Processing Social Media Text for the Quantamental Analyses of Cryptoasset Time Series

Andrew Peter Burnie

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

February 20, 2020

I, Andrew Peter Burnie, 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 work.

Abstract

This thesis analyses social media text to identify which events and concerns are associated with changes between phases of rising and falling cryptoasset prices.

A new cryptoasset classification system, based on token functionality, highlights Bitcoin as the largest example of a ‘crypto-transaction’ system and Ethereum as the largest example of a ‘crypto-fuel’ system. The price of ether is only weakly correlated with that of bitcoin (Spearman’s rho 0.3849).

Both bitcoin and ether show distinct phases of rising or falling prices and have a large, dedicated social media forum on Reddit. A process is developed to extract events and concerns discussed on social media associated with these different phases of price movement. This innovative data-driven approach circumvents the need to pre-judge social media metrics.

First, a new, non-parametric Data-Driven Phasic Word Identification methodology is developed to find words associated with the phase of declining bitcoin prices in 2017-18. This approach is further developed to find the context of these words, from which topics are inferred. Then, neural networks (word2vec) are applied to evolve analysis from extracting words to extracting topics. Finally, this work enables the development of a framework for identifying which events and concerns are plausible causes of changes between different phases in the ether and bitcoin price series.

Consistent with Bitcoin providing a form of money and Ethereum providing a platform for developing applications, these results show the one-off effect of regulatory bans on bitcoin, and the recurring effects of rival innovations on ether price. The results also suggest the influence of technical traders, captured through market price discourse, on both cryptoassets. This thesis demonstrates the value of a quantamental approach to the analysis of cryptoasset prices.

Impact Statement

The first benefit of this research is to develop a user-friendly cryptoasset classification system based on token functionality. This has been published in the peer-reviewed journal Ledger [40], and formed part of the written evidence submitted to the UK Parliament Digital Currencies Inquiry to inform public policy on cryptoassets, in conjunction with Eversheds Sutherland (International) LLP [100].

The second impact is to quantitatively assess social media discussion forums to identify what events and concerns are associated with major shifts between different phases of price. The benefit of this is that it necessitated the development of new methodologies that recognise the need for non-parametric analyses to quantitatively examine discussion forums. This moves the debate from previous analyses of volume and sentiment to associating changes in price with specific events and concerns. This starts with Data-Driven Phasic Word Identification (DDPWI; see Chapter 5), and then uses word2vec neural networks to evolve from finding 'price dynamic words' to topics (see Chapter 6). It demonstrates the benefits of data derived from social media discussion forums over alternative sources such as web search or Twitter data used in previous studies. Rather than pre-judging potential causes of movement that are then tested, these data-driven approaches discover relevant events and concerns from social media text. These methodologies could be applied to other cryptoassets and more generally to other research areas where there is a time series and a relevant social media text source.

Outside academia the emphasis has been on developing trading algorithms to predict cryptoasset price. These have used prejudged metrics and ignored the inherently phasic nature of the price series, with the possibility that causal effects may vary over time. The impact of this research is that it identifies the limitation of this approach by showing that there are both recurring events and unanticipated, one-off, ‘black swan’ events associated with phasic shifts in price. The results differ between Bitcoin and Ethereum, with the exception of speculation (see Chapter 7), which is consistent with their different token functionality (see Chapter 4).

The impact of the research has been brought about through publications and conference proceedings to international academics at SIGIR and to the FinTech industry (see Section 1.5). This included three peer-reviewed, open-access articles [40,45,46]. The correlation analyses presented at the Cryptocurrency Research Conference 2018 has been cited 9 times [39]. The article on cryptoasset classification has been downloaded 3,024 times [40], DDPWI [45] 776 times and the word2vec topic modelling technique [43] 132 times (all by 22nd January 2020).

Acknowledgements

I would like to express my gratitude to my supervisors Professor Emine Yilmaz and Professor Tomaso Aste for their guidance and critical feedback, which were invaluable in encouraging me to take my ideas to the next level and in determining where the research contained within this thesis might be published for maximum impact. I would also like to extend my gratitude to my examiner in the PhD transfer viva, Professor Suzy Moat, who provided greatly appreciated recommendations and helpful advice in guiding the direction of this research.

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and Turing award number TU/C/000028. I would like to thank The Alan Turing Institute for their financial support as well as for the provision of lectures, workshops and facilities that helped to stimulate the research in this PhD. I would also like to extend this thanks to University College London which provided a fantastic environment in which to pursue a PhD.

I would like to provide a special thanks to my brother, James Burnie, Head of Blockchain and Cryptoassets (United Kingdom) at Eversheds Sutherland (International) LLP. James's enthusiasm for cryptoassets and innovation in financial technology sparked my interest in pursuing research in this exciting field; his expertise was a valuable source of feedback, particularly on the cryptoasset classification; and he was an invaluable liaison with Eversheds Sutherland (International) LLP.

Finally, I would like to thank my parents, Professor James Burnie and Professor Ruth Matthews, for inspiring my lifetime love for science and data that motivated me to pursue this PhD.

Table 1: Abbreviations used in Thesis

Abbreviation	Text
AODE:	Averaged One-dependence Estimators
ARDL:	Autoregressive Distributed Lag
ARIMA:	Auto-Regressive Moving Average
ARIMAX:	Extended version of ARIMA that includes other predictors.
EC:	Empirical Conditional Model
ECM/VECM:	Error Correction Model / Vector Error Correction Model
EEMD:	Ensemble Empirical Mode Decomposition method
ENET:	Elastic-Net regularized regression method
EWMA:	Exponential Weighted Moving Average
GBT:	Gradient Boosted Tree
GDA:	Gaussian Discriminant Analysis
GLM:	Generalised Linear Model
GP:	Gaussian process based regression
HMM:	Hidden Markov Model
ICO:	Initial Coin Offering
LASSO:	Least Absolute Shrinkage and Selection
LDA:	Linear Discriminant Analysis
QDA:	Quadratic Discriminant Analysis
LIWC:	Linguistic Inquiry and Word Count framework
LR/WLR:	Logistic Regression / Weighted Logistic Regression
PCA:	Principal Component Analysis
RF:	Random Forest
STR:	Structured Time Series Model
STRX:	STR plus regression terms on external features similar to ARIMAX
SVM/SVR:	Support Vector Machine / Support Vector Regression
VADER:	Valence Aware Dictionary for sEntiment Reasoning
VAR:	Vector Autoregression
XGT:	Extreme gradient boosting
	Evaluation Metrics
RMSE:	Root Mean Square Error
MAE:	Mean Absolute Error
MAPE:	Mean Absolute Percentage Error
FEVD:	Forecast-Error Variance Decomposition
	Correlation Metrics
PMCC:	Pearson's Product Moment Correlation Coefficient
SR:	Spearman's Rho
KT:	Kendall's Tau
VIF:	Variance Inflation Factor
	Neural Networks
BNN:	Bayesian Neural Networks
CNN:	Convolutional Neural Network
FFN:	Feedforward Neural Network
GASEN:	Genetic Algorithm based Selective Neural Network Ensemble
GRU:	Gated Recurrent Unit
LSTM:	Long Short-Term Memory
RNN:	Recurrent Neural Network
EEMD-ELMAN:	applies EEMD then RNN
RRL:	Recurrent Reinforcement Learning
	Variants on ARCH
ARCH:	Auto-Regressive Conditional Heteroskedasticity
GARCH:	Generalized Auto-Regressive Conditional Heteroskedasticity
EGARCH:	Exponential GARCH
AR-GARCH:	Asymmetric Power GARCH
AR-CGARCH:	Asymmetric Power Component GARCH
BEGARCH:	GARCH but lets conditional log-transformed volatility be dependent on past values of a t-distribution score
	Regulatory Bodies
CFTC:	Commodity Futures Trading Commission
EBA:	European Banking Authority
ESMA:	European Securities and Markets Authority
FCA:	Financial Conduct Authority
FINMA:	Swiss Financial Market Supervisory Authority
SEC:	United States Securities and Exchange Commission

