predict price movements by harnessing the promising indicators discovered in the previous chapter. Chapter 6 investigates whether the topic-based content of relevant social media discussion can be indicative of upcoming upward or downward price movements. Chapter 7 summarises the results of the research and outlines potential future extensions.

1.4 Related publications

The following publications resulted from the work in this thesis:

- R. Phillips and D. Gorse, "Cryptocurrency price drivers: Wavelet coherence analysis revisited", PLoS ONE, vol. 13, no. 4, 2018.
    - The above paper resulted in an approach for an interview to discuss its results and to provide general views on the cryptocurrency market: Goborov, A. (2018). Crypto Prices Can Be Predicted, Says Science. [online] U Today. Available at: https://u.today/crypto-prices-can-be-predicted-says-science [Accessed 18 Jan. 2019].

- R. Phillips and D. Gorse, "Predicting cryptocurrency price bubbles using social media data and epidemic modelling", IEEE Symposium Series on Computational Intelligence (SSCI), 2017.

- R. Phillips and D. Gorse, “Mutual-excitation of cryptocurrency market returns and social media topics”, 3rd International Conference on Knowledge Engineering and Applications (ICKEA), 2018.
    - Awarded one of three awards: Best Presentation (includes free entry to ICKEA 2019).

Chapter 2

Background and Literature Review

This chapter introduces necessary background and discusses literature relevant to themes present in this research. The chapter gives an introduction to cryptocurrencies, and how they are traded, before examining existing attempts to predict their prices. Particular attention is paid to past uses, documented within the literature, of social media to predict cryptocurrency prices. This chapter covers background and literature of relevance throughout the thesis. Where the background or literature is relevant only to an individual chapter, it is presented in that chapter.

2.1 An introduction to cryptocurrencies

In 2008 a pseudonymous author, under the name Satoshi Nakamoto, posted a paper outlining ‘a purely peer-to-peer version of electronic cash [that] would allow online payments to be sent directly from one party to another without going through a financial institution.’ [1]. Key points in that statement that were developed further in the paper were Bitcoin's aim to be both an electronic alternative to traditional fiat (state-backed) currency, and a new way to transact value without going through traditional payment intermediaries. Given these aims, the introduction of Bitcoin was well timed, although likely unintentionally, coinciding with the global financial crisis, which led some to mistrust traditional financial systems and government policy [13].

Nakamoto's paper proposed three components which, when combined, allowed the system to function without a central party in control: (1) a unit of currency (a Bitcoin); (2) a mechanism for exchanging the currency (the Bitcoin network); and (3) a mechanism for recording transactions (the blockchain). The following paragraph gives a high-level overview of how these components come together to allow the Bitcoin network to function; a more comprehensive explanation can be found in the original Bitcoin paper [1] or in the growing literature dedicated to blockchain design, for example [14] and [15].

The function of the Bitcoin network can be illustrated via the example of a transfer of an amount of Bitcoin. A participant, user A, can transfer ownership of an amount of Bitcoin to another user, user B, by generating a transaction. The transaction is signed by user A’s private key which is used to prove they have ownership of the Bitcoin; the transaction also includes user B’s public address. The transaction is then broadcast to the network where it is validated by other network participants (called *‘miners’*) who compete to validate transactions in return for potential rewards; such a mechanism is called *proof of work* and is one of many possible decentralised consensus mechanisms. Once validated, the transaction is written to the blockchain, a publicly visible ever-expanding list of all transactions. Due to the public recording of all transactions, miners can check that Bitcoins are not being double spent (the same Bitcoin being transferred twice), thus removing the need for a third party, such as a bank, to do such validation.

There is no need for a centralised party to ensure the correct functioning of the network because the miners are competing to do so; they are incentivised by rewards in the form of newly generated Bitcoins and optional transaction fees on each transaction. The system is decentralised in the sense that there is no single party responsible for the validity of the network and no single point of failure. If one miner is removed from the network, the network will still function. If one miner, with below 50% of mining power, tries to lie to the network about the validity of a transaction, the other participants will be able to check this and easily overrule the malicious miner. The level of decentralisation depends on the mining power that competing miners contribute; if for example, a cryptocurrency had one single miner, it would be centralised. It has been found that miners cluster together geographically [16]—likely in data centres [17] and commonly in Chinese ‘mining farms’ [18].

Although Bitcoin was the first cryptocurrency, a number of other cryptocurrency projects were launched over the following years. In 2013 Vitalik Buterin released a whitepaper for Ethereum [19], which suggested a new blockchain system that included a Turing-complete programming language, enabling users to build logic to be run on the blockchain. Ethereum not only allowed ownership to be recorded, as is the case within Bitcoin, but also allowed code to be executed. This allowed developers to design and deploy pieces of code, named *smart contracts*, that could run on blockchain nodes anywhere in the world. Smart contracts are pre-defined code specified agreements, automatically enforced by blockchain miners without relying on the trust of a single party or authority (for a more detailed description of smart contracts, see [20]). A smart contract, or a set of smart contracts, designed for a particular purpose is commonly described as a decentralised application (or dApp—although different capitalisation styles are commonly used). A detailed survey of available dApps, as of late 2018, can be found in [21].

One common use of smart contracts is the creation of tokens running on top of the Ethereum blockchain, an example being the GNT token. These tokens are essentially their own cryptocurrency and have their own amount of supply and their own value. By following a set of standards (for example, ERC-20), it is possible for anyone to design and create a token which can be stored in an Ethereum wallet[3]. It has become common for projects building a dApp on top of Ethereum to issue an associated token which is somehow connected with the use of the dApp (for example, paying for the use of the dApp). A token on its own may have no value, but the association of a token with the use of a particular dApp may give it value (for example, GNT is associated with the use of the Golem network[4], a dApp running on Ethereum).

The ease of creating tokens has caused a proliferation of tokens in recent years. Commonly associated with the creation of a token is an Initial Coin Offering ('ICO') whereby a new token is created by a cryptocurrency project team and exchanged for more established cryptocurrencies or fiat currency. A team, usually at the early stages of a project, will outline what their technology is and how the token will be used within that technology (as mentioned above; this is what gives the token at least *perceived* value and sets the terms of how participants can

[3] Almost all blockchain systems have a wallet component. Wallets can be used to store cryptocurrencies. Wallets, and therefore the cryptocurrencies they hold, are in the control of whoever has the private key to the wallet.
[4] https://golem.network/

acquire the new token. The team usually retains a percentage of the new tokens and receives more established tokens to fund their work.

An associated area of research has been the investigation of the attributes of a successful ICO. It has been reported that successful ICOs are more likely to have public code repositories and operate pre-sales [22] [23], and also have high levels of community engagement and previous venture capital investment [23]**.** Once a project has met its funding goals, it is likely to release the tokens to participants, and pursue the listing of the token on trading exchanges.

The growth of interest in ICOs is likely to have partially fuelled the very significant growth of the overall cryptocurrency market in 2017 [24]. In 2017, the market capitalisation of the whole cryptocurrency market rose from $18 billion at the start of the year to $600 billion at the end. The price rises created huge interest from those previously unaccustomed to cryptocurrencies or blockchain technology. For example in May 2017 the ICO for Basic Attention Token (BAT) was fully funded within two Ethereum blocks (30 seconds), raising around $35 million. The speed at which the ICO completed created a sense of urgency, which resulted in participants scrambling to get into future ICOs, some of which managed to sell out equally quickly.

Again in 2017, blockchain forks also became increasingly popular. As many cryptocurrency and blockchain projects are open source, anyone can copy and change the code, and if they can get the backing of enough miners, create their own version of the blockchain (for more details on the technicalities, see [25]). By splitting a blockchain in two, any cryptocurrency associated with the original fork would then exist on the second one. Often ignoring the technical reasons for a fork, general cryptocurrency market speculators would buy cryptocurrencies pre-fork so that they could have two cryptocurrencies (from the original and new chain) after the fork. Although it should be expected that the combined value of the assets after the fork would be, at most, equal to the value before the fork, the post-fork combined valuation was often more. It has been found that a fork is sometimes conducted by people previously outside the original project's community [26]; forks also appear sometimes to be aimed at the creation of additional value rather than bringing significant technical changes.

Both ICOs and forks are likely to have increased the overall cryptocurrency market capitalisation. Participation in an ICO usually requires Ethereum or Bitcoin. Fiat would be used to purchase the required Bitcoin or Ethereum (increasing the price), and these cryptocurrencies would be sent to the ICO to be locked up for a certain time, depending on the details of the ICO. Every time a new ICO or fork was added to the cryptocurrency data provision websites, there would be an increase in overall market capitalisation.

After reaching a peak cryptocurrency market capitalisation of around $830 billion in early January 2018 prices started to decline and did so for most of the year (barring a few short upward rallies). The total market capitalisation ended 2018 at $125 billion. Some previously funded projects closed down, disappeared, or stopped developing new code. Furthermore, the interest in new projects and ICOs has declined significantly; in 2018, 58% of ICOs failed to raise capital, disappeared, or refunded their participants [27].

However prices should not be considered the only indicator for the overall health of the cryptocurrency ecosystem. Attendance at cryptocurrency conferences is increasing, and the development activities of many projects continue (mostly those projects that set out with a technical goal in mind, rather than financial gains). Whereas ‘HODL’ was a term formerly commonly used by the cryptocurrency community (referring to holding onto your cryptocurrencies—see [28] for origin), many have now adopted the new term ‘BUIDL,’ urging those in the space to focus on building technology rather than waiting for prices to go up.

2.1.1 Cryptocurrencies as money

The high volatility seen over 2017 and 2018 and also in previous years raises the important question: does an asset with such volatility have the attributes required of money? Though there are a plethora of cryptocurrency and token designs (as mentioned in the previous section), the focus of this section will be on Bitcoin. This choice is made because Bitcoin is the most well-known cryptocurrency, meaning it may be the closest to mainstream adoption. For example it is the coin most commonly focused by regulators who are looking to understand what cryptocurrencies *as a whole* are[5]. Furthermore, the original aim of Bitcoin was to function as alternative money, whereas many other cryptocurrencies do not have this as a stated aim. To

[5] The choice of regulators to focus largely on Bitcoin when considering cryptocurrencies as a whole has its disadvantages – as discussed later in Section 2.1.2.

decide if Bitcoin can currently be considered as money it should be first understood what attributes an asset needs to have to be considered as money. Typically, there are three: store of value, unit of account, and medium of exchange. The applicability of Bitcoin to each one is detailed below:

1. **Store of value.** An asset meets this requirement if it retains its purchasing power over a sustained period. It is thus possible to save the asset for some time, then retrieve it and exchange it for a predictable amount of goods. Traditional examples include most fiat currencies (excluding currencies with high inflation), bonds and precious metals. The extreme volatility (orders of magnitude larger than seen in the foreign exchange markets [29]) of the price of Bitcoin limits its ability to be a reliable store of value and thus hinders its ability to meet this requirement, as well as the two that follow.

2. **Unit of account.** Assets meet this requirement if they provide a unit of measurement for recording and comparing value. It is common for those retailers who do accept Bitcoin as payment to either 1) rapidly and automatically update any Bitcoin price quotes to maintain a stable fiat currency price, or 2) quote the price in a stable fiat currency and then calculate the equivalent amount of Bitcoin at checkout (with the condition that the user has a limited time to purchase, or get a requoted price). These pricing mechanisms suggest Bitcoin is a poor unit of account [30].

3. **Medium of exchange.** Assets meet this requirement if they can be used as an intermediary in exchange, to avoid bartering directly for the exchange of two goods. Bartering requires both parties to value the item the other has. An asset meeting the ‘medium of exchange’ requirement can be used to value another item (in terms of the asset), and can then be spent by the receiver on items they desire. Bitcoin’s suitability as a medium of exchange is a complex topic, and includes the consideration of a number of factors which are discussed below:

    (a) **Real-time transactions.** It has become commonplace for those accepting Bitcoin transactions to require six block confirmations before a block of transactions is considered as entirely secure; with an average confirmation time of 10 minutes this requires on average 60 minutes. This delay reduces the possibility of real-time transactions without trust in the counterparty (some retailers are happy to trust the counterparty that a transaction is valid before block confirmations [31]).

(b) **Scalability.** Bitcoin blockchain blocks are currently limited in size to 1MB [1]. Given each transaction takes up a certain amount of space, a certain number of transactions can fit into each block, with blocks occurring, on average, every 10 minutes. This limitation of the maximum number of transactions that can occur, on average, every ten minutes means that the Bitcoin network is able to handle an estimated maximum of 7 transactions per second [32], with problems if any usage larger than this is required. One issue has already arisen over the last few years: there have been cases where a backlog of *pending transactions* build up. These are transactions sent to the network but not yet included in a block (the transactions being not yet verified). Miners choosing transactions to verify will choose the transactions that are willing to pay the miner a fee for mining them (the second source of income for the miners on top of the *block reward* paid to them by the Bitcoin network). As a result, in periods of congestion, it is likely those transactions paying a lesser fee will remain pending [33].

(c) **Energy consumption.** Bitcoin's proof of work consensus mechanism uses vast amounts of energy. Each miner is using computational power in an attempt to be the first miner to mine each block. For every block solved, one miner gets paid the reward; all other computation, by other miners, results in nothing. Estimates have indicated that the energy cost of the computation has been as much as $5 per transaction [34], and that the energy consumption of the Bitcoin network is equivalent to the energy consumption of some countries [35]. This high energy consumption has been seen as an obstacle for retailers considering adopting Bitcoin as a payment while wishing to remain as energy efficient and socially accountable as possible.

(d) **The use of Bitcoin (and other major cryptocurrencies) in ICOs.** As mentioned earlier, it is common for ICOs to receive Bitcoin or Ethereum in exchange for the token being generated. Sometimes one or both of these cryptocurrencies are the only way to fund a particular ICO (no other cryptocurrency or fiat currencies being allowed). The use of Bitcoin and Ethereum for these purposes may cause them to better meet the medium of exchange attribute, at least in this limited context, as they are here exchangeable for access to new ICOs.

Having reviewed the three properties—store of value, unit of account, and medium of exchange—commonly expected of money, it is clear that Bitcoin does not currently display them to a significant extent, outside of the ICO context. The ability of Bitcoin to exhibit these attributes may change over time, as at a particular point in time, or in a particular situation, any asset may meet the above criteria depending on the context of its use (e.g. cigarettes met all three of the above criteria within prisoner of war camps [36]).

Another aspect raised during a speech by the Head of Research at the Bank for International Settlements [37] is the need for finality of payments. It was argued that, though it is almost always the case that being included in the blockchain means a payment is confirmed, this certainty is never equal to that in traditional financial systems. This issue arises as the blockchain is the result of the miners agreeing on what is valid and what is not valid, and some miners may intend to rewrite history. It was pointed out that less popular cryptocurrencies may be more at risk of this, due to fewer miners securing their blockchain. An example of a miner rewriting history occurred in January 2019 when the Ethereum Classic blockchain (a fork of the more well-known Ethereum blockchain) was reorganised by a single miner that was temporarily able to control over 51% of the mining power. This control of the majority of mining power allowed the miner to dictate what state of the blockchain was valid. The attacking miner reorganised a number of blocks to introduce double spends of the Ethereum Classic cryptocurrency. The total amount of cryptocurrency gained by the attacker was valued at over $1 million[6].

[6] https://blog.coinbase.com/ethereum-classic-etc-is-currently-being-51-attacked-33be13ce32de

As well as academic work examining whether cryptocurrencies meet the attributes of currencies, there has been additional work by regulatory agencies, governments and parliamentary committees. For example in a report [38] produced at the request of the Economic and Monetary Affairs Committee of the European Parliament it was argued that cryptocurrencies were not exhibiting the attributes expected of money, and thus cryptocurrencies should be considered as speculative assets rather than currencies. Such an argument has implications, as the definition chosen by a jurisdiction is likely to influence the regulation that is applied. The following section reviews cryptocurrency regulation around the world.

2.1.2 Regulatory involvement and oversight

Although parliaments, governments, and regulatory agencies have already started investigating cryptocurrencies, with some issuing advice and regulations (to be reviewed in this section), many participants feel more regulatory certainty is required (e.g. clearer certainty on stances on cryptocurrency / more comprehensive and static guidance). This desire for increased regulatory certainty can be seen in a law firm's survey [39] of cryptocurrency market participants (including investors, traders and executives, among others): 84% of those who responded expressed a desire for greater regulatory certainty[7]. Also, 86% responded that the industry should self-police by developing common voluntary standards; as of December 2018; this is something that has been recognised as happening [40]. It is not only participants that feel the need for increased regulatory oversight and guidance, but regulators as well. The U.S. Securities and Exchange Commission (SEC) has listed cryptocurrencies as their 'top examination priority' for 2019 [41]. The UK government (via HMRC) is in the process of releasing more guidance material on the topic: on 19th December 2018 a report [42] was released advising individuals on the tax implications of cryptocurrencies, based on the current law, and guidance for businesses and companies is to follow in a later report.

Although there may be a desire by both participants and regulators for more regulatory involvement and regulatory certainty it is likely that the novelty and evolving nature of the

[7] It should be noted that these high percentages are possibly biased towards displaying the feelings of the type of cryptocurrency market participants who are in contact with and willing to respond to a survey by a law firm.

ecosystem has delayed regulation in many countries. Many countries have opted to wait for a better understanding of cryptocurrencies before applying regulation. The results of quickly applied regulation can be seen by examining the state of cryptocurrency regulation in the US. Multiple agencies, wanting to correctly and quickly regulate cryptocurrencies have developed different definitions of what a cryptocurrency is, and therefore, who should regulate them [43]. For example in the US cryptocurrencies have been classified as currencies (by FinCen), as property (by I.R.S), as commodities (by CFTC), and *mostly* as securities (by the SEC)[8]. Anyone participating in cryptocurrency markets needs to adhere to the appropriate regulation, and this overlapping regulation can overburden and cause complication. It has been suggested that this confusion probably arises from many of the agencies bundling all cryptocurrencies under the term ‘virtual currency’, and not understanding the intricacies and differences of each different cryptocurrency [44]: once one particular cryptocurrency appears to fall within the regulatory reach of a particular agency, all are assumed to do so. This has resulted in many regulators classifying cryptocurrencies, as a whole, to fall within the area within which they have oversight. The SEC is one agency that does not use the ‘virtual currency’ umbrella and reviews individual cryptocurrencies separately. They have ruled that Bitcoin and Ethereum are not securities, so not under the SEC’s regulatory scrutiny, but that it is likely most cryptocurrencies arising from token sales are securities due to the way they are initially marketed [45].

Even if there were regulatory certainty within one country, regulators in that country might face problems enforcing their policies due to the anonymity cryptocurrencies provide and the ability to send the cryptocurrencies anywhere in the world without the approval of government regulated entities (for example, the traditional banking systems). Cryptocurrency related companies, including the trading exchanges, prefer to locate in countries whose regulations are most agreeable to their business practices. For example it was reported in December 2018 that 38% of fiat-supporting service providers had chosen to close a location as a result of changes in that location's regulatory environment [40]. Indeed, cryptocurrency companies can find themselves treated very differently, depending on their location, as the level of regulation and acceptance varies greatly around the world (for examples, see [46]**).** An academic study on the

[8] Note that these are all federal level regulatory agencies, it gets even more complicated if considering that individual US states also have their own laws and viewpoints on how to regulate cryptocurrencies.

topic used even more granular groupings to categorise countries, including the categories: 1) ignoring; 2) monitoring; 3) recommending (for or against); 4) guiding; 5) integrating their use or banning them [47].

2.1.3 How cryptocurrencies are currently used

As well as issuing guidance, advice, and regulation, certain government agencies are also actively monitoring the use of cryptocurrencies via their public blockchains. One such agency, the U.S. Drug Enforcement Administration (DEA), which has reportedly been monitoring blockchain activity for many years, commented in August 2018 that price speculation had replaced illegal activities as the main purpose of most transactions [48].

In similar academic research analysing usage patterns, researchers have investigated the motivations of those becoming involved in Bitcoin to try and understand how the majority of users view it. One study attempted to decipher new users' intentions when first purchasing Bitcoin [4]. It was found that increasing Wikipedia views (used as a proxy for monitoring those discovering Bitcoin) and increasing trading volume on private off-blockchain exchanges are not mirrored by equivalent increases in transaction volume on the Bitcoin network. This suggests that users tend to leave their Bitcoins with the trading exchange and do not use the Bitcoin network (to either store the Bitcoins in their own wallet or transfer ownership to someone else in exchange for goods or services). Given that the Bitcoins are left on the trading exchange, and the only action a user can take is to sell them at a later time, this is taken as a strong indication that many new users buying Bitcoins are doing so as an investment.

Using an early dataset, up to the end of 2013, snapshots of Bitcoin's public ledger were taken [49]. These snapshots included information on the number of Bitcoins associated with a particular wallet address and the number of historical transactions in and out of the address up to the snapshot point. Addresses (and therefore users of the Bitcoin network) were categorised into the following types: *active investor*, *passive investor*, *currency user*, *tester*, *miner* and *hybrid user*, based on a rule-based approach. It was observed that the number of *investors* (both *active* and *passive*) and *hybrid users* had grown over time while the number of *currency users*, *miners* and *testers* had decreased. It was also seen that *passive investors* and *miners*, while small in numbers, hold large amounts of Bitcoin compared to other groups, suggesting that a few accounts hold large amounts of Bitcoins. Again this analysis points towards users' involvement with Bitcoin being of a speculative nature. In a separate study, regression analysis found that Bitcoin market returns were driven by internal buyers and sellers, and had little relationship with other factors considered, again demonstrating Bitcoin's speculative nature [5].

2.2 Cryptocurrency trading markets

It is unlikely that a user's wallet will be sent cryptocurrency with nothing in return and so cryptocurrency exchanges exist. Exchanges provide an entry point for those who wish to purchase cryptocurrency simply: the user can send money to the exchange (most of the time fiat deposits and withdrawals are done via bank transfer [40]) and then purchase the cryptocurrencies offered by the exchange. Users can also trade one cryptocurrency for another. Prices are set by other users participating on the exchange, usually using a limit order book based system (for more details on the function of limit order book based systems, see [50]).

A user having cryptocurrency in their exchange account (for example, having just bought some Bitcoin) usually means that the exchange has taken custody of the cryptocurrency for the user. If the user wishes to take ownership themselves, this is done by the user initiating a withdrawal from the exchange. Due to the large amount of cryptocurrencies held by individual exchanges (the cryptocurrencies being placed into the exchange's custody by the exchange's millions of users), exchanges are a target for hackers looking to steal those cryptocurrencies. Over the previous ten years, the number of exchange failures, mostly caused by security breaches [51], has been high. It is claimed that 45% of exchanges that had previously existed had failed by 2013 [52], and on many occasions such failure resulted in users losing funds. The exchange ecosystem (i.e. which exchanges exist and are traded on) changes frequently due to sudden exits by exchanges due to security breaches and also changing regulation. For example, Mt. Gox handled the most Bitcoin trading volume pre-2014 but then collapsed due to security breaches. When exchanges exit this opens up opportunities for new exchanges to take their place. For example, Binance, a relatively new trading exchange having launched in June 2017, had one of the top ten trading volumes within 5 months of launching [53], and by October 2018 reportedly had the largest USD trading volume of all exchanges [54].

Due to the security concerns raised by centralised exchanges, decentralised exchanges have been proposed (for example, in [55]), implemented, and launched. These exchanges are usually based on smart contracts (they are an example of a dApp) and allow direct trading between participants who keep custody of their cryptocurrencies until atomically traded for other cryptocurrencies. However in October 2018 a report [54] documented that the top five decentralised exchanges only contributed, on average, 2.4 million USD to the daily cryptocurrency market trading volume (just 0.4% of all trading volume).

All cryptocurrency exchanges are open continuously, with no evening or weekend closures. Each exchange contributes separately to the overall price discovery of cryptocurrency prices [56], with each exchange having its own market price. The presence of traders looking to exploit exchange arbitrage strategies means more substantial price differentials between exchanges are likely to be reduced, resulting in relatively similar market prices on all exchanges, as seen in Figure 1.

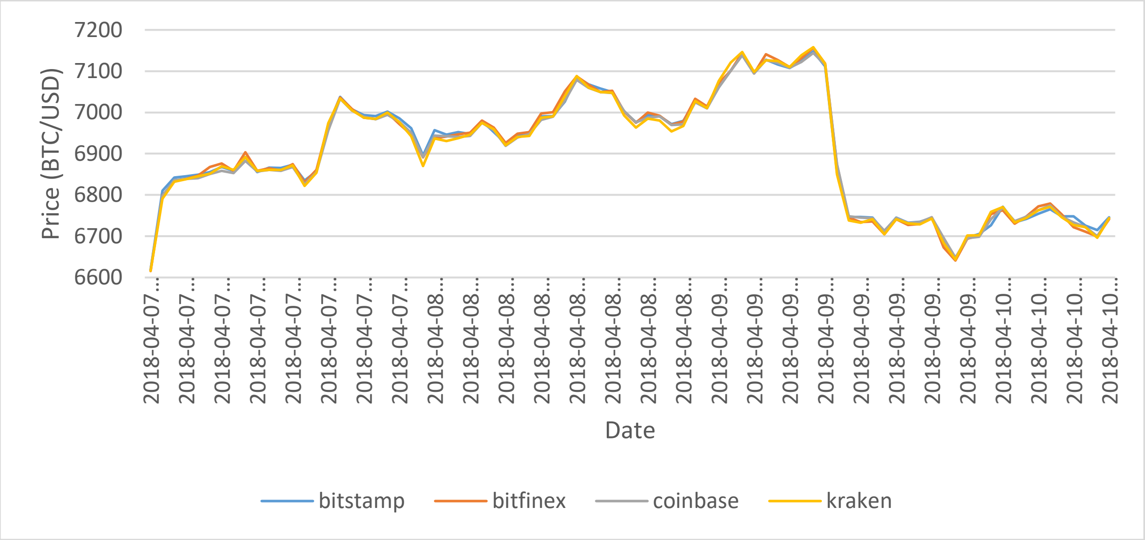

Figure 1. Similar hourly exchange prices over several days

When the Bitcoin price is analysed, it is common to either examine the price on one exchange or take an aggregate price from a number of exchanges. The dynamics of the price have been explored, finding similarities with more traditional markets, including that the Bitcoin price follows the unit-root law, fat-tail phenomenon, and exhibits volatility clustering [57]. Patterns in trading volumes have also been observed. Many exchanges (although not all) have been seen to show intraday seasonality in their trading volumes (similar, but to a lesser extent, to what is seen in the forex markets), whereby trading volume for particular markets (BTC/USD, BTC/EURO etc.) is higher when their respective region of the world is awake [58] [59]. Price clustering has also been documented, whereby the price of Bitcoin against USD appears to cluster around round numbers (and to a lesser extent, decimals of .50 and .99) [60]. As well as looking at USD pricing, as in [60], another study [61] extended the analysis by looking for clustering of prices when cryptocurrencies are priced in other cryptocurrencies. It was found that the same round number clustering occurs. For example, 35% of Litecoin prices were priced in 100 Satoshi increments (a Satoshi is a one hundred millionth of a single Bitcoin). It was also found there was clustering around other round numbers and just above or below round numbers.

Price and trading-related data can provide enough information for those looking to speculate on future prices, and a common way of doing so is by providing this data as input to particular machine learning models. For example one study evaluated the relative success of two machine learning techniques, an RNN and an LSTM, in their attempts to predict the Bitcoin price [62]. The LSTM, while outperforming the RNN, was reported to only have 52% accuracy**.** The authors however noted that the use of accuracy as a performance measure may be misleading especially when Bitcoin was generally increasing in value (e.g. a model that simply predicts up could achieve a very high percentage accuracy if applied to a dataset where the Bitcoin price trends upwards). One technique to combat this would be to express prices in Bitcoin (if considering altcoins) or to evaluate the predictions as part of a trading strategy against a benchmark strategy. Both of these techniques were used in another study while evaluating separately an LSTM and two models based on gradient boosting decision trees, supplied with price related inputs, in their ability to predict future prices [63]. If proven profitable on the necessary historical data, machine learning models similar to these are likely to be deployed by speculators looking to profit from price movements. Indeed, artefacts extracted from the market-related data and other findings indicate the prominence of such speculators, and the following paragraphs cover these findings.

It was noted earlier, in Section 2.1.3, that the characteristics of use and ownership of cryptocurrencies mirror the characteristics of ownership of speculative investments. This finding was achieved by analysing blockchain (‘on-chain’) activity. Having since that earlier section considered the markets in which cryptocurrencies are traded, it is discussed below whether trading market dynamics (occurring on exchanges, and hence, mostly ‘off-chain’) indicate the same speculative behaviour.

Short term price movement patterns have been observed to reoccur over time [64]**,** indicating the existence of trading strategies following short-term indicators (for example, technical analysis) or the presence of trading bots following pre-defined strategies. Another study found short term changes in trading activity around significant cryptocurrency events [65]. These changes in trading activity suggest the presence of traders entering and exiting positions based on speculation rather than on fundamental value. In addition artificial models including speculative actors have been able to reproduce price movements (and price characteristics) of Bitcoin reasonably well [57], suggesting the existence of similar speculators in the real market. This work used an agent-based artificial simulation where heterogeneous agents bought and sold Bitcoins using strategies dependent on the type of agent. The agents were classified as *random traders* and *chartists*. The random traders would submit buy and sell orders for random (presumed exogenous) reasons. The chartists (the agents of more interest here) represented speculators who were programmed to follow trends—when the price is rising, the chartists issue buy orders as they anticipate a continued increase in price (trend following). Similar results were found in another study [66] that tested a variety of trading strategies on real market data. This study finds strong evidence of price momentum, resulting in the authors' price momentum trading strategies outperforming the more conservative risk-based strategies. The authors suggest that the observed price momentum is unlikely to be explained by changing fundamentals (traded on by fundamental traders) and is more likely to be the result of speculators following trends.

As well as the existence of trend-following speculators, one study found significant peer influence between traders within cryptocurrency markets, suggesting the occurrence of a substantial amount of speculation [67]. The authors of this study created automated trading bots to submit thousands of scheduled orders over six months. The authors monitored trading activity by other market participants after such activity and compared it to periods without such submitted orders. The authors found buy orders led to further buying activity, which suggested traders are susceptible to being influenced by their peers. The authors argued that this peer influence is a result of a large number of speculators in the market, who may judge the value of a cryptocurrency by the demand shown by other participants rather than calculating what they believe is the cryptocurrency's intrinsic value. Another study [68] found a similar peer influence between cryptocurrency investors. This time the peer influence was over longer time horizons and varied over time, resulting in investors herding to the same cryptocurrency projects; it was hypothesised that investors were mimicking the investment decisions of other investors. The authors suggested the observed herding behaviour could be one of the causes of the speculative cryptocurrency bubbles commonly reported on in mainstream news and also detected in academic studies. For example, one study [12] applied a bubble-detecting statistical test and from this argued that fluctuations in the Bitcoin price are not suggestive of constant fundamental value and that the price encompasses a significant speculative component. More literature relating to price bubbles, and their existence in cryptocurrency markets, is reviewed in the appropriate chapters later in this thesis. (Both Chapter 4 and Chapter 5 review appropriate literature relevant only to those chapters, respectively.)

Speculation by investors has also been confirmed as influencing the Bitcoin price through the application of wavelet coherence [69] [9]. Relationships between the Bitcoin price and a number of different factors over the short, medium and long term were considered. It was found that the Bitcoin price is strongly influenced by the interest of investors (a good indicator of demand). The interest of investors, which may be hard to monitor (due to it being private to an individual investor), was in this case estimated from online indicators (relevant Google search volumes and Wikipedia views). In addition to [69] a number of other studies have investigated the use of online indicators—including social media indicators—for their predictive power within cryptocurrency markets. These studies are summarised in the section below, with a focus on the specific indicators used.

2.3 Use of online indicators in cryptocurrency price prediction

2.3.1 An introduction to using online indicators in financial price prediction

Word of mouth has been shown to be an important factor in investment decisions; an investor's peers are more likely to invest in the same financial instrument [70], and share the same investment characteristics [71]. It has also been seen that neighbours' investment decisions are

[9] This work is considered comprehensively later in section 4.1.2 due to its relevancy to that particular chapter.

likely to be linked [72]. In an increasingly digital age, peers may now be online and 'neighbours' geographically located around the world.

Following seminal and heavily publicised work detailing how Twitter sentiment can be predictive of stock market movements (the Dow Jones Industrial Average) [73], multiple avenues of work have explored the ability of social media and other online indicators to predict stock market movements. There exists a broad range of methods—spanning economics, data mining, natural language processing, and machine learning—to predict a range of financial assets, as summarised by [6]. As seen in [6] it is common to use trading strategies to validate indicator driven models. However an often overlooked precursor to the generation of such trading strategies is to first determine whether actual relationships between indicator and price are present, and to discover if the indicators are leading the price, as noted by [74].

Inspired by such work, prediction of cryptocurrency markets via online indicators has now become popular. The following subsections summarise where existing work has utilised online indicators, including social media indicators, to predict cryptocurrency markets.

2.3.2 Use of Wikipedia views as a factor

One such indicator, the number of Wikipedia views for a particular cryptocurrency, has been found to exhibit a bidirectional relationship with price [75]: i.e. not only do price changes lead changes in the number of Wikipedia views but, more importantly for prediction, changes in the number of Wikipedia views also appear to lead price changes. It has been suggested that monitoring Wikipedia views may provide different insights to other online factors (social media usage/sentiment) as Wikipedia views may provide a footprint of new users learning about a cryptocurrency [4]. However another study [76] found only a weak relationship between Wikipedia views and the Bitcoin price, with many of the 20 other factors in this study exhibiting stronger relationships. In addition the strength (and, interestingly, the direction) of the relationship between Wikipedia views and Bitcoin price was inconsistent depending on which of the (multi-year) time periods was being considered.

2.3.3 Derivation of factors from discussion forums

Discussion forums have been investigated to determine whether valuable price-related information can be retrieved. In the case of [77], the application of a topic modelling technique allowed the evolution of topics on the *bitcointalk.org* forum to be monitored. A number of topics were identified, including those labelled by the authors as relating to *blockchain*, *illegal* and *investment*. The topic categorised as *China* was found to have significant Granger causality [78] with the Bitcoin price, suggesting the Bitcoin price was responsive to China-related events (the data used included a period where Chinese cryptocurrency trading exchanges were reported to facilitate over 95% of global Bitcoin volume). While in [77] Granger causality was checked only in one direction (topic occurrence to price) and not in the other (price to topic occurrence), so it could not be determined whether the relationship discovered is unidirectional or bidirectional, topics such as *blockchain, altcoin* and *transaction* were found to have a high Granger causality with Bitcoin transaction count. Another study [79] looked at the content of forum posts on a range of cryptocurrency forums but instead of extracting topics from the content extracted sentiment. The sentiments were categorised from very negative to very positive, using the VADER Python library. A Granger causality test showed significant causality between sentiment and lagged prices (6-7 days later), and a profitable trading strategy was generated from the relationship.

As well as retrieving information using topic modelling and sentiment of forum content, previous work has considered the connectedness of *bitcointalk.org* forum users [80]. Those users that stimulate discussion from their posts (measured by the number of times their posts were directly replied to) were more likely to share accurate information relating to future price changes. The "connectedness of authors" metric used in this work is similar to indicators used in another study [81], where the authors aimed to monitor “collective sensemaking” of discussion associated with particular cryptocurrencies on *bitcointalk.org*. Examples of metrics included the amount of conversational turn-taking and exposure to various topics (as a proxy for the user having diverse interests). It was found that discussions showing high levels of collective sensemaking were associated with technical and well-documented coins, and discussions showing hype and naivety were associated with less technical but more volatile cryptocurrencies. Although not predictive of future prices, these findings, relating to characteristics of community discussion around particular cryptocurrencies, could be used as inputs to a trading strategy.