Contents

1 Introduction **18**

1.1 Research Background and Context . . . 18

1.2 Cryptoasset or Cryptocurrency? . . . 24

1.3 Research Objective . . . 25

1.3.1 Delineating the System to be Analysed . . . 25

1.3.2 Characterising the Dataset and Methodology for the Quantitative Analysis . . . 25

1.3.3 Fundamental Analysis . . . 25

1.3.4 Quantitative Analysis of Social Media . . . 25

1.3.5 Comparative Analysis . . . 25

1.4 Thesis Outline . . . 26

1.5 Contributions . . . 29

1.5.1 Publications . . . 29

1.5.2 Other Contributions . . . 33

2 Literature Review **34**

2.1 Cryptoasset Heterogeneity . . . 36

2.1.1 Price Co-Movement . . . 36

2.1.2 Classification . . . 38

2.2 Cryptoasset Selection . . . 41

2.2.1 Highest market capitalisation and liquidity . . . 41

2.2.2 Entity-independent . . . 43

2.2.3 Sufficiently large, publicly available database . . . 44

2.2.4 Tokens are of a different type 45
2.2.5 Selection of Bitcoin and Ethereum 45
2.3 Datasets 47
2.3.1 Internet Activity Volume 47
2.3.2 Sentiment 52
2.3.3 Topic Modelling 57
2.3.4 Limitations to Examining Sentiment and Topics 59
2.3.5 Words 60
2.3.6 Selection of Reddit Data 60
2.4 Methodologies 78
2.4.1 Forecasting the Cryptoasset Price 78
2.4.2 Causal Inference 82
2.4.3 Selection of Analytic Approach 87

3 Cryptoasset Price Co-movement 92

3.1 Introduction 92
3.2 Data Preparation 93
3.2.1 Choice of Cryptoasset 93
3.2.2 Choice of Dataset 93
3.2.3 Preparing the Price Data 94
3.3 Methodology 94
3.3.1 SR and KT Formulae 94
3.3.2 Correlation Networks 95
3.3.3 Software 95
3.4 Results 96
3.4.1 Cryptoassets Selected 96
3.4.2 Correlation Values 97
3.5 Discussion 101

4 Cryptoasset Classification and Analysis of Non-Conventional Fundamentals 103

4.1 Introduction 103
4.2 Methodology 105
4.2.1 Scope 105
4.2.2 Determining Financial Significance 105
4.2.3 Criteria Applied to Each Cryptoasset 106
4.3 Results 106
4.3.1 Dataset 106
4.3.2 Classification 108
4.4 Implications for the Analysis of Fundamentals 113
4.4.1 Determining a 'better' form of money 115
4.4.2 Forks 116
4.4.3 Token Supply 117
4.4.4 Entity Dependence 117
4.5 Limitations to the Analysis of Fundamentals 117
4.6 Discussion 118
4.6.1 Classification 118
4.6.2 Analysis of the Fundamentals for Bitcoin and Ether 121

5 Words Associated with Bitcoin Price Phases 123

5.1 Introduction 123
5.2 Data Preparation 125
5.2.1 Data Sources 125
5.2.2 Engineering Word Frequency Data from Reddit Submissions Text 126
5.3 Methodology 128
5.3.1 Identifying Words by Absolute Frequency 128
5.3.2 Identifying Words by Relative Frequency 128
5.3.3 Context of Price Dynamic Words 129
5.4 Results 130
5.4.1 Reddit Submissions Descriptive Statistics 130
5.4.2 Most Frequent Words by Absolute Frequency 130

5.4.3 Comparing Word Frequencies Across Stages and Identifying Price Dynamic Words . . . 133
5.4.4 Context of Price Dynamic Words . . . 134
5.5 Discussion . . . 141

6 Topics associated with Phasic Shifts in Price 143
6.1 Introduction . . . 143
6.2 Framework . . . 144
6.3 Experiments in Topic Modelling Optimisation . . . 144
6.3.1 Datasets . . . 144
6.3.2 Model Variants . . . 145
6.3.3 Evaluation Metrics . . . 145
6.3.4 Evaluating Model Variants . . . 146
6.4 Interpretation . . . 151
6.5 Discussion . . . 152
6.6 A Comparison of DDPWI with Word2vec-based Topic Modelling . 153

7 Cryptoasset Phasic Shifts and Causality 154
7.1 Introduction . . . 154
7.2 The Causality Framework . . . 155
7.3 Data preparation . . . 158
7.3.1 Dataset . . . 158
7.3.2 Dividing the price series into phases . . . 158
7.3.3 Text preparation . . . 161
7.3.4 Measuring frequency . . . 167
7.4 Methodology . . . 167
7.4.1 Mono-phase analysis . . . 167
7.4.2 Multi-phase analysis . . . 168
7.4.3 Context of concepts . . . 169
7.5 Results . . . 170
7.5.1 Comparison of Bitcoin and Ethereum price phases . . . 170

7.5.2 Mono-phase concepts and their context 173
7.5.3 Multi-phase concepts and their context 177
7.6 Coherence with known facts 182
7.7 Discussion 184
7.7.1 Limitations of Causality Analysis 185
7.7.2 X is a Response to Y 186
7.7.3 X and Y are Symptoms 186
7.7.4 X Causing Y Requires a Catalyst 187

8 Quantamental Analysis of Bitcoin and Ethereum 189
8.1 Comparison of Quantitative Analysis Results 189
8.1.1 Bans 189
8.1.2 US Dollars 190
8.1.3 Tax 191
8.1.4 Is Ethereum Different? 191
8.2 Quantamental Analysis 192
8.2.1 Bitcoin as Money 192
8.2.2 Ethereum for Application Development 193
8.3 Conclusion 194

9 Conclusion 195
9.1 Addressing the Research Questions 195
9.2 Future Work 199

Bibliography 200

List of Figures

1.1 Daily ether and bitcoin price from 1 January 2017 to 14 May 2019 . 19

3.1 Correlation network diagrams depicting the Spearman's rho correlations between different cryptoasset returns. 98

4.1 A fundamental analysis assessment framework for cryptoassets. . . 114

5.1 Daily bitcoin price from 1 January 2017 to 3 December 2018. 124

5.2 Words in at least 5% submissions in all stages 131

5.3 Words in at least 5% submissions in at least one stage but not all . . 132

5.4 Usage and context of the word 'bitcoin' 137

5.5 Usage and context of the word 'ban' 138

5.6 Usage and context of US Dollar mentions 139

5.7 Usage and context of the word 'tax' 140

6.1 Groups Rising in Frequency from Stages 1 to 2. 149

6.2 Groups Falling in Frequency from Stages 1 to 2. 149

7.1 The causality framework 157

7.2 Comparison of ether and bitcoin US Dollar Price Local Extrema . . 160

7.3 Frequency data for mono-phase concepts in the case of Bitcoin . . . 174

List of Tables

1 Abbreviations used in Thesis 8

2.1 Number of subscribers for Reddit subreddits dedicated to a cryptoasset 44

2.2 Summary of why Bitcoin and Ethereum are selected for analysis . . 46

2.3 Internet Activity Literature Review 2013 – 2014. 63

2.4 Internet Activity Literature Review 2015 part 1 (datasets ending in 2014). 64

2.5 Internet Activity Literature Review 2015 part 2 (datasets ending in 2015). 65

2.6 Internet Activity Literature Review 2016. 66

2.7 Internet Activity Literature Review 2017 part 1 (datasets ending before 2017). 67

2.8 Internet Activity Literature Review 2017 part 2 (datasets ending in 2017). 68

2.9 Internet Activity Literature Review 2018 part 1 (datasets ending before May 2017). 69

2.10 Internet Activity Literature Review 2018 part 2 (datasets ending in May or June 2017). 70

2.11 Internet Activity Literature Review 2018 part 3 (datasets ending after June 2017 and before 2018). 71

2.12 Internet Activity Literature Review 2018 part 4 (datasets ending in 2018 and before May 2018). 72

2.13 Internet Activity Literature Review 2018 part 5 (datasets ending in May 2018 and after). 73

2.14 Internet Activity Literature Review 2019-20 part 1 (datasets ending in 2017). 74

2.15 Internet Activity Literature Review 2019-20 part 2 (datasets ending in 2018 and before August). 75

2.16 Internet Activity Literature Review 2019-20 part 3 (datasets ending in or after August 2018 and before 2019). 76

2.17 Internet Activity Literature Review 2019-20 part 4 (datasets ending in 2019 or dates unstated). 77

2.18 Forecast Models Literature Review 2014 – 2016. 88

2.19 Forecast Models Literature Review 2017. 89

2.20 Forecast Models Literature Review 2018 (datasets ending in 2017). . 90

2.21 Forecast Models Literature Review 2018 (datasets ending in 2018). . 91

3.1 Cryptoassets, abbreviations and data availability 96

3.2 Spearman's Rho (SR) correlations between the daily percentage change in prices from 9 November 2017 to 6 March 2018 99

3.3 Spearman's Rho (SR) correlations between the daily percentage change in prices from 9 September 2016 to 6 March 2018 100

4.1 Market capitalisation and liquidity for selected cryptoassets 107

4.2 Amount raised at ICO for selected cryptoassets 107

4.3 Allocation of cryptoassets across the different groups 112

5.1 Descriptive statistics for Reddit Submissions (1 January 2017 to 3 December 2018). 130

5.2 Chain of most frequent words associated with price dynamic words . 135

6.1 Descriptive statistics for Reddit Submissions (1 January 2017 to 22 January 2019). 145

6.2 Grouping 586 Words Falling from Stages 1 to 2 147

6.3 Grouping 129 Words Rising from Stages 1 to 2 147

6.4 Grouping 40 Words Falling from Stages 2 to 3 147

6.5 Grouping 83 Words Rising from Stages 2 to 3 147

6.6 Grouping 63 Words Rising from Stages 3 to 4 147

7.1 Replacing currency code with associated name 162

7.2 Ethereum concepts standardised. 163

7.3 Bitcoin concepts standardised 163

7.4 Cryptoasset concepts standardised. 164

7.5 Finance concepts and nationalities standardised. 165

7.6 Lemmatised and stemmed words standardised. 166

7.7 Phases in the bitcoin and ether price being compared 171

7.8 Word2vec-based topic modelling specifications 172

7.9 Descriptive statistics for phases in the bitcoin and ether price series . 172

7.10 Top five words occurring with each Bitcoin mono-phase concept . . 175

7.11 Top five words occurring with each of Cardano, Tron and Ripple ('rippl') compared with the Bitcoin mono-phase concept 'cardano'/'eo'/'iota'/'rippl'/'stellar'/'tron' 176

7.12 Top ten words occurring with Binance ('binanc') compared with the Bitcoin mono-phase concept 'binanc'/'hitbtc' 176

7.13 Top five words occurring with each Bitcoin falling-price, multi-phase concept 178

7.14 Top five words occurring with each Ethereum rising-price, multi-phase concept 179

7.15 Top five words occurring with each Ethereum falling-price, multi-phase concept 180

7.16 Change in frequency for Bitcoin multi-phase concepts 181

7.17 Change in frequency for Ethereum multi-phase concepts 181

Chapter 1

Introduction

1.1 Research Background and Context

In less than a decade, a single bitcoin token rose from having no price to becoming worth 19,498.68 US Dollars (from launch in 2008 to 16 December 2017) [26]. This increase in value came with an escalating belief in what cryptoassets could achieve; an evolution in purpose encapsulated by the arrival of platforms that facilitate blockchain-supported application development, such as Ethereum [97].

Whilst Bitcoin removed intermediaries to decentralise online payments [211], so as to create 'The Best Money in the World' [53], the remit of blockchain technology has subsequently broadened with its impact being compared with the internet [48]. Specific use-cases that have been explored include those in file storage [178], online voting [82], shareholder rights management [170, 247] and even decentralising the organisation of entire firms [97]. Blockchain technology has been advocated as enabling the automation of regulatory work flows in the movement of physical goods [190], and as providing a means of reducing the administrative burden, raising transparency and enabling automation in shareholder rights management [170, 247].

Enthusiasts attracted to the rising price and potential of cryptoassets face, however, a typically volatile price. There are phases of optimism where prices can rise to a multiple of the initial value and phases of pessimism where prices can fall to less than half the initial value. Figure 1.1 shows how ether prices rose 170-fold in just over a year (1 January 2017 to 13 January 2018) and fell 73% in a few months (13 January to 6 April 2018) [98]. Bitcoin prices rose twenty-fold in less than a year (1 January to 16 December 2017) and fell 65% in just under two months (16 December 2017 to 5 February 2018) [26].

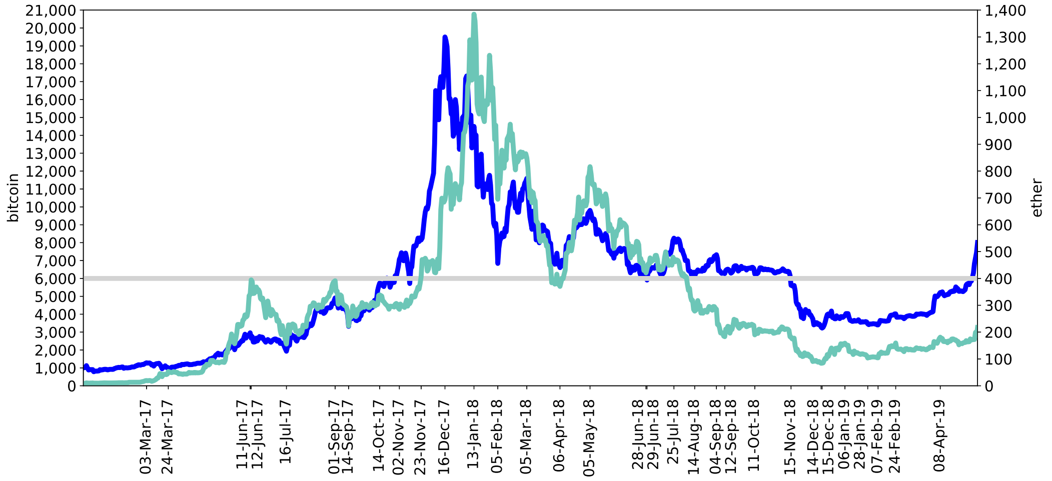


Figure 1.1: US Dollar ether and bitcoin Price from 1 January 2017 to 14 May 2019. Bitcoin series is in blue and the price is given by the left axis. Ether series is in light green and the price is given by the right axis. The horizontal line represents identified support or resistance price levels which were 400 US Dollars for ether and 6000 US Dollars for bitcoin. The labelled dates on the x-axis are dates where there was a bitcoin or ether local maxima or minima, or where the horizontal line was breached. Bitcoin prices sourced from Blockchain Luxembourg S.A. [26] and ether prices from Etherscan [98]

Regulators are concerned that investors may lack information for avoiding the large losses associated with this volatility [102]. The typically decentralised structure of cryptoassets (such as with Bitcoin and Ethereum) means a lack of a well-defined management structure which could be held to account and deficiency of well-balanced reports being issued that are audited by regulated entities. A survey commissioned by the Financial Conduct Authority (FCA) found that, before purchasing, although 50% consumers performed general research on a cryptoasset, only 3% consumers discussed their investment with a financial advisor [101].

By comparison, a retail investor considering purchasing shares in a public company has access to detailed information on the company's financial and operational performance which directors are legally obliged to report on a regular basis. Share-holders can directly ask questions to a well-defined management and can hold this management to account, typically through Annual General Meetings and Extraordinary General Meetings. Investors can make an informed decision often with the help of a qualified financial advisor and analysts' notes. Even in the case of internet start-up companies, where financial statements have been found to explain only about a third of the variation in price/sales ratio [42], there are still financial advisors, analyst notes and the capacity to question a well-defined management.