Another study, comparing two data sources, noted that internet forums seek discussion of ideas whereas Twitter is more about the propagation of single messages [82]. Discussion of ideas, as on a forum, appears inherently more connected ('sensemaking') than propagation of single 140 character messages, and this may have been one of the reasons for the paper's finding that sentiment derived from *bitcointalk.org* had a stronger relationship with future Bitcoin prices than the sentiment derived from Twitter.

2.3.4 Use of factors from additional specialised sources

News sites dedicated to Bitcoin have also been explored [83]. Although it was noted in this paper that such sites might not be read in their entirety by all those trading Bitcoin, news sites provide a good summary of the events occurring within the ecosystem, and their content can be analysed for sentiment. A bespoke sentiment classification system was designed, in conjunction with contextual word lists. A number of different rules were used including the use of negators (words like 'but' and 'not') to flip the polarity of a sentiment and intensifiers (words like 'very') to add or subtract further weight to sentiment. In addition the sentiment associated with a headline was weighted higher than that provided by the body of an article. Validation of the model was provided via a trading strategy. On a day with positive news a long position was entered, and then a short position was entered for the next two days (based on the assumption the market had overreacted). Opposite actions were taken for each step on days with negative news. The trading strategy proposed did not achieve returns that could be considered a successful validation of the relationship between sentiment and price. However this could have been due to the design of the trading strategy; the introduction of a trading strategy without preceding confirmation of a relationship obscures what part of the work is unsuccessful.

Other novel data sources have also been explored, including monitoring levels of technical progress and 'innovation' of a cryptocurrency project. One way of doing so, due to the open-source nature of many cryptocurrency projects, is to track the software development progress on the project's GitHub code repository [84].

2.3.5 Use of Google search volumes as a factor

The relationship between Google search volumes (often termed "*Google Trends*" due to the name of the service that provides this data) and cryptocurrencies has been explored in the literature. It was found that when prices are high (above trend), increasing search volumes push prices higher [75]. In contrast, when prices are low (below trend), increasing search volumes push prices even lower. The authors comment that the observed behaviour forms an environment suitable for bubble-like behaviour.

One study reported that Google search volumes led Bitcoin trading volume [85]. Using Granger causality, the authors found a unidirectional relationship, where search volumes were found to lead trading volume, but not the other way round. Using cross-correlation with lags -5 to 5 days, the strongest correlations were between Google search volumes on day T and trading volumes on day T+3. One possible reason for this is that Google search volumes may capture new interest from people looking to buy Bitcoin. Once they have searched for Bitcoin and found a trading exchange, they need to fund their account with capital to be able to make a purchase. Bank transfers to exchanges are not instantaneous and take a few days. However, these results differ from a more recent study [86]**,** which uses a larger duration of data, where it is found that trading volume Granger-causes search volumes but not vice-versa (i.e. both studies are reporting Granger causality but in different directions). The possible relationship between Google search volumes and market activity is thus, currently, unclear.

2.3.6 Use of Twitter activity and/or content as factors

In a comparison between the indicative power of Google search volumes and Twitter submission volumes and sentiment, it was found that Google search volumes exhibited higher cross-correlation with the associated cryptocurrency price than the Twitter-based metrics [87]. Twitter is a common source of social media data in the pursuit of financial asset prediction, and the prediction of cryptocurrency markets is no exception as numerous studies have used Twitter as a data source, with varying success. One example study considered the strength and polarisation of opinions rather than looking at absolute volumes of Twitter posts ("tweets") containing different sentiments [88]. It was found an increase in the polarisation of sentiment (disagreement of sentiment) preceded a rise in the price of Bitcoin. Their results were validated through the generation of a profitable trading strategy back-tested over an unseen test set.

In other work, tweets have been categorised into “positive”, “negative” and “uncertain” based on the matching of occurrences of words in the tweets with occurrences of words in pre-defined wordlists [89]. The Pearson correlation highlighted a significant positive correlation between the sentiment of tweets and daily close price and trading volume. However Granger causality did not confirm a leading relationship between tweets and price changes. The work demonstrated that, with the configuration outlined, Twitter sentiment might not be able to *predict* cryptocurrency movements, but that sentiment within the platform mirrors (is responsive to) market movements. The results could be improved by using a wordlist customised for the specific context, for two reasons. First, it has been shown that word lists generated for multi-disciplinary use may misclassify words used in a financial context—in particular, many words classified as negative may not be negative in a financial context [90]. Second, those interested in cryptocurrency have developed their own terminology/jargon (e.g. “moon” is used to refer to a meteoric price rise), much of which may prove to be more informative than more usually expected words, due to the frequency and consistency of its use.

It should be noted that many studies focus on the production of trading strategies and often overlook the explicit details of any raw relationships present between online activity indicators and associated cryptocurrency prices. As well as investigating raw relationships between unexplored indicators and associated cryptocurrency prices, the work in Chapter 4 will also aim to clarify and contextualise certain relationships which have already been reported. Chapter 5 will then introduce a methodology for predicting bubbles, using the first application of an epidemic-based bubble model within a cryptocurrency context. As can be seen from the above examples there are a plethora of online data sources from which to retrieve information that might pertain to future cryptocurrency market movements, however one platform, Reddit, has yet to be analysed. The work contained in this thesis also introduces Reddit and validates it as a viable and promising data source; the next chapter will analyse a number of possible social media data sources available, including Reddit, and provide justification for the final data sources chosen.

Chapter 3

Data Sources and Acquisition

3.1 Choice of social media data sources

The popularity of social media services has exploded over the last decade; such services, although varying in exact functionality, generally allow users to share content (including text, video and pictures) with one another. There are now hundreds of services—accessed either via a web browser or mobile application—that fall under the broad umbrella of social media. In this chapter, three possible social media platforms are evaluated for use in later work. Twitter is naturally considered due to it being the most commonly chosen social media platform for similar analysis (although as seen in Section 2.3.6, previous cryptocurrency-specific use has had varying success). Alongside Twitter, two other social media platforms are considered: Reddit and Facebook. Both Reddit and Facebook are novel and unexplored in cryptocurrency price prediction related literature. The characteristics of each platform are examined, then accessibility of data is considered before a final decision is made as to which data sources should be used.

3.1.1 A comparison of the design implications of the considered platforms

Firstly, message-based restrictions of all three platforms are considered. Message length constraints could inflate the number of messages on each platform. Twitter submissions are limited to 140 characters[10]. Such restriction results in two issues when a user wishes to convey a message that requires a longer body of text: 1) the limit prompts users to split their message into a number of successive submissions, inflating the number of submissions on a particular topic; 2) it causes users to spend time adapting their content to adhere to the limit[11]. This potentially prompts users to use other social media platforms for longer messages. Neither Reddit nor Facebook has such restrictive constraints. Another mechanism likely to inflate messages on Twitter is the retweet mechanism. Retweeting a submission reposts it again from a second user's account rather than the original author's account (the retweeting user has the ability to add an extra message, but commonly this is not done). Facebook provides the ability to share a post from another author onto a personal timeline, again with the option to add content. Reddit doesn't have an equivalent feature, all content on an individual subreddit is a unique contribution.

Secondly, the amount of spam on each platform is considered; the existence of spam messages will increase the number of submissions while reducing the quality of the content. Twitter has a large number of spam messages submitted by automated accounts ("bots") [91]. A wide spectrum of niche communities are targeted by the bots including cryptocurrency communities. A similar volume of spam messages is not seen within cryptocurrency communities on Reddit. As Reddit is meant for discussion threads, any content not tailor-made to the thread (or wider subject area) is identifiable to most users as spam; users are likely then to *downvote* the content until it is not visible to future users (which happens once a certain threshold is met; approximately an overall *upvote*/*downvote* score below 0). The lack of visibility achieved by spam messages on Reddit reduces the likelihood that further spam messages will target the platform. The volume of spam on Facebook groups depends on the level of activity of the moderators of that group. The large number of spam messages seen on Twitter hinders data mining attempts; the spam messages create background noise and desensitise overall sentiment and volume counts on Twitter in relation to cryptocurrency market events (sentiment/volumes are less connected to market events). This has been suggested as the reason for worse results (in a cryptocurrency specific context) achieved using Twitter data when compared with other data sources [92].

[10] In September 2017, Twitter trialled allowing a limited number of users to use up to 240 characters. Even though an increase in allowed characters, the limit is still restrictive.
[11] https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html

Thirdly, the volume of discussion relating to cryptocurrencies is considered. The volume on Twitter dwarfs the volumes of submissions on Reddit and Facebook. However, due to the character limit (causing users to split messages between submissions), the retweet mechanism (inflating the number of submissions even though no additional content is being added), and the existence of the huge amount of spam messages on Twitter, the value of information in these submissions is expected to be low. In the pursuit of valuable information, quality of submissions may be more important than quantity, implying Reddit or Facebook could be more informative platforms.

Fourthly, the characteristics of the cryptocurrency community on each platform are considered. Interestingly, the characteristics of real (non-bot) Twitter users communicating about Bitcoin suggests they would be suited to social media platform with a more topic-based structure. In a study attempting to investigate Bitcoin user characteristics [93], it was found that those users communicating about Bitcoin behaved differently to the majority of Twitter users, in that they were not engaging in general social interaction but were focusing (almost entirely) on their specific area of interest, i.e. cryptocurrencies. More specifically they were less likely to mention family, friends, religion and sex, and had less observable interaction with other users. The findings of the study would suggest that other more subject-oriented platforms may be better suited to those interested in discussing cryptocurrencies. Both Reddit and, to an extremely limited extent, Facebook offer subject-oriented areas through *subreddits* and *groups*, respectively.

Finally, the type of content disseminated on each platform is considered. Twitter and Facebook allow images and videos within submissions without text—in fact statistics show that submissions to Twitter with images get 150% more retweets, highlighting the popularity of this type of content[12]. Reddit submissions always have some textual element, but can also include images and videos in certain situations. Although unintended, this aids data mining within Reddit, as the text-based content is more easily analysed than image- and video-only content (which are possible on Facebook and Twitter). It is likely that images and videos would be ignored in most unspecialised data mining approaches.

[12] http://www.adweek.com/digital/twitter-images-study/

3.1.2 A comparison of the accessibility of data on the considered platforms

The accessibility of data has considerable practical implications. Accessibility is important because: 1) it aides the current work (unhindered access being the ideal, unrestricted by budget and data availability constraints); 2) it enables other researchers to reproduce and extend the work presented here, with greater ease and likelihood than if paid data was required. Table 1 summarises the accessibility characteristics discussed further below, with a focus on historical data (which is more useful than real-time data as extended data sets can be used).

Table 1. Data accessibility comparison between three social media platforms

Platform	Twitter	Reddit	Facebook
Proportion of live data available	1% (public use) 10% (academic use)	100%	100%
Historical data availability	0% (free); Previously, $600 - $1000 per month of term-based data In March 2018, a new access mechanism was introduced. Approximately $1 per 30 days of count-based daily data points (term-based data)	100% (free)	5-10% (free); More available if paid for.

Formerly, historical Twitter data had been accessible via a number of third-party data providers. These companies purchased data in bulk from Twitter and then resold filtered variants of the data to end-customers (filtering being usually term-dependant, e.g. all submissions containing “bitcoin”). Two companies were approached during an earlier phase of the current work and quoted $600 and $1000 per month of term-based data. Academic discounts were insignificant or non-existent. The expense of historical data from Twitter prompted researchers requiring such data to build systems which monitor the real-time API and store the retrieved data (hence, building up a store of historical data). Naturally, data could only be recorded from the point in time that the initial idea to use Twitter had been devised. This resulted in short data periods being used; for example, some cryptocurrency prediction studies using Twitter data used 104 days [89], 71 days [89]**,** 60 days [94] and 91 days [82], respectively. However, recently (March 2018), a new access mechanism has been introduced, discussed in Section 3.3.5. The cost of the new mechanism is approximately $1 per 30 count-based historical data points (e.g. if minute data is chosen, 30 minutes of count-based data is $1 whereas if daily data is chosen, 30 days of count-based data is $1). In the situation where daily count data is required, this is much more affordable than previous options (retrieving raw tweet data for any reasonable duration being still out of budget for academic research).

Historical Reddit submissions can be retrieved from the public API. A time-based window of submissions can be specified using *start* and *end* parameters. Constraints are placed on the number of results returned. The whole history of submissions cannot be retrieved in one go (this is expected as the space requirements of such a response would be unworkable). The whole history of submissions can instead be retrieved by adjusting the *start* and *end* parameters and working systematically through the date period of interest (more details in Section 3.3.1).

Historical Facebook group submissions had previously been available in their entirety using the *posts* endpoint of Facebook's developer API. As of September 2017, however, users of this endpoint started reporting an issue[13] where although the most recent posts were retrieved in their entirety (hence 100% accessibility of *near* real-time data in Table 1), after a certain number of results were reached only 5 to 10% of further posts were being returned. Facebook engineers reported that this issue would not be resolved soon, and advised using a different endpoint. However the suggested endpoint requires the requester to be the owner (administrator) of wherever data is being requested from (for example, administrator of a particular group). As a result, in practice, only 5-10% of historical data can be retrieved for free from Facebook. Facebook recently acquired a social media monitoring company called CrowdTangle; historical Facebook data can be purchased from the company, but prices are not publically available, and bespoke

[13] https://developers.facebook.com/bugs/1838195226492053/

customer accounts are only set up after consultation with a member of the company. It is expected that the pricing of data from CrowdTangle will be significant, and also that only summaries of data would be retrievable (data would not be available in its rawest form).

3.1.3 Final decision on social media data platforms

After comparison of the three data sources (as detailed in sections 3.1.1 and 3.1.2), Reddit appears the most promising platform for the analysis in this work. The reasons for this are, in summary: 1) there is less cryptocurrency-related spam on Reddit (and thus proportionally more quality content); 2) the topic-based structure is beneficial to data mining pursuits; 3) Reddit data is relatively accessible. However alongside Reddit, data from Twitter will also be analysed. Although some qualitative/design-based concerns have been raised in this section, it is still of interest to include Twitter in the initial quantitative analysis due to its prominence in previous similar work. As well as these two social media sources (Reddit and Twitter), factors will also be derived from other online sources (Wikipedia and Google). Given the novelty of Reddit as a data source, the following section provides a more comprehensive introduction to Reddit.

3.2 An introduction to Reddit

Reddit, an online social media platform, is a collection of communities dedicated to the discussion of different subjects. It was launched in June 2005, amassing over 250 million users and achieving 8 billion page views per month. Unlike other social networks where the focus is mostly on social interaction with individuals with which one already has a shared connection, people on Reddit congregate together based on their shared interest in a particular topic. Different topics have their own *subreddit* (/r/politics and /r/london are two examples of the 850,000 subreddits in existence). Figure 2 shows an example subreddit page layout.

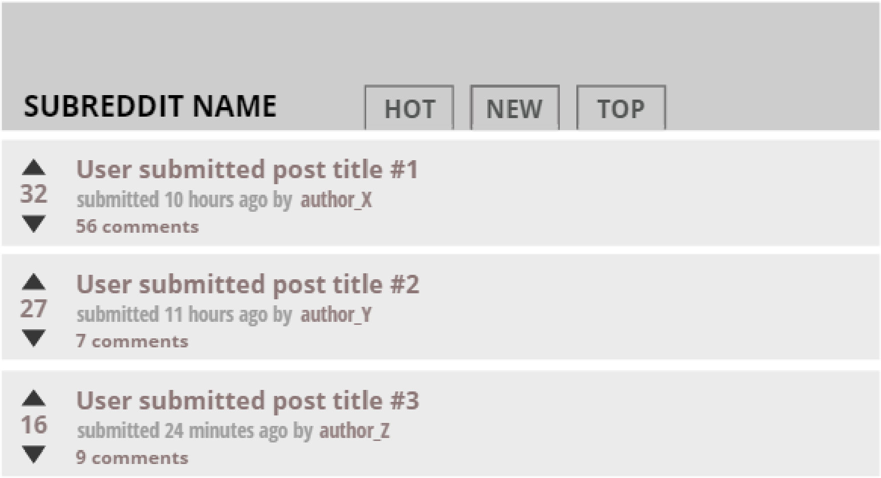
SUBREDDIT NAME
HOT
NEW
TOP
32
User submitted post title #1
submitted 10 hours ago by author_X
56 comments
27
User submitted post title #2
submitted 11 hours ago by author_Y
7 comments
16
User submitted post title #3
submitted 24 minutes ago by author_Z
9 comments

Figure 2. Example subreddit layout

On accessing a subreddit, a list of *posts* is visible. Posts can be links to content elsewhere on the internet or original content produced by the *author* of the post. Users can *subscribe* to a particular subreddit. Posts made in that subreddit and other subreddits the user has subscribed to will appear on the user's personalised Reddit homepage. Clicking on the comments button under a particular post will navigate the user to a page showing an individual post and any comments submitted in response to it, as seen in Figure 3.

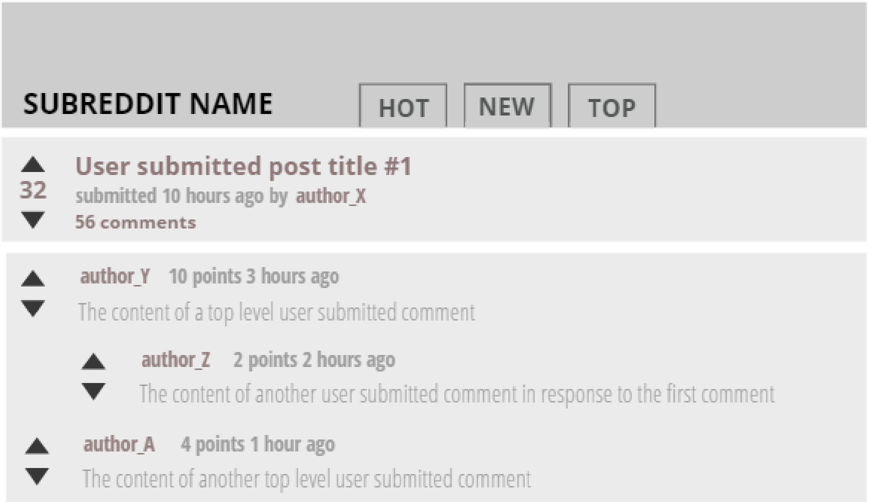
SUBREDDIT NAME
HOT
NEW
TOP
32
User submitted post title #1
submitted 10 hours ago by author_X
56 comments
author_Y 10 points 3 hours ago
The content of a top level user submitted comment
author_Z 2 points 2 hours ago
The content of another user submitted comment in response to the first comment
author_A 4 points 1 hour ago
The content of another top level user submitted comment

Figure 3. Example post layout

Users can submit a *comment* to respond to a post or another, previous, comment. Long *discussion threads* can thus result from individual posts or comments. Posts and comments can be *upvoted* and *downvoted* by other users; this mechanism affects the visibility of the content.

Most cryptocurrencies have their own subreddit. Subreddits are commonly used by the development teams of a particular cryptocurrency to communicate with the community, to engage in debate over technical issues [95], and to distribute news. There have also been cases where time-sensitive news (hacks/code bugs) have first appeared as a Reddit post by a community member before being discussed publically or announced by the development team[14]. Reddit themselves have been seen to embrace the growth of cryptocurrency communities on their platform; it has been observed that intermittently during 2018 the Reddit Android app mentioned cryptocurrencies in its title.[15]

3.3 Data acquisition

3.3.1 Reddit message data

As mentioned above, in Section 3.1.2, the history of content on Reddit can be retrieved from publicly accessible APIs. Reddit offers their own API; however, the endpoints they offer are targeted towards programmatic interaction (for example, for third-party applications to allow users to interact with their Reddit accounts) rather than data retrieval. A common choice and more suited to extensive data retrieval is to use an external service called pushshift.io, built by Reddit moderators, which contains an API allowing access to the full history of content on Reddit. Its dataset is updated in real time, hence always expanding, as more content is submitted to Reddit.

The pushshift.io API has a number of different endpoints allowing for a range of queries to request different types of data. The API always returns results in JSON format with the most recent results matching the input query being returned first, unless another custom ordering is specified in the query. There is a limit of 500 results per query (the maximum number of results returned to one query). As an example, it is possible to query for the most recent 500 submissions (occurring anywhere on Reddit) with a specific word by using the *q* parameter (e.g. https://api.pushshift.io/reddit/search/submission/?q=bitcoin&size=500). Content occurring within particular subreddits is the main interest of this work. Therefore the *subreddit* parameter is

[14] https://www.reddit.com/r/ethereum/comments/4oi2ta/i_think_thedao_is_getting_drained_right_now/
[15] https://www.reddit.com/r/CryptoCurrency/comments/84pqjr/the_reddit_android_app_changed_its_title_to/

used which returns the most recent posts from a particular subreddit. (e.g. https://api.pushshift.io/reddit/search/submission/?subreddit=bitcoin). Additional parameters can also be added to queries including *before* (filtering results to be before a certain Unix timestamp) and *after*. For example, the following example query: http://api.pushshift.io/reddit/search/submission/?subreddit=bitcoin&after=1546300800&before=1546847893&size=500 filters for submissions to the bitcoin subreddit after 01-01-2019 and before 07-01-2019, with the most recent 500 being returned.

As mentioned in the previous paragraph, there is a limit to how many results can be returned for one query, meaning that it is impossible to retrieve the whole history of a subreddit with one query. The before and after parameters can however be used to iteratively retrieve all the submissions over the entire history of a subreddit. An initial method that might be considered could combine them into a loop which is designed to query one day at a time programmatically. This, however, would not capture everything if there are more submissions in an individual day than the number allowed to be returned; in other words, this approach would return only the latest (most recent) 500 submissions per day to a particular subreddit.

To overcome this, rather than querying day by day, custom date ranges can be designed based on the data retrieved, with the latest point of interest in the second query being programmatically set to be the timestamp of the earliest submission returned in the first query made, and so on. The queries would repeat, retrieving the most recent 500 submissions before each timestamp provided (i.e. the *before* parameter moves iteratively backwards through the data, collecting all data points).

To undertake this process, a Python script is created to iteratively retrieve all submissions to a particular subreddit of interest. As with all APIs it is good practice to not unintentionally overload the API by querying it too frequently in a short time span; therefore a short delay occurs between each query. Upon receiving the results for each query (a list of submissions), each submission is processed (filtering out unrequired metadata) and stored sequentially into a separate CSV file for each subreddit considered (for example, bitcoin_submissions.csv and bitcoinmarkets_submissions.csv). Alongside the body (text/content) of submissions, the following details are also recorded: submission timestamp, score (based on upvotes and downvotes), and author. From the stored details of each submission (and aggregation of multiple submissions over specified time periods) further metrics and derived information can be retrieved. The choice of which metrics and derived information to retrieve depends on what is needed for a particular purpose and will be covered separately in each experiment.

3.3.2 Reddit subscriber data

Subscriber growth is harder to track than the other metrics mentioned above. Only the current subscriber count is displayed for a particular subreddit, and historical data cannot be rebuilt retrospectively as 1) the act of subscribing does not have a visible historical imprint; and 2) historical subscriber counts are not available from the API. A third-party website, RedditMetrics (http://redditmetrics.com/), has been retrieving and storing real-time subscriber counts since 2012. The data provided by RedditMetrics is not available in downloadable form (e.g. CSV file or JSON file) and is only available in chart visualisations on their web pages. However, the raw data inside each chart is delivered to the web page via a JSON object which is visible on the client-side via the *view source* option. A Python script was created to retrieve and parse the data and then store the raw data values into a CSV file.

3.3.3 Wikipedia views

Most cryptocurrencies have their own Wikipedia page providing an introduction to the cryptocurrency. There is not one single location for Wikipedia views data covering the historical data interval required. Wikipedia views data from the start of 2015 onwards can be retrieved using the official *mwviews* Python library which connects to Wikipedia's pageview API. Previous historical daily data can be retrieved in one-month buckets from a separate website (http://stats.grok.de). Data was programmatically retrieved here from both sources and then merged to produce a single time series.

3.3.4 Google search volume

The volume of searches for particular terms is retrieved from the Google Trends service, a service provided by Google. Search volumes returned from Google Trends are scaled from 0 to 100, where 100 represents the highest search volume within the time frame queried. In this work, the search term used is the name of each cryptocurrency, for example the volume of searches for "Bitcoin".

Google Trends returns data with different granularity depending on the historical time interval queried: daily search volumes are returned for queries of duration under 90 days, and weekly search volumes for queries of duration over 90 days. However, it is possible to reconstruct daily data for long time intervals using a combination of daily and weekly data. The method is described and validated by [9]. In this method daily data is retrieved in buckets of under 90 days, and weekly data is also retrieved for the complete time interval of interest. Then, using the daily data, the percentage change of each day in a week from the first day of the week is calculated; these percentage changes are finally applied to the weekly data to build a daily time series over a more extended period.

3.3.5 Twitter data

Twitter has recently redesigned their mechanism of provision of historical data. Previously gaining access to historical Twitter data involved purchasing it from certified reseller companies (who commonly quoted amounts in the tens of thousands of dollars). Recently, however, Twitter has devised a new mechanism to retrieve the data directly from them via paid API access, which for certain needs (depending on the data required) is much more affordable.

Interested parties are able to access a number of APIs which provide access to different datasets coming from the Twitter platform. For the work presented within this thesis, the 'Search API' is the most relevant API as it allows historical tweets to be retrieved based on matching particular user-specified filters. Count-based daily data was queried from the paid 'Search API' endpoint. The endpoint only returns 30 results for each query (e.g. if daily counts are requested, approximately a month of counts are returned), meaning a number of queries (at a cost of $1 each) are required spanning the whole required data duration, for each cryptocurrency.

3.3.6 Cryptocurrency market data

Earlier academic work in the area (2010 – 2016) tends to focus on Bitcoin, due to Bitcoin being the first, best-known and largest valued cryptocurrency. However, over the years, other cryptocurrencies have risen to prominence, and academic work has started to reflect this with other major cryptocurrencies being considered alongside Bitcoin. Analysing more cryptocurrencies than Bitcoin alone has the advantage that findings for one cryptocurrency can be compared with other cryptocurrencies to see if the findings hold across a broader spectrum. In this work, Bitcoin, Ethereum, Litecoin and Monero are chosen for evaluation. This choice is made based on the public exposure of these cryptocurrencies, their high market capitalisations and high liquidity on exchanges and their relatively large communities/followings.

The experiments undertaken in this work could be extended, without modification, to a wider universe of cryptocurrencies, assuming those cryptocurrencies have an active subreddit from which information can be retrieved and that they are traded on exchanges (in contrast to ICOs which are not traded during the ICO period). Having chosen the cryptocurrencies to consider, the location from which to retrieve market data (price and volume) is discussed next.

Sourcing and processing daily cryptocurrency market data

When examining traditional financial markets (e.g. equities and commodities), work is often required to pre-process data to avoid spurious correlations caused by exchange holidays and other intervals when trading is not possible. Cryptocurrency markets are unusual in the sense that they operate 24 hours a day, 7 days a week, with no planned closures, and as such, this additional pre-processing should not be required. However, although cryptocurrency exchanges do not have planned closures, they are prone to unscheduled outages where trading is not possible on a particular exchange. In addition cryptocurrency trading exchanges have historically been notoriously bad at remaining operational [52]. For these reasons aggregated trading-related data from a number of exchanges is used, where possible.

Aggregated data is retrieved from two well-regarded cryptocurrency data providers: 1) BraveNewCoin, a provider of data and market research for cryptocurrency markets; 2) CryptoCompare, a data and statistics provider, as well as a community platform, for cryptocurrency market data. Aggregated indices are used from both sites; the historical End of Day (EOD) aggregated index is used from BraveNewCoin, and the CCCAGG index is used from CryptoCompare. The choice of source (either BraveNewCoin or CryptoCompare) to use for an individual experiment depends on the time duration of data required. A comprehensive outline of each index's constituents and calculation methodology can be found in the associated documentation provided online for both BraveNewCoin[16] and CryptoCompare[17].

Sourcing and processing intraday cryptocurrency market data

For situations where intraday data is required, the daily benchmarks detailed above are not appropriate. For intraday data, it was decided to retrieve tick data which can then be aggregated up to the required time interval (e.g. to standard open-high-low-close (OHLC) candlesticks of 1, 5, 30 or 60-minute intervals). Cross-exchange (source-aggregated) tick data is not possible, and if attempted to be implemented would produce nonsensical price data (the tick data would constantly vary due to price differentials on different exchanges). It therefore needs to be decided from which sole exchange to retrieve tick data.

Bitfinex is chosen as the source of tick data for a number of reasons. Firstly, due to its longevity of existence and its relatively large historical market share (at times, it has held the status of being the largest cryptocurrency exchange, based on trading volume). Secondly, it has a reasonably well documented (https://docs.bitfinex.com/docs/ws) and reliable API. Thirdly, it has high enough trading volume to enable the presence of arbitrage bots; the presence of exchange arbitrage bots will ensure the prices on Bitfinex reflect prices elsewhere (the bots will buy on the cheaper exchange and sell on the more expensive exchange until any profitable price difference is removed). This means Bitfinex can be considered as a sufficiently accurate proxy for the prices seen by the whole market.

[16] https://bravenewcoin.com/api/digital-currency-historical-data/
[17] https://www.cryptocompare.com/media/12318004/cccagg.pdf

Bitfinex provides access to multiple market data feeds including the *order book feed, ticker feed* and *trades feed*. To start receiving data from one of the feeds a *subscribe* message (specifying the particular feed of interest) needs to be sent to Bitfinex. In the current work, the *trades feed* is of interest; this feed sends a new message every time a trade occurs and includes details such as the *price*, *size* and *time*. All messages sent by the feeds are structured as JSON objects. Upon receiving each message, the data is extracted, and messages are sequentially appended to a CSV file. Over time, this builds up a history of all trades occurring on Bitfinex at the finest possible granularity. Once analysis of this data is required, the data can be aggregated into standard open-high-low-close (OHLC) candlesticks of the desired aggregation interval.

3.4 Preliminary factor derivation and data analysis

3.4.1 Choice and derivation of initial factors

Having outlined the online platforms to be used and how data will be retrieved from these platforms, the next step is to consider how raw data from these platforms will be converted into potentially meaningful metrics.

Reddit derived factors

Using raw submission data (and the attached metadata), a number of activity-based metrics can be derived. Table 2 outlines the factors derived from usage of a particular cryptocurrency subreddit (e.g. r/Ethereum). The *posts per day* factor is chosen as it is equivalent to a metric commonly used in the existing social media literature: the volume of submissions that relate to a topic (e.g. the volume of Twitter submissions per day relating to Bitcoin [9]). Whereas an increase in *posts per day* can be generated by existing community members becoming more active, *subscriber growth* records new users subscribing to updates from the subreddit. *New authors* captures new users joining the community, specifically ones who rather than simply subscribing to news from the subreddit have started contributing their own content.

Table 2. Social media (Reddit) factors considered

Factor	Description
Posts per day	The number of posts made on a particular subreddit, per day. This factor does not include comments made in response to particular posts.
Subscriber growth	The number of new subscribers that a subreddit receives, per day. Subscribing to a subreddit means that posts made to that subreddit show up in a user's personalized home page.
New authors	The number of new authors posting on a particular subreddit, per day. These authors may be new Reddit accounts or existing accounts who have previously posted only on other subreddits and not the particular subreddit being examined.

The *posts per day* indicator will be investigated in preference to measuring comments per day as examples exist where huge numbers of comments are generated that are unrelated to market activity. For example, people occasionally give away small amounts of cryptocurrency to everyone who comments with their public blockchain wallet address, which causes a huge spike in comments. (Preliminary work was in fact conducted with comments per day but, as was expected, showed less significant relationships than *posts per day* and price.) Both *posts per day* and *new authors* are retrieved from aggregating and processing timestamped message data (retrieval steps documented in Section 3.3.1). *Subscriber growth* is retrieved from a third-party data source, as outlined in Section 3.3.2. While initially the work of this thesis considers and evaluates the previously described count-based metrics, the content of individual submissions is also considered, in Chapter 6.

Twitter data

As discussed in Section 3.3.5, count-based data is retrieved from Twitter. This shows the volume of tweets containing a certain term aggregated over a certain time duration. In this case, volume of Twitter submissions per day relating to particular cryptocurrencies is considered. This is a commonly used metric, and its use can be seen in previous work (for example, [9]).

Google Trends and Wikipedia data

Both Google Trends and Wikipedia allow for the retrieval of count-based metrics relating to activity on their respective platforms. In the case of Google Trends, the data shows the quantity of searches for particular terms. In the case of Wikipedia, the data captures the number of views of a particular Wikipedia page. These are the common metrics used when these platforms are considered in the existing literature (e.g. [69] [75] [96]).

Now that several online indicator based metrics have been designed, it is beneficial to undertake statistical analysis on these metrics. This preliminary data analysis can provide further understanding of the data's properties, characteristics, and any relationships which may impact or influence the results of complex modelling.

3.4.2 Stationarity

Visual inspection of the time series under consideration indicates that they clearly exhibit trending patterns (general growth over the datasets, predominately caused by growth in interest in cryptocurrencies). Such trends suggest non-stationarity which can cause issues in the generation of descriptive statistics relating to a time series, and also issues when applying financial models to predict them (the assumptions made by a number of standard financial forecasting models are violated when applied to non-stationary data). An augmented Dickey-Fuller test (ADF) test can be used to quantify whether a time series can be considered as stationarity. It has a null hypothesis that a unit root is present in the time series under consideration and an alternative hypothesis of stationarity. The results of the test are displayed in Table 3.

Table 3 P-values for the augmented Dickey-Fuller test on the raw online indicator and cryptocurrency price series; grey shading signifies where the null hypothesis can be rejected (<0.01) in favour of stationarity.

	Bitcoin	Ethereum	Monero	Litecoin
Posts per day	0.0311	0.0087	0.1246	0.0001
New authors	0.0001	0.0000	0.0020	0.0002
Subscriber growth	0.0000	1.0000	0.0619	0.071
Wikipedia views	0.0000	0.9988	0.1820	0.1161
Google trends	0.0033	1.0000	0.0918	0.0002
Twitter volume	0.8632	0.9805	0.2421	0.0192
Close (price)	0.9984	1.0000	0.9988	0.6346

It can be seen from Table 3 that the majority of the considered time series are not stationary. When time series are non-stationary, it is usually possible to derive related time series which are stationary; one such way is to calculate returns between sequential data points. Returns are generally stationary because they tend to be mean-reverting (returns tend to oscillate above and below a constant mean) and the magnitude of returns, despite numerous spikes, is likely to be relatively constant over time. Here log returns are chosen (Returns(i, j) = $\log\left(p_i/p_j\right)$ where p_i is the price at i and j is the timestep before i). Log returns have a number of characteristics which make them a common choice in financial time series modelling, including that they are additive, making them easier to compute. Table 4 shows a unanimous rejection of the null (non-stationary) hypothesis illustrating the stationary nature of the returns for each time series.

Table 4 P-values for the augmented Dickey-Fuller test on log returns of the online indicator and cryptocurrency price series; grey shading signifies where the null hypothesis can be rejected (<0.01) in favour of stationarity.

	Bitcoin	Ethereum	Monero	Litecoin
Posts per day	<.001	<.001	<.001	<.001
New authors	<.001	<.001	<.001	<.001
Subscriber growth	<.001	<.001	<.001	<.001
Wikipedia views	<.001	<.001	<.001	<.001
Google trends	<.001	<.001	<.001	<.001
Twitter volume	<.001	<.001	<.001	<.001
Close (price)	<.001	<.001	<.001	<.001

3.4.3 Pearson correlation

Pearson correlation between two time series is represented by a number between 1 and -1. Numbers near 1 indicate a strong positive correlation, meaning upward movements in one time series tend to be relatively aligned with upward movements in the other time series, and likewise with downward movements. Numbers near -1 indicate strong negative correlation (as one time series experiences upward movements the other experiences downward movements).

Although similar work (for example, [87]) has performed correlation analysis on price data, it is generally recommended to use returns instead as returns can usually be assumed to be stationary, which is required to be able to achieve meaningful results [97]. As such, in this work, Pearson correlation is calculated on log returns (already confirmed as stationary in Section 3.4.2). Table 5 shows the Pearson correlation between a number of cryptocurrency prices and associated online factors.

Table 5 Pearson correlation between cryptocurrency price and associated online indicator.

	Bitcoin	Ethereum	Monero	Litecoin
Posts per day	0.0246	-0.0047	0.0973	0.0610
New authors	0.0092	-0.0275	0.1329	0.0534
Subscriber growth	-0.0198	0.1437	0.1129	0.0713
Wikipedia views	0.0118	0.1036	0.1875	0.1214
Google trends	0.0160	0.1468	0.0526	0.1457
Twitter volume	0.0230	0.0286	0.0408	0.1057

It can be seen in Table 5 that all correlations are weak, some considerably weaker than others (for example, the correlation between Bitcoin and its associated online indicators generally appears weaker than the correlation between other cryptocurrencies and their associated online indicators). There may be several reasons for the weak correlation seen.

Firstly, if there is no consistency in the direction of the correlation during the time period, this is likely to weaken the overall correlation over the entire duration. Two hypothetical situations can be used to illustrate why a lack of consistency might occur. For the first situation, imagine a circumstance where unexpected good news comes out about a cryptocurrency causing both price and *posts per day* increases. This would result in a positive correlation over this time period. For the second situation, imagine part of a cryptocurrency's ecosystem is hacked. Again it is likely *posts per day* will increase (as people disseminate and discuss the news, its implications and ways to alleviate it), but the price is likely to fall as news of the hack spreads. This would result in a negative correlation over this time period.

Secondly, other existing work (examining the relationship between Twitter and stock prices) found similarly weak correlations, and the authors hypothesised that correlation measured over an extensive data set is weakened by periods of low social media activity [98]. Such periods of low activity are especially present within the cryptocurrency related data sets used here as the datasets used begin early in each cryptocurrency's life when activity is low. As each cryptocurrency's community has grown on average over time, it is possible to (relatively crudely) test the hypothesis of [98] by splitting each cryptocurrency dataset into two and calculating the Pearson correlation on each half of the data. The results of this are displayed in Table 6.

Table 6 Pearson correlation in two different time periods (first half of the Ethereum dataset and second half of the Ethereum dataset) between online indicators and the Ethereum price

	Ethereum 08-08-2015 to 26-03-16	Ethereum 26-03-16 to 31-05-17
Posts per day	-0.0490	0.0664
New authors	-0.1134	0.1435
Subscriber growth	0.0530	0.3902
Wikipedia views	0.0765	0.4325
Google trends	0.1444	0.5515
Twitter volume	0.0136	0.0627

It can be seen that during the second half of the dataset each correlation is stronger than in the first half of the dataset, suggesting the correlation has strengthened as activity has increased. Separately, it is of interest that the correlation between price and both *posts per day* and *new authors* has changed from being negative to positive, further emphasising the earlier point that a lack of consistency in correlation is expected and that correlation depends on the particular events occurring within a certain time period.

This finding, that correlation is lower when there is low social media activity, would prompt caution when investigating cryptocurrencies with limited communities (hence being cryptocurrencies with low social media activity). There are also other reasons to be cautious when attempting to use social media to predict smaller cryptocurrency projects. For example, they are likely to exhibit lower liquidity meaning market orders of any size may—inadvertently or advertently—cause considerable and unpredictable changes in the price. Smaller social media communities can more easily be manipulated scale by sock puppet and shill accounts[18].

Although correlation can be used to discover relationships between time series, it doesn't show whether one time series is leading another; Granger causality can however be used to investigate this. Due to the time-evolving nature of relationships already seen, rolling Granger causality will be used to investigate whether Granger causality changes over time.

[18] Sock puppet account: an online identity used for purposes of deception. Shill account: An account set up to engage in covert advertising.

3.4.4 Granger Causality

Variable A *Granger causes* variable B if including variable A in forecasts improves the accuracy of predicting variable B, compared to only using data from variable B [78]. Granger causality has already been applied in literature relating to cryptocurrency market prediction [77]. One example investigates Granger causality between discussion on online forums and price and transaction count, respectively; it was found that discussion related to China Granger caused the Bitcoin price. It should be noted that variable A may Granger cause variable B without this relationship being one of true causation, as a third variable (variable C) may be causing both variables to change but with variable A leading the variable B. Although not proof of true causation, Granger causality identifies relationships suitable for prediction. There are four possible results when applying Granger causality (posed in the context of this work):

1. Unidirectional Granger causality from the online indicator to the cryptocurrency price. Changes in the online indicator are likely to precede changes in the cryptocurrency price.

2. Unidirectional Granger causality from the cryptocurrency price to the online indicator.

Changes in the cryptocurrency price are likely to precede changes in the online indicator.

3. Bidirectional (or positive feedback) causality. Changes in one variable are likely to precede changes in the other variable and vice versa.

4. No Granger causality of any sort (unidirectional or bidirectional).

Granger causality requires the time series under consideration to be stationary. Again log returns are used due to their stationary nature, seen in Table 4, and their common choice elsewhere (for example, in [99]). Rolling Granger Causality applies a Granger causality test to subsets of the data rather than unconditionally to the whole dataset. A fixed sized rolling/moving window (with, in this work, a window length of 100 data points) moves sequentially from the beginning to the end of the data. Rolling Granger causality has been applied in this way in a number of contexts (for example, [99] and [100] uses rolling Granger causality to identify relationships between regional stock markets). Figure 4 shows p-values of the rolling Granger causality tests between a number of online indicators and the associated cryptocurrency price.

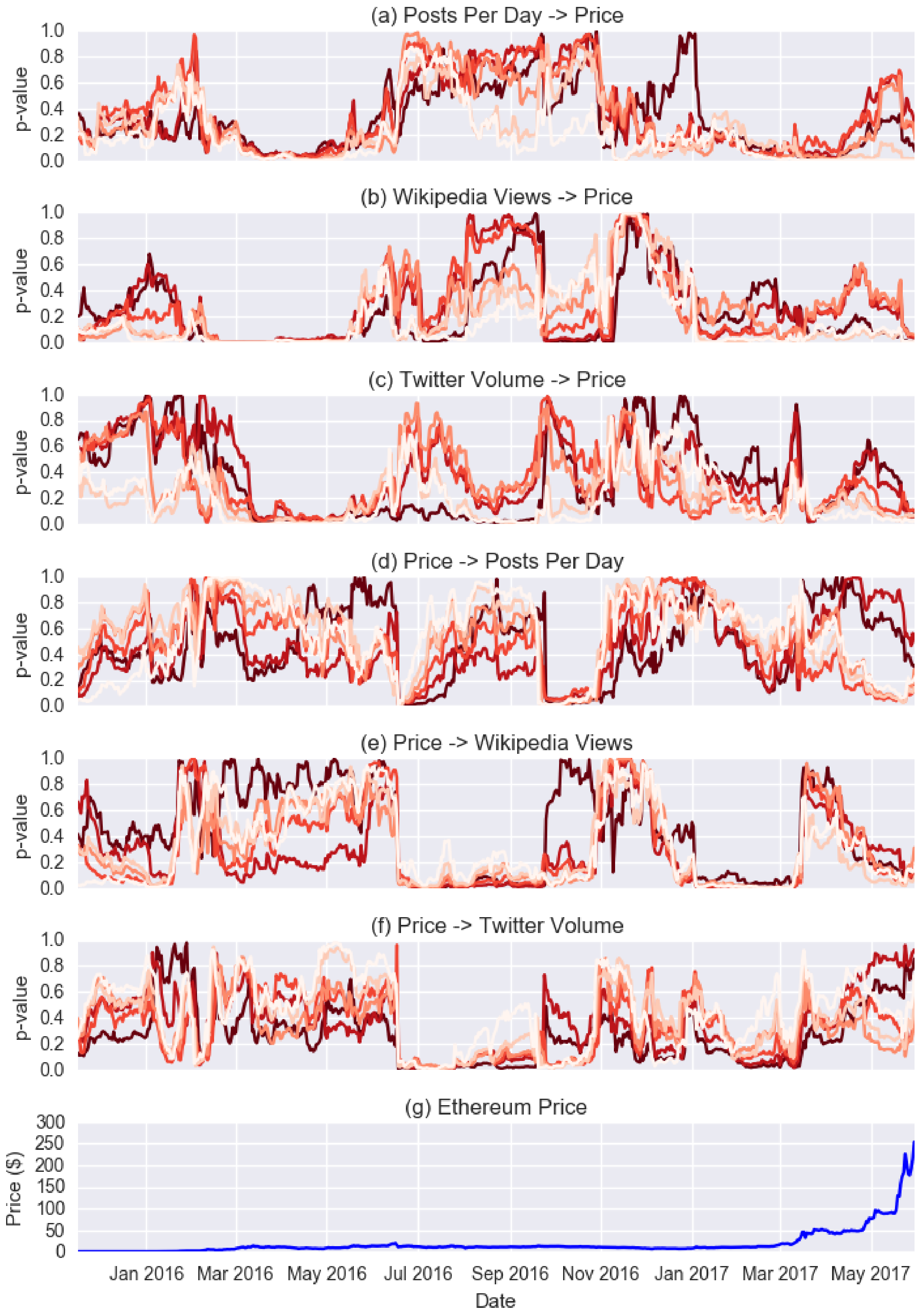

Figure 4. Rolling Granger causality p-values for the relationship between a number of online indicators and the Ethereum price; colours range from dark red (1 day lag) to light red (6 day lag).

A low p-value (shown on the y-axis) suggests the null hypothesis (no Granger causality) can be rejected in favour of Granger causality from one variable to the other. For brevity, only a subset of indicators are considered in relationship to Ethereum. The factors chosen include one Reddit factor (*posts per day*), one Twitter factor (Twitter volume) and one miscellaneous online indicator (Wikipedia views). This choice results in 6 Granger causality diagrams to be examined rather than the 48 diagrams that would result if all combinations were considered (48 = 4 cryptocurrencies multiplied by 6 factors multiplied by 2 directional graphs (factor to price and price to factor) for each relationship). Although only a subset of relationships are documented here, other relationships were also generated for validation and produced similar time-dependent results.

Considering factor to price causality, there are periods where all considered indicator to price relationships ((a), (b) and (c) in the figure above) experience a strengthening of Granger causality. Between February 2016 and July 2016, all factor to price relationships experience a period of low p-values indicating a more significant Granger causality. This is especially of interest as it is unidirectional (ideal for forecasting purposes) and it is during a period where the Ethereum price rises from around $2 to around $15, the most significant rise it had undergone to that point. Likewise, towards the end of the data period considered, significant Granger causality is seen from the indicators (especially posts per day) to price. Again this occurs during a period where the price undergoes a considerable increase (as seen in (g), above, where the price rises from around $15 in March 2017 to around $250 in June 2017).

Time-dependent relationships can again be seen when considering price to indicator causality. However, it appears possible to link the changes over time with significant external events whose occurrence appears to alter the dynamics of any relationship present. For example, all three price to indicators relationships ((d), (e) and (f) above) exhibit a sudden drop in p-values around June 2016 and move into a period where price Granger causes the indicator movements. This coincides with a major event, the hack of the first well-publicised application built on the Ethereum network (the DAO), which had a long-lasting effect on the Ethereum project and on the price of its cryptocurrency. During the subsequent months, it can be seen that price Granger causes the online indicators, especially Wikipedia views and Twitter volume. This may be because technical progress and adoption were limited during this period, and the only times the network was generating external interest (Wikipedia views, especially, being a good indicator of *new* interest [4]) was due to price changes. Another sudden change can be seen in March 2017 where price to factor relationships lose any previous Granger causality (especially apparent for Wikipedia views and Twitter volume). This coincides with a period where Ethereum prices consolidated after a previous dramatic rise in price; during this period, Wikipedia views and Twitter volume dropped. Due to the small price changes during this period, the price changes had little possibility of generating new interest (something captured by Wikipedia views and, to a lesser extent, Twitter volume).

Although reasons for these dramatic changes in Granger causality can be theorised, the need to investigate the raw data to have confidence in the varying Granger causality results may suggest other techniques may be better at capturing the dynamic nature of any relationship present. The next section, as well as summarising the findings of this preliminary data analysis, outlines other ways in which more clarity can be added to the current results.

3.4.5 Summary of preliminary data analysis

In this preliminary data analysis section, a number of statistical tests were undertaken, the findings of these tests being summarised in this section. Firstly, an ADF test was used to discover that many of the raw data time series were non-stationary. It was noted that being non-stationary would cause a weakening of the reliability of results in further statistical tests as the work proceeded, so log returns were calculated, and these were found to be unanimously stationary. Due to this, log returns will be used, where required, in the remainder of the work presented here rather than raw time series data.

Pearson correlation was calculated between the log returns of the considered online indicators and the associated cryptocurrency price series. Correlation was generally found to be weak and dependent on the data period used. It is hypothesised that the correlations exhibited are weakened due to being unable to capture the dynamic nature of relationships present. Isolated situations were demonstrated where the same online indicator could be either positive or negatively correlated with the price, depending on the particular event. Calculating an unconditional correlation over the whole dataset does not capture this well as it is weakened by periods of positive correlation offsetting periods of negative correlation, and vice versa. This prompted the use of rolling Granger causality to capture changing relationships, and also identified two further investigations to be undertaken: 1) a technique needs to be applied to validate that both significant positive and significant negative correlations are occurring (to be done in Chapter 4); 2) rather than looking at count-based metrics like those introduced here (e.g. *posts per day*, *new authors* per day, etc.) one way to find metrics that may have a consistent directional relationship with the price would be to look at the content of submissions. In other words, although *posts per day* may be positively or negatively correlated with price, depending on the occasion, *posts per day* relating to a particular topic may be consistently correlated in one direction with the price. Analysis of content related metrics is conducted in Chapter 6.

The use of rolling Granger causality allowed for time-dependent relationships to be identified. Of most interest, it was found that the indicators appear to undergo periods where there is Granger causality towards the price. As well as the periods of Granger causality from online indicator to price, there are also periods of bidirectional relationships and periods where there are no relationships. The next chapter investigates the cause of time-specific relationships (in other words, why relationships are sometimes present but not present at other times); to achieve this a tool is applied (wavelet coherence) that can better track and visualise the complex (and changing) relationships than the already examined Granger causality.

The next chapter will also consider relationships occurring over different time durations. In the correlation analysis discussed in this chapter only same day correlations were considered; however this does not capture longer-term relationships.

Chapter 4

Cryptocurrency Price Drivers: Wavelet Coherence Analysis Revisited

As discussed towards the end of the literature review (Chapter 2), a number of relationships between online indicators and cryptocurrency prices have previously been identified. In these works the authors typically demonstrated the existence of such relationships through the generation of (profitable) trading strategies. However, a complex (and potentially rule-based) trading strategy may obscure the real relationship between individual factors and related price. Here, it is investigated whether raw relationships between online indicators and cryptocurrency prices exist and whether they are leading relationships (whereby the online indicator leads the price) rather than lagging relationships. In addition, it is investigated whether relationships exist over the short, medium and long term, providing knowledge to influence the holding time of later prediction strategies.

It is apparent from one previous study [69] that relationships between particular factors and the Bitcoin price are not consistently present. It is here hypothesised that a cryptocurrency's relationship with potentially relevant online indicators is dependent on the market regime. Market regimes have previously been observed in cryptocurrency markets, for example, bull and bear markets [101], and also more commonly the presence of speculative bubbles. This observation about changing market regimes should shed also more light on the changing (seemingly time-dependent) relationships observed in the preliminary data analysis described in the previous chapter (Chapter 3).

4.1 Background and related work

4.1.1 Speculative bubbles within cryptocurrency markets

Accompanying the rise in mainstream news pertaining to cryptocurrencies has been a rise in articles discussing the possibility of whether cryptocurrency prices are undergoing a price bubble. However, one detail is often overlooked—what exactly is an asset price bubble? Among experts within both academia and industry, there is no general agreement on the exact definition of a price bubble, with many authors providing their own interpretation [102]. Furthermore, interpretations can either be from an economic perspective (e.g. understanding the causes of bubbles) or from a statistical/econometric perspective (usually detecting the presence of time series movements indicative of a bubble rather than the exact cause). Most current work within a cryptocurrency-specific context focuses on the statistical/econometric detection of bubbles; this section reviews such work as the detection of bubbles is required as a part of this chapter.

One common method to detect whether an asset is in a bubble regime is to examine whether the asset's price is deviating from the asset's 'fundamental value'. This means that before identifying bubbles, an asset's fundamental value needs to be identified. The fundamental value of a stock is usually judged as the present value of future cash flows the owner of the asset is expected to receive [103]. Using this methodology for most cryptocurrencies would be challenging as they lack clearly identifiable future cash flows [12]. One study, in fact, conjectured—rightly or wrongly—that the fundamental value of Bitcoin is zero [12]. Thus alternative bubble-detection approaches which avoid identification of an asset's fundamental value have become popular within cryptocurrency-related research, as detailed below.

The price movements of Bitcoin preceding 2014 have become a common focus given the substantial price rise then observed. Several statistical/econometric-based bubble-detection techniques have been applied to identify the presence of at least one bubble during this period [12]. One particular subset of statistical/econometric-based tests [104] exhibits the ability to timestamp when an asset price series enters and exits the bubble regime. The application of such tests during this pre-2014 data period identifies the starting and ending point of a number of short-lived bubbles and then the existence of three larger bubbles towards the end of 2013, each lasting between 66 and 106 days [12]. More recently, a sample period between December 2016 and January 2018 was examined [105]. The authors combined a number of well-known bubble detection tests to create a larger bubble detection system, finding evidence of bubble-like price movements for Ethereum in June 2017 and for Bitcoin in December 2017.

Instead of relying solely on price movements to detect bubbles, one recent study looked at valuing Bitcoin based on the number of active users [106]. To achieve this, the authors harnessed Metcalfe's law, which states that the value of a network (in the original case, a telecommunication network) is proportional to the square of the number of nodes [107]. The number of active users of Bitcoin was quantified by a proxy examining the number of active blockchain addresses. The authors noted that this proxy is not perfect as one user can own multiple blockchain addresses and address rotation is common. In addition substantial blockchain activity can appear and disappear based on fads (e.g. blockchain-based games such as CryptoKitties[19]). Once a value for Bitcoin is quantified via Metcalfe's law, the price can then be examined to identify periods where it diverges from such a value. The authors identified four significant bubble-like regimes in the history of Bitcoin (2010 to 2018) where price diverges from the value anticipated by Metcalfe's law.

However, users of the blockchain may not be the parties actively speculating on the associated cryptocurrency prices, and thus other usage metrics may better capture this. Another study investigating cryptocurrency price bubbles instead considers users on social media [9] (who may, in fact, be the speculators driving the price). The authors found the relationship between price and Twitter submissions acts as an amplification mechanism; such amplification mechanisms are a commonly identified component of the propagation of speculative bubbles [108]. To be explicit, a positive feedback loop is identified whereby firstly price increases cause search volume to increase, which in turn cause mentions on Twitter submissions to increase, with this, in turn, causing a further increase in price.

[19] https://media.consensys.net/the-inside-story-of-the-cryptokitties-congestion-crisis-499b35d119cc

4.1.2 Temporal factor relationships

In an attempt to understand what factors can be indicative of future price movements, one (previously mentioned) study applied a technique called wavelet coherence to several factors of interest to investigate each factor's relationship with the Bitcoin price [69]. A number of factors were considered, including trading volume, several blockchain statistics (e.g. mining power, difficulty and number of transactions), Google search volumes and Wikipedia views. Due to the visualisation provided by wavelet coherence, it is possible to inspect how any identified relationships changed over time. Although relationships were identified between both Google searches and Wikipedia views and the price, these seemed to appear and disappear at different points in time, and it was beyond the scope of the investigation in [69] to understand why this is.

More generally, the application of wavelet coherence allows for the identification of time-evolving relationships occurring over the short, medium and long term. Wavelet coherence has become a common choice in the financial literature to inspect relationships within stock indices [109], commodities [110], cross-asset behaviour [111], and between social media and stock prices [112]. Wavelet-based analysis has been used to identify co-movement between Bitcoin and, separately, global uncertainty [113] and regional markets [114].

The work in this chapter will revisit and extend the application of wavelet coherence of [69] using a longer data period and additional factors; this will provide information on whether relationships, especially with newly introduced factors, are leading or lagging. A timestamping bubble detection technique is applied to the price time series, as is common elsewhere in similar contexts [12] [105]; this statistical/econometric technique detects bubbles by examining price series movements, rather than devising an economic model/theory to understand why a bubble is occurring. This method is suitable for the current work in this chapter as only the detection of bubble regimes is required. The combination of wavelet coherence with bubble detection will be used to determine whether factor relationships are dependent on the current market regime (i.e. bubble or non-bubble regime). Additionally, relationships between different cryptocurrency price series are investigated to see if there are relationships between cryptocurrencies; this is of interest for trading strategies which involve multiple cryptocurrencies.

4.2 Methodology

4.2.1 Wavelets

A comprehensive explanation of wavelet methodologies can be found for example in [109], [110], [115]; this section aims to provide an overview based on the presentation in these papers.

Wavelets are wavelike functions used to transform signals into a representation which has time and frequency domain components. Visually wavelets appear as wave-like oscillations with an amplitude that starts at zero, increases, then returns to zero. Another way to consider a wavelet is as a bandpass filter, which can be applied to a time series under investigation, letting through only components of the time series within a finite range of frequencies to different extents depending on the energy spectrum of the wavelet. Wavelets take the form

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)$$

The *u* parameter specifies the location of the wavelet. The scale parameter *s* refers to the width of the wavelet, indicating how stretched or dilated the wavelet is while retaining the same wavelike shape. Larger values of *s* increase the width of the wavelet, and therefore more of the observed time series is considered, but granularity of the observation is reduced meaning a higher-level view of the time series is taken. Low scales will allow for analysis of (higher frequency) short-term dynamics of the time series under consideration, whereas high scales will allow for analysis of (lower frequency) long-term dynamics. If the wavelet and time series follow a similar pattern at a specific temporal location and scale, then a large transform value is generated. If the wavelet function is applied in a continuous fashion, as done in this work, this is referred to as *continuous wavelet transform*. The continuous wavelet transform is defined as

$$W_x(\mathrm{u,s}) = \int_{-\infty}^{+\infty} x(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)\,dt$$

where ψ^* is the complex conjugate of ψ. There are many examples of functions that can be categorised as a wavelet. As it has been used in similar previous financial applications [109] [110], the Morlet wavelet will be used here. It is made up of a normalisation factor, complex sinusoid, and Gaussian bell curve. It is essentially a sine wave multiplied point by point by a Gaussian. The Morlet wavelet is defined as

$$\psi^M(t) = \frac{1}{\pi^{1/4}} e^{i\omega_0 t} e^{-t^2/2}$$

where ω_0 is chosen to be 6, a good choice for feature extraction purposes [115] which is a commonly chosen value in similar work [109] [110]. Continuous wavelet transforms are useful when considering a time series and breaking down and examining its constituent waveforms. It is also possible to use another wavelet transform, the *cross wavelet transform*, to examine two time series with the aim of identifying locations where similar correlations with a particular wavelet exist. This is defined for two continuous wavelet transforms, $W_x(u,s)$ and $W_y(u,s)$, as

$$W_{x,y}(u,s) = W_x(u,s)\, W^*{}_y(u,s)$$

where * denotes the complex conjugate. Regions that have high values in both continuous wavelet transforms will result in high cross wavelet power ($|W_{x,y}(u,s)|$).

As in previous work [69] [109] [110], it is of more interest here whether the time series being considered co-move than whether they produce large cross wavelet transform values, and *wavelet coherence* is utilised for this purpose. Wavelet coherence is defined as

$$R^2(u,s) = \frac{\left|S\left(s^{-1}W_{x,y}(u,s)\right)\right|^2}{S(s^{-1}|W_x(u,s)|^2)S\left(s^{-1}\left|W_y(u,s)\right|^2\right)}$$

where S is a smoothing operator applied in both the time and frequency domain (the smoothing operator used in this work is described by [115]). Wavelet coherence is the ratio of the cross wavelet power to the product of the individual wavelet power, comparable to the squared coefficient of correlation; essentially this is providing the correlation coefficient around each moment in time and for each frequency. It can be used to identify regions in time-frequency space where the two time series being examined move in a similar way, though they do not necessarily display high power. A map of phase differences between the signals can also be obtained. This can be used to identify the lag between the two time series (which series is leading and which series is lagging).

4.2.2 Further details and interpretation of wavelet coherence scalograms

Figure 5 shows an example wavelet coherence scalogram (the wavelet coherence scalogram for Bitcoin and Litecoin which will be analysed later). All following scalograms use the cross wavelet and wavelet coherence software provided by A. Grinsted [115].

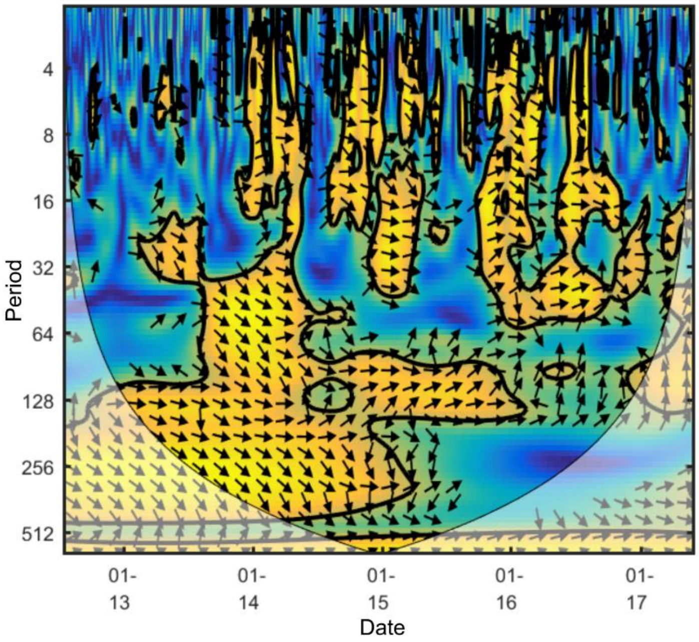

Figure 5. Example wavelet coherence scalogram

The horizontal axis in the above figure shows the time; relationships positioned towards the leftmost area of a diagram occurred at the start of the data interval considered, and those at the rightmost end occurred at the end of the data interval considered. The vertical axis shows the period; lower period bands (higher frequencies) are shown near the top and higher bands (lower frequencies) are near the bottom. Lower bands would be of interest to investors with short term horizons, whereas higher bands would be of interest to investors with longer-term horizons.

Wavelet coherence plots as illustrated above highlight areas in the time-frequency space where the two series co-move. The warmer the colour, the higher the coherence (which can be interpreted as correlation) at that location in the time-frequency space. The colours used in this work range from dark blue (0, no coherence) to yellow (1, strong coherence). Statistically significant areas of coherence are surrounded by a thick black line.

The direction of the *oriented arrows* displays two things: the correlation, and which of the two time series is leading the relationship at that point. An arrow pointing left is *anti-phase*, meaning that the two time series are negatively correlated at this location. An arrow pointing right is *in-phase* meaning the two time series are positively correlated at this location. A downward arrow means the first time series is leading the second whereas an upwards arrow means the second time series is leading the first. In Figure 5 it is possible for example to see arrows pointing southeast during 2013/2014 and the period band of 256-512 days; this can be interpreted as the two time series being positively correlated at that time, with the first series (Bitcoin price) leading the second series (Litecoin price). In the later scalograms that include an online factor and price the online factor will always be the first time series and the price series the second, meaning a downward arrow will indicate that the factor is leading the price.

At each point information from neighbouring data is used. As the time series considered are finite, the areas at the start and end of the data (especially at higher period bands) will not have all the data required. One solution to make computation possible, chosen here, is to pad the time series with zeros where required. However, the zero padding will impact the reliability of the results. It is standard to use a *cone of influence* to represent this difference in reliability of results. Pale colours represent those areas outside the cone of influence (an example cone being visible in Figure 5), with less reliable results. Higher period bands require more data for computation resulting in the cone shape.

4.2.3 Bubble detection using the GSADF test

In order to provide a methodology to detect bubbles in time series, Phillips, Wu, and Yu [116] proposed the *supremum augmented Dickey-Fuller* (SADF) test. This applies a series of right-tailed unit root tests to expanding windows of a time series (with a fixed start date), defined by

$$\mathrm{SADF}(\mathrm{r}_0) = \sup_{\mathrm{r}_2 \in [\mathrm{r}_0,1]} \mathrm{ADF}_0^{\mathrm{r}_2}$$

where r_2 is the final data point to be considered in each window, starting at r_0 which is a fraction representing the smallest allowed window size and expanding to 1 (the complete data set). The SADF test finds the largest ADF statistic from all the windows considered. If this value exceeds a critical value, the null hypothesis can be rejected, and it is deemed the series displays explosive behaviour in at least one of the windows (taken as an indication of a bubble occurring).

Although this test successfully detects single isolated bubbles, Phillips, Shi, and Yu [104] acknowledge it may suffer from reduced discriminatory power when applied to time series with multiple occurrences of bubbles. To overcome this weakness, a further enhancement was proposed, as a new method, called a *generalized supremum ADF* (GSADF) test. This test allows both the start and end points of data subsets to vary, which in turn enables the identification of multiple bubble regimes in one observed time series. The GSADF test is defined by

$$\mathrm{GSADF}(r_0) = \sup_{\substack{r_2 \in [r_0, 1], \\ r_1 \in [0, r_2 - r_0]}} \mathrm{ADF}_{r_1}^{r_2}$$

Whereas in the original SADF test the starting value of the window, r_1, was fixed to 0, in the GSADF test the starting point can now vary from 0 to $r_2 - r_0$ (this is the last possible starting point, near the end of the data set, that allows the test to be run on the minimum window size).

A further methodology was introduced in [116] that gave better results compared to SADF when using a backward expanding window; this was denoted *backward SADF* (BSADF). This performs the same supremum ADF test, but with a fixed ending point, r_2, and backwards expanding window:

$$\mathrm{BSADF}_{r_2}(r_0) = \sup_{r_1 \in [0, r_2 - r_0]} \mathrm{ADF}_{r_1}^{r_2}$$

Combining the BSADF with the GSADF test allows the r_2 value to vary while still using a backward expanding window. r_2 starts at the smallest possible window size, and moves one point at a time towards the end of the time series.

$$\mathrm{GSADF}(r_0) = \sup_{r_2 \in [r_0, 1]} \mathrm{BSADF}_{r_2}(r_0)$$

The GSADF method can be used to date stamp the start and end of bubble regimes. At each point of r_2, the BSADF statistic is generated. The start of a bubble is defined as the first r_2 value that generates a BSADF value larger than the appropriate critical value (the null hypothesis of a unit root in the time series is rejected in favour of a mildly explosive alternative). The end of the bubble is the first r_2 after the start point such that the BSADF statistic is smaller than the critical value. Critical values are obtained via Monte Carlo simulation of a random walk (Wiener process) whereby the random walk is generated by the partial sums of N(0,1). Generation of these values for the current work proved to be computationally expensive. A cloud-based infrastructure was used, enabling the work to be parallelised and provided a speed up of around 46 times compared to calculating the values on a single CPU. Convenient integration between Matlab and Google Cloud was achieved by using software called Techila Technologies.

As noted earlier in Section 4.1.1, there is not a widely accepted or consistent definition of the term “bubble”. The GSADF test used here assumes a bubble is any time series interval which deviates from a random walk to become explosive.

4.2.4 Data pre-processing

Raw time series can be multi-modal. This is especially apparent for financial asset price time series, as prices are likely to locate around psychological supports and resistances [117]. Multi-modal distributions are not ideal for use in wavelet analysis, and it is advised to transform the time series to avoid such distributions [115]. Log returns (shown to be stationary in Section 3.4.2) are used instead of the raw time series. Log returns are commonly used elsewhere within wavelet coherence work [109] [110]) and result in unimodal distributions nearer the normal distribution. The same transformation is applied to all online metric time series, to the same effect. As a result all the time series under examination can be considered as growth rates rather than absolute amounts; this is an important design decision as one would expect peaks in growth rates to lead peaks in absolute values (and as such could be interpreted wrongly as a leading relationship, if one time series was growth rates and another absolute values).

Figure 6 shows the raw price series evolution for each cryptocurrency considered (prior to conversion into log returns); it is plain such price series would be unlikely to display stationarity and that a data transformation would be required.

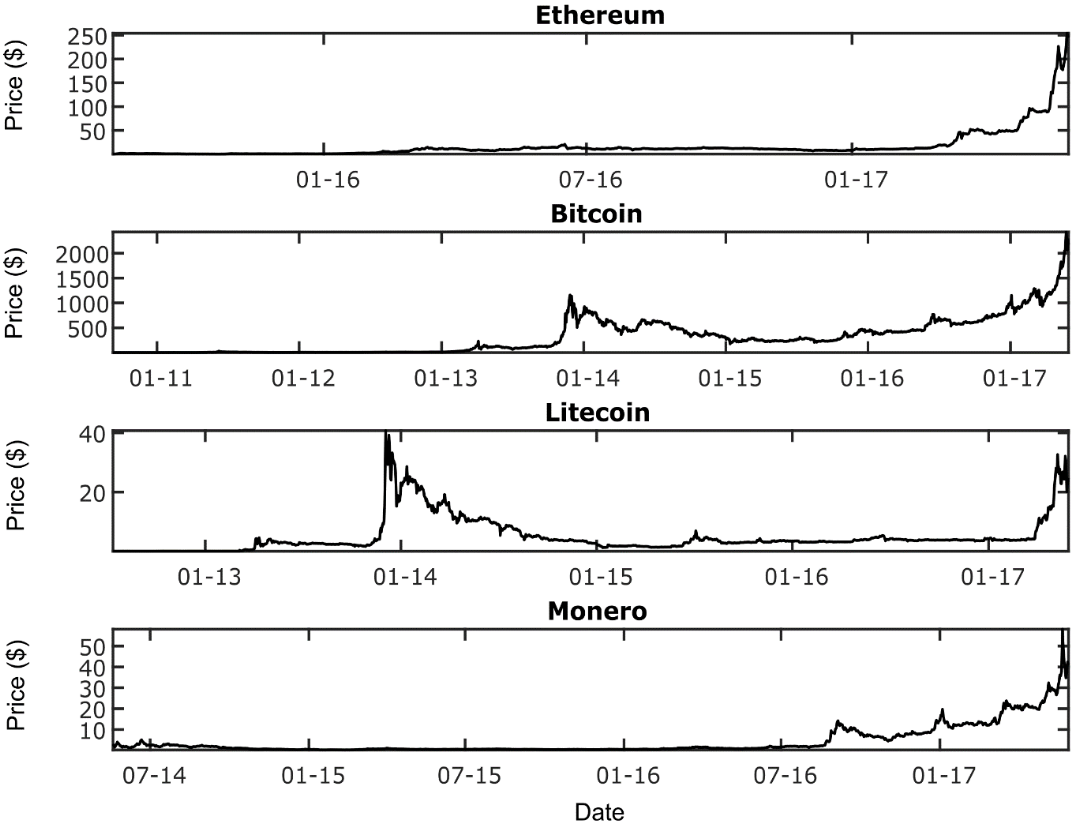

Figure 6. Price series for each cryptocurrency considered (each cryptocurrency priced in USD)

4.3 Results: Coherence between cryptocurrencies and online factors

Figure 7 and Figure 8 present the wavelet coherence scalograms between the different cryptocurrency and factor combinations. Each column contains scalograms for a different cryptocurrency; each row contains scalograms for a different factor. Looking down a column shows how a certain cryptocurrency is associated with different factors. The red shaded areas indicate locations within a cryptocurrency's price time series that have been identified as in a bubble-like regime, using the GSADF bubble test [104] described previously. It should be noted that the dark blue areas between 2010 and 2012 for Bitcoin *subscriber growth*, Google Trends, and Wikipedia views are due to a lack of data for these metrics prior to 2012.