The problem is compounded by cryptoassets being a 'new asset class' [48] where what drives the price dynamics and so what information might be relevant could be unique. Baur et al [16] showed that not only were Bitcoin returns uncorrelated with traditional asset classes (such as currencies, stocks, bond, commodities) but that no other asset exhibited such weak correlations with other assets across the board. Both Liu and Tsyvinski [183] as well as Burniske and Tatar [48] corroborate this existence of a low correlation between cryptoasset prices and other asset classes.

The main information sources available to the cryptoasset investor are those online. These include sources provided by the cryptoasset's developers such as the cryptoasset's website and whitepapers published to explain the cryptoasset, as well as any third-party explanations or reviews (such as those available on `www.coindesk.com` or `blockgeeks.com`). Whilst these may help to build an understanding of the cryptoasset, investors will also need regular updates as to key events that may affect the cryptoasset price. An FCA-commissioned survey found social media to be the most popular such source of news and information. In-depth interviews revealed how this was motivated by a distrust of mainstream media, which was perceived as having an 'agenda' due to its link with the 'establishment' [241]. Social media forums both contain updated posts on the cryptoasset and also enable the investor to post questions to the wider community when a concern arises or to help further improve the investor's understanding of the cryptoasset.

Hence, the investor in cryptoassets is presented with a variety of sources providing information, of which only a portion of this information might be directly relevant to the price. Of the daily submissions being posted on social media, only a few, if any, may actually influence traders to buy or sell. There is thus a need to filter the information available to extract those insights that are most relevant to informing cryptoassets investors in avoiding large losses.

One approach would be to train a model to predict the future change in the cryptoasset price. This would require data on the historic price and other variables that are felt to be predictive of the future price. Data could be extracted from providers such as Blockchain Luxembourg S.A. [26] and Etherscan [99]. A modelling framework would then be selected to link the predictors with the future price such as a neural network or random forest. The parameters of the model could then be tuned to minimise the forecast error. Holders of cryptoassets could then seek to avoid large losses by selling their holdings whenever a drop in the cryptoasset price is predicted.

The problem with such an approach is the key assumption that the future is like the past [271]. The tuning of the parameter values in the model would have to be conducted using historic data, and so the future data would have to be similar to the historic data for the trained model to be reliable going forward. This would not be the case if a variable that drove price in the historic data no longer had an effect on price or, more generally, if the relationships between the predictors and price had substantially changed across time.

This suggests that any reliable predictive modelling of price would need an understanding of what features had a robust association with price and the nature of such associations. Bengio et al [20] determined that this meant finding what are the true cause-effect relationships. This could potentially guide the creation of forecasting models that are more accurate when presented with new data or, at least, provide information as to the limitations forecasting models face when applied to cryptoassets. Knowing what events or concerns are relevant to price may also help investors in deciding whether the occurrence of an event, possibly reported on social media, is something that can be safely ignored or is of a nature as to justify a concern for imminent losses.

Hence, this thesis explores what causes the prices of cryptoassets to change across time. The thesis first considers the theoretical attributes of the cryptoasset token that provide a justification for the token's non-zero value – what this thesis terms the 'fundamentals' of the asset. For assets such as bonds and equity, holding the asset generates a cash return which can be used to price the asset by calculating the net present value of the expected returns [124]. Cryptoassets typically lack such returns and so there is a need to look for more non-conventional fundamentals. This involves developing a classification of different cryptoasset types and then deriving what benefits a participant receives from holding each type of cryptoasset other than profit from an increase in price.

Changes in the fundamentals present a theoretical cause of observed price fluctuations across time. However, what in theory might affect price may not have an effect in practice, particularly when the influence of fundamentals on price might be small compared with that of speculation (discussed further in Section 4.5). Hence, there is a need for quantitative analysis to find events and concerns that occurred in association with changes in the price movement. These provide empirically-supported potential causes of price movement that help to check the relevance of the identified, theoretical fundamentals.

The quantitative analyses applied evolve from finding words associated with a phase in the bitcoin price to extracting plausible causes of bitcoin and ether price movements. This evolution uses word2vec neural networks to change from identifying words to topics, and criteria derived from healthcare epidemiology literature to develop a framework for considering causality. These analyses explore the role of social media forums as a data-source. Social media is the most popular source of news and information on cryptoassets [241], with text available from large Reddit subreddits that are dedicated to specific cryptoassets. These include a subreddit on Bitcoin with over one million subscribers (`https://reddit.com/r/bitcoin`) and a subreddit on Ethereum with over 400,000 subscribers (`https://www.reddit.com/r/ethereum`) (see Section 2.3.6 in Chapter 2).

Following epidemiology literature, observational data can provide evidence that favours a causal link between two variables over other plausible types of relationship [223, 251], but observational data cannot prove a causal connect between two variables. For example, a variable and price may be found to move together across time but this could be because a third, unknown factor varied across the dataset causing both the variable and the price to change in a way that generated the observed co-movement [223]. Hence, whilst quantitative analyses help to evaluate the practical relevance of identified fundamentals, knowing what the theoretical fundamentals are also helps in evaluating if the results identified by the quantitative analyses 'make sense' [10] and are 'plausible' [35].

Hence, overall, this thesis pursues a 'quantamental' strategy that applies quantitative analyses to social media data ('quant-'), considers the theoretical fundamentals underpinning price ('-amental') and compares the results. This improves our understanding of what were the most plausible causes of cryptoasset price variation across 2017-18. The terminology 'quantamental' follows that used in the equity investing literature [10, 38] which recommends [10] a 'quantamental' approach in combining the best insights from both an analysis of the data (quantitative analysis) and a consideration of the fundamentals.

1.2 Cryptoasset or Cryptocurrency?

This thesis examines tokens that are: an entirely digital store of value, publicly available and supported by a blockchain. Historically such assets have been referred to as ‘cryptocurrencies’ [48], as these systems typically sought to provide a form of currency [65, 211]. There has recently been a shift to describing such assets as ‘cryptoassets’ instead. Burniske and Tatar criticised the term ‘cryptocurrency’ for failing to capture the potential of these assets. In their view, ‘currency’ captures only one use out of a spectrum of applications being examined, with some blockchain-supported assets being launched that are not intended for use as a currency [48]. Central banks have also criticised the term ‘cryptocurrency’ for exaggerating the potential of blockchain-supported assets, as, in their view, assets such as bitcoin function poorly as a form of money [52]. Both view-points agree that the term ‘cryptoasset’ is less misleading than ‘cryptocurrency’ as terminology, and so this thesis will use the term ‘cryptoasset’ throughout.

1.3 Research Objective

This is to increase our knowledge of what determines the value of cryptoassets across time. This thesis meets this research objective by asking a series of research questions.

1.3.1 Delineating the System to be Analysed

1. Should cryptoasset price series be analysed individually or in aggregate?

2. Which cryptoassets are to be analysed?

1.3.2 Characterising the Dataset and Methodology for the Quantitative Analysis

3. What social media data are to be used?

4. What analytic approach is to be applied?

1.3.3 Fundamental Analysis

5. What benefit does a participant receive from holding a cryptoasset token and how might this influence the value of the token?

1.3.4 Quantitative Analysis of Social Media

6. What words were associated with the phase of volatile but overall falling bitcoin prices 2017-18?

7. How can we evolve the results from words associated with phases to topics associated with phasic shifts in the bitcoin price?

8. How can we evolve the analysis to find potential causes of phasic shifts in the bitcoin and ether price?

1.3.5 Comparative Analysis

9. How do the results for Bitcoin and Ethereum compare? Are the insights for each cryptoasset shared or unique?

1.4 Thesis Outline

This section summarises what content is in which chapter of the thesis, and explains which sections answer the different research questions delineated in Section 1.3.

Chapter 2 reviews the related literature and justifies the selection of Bitcoin and Ethereum for analysis:

- In Section 2.1, previous studies that explored cryptoasset heterogeneity from a quantitative and qualitative perspective (relevant to research question 1) are first examined, with the identified gaps in this literature being addressed in Chapters 3 and 4:
  - Section 2.1.1 examines quantitative heterogeneity, reviewing previous studies that examined the association between different cryptoasset prices.
  - Section 2.1.2 examines the qualitative heterogeneity, detailing previous classifications of cryptoassets.
- Section 2.2 justifies why Bitcoin and Ethereum are selected for the quantitative analyses (research question 2). Such analyses require a dataset and a methodology.
- Section 2.3 critically reviews the literature on different internet metrics that have been used to understand cryptoasset price variation. The dataset to be analysed is then specified (research question 3).
- Section 2.4 reviews the different methodologies that could be applied in the quantitative analyses. This examines the utility of forecasting models and considers the different causal inference methodologies that have been applied to cryptoasset price data and in other research areas, such as equity markets. This is used to justify the characteristics of the quantitative analytic methodologies applied (research question 4).

The next two chapters continue the preparation process, meeting those deficiencies identified in the literature in answering whether cryptoassets should be analysed as a group or separately (research question 1):

- **Chapter 3** analyses the correlations between cryptoasset prices that supports cryptoassets being heterogeneous in the movement of prices across time. This also supports comparing Bitcoin with Ethereum.
- **Chapter 4** develops a cryptoasset classification that further supports heterogeneity among cryptoassets and, in particular, a distinctiveness between Bitcoin and Ethereum.

Chapter 4 also provides an assessment framework based on the cryptoasset classification. This explains the implications of the classification regarding which variables may affect the prices of different types of cryptoasset, based on the benefit from holding cryptoasset tokens (research question 5). These are the fundamentals against which the results of the quantitative analyses are compared.

The next chapters detail separate studies that quantitatively analyse the link between social media discussions and the cryptoasset price:

- **Chapter 5** examines what words were associated with the phase of volatile but overall falling bitcoin prices 2017-18 (research question 6).
- **Chapter 6** evolves analysing words into analysing topics (research question 7).
- **Chapter 7** builds on these publications to establish the potential causes of phasic shifts in the bitcoin and ether price (research question 8). This finds:
  - plausible causes of a single phasic shift in price
  - plausible causes of rising prices
  - plausible causes of falling prices

Chapter 8 provides the comparative analysis that connects the results of the quantitative analyses with the theoretical fundamentals to determine what are the plausible causes of phasic shifts in cryptoasset prices, comparing Bitcoin with Ethereum (research question 9).

Chapter 9 concludes by describing how the thesis addresses each of the research questions and future work.

1.5 Contributions

1.5.1 Publications

1. **Andrew Burnie, James Burnie, and Andrew Henderson. Developing a Cryptocurrency Assessment Framework: Function over Form. *Ledger*, 3, July 2018. Available at: `https://doi.org/10.5195/ledger.2018.121`.**

    - A new cryptoasset classification is developed that splits different cryptoassets into three types (‘crypto-transaction’, ‘crypto-fuel’ and ‘crypto-voucher’) where the constituent cryptoassets for each type have tokens that share common functionality.
    - The functionality of a type of token implies benefits that the holder receives from owning that type of token. Hence, this publication also specifies an assessment framework that details what the implied benefits are so that the fundamentals underpinning the prices of different types of cryptoasset can be delineated.
    - This publication is discussed further in Chapter 4.

2. **Andrew Burnie. Exploring the Interconnectedness of Cryptocurrencies using Correlation Networks. In *Cryptocurrency Research Conference 2018*. Anglia Ruskin University, Cambridge, UK, 24 May 2018. Available at: `https://arxiv.org/abs/1806.06632`.**

    - The co-movement in cryptoasset prices is examined and displayed as a correlation network, suggesting differences in the price movement across different cryptoassets.
    - This publication is discussed further in Chapter 3.

3. **Andrew Burnie, Andrew Henderson, and James Burnie. Putting Names to Things: Reconciling Cryptocurrency Heterogeneity and Regulatory Continuity. *Journal of International Banking and Financial Law (JIBFL)*, 33(2): 83-86, February 2018.**

    - One of the criteria used in motivating the selection of Bitcoin and Ethereum for analysis is that the cryptoasset is not ‘entity-dependent’. This publication defines this concept.
    - This is discussed in Section 2.2.2

4. **Andrew Burnie and Emine Yilmaz. Social media and Bitcoin Metrics: Which Words Matter. *Royal Society Open Science*, 6, 2019. Available at: `https://doi.org/10.1098/rsos.191068`.**

   - The new Data-Driven Phasic Word Identification methodology is proposed and applied in a non-parametric, statistical analysis of social media discussions. This extracts what words are associated with the phase in 2017-18 bitcoin price time series when prices were falling overall but also particularly volatile.
   - New approaches are also developed to determine the context of the extracted words, by identifying the words used with that word as well as sentiment.
   - This requires a new word frequency dataset that is publicly available [44].
   - This publication is discussed further in Chapter 5.

5. **Andrew Burnie and Emine Yilmaz. An Analysis of the Change in Discussions on Social Media with Bitcoin Price. In *42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France, 21-25 July 2019. Available at: `https://doi.org/10.1145/3331184.3331304`.**

   - A topic modelling methodology is developed based on neural networks (word2vec). This enables finding topics associated with phasic shifts in the bitcoin price.
   - This publication is discussed further in Chapter 6.

6. **Andrew Burnie, Emine Yilmaz, and Tomaso Aste. Analysing Social Media Forums to Discover Potential Causes of Phasic Shifts in Cryptocurrency Price Series. *Frontiers in Blockchain*, 3:1, 2020. Available at: `https://doi.org/10.3389/fbloc.2020.00001`**

   - A new causality framework is developed that discovers potential causes of rising prices, falling prices and a major phasic shift in the cryptoasset price movement.
   - The results for Bitcoin are compared with Ethereum to answer whether the insights for one cryptoasset are unique to that system or shared across cryptoassets.
   - This requires a new topic frequency dataset that is publicly available [47].
   - This publication is discussed further in Chapter 7.

7. **Eversheds Sutherland (International) LLP. Eversheds Sutherland (International) LLP - written evidence. *UK Parliament Treasury Committee Digital Currencies Inquiry*, May 2018, DGC0020, Available at: `http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/treasury-committee/digital-currencies/written/81375.pdf`.**

   - The cryptoasset classification described in the Ledger publication was included as part of submitted evidence to the UK Parliament Treasury Committee's Digital Currencies Inquiry.
   - The classification is presented in Chapter 4.

1.5.2 Other Contributions

1. **Andrew Burnie, Andrew Henderson and James Burnie. ICOs and Cryptocurrency. In *3rd annual Eversheds Sutherland Digital Financial Services and Fintech Conference*. London, UK, 21 November 2017.**

    - Andrew Burnie discussed the history of cryptoassets, the rationale for holding cryptoassets Bitcoin and Ethereum and why cryptoassets are launched by Initial Coin Offering (ICO).
    - The research for this talk helped to inform the cryptoasset classification discussed in Chapter 4.

2. The literature review covering studies associating measures of online activity with variations in bitcoin and ether prices (see Section 2.3)

3. The literature review of causal inference methodologies (see Section 2.4).

4. The comparison of results from quantitative and fundamental analyses in identifying events that are best supported as causing phasic shifts in the cryptoasset price movement (see Chapter 8).

Chapter 2

Literature Review

The Literature Review examines the extent the existing literature can be used to address the research questions stated in Section 1.3. This describes the state of the art, its limitations and the opportunities for research that this thesis will subsequently examine.

The insights generated are of relevance to all the research questions presented in Section 1.3 but are of greatest relevance to research questions 1 to 4. The literature is used to directly answer research questions 2, 3 and 4 and contributes to addressing research question 1, which requires further studies, detailed in Chapters 3 and 4. Determining the dataset and analytic approach (questions 3 and 4) informs the quantitative analyses detailed in Chapters 5 to 7 (addressing questions 6 to 8). The new classification created to address research question 1 also informs the fundamental analyses used to address question 5. The fundamental and quantitative analyses results together inform the comparative analyses in Chapter 8 that examines question 9.