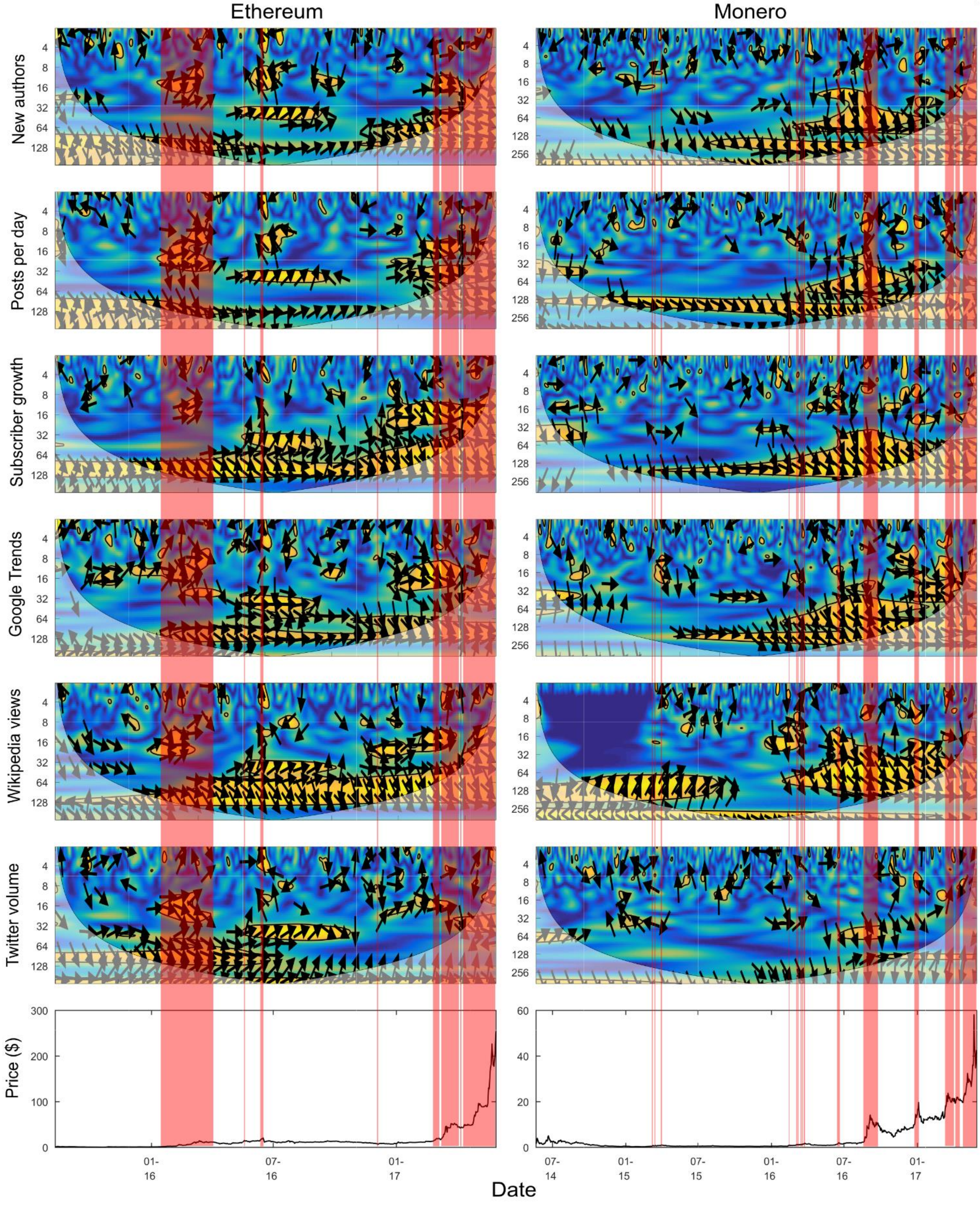

Figure 7. Wavelet coherence scalograms between online factors and price (with GSADF test bubble overlay) for Ethereum and Monero

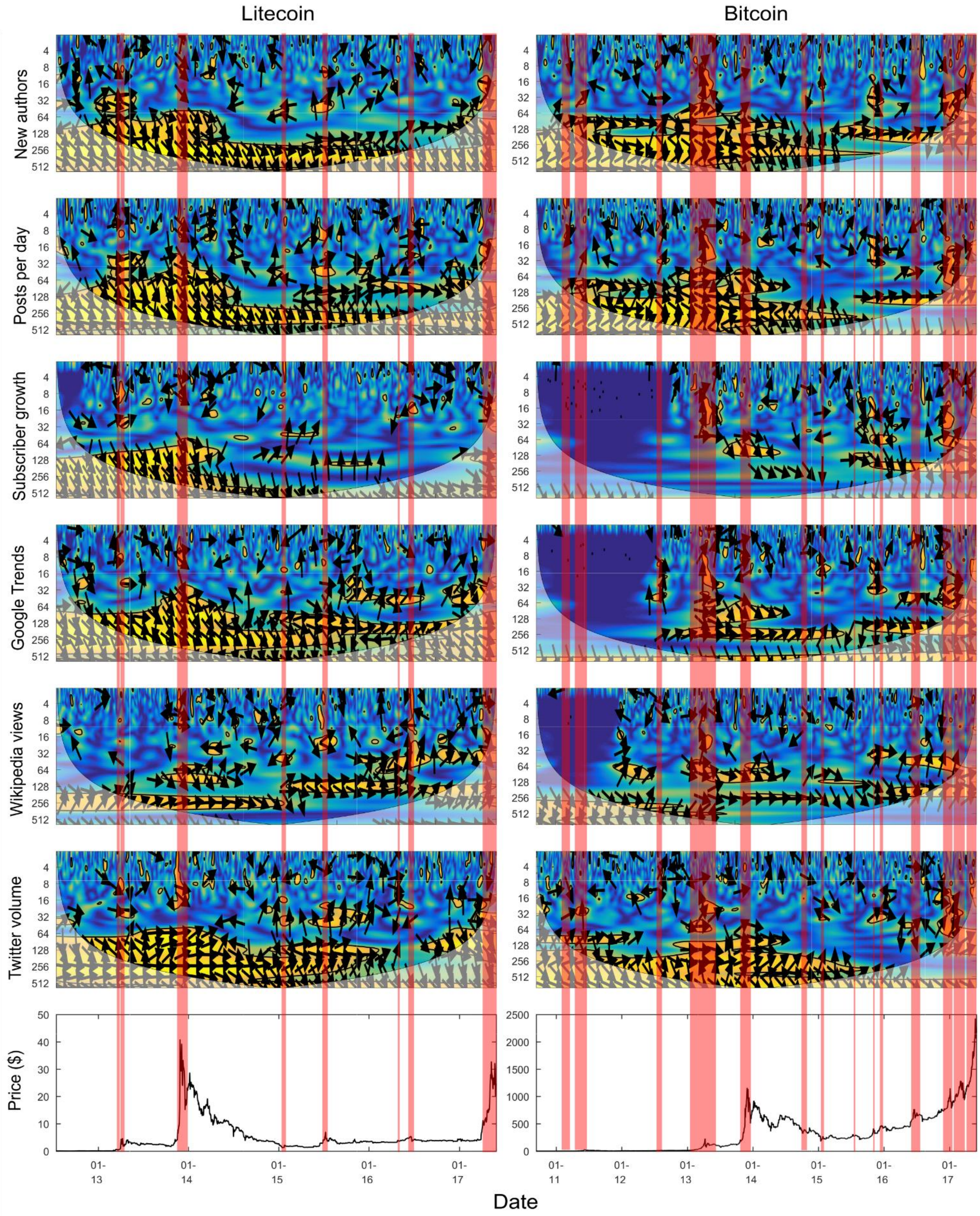

Figure 8. Wavelet coherence scalograms between online factors and price (with GSADF test bubble overlay) for Litecoin and Bitcoin.

For the sake of clarity an explicit definition of short, medium, and long term is required. In this work, *short term* refers to the 2-4 and 4-8 day period bands. *Medium-term* refers to the 8-16 and 16-32 day bands. *Long term* will be used to refer to the 32-64, 64-128, 128-256 and 256-512 day bands. The short, medium, and long term bands will be considered separately to begin with, and then considered collectively alongside the results of the GSADF bubble test.

4.3.1 Short term relationships

Although short term relationships are erratic and sparse, this period band contains examples of negative—although usually extremely fleeting—relationships (shown by leftward facing arrows). The relationships link online activity increases to price falls (the converse is not observed). It is not surprising that occasionally discussion is associated with price falls, as negative events (e.g. blockchain bugs, and exchange hacks) are newsworthy in the community.

One example is the negative correlation that occurs between Ethereum and its associated factors around June 2016 (left facing arrows at the top and just left of the horizontal middle of the Ethereum scalograms). During this time interval, one of the best-known applications at the time, the DAO, built on top of the Ethereum environment, was hacked. It can be seen that all factors are negatively correlated in the short term with the price during this time interval. As a result of the uncertainty generated by the hack, price dropped sharply, but activity on social media and interest increased (causing the negative correlation). The negative relationship can be seen during the 2-4 day band for all factors.

In the short term, situations occur where the factors lead the price and where the factors lag the price. In many cases, the factor lags the price in the short term (seen by upward facing arrows near the top of each scalogram). This is understandable given short term changes appear likely to be the result of particular events, as discussed above. It is likely the market price will reflect the event quicker than social media; social media may experience a longer interval of discussion and activity relating to the original event and resulting price change.

The erratic and spare nature of relationships in the short term may demonstrate that short term price changes are caused by a confluence of factors and that online metrics may not be the most interrelated factor with price changes in the short-term. The following examples are given to show what factors can effect cryptocurrency prices in the short-term; both examples are unrelated to the adoption-related online metrics considered in this work. As a first example, it is common within cryptocurrency markets for intraday traders to follow technical analysis pattern-based trading strategies. Enough traders following these will cause short term price changes based on the indicators they are watching (if enough traders buy believing the price will go up, this will become self-fulfilling). As a second example, as will be documented later in Section 4.4, there exist isolated periods of short-term coherence between different cryptocurrency prices. Examining cryptocurrency-specific online metrics without regard to the general cryptocurrency ecosystem may not provide a complete picture. For example, if a favourable news article occurs for, say, Ethereum, the price of Ethereum may go up, while the price of Bitcoin may go down, as people sell Bitcoin to buy Ethereum. This short-term movement of the Bitcoin price may be unexplainable by Bitcoin-related online metrics.

4.3.2 Medium-term relationships

Relationships in the medium term are much less erratic than those observed in the short term. Considering Figure 7 and Figure 8 together, there are distinct patches of strong relationships separated by substantial areas with no relationship present. The relationships are predominately positively correlated, with the clearest exception being the Ethereum DAO hack (June 2016) discussed above, which displays negative medium-term correlation for the *new authors* and *posts per day* factors (seen in the 8-16 day band just left of the horizontal middle of the Ethereum scalograms). In the medium term, there is improved consistency (compared to the short term) as to whether the factor or price is leading, with in many cases co-movement being observed where both time series change around the same time. It is observed that during the prolonged bubble regimes (especially in the case of Ethereum and Monero), there are frequently periods where the online indicators are leading the price (with the exception of Twitter volume, which is discussed in the next paragraph). A further consideration is given to the bubble regimes in Section 4.3.4, which also provides an explanation for the temporal emergence of medium-term relationships.

Relationships between volume of discussion on Twitter and the associated cryptocurrency price are seen in similar medium-term regions as relationships exhibited by the other considered indicators. However, in many cases Twitter volume is not leading the price and, in fact, for much of the time is lagging the price. For example almost all significant relationships exhibited during this interval with the Monero price occur where Twitter volume lags the price.

4.3.3 Long term relationships

Longer term relationships appear more consistent over time and do not appear directly affected by individual news items. Almost all long term relationships are consistently positive when they exist, suggesting a positive long term relationship between price and online activity. The lack of consistency of Wikipedia views and consistency of Reddit factors in leading the prices indicate that the Reddit derived factors are better predictive indicators in the long term. *Posts per day*, *new authors*, and *subscriber growth* (all the metrics derived from Reddit) are predominately leading the price in the long term (shown by largely downward oriented arrows). In contrast, Google Trends has more locations where there is no obvious leader, and Wikipedia views has more variation than the other factors; there is no consistent leader in the relationships with Bitcoin and Litecoin. For the other two cryptocurrencies considered there are intervals where Wikipedia views significantly lag the Monero price, but in contrast, Wikipedia views lead the Ethereum price throughout the data interval considered.

Interestingly, in the long term, although Twitter volume experienced earlier significant relationships with cryptocurrency prices (most notably the 2013/2014 Bitcoin and Litecoin bubble and subsequent decline), it has not exhibited as many recent relationships (when compared to the other indicators). This is seen, for example, by large areas of blue towards the bottom right of the Ethereum (post 09/2016) and Bitcoin Twitter scalogram (post 01/2016). Potentially, this is because the term 'Bitcoin' is now used commonly when referring to the whole of cryptocurrency market/ecosystem, so tweets may have continued to rise over 2015/2016 while the Bitcoin price in contrast went through a period of consolidation. In addition hundreds of other cryptocurrencies were launched post-2016, increasing the breadth of the market. This increased the size of the ecosystem not specifically Bitcoin-related, while being possibly erroneously labelled as ‘Bitcoin’. This effect would occur for all indicators, but seems especially apparent for the Twitter-derived metric, potentially suggesting that the Twitter-derived metric is capturing a different kind of interest/different kind of user (possibly in terms of prior knowledge) than metrics derived from Reddit. The absence of significant long term relationships between Twitter volume and the

Monero price should also be noted. This may be caused by Monero being a slightly more, and possibly increasingly[20], niche cryptocurrency than the other cryptocurrencies considered (Bitcoin, Litecoin and Ethereum can be purchased on Coinbase, the largest retail facing cryptocurrency exchanges, while Monero cannot). The lack of market-related discussion relating to a more niche cryptocurrency on Twitter again suggests Twitter is capturing a different user demographic (e.g. mainstream cryptocurrency interest) than the other data sources considered.

4.3.4 Bubble regimes and changing factor relationships

Looking at the bubble regimes (shaded red areas) identified by the GSADF test, it appears there is a strengthening of the medium term—and to some extent long term—coherence relationships within the time intervals identified as being bubble-like regimes; this can be justified intuitively by considering that interest is likely to rise as price rises. This result echoes other work which found that social media factors and price are likely to exhibit positive feedback loops [9], whereby increasing social media usage causes the price to increase and vice versa, reinforcing one another. The strengthening of medium-term relationships can be seen, to different extents, for all of the factors considered. An example of this is Ethereum between January 2016 and April 2016 (seen in the leftmost red shaded area of the Ethereum scalograms) where during a prolonged interval identified as a bubble, positive coherence forms between all factors (most prominently in *posts per day*) and the price.

Long term relationships also strengthen, to some extent, around areas indicated as bubbles. The previously observed long term relationship between Google Trends and Bitcoin price [69] can also be seen here, between late 2012 and 2014 (period band 64-256). With the benefit of extra data it can be observed that the relationship disappears around 2014 (for lower period bands) and 2015 (for higher period bands), before the relationships start occurring more consistently in 2016 and 2017 (a region with a number of bubbles identified). The previously

[20] Over the last two years, there have been several other projects that have captured mainstream interest and received significant investment. They have possibly diverted interest proportionally away from Monero. For example, the respective market capitalisations of Ripple, EOS and TRON have overtaken the market capitalisation of Monero.

observed relationship between Wikipedia views and Bitcoin observed in 2013 (64-128 band), disappears before again returning in mid-2016 and 2017.

Interpretation of visual scalograms is subjective, so it is desirable to use a more quantifiable way to validate the strengthening of coherence in bubble regimes. Figure 9 shows the wavelet coherence (as introduced in Section 4.1.2) over time for the different period bands, in the case of the *new authors* factor for Ethereum. Coherence values, plotted on the vertical axis, vary between zero and one.

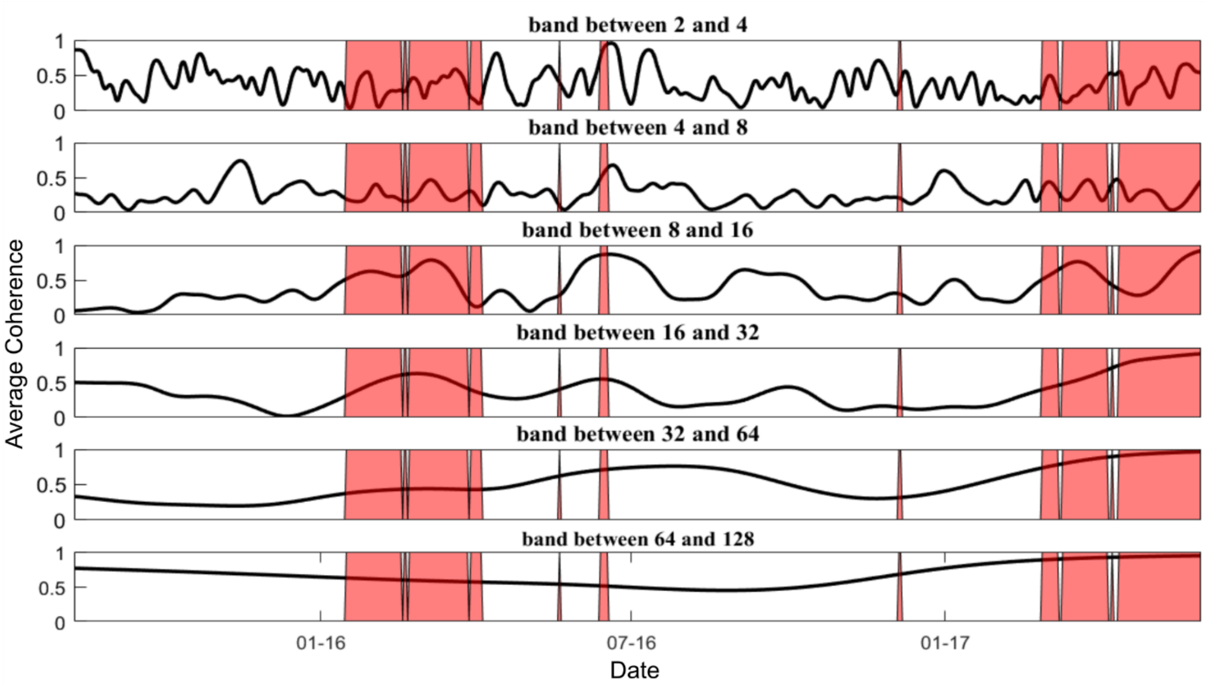

Figure 9. Wavelet coherence between Ethereum new authors and price decomposed for different period bands (with GSADF test bubble regimes shaded red)

It can be seen from Figure 9 that coherence in the short run is erratic throughout the time interval analysed and that there is little appreciable difference between the bubble and non-bubble regimes. However in the medium term (8-16 and 16-32 days), coherence generally peaks around areas where bubbles have been identified in the price series. The longer term relationship, though, is less dependent on whether the price is in a bubble phase.

Although analysis of a single factor and cryptocurrency combination, as above, is of interest, more general findings can also be pursued. Figure 10 shows, for each cryptocurrency and factor combination, the mean coherence values during the bubble and non-bubble regimes. Each horizontal subplot shows a different coherence period band, from the lowest period band (2-4 days) at the top to the highest period band (256-512 days) at the bottom. As the duration of data for each cryptocurrency varies, certain ranges are left blank when that cryptocurrency does not have enough data to produce values for such bands.

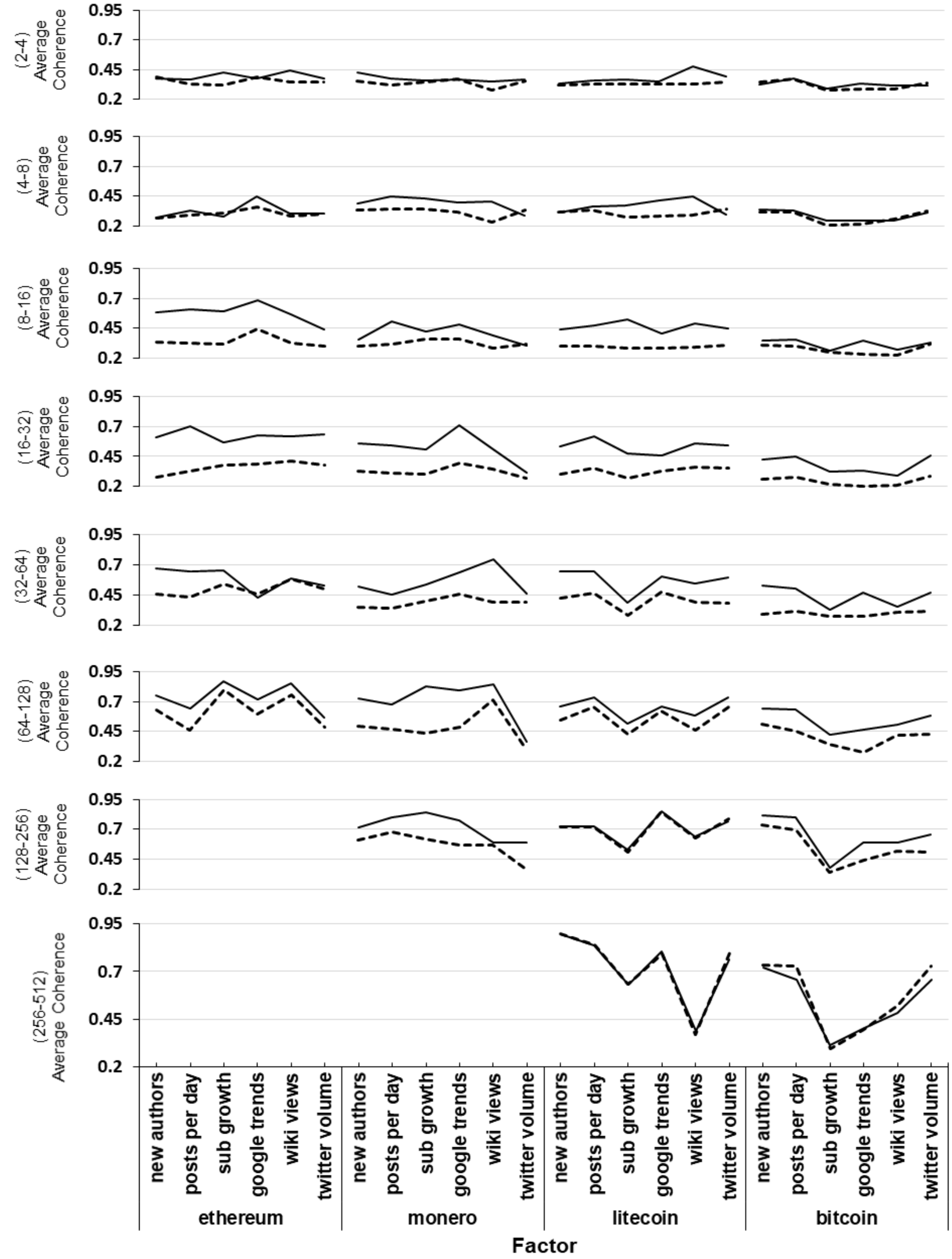

Figure 10. Visualisation of the average wavelet coherence values for bubble (solid) and non-bubble (dashed) regimes decomposed by period band

From Figure 10 it can be seen that, for all cryptocurrency/factor combinations, there is very little difference in coherence values between the bubble and non-bubble regimes in the 2-4 day band. In the 4-8 band, some differences are observed, but without consistency (there are occurrences of bubble regime coherence values being below the non-bubble regime values). In the 8-16 and 16-32 day period bands, large differences can be seen in the coherence values between the bubble and non-bubble regime (for all factors), with the bubble regime coherence being consistently above the non-bubble regime coherence. Ethereum exhibits the largest medium term (8-16 and 16-32) differences in coherence values between its factors for bubble and non-bubble regimes. The differences observed start to reduce as the period bands get larger (with the exception of Monero which exhibits longer-term differences). Almost all impact of the bubble regime has disappeared by the 256-512 data band (for those cryptocurrencies with enough data to generate results), where very similar values are seen for the bubble and non-bubble regimes. It can in addition be observed from Figure 10 that as the period band considered increases, the overall (bubble and non-bubble) coherence values generally get stronger, suggesting online factors have a medium to long term link with price.

Bitcoin’s coherence values appear noticeably less affected by bubble and non-bubble regimes, especially over short and medium terms (2-4, 4-8, 8-16 and 16-32). The non-bubble coherence values are similar to those of the other cryptocurrencies, but the bubble regime values do not reach a similar magnitude to the other cryptocurrencies.

To validate whether the coherence values observed in the bubble and non-bubble regimes are statistically different, a two-sample one-tailed t-test is conducted (for each cryptocurrency/metric pair). A one-tailed test is chosen as it is only of interest whether the coherence values in the bubble regime are statistically larger than in the non-bubble regime. The null hypothesis is that there is no statistically significant difference between the coherence values in bubble and non-bubble regimes and the alternative hypothesis is that the coherence values in the bubble regime are statistically larger than the non-bubble regime. The t-test p-values are listed in Table 7. The cells with a p-value smaller than 0.01 are highlighted grey; in such cases, the null hypothesis can be rejected in favour of the alternative hypothesis.

Table 7. T-test *p-values (for each period band of each cryptocurrency / metric pair)*

		Period band							
Cryptocurrency	Metric	2-4	4-8	8-16	16-32	32-64	64-128	128-256	256-512
Ethereum	New authors	0.200	0.323	0.000	0.000	0.000	0.000		
	Posts per day	0.017	0.001	0.000	0.000	0.000	0.000		
	Subscriber growth	0.000	0.005	0.000	0.000	0.000	0.000		
	Google trends	0.219	0.000	0.000	0.000	0.068	0.000		
	Wikipedia views	0.000	0.023	0.000	0.000	0.420	0.000		
	Twitter volume	0.067	0.210	0.000	0.000	0.019	0.000		
Monero	New authors	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	Posts per day	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	Subscriber growth	0.170	0.000	0.000	0.000	0.000	0.000	0.000	
	Google trends	0.282	0.000	0.000	0.000	0.000	0.000	0.000	
	Wikipedia views	0.000	0.000	0.000	0.000	0.000	0.000	0.103	
	Twitter volume	0.264	0.000	0.139	0.000	0.000	0.000	0.000	
Bitcoin	New authors	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.070
	Posts per day	0.146	0.029	0.000	0.000	0.000	0.000	0.000	0.000
	Subscriber growth	0.059	0.001	0.162	0.000	0.000	0.000	0.002	0.000
	Google trends	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.093
	Wikipedia views	0.001	0.082	0.000	0.000	0.000	0.000	0.000	0.000
	Twitter volume	0.011	0.056	0.053	0.000	0.000	0.000	0.000	0.000
Litecoin	New authors	0.256	0.454	0.000	0.000	0.000	0.000	0.209	0.395
	Posts per day	0.031	0.016	0.000	0.000	0.000	0.000	0.380	0.415
	Subscriber growth	0.002	0.000	0.000	0.000	0.000	0.000	0.175	0.364
	Google trends	0.051	0.000	0.000	0.000	0.000	0.009	0.143	0.023
	Wikipedia views	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.027
	Twitter volume	0.002	0.000	0.000	0.000	0.000	0.000	0.050	0.001

It can be observed that in the short term (2-4 and 4-8 day period band) there is no consistency in results; in some cases the null hypothesis can be rejected and in some cases it cannot. In the medium term there is more consistency in rejection of the null hypothesis in favour of bubble regime coherence values significantly exceeding the non-bubble regime values. In the long term, the proportion of instances exhibiting statistical significance reduces, with the majority of cases in the 256-512 band not being a rejection of the null hypothesis. This reduction of statistically significant differences when considering longer-term periods further emphasises the point that it is the medium term in which coherence tends to strengthen during bubble regimes.

4.4 Results: Coherence between different cryptocurrencies

An interesting separate avenue to explore is the wavelet coherence between different cryptocurrencies, allowing any relationships between different cryptocurrencies to be detected and documented. Relationships between different cryptocurrencies would be of interest for those searching for diversification, which could be a factor in the trading strategies proposed later.

Figure 11 (a) shows many significant positive correlations between Bitcoin and Litecoin. This is an expected relationship given Litecoin is technically very similar to Bitcoin (Litecoin is essentially Bitcoin with faster block confirmations). Overall, there is no clear leader in the relationship. However, during the interval of the late 2013 price bubble (where Bitcoin and Litecoin reached around $1000 and $40 respectively), it can be seen that Bitcoin is leading Litecoin (slightly downward facing arrows across all periods).

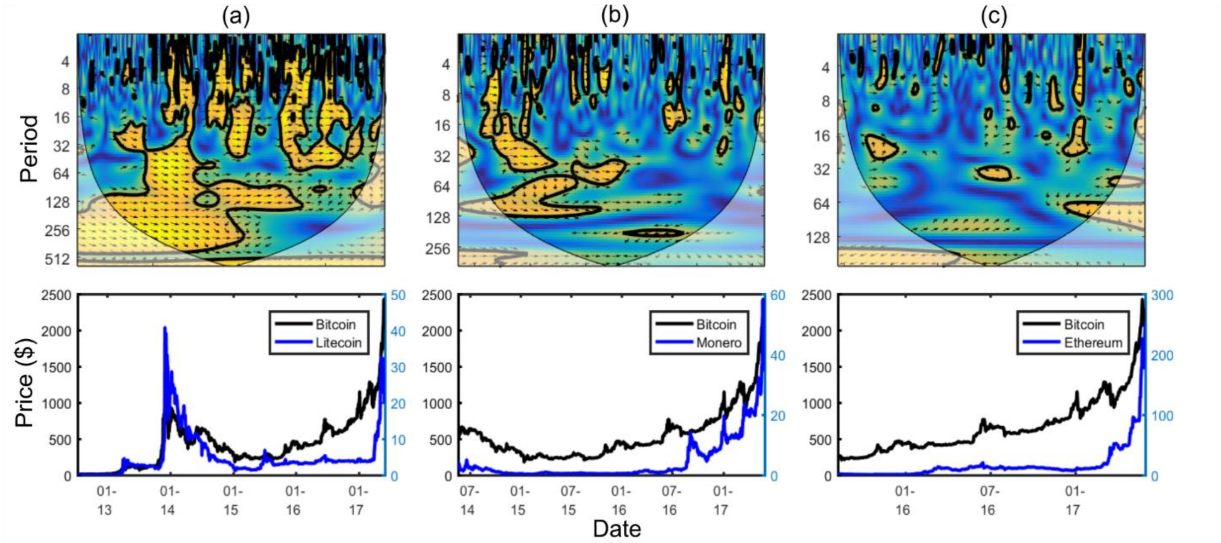

Figure 11. Wavelet coherence plots between (a) Bitcoin and Litecoin prices; (b) Bitcoin and Monero prices; (c) Bitcoin and Ethereum

The longer term relationship varies over time. After exhibiting strong positive correlation in 2013 (where prices rose) and 2014 (where prices fell for a sustained interval), the long term relationship between Bitcoin and Litecoin starts to break down around the middle of 2015. It can be seen in the accompanying price plot that at this point the Bitcoin price starts to recover gradually, whereas the Litecoin price does not.

Over the short and medium term, there are frequent intervals of positive correlation between Bitcoin and Litecoin. There is a limited interval, around March 2017, where a weakened (less significant) relationship exists in the short term (the top right of Figure 11 (a)), where a positive relationship had previously existed. This lack of positive relationship suggests the price movements decoupled. Two reasons could contribute to this decoupling. In early March 2017, the SEC gave its long-awaited (in practice negative) decision on a Bitcoin ETF, but it appears this had little impact on Litecoin—Litecoin was potentially even used as a hedge against the resulting Bitcoin price changes. In late March, percentage support for a Litecoin technical enhancement (SegWit) increased beyond the threshold percentage required for adoption around the same time as significant increases in the Litecoin price. The adoption of this change would temporarily reduce the similarity between Bitcoin's and Litecoin's technology (Bitcoin has since also adopted SegWit).

Figure 11 (b) and (c) show less consistent relationships with Bitcoin. Figure 11 (b) shows that Monero nearer its inception was significantly impacted by Bitcoin price changes (positive correlation with Bitcoin leading the price changes (seen towards the left of Figure 11 (b)), with co-movement over the short, medium, and long terms. In 2016, Monero had a number of positive developments which may have led to its price behaviour decoupling from Bitcoin's. For example, on August 22nd 2016, AlphaBay Market, a dark-net market, announced they would start accepting Monero-based transactions. Integration announcements from other dark-net markets also occurred around this time prompting mainstream media coverage. Furthermore, as Monero grows, a lack of long-term co-movement is understandable due to Monero having very different objectives to Bitcoin (unlike Litecoin and Bitcoin which have very similar objectives); Monero focusses primarily on the privacy of those transacting whereas Bitcoin does not.

There is a lack of longer-term relationship between the Bitcoin and Ethereum price (Figure 11 (c)). Although there are limited areas of co-movement, there is no clear pattern.

However, the short term exhibits brief intervals of co-movement. It is likely that events affecting the cryptocurrency environment as a whole will have similar (short-term) effects on all cryptocurrencies. One example in early January 2017 can be examined to demonstrate this. Following weeks of increasing Bitcoin prices (and high volatility), on January 6th 2017 the People’s Bank of China (PBOC) issued a statement expressing their concern regarding Bitcoin’s recent price volatility, and reminding cryptocurrency exchanges that they must operate within the laws and regulations of China. This caused cryptocurrency markets to speculate that a tightening of Chinese regulations was imminent—this was especially significant as at the time Chinese trading was reported to be around 95% of global Bitcoin trading volume. The price of many cryptocurrencies decreased during this period. This example highlights how individual events have a similar impact on a number of cryptocurrencies (and hence, short-term positive coherence). The resulting positive coherence can be seen on a short-term horizon for both Monero (Figure 11 (b)) and Ethereum (Figure 11 (c)) around early 2017 (the strips of yellow, touching the topmost border towards the top right of the scalograms).

Overall, it is found cryptocurrencies experience short term intervals of co-movement, sometimes caused by sector-wide news and cross-market contagion, though correlation is likely to depend on the nature of the causal event and market environment. Certain cryptocurrencies may be more linked in the medium and long term due to similar technical aspirations.

4.5 Discussion

In this chapter time-evolving relationships between a number of online indicators and associated cryptocurrency prices are evaluated, as well as relationships between cryptocurrencies. Short term relationships between online indicators and associated cryptocurrency prices appear sparsely and erratically. This sparsity of short term correlation and its erratic nature (either positive or negative correlation is displayed) mirrors the results seen in the preliminary data analysis (Section 3.4.3) whereby correlation direction (either positive or negative correlation) switches depending on time period considered (seen in Table 6) which results in a weakening on overall time independent correlation (seen in Table 5). This suggests other factors may be more predictive of cryptocurrency prices in the short term—possibly market-related events (e.g. unexpected shocks) and patterns stemming from technical analysis. Medium-term relationships appear and disappear over time, and the relationships strengthen during bubble regimes. Long term relationships between indicators and price appear more consistently positive. It is hypothesised here that this long term relationship may be due to another factor - technical progress. As a project makes technical progress, a community will form over time, increasing online activity and also demand, and hence price of the particular cryptocurrency.