The layout of the current chapter is subsequently detailed, relating each section to the specific research question being addressed.

As there were 1350 cryptoassets [66] on 18 December 2017, research question 1 was whether to analyse cryptoassets together as a single unit or to analyse each cryptoasset individually. This question is approached in Section 2.1 from a quantitative and qualitative perspective. Section 2.1.1 reviews studies on the comovement between different cryptoasset prices which provides the quantitative perspective. Section 2.1.2 examines the qualitative perspective by reviewing the different classification systems that have been proposed. These reviews inform two subsequent studies conducted in this thesis: Section 2.1.1 influences how the correlations between cryptoasset prices are analysed in Chapter 3 while Section 2.1.2 substantiates the need for the new cryptoasset classification developed in Chapter 4.

These studies supported the finding that cryptoassets largely acted as heterogenous entities and this raised the question of which cryptoassets to select for further analysis (research question 2). Section 2.2 uses four criteria for selection. Applying these criteria to the literature supports studying Bitcoin and Ethereum further. The subsequent literature review focusses on these two cryptoassets.

Research question 3 asked what data should be used and question 4 what methodology should be applied to the data. This thesis innovates in two areas: in the choice of internet metrics and in the choice of methodology in linking changes in these metrics with variations in the cryptoasset price. Literature reviews are thus conducted examining the dataset and methodology separately.

Section 2.3 examines previous literature to understand what internet metrics have been used to analyse the cryptoasset price. This starts with relatively simple metrics, specifically the volume of internet activity (Section 2.3.1) which includes measures such as number of Google searches and Wikipedia page views relevant to a cryptoasset. The evolution in the literature is then followed towards the more involved considerations of sentiment (Section 2.3.2) and topic popularity (Section 2.3.3). The limitations of these studies is that they provide limited information and often rely on personal judgement (Sections 2.3.1.5 and 2.3.2). Section 2.3.6 explains the choice of a Reddit submissions dataset and what information is extracted from this text (answering research question 3). This is followed by Tables 2.3 to 2.17 that summarise the associated literature.

Section 2.4 takes a different perspective – examining the methodologies that could be applied to the dataset (addressing research question 4). This begins with forecast models (Section 2.4.1), providing examples in the cryptoasset literature which are then summarised in Tables 2.18 to 2.21. Forecast models are insufficient in elucidating what is associated with the cryptoasset price. A broad perspective is then applied in reviewing causal inference methodologies. This includes papers from areas such as Earth Sciences [254] and equity market research [27, 28, 195]. The published literature is focussed on testing rather than extracting features that potentially caused subsequent fluctuations in an analysed variable.

Having considered the limitations in existing methodologies detailed in the literature, Section 2.4.3 explains the analytic approach applied in this thesis. This thesis will extract words and groups of similar-meaning words from social media data associated with the different phases in the cryptoasset price movement, and will then explore causality.

2.1 Cryptoasset Heterogeneity

2.1.1 Price Co-Movement

This section examines cryptoasset heterogeneity from a quantitative perspective – reviewing studies that have analysed the extent the prices of different cryptoassets move together across time.

Past studies frequently tested for linear associations between cryptoasset prices [17, 61] or left unspecified whether the 'correlation' metric used was non-parametric [94, 111]. The most extensive of these studies [61], studying data from 2013-16, found, using time-series analyses, that the Bitcoin-altcoin price relationship was significantly stronger in the short-run than the long-run; in the long-run, macro-financial indicators (oil price, gold price, NASDAQ Composite and the 10-Year Treasury Constant Maturity Rate) determined the altcoin price more than Bitcoin did. The authors make the point that given the dominant position of Bitcoin, it would be expected that it would be both the preferred medium of exchange and the preferred investment asset [61]. The emergence of this 'winner-take-all' dynamic was also observed by Gandal and Halaburda as the market matured between 2013 and 2014 [111].

An alternative approach is to use transfer entropy to measure the association between cryptoasset prices. Transfer entropy is 'model-free' in the types of relationship tested [195]. Whilst entropy measures the uncertainty in cryptoasset prices [224], transfer entropy evaluates whether one cryptoasset price depends on another [256].

Dimpfl and Peter developed a 'group transfer entropy' [84] approach to evaluate whether cryptoasset prices were dependent on the other cryptoasset prices in a group. Simulation data supported that group transfer entropy had value over linear modelling in capturing non-linear relationships and in being more robust to extreme outliers [84]. Applying group transfer entropy to cryptoasset price data suggested that all cryptoasset prices were inter-related, a result that could not be replicated through linear modelling. This result was robust to removing outlier data around the dates of Bitcoin forks. Applying transfer entropy to the residuals of the linear models supported the relationships existing between different cryptoasset prices being non-linear. Hence, the results of Dimpfl and Peter suggested a non-linear relationship between cryptoasset prices. The strength of the associations was left unclear.

In Chapter 3, Spearman's Rho (SR) and Kendall's Tau (KT) will be applied as non-parametric correlation measures capable of measuring the strength of the monotonic relationship between different cryptoasset prices [299]. The use of correlation metrics has a further advantage over transfer entropy in not requiring the data to be discretised into bins of a few different possible values [256].

The sample of cryptoassets analysed in Chapter 3 includes only the most financially important cryptoassets. This compares with Osterrieder et al [219] who excluded Ethereum from their sample and Aste [9] who applied KT correlation to compare price series across 1944 cryptoassets. Aste [9] found an average correlation value of 0.4 and, on average, a cryptoasset was significantly correlated with 300.7 other cryptoassets. Focussing the sample of cryptoassets on only the most important and displaying the results in a correlation network both help in enabling a more detailed comparison of the correlation values between the different cryptoasset prices. This correlation data will be used to justify whether to analyse cryptoassets separately or in aggregate (research question 1).

2.1.2 Classification

This section analyses cryptoasset heterogeneity from a qualitative perspective by reviewing previous cryptoasset classifications that have been proposed.

One approach is to adopt an existing regulatory framework, such as those provided by the European Banking Authority (EBA), European Securities and Markets Authority (ESMA), Financial Conduct Authority (FCA), Swiss Financial Market Supervisory Authority (FINMA) or the United States Securities and Exchange Commission (SEC). Such frameworks match cryptoassets against existing regulation to determine which systems should be subject to varying degrees of regulator oversight. Those tokens deemed subject to more regulatory oversight have been referred to as 'investment' (ESMA and EBA), 'security' (FCA and SEC) or 'asset' (FINMA) tokens, compared with the less regulatory oversight applicable to 'utility' tokens (SEC, ESMA, EBA and FCA) and 'exchange' (FCA) or 'payment' (ESMA, EBA and FINMA) tokens, which provide a 'means of exchange' (FCA) [73,92,102,107,257].

Applying a regulatory framework faces the practical issue of deciding which regulatory body's framework to use. The EBA, ESMA, FCA, FINMA and SEC approaches varied both in the terms used and in how to define the investment/security/asset token [92,102,107,257]. Even within the one country different regulators may exist who propose different classifications. For example, in the US, the SEC split cryptoassets between lightly regulated 'utility' tokens and 'security' tokens, where there was a 'reasonable expectation of profits to be derived from the entrepreneurial or managerial efforts of others' [73]. The US Commodity Futures Trading Commission's (CFTC) instead split all virtual tokens into commodities or derivatives [136].

This thesis is primarily interested in cryptoasset heterogeneity from the perspective of whether what causes the price to change is the same or fundamentally different across cryptoassets. Those considerations that are important from a regulation perspective may not match price-relevant issues. For example, the FCA grouped ether with bitcoin as exchange tokens as both face similar regulatory treatment [102]. That Ethereum was designed for more than payments [97], unlike Bitcoin [211], was not relevant to determining the comparative regulatory treatment of these cryptoassets. This is similarly an issue for the subset of classifications designed to guide developers seeking to optimally use distributed ledger technology; a review of such classifications is provided by Ballandies et al [13].