Of the cryptocurrencies considered, it is observed that bubble regimes have the least effect on relationships between Bitcoin-related indicators and the Bitcoin price. There are several possible reasons for Bitcoin’s smaller coherence change in bubble regimes. Firstly, Bitcoin has always been the most well-known cryptocurrency, and so online activity that appears related to it may actually relate to cryptocurrencies in general (rather than being specific to Bitcoin), resulting in less of a relationship between this perceived activity and the Bitcoin price. Secondly, given that Bitcoin is the largest cryptocurrency, by market valuation and trading volume, online speculators may have less opportunity to influence the price with their buying or selling pressure. Thirdly, the Bitcoin subreddit considered in this work (/r/Bitcoin) is commonly used as a platform for the community to debate a variety of contentious scaling solutions that would enable the Bitcoin network to process more transactions concurrently. The amount of activity devoted to debating scalability would be unlikely to change dramatically in relation to price changes.

In the medium and longer term, the Reddit related indicators are more consistently leading than Google Trends and Wikipedia view indicators. In particular both Bitcoin and Litecoin show no clear leader between these factors and the price. In addition it can be noted that Wikipedia views lag the Monero price but lead the Ethereum price during the data period.

The work in this chapter also considers the volume of discussion on Twitter, due to its previous extensive use in the literature. Similar relationships to those that exist with other factors are found: in the short term, relationships between Twitter volume and associated cryptocurrency price are sparse and erratic; in the medium term, relationships strengthen during bubble regimes; in the long term, there exist more consistent positive relationship with price. Despite the aforementioned similarities with other indicator to price relationships, there are a number of differences when considering Twitter data. For example, Twitter has almost no relationship—predictive or otherwise—with the least mainstream cryptocurrency considered in this work (Monero). Furthermore, Twitter has been already shown several times to be useful in the prediction of price dynamics around the 2013/2014 Bitcoin bubble. The relationship between Twitter and Bitcoin is confirmed in the wavelet coherence work in this chapter; however it is observed that towards the end of the data period considered there are less significant relationships between Twitter and price than observed between Reddit derived metrics and the price.

The work presented in this chapter has introduced and confirmed Reddit as a valuable data source within cryptocurrency price prediction related work, as meaningful leading relationships have been identified. The Reddit derived factors will be used in the following chapters due to their leading nature identified here, their novelty and the relatively more frequent occurrence of relationships. It should also be noted that although the count-based metrics used in this chapter can be retrieved for a reasonable price, retrieval of raw historical Twitter submission data (as would be required in Chapter 6, if Twitter were to be used) is still out of the budget of most researchers (as described previously in Chapter 3).

The strengthening of medium term relationships between the considered online indicators and associated cryptocurrency price during bubble regimes is a new addition to the literature and provides an explanation for the appearance and disappearance of relationships seen in previous work [69]. The bubble regime dependent relationships identified here suggest that these indicators may provide effective predictive power in the prediction of cryptocurrency price bubbles, and the next chapter will consider this in further detail, progressing from the simple identification of factors seen here to using them to make predictions.

As well as using the Reddit derived indicators in the next chapter, the knowledge acquired here that these indicators have more bubble dependent dynamics in the medium and long term (rather than the short-term) will factor into both the choice of data granularity and also into the parameters of the model (to promote slightly longer holding periods).

Chapter 5

Predicting Cryptocurrency Price Bubbles using Social Media Data and Epidemic Modelling

The previous chapter identified relationships that exist between indicators derived from Reddit usage and the prices of associated cryptocurrencies. It was found that relationships strengthen during bubble regimes of the price series. Whereas the previous chapter focused on identification of relationships, this chapter focuses on using these relationships for prediction. In addition, whereas the previous chapter documented historical bubbles by considering price movements alone, this chapter detects bubbles by monitoring changing fundamentals unique to cryptocurrencies; these fundamentals are related to community size and interest around a particular cryptocurrency. To achieve this end, this chapter introduces a bubble-detecting model using the previously considered social media indicators as input. Given social media can be considered as online word-of-mouth [118], a model [119] is used which tracks the epidemic-like spread of an investment idea. To validate the predictions generated by the model, a trading strategy is built and tested on historical data.

5.1 Background and related work

5.1.1 Epidemic-based asset bubble modelling

Investment ideas have been shown to spread from a person to those they are connected with [70], on a street-by-street level [120] and on a city-wide level [121]. One approach to track the spread of ideas, including investment ideas, is to repurpose models originally designed by epidemiologists to predict the spread of disease epidemics [122].

One of the simplest epidemic models (named the SIR model[21]) has an infection rate (the rate the disease spreads from infected people to susceptible people) and a removal rate (the rate at which infected people either recover or are removed from the population). The spread of an investment idea can be tracked with a slight modification to this definition: the infection rate represents the communication of an idea, and the removal rate represents the loss of interest in the idea. For example, a modified SIR model tracked investors trading Finnish stocks between 1995 and 2003 [123]; communication between investors was identified as having a considerable influence on their investment choices. Another application of a model inspired by epidemiology was used to simulate housing price bubbles [124]. Particular attention was given to considering if all booms are followed by busts. Dynamics were replicated from the modelling of infectious diseases in the following way: a number of agents existed in the simulation with varying prior beliefs (in this case, the idea of an impending bust), and the ability to convert others to their belief. Other authors [125] examined an explicit example of a housing bubble, the pre-2007 US housing price bubble, to determine whether social contagion fueled the increase in prices. The authors found the extent to which an idea spreads depends on the relative influence of more sophisticated investors compared to less sophisticated investors.

Asset price bubbles are likely to be caused by a confluence of contributory factors, with the aforementioned social contagion of an investment idea being only one component. An event—one hailed as revolutionary or the bringer of a new era—may be required to encourage

[21] SIR stands for the three population categories that are part of the model. S = the susceptible population. I = the infected (or infectious) population. R = the recovered (or immune) population. At any one time, each member of the population is part of one of these categories.

initial enthusiasm [126]. One author, Shiller [122], describes further factors whereby news (the first factor) of price rises (another factor) increases investors' enthusiasm (another factor) which spreads to new investors (social contagion); this causes further price rises and thus increases enthusiasm. Although the social contagion is only one component, if monitored accurately it can be predictive of the scale and timing of the bubble [9].

Finally, Shiller [122] comments that a burst phase, though fitting with the metaphor of a bubble, is not essential to the formation of a bubble, and notes that history shows this burst phase does not always occur, or that if it does occur it can then be followed by a continued boom. This avenue of thought was also shared when modelling housing bubbles [124], as the model allowed for some booms not to be followed by busts.

5.1.2 Challenges when applying epidemic models to asset bubble modelling

The geographical proximity of investors is commonly used as a proxy to infer their level of communication (for example, in [123] and [126]). However, such geographical proxies are not always ideal [127], especially due to the increasing connectedness of the world enabled by technology. Use of such proxies is the result of the difficulty in obtaining data on word-of-mouth communication (researchers being unlikely to know what people talk about). Novel data sources have also been turned to in the pursuit of an accurate data source. For example, one study uses police records of an invite-only Ponzi scheme [128] to retrieve inviter-invitee relationships. These inviter-invitee relationships allow the authors to monitor how the investment idea spread between participants. Although insightful, the data source is not transferable to investigation into the spread of ideas related to other (non-Ponzi) investment ideas. One study [129] removes the need for acquiring accurate real data by examining agent-based models on simulated data.

Rather than relying on proxies, police records or agent-based models to provide indirect information on the interaction between people, the advent of social media allows that communication to be viewed directly; it is fortuitous that the means by which communities are now more often manifested is one that leaves easy traces to follow. As much of the data is public, analysing social media has provided a new way to track the epidemic-like spread of an idea using real data. For example, one study used epidemic models to track the diffusion of time-sensitive events (e.g. breaking news items) among Twitter users [130].

The accuracy of epidemic models when modelling idea spread is limited due to word-of-mouth communication having high mutation rates [122] (people getting the details wrong). It has been noted that to have a low mutation rate an idea needs to have a good story associated with it. Two metrics have been designed to evaluate whether ideas have this story-potential to spread and result in bubbles. Firstly, the "Chinese Whispers" theorem, which identifies necessary conditions for diffusion of an idea, and secondly, the "Made to Stick" theorem, which judges the likelihood that an individual will incorporate the idea into their worldview [131]. In a separate area of the literature, it is observed that trading strategies with extremely high returns naturally offer those undertaking them more opportunity to broadcast them (meeting the conditions of the "Chinese Whispers" theorem) [127] —and cryptocurrencies provide an opportunity for these extremely high returns. In addition, if tracking word-of-mouth through online social media communication, the mutation rate is expected to be lower due to the ability to copy and paste and to provide a link to a website that all participants can inspect [122] (meeting the conditions of the "Made to Stick" theorem).

5.1.3 An alternative epidemic modelling technique

Alongside the well-known SIR model, a number of alternative epidemic modelling techniques have risen in popularity, including spatial models, regression techniques, agent-based simulations and network models (a comprehensive review of current epidemic detection methodology can be found elsewhere, for example [132]). One subset of these uses hidden Markov models (HMM). HMM-based models have the advantage that pre-defined epidemic state characteristics do not have to be specified by a domain expert, and the model learns state classifications from the data provided [119]. One such HMM was applied to differenced infection rates to classify influenza time series data into epidemic and non-epidemic states [119]. Differenced incidence rates were used to detrend the data and thus allow autoregressive modelling within the HMM[22] (explicit details are provided later within Section 5.2). The model worked well in an online environment (when it received new data points one by one[23]) and, at each timestamp, produced a probability of the system being in the epidemic state. This model has since been applied successfully to Twitter submissions to provide a novel approach to categorizing submissions as 'trending' and 'non-trending' [133].

5.1.4 Previous similar work in the cryptocurrency prediction domain

Although no known prior research has considered how epidemic models can be used to model cryptocurrency market dynamics, one paper does approach cryptocurrency markets from an ecological perspective and is mentioned here due to the relative closeness (and real-world natural application) of ecology and epidemic modelling. This work [134] repurposes a well-established model originally designed to model genetic evolution and applies the model to competition between cryptocurrencies. In the original context, the model contains individuals who are part of one of many species. In the repurposed context, individuals represent a certain amount of USD and species represent the cryptocurrencies that can be invested in; an individual being part of a species thus translates to a certain amount of USD being invested in a particular cryptocurrency. Over a number of iterations (called generations in the original model), individuals are randomly assigned to either existing species or to a new species added that iteration. The growth and decline of species in the repurposed model represents investment moving between existing cryptocurrencies and also (the 'new species' case) to new cryptocurrencies. The authors found that the model produced a number of similar characteristics to those observed in the cryptocurrency market, relating to the fluctuations of rankings of competing cryptocurrencies. These results encourage the further application of exogenous models (modelling derived from situations outside of trading markets) to the cryptocurrency specific domain.

[22] Separately, and relevant to the purpose here, auto-regression has been demonstrated as a good way to model an asset price bubble [167]

[23] Parallels can be drawn between the arrival of influenza data one data point at a time and the arrival of trading data.

The work in this chapter will investigate whether it is possible to examine patterns in social media use to detect cryptocurrency price bubbles using a technique designed for epidemic modelling. Specifically, an epidemic-detecting hidden Markov model [119] will be used, due to its ability to model auto-regressive situations [119] and its previous success identifying particular tweets as trending [133]. This work aims to contribute not only to the cryptocurrency related literature, but also to the speculative asset bubble literature by demonstrating that social media can be used as an input to epidemic-based bubble models.

5.2 Methodology

As stated above, an HMM is chosen to be used to detect epidemic and non-epidemic states of social media usage and trading volume, due to its successful use in influenza epidemic outbreak prediction [119] where bubble-like behaviour is seen in relation to the number of individuals infected. An HMM has a number of underlying hidden states, which are transitioned between. Each state has associated possible observations. Given an observed series of data, an HMM can be used to identify the most likely hidden state the model is in at each data point. The particular configuration of the HMM used in this case is outlined in the following subsections.

5.2.1 Number of hidden states

The model uses two hidden states, epidemic and non-epidemic, which are unobserved. E_t is an unobserved random binary variable to denote whether the system is in the epidemic state (1) or not (0) at time t.

5.2.2 Observation probability distribution

The hidden states ($E_t = 1$ and $E_t = 0$) have associated emission probabilities. Emission probabilities give the likelihood of seeing particular output values and can be sampled from different distributions depending on which state the system is in. The actual output values are retrieved from differenced time series data (for example, for one of the social media indicators) where I_t represents the difference between the time series values at time t and t-1. The model definition specifies that the conditional probability distribution of I_t is sampled from either an autoregressive process of order 1 (AR(1)) for the epidemic state ((1a) and (1b) below) or sampled from a Gaussian white noise distribution for the non-epidemic state ((2a) and (2b) below). Essentially, the epidemic state has interrelated changes, whereas the non-epidemic state has small random changes. Hence the emissions conditional distribution is defined as

$$I_1 \mid (E_1 = 1) \sim N(0, \sigma_1^2), \qquad (1a)$$

$$I_t \mid (E_t = 1) \sim N(\rho * \; I_{t-1}, \sigma_1^2), \qquad (1b)$$

$$I_1 \mid (E_1 = 0) \sim N(0, \sigma_0^2), \qquad (2a)$$

$$I_t \mid (E_t = 0) \sim N(0, \sigma_0^2). \qquad (2b)$$

5.2.3 Transition probabilities

The HMM transitions between hidden states according to a transition probability matrix, which gives the probabilities of transitioning from one hidden state to another. Transitions exhibit the Markov property, whereby the transition probability depends only on the current state, and the state history is forgotten. $P_{k,l}$ denotes the probability of transitioning to state l at time $t+1$ given that the current state is k at time t, i.e.

$$P_{k,l} = P(E_{t+1} = l \mid E_t = k).$$

5.2.4 Initial state distribution (parameter priors)

Prior parameter distributions are specified based on an understanding of the context. The prior parameter definitions below ensure that σ_1 (the sigma (standard deviation) associated with an epidemic state) has a higher prior value than σ_0 (the sigma associated with a non-epidemic state). Uniform distributions are chosen for the standard deviations, as suggested by Gelman [135]:

$$\theta_{low} \sim Unif(a, b),$$

$$\theta_{mid1} \sim Unif(\theta_{low}, b),$$

$$\theta_{mid2} \sim Unif(\theta_{mid1}, b),$$

$$\theta_{high} \sim Unif(\theta_{mid2}, b),$$

$$\sigma_0 \sim Unif(\theta_{low}, \theta_{mid1}),$$

$$\sigma_1 \sim Unif\left(\theta_{mid2}, \theta_{high}\right).$$

In the above a and b are hyper-parameters. The prior value for b is set as the maximum difference between two successive time series points. The prior value for a (set to be 1/10 of b) represents a lower bound on the non-epidemic state standard deviation. The a parameter is used to attempt to ensure that the standard deviation does not converge to 0. The prior values for the remaining parameters are defined as in [119]:

$$\rho \sim Unif(-1,1),$$

$$P_{0,0} \sim Beta(0.5,0.5),$$

$$P_{1,1} \sim Beta(0.5,0.5).$$

Using $Beta(0.5,0.5)$—which looks like a ‘U’ shape on a graph—as a prior parameter for the probability of remaining in a state encourages a switch of state almost instantly (advantageous in the case of a wrong signal) or a longer persistence in a particular state. In the context of the present work this longer persistence in a particular state will result in a longer persistence of any trading positions (minimising trading fees).

Once the model and priors have been established, estimates of the optimal parameter values can be found using expectation maximisation (EM), a commonly used process whose effectiveness has been demonstrated by its use in multiple applications. EM is an iterative process to find maximum likelihood parameters given an observed set of data, the parameters converged upon being those that provide the best fit with the observed data. The observed data utilised in the EM process depends on the position of the moving window (Section 5.2.6 discusses how the data is partitioned into multiple moving windows, and how state probabilities are retrieved from the model).

The subsections above have considered the design of this particular HMM; the subsections below move on to consider how the HMM is applied as part of this experiment.

5.2.5 Data used

The previous chapter discovered bubble-dependent relationships between Reddit-based metrics and associated cryptocurrency prices (e.g. that relationships strengthen in bubble regimes). As such, these metrics (*new authors*, *subscriber growth* and *posts per day)* will be used as inputs to the HMM. Given that the previously discovered relationships strengthen in the medium and long term, daily data is selected here rather than a higher granularity (e.g. intraday data). It is true that intraday data (such as 1 minute or 5 minute candlestick data) is commonly used for trading strategies. However the previous chapter showed that in the short term the Reddit-based metrics do not have a consistent or clearly discernible relationship with price changes, justifying the choice of daily data. The experimental period chosen here is April 2015 to September 2016, with the exception of Ethereum. Ethereum was first listed on trading exchanges on August 8th 2015, and so the Ethereum time series will start from August 2015.

As well as the three indicators derived from social media (specifically, Reddit), trading volume is added as a fourth input to the HMM, as a confirmatory signal; while most discussion on cryptocurrency subreddits pertains to, and may result in, further price movements, there are occasionally cases where a large amount of discussion is associated with some unrelated topic. Volume is therefore important to confirm that social media activity relates to market activity[24].

[24] A model without volume input was constructed, and in fact the resulting trading strategy proved somewhat more profitable than that described. The profit-depressing effect of volume is because volume tends to lag social media usage in moving to an epidemic state, and therefore positions are entered later. However, the risk benefit of the additional volume input was considered to outweigh this lessened profitability.

As discussed in Section 5.1.3 the HMM in [119] added to the previous epidemic-detecting HMM literature by using differenced data rather than unmodified influenza data, which enabled use of an AR(1) process on the de-trended data. All time series here will be differenced, to the same effect, with the positive by-product that immediate changes can be recognised quicker than if absolute values were used.

5.2.6 Moving window and state probability

Data is grouped into windows of length 100 data points. Based on preliminary examination of the cryptocurrency price series this length is sufficiently long to encompass typical bubble and non-bubble regimes but short enough to be computationally viable. These windows are moved forward in time according to the mechanism used in [136] (new pieces of data being added to the window as the oldest data is removed). This moving window approach means that the model is always considering the most recent data as it becomes available, as visually outlined in Figure 12.

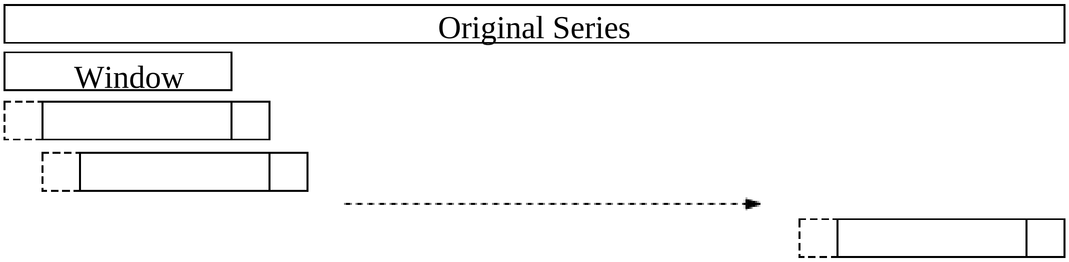

Figure 12. Dynamic moving window

In each window location, emission and transmission parameters are estimated[25]. The HMM can output the probability of being in the epidemic state at each time series point in the window. It is important to note that the final point in each moving window can be regarded as the current point (as no future data can be seen) and the probability of being in the epidemic state at only this point is retrieved (thus not allowing for state categorisations to use future data). A similar technique has been used in [137] to train an HMM on a moving window of data (while avoiding look-ahead bias), and also in the context of financial time series prediction [138].

[25] Each new window location has a new initialisation of parameters and a new fitting process via expectation maximisation. The current system does not hand over parameters from one window location to another.

In this work, as the window moves forward, a sequence of epidemic state probabilities are generated. The probability of being in the epidemic state can then be used within a trading strategy, as will be described in Section 5.2.9.

5.2.7 Preventing local maxima in the parameter fitting process

The expectation maximization fitting process is susceptible to converging to local maxima. To overcome this, for each window of data 20 repeated parameter fittings are completed, and the fit that achieves the highest likelihood is then chosen as the final model. This multiple trial approach was used by Chan [139] while attempting to use HMMs to detect different regimes within trading markets.

5.2.8 Use of distributed computation

The parameter fitting process, combined with the moving window approach described above (where in this case windows containing near-identical data points need to be analysed), takes considerable time to converge on a solution. Researchers facing similar situations have parallelized the computation and executed the computation on cloud-based high-performance infrastructure [140][26]. One such tool to provide easy integration with cloud parallelization is provided by Techila Technologies and enables connectivity with Amazon Web Services and Google Cloud. This functionality was used in the current work to run the parameter fitting process on multiple moving windows in parallel on a distributed grid of processors.

Code originally intended to be run locally needed to be modified to work with the Techila environment. The environment consists of three main components that are briefly covered here. The *Techila SDK* is installed locally (on the machine issuing tasks) and enables applications to communicate with the Techila distributed environment. The *Techila Server* runs on a virtual machine, receiving tasks and distributing them to the *Techila Workers*. The Techila Workers

[26] Other solutions (including modification of the parameter fitting process) are possible and would have been explored if parallelisation had not produced viable execution time.

provide the computing power. Projects are split up into jobs, and each Techila Worker executes a separate job in parallel. Figure 13 shows a demonstration of the parallelisation achieved.

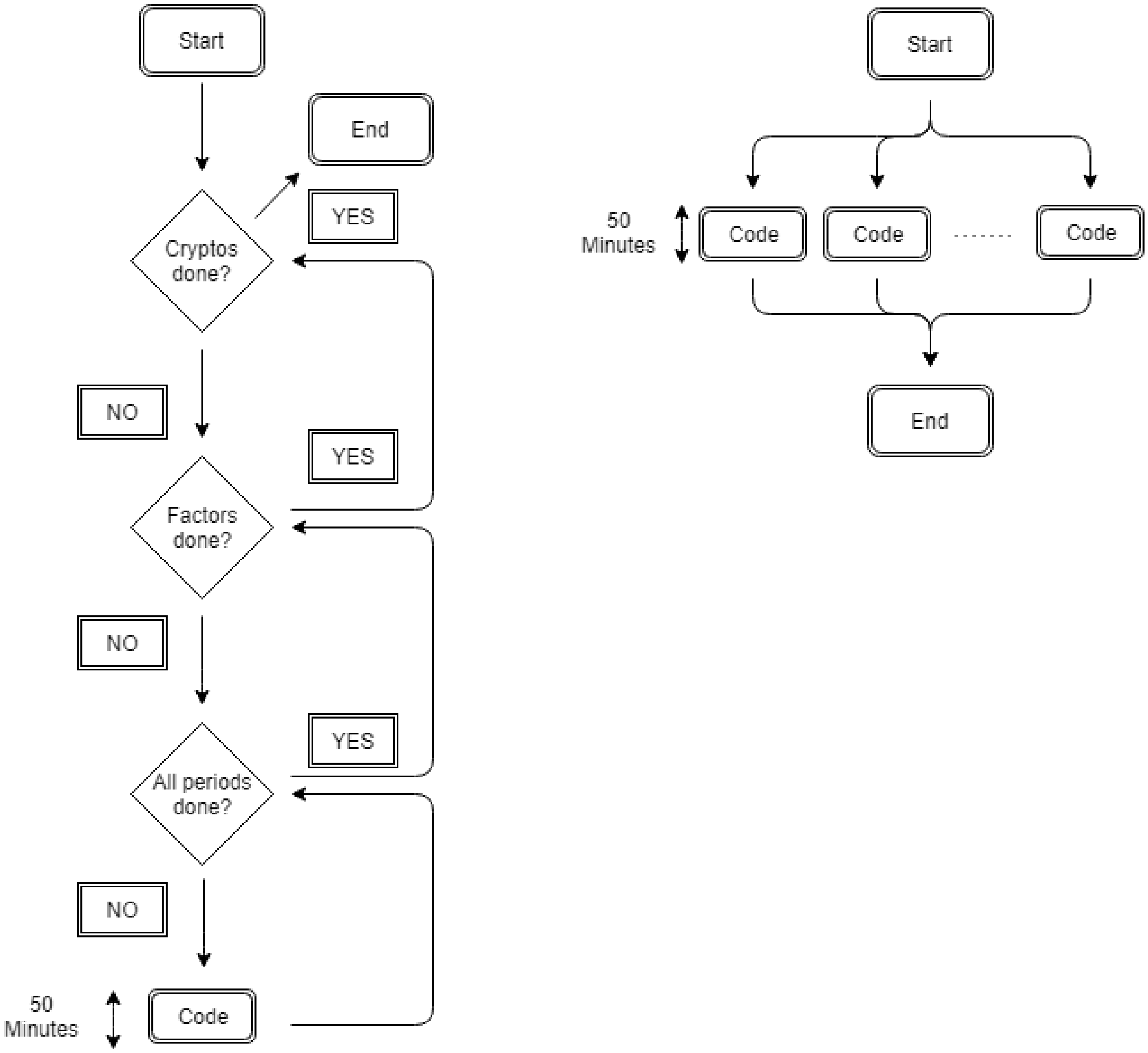

Figure 13. Execution paths for local vs parallelised for-loop

The non-parallelised for-loop (left-hand side of Figure 13) demonstrates the execution path when no parallelisation is present. The section marked ‘code’ (at the bottom left-hand side of the diagram) runs the parameter fitting process; this takes approximately 50 minutes running on a single CPU. In non-parallelised code, this is inside three nested for-loops (one for-loop for the cryptocurrencies, one for-loop for the factors and one for-loop for the moving window intervals). For a basket of 4 cryptocurrencies, with 4 indicator factors and 350 time periods, there are 5,600 sequential executions of the ‘code’ block. The parallelised for-loop (right-hand side of Figure 13) demonstrates the execution path when parallelisation is present, the code being run simultaneously on a number of Techila workers. Executing a single ‘code’ block still takes approximately 50 minutes however multiple executions happen at once, in parallel. Each job is supplied with the appropriate input parameters as though it is a separate iteration of each for-loop. This enables an overall job execution time to be reduced to the aforementioned 50 minutes (with some setup time), assuming for this example that the number of workers is the same as the number of jobs.

Table 8 shows the speedup in execution time achieved (on average for a particular cryptocurrency/factor pair). 64 workers were used and the number of parameter fittings (individual jobs) to run exceeded the available 64 workers. This meant that not all parameter fittings could be run at the same time and individual workers were allocated their next job after completing a particular job. The ‘single core execution time’ row shows the job execution time if executed on a single CPU, a good approximation for the execution time if run on a local machine. Parallelisation results in the parameter fitting being completed approximately 55.1 times quicker than if using single CPU execution, a now-viable amount of time.

Table 8. Execution time for local vs parallelised execution

Metric	Result
Single core execution time	13 days 17 hours 50 minutes
Parallelised time	0 days 5 hours 58 minutes
Acceleration factor	55.1x

5.2.9 Trading strategy

Assessing the directional accuracy of predictions may be insufficient to assess their value. For example the correctly predicted movements might frequently be small ones, possibly so small that trading costs would erode their profitability, while incorrect predictions could at the same time lead to large losses. A more persuasive way to validate the predictive power of the system is to convert the predictions of the HMM (that the system is in a state classed as either epidemic or non-epidemic) into a realistic trading strategy [6], and assess its performance. In this strategy entry into an epidemic state is considered a buy signal, and exit from the epidemic state is considered a sell signal (to close the position and no longer hold the asset).

Figure 14 shows a simulated example of the strategy. Figure 14 (upper) shows the probability of the HMM being in the epidemic state at each time point. As the probability goes above 0.5 at time T1, the HMM is more likely to be in an epidemic state than not, which is considered an entry point for the trading strategy, and conversely, as the probability drops below 0.5 at time T2, the HMM is now more likely to be in the non-epidemic state, which is considered to be an exit signal. In Figure 14 (lower), a corresponding simulated asset price time series is shown with the recommended entry and exit locations.

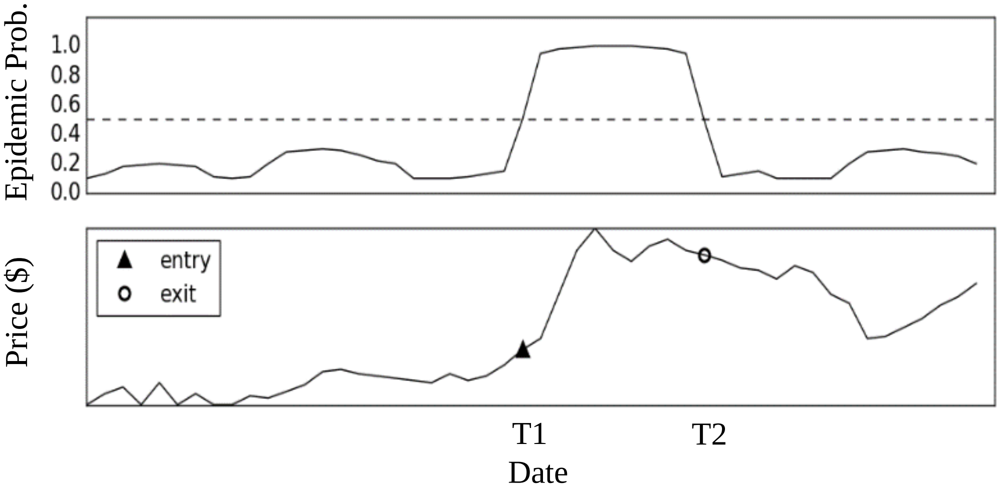

Figure 14. Illustrative trading strategy entry and exit points (lower) based on probability of epidemic state (upper).

A separate HMM exists for each input factor, so several epidemic probabilities are produced at each timestamp. These can be combined into an overall prediction using either of the following mechanisms (both are considered in the results section to follow):

1. *Unanimous voting*. A system where each HMM votes 'epidemic' or 'not epidemic'. The overall system needs to achieve consensus before an epidemic state is signalled.
2. *Averaging*. The individual probabilities are averaged. If the aggregated probability is above 0.5, the overall system is considered in an epidemic state.

Trading strategies can hold multiple assets at the same time. In this work, funds are allocated as follows: if one cryptocurrency is being signalled as epidemic, all the funds are allocated to this cryptocurrency; however if multiple cryptocurrencies are simultaneously signalled as epidemic the funds are split between the cryptocurrencies equally. The weightings are updated in the context of other positions. For example, when one cryptocurrency is no longer signalled as being in the epidemic state, the position in that cryptocurrency is closed, and the funds are reallocated to other open positions, purchasing additional units of these currencies.

A back-testing framework was also built. Back-tests simulate a trading strategy on historical data to determine its performance. The objective of a back-test is to produce results as close as possible to the results that would have been achieved if the strategy had been trading real money over the tested period. As such, standard cryptocurrency exchange transaction fees have been included (chosen as 0.2% per transaction) in the simulations.

5.2.10 Benchmark strategy

To help assess the performance of the above trading strategy it is useful to define a benchmark strategy to which it can be compared. An equally weighted buy and hold strategy is used for this purpose. In this strategy a total notional amount (here, $1,000) is divided by the number of assets being considered, with the assets being bought on the first day of the back-testing period and held until the last. Buy and hold is generally regarded as a difficult benchmark to beat; this is especially true in the case of cryptocurrencies, for which prices have notoriously soared over short time horizons.

5.3 Results

This section first examines the HMM state probabilities and then validates the utility of these outputs by use of the multi-asset trading strategy defined above. Finally, the HMM parameters converged upon are analysed to better understand the reasons for the system's success.

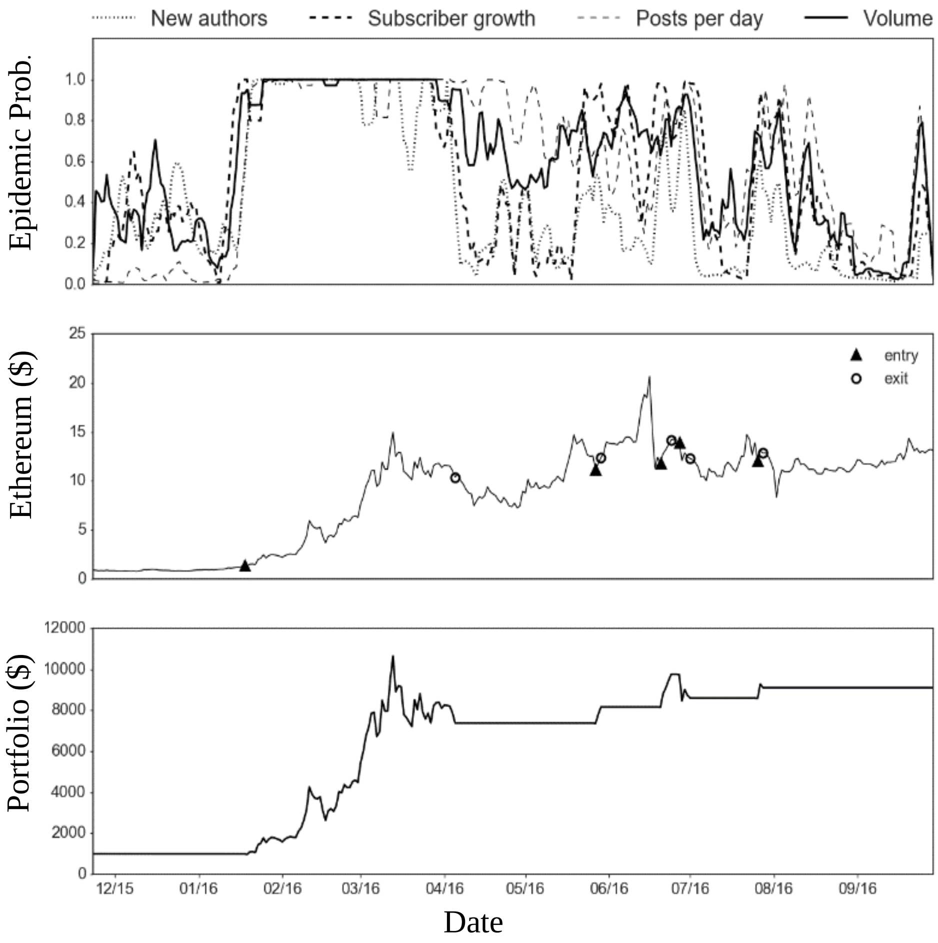

Figure 15. Epidemic probabilities and trading strategy for Ethereum

As an example Figure 15 (top) displays the probability of being in the epidemic state for the indicators relating to Ethereum, while Figure 15 (middle) shows how the price series for Ethereum evolves during the same time period. The best example of a bubble seen in the data period used in this chapter occurs for the Ethereum price between January and April 2016 (the price rising from around $1.50 to around $11); during this period all four indicator probabilities are consistently signalling the epidemic state, as would be expected.

The entry and exit markers are placed in Figure 15 (middle) to indicate when the overall system moves into (buy signal) and out of (exit signal) the epidemic state (using the unanimous voting methodology detailed in Section 5.2.9). The profitability of these signals will be considered later in the trading strategy commentary.

Occasionally, false positives are seen (the system identifying epidemic related usage while the price series is not going through a bubble regime). This is due to sharp downward price movements causing brief epidemic-like social media and trading volume activity. One example of this can be seen in June 2016, where the Ethereum price dropped from above $20 to just above $10. This resulted from the hack of an application built on Ethereum called the DAO [141], causing panic, which was reflected in the price, and on social media for a number of weeks afterwards. The system has, in fact, profited from these periods due to buying at a low price—but they still represent false positives. To overcome this, the model could also use price related conditions (e.g. a requirement that price is rising) to avoid signalling an epidemic after crashes. Further discussion of this situation and how to overcome it is provided in Section 5.4.

The key advantage of the system presented here is that trading positions are only entered for short periods when price rises are expected. For 222 days out of the 313 days in the tested period the strategy does not have a position (as shown by the entry and exit points in Figure 15 (middle) and the portfolio value in Figure 15 (bottom)). A multi-asset strategy, as used below, can at such a time allocate more funds to buying other cryptocurrencies signalled as being in their epidemic state. The next section examines the profitability of this.

5.3.1 Trading strategy: Performance

This section shows the results of the multi-asset trading strategy described in Section 5.2.9, which aims to allocate money to buy whichever cryptocurrencies are signalled as being in an epidemic state (using unanimous voting). Figure 16 shows the price series for each cryptocurrency considered here, with their entry and exit points.

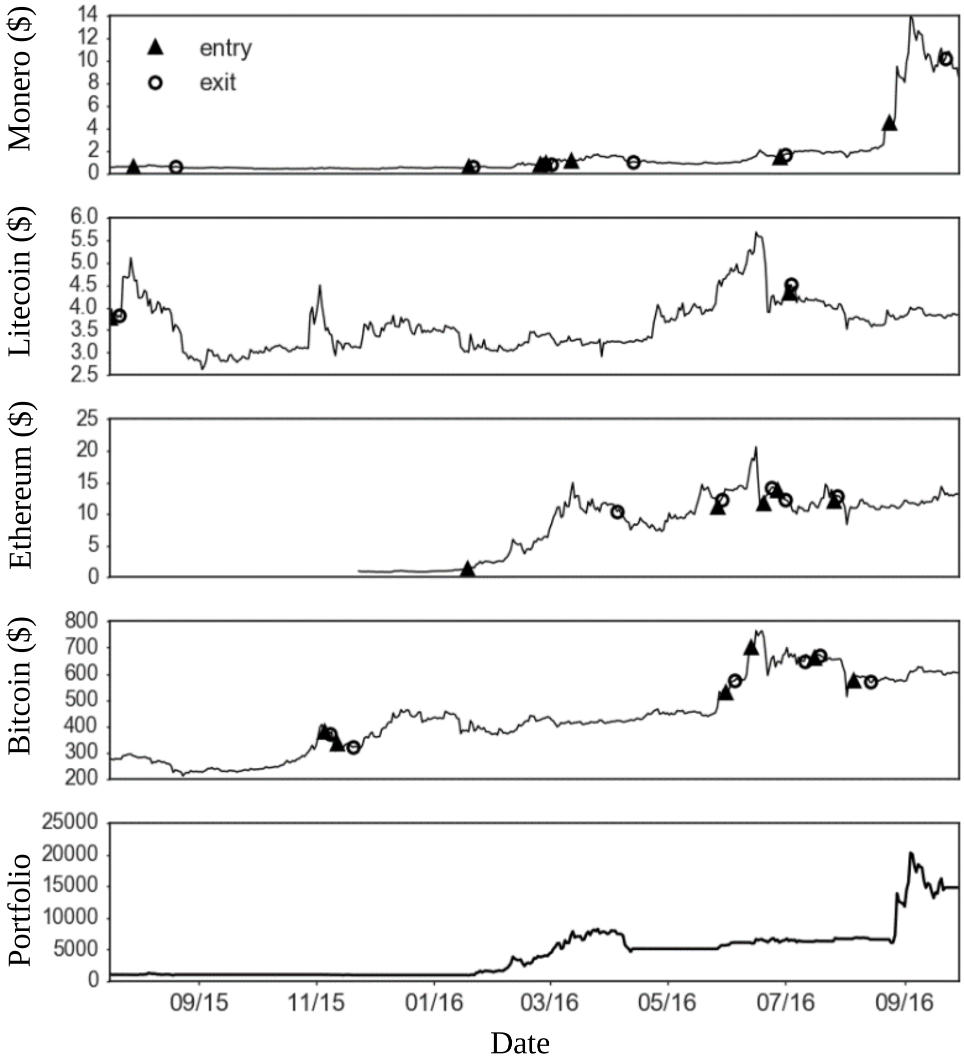

Figure 16. Price series, and entry and exit points for each cryptocurrency, and overall portfolio value (last).

As can be seen in Figure 16 (top), Monero undergoes a sudden and substantial price rise near September 2016 as the price rises from around $2.20 to around $12; this is the second best example of a bubble in this particular data period, after the Ethereum bubble already mentioned (January-April 2016). The strategy clearly profits from the Monero bubble. However, Litecoin has no comparably sized price rises to Monero or Ethereum, while Bitcoin shows a general sustained price rise throughout the testing period but also some periods of increased growth from which the system can profit. Figure 16 (bottom) shows the portfolio value over the tested period. The portfolio starts with $1000 and can be seen to make most profit in the time period around the Ethereum bubble (January-April 2016) and the Monero bubble (near September 2016).

Table 9 shows the current unanimous voting HMM strategy evaluated against common trading strategy metrics. It can be seen that the strategy outperforms the buy and hold strategy on all metrics including the Sharpe and Sortino ratios. It also has a smaller (better) percentage drawdown occurring over a shorter (better) duration. The final column in Table 9 shows the profitability of a variant version of the HMM strategy which will be discussed in the next section.

Table 9. Performance comparison of trading strategies

Metric	Buy and hold	HMM Strategy (unanimous voting)	HMM Strategy (averaging)
Ending portfolio (starting $1000)	$7,939	$14,804	$8,751
Returns	693.9%	1380.4%	775.1%
Sharpe ratio	1.77	1.93	1.48
Sortino ratio	2.63	2.64	2.29
Position number	4	20	33
Maximum drawdown (%)	50.03%	35.19%	32.68%
Maximum drawdown (duration)	47 days	12 days	12 days

5.3.2 Trading strategy: voting vs averaged probabilities

An alternative to the unanimous voting used up to this point is to take the average of the factor-specific epidemic probabilities instead. Figure 17 (top) visualises the resulting epidemic probability of the system in the case of Ethereum, and Figure 17 (bottom) displays the trades made by the averaging method (using the default epidemic threshold of 0.5).

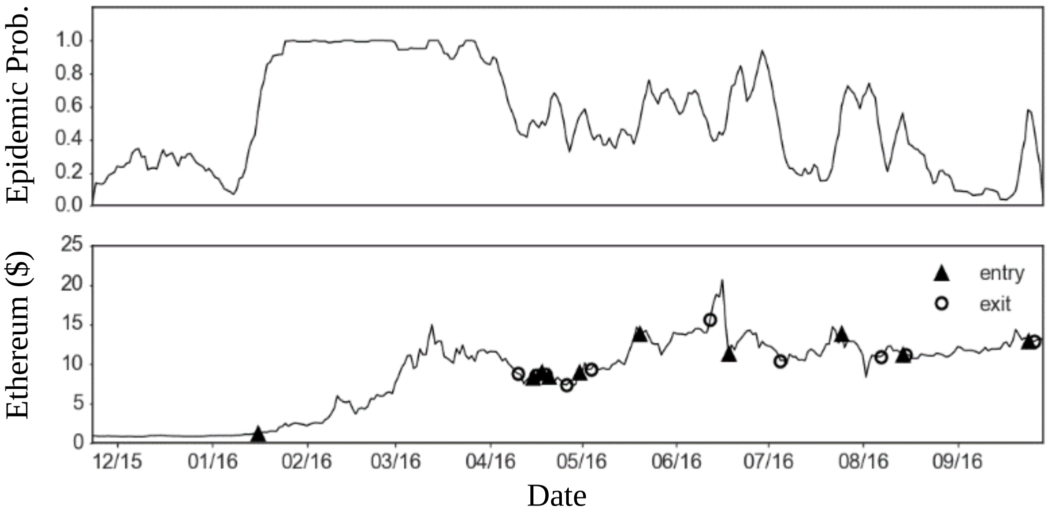

Figure 17. Price series, and entry and exit points (lower) for Ethereum based on aggregate probability of epidemic state (upper).

During May 2016 the epidemic probability remains around 0.5, causing the averaging variant to repeatedly enter and exit positions, which is not ideal given the transaction fees associated with trading. Averaging (with the current epidemic/non-epidemic threshold of 0.5) takes 33 positions for the multi-asset strategy, instead of 20 for the voting method, as shown in Table 9. Although the averaging method still outperforms buy and hold, the system's profitability is reduced greatly compared to unanimous voting.

Investigation into the impact of changing the epidemic/non-epidemic threshold suggests an increase in the threshold decreases the number of trades while increasing the overall profitability of those trades, thus improving overall strategy performance. Further work could explore setting a threshold via an optimization process examining the profitability of different threshold values on historical data (similar to [142] and [143] which investigate the robustness of training trading systems that use committee based voting mechanisms similar to that proposed here).

5.3.3 Parameters converged upon

Before state probabilities can be retrieved, the HMM is trained on previous data observations in order to provide estimated values for a number of parameters (as outlined earlier in Section 5.2). Table 10 shows the values converged upon for one example cryptocurrency/factor combination:

the *new authors* time series on the Ethereum subreddit (similar characteristics are found for the above parameters when examining other cryptocurrency/factor combinations). These values have for simplicity been averaged over those generated from all training window periods.

Table 10. Posterior mean of parameters for Ethereum/New Authors

Parameter	Posterior mean
ρ	0.80
σ_0	0.82
σ_1	3.20
$P_{0,0}$	0.86
$P_{1,1}$	0.72

The positive value of ρ shows the positive impact each data value has on the next, once in the epidemic state, and justifies the use of an autoregressive process. The values of $P_{0,0}$ and $P_{1,1}$ suggest that once the HMM is in a particular state it is likely to remain in that state at the next data observation; this is expected as it is likely that epidemic states will continue for a number of data points. It is also advantageous for the trading strategy, as when receiving persistent signals the strategy does not change positions too frequently. The model is slightly more likely to exit the epidemic state ($P_{1,0} = 0.28$) [27] than it is to exit the non-epidemic state ($P_{0,1} = 0.14$), reflecting the fact that epidemic states are expected to be shorter-lived than non-epidemic states. The sigma associated with the epidemic state, σ_1, is larger than the sigma associated with the non-epidemic state, as intended by the prior parameter choices.

The above values in Table 10 were time-averaged; however it should be noted that variation in parameter values can occur depending on the data period used for training. During the large epidemic period seen in the Ethereum price (between January and April 2016—Figure 15), the probability of remaining in the epidemic state $P_{1,1}$ approaches 1. Such variability demonstrates the advantage of using a moving window trained on the most recent data.

[27] $P_{1,0}$ represents the probability of being in the epidemic state (1) and moving to the non-epidemic state (0).

5.4 Discussion

This chapter has introduced an epidemic bubble-detection mechanism using social media data as inputs. In doing so, the work has demonstrated how epidemic detection techniques can be applied to social media data to predict cryptocurrency price bubbles. To achieve this, an HMM methodology originally designed to detect influenza outbreaks is applied to community-based social media indicators and trading volume, relating to certain cryptocurrencies, categorizing usage into epidemic and non-epidemic states. The utility of state probabilities were validated by transforming them into a profitable trading strategy that outperformed a comparable benchmark. It is notable that no price-related trading signals were used in the trading strategy; only social media usage and trading volume were considered.

Regarding the parameters converged upon by the model, it is of interest that the $P_{1,1}$ and $P_{0,0}$ parameters (the parameters giving the probability of remaining in the same state) converged upon are values that promote remaining in a particular state for a number of time periods (including $P_{1,1}$ reaching 1 for certain bubble-like intervals of the data period). Given the knowledge obtained in the previous chapter that social media activity is more consistently correlated in the medium and longer term with cryptocurrency prices, it is advantageous that there is a level of persistence of states (rather than continuously switching states).

Although the HMM-based prediction model is validated as profitable and outperforms a comparable benchmark strategy, false positives are occasionally generated (bubble regimes being identified where there are none). As noted earlier, in Section 5.3, one mechanism to overcome this would be to add a condition to the trading model that the price should be rising before a position is entered. However, the root cause of these false positive signals is that the indicators being used are count-based (as opposed to content-based) meaning all usage is currently detected without separation into positive or negative usage.

Analysis of submission content would allow topic-based signals to be extracted from social media submissions, allowing for further understanding of what kind of usage is occurring. Such content-based indicators may provide different insights into future cryptocurrency price movements to the current count-based indicators. The relationship between content of social media submissions and associated cryptocurrency prices is investigated in the next chapter.

Chapter 6

Mutual-Excitation of Cryptocurrency Market Returns and Social Media Topics

The previous two chapters have considered quantitative count-based metrics (*posts per day*, *new authors* and *subscriber growth*) to monitor user interest relating to a particular cryptocurrency. As discussed at the end of the previous chapter these count-based factors do not discriminate between different types of usage (i.e. all usage is displayed as a single count number, for example posts per day, without regard for the content of those posts) enabling the generation of false positives. It was also shown in Chapter 4 that the count-based metrics had poor predictive power in the short-term, potentially due to the inability to decipher the root cause of short-term spikes in usage.

The current literature has not explored the role of specific topics of discussion within cryptocurrency subreddits. It is hypothesised here that these topics of discussion may provide additional predictive power in relation to upcoming price movements. Using content-based metrics allows the investigation of relationships occurring over a shorter duration—whether price movement affects topic discussion, and whether topic discussion affects price movement. The following section provides background first on how topics can be retrieved from social media content, and then as to how relationships between these topics and price movements can be deciphered. The work in this chapter focuses first on the predictive power of discussion topics in the short term and then, for completeness, more general longer-term relationships between topic occurrence and cryptocurrency market prices are also considered.

6.1 Background and related work

6.1.1 Topic modelling

Although applied commonly elsewhere (for example, [144]), topic modelling techniques have only recently been applied to Bitcoin-related discussion, in this case sourced from a forum dedicated to Bitcoin (bitcointalk.org) [77] [145]. In the case of [145], the application of dynamic topic modelling (explained later in Section 6.2.1) allowed the evolution of topics, and terms within those topics, to be tracked over time. Results showed how discussion relevant to certain related technologies has changed over time. For example in one of the discovered topics manually labelled by the authors as relating to '*Bitcoin mining'* (the technical process by which blockchain transactions are validated), the term *CPU* was common earlier in the dataset, whereas terms relating to superior technology such as *GPU* increased in popularity over time. The same data source (*bitcointalk.org*) has been used in other topic modelling work [77]. Granger causality was applied to discovered topics to investigate whether there were relationships present between the occurrence of particular topics and statistics relating to Bitcoin; it was found, for example, that the topic related to *China* had a significant Granger causality with the Bitcoin price.

6.1.2 Hawkes model

Hawkes processes [146] model situations where the occurrence of an event increases the probability of subsequent events. Since their introduction they have been applied to model a range of event-based situations, including, in early work, the occurrence of earthquakes [147]. In more recent work, the application of Hawkes models within the separate fields of finance and social media has become popular. Within finance, Hawkes models have been used to provide an understanding of a variety of dynamics, for example the occurrence of financial contagion between different markets [148]. Recently, a Hawkes model has been applied to stock market returns and news article sentiment [149], in the first known application of a Hawkes model to the joint modelling of financial markets and news. Interactions between four types of events were considered: positive and negative market return events, and positive and negative news sentiment events. The methodology allowed for several findings, including a linkage of positive (negative) returns with positive (negative) sentiment.

Separately, Hawkes models have been used to model interactions between a number of social media sources. One relevant recent application considered the arrival of user submissions on three social media websites—Twitter, Reddit and 4chan—and achieved an understanding of the influence the platforms have on one another in the propagation of political news [150].

The combination of a Hawkes model with topic modelling allowed examination of the self-excitation and mutual-excitation of regional discussion topics (originating from Los Angeles) on Twitter [151]. In this work topics were extracted from a corpus of submissions using non-negative matrix factorisation; when the proportion of a topic in a submission was above 0.1, this was classified as an occurrence of that topic. This classification allowed a time series of topic occurrences (for selected topics) to be generated, upon which a Hawkes model was applied to decipher hidden relationships between the topics. For example, one relationship found was that topics relating to holidays preceded topics relating to basketball, but that topics relating to basketball did not precede topics relating to holidays.

The work described in Section 6.2 and Section 6.3 of this chapter retrieves discussion topics from social media content using dynamic topic modelling, chosen for its ability to track the time-varying interest in different topics. Once topics have been retrieved a Hawkes model is used to decipher hidden interactions between topics and cryptocurrency market prices. Through the combination of dynamic topic modelling and Hawkes models, it is possible to examine which topics precede positive and negative price movements. Section 6.4 explores more general longer-term relationships between topics and market regimes.

6.2 Methodology

6.2.1 Topic modelling

Topic modelling involves using statistical models to discover themes occurring within a corpus automatically; the aim is to find a distribution of words in each topic and the distribution of topics in each document. A *topic* can be considered as a probability distribution over a collection of words, e.g. a topic relating to *football* (soccer) is more likely to contain the words *goal* and *offside* than a topic relating to *cricket*. Since its introduction in 2003 [152], LDA (short for Latent Dirichlet Allocation) has become a popular unsupervised learning technique for topic modelling—through its application, it can cluster terms that appear together across a large corpus into coherent topics. LDA assumes each document contains multiple topics to different extents. The generative process by which LDA assumes each document originates is described below:

1. Choose N ~ Poisson(ξ).
2. Choose θ ~ Dir (α).
3. For each of the N words W_n:
    a. Choose a topic Z_n~ Multinomial(θ).
    b. Choose a word W_n from $p(W_n\,|Z_n, \beta)$, a multinomial probability conditioned on the topic Z_n.

Essentially, for each document, the number of words, N, to generate is chosen (step 1). The process then randomly chooses a distribution over topics, θ (step 2). Then for each word to be generated in the document, the process randomly chooses a topic, Z_n, from the distribution of topics (step 3a), and from that topic chooses a word, W_n, using the distribution of words in the topic (step 3b).

The variables of interest are $\theta_{d,k}$ (the distribution of topic k in document d) and β_k (the distribution of words in topic k). These are latent (hidden) parameters that can be estimated (for a particular dataset) via inference. The generic details of the inference process are described in

[152] for LDA and [153] for dynamic topic models (discussed below). Inference allows for retrieval of per-document topic distributions and per-topic word distributions.

In standard LDA, there is no understanding of either the ordering of words within a document or the ordering of documents within a corpus. The set of topics that make up a particular document does not affect the set of topics that make up the next document. In some contexts, this may not be appropriate. For example, email threads, global news, or in the case presented here, messages on social media, are all examples where there are likely to be temporal trends in topics discussed. As an extension to LDA, a *dynamic topic model* was introduced in 2006 [153]. In a dynamic topic model, there is still no understanding of the order of words in a document, but the order of documents in the corpus is now accounted for, meaning a sequentially organised corpus can be examined for evolving topics. To achieve this, data is divided into time slices over which topics can evolve. It is assumed that topics appearing in one time slice are influenced by topics appearing in the previous time slice; a more detailed definition is provided in [153].

In the current work, the corpus is the collection of timestamped user submissions to a particular subreddit, each submission being considered as a separate document. As done commonly elsewhere (for example, in [152]), the corpus is pre-processed before topic modelling is applied. Stop words (commonly used words such as “the”) are removed. Part-of-speech (POS) tagging is used to categorise words into types; nouns and adjectives are maintained while other types are removed. This filtering was decided upon on the basis of both preliminary work, and because it has been shown elsewhere that reducing a corpus to nouns can improve topic modelling results [154]. Finally, words appearing in less than 20 documents or more than 50% of documents are removed; such removals are commonly done elsewhere [153]. If the proportion of a particular topic in a submission is above 0.1, this is classified as an occurrence of the topic, as in [151]. Once distinct topics have been identified by topic modelling, a time series of topic occurrence can be generated.

A subset of topics are identified (documented below in Section 6.3.1), based on their relevance and relative coherence (other less coherent topics are not analysed further). Highlighting only a subset of topics is common in topic modelling research (for example, in [145]). To evaluate short term predictive power, these chosen topics are then analysed in a Hawkes model, alongside market prices. This approach, to consider a subset of topics, has been used elsewhere when applying Hawkes models to topic modelling results [151]. The approach makes the assumption that the selected components (topics and market prices) exist in isolation (and ignores any explicit relationship with other factors not included in the model). This assumption is suitable for the purpose of this analysis, to decipher how price changes relate to these chosen topics.

6.2.2 Hawkes model

A comprehensive explanation of Hawkes models can be found for example in [146]; the below provides an overview which focuses on how they are applied here. Hawkes models can be used to decipher the interaction dynamics between a group of K processes where the K processes can be considered as an implicit latent network; although connections between processes cannot be directly observed, the connections can be inferred from the temporal patterns of events (emissions) occurring on each process, k. Events are specified depending on the context. An event is essentially a jump in time series values; for example, a jump in market returns or a jump in the discussion of a topic. A definition of the event types relevant to this work will follow later in this subsection. The occurrence of an event on a particular process can cause an impulse response (increasing the likelihood of further events) on a) that process itself (self-excitation) and on b) other processes (mutual-excitation). Given events occurring on a number of processes, the application of a Hawkes model can quantify previously hidden connections between the processes, applied here with the aim of deciphering how topics are related to one another, and how price changes are related to topic occurrence. After being fit to the data, the Hawkes model will contain weights representing the directional strength of any interaction between processes; these weights can be considered as the expected number of events on process B resulting from an event on process A. Figure 18 shows three example processes: market returns and two topics, X and Y.

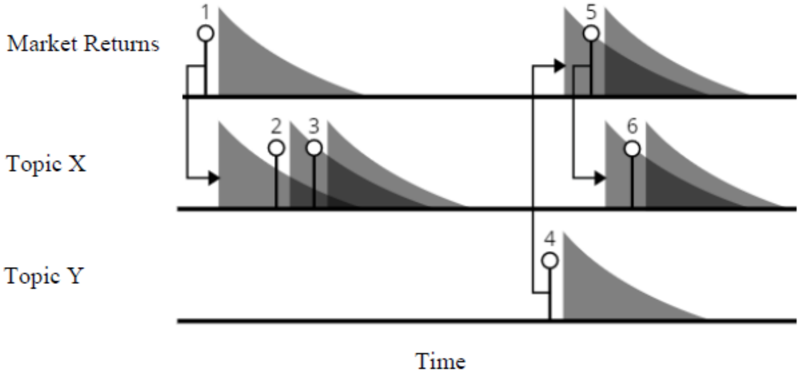

Figure 18. Example events (vertical line with open circle) and impulse responses (grey shading) on three processes

When relationships exist, they can be unidirectional or bidirectional. Each process has a background rate describing the rate of arrival of independent events (events not triggered by preceding events). Event 1 in Figure 18 (a jump in market returns) is an example of an occurrence of an independent event. This causes an impulse response on its own time series (self-excitation) and on the Topic X time series (mutual-excitation). The impulse response increases the likelihood of events on these time series. Event 2 (a jump in topic discussion) occurs, and the resulting self-excitation prompts event 3. Event 4 then occurs on Topic Y's time series. This causes an impulse response on the market returns time series which prompts event 5; the resulting excitation prompts event 6. Overall, Topic X appears responsive to price changes and Topic Y events appear to precede price events. The resulting weight $W_{Returns \rightarrow X}$ would be a number between 0 and 1, and $W_{X \rightarrow Returns}$ would be 0 implying no relationship in that direction. Other relationships can be deciphered similarly.

Hawkes models are most commonly applied to derived time series representing the occurrence of significant events (jumps/extreme changes) in the original time series, rather than to the original time series. As such, steps need to be taken to identify significant events. The occurrence of topics is aggregated into fifteen-minute buckets (Δt = 15). This interval is small enough to avoid having too many overlapping events (for example, both a jump in market returns and a jump in topic occurrence) occurring within the same time bucket, and large enough to find mutual-excitation between buckets. Wider buckets (for example, one hour) are likely to group a number of events into a single bucket, losing the exact ordering of events. The same bucket size was chosen for similar work [149] after smaller intervals (2 and 5 minutes) failed to find excitation between processes. Instead of using absolute values (for example, the count of submissions containing a particular topic within that time bucket), log-returns between buckets are taken here, as is commonly done elsewhere (for example, [148]). Jumps should be specified such that not every (non-zero) log-return is considered an event. A critical value is specified such that log-returns above this value are considered a jump, hence generating a time series of events to be considered by the Hawkes model. It was found that using $\Delta t = 15$ and the 99th percentile of returns meant that 93% of events are non-overlapping (a similar percentage was seen in [150]). The maximum time for which an individual event can have an effect was chosen as one day (dt_max = 96 buckets). One reason for this choice is that cryptocurrencies are a globally traded market, and also often participation is done in a person's spare time, so it cannot be assumed all news will be acted upon instantly as it happens. Experiments with variations of dt_max gave similar results. Inference of parameters (based on the event-based data provided to the model) is achieved via Gibbs sampling, as detailed fully in [155].

6.3 Intraday results

6.3.1 Topic modelling

Table 11 shows notable topics selected for their coherent cryptocurrency-related content (coherent/coherence, in this context, refers to whether a set of words form a meaningful topic). These topics have been manually labelled, as is common in topic modelling (for example, in [77] [152]). The most probable words in each topic are retrieved from the final point in the dataset and displayed; although the probability of words (and thus the most probable words) within a topic varies gradually over time, the gist of the topic remains the same.

The balance between technical and non-technical discussion differs between Ethereum and Bitcoin. Of the 30 identified topics on each subreddit, /r/Ethereum contains only three topics that could be considered price or trading related, while /r/Bitcoin contains twelve. It is hypothesised that this occurs for two reasons: 1) Bitcoin aims to be a currency (and hence price is a big part of it); 2) /r/Ethereum more actively discourages such discussion.

Many topics contain acronyms commonly used in cryptocurrency communities. For example, Topic 3 (on /r/Bitcoin) contains *btc* (Bitcoin), *eth* (Ethereum), *cap* (market capitalisation), *bch* (Bitcoin Cash), *btg* (Bitcoin Gold), and *ath* (all time high), while Topic 6 (on /r/Ethereum) contains *pow* (proof of work) and *pos* (proof of stake).

Table 11. Selected topics from each subreddit

	#	Label	Most probable words
/r/Bitcoin	1	Mainstream adoption	site, dip, interested, website, Amazon, article, company, Google, group, page
	3	Trading terms / Bitcoin alternatives	btc, market, eth, cap, ratio, fork, trade, bch, btg, ath
	20	Substantial price movement	pump, moon[a], dump, sorry, list, quick, dude, random, it'll, way
/r/BitcoinMarkets	4	Downward price movement	big, crash, dip, bubble, huge, major, part, scam, correction, scale
	17	Risk / investment vs trading	trading, risk, everyone, worth, trade, plan, way, advice, strategy, investment
	26	China / announcements	hope, statement, announcement, Chinese, list, announce, official, right, audit, illegal
/r/Ethereum	6	Consensus mechanisms	pos, pow, mining, stake, day, proof, security, network, mine, energy
	8	Hacks / Nervousness	money, attack, wait, someone, long, term, way, internet, short, iota
	23	Fundamental cryptocurrency value	value, crypto, currency, cash, eth, fiat, price, coin, market, news
/r/EthTrader	4	Trading terms	big, mean, moon[a], ratio, support, dip, break, chart, line, joke
	7	Future investments	money, next, crypto, real, devcon, link, investment, year, lot, half
	24	Mainstream adoption / app development	hope, private, key, site, Google, Amazon, bittrex, code, trust, app

a) The term *moon* may seem out of context, however it is cryptocurrency-specific jargon.

6.3.2 Hawkes model

Figure 19 and Figure 20 show the strength of connections ($W_{vertical \rightarrow horizontal}$) between the considered processes for Bitcoin and Ethereum respectively. These weights are extracted from the Hawkes model after fitting to the dataset. Weights are displayed from the vertical to the horizontal axis; for example, the bottom left cell on Figure 19 shows the weight value from Bitcoin price decrease (negative) events to /r/Bitcoin Topic 1 events.

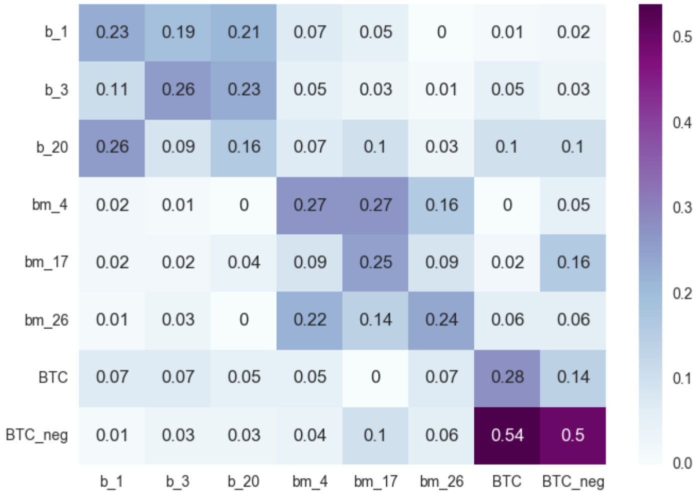

Figure 19. Weight values extracted from Hawkes model for Bitcoin-related topics and price movements; 'b_*' refers to the /r/Bitcoin subreddit and corresponding topic number '' in Table 11. BTC represents positive Bitcoin price shocks and BTC_neg represents negative Bitcoin price shocks.*

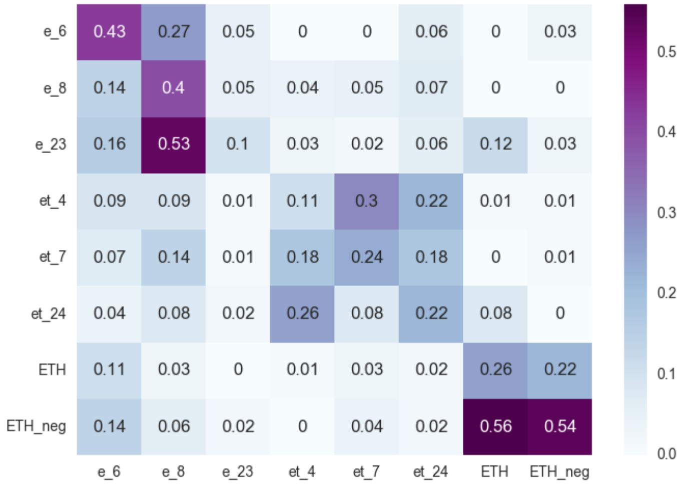

Figure 20. Weight values extracted from Hawkes model for Ethereum-related topics and price movements; 'e_*' refers to the /r/Ethereum subreddit and corresponding topic number '' in Table 11. ETH represents positive Ethereum price shocks and ETH_neg represents negative Ethereum price shocks.*

There is a general pattern of stronger mutual-excitation between topics within a particular subreddit than between topics across different subreddits, observable by the cells having higher weights. Submissions are likely to prompt other submissions (in the same subreddit) as people reply to one another. Notable relationships are discussed below, starting with relationships between topics, and then moving on to relationships observed involving market returns.

The occurrence of a particular topic can influence future occurrences of topics, to varying extents. Self-excitation (seen along the diagonal) is generally stronger than mutual-excitation (seen in non-diagonal cells)—understandably as discussion of a topic is likely to prompt further discussion of the same topic as people reply to each other. A significant non-diagonal relationship is the mutual-excitation between discussion of *'Downward price movement'* (/r/BitcoinMarkets topic 4) and discussion of *'Risk / investment vs trading'* (/r/BitcoinMarkets topic 17), evidenced by $W_{bm\_4 \to bm\_17} = 0.27$. It appears that jumps in discussion of downward price movements prompt people to discuss how they invested for the long-term (rather than actively trading) and are hence less sensitive to downward price movements. Elsewhere, discussion events relating to *'Fundamental cryptocurrency value'* (/r/Ethereum topic 23) are less likely to be triggered by other topics on the subreddit (seen by the smaller weights in the e_23 column), possibly because this is a distinct topic separate from the technical discussion.

It is however of greater interest to explore the relationships between topics and price movements, since discovered relationships could potentially be used in trading. These relationships can be examined in the last two columns of each matrix. The topic *'Substantial price movement'* (/r/Bitcoin topic 20) has a stronger mutual-excitation with price events than most other topics, but does not indicate whether forthcoming price movements are positive or negative (as both $W_{b\_20 \to BTC}$ and $W_{b\_20 \to BTC\_neg}$ = 0.1). This topic can hence be considered as being indicative of price volatility. *'Downward price movement'* (/r/BitcoinMarkets topic 4) has mutual-excitation with future negative price movements (based on $W_{bm\_4 \to BTC\_neg} = 0.05$) and has no relationship with future upwards price movement events ($W_{bm\_4 \to BTC} = 0$). *'Risk / investment vs trading'* (/r/BitcoinMarkets topic 17) events has significant mutual-excitation with negative price movements; this, combined with the observation that negative price movements can excite this topic, suggests that when a downward price movement occurs users start to attempt to reassure others (and themselves) that they are invested for the long-term. However this topic appears to precede further price declines (based on $W_{bm\_17 \to BTC\_neg} = 0.16$, which is higher than any other topic-to-negative-price movement relationship for Bitcoin). Regarding the Ethereum-related subreddits, discussion of *'Fundamental cryptocurrency value'* (/r/Ethereum topic 23) has mutual-excitation with positive price increases (based on $W_{e\_23 \rightarrow ETH} = 0.12$, which is higher than any other topic to price movement relationship for Ethereum). While many topics on the trading subreddit relate to price movements and the value of Ethereum, very few topics on the technical subreddit relate to this, which might tend to add significance when it is discussed. Finally, discussion of *'Mainstream adoption/app development'* (/r/EthTrader topic 24) precedes price rise (versus price fall) events; this topic may relate to news or speculation of adoption, or may be indicative of overall positive sentiment.

As discussed briefly above (in relation to /r/BitcoinMarkets topic 17), price movements can also influence topic discussion. Relationships between price and topics can be examined in the last two rows in each matrix. For example price increase events are likely to lead to discussion events of *'Mainstream adoption'* (/r/Bitcoin topic 1) (compared to price decrease events, which do not show evidence of such a strong relationship), as it is likely that if any news of such adoption comes out, the markets will react immediately and then the news will be discussed on social media for a period after it. In contrast, both positive and negative price movements are likely to precede discussion of *'consensus mechanisms*' (/r/Ethereum topic 6). Ethereum developers have been working on transitioning Ethereum from proof of work to proof of stake, so any news on this—positive or negative—can cause major price movements. In this case, the social media discussion events appear to lag price change events, suggesting the market is gaining awareness of this news from a source other than Reddit (possible alternative sources include GitHub progress and public developer conference calls). Although topic discussion is lagging in this case, the topic is still strongly associated with price events. Intuitively, negative price movements are more likely than positive price movements to trigger discussion of *'Hacks / Nervousness'* (/r/Ethereum topic 8), in line with the observation $W_{ETH\_neg \rightarrow e\_8} = 0.06$ compared to $W_{ETH \rightarrow e\_8} = 0.03$). When a hack occurs it is likely the market will react quicker than social media, as demonstrated by the following example. On July 19th, 2017, an exploit was found in wallet software used by some to store their Ethereum, allowing an attacker to steal funds. Due to the uncertainty caused, the Ethereum price dropped approximately 15% over the first few hours. Social media discussion extended over 24 hours and beyond (first, news of the attack, then a few hours later actions taken to protect vulnerable wallets, then a post-mortem of the exploit published the next day—all causing events detected in this work).

Finally, price movements can influence the likelihood of future price movements. These relationships can be seen in the four cells at the bottom right of each matrix. For both Bitcoin and Ethereum there is strong self-excitation for both positive and negative returns; however for both cryptocurrencies, negative returns are more self-exciting than positive returns. This might result from: 1) negative returns inducing panic; 2) negative returns triggering stop losses, which can cause further negative returns. A Bitcoin price increase event is two times more likely to generate a further price increase event than generate a price decrease event (based on $W_{BTC\rightarrow BTC} = 0.28$ compared to $W_{BTC\rightarrow BTC\_neg} = 0.14$), a much larger ratio than seen for Ethereum, indicating Bitcoin is more trend following (for upwards price movements) than Ethereum.