Numerous conflicting cryptoasset classifications have been proposed as alternatives to a regulatory framework. An early approach was the 'ontological' classification proposed by Herbert and Stabauer [133] in 2016. This was substantiated by actual cryptoassets, but was disadvantaged in not allowing for the proliferation of new token types since that study [24] and recent developments fundamental to most new financially significant cryptoassets, e.g. the use of ICOs. Only three (Bitcoin, Ethereum and Ripple) of the cryptoassets covered remained in the top ten financially important by 2018. The more recently proposed classifications are typically limited in not clearly detailing how the classification was developed [48, 50, 80, 128, 184, 263, 270, 284], and/or how investors might use the classification [48, 50, 80, 196, 263, 270].

A particularly influential 'taxonomy' intended for investors was that of Burniske and Tatar who popularised the word 'cryptoasset' [48]. It was cited by CryptoCompare [196] in their classification, which considered a breadth of attributes across regulation, industry classification, rationales for holding tokens and 'economic value drivers.' Burniske and Tatar differentiated cryptoasset types according to what was being provisioned, using the taxonomy to illustrate how cryptoassets had diversified from 'cryptocurrencies' that did not provide a resource (including Bitcoin) to a universe of 'cryptoassets' that provided 'raw digital resources' ('cryptocommodities') or finished products ('cryptotokens') [48]. This taxonomy relies on subjectivity on the part of the user to differentiate between whether the digital resource was closer to a raw resource or finished good/service. Furthermore, the taxonomy was not based on the intrinsic characteristics of the token and so the relative benefit for the investor of owning different cryptoasset types is unclear.

A new classification is developed in this thesis that compares cryptoassets with each other reflecting the uniqueness of cryptoassets as a 'new asset class' [16, 48, 183]. This defines three different categories to which a cryptoasset can belong rather than just specifying important issues that cryptoasset investors should consider [128, 280]. The classification focusses on the characteristics of the tokens being bought or sold which is particularly relevant to buyers or sellers of the cryptoasset. This compares with prior considerations of regulatory treatment [73, 92, 102, 107, 257], the digital resource provisioned [48], the market sold to [171] or the technology stack [13]. This classification will be generated by comparing a defined group of cryptoassets using specified criteria. The results will be contrasted against both the classifications from Burniske and Tatar and CryptoCompare to assist in understanding the value and distinctiveness of the new system (see Chapter 4).

The characteristics of a token implies benefits for the holder of that token that may vary across time and so which may affect the price. The classification is thus used to inform an analysis of the fundamentals for each cryptoasset type that specifies the theoretical causes of price variation.

2.2 Cryptoasset Selection

Chapters 3 and 4 support variation in movement of different cryptoasset prices across time (Chapter 3) and variation in the characteristics of the tokens of different cryptoassets (Chapter 4). Cryptoassets are thus selected to be analysed individually. The final choice of which cryptoasset should be examined was made on the basis of four criteria using information derived initially from the review of published literature, websites and cryptoasset whitepapers and supplemented by the analyses performed in Chapter 4. The criteria are:

1. Consistently in the top ten by market capitalisation and liquidity (see Section 2.2.1).
2. Entity-independent (see Section 2.2.2).
3. There exists a sufficiently large, publicly available database during the time period examined to make statistical analyses feasible (see Section 2.2.3).
4. The tokens are of a different type (see Section 2.2.4 and Chapter 4).

2.2.1 Highest market capitalisation and liquidity

Three metrics for comparing the size of cryptoassets are widely available publicly: price, market capitalisation, circulating supply and liquidity [63, 64, 67–69, 105, 106]. Comparing cryptoassets according to price can be misleading because if the supply of cryptoasset tokens is low, buyers may offer a high price for tokens even if the actual use of the cryptoasset is limited. Instead of price, market capitalisation and liquidity are used to compare cryptoassets.

Market capitalisation is the price of a token multiplied by the circulating supply of tokens. Circulating supply deducts from total supply publicly unavailable tokens. This metric is used because it directly measures the value of investments held by the general public in the cryptoasset, and so focusses analyses on the most financially significant cryptoassets.

For some cryptoassets, a proportion of once publicly accessible tokens may have become, in practice, inaccessible. This could happen if owners lose access to their wallets or if the tokens of the cryptoasset are being hoarded [282]. This could lead to market capitalisation giving a misleading impression of the true amount being invested in a cryptoasset.

Liquidity is thus also considered, as measured by transaction volume over the last 24 hours. Low liquidity is used to indicate that token inaccessibility was more of an issue. The fewer tokens that are for sale, the lower the transaction volume is likely to be. Liquidity is further important because a lack of liquidity implies that traders can only buy and sell the cryptoasset slowly and at great cost, inhibiting adoption [282, 296].

Analyses focus on the cryptoassets with the highest market capitalisation and liquidity. This is because the prices of smaller cryptoassets are likely to be more volatile and shaped by random noise.

The price of a cryptoasset with a smaller market capitalisation is more susceptible to price manipulation [48]. As more of the cryptoasset can be bought at a lower price, less money is required to buy a large proportion of the cryptoasset. A price manipulator could combine such a purchase with hype on social media to generate the illusion of enthusiasm for the asset. This may cause others to buy, pushing the price up further, at which point the price manipulator sells at a profit. Such pump-and-dump schemes have been described as ‘common’ [48] among smaller cryptoassets and may act as a source of unpredictable, random noise.

A small userbase also suggests fewer buyers and sellers at any given point in time. This means that buyers will likely need to increase prices more to induce sufficient supply and sellers will need to reduce prices more to encourage sufficient demand. This induces greater price volatility over time.

A smaller cryptoasset is further likely to be listed on fewer exchanges. Exchanges have a lower incentive to offer a cryptoasset for purchase if only a small number of tokens are bought or sold at any one time. This suggests that smaller cryptoassets will be more dependent on specific exchanges, and so the price of the cryptoasset will be more influenced by difficulties faced by a single exchange. This could include that exchange being hacked, going into administration or trading ceasing because of technological difficulties faced by the exchange.

Three websites were examined at multiple timepoints to obtain data on market capitalisation and liquidity: coinmarketcap.com at 14:27 on 4 October 2017 [68], 15:48 on 30 October 2017 [69], and 10:27 on 18 December 2017 [67]; coincap.io at 15:58 on 30 October 2017 [63] and 10:28 on 18 December 2017 [64]; and onchainfx.com at 15:58 on 30 October 2017 [105] and 10:28 on 18 December 2017 [106]. The websites coincap.io and onchainfx.com were corroborated by coinmarketcap.com except onchainfx.com excluded Tether from its rankings. The rankings were then updated at 20:40 on 17 January 2019, using coinmarketcap.com [70]. Where rankings were inconsistent, only cryptoassets in both lists were considered.

This identified the following five cryptoassets which consistently had the largest market capitalisations and liquidity: Bitcoin, Ethereum, Bitcoin Cash, Ripple and Litecoin. Considering data sourced at 10:27 on 18 December 2017 from coinmarketcap.com [67], Bitcoin had over half of the market share of cryptoassets (about 54%) and Ethereum had the second largest market share (about 12%), which is over twice as high as that of Bitcoin Cash (about 5%). Bitcoin had the highest liquidity in terms of transaction volume (over 13 billion US Dollars) with Ethereum coming second (over 2 billion US Dollars), which was substantially higher than Litecoin (1.2 billion US Dollars).

2.2.2 Entity-independent

Entity-dependence relates to when a cryptoasset system is reliant on a small number of operators to function [41]. Because the price of tokens in such systems is likely to depend on the entities the system depends on, these price series are likely to behave distinctively and so this thesis avoids systems with a clear entity-dependence.

The US Securities and Exchange Commission (SEC) saw entity-dependence as important from the perspective that it would not be ‘meaningful’ [139] to determine that a specific entity should issue disclosures if a network were truly decentralised, and so the tokens in such systems should not be labelled as securities [139]. The SEC specified that both Bitcoin and Ethereum were examples of decentralised networks, with Bitcoin having been so ‘perhaps from inception’ [139]. Bitcoin, the classical cryptoasset, was launched as an entity independent form of currency: a 'peer-to-peer version of electronic cash' that would enable online payments without the need for intermediates or the oversight of a central bank [211].