So far, the work in this chapter has extracted a number of cryptocurrency-related discussion topics and explored the relationship between these topics and the price of the associated cryptocurrency. The content-based analysis introduced here has been shown to provide predictive power in the short term (<24 hours) and, importantly, has found some indicators (occurrence of topics) that are indicative of either upwards or downwards price movements (and not vis-versa). A number of topics were shown to precede price changes over the short term: for example, discussion of *'Fundamental cryptocurrency value'*, on the otherwise technical subreddit /r/Ethereum, precedes a positive return event; that discussion of *'Substantial price movement'* on /r/Bitcoin is indicative of price volatility; and that discussion of *'Risk / investment vs trading'* on /r/BitcoinMarkets, precedes a negative return event. This knowledge of the discussion topics that indicate short term price movements would be a useful component of any manual or automated trading strategy.

6.4 Topic occurrence in different regimes

Having discovered quantifiable short term (intraday) relationships between particular topics and price movements, an obvious progression is to investigate whether longer-term relationships are present between topics and cryptocurrency price movements. To investigate this, this subsection considers whether the occurrence of particular topics changes depending on the market regime. Naturally, people holding investments in particular cryptocurrencies will experience a number of emotions relating to price movements; thus the presence of particular market regimes potentially impacts their discussion online. Although this subsection doesn’t create a trading strategy or price prediction model, its findings provide additional understanding of cryptocurrency community’s intentions, motivations and emotions as the market moves through different regimes. To achieve this, the same topic modelling process is run over an extended period of data (given that longer term relationships will naturally exist over longer durations, and also this additional analysis was conducted after the original analysis, more data at this point existed); in fact the entire history of submissions to a particular subreddit is now considered, from inception of the subreddit to the data of this analysis[28]. The frequency of topic occurrence (the number of posts containing each topic), on each subreddit, is then retrieved per day. This is then converted into relative topic occurrence rather than absolute topic occurrence; the absolute number of posts (*posts per day*) fluctuates and has its own relationship with price. It would be expected that as the absolute number of posts fluctuate, so would the absolute number of topic occurrences to some extent; thus using absolute topic occurrence would hide the true relationship between topic occurrence and price, by including *posts per day* relationships. This is not an issue when using relative topic occurrence.

Given that Hawkes models are good at modelling reactions (i.e. one event causing an almost instantaneous reactionary event) they are not well suited for modelling the longer term relationships now being explored. Rather than using a Hawkes model, this section considers whether the occurrence of particular topics strengthens during particular regimes. To do this, time periods are classified into particular regimes: bubble and non-bubble and, separately, bull and bear. The differing levels of discussion of a topic in different regimes can then be evaluated for statistically significant differences. This allows for an investigation as to whether topical discussion changes have been observed in different price regimes, and provides an understanding of the mindset of market participants in each regime.

[28] As topics do not have a permanent link to a particular topic number (produced by LDA), it is likely that although the same, or at least similar, topics will be discovered to the preceding work in this chapter, topics will have different arbitrary numberings from topic 01 to topic 30. To avoid confusion, topic numbers will not be used in this section, but instead they are referred to by their overarching coherent subject (for example, ‘trading exchanges’ or ‘news’).

6.4.1 Topic occurrence in different regimes: bubbles

The bubble and non-bubble classification is achieved by applying the GSADF bubble test (previously introduced in Section 4.2.3). This test is used rather than the bubble detection model designed, used and evaluated in Chapter 5 because the bubble detection model designed in Chapter 5 relies on social media data as an input. If used here, it outputs would be used for further analysis of the same social media data, meaning there is potential to reduce the clarity of the results; this is avoided by using a non-social media derived bubble test.

An example of the results for the Bitcoin subreddit can be seen in Figure 21, in which certain topics are highlighted on the x-axis (these being the topics for which the bubble regime values exceed the non-bubble regime values).

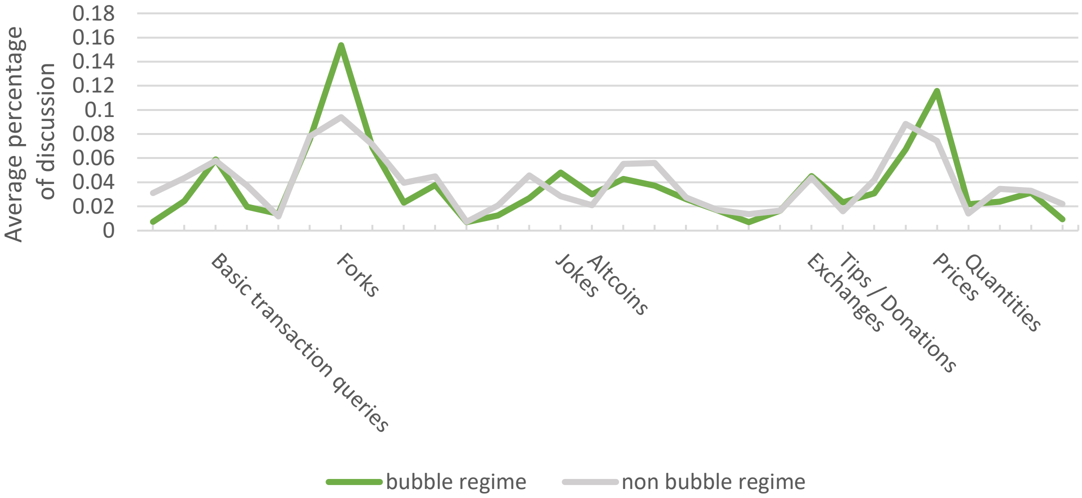

Figure 21. Average percentage of discussion dedicated to particular topics on the Bitcoin subreddit

It can be seen from Figure 21 that discussion of forks and prices increased the most during bubble regimes. After these, topics containing jokes and topics containing altcoins appear to undergo the next most prominent increase, with topics relating to tips/donations, basic transactional queries, exchanges, and quantities of bitcoin owned undergoing the smallest increase. Although the above figure provides a good visualization of whether particular topic occurrence increases in bubble regimes, it does not show whether any increase is statistically significant. To check this, a two sample one-tailed t-test is used, with the results of this test (p-values) shown in Table 12 for all subreddits. The cells with a p-value smaller than 0.01 are highlighted grey; in such cases, the null hypothesis can be rejected in favour of the alternative hypothesis (that the bubble regime values are statistically larger than the non-bubble regime values).

Table 12 T-test p-values (topics with statistically significant increase in bubble regimes)

Subreddit	Topic	P-value
Bitcoin	Basic transaction queries	0.3542
	Forks	0.0000
	Jokes	0.0000
	Altcoins	0.0000
	Exchanges	0.3170
	Tips / donations	0.0008
	Prices	0.0000
	Quantities	0.00321
BitcoinMarkets	Forks	0.0000
	Bubbles	0.3801
	News	0.0000
	Altcoins	0.0010
	Exchange issues	0.4289
	Automated trading	0.1229
	Profit & loss	0.4917
Ethereum	Basic transactions # 1	0.0045
	Basic transactions # 2	0.0000
	Basic transactions # 3	0.0000
	Security	0.0267
	ERC token	0.4751
	Fundamental value	0.4100
	Web addresses	0.1801
	Altcoins	0.0015
	Adoption	0.4588
EthTrader	Basic transactions	0.3495
	Taxes	0.2025
	Banking systems	0.3601
	Exchanges	0.0049
	Tips / donations	0.0022
	Jokes	0.0005
	Future plans	0.0064

It can be observed from Table 12 that a number of topics undergo statistically significant increases during bubble regimes, with certain topics, now to be discussed, being of particular interest. Firstly, discussion of basic transactional queries increases in bubble regimes for Bitcoin, Ethereum and EthTrader; however this is only classed as a statistically significant increase in the Ethereum subreddit where, in fact, discussion relating to basic transactional queries is spread across three separate topics, all of which have a significant increase. It is likely that the observed increases occur because increasing prices during a bubble regime are paired with new interest in a cryptocurrency; these new users are likely to ask basic transactional questions while getting set up with a wallet to hold their cryptocurrency.

Interestingly, although price discussion increases significantly during a bubble on Bitcoin, it does not increase significantly on BitcoinMarkets, Ethereum and EthTrader. For the Ethereum subreddit the reason is clear: discussions relating to prices are not allowed at any point on Ethereum due to subreddit specific rules. Elsewhere the difference observed may be because price is a frequent discussion topic on the trading subreddits (BitcoinMarkets and EthTrader) so any jump during a bubble is proportionally less, or because price discussion is included as part of other discussions (for example, the ‘Profit & loss’ topic on BitcoinMarkets).

Discussion of forks also has a statistically significant increase during bubble regimes on Bitcoin and BitcoinMarkets. A number of forks occurred during the bull (potential bubble) regime towards the end of 2017, which indicates a correlation, but not necessarily causation between one and the other. There is however a chance, for Bitcoin, that forks further fuelled the upwards prices. It is known that speculators understood that Bitcoin forks could provide them with the opportunity to acquire a second Bitcoin-like currency. Such speculators have been seen to buy Bitcoin preceding a fork to retrieve the new forked asset; this movement of funds prior to a fork pushes up the Bitcoin price (and as such, further fuels any bubble). The topic of forks also exists on the Ethereum subreddits. However it doesn’t strengthen during bubble regimes, the reason for this being discussed later in Section 6.4.2.

Finally, it can be noted that the topic labelled ‘Jokes’ undergoes a significant increase on both the Bitcoin and EthTrader subreddits during bubble regimes. This topic encompasses memes (memes, in this context, are images, usually containing a joke, which spread rapidly over the internet) which have become a common part of both communities (so much so the overuse of memes has been commented upon by project founders[29]). As prices rise, so does the number of memes being posted on Bitcoin and EthTrader. Such posts are against the rules of the BitcoinMarkets and the Ethereum subreddit, hence a relationship is not seen on these subreddits. An increase in jokes and memes during bubble regimes could highlight the hysteria and euphoria stage of a bubble regime (the final stage before a crash); an interesting avenue of future work could investigate whether any predictive power could be gained from tracking the use of memes in cryptocurrency communities.

6.4.2 Topic occurrence in different regimes: bull and bear markets

The same process is undertaken as in the preceding subsection; however instead of classifying time series points into bubble and non-bubble regimes time series points are classified into bull and bear markets. To identify time series points that fall within bull and bear regimes, the hidden Markov model outlined in Chapter 5 was redesigned and simplified; it here had one state, the bull state, with positive average returns and another state, the bear state, with negative average returns and higher variance of returns. Empirical evidence from previous applications of hidden Markov models shows that bear markets exhibit higher variance than bull markets [156]. As such, a new structure and starting parameters were designed to capture bull and bear regimes, so as to mirror other hidden Markov models used for this purpose [157] [158]. The hidden Markov model used here was able to output a prediction as to whether each time series point was part of a bull or bear regime, such points then being visually inspected to check that regime classification was occurring reasonably accurately. Due to the observed likelihood of remaining in the same hidden Markov state exceeding the likelihood of switching state, it is likely bull and bear markets will be present for longer than single time points.

It is likely bubble regimes would in fact be a subset of bull markets. Thus it is expected there would be an overlap between topics occurring in a bubble regime and bull market; any such overlap is commented upon where it exists. Table 13 shows those topics which were observed to have statistically significantly increased during bull market regimes.

[29] https://twitter.com/VitalikButerin/status/945988644661207040

Table 13. T-test p-values (topics with statistically significant increase in bull regimes)

Subreddit	Topic	P-value
Bitcoin	Forks	0.0000
	Banking systems	0.0002
	Blockchain	0.0000
	Tips / donations	0.0000
BitcoinMarkets	Exchange issues	0.0038
	Banking systems	0.0088
	Bubbles	0.0000
	News	0.0005
	Quantities	0.0005
Ethereum	Decentralisation	0.0004
	Basic transactions # 1	0.0000
	Basic transactions # 2	0.0058
	Basic transactions # 3	0.0000
	Ethereum games	0.0002
	Security	0.0001
	Wallet setup	0.0023
EthTrader	Banking systems	0.0033
	Tokens / ICOs #1	0.0004
	Tokens / ICOs #2	0.0017
	Exchanges	0.0001
	Tips / donations	0.0058
	News	0.0077
	Jokes	0.0000
	Ratio between BTC and ETH	0.0002
	Future plans	0.0008

It is observed that the discussion of ‘tips/donations’ undergoes a statistically significant increase in both bull and bubble regimes in the Bitcoin and EthTrader subreddits. Although users may be discussing real-world philanthropic uses of their appreciating funds (assuming they own cryptocurrencies which are increasing as part of a bull regime), given the individual words identified in these topics (e.g. ‘tip’, ‘bitcointip’, ‘changetip’) it is more likely that the users are donating to other users on Reddit. Through the use of automated programs, users of Reddit are able to tip others if they appreciate their posts. To do so, the user sending the tip replies to the user they wish to tip with a phrase a pre-configured bot is programmed to detect and operate on: ‘for example,!tip 0.003 /u/tipjarbot’. This sends the user to be tipped a certain amount of cryptocurrency from a wallet owned by the user who is tipping.

The two topics related to tokens/ICOs undergo a statistically significant increase during bull regimes on the EthTrader subreddit. The most common way to fund cryptocurrency projects with associated tokens is to participate in an ICO, hence the appearance of both tokens and ICOs within the same topic clusters. It is likely these topics increase in bull markets because while the ecosystem is prosperous there will be more interest in further new investment opportunities.

Discussion of Ethereum-based games undergoes a statistically significant increase during bull regimes on the Ethereum subreddit. It is likely this increase was caused by the launch of the best-known blockchain based game, CryptoKitties, in November 2017. This was while ETH was undergoing a bull regime, and use of CryptoKitties resulted in Ethereum experiencing an all-time high volume of transactions and struggling with transaction congestion due to the number of transactions occurring (one rare CryptoKitty was sold for $100,000 in December 2017). It is likely that much social media discussion was generated around the sale and the congestion caused by CryptoKitties, thus likely causing the correlation of increased topic discussion and bull regimes, although not necessarily suggesting any causation between the two. This explanation suggests that any correlation between increased discussion of Ethereum-based games and bull market regimes is probably caused by chance—both happening to occur during the same prolonged period.

Discussion relating to banking systems undergoes a statistically significant increase during bull regimes on Bitcoin, BitcoinMarkets and EthTrader. (It was also seen to increase during bubble regimes on EthTrader; however the increase was not large enough to be statistically significant.) The finance industry is well positioned to encourage or discourage the use of cryptocurrencies and related blockchain technology—for example, some banks don’t let customers send money to buy cryptocurrencies—and as such any announcement they make is important and discussed widely. The finance industry has also been one of the fastest to contribute to the cryptocurrency development ecosystem and to suggest real-world applications using blockchain technology, with J.P. Morgan developing Quorum, Thomson Reuters developing BlockOne ID, and Ripple touting a project with the Bank of England.

Finally, as previously seen for bubble regimes, basic transactional queries undergo a statistically significant increase in bull regimes on the EthTrader subreddit. A similar topic also undergoes a significant increase, ‘wallet setup,’ and it is likely both these topics would be initiated by new users. A proposal for a future extension relating to the investigation of topic discussion generated by new users is provided at the end of this chapter.

Topic increases for bear markets can be checked in a similar way as for bull markets, and Table 14 shows those topics that undergo statistically significant increases during bear market regimes.

Table 14. T-test p-values (topics with statistically significant increase in bear regimes)

Subreddit	Topic	P-value
Bitcoin	China	0.0031
	Mining	0.0003
	Bad language #1	0.0000
	Bad language #2	0.0000
	Security	0.0079
	Exchange issues	0.0000
	Price drops	0.0000
BitcoinMarkets	Price drops	0.0000
Ethereum	Forks/DAO #1	0.0000
	Forks/DAO #2	0.0072
	Forks/DAO #3	0.0000
	Governments	0.0049
EthTrader	Price drops	0.0000
	Deleted messages	0.0017
	Bad language	0.0007
	Forks/DAO	0.0000

It is observed that discussion relating to China increases in bear markets. China has slowly applied more and more restrictions to cryptocurrency market operation within its borders; generally, the majority of such announcements have been viewed negatively, leading to a link with bear markets/price decreases). In particular after most of the Chinese announcements applying restrictions (and sometimes preceding leaks of the announcement) there has been a resulting drop in the Bitcoin price[30]. Likewise, any announced government decision, from any country, can often be highly impactful, which accounts for the 'Government' topic also existing on the Ethereum subreddit. A possible reason for the Government topic being discussed more during bear regimes is that negative decisions are more likely to create headlines than positive decisions. (Negative announcements are bans/restrictions, whereas positive government actions are usually gradual steps in the right direction).

Naturally, bear regimes also have increased discussion of price drops (present on Bitcoin, BitcoinMarkets and EthTrader). These discussions are not however present on the Ethereum subreddit, as price discussion is not allowed as per the rules of the subreddit, and any price discussion is redirected to EthTrader. It is seen that the discussion of mining increases in bear markets on Bitcoin. An increase was not observed in the other regimes considered. This could be because when people aren't making money by simply holding the cryptocurrency they look for other avenues of income, and one such avenue is mining.

Whereas forks had previously been associated with bull regimes for Bitcoin, it is observed that for Ethereum there is an increase in discussion related to forks in bear regimes (on both Ethereum and EthTrader). Why the reason for this difference? The main fork to occur to Ethereum was the result of a security breach where the DAO application was hacked; the fork was applied to reacquire the funds obtained by the hacker. The DAO hack caused people to question the future of the Ethereum network and resulted in declining prices. In addition there was division among the community as to how to handle the changes, and uncertainty whether the proposed changes would work, and which fork would have consensus majority. A lot of discussion occurred prior to the fork while the price moved through a bear regime (triggered by the hack). Whereas Bitcoin forks are seen as an opportunity to profit by some, the Ethereum fork was the result of security breaches and was associated with uncertainty, thus causing the observed historical link between Ethereum forks and bear markets.

A topic named 'Deleted messages' increases during bear regimes on EthTrader. Deleted messages can occur for two reason: firstly, a user may opt to remove their message, (though this

[30] News pertaining to China appears to have more impact on Bitcoin than Ethereum, possibly due to the dominance of Bitcoin miners located within China and the historical perception that the majority of Bitcoin trading volume occurred within China. Ethereum has proportionally less trading volume and mining occurring within China.

rarely occurs), and secondly, a moderator may remove a message if it breaches the subreddit rules. Again around the period of the DAO hack, many accounts which had not previously used the EthTrader subreddit arrived at the subreddit to submit messages predicting the downfall of Ethereum. These accounts were either newly registered accounts to Reddit or accounts that had been active within other cryptocurrency communities, but not active on the EthTrader subreddit. The messages were submitted in a manner that could be described as spam or 'trolling,' and many of these messages were removed by moderators. Deleted comments could in general represent periods of controversy or uncertainty that are being capitalized upon by members of other cryptocurrency communities or speculators wanting to drive the price down, thus creating a link between deleted messages and bear markets.

6.4.3 Topic occurrence in different regimes: summary

Section 6.4 identified a number of topics whose occurrences strengthened during historical bubble, bull and bear regimes, and discussed possible reasons for these changes. Such analysis gives an idea of what different cryptocurrency communities care about, what they discuss during different market regimes and in part what motivates them to submit comments on social media (some users participating in certain topics and not others). The commentary also highlighted when increased topic discussion might be due to particular events (such as a hack, or government announcement) that occurred at the same time as a particular regime. Sometimes the relationships observed were the same across both cryptocurrencies (for example, discussion of price drops increasing in bear markets for both Bitcoin and Ethereum) and sometimes they were different (for example, discussion of forks was seen to increase in bull markets for Bitcoin but in bear markets for Ethereum). Certain discussion topics were directly market-related and others were not.

The presence of several fork/DAO related topics demonstrates the large relative impact of the DAO hack on the Ethereum community, whereas the presence of a topic related to China for Bitcoin could indicate a stronger link between China and Bitcoin than between China and Ethereum. Interestingly, although EthTrader is a community intended for traders, it appears many members are holders (or at least, extremely passive traders holding long term positions) of the cryptocurrency rather than active traders. Traders are able to profit from both upwards and downwards movements of a cryptocurrency, so shouldn't show substantially different emotions in each regime. However happiness appears to increase in bull markets (fantasizing about future plans, an increase in jokes/memes) and unhappiness and frustration in bear markets (bad language and deleted messages in bear markets). Finally, it was observed that certain relationships arise (or don't arise) because of subreddit-specific rules relating to what can (and can't) be posted; this secondary finding highlights an important consideration for any work involving Reddit communication in the future.

6.5 Comparison & discussion

Comparison of the shorter term relationships detailed in Section 6.3 and the more general longer term relationships detailed in Section 6.4 can provide an understanding as to whether any discovered topics have a consistent relationship with price, spanning different durations. These are potentially the topics of greatest importance to cryptocurrency communities and trading positions taken from changes in these topics are likely to be more secure when left open (when investments are held over time) than topics that have a short term and contradictory long term relationship.

The remainder of this comparison focuses on a number of topics of interest. Discussion topics relating to jokes and, separately, tips/donations were found to exhibit a longer term relationship with the Bitcoin price, but no shorter-term relationship was found. The lack of a shorter-term relationship here is potentially because these topics are not inherently linked with price movements (it would be almost nonsensical for a shock/jump in one to result a change in the other). The cause of the longer term relationship (the appearance of the topics in bubble and bull markets) is potentially because the topics indicate of overall positive longer-term sentiment. A similar dynamic was found with discussion relating to 1) forks and the Bitcoin price and 2) future investments and plans and the Ethereum price. In both cases there was no short term relationship with price changes in either direction, though a longer term relationship with increasing prices was observed. Discussion topics relating to China were found to exhibit both a longer term and shorter term relationship with the Bitcoin price, being in both cases linked with negative price movements. This suggests breaking news relating to China has the potential to move the market downwards almost immediately and also be linked with longer-term downwards price movements (movement or continuation of a bear regime). A similar dynamic was found with discussion relating to hacks/the DAO and the Ethereum price; in both the shorter and longer term both of these exhibited relationships with falling prices.

A number of additional ideas for future work arise from the findings of this chapter:

1. It could be investigated which topics new users to a subreddit discuss, and how this might change over time. What topics are prompting their arrival, and what topics are they afterward engaging with? The relationship between these sets of topics could indicate whether they plan on staying. Future work in this area could also highlight those topics that are good at capturing new user interest, which would be important to cryptocurrency projects looking to grow their user base.
2. It could be investigated from where a particular user has originated—whether they are a new account, or an existing account that has been active on other cryptocurrency subreddits, as again this potentially provides an indication of their motivations. Findings here could be combined with the *new authors* metrics (analysed in Chapter 4 and Chapter 5) as an optimization, to consider only *new authors* believed to be arriving at a subreddit with good intentions and who are expected to remain part of the community.

Further research that arises from the entire body of work contained in this thesis will be discussed in the following, concluding chapter.

Chapter 7

Conclusions and Future Work

In this final chapter, the main contributions of the research are summarised prior to proposals being given as to how the work can be extended in the context of the dynamic and fast-moving area of cryptocurrency research.

7.1 Discussion and summary of contributions

The main objective of this research was to investigate whether online indicators, especially those from social media, could be harnessed to gain information relevant to, and predictive of, cryptocurrency price movements. At the beginning of this thesis, the introduction suggested qualitative reasons for a link between social media and cryptocurrency markets. Through the work of this thesis, quantitative evidence has been presented to demonstrate that such relationships—although complex and time-evolving—do exist. The work has provided successful ways to gain information from social media that are predictive of upcoming price movements over different durations—short, medium and long term. Such findings would be of interest to those looking to profit from trading cryptocurrency markets (individual traders and hedge funds) and also to the other cryptocurrency market participants and researchers mentioned in the introduction (e.g. technology companies and central banks). As well as providing mechanisms to gain insights into upcoming cryptocurrency price movements, the work here has allowed for knowledge to be gained about cryptocurrency community dynamics, and about the interactions of individuals in the cryptocurrency space while experiencing different price movements, something of interest when understanding cryptocurrency ecosystems dynamics.

Considering the use of social media as a data source, the work introduced influential factors derived from the social media platform Reddit, a platform previously unexplored within cryptocurrency price prediction literature and often overlooked in broader social media data mining applications. Through the successful results achieved here, Reddit has been shown to be a viable data source for cryptocurrency market-related research, and it is hoped that this work prompts further use of Reddit as a data source.

The work achieved its objectives, which in overview were: 1) the identification of a suitable data source, and of factors from that data source, relevant to the prediction of prices and identification of the characteristics of relationships present; 2) the use of these factors to develop a model predictive of medium and long-term cryptocurrency price bubbles; 3) analysis of the content of social media submissions to predict short-term, but major, price movements. The work undertaken in these areas is discussed more thoroughly below.

The research started by reviewing existing literature, where it became apparent that research studies were reporting differing correlations between online indicators and cryptocurrency prices (with respect to both strength and direction). In the preliminary work described in Section 3.4, it was demonstrated that these varying correlations were due to different data periods being used; it was found both strength and direction of correlation varied depending on the particular time interval considered. In addition it was noted that the weak unconditional correlation seen in certain previous studies was likely to be caused by periods of positive and negative correlation offsetting one another. This early finding indicated that a technique of analysis needed to be used that could represent the evolving nature of such relationships. It also provided a first indication that count-based metrics (e.g. *posts per day*, *new authors* per day etc.) captured all usage—positive or negative—and that a possible avenue of research would be to investigate the content of social media to attempt to gain more information.

Chapter 4 examined the evolving nature of complex relationships, using wavelet coherence analysis. Wavelet coherence allowed the time-evolving dynamic relationships to be visualised over time, and considered at different durations: short, medium and long term. In addition the application of such a technique was an important first step as it provided a more concrete understanding of the relationships between social media and cryptocurrency prices before moving on to using a trading strategy, which has the potential to obscure the full relationships present. It was noted from observing the scalograms contained in a previous study using wavelet coherence [69] that there were periods where relationships appeared and disappeared. It was hypothesised in the work in this thesis that the relationships were changing depending on the market regime. Due to speculative bubbles being commonly linked to cryptocurrency markets (as seen in the literature review of previous academic work and news articles), a statistical technique for the detection of bubble market regimes was therefore used. The main finding was that medium and long term correlations between online indicators and associated cryptocurrency prices strengthen during bubble regimes. As noted above, the appearance and disappearance of factor and price relationships are visible in the existing literature. However, the contribution here is an explanation of why these observed relationships appear and disappear at particular times (the relationships strengthen during bubble regimes). In the short term, it was shown that relationships between online indicators and cryptocurrency prices were sparse and, when they did occur, erratic (frequently switching between positive or negative correlation). Both the sparsity and frequently switching polarity of the short term relationships mirrored what was seen in the preliminary work section, whereby the direction of correlation depended on the date range considered. This finding suggested that count-based social media indicators, although shown to display more consistent relationships in the medium and longer term, may not necessarily provide the same predictive power over short term durations. As well as the newly introduced factors from Reddit, factors previously seen in other research such as Wikipedia views and Google search volumes were also considered. In the periods where persistent relationships were present (the medium and long term), the Reddit related indicators were shown to be more consistently leading than the Google Trends and Wikipedia views indicators, highlighting the predictive value of Reddit based indicators. A number of these indicators were shown to exhibit relationships with cryptocurrency prices, acting as the equivalent of fundamentals for companies. To be more explicit, whereas companies have fundamentals such as cash flow and revenue which can be analysed to understand the health of a company, it is likely that cryptocurrency projects, being the unique asset class that they are, have their own set of fundamentals which can be analysed; these community-related indicators were shown to be good candidates as fundamentals useful to value a cryptocurrency project.

Having identified that relationships were present between several promising new indicators and cryptocurrency prices, the next chapter, Chapter 5, focused on using the newly introduced indicators in a predictive model. A model originally designed to detect influenza epidemic outbreaks was repurposed and applied to the newly introduced indicators to make predictions of epidemic-like regimes. The work successfully demonstrated a means by which epidemic detection techniques can be applied to social media data to predict cryptocurrency price bubbles. The predictive power of the model was validated through the generation of a trading strategy which was tested on historical data using a purpose-built backtester. The strategy was profitable and, more importantly, outperformed a challenging benchmark strategy, thus demonstrating the value of the predictions. Most notable was that price was not used as an input to the model and resulting trading strategy; this is important because the use of price would have allowed signals to be generated from price movements so that it would not have been clear how much the social media related indicators were really contributing to the success of the model. The successful use here of social media in an epidemic-based asset bubble model provides a further contribution to the broader literature, in terms of providing a new data source for epidemic-based speculative bubble detection models. It was observed from reviewing the epidemic-based speculative asset bubble literature that often imperfect proxies are used to represent the unknown true extent of word-of-mouth communication [125]; such proxies included geographical proximity of investors, using Ponzi scheme records and using agent-based simulations. Here, social media communication, a much more direct measurement of interpersonal communication, was successfully used to track interest in an investment idea and, for the first time in the literature, in an epidemic-based bubble detection mechanism. Analysis of the individual signals provided by the model demonstrated that occasional false positives were generated. On the evaluation of a selection of the false positives, the root cause of those considered was that the indicators being used were detecting epidemic-like negative social media usage (e.g. increased discussion due to a hack or sudden price decline). This was similar to the reason in Chapter 4 for why count-based metrics had poor predictive power in the short-term (potentially due to the inability to decipher the root cause of short-term spikes) and the observation in the preliminary work section that count-based indicators can be positively or negatively correlated to prices, depending on the event. This recurring disadvantage associated with purely count-based indicators prompted the consideration of content-related analysis in the following chapter.

In Chapter 6, discussion topics were extracted from the content of social media submissions using dynamic topic modelling (a time-evolving variant of the more commonly used technique, LDA). A Hawkes model was then used to decipher relationships between these discussion topics and cryptocurrency prices in the short term, a time period in which count-based factors were shown to provide little predictive power. A number of discussion topics were shown to precede same day major (upwards and/or downwards) price movements. Importantly, certain discussion topics were found to precede only positive, or only negative, price movements, thus achieving what count-based factors did not (count-based factors exhibited non-discriminatory relationships with both positive and negative price movements in the short-term). This demonstrated the additional information that can be retrieved from looking at the content of submissions and moreover was the first such analysis of cryptocurrency discussion on Reddit. Having found such short-term relationships Section 6.4 then investigated whether longer-term relationships are also present. Market regimes were used to represent longer term movements of the price. It was found that particular topics undergo statistically significant increases during historical bubble, bull and bear regimes, and the possible reasons for these changes were discussed. The findings of this section, although not directly predicting price movements, provides additional understanding of why different parties participate and the emotions they experience during different market regimes. Finally, it was observed that certain content-related relationships were present because of subreddit-specific rules that restricted how users interacted while on that subreddit; this finding is especially useful given that the current work is a first introduction of Reddit for cryptocurrency prediction. The work here acts as a pointer for anyone wishing to use data derived from Reddit (especially content-related data), noting that certain characteristics may be present due to rules affecting individual subreddits.

7.2 Future work

Due to the novelty of the area and the promise of the work so far conducted, a range of extensions and advancements could be suggested, as described below.

7.2.1 Proposal 1: A focus on the cause of the end of bubble regimes

Existing cryptocurrency literature has so far focused mainly on tracking bubbles and discovering what causes their initiation and amplification; this includes the work in Chapter 5, which tracked cryptocurrency price bubbles (and led to a profitable trading strategy) by detecting epidemic-like usage of social media. The work conducted in Chapter 5 identified the end of a social-media derived epidemic regime as a signal a bubble was ending; this provided successful tracking and prediction of the bubble regime ending, which though was suitable for the purpose of trading did not explain *why* the bubble regime ended. Research has yet to investigate the causes of the end of cryptocurrency bubbles. Determining the cause of the end of bubble regimes would potentially allow for earlier detection of this end-phase and would be particularly useful for trading strategies as the end of a cryptocurrency bubble can result in dramatic price movements. It is expected that social media, especially Reddit, could be an ideal data source to investigate this (as detailed further below).

In many historical bubbles, a *trigger event* leads to this transition into a *crisis phase* and the end of a bubble. Although these isolated events may appear relatively inconsequential compared to an entire financial market/system, they can manifest unexpectedly to have a much more significant market-wide impact. For example, it has been suggested that the burst of the internet bubble was due to the relatively minor announcement that no patent could be granted to the human genome project [159]. As a second example, the burst of the housing bubble in 2008 (and subsequent recessions seen around the world) has been attributed to the subprime mortgage crisis, which, although well known, only contributed to 4% of the overall mortgage market [159]. A trigger event within the context of cryptocurrency bubbles was the collapse of Mt. Gox, which was undoubtedly the cause of the end of the 2013/2014 Bitcoin price bubble; not only did people lose access to a trading platform and lose ownership of their cryptocurrencies, but the collapse also caused a general distrust and disbelief in cryptocurrencies.

Investigation of potential trigger events experienced within cryptocurrency markets, together with an attempt to decipher whether certain types of events preceded significant periods of negative outlook or price downtrends, would therefore be of interest. The current work has indicated that Reddit could be a suitable platform to identify significant empirical events and subsequent sentiments; all historical Reddit submissions are timestamped and discussion related to such events can be tracked (to quantify the relative importance of each event). As well as identifying trigger events from Reddit based data, sentiment-based analysis could identify whether certain types of events generated short, medium and long-term periods of negativity. The introduction of sentiment analysis leads onto the next proposed advancement of the research.

7.2.2 Proposal 2: Combining sentiment analysis with topic occurrence

The work of Chapter 6 found that the occurrence of particular discussion topics was predictive of associated cryptocurrency price movements in the short term. It is expected that the predictive power could be improved further by including an understanding of whether the topics were being discussed positively or negatively. To achieve this, the understanding of discussion topics achieved in Chapter 6 could be combined with sentiment analysis techniques, allowing sentiment relating to particular topics to be identified. Following this step, the same analysis as in Chapter 6 could be undertaken (i.e. generation of a Hawkes model) to investigate whether certain sentiment-based topics were particularly indicative of future price changes. For example, negative sentiment towards China could be more indicative of price changes than positive sentiment towards China. Similarly, positive or negative sentiment towards China could have more impact than positive or negative sentiment towards topics relating to another country.

The aim would be to improve both 1) the results of the topic modelling undertaken in this work and 2) the results of traditional sentiment techniques already applied in the literature, with no regard being given to discussion topics. The predictive power of such a combined technique could be compared to prediction mechanisms using solely topic modelling or sentiment analysis.

Combinations of sentiment analysis and topic modelling have been explored in the broader social media mining literature [160] [161], and such research would provide a foundation upon which a methodology could be built, and to which results could be compared. Such a combination has not been explored within cryptocurrency markets or to make cryptocurrency price predictions.

7.2.3 Proposal 3: User specific reputation enhancement

The third proposal would aim to improve the quality of information retrieved from social media and hence optimise both current and future social media mining results.

Work so far done both here and in the cryptocurrency-related literature considers volume and content of relevant submissions on social media; however, usually no weight is placed on the background or influence of the individual authors generating such content. Information related to the author could be included to attempt to increase confidence in the significance of specific content or usage. Online cryptocurrency communities are made up of a broad spectrum of authors, from those who are new to cryptocurrencies to the creators of billion-dollar projects. Knowledge of a user's characteristics may allow for an understanding of the accuracy of the content they produce and the influence of their content. For example, a high profile knowledgeable user may produce more insightful/informative content or discuss more promising future projects than newer users (it was already identified in Section 6.4 that the presence of new users, compared to existing users, may result in different topics being discussed).

One technique used in the existing social media mining (not-cryptocurrency specific) literature to automatically assess the influence of a user considers the social connections of a user. It has been observed that those users with a wider online following have higher potential to create influential and accurate content [162]. Common quantifiable factors used to assess a user's influence are their number of followers, retweets by other users, and tweets [163], and such work does show improved results compared to sentiment analysis techniques that take all content as equal.

However, in certain niches, malicious actors try to inflate their perceived influence (which would render the above metrics inaccurate). Malicious actors have been observed in cryptocurrency communities; usually, these accounts are set up to serve a particular purpose such as skewing the discussion of controversial technical decisions known to have divided cryptocurrency communities [95], or to promote scams. This inflation of a user's perceived influence can be achieved with relative ease; a number of services exist online providing different packages, such as buying X number of followers for Y dollars. This influence manipulation is not specific to cryptocurrencies; both academia and the mainstream news have shown interest in this topic due to the use of social media bots in the 2016 US presidential elections (where bots are alleged to have generated 1/5 of election-related social media content) [164].

Given the existence of bots and influence manipulation, one technique to evaluate a user's credibility on a topic is to evaluate their previous sentiment towards the topic. For example, accounts set up solely to support one cryptocurrency are likely to show unwavering support (positive sentiment). Such understanding of an account's prior sentiment in relation to a particular topic has been shown to improve the accuracy of future sentiment based analysis [165]. In a cryptocurrency context, this account-based prior sentiment would be useful for a number of reasons: 1) users are prone to *'talk their book'* (promote the investments they own) making it likely that a social media user with previous positive sentiment about a particular cryptocurrency will continue to show positive sentiment (while the user still owns that cryptocurrency); 2) accounts that have shown previous varying sentiment towards a cryptocurrency may be identified as providing unbiased content; 3) accounts showing positive sentiment towards a cryptocurrency may show negative sentiment towards competing cryptocurrencies. The ability to track individual user-level sentiment towards particular cryptocurrencies could be used to filter out as less-useful sources those who show unwavering support towards a particular project and track how sentiment displayed by individual users changes over time. This improved understanding of sentiment should provide greater predictive power within cryptocurrency markets.

7.2.4 Proposal 4: Mechanisms for creating stable cryptocurrency prices

As outlined in the introduction to this thesis (Chapter 1), there are a number of types of participants in cryptocurrency markets each participating for their own reasons. Although traders may favour the current volatility of cryptocurrency markets—and can profit from this volatility by predicting price movements, including via indicators derived from social media, as demonstrated herein—it was noted there are other participant types, including central banks and technology companies, who would favour less volatility. Use of a cryptocurrency by such parties is important for the success of the cryptocurrency and for the blockchain ecosystem in the long term. The final proposal takes a higher level view of the research that would be needed to move forward the cryptocurrency ecosystem: research into producing stable valued cryptocurrencies.

Although Bitcoin, Ethereum and other already existing cryptocurrencies are unlikely to change their internal mechanisms to ensure stability, new cryptocurrencies could be designed with stability in mind. Cryptocurrencies which aim for price stability are termed *stablecoins*. Design mechanisms for such cryptocurrencies could be viewed as a vital research area upon which future adoption of the cryptocurrency ecosystem might to some extent rely. Stablecoins create an economic reason as to why their value, and purchasing power, should remain stable. The findings documented in this thesis, although not providing stability, at least would help predict periods of stability (or conversely, periods which lack stability) in existing cryptocurrencies. It is likely that having an understanding of the price dynamics of current speculative cryptocurrencies would help guide decisions when designing new more stable cryptocurrencies.

Returning to the properties to be expected of money (which Bitcoin was judged not to meet well, earlier in Section 2.1.1), it is likely that a hypothetical stablecoin that functions as desired should be: 1) a store of value (as it is stable, by design); 2) a unit of account, as goods priced in the stablecoin would not have to rapidly update their prices to reflect changes in the stablecoin’s price; 3) likely to be a better medium of exchange than volatile assets. The last would require the currency to be able to purchase goods and require belief by vendors and sellers of goods that it is an adequate currency to accept; their confidence in the currency is likely to be helped by the currency’s ability to meet 1) and 2).

Such a hypothetical perfectly functioning stablecoin does not yet exist. Stablecoin mechanisms have been proposed, usually by industry or hybrid industry / academic work, and have had varying levels of success. Mechanisms for ensuring cryptocurrency stability fall into three main categories: 1) fiat-based collateralisation, 2) cryptocurrency-based collateralisation, and 3) non-collateralised algorithmic-driven stability. Fiat-based collateralisation involves issuing a cryptocurrency backed by an underlying asset held by a trusted custodian, Tether being the best-known fiat-based collateralised cryptocurrency. Cryptocurrency-based collateralisation involves issuing some amount of a stablecoin by locking up other cryptocurrencies as collateral, with an example of a cryptocurrency-collateralised stablecoin being DAI[31]. Non-collateralised algorithmic-driven stability involves issuing a cryptocurrency whose supply increases and decreases automatically based on demand fluctuations, an example project being Carbon[32].

As more stablecoin research is done, and implementations emerge, their regulatory standpoint, societal implications and associated trading strategies (including game theory considerations) are likely to present their own fields of research. Regarding the trading strategies, any widespread use of a stablecoin is likely to introduce a new form of trading strategy—at a high level either betting with the system (allowing the system to maintain its planned stability and being rewarded for doing so) or betting against the currency (similarly to how speculators bet against the Bank Of England's currency peg in 1992, leading to its demise [166]).

7.3 Concluding remarks

The prediction of the movement of financial markets is, and will remain, highly challenging, influenced as it is by imponderable and random factors. Yet no-one would doubt that financial markets are fundamentally a manifestiation of human behaviours such as euphoria, fear, and greed. Social media have made certain human behaviours easy to monitor and analyse; cryptocurrencies are unique as assets in that those involved in online communities (specifically, Reddit) overlap very substantially with those involved in developing the projects and in other activities that shift prices. Thus at this point in time observation of social media behaviour, via Reddit, is opening up new and effective avenues for the pursuit of profit.

[31] https://makerdao.com/en/
[32] https://www.carbon.money/

Bibliography

[1] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008. [Online]. Available: http://www.bitcoin.org/bitcoin.pdf. [Accessed 07 02 2019].

[2] G. Danezis and S. Meiklejohn, “Centrally banked cryptocurrencies,” in *Network and Distributed System Security Symposium (NDSS),* San Diego, California, 2016.

[3] E. Mokhtarian and A. Lindgren, “Rise of the crypto hedge fund: Operational issues and best practices for emergent investment industry,” *Stanford Journal of Law, Business, and Finance,* vol. 23, no. 1, pp. 112-158, 2018.

[4] F. Glaser, K. Zimmermann, M. Haferkorn, M. C. Weber and M. Siering, “Bitcoin - asset or currency? Revealing users' hidden intentions,” in *Proceedings of the European Conference on Information Systems (ECIS),* Tel Aviv, Israel, 2014.

[5] C. Baek and M. Elbeck, “Bitcoins as an investment or speculative vehicle? A first look,” *Applied Economics Letters,* vol. 22, no. 1, pp. 30-34, 2015.

[6] M. Nardo, M. Petracco and M. Naltsidis, “Walking down Wall Street with a tablet: A survey of stock market predictions using the web,” *Journal of Economic Surveys,* vol. 30, no. 2, pp. 356-369, 2015.

[7] “'Pump-and-dumps' and market manipulations,” U.S. Securities and Exchange Commission, 25 06 2013. [Online]. Available: https://www.sec.gov/fast-answers/answerspumpdumphtm.html. [Accessed 05 02 2019].

[8] J. Papp, “A medium of exchange for an internet age: How to regulate bitcoin for the growth of e-commerce,” *Pittsburgh Journal of Technology Law & Policy,* vol. 15, no. 1, 2014.

[9] D. Garcia, C. J. Tessone and P. Mavrodiev, “The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy,” *Journal of The Royal Society interface* , vol. 11, 2014.

[10] A. B. Perkins and M. C. Perkins, The internet bubble: Inside the overvalued world of high tech stocks, New York: Harper Business, 1999.

[11] G. Garino and L. Sarno, “Speculative bubbles in U.K. house prices: Some new evidence,” *Southern Economic Journal,* vol. 70, no. 4, pp. 777-795, 2004.

[12] E. T. Cheah and J. Fry, “Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin,” *Economics Letters,* vol. 130, pp. 32-36, 2015.

[13] L. Swartz, “What was Bitcoin, what will it be? The techno-economic imaginaries of a new money technology,” *Cultural Studies,* vol. 32, no. 4, pp. 623-650, 2018.

[14] W. Gao, W. G. Hatcher and W. Yu, “A survey of blockchain: techniques, applications, and challenges,” in *27th International Conference on Computer Communication and Networks (ICCCN),* Hangzhou, China, 2018.

[15] P. Tasca and C. Tessone, “Taxonomy of blockchain technologies. Principles of identification and classification,” *SSRN,* 2018.

[16] G. Hileman and M. Rauchs, “Global cryptocurrency benchmarking study,” *SSRN,* 2017.

[17] A. E. Gencer, S. Basu, I. Eyal, R. van Renesse and E. G. Sirer, “Decentralization in Bitcoin and Ethereum networks,” in *Financial Cryptography and Data Security,* Curaçao, 2018.

[18] S. Chow and M. E. Peck, “The Bitcoin mines of China,” *IEEE Spectrum,* vol. 54, no. 10, 2017.

[19] V. Buterin, “Ethereum white paper: A next-generation smart contract and decentralized application platform,” 2013. [Online]. Available: https://github.com/ethereum/wiki/wiki/White-Paper. [Accessed 05 02 2019].

[20] C. D. Clack, V. A. Bakshi and L. Braine, “Smart contract templates: Foundations, design landscape and research directions,” *arXiv,* 2017.

[21] W. Cai, Z. Wang, J. B. Ernst, Z. Hong, C. Feng and V. C. M. Leung, “Decentralized applications: The blockchain-empowered software system,” *IEEE Access,* vol. 6, 2018.

[22] S. Adhami, G. Giudici and S. Martinazzi, “Why do businesses go crypto? An empirical analysis of initial coin offerings,” *Journal of Economics and Business,* vol. 100, pp. 64-75, 2018.

[23] S. T. Howell, M. Niessner and D. Yermack, “Initial coin offerings: Financing growth with cryptocurrency token sales,” *NBER Working Paper No. 24774,* 2018.

[24] D. A. Zetzsche, R. P. Buckley, D. W. Arner and L. Föhr, “The ICO Gold Rush: It’s a scam, it’s a bubble, it’s a super challenge for regulators,” *Harvard International Law Journal,* vol. 63, no. 2, 2019.

[25] C. Berg, S. Davidson and J. Potts, “Institutional discovery and competition in the evolution of blockchain technology,” *SSRN,* 2018.

[26] S. Azouvi, M. Maller and S. Meiklejohn, “Egalitarian society or benevolent dictatorship: The state of cryptocurrency governance,” in *The 5th Workshop on Bitcoin and Blockchain Research (Financial Cryptography and Data Security),* 2018.

[27] “The state of the token market,” 2018. [Online]. Available: https://www.fabric.vc/s/State-of-the-Token-Market-2-FINAL.pdf. [Accessed 05 02 2019].

[28] C. Dierksmeier, “Just HODL? On the moral claims of Bitcoin and Ripple users,” *Humanistic Management Journal,* vol. 3, no. 1, pp. 127-131, 2018.

[29] D. G. Baur and T. Dimpfl, “Realized Bitcoin volatility,” *SSRN,* 2017.

[30] S. Lo and C. J. Wang, “Bitcoin as money?,” in *Current Policy Perspectives,* Federal Reserve Bank of Boston, 2014.

[31] A. Zohar, “Bitcoin: Under the hood,” *Communications of the ACM,* vol. 58, no. 9, pp. 104-113, 2015.

[32] E. K. Kogias, P. Jovanovic, N. Gailly, I. Khoffi, L. Gasser and B. Ford, “Enhancing Bitcoin security and performance with strong consistency via collective signing,” *Proceedings of the 25th USENIX Security Symposium,* 2016.

[33] K. Baqer, D. Y. Huang, D. McCoy and N. Weaver, “Stressing out: Bitcoin 'stress testing',” *Financial Cryptography and Data Security,* vol. 9604, pp. 3-18, 2016.

[34] T. Aste, “The fair cost of Bitcoin proof of work,” *SSRN,* 2016.

[35] K. J. O’Dwyer and D. Malone, “Bitcoin mining and its energy footprint,” *25th IET Irish Signals & Systems Conference 2014 (ISSC),* 2014.

[36] R. A. Radford, “The economic organisation of a P.O.W. camp,” *Economica,* vol. 12, no. 48, pp. 189-201, 1945.

[37] H. S. Shin, “Cryptocurrencies and the economies of money,” 24 6 2018. [Online]. Available: https://www.bis.org/speeches/sp180624b.pdf. [Accessed 19 1 2019].

[38] G. Claeys, M. Demertzis and K. Efstathiou, “Cryptocurrencies and monetary policy,” Policy Department for Economic, Scientific and Quality of Life Policies (European Parliament), 2018.

[39] “2018 cryptocurrency survey,” 2018. [Online]. Available: https://www.foley.com/files/uploads/Foley-Cryptocurrency-Survey.pdf. [Accessed 05 02 2019].

[40] M. Rauchs, A. Blandin, K. Klein, G. Pieters, M. Recanatini and B. Zhang, “2nd global cryptoasset benchmarking study,” SSRN, 2018.

[41] “2019 examination priorities,” 2019. [Online]. Available: https://www.sec.gov/files/OCIE%202019%20Priorities.pdf. [Accessed 05 02 2019].

[42] HM Revenue & Customs, "Policy paper: cryptoassets for individuals," 19 12 2018. [Online]. Available: https://www.gov.uk/government/publications/tax-on-cryptoassets/cryptoassets-for-individuals. [Accessed 19 01 2019].

[43] A. Sotiropoulou and D. Guegan, "Bitcoin and the challenges for financial regulation," *Capital Markets Law Journal,* vol. 12, no. 4, pp. 466-479, 2017.

[44] R. C. Goforth, "U.S. law: Crypto is money, property, a commodity, and a security, all at the same time," *Journal of Financial Transformation, Forthcoming,* 2018.

[45] W. Hinman, Interviewee, *Digital asset transactions: When howey met gary (plastic).* [Interview]. 14 6 2018.

[46] "Regulation of cryptocurrency in selected jurisdictions," 2018. [Online]. Available: https://www.loc.gov/law/help/cryptocurrency/regulation-of-cryptocurrency.pdf. [Accessed 07 02 2019].

[47] J. Lansky, "Possible state approaches to cryptocurrencies," *Journal of Systems Integration,* vol. 9, no. 1, 2018.

[48] C. Russo, "Bitcoin speculators, not drug dealers, dominate crypto use now," Bloomberg, 7 8 2018. [Online]. Available: https://www.bloomberg.com/news/articles/2018-08-07/bitcoin-speculators-not-drug-dealers-dominate-crypto-use-now. [Accessed 16 1 2019].

[49] D. G. Baur, K. Hong and A. D. Lee, "Bitcoin: Medium of exchange or speculative assets?," *Journal of International Financial Markets, Institutions and Money,* 2017.

[50] D. M. Gould, M. A. Porter, S. Williams, M. Mcdonald, D. J. Fenn and S. D. Howison, "Limit order books," *Quantitative Finance,* vol. 13, no. 11, pp. 1709-1742, 2013.

[51] P. McCorry, M. Möser and S. T. Ali, "Why preventing a cryptocurrency exchange heist isn't good enough," in *Security Protocols 2018*, Cambridge, 2018.

[52] T. Moore and N. Christin, "Beware the middleman: Empirical analysis of Bitcoin-exchange risk," *Financial Cryptography and Data Security,* pp. 25-33, 2013.

[53] "Global cryptocurrency market report," 2018. [Online]. Available: https://media.ibinex.com/docs/Global_Cryptocurrency_Market_Report_2018.pdf. [Accessed 05 02 2019].

[54] "CryptoCompare October 2018 exchange review," 2018. [Online]. Available: https://www.cryptocompare.com/media/34836036/cryptocompare_exchange_review_october_2018.pdf. [Accessed 05 02 2019].

[55] M. Zima, "Coincer: Decentralised trustless platform for exchanging decentralised cryptocurrencies," in *Network and System Security,* Helsinki, 2017.

[56] M. Brandvold, P. Molnár, K. Vagstad and O. C. A. Valstad, "Price discovery on Bitcoin exchanges," *Journal of International Financial Markets, Institutions and Money,* vol. 36, pp. 18-35, 2015.

[57] L. Cocco, G. Concas and M. Marchesi, "Using an artificial financial market for studying a cryptocurrency market," *Journal of Economic Interaction and Coordination,* vol. 12, no. 2, pp. 345-365, 2017.

[58] T. Dimpfl, "Bitcoin market microstructure," *SSRN,* 2017.

[59] A. Eross, A. Urquhart, F. McGroarty and S. Wolfe, "The intraday dynamics of Bitcoin," *SSRN,* 2017.

[60] A. Urquhart, "Price clustering in bitcoin," *Economics Letters,* vol. 159, pp. 145-148, 2017.

[61] B. Hu, T. McInish, J. Miller and L. Zeng, "Intraday price behavior of cryptocurrencies," *Finance Research Letters,* 2018.

[62] S. McNally, J. Roche and S. Caton, "Predicting the price of Bitcoin using machine learning," in *Euromicro International Conference on Parallel, Distributed and Network-based Processing,* Cambridge, 2018.

[63] L. Alessandretti, A. ElBahrawy, L. M. Aiello and A. Baronchelli, "Anticipating cryptocurrency prices using machine learning," *Complexity,* 2018.

[64] R. Selmi, A. K. Tiwari and S. Hammoudeh, “Efficiency or speculation? A dynamic analysis of the Bitcoin market,” *Economics Bulletin,* vol. 38, no. 4, pp. 2037-2046, 2018.

[65] W. Feng, Y. Wang and Z. Zhang, “Informed trading in the Bitcoin market,” *Finance Research Letters,* vol. 26, pp. 63-70, 2017.

[66] H. Yang, “Behavioral anomalies in cryptocurrency markets,” *SSRN,* 2018.

[67] P. M. Krafft, N. D. Penna and A. Pentland, “An experimental study of cryptocurrency market dynamics,” in *ACM Conference on Human Factors in Computing Systems,* Montréal, Canada, 2018.

[68] E. Bouri, R. Gupta and D. Roubaud, “Herding behaviour in cryptocurrencies,” *Finance Research Letters,* 2018.

[69] L. Kristoufek, “What are the main drivers of the bitcoin price? Evidence from wavelet coherence analysis,” *PLoS ONE,* 2015.

[70] L. Bursztyn, F. Ederer, B. Ferman and N. Yuchtman, “Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions,” *Econometrica,* vol. 82, no. 4, 2014.

[71] R. Z. Heimer, “Peer pressure: Social interaction and the disposition effect,” *The Review of Financial Studies,* vol. 29, no. 11, pp. 3177-3209, 2016.

[72] V. K. Pool, N. Stoffman and S. E. Yonker, “The people in your neighborhood: Social interactions and mutual fund portfolios,” *The Journal of Finance,* vol. 70, no. 6, 2014.

[73] J. Bollen, H. Mao and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science,* vol. 2, no. 1, pp. 1-8, 2011.

[74] I. Zheludev, R. Smith and T. Aste, “When can social media lead financial markets?,” *Nature: Scientific Reports,* vol. 4, 2014.

[75] L. Kristoufek, “Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era,” *Nature: Scientific Reports,* vol. 3, no. 1, 2013.

[76] T. Panagiotidis, T. Stengos and O. Vravosinos, “On the determinants of bitcoin returns: A LASSO approach,” *Finance Research Letters,* vol. 27, pp. 235-240, 2018.

[77] Y. Kim, J. Lee, N. Park, J. Choo, J. Kim and C. Kim, “When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation,” *PLoS ONE,* vol. 12, no. 5, 2017.

[78] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica,* vol. 37, no. 3, pp. 424-438, 1969.

[79] Y. B. Kim, J. G. Kim, W. Kim, J. H. Im, T. H. Kim, J. S. Kang and C. H. Kim, “Predicting fluctuations in cryptocurrency transactions based on user comments and replies,” *PLoS ONE,* vol. 11, no. 8, 2016.

[80] P. Xie, H. Chen and Y. J. Hu, “Network structure and predictive power of social media for the Bitcoin market,” *SSRN,* 2017.

[81] E. Jahani, P. M. Kraft, Y. Suhara, E. Moro and A. Pentland, “ScamCoins, s*** posters, and the search for the next Bitcoin™: Collective sensemaking in cryptocurrency discussions,” *Proceedings of the ACM on Human-Computer Interaction,* vol. 2, 2018.

[82] F. Mai, Z. Shan, Q. Bai, X. Wang and R. H. Chiang, “How does social media impact Bitcoin value? A test of the silent majority hypothesis,” *Journal of Management Information Systems,* vol. 35, no. 1, pp. 19-52, 2018.

[83] V. Karalevicius, “Using sentiment analysis to predict interday Bitcoin price movements,” *The Journal of Risk Finance,* vol. 19, no. 1, pp. 56-75, 2018.

[84] S. Wang and J. P. Vergne, “Buzz factor or innovation potential: What explains cryptocurrencies’ returns?,” *PLoS ONE,* vol. 12, no. 5, 2017.

[85] M. Matta, I. Lunesu and M. Marchesi, “The predictor impact of web search media on Bitcoin trading volumes,” in *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management,* Lisbon, 2015.

[86] A. Urquhart, “What causes the attention of Bitcoin?,” *Economics Letters,* vol. 166, pp. 40-44, 2018.

[87] M. Matta, I. Lunesu and M. Marchesi, “Bitcoin spread prediction using social and web search media,” *DeCAT 15: Proceedings of the Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services,* 2015.

[88] D. Garcia and F. Schweitzer, “Social signals and algorithmic trading of Bitcoin,” *Royal Society Open Science,* vol. 2, no. 9, 2015.

[89] J. Kaminski, “Nowcasting the Bitcoin market with Twitter signals,” *arXiv,* 2016.

[90] T. Loughran and B. Mcdonald, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” *The Journal of Finance,* vol. 66, no. 1, pp. 35-65, 2011.

[91] Z. Gilani, R. Farahbakhsh and J. Crowcroft, “Do bots impact Twitter activity?,” *Proceedings of the 26th International Conference on World Wide Web Companion,* pp. 781-782, 2017.

[92] M. Laskowski and H. M. Kim, “Rapid prototyping of a text mining application for cryptocurrency Market intelligence,” *Information Reuse and Integration (IRI),* 2016.

[93] I. Hernandez, M. Bashir, G. Jeon and J. Bohr, “Are Bitcoin users less sociable? An analysis of users’ language and social connections on Twitter,” *HCI International 2014 - Posters’ Extended Abstracts,* pp. 26-31, 2014.

[94] A. Jethin, H. Daniel, N. John and I. Juan, “Cryptocurrency price prediction using Tweet volumes and sentiment analysis,” *SMU Data Science Review,* vol. 1, no. 3, 2018.

[95] A. Haynes, “Author attribution in the Bitcoin blocksize,” 2015. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1004.3077. [Accessed 07 02 2019].

[96] J. Seungmin, R. Ali and A. V. Vlasov, "Cryptoeconomics: Data application for token sales analysis," *International Conference Information Systems 2017 Special Interest Group on Big Data Proceedings,* 2017.

[97] R. Schäfer and T. Guhr, "Local normalization: Uncovering correlations in non-stationary financial time series," *Physica A: Statistical Mechanics and its Applications,* vol. 389, no. 18, pp. 3856-3865, 2010.

[98] G. Ranco, D. Aleksovski, G. Caldarelli and M. Grča, "The effects of Twitter sentiment on stock price returns," *PLoS ONE,* vol. 10, no. 9, 2015.

[99] J. Aaltonen and R. Östermark, "A rolling test of granger causality between the Finnish and Japanese security markets," *Omega,* vol. 25, no. 6, pp. 635-642, 1997.

[100] K. Smith, J. Brocato and J. Rogers, "Regularities in the data between major equity markets: evidence from Granger causality tests," *Applied Financial Economics,* vol. 3, no. 1, pp. 55-60, 1993.

[101] M. Balcilar, E. Bouri, R. Gupta and D. Roubaud, "Can volume predict Bitcoin returns and volatility? A quantiles-based approach," *Economic Modelling,* vol. 64, pp. 74-81, 2017.

[102] R. Shiller, "Speculative asset prices," *American Economic Review,* vol. 104, no. 6, pp. 1486-1517, 2014.

[103] J. Scheinkman and W. Xiong, "Overconfidence and speculative bubbles," *Journal of Political Economy,* vol. 111, no. 6, 2003.

[104] P. Phillips, S. Shi and J. Yu, "Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500," *International Economic Review,* vol. 56, no. 4, pp. 1043-1078, 2015.

[105] M. Bianchetti, C. Ricci and M. Scaringi, "Are cryptocurrencies real financial bubbles? Evidence from quantitative analyses," *SSRN,* 2018.

[106] S. Wheatley, D. Sornette, T. Huber, M. Reppen and R. N. Gantner, “Are Bitcoin bubbles predictable? Combining a generalized Metcalfe's Law and the LPPLS model,” *arXiv,* 2018.

[107] B. Metcalfe, “Metcalfe's Law after 40 years of ethernet,” *Computer,* vol. 46, no. 12, pp. 26-31, 2013.

[108] D. Sornette and R. Woodard, “Financial bubbles, real estate bubbles, derivative bubbles, and the financial and economic crisis,” *Econophysics Approaches to Large-Scale Business Data and Financial Crisis,* pp. 101-148, 2010.

[109] C. Aloui and B. Hkiri, “Co-movements of GCC emerging stock markets: New evidence from wavelet coherence analysis,” *Economic Modelling,* vol. 36, pp. 421-31, 2014.

[110] L. Vacha and J. Barunik, “Co-movement of energy commodities revisited: Evidence from wavelet coherence analysis,” *Energy Economics*, vol. 34, no. 1, pp. 241-7, 2012.

[111] M. Madaleno and C. Pinho, “Wavelet dynamics for oil-stock world interactions,” *Energy Economics,* vol. 45, pp. 120-133, 2014.

[112] Y. Xu , Z. Liu, J. Zhao and C. Su, “Weibo sentiments and stock return: A time-frequency view,” *PLoS ONE,* vol. 12, no. 7, 2017.

[113] E. Bouri, R. Gupta, A. Tiwari and D. Roubaud, “Does Bitcoin hedge global uncertainty? Evidence from wavelet-based quantile-in-quantile regressions,” *Finance Research Letters,* 2017.

[114] S. J. Lim, “Exploring portfolio diversification opportunities in Islamic capital markets through bitcoin: evidence from MGARCH-DCC and wavelet approaches,” *Munich Personal RePEc Archive,* 2017.

[115] A. Grinsted , J. C. Moore and S. Jevrejeva, “Application of the cross wavelet transform and wavelet coherence to geophysical time series,” *Nonlinear Processes in Geophysics.,* vol. 11, no. 5/6, pp. 561-566, 2004.

[116] P. Phillips, Y. Wu and J. Yu, “Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values?,” *International Economic Review,* vol. 52, no. 1, pp. 201-226, 2011.

[117] J. Mitchell, “Clustering and psychological barriers: the importance of numbers,” *Journal of Futures Markets,* vol. 21, no. 5, pp. 395-428, 2001.

[118] S. Kucukemiroglu and A. Kara, “Online word-of-mouth communication on social networking sites: An empirical study of Facebook users,” *International Journal of Commerce and Management,* vol. 25, no. 1, pp. 2-20, 2015.

[119] M. A. Martínez-Beneito, D. Conesa and A. López-Quílez, “Bayesian Markov switching models for the early detection of influenza epidemics,” *Statistics in Medicine,* vol. 27, no. 22, pp. 4455-4468, 2008.

[120] Z. Ivković and S. Weisbenner, “Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices,” *The Review of Financial Studies,* vol. 20, no. 4, pp. 1327-1357, 2007.

[121] H. Hong, J. Kubik and J. Stein, “Social interaction and stock-market participation,” *The Journal of Finance,* vol. 59, no. 1, 2005.

[122] R. J. Shiller, Irrational exuberance, Princeton University Press, 2014.

[123] S. Shive, “An epidemic model of investor behavior,” *The Journal of Financial and Quantitative Analysis,* vol. 45, no. 1, pp. 169-198, 2010.

[124] C. Burnside, M. Eichenbaum and S. Rebelo, “Understanding booms and busts in housing markets,” *Journal of Political Economy,* vol. 124, no. 4, 2016.

[125] P. J. Bayer, K. Mangum and J. W. Roberts, “Speculative fever: Investor contagion in the housing bubble,” *NBER,* 2016.

[126] N. D. Pearson, Z. Yang and Q. Zhang, “Evidence about bubble mechanisms: Precipitating event, feedback trading, and social contagion,” *7th Miami Behavioral Finance Conference,* 2016.

[127] D. Simon and R. Heimer, “Facebook finance: how social interaction propagates active investing,” *AFA San Diego Meetings Paper,* 2013.

[128] V. Rantala, “How do investment ideas spread through social interaction? Evidence from a ponzi scheme,” *6th Miami Behavioral Finance Conference,* 2015.

[129] D. Hong and H. G. Hong, “An epidemiological approach to opinion and price-volume dynamics,” *AFA Chicago Meetings Paper,* 2012.

[130] F. Jin, E. Dougherty, P. Saraf, Y. Cao and N. Ramakrishnan, “Epidemiological modelling of news and rumours on Twitter,” *Proceedings of the 7th Workshop on Social Network Mining and Analysis,* 2013.

[131] B. Markey-Towler, “Narratives and Chinese whispers: ideas and knowledge in bubbles, diffusion of technology and policy,” *SSRN,* 2017.

[132] C. I. Siettos and L. Russo, “Mathematical modeling of infectious disease,” *Virulence,* vol. 4, no. 4, pp. 295-306, 2013.

[133] S. Abdullah and X. Wu, “An epidemic model for news spreading on Twitter,” *IEEE 23rd International Conference on Tools with Artificial Intelligence,* 2011.

[134] A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras and A. Baronchelli, “Evolutionary dynamics of the cryptocurrency market,” *Royal Society: Open Science,* vol. 4, no. 11, 2017.

[135] A. Gelman, “Prior distributions for variance parameters in hierarchical models,” *Bayesian Analysis,* vol. 1, no. 3, pp. 515-534, 2006.

[136] S. H. Park, J. H. Lee, J. W. Song and T. S. Park, “Forecasting change directions for financial time series using hidden Markov model,” *International Conference on Rough Sets and Knowledge Technology,* pp. 184-191, 2009.

[137] W. Sherchan, S. Nepal and A. Bouguettaya, “A trust prediction model for service web,” *IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications,* 2011.

[138] Y. Zhang, “Prediction of financial time series with hidden Markov models,” *M.Sc. thesis, Simon Fraser University,* 2004.

[139] E. Chan, Machine trading, Hoboken, New Jersey: John Wiley & Sons, 2017, pp. 101-105.

[140] J. Yang, B. Lin, W. Luk and T. Nahar, “Particle filtering-based maximum Likelihood estimation for financial parameter estimation,” *24th International Conference on Field Programmable Logic and Applications (FPL),* 2014.

[141] V. Dhillon, D. Metcalf and M. Hooper, “The DAO Hacked,” in *Blockchain Enabled Applications*, Berkeley, CA, Apress, 2017, pp. 67-78.

[142] W. Yan and C. Clack, “Evolving robust GP solutions for hedge fund stock selection in emerging markets,” in *Genetic and Evolutionary Computation Conference (GECCO),* 2007.

[143] W. Yan and C. Clack, “Diverse committees vote for dependable profits,” in *Genetic and Evolutionary Computation Conference (GECCO),* 2007.

[144] R. Alghamdi and K. Alfalqi, “A survey of topic modeling in text mining,” *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 6, no. 1, 2015.

[145] M. Linton, E. G. S. Teo, C. Y. Chen and W. K. Härdle, ““Dynamic topic modelling for cryptocurrency community forums”,” *SFB Discussion Paper,* 2016.

[146] A. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” *Biometrika,* vol. 58, no. 1, pp. 83-90, 1971.

[147] L. Adamopoulos, “Cluster models for earthquakes: Regional comparisons,” *Journal of the International Association for Mathematical Geology,* vol. 8, no. 4, pp. 463-475, 1976.

[148] Y. Aït-Sahalia, J. Cacho-Diaz and R. Laeven, “Modeling financial contagion using mutually exciting jump processes,” *Journal of Financial Economics,* vol. 117, no. 3, pp. 585-606, 2015.

[149] S. Yang, A. Liu, J. Chen and A. Hawkes, “Applications of a multivariate Hawkes process to joint modeling of sentiment and market return events,” *Quantitative Finance,* vol. 18, no. 2, pp. 295-310, 2017.

[150] S. Zannettou, T. Caulfield, E. Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini and J. Blackburn, “The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources,” in *Internet Measurement Conference,* London, 2017.

[151] E. L. Lai, D. Moyer, B. Yuan, E. Fox, B. Hunter, A. L. Bertozzi and P. J. Brantingham, “Topic time series analysis of microblogs,” *IMA Journal of Applied Mathematics,* vol. 81, no. 3, 2016.

[152] D. Blei, A. Ng and M. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[153] D. Blei and J. Lafferty, “Dynamic topic models,” *International Conference on Machine learning,* pp. 113-120, 2006.

[154] F. Martin and M. Johnson, “More efficient topic modelling through a noun only approach,” *Australasian Language Technology Association Workshop,* pp. 111-115, 2015.

[155] S. W. Linderman and R. P. Adams, “Discovering Latent network structure in point process data,” *International Conference on Machine learning,* 2014.

[156] J. Maheu and T. McCurdy, “Identifying bull and bear markets in stock returns,” *Journal of Business & Economic Statistics,* vol. 18, no. 1, pp. 100-112, 2000.

[157] L. De Angelis and L. J. Paas, “A dynamic analysis of stock markets using a hidden Markov model,” *Journal of Applied Statistics,* vol. 40, no. 8, pp. 1682-1700, 2013.

[158] S. B. Ramos, J. K. Vermunt and J. G. Dias, “When markets fall down: Are emerging markets all equal?,” *International Journal of Finance and Economics,* vol. 16, pp. 324-338, 2008.

[159] G. M. Constantinides, M. Harris and R. M. Stulz, “Bubbles financial crises and systemic risk,” in *Handbook of the Economics of Finance*, Oxford, Elsevier, 2013, p. 1245.

[160] Y. Rao, Q. Li, X. Mao and L. Wenyin, “Sentiment topic models for social emotion mining,” *Information Sciences,* vol. 266, pp. 90-100, 2014.

[161] C. Lin and Y. He, “Joint sentiment/topic model for sentiment analysis,” *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 375-384, 2009.

[162] A. B. Eliacik and N. Erdogan, “Influential user weighted sentiment analysis on topic based microblogging community,” *Expert Systems with Applications,* vol. 92, pp. 403-418, 2018.

[163] A. Bukhari, U. Qamar and U. Ghazia, “URWF: User Reputation based Weightage Framework for Twitter micropost classification,” *Information Systems and e-Business Management,* vol. 15, no. 3, pp. 623-659, 2017.

[164] A. Bessi and E. Ferrara, “Social bots distort the 2016 US presidential election online discussion,” *First Monday,* vol. 21, no. 11, 2016.

[165] M. Alrubaian, M. Al-Qurishi, M. Al-Rakhami, M. Mehedi and H. A. Alamri, “Reputation-based credibility analysis of Twitter social network users,” *18th IEEE International Conference on Computational Science and Engineering (CSE2015),* vol. 29, no. 7, 2016.

[166] W. Bonefeld and P. Burnham, “Britain and the politics of the European exchange rate mechanism 1990–1992,” *Capital & Class,* vol. 20, no. 3, pp. 5-38, 1996.

[167] E. Shtatland and T. Shtatland, “Another look at low-order autoregressive models in early detection of epidemic outbreaks and explosive behaviors in economic and financial time series,” in *Northeast SAS Users Group Conference (NESUG)*, 2008.