Ripple is excluded from analysis because Ripple is entity-dependent on Ripple Labs [294]. Of all ripple, 61% is owned by Ripple Labs [294]. This has been placed under escrow, but Ripple Labs still receives 1% of total ripple per month, over which it has full discretion [294]. If Ripple Labs decides to sell its accumulated supply, this could skew the price of ripple significantly. Such discrete decisions to sell are likely to be difficult to model.

This dependence of Ripple on Ripple Labs has led to concerns as to whether Ripple is fact a security. The SEC has cautioned that exchanges could face penalties for listing unregistered securities. As a result, the major exchanges (Coinbase and Gemini) do not list Ripple [197]. Hence, Ripple was also excluded because its tokens are not as publicly available as Bitcoin and Ethereum.

2.2.3 Sufficiently large, publicly available database

Bitcoin, Ethereum, Litecoin and Bitcoin Cash have large associated discussion forums that could be used in social media analysis. Table 2.1 shows that these include subreddit forums on Reddit that have thousands of subscribers.

Table 2.1: Number of subscribers for Reddit subreddits dedicated to a cryptoasset, as taken from the subreddit website at 11:11 (GMT) 15 January 2020. The subreddit can be found by appending the provided subreddit name to the end of the URL `https://reddit.com/`.

	subreddit	Number of Subscribers
Bitcoin	`r/bitcoin`	1,248,690
Ethereum	`r/ethereum`	449,569
Litecoin	`r/litecoin`	210,340
Ripple	`r/Ripple`	208,575
Bitcoin Cash	`r/Bitcoincash`	47,723

2.2.4 Tokens are of a different type

Both Litecoin and Bitcoin Cash were created as forks of the Bitcoin codebase [286], resulting in similarities in these cryptoassets that are reflected in classifications grouping these three cryptoassets into the one type. CrytoCompare described all three systems as having ‘payment’ tokens [196], whilst Burniske and Tatar saw Bitcoin and Litecoin as both examples of true ‘cryptocurrencies’ [48]. The classification developed in Chapter 4 specifically finds the functionality of the tokens of these three cryptoassets to be sufficiently similar to justify referring to them all as ‘crypto-transaction’ systems.

In contrast, Ethereum was not launched as a fork of Bitcoin [286] so as to enable extensions to the functionality of the ether token [97]. This results in ether tokens being seen as distinct from Bitcoin, Litecoin and Bitcoin Cash, with ether tokens being labelled alternatively as ‘crypto-fuel’ (Chapter 4), ‘utility’ (CryptoCompare [196]) or ‘cryptocommodity’ (Burniske and Tatar [48]) tokens.

Hence, Ethereum is selected for comparison with Bitcoin on the basis of the distinctiveness of its tokens. How Ethereum and other ‘crypto-fuel’ systems differ from ‘crypto-transaction’ systems such as Bitcoin, Litecoin and Bitcoin Cash is discussed further in Chapter 4.

2.2.5 Selection of Bitcoin and Ethereum

The conclusion from a review of the literature, websites and whitepapers was that this thesis will examine the features associated with the valuations of bitcoin and ether (Ethereum’s token).

The thesis will initially focus on Bitcoin (the largest cryptoasset) to develop and refine the methodologies. Bitcoin had 4.5 times the market capitalisation and 6 times the liquidity of the next largest cryptoasset (Table 2.2). The bitcoin price series also followed three distinct phases of movement across 2017-18 that enabled comparison of the central phase of overall falling, but volatile, prices with before and after (discussed further in Section 5.1).

Ethereum is then used as a comparator to understand the extent identified potential causes of phasic shifts in cryptoasset prices are shared across different cryptoassets. Ethereum has tokens with a distinct functionality compared with Bitcoin. The market capitalisation and liquidity for Ethereum is also second to Bitcoin and about twice that of the next largest cryptoasset (Table 2.2). Ethereum has also been found to be entity-independent (see Section 2.2.2) and has at least one social media forum on Reddit with hundreds of thousands of subscribers (Table 2.1).

Table 2.2: Summary of why Bitcoin and Ethereum are selected for analysis. The cryptoassets shown are consistently important cryptoassets by market capitalisation and liquidity (see Section 2.2.1 for details). Market capitalisations and liquidity were sourced at 10:27 on 18 December 2017 from coinmarketcap.com [67]. Market capitalisations are stated in US Dollars and as percentage of all cryptoassets ('Market Share'). Liquidity is US Dollar transaction volume over the last 24 hours. Ripple is excluded from analysis because it is not found to be entity-independent (Section 2.2.2), whilst Bitcoin Cash, Ripple and Litecoin are excluded because the function of the token is not sufficiently distinctive compared with Bitcoin (Section 2.2.4).

	Market Capitalisation	Market Share	Liquidity	Entity-Independent	Distinct Function
Bitcoin	318.6 bn	53.87%	13.1 bn	yes	
Ethereum	69.6 bn	11.79%	2.1 bn	yes	yes
Bitcoin Cash	31.5 bn	5.34%	0.9 bn	yes	no
Ripple	28.8 bn	4.87%	1.1 bn	no	no
Litecoin	17.3 bn	2.93%	1.2 bn	yes	no

2.3 Datasets

The literature review considers the datasets available by subdividing these into those dependent on the volume of internet activity, sentiment, topics and words. The preference for Reddit data is then considered in detail.

2.3.1 Internet Activity Volume

2.3.1.1 Google Search

For traditional asset classes, such as equity, internet activity (measured by Google search volumes) has been used as a proxy for public interest and matched with changes in market behaviour. A correlation has been found between Google searches [121] and cumulative weekly stock transaction volume [234] and between searches and stock market moves [233]. This use of Google searches as a proxy for public interest has subsequently been extended to Bitcoin (see Tables 2.3 to 2.17).

Typically, a positive correlation was identified between Google search volumes and bitcoin price returns, which suggested that a higher Google search volume for 'Bitcoin' tended to occur with larger bitcoin price rises. This was supported by results from linear regression [167, 168, 173, 183, 230], cross-correlation analysis [198], Pearson's Product Moment Correlation Coefficient (PMCC) testing [2], Spearman's Rho (SR) testing [83, 86] and Copula-based Granger Causality in Distribution testing [79]. Cai et al [49] applied fixed-effects panel regression across 268 cryptoassets and found a positive correlation between Google search volumes and price across cryptoassets. Applying Multifractal Detrended Cross-correlation Analysis [304] and three different machine learning feature selection algorithms [57] supported the existence of an association between bitcoin price and Google search volumes.

There was disagreement on whether the association between Google searches and price occurred over a short or long time frame. Bouoiyour and Selmi found Google searches to be predictive in the short-run but not in the long-run [29]. Sovbetov [266] found that Google searches had only a long-term association with the price of bitcoin and ether at the 1% significance level. Rebane et al [242] compared different predictors in forecasting the bitcoin price with seq2seq RNN models; Google search volumes enhanced long-term forecasts but were detrimental to forecasts in the short-term.

The association discovered between Google search and price depended on the dates examined. Kristoufek [173] found evidence to suggest that the positive correlation relied on including days in the dataset when the price was high and positive news events common. Wavelet analysis [174] suggested that prices led searches up to June 2012 whilst from January to April 2013 the relationship was reversed. Panagiotidis et al identified that higher Google search volumes preceded higher bitcoin returns when Google search volumes were above the 7-day moving average in a shorter dataset (18 July 2010 to 30 September 2016) and below the moving average in an extended dataset (18 July 2010 to 31 August 2018) [221]. Li and Wang [182] found limited evidence for Google search volumes having a short-term impact on price (p-value less than 10%) in an earlier dataset (1 January 2011 to 31 December 2013) but not in a later dataset (1 July 2013 to 31 December 2014). Poyser [232] discovered evidence to suggest that the nature of the association varied both across time and according to which country's Google search volume was analysed.

In some studies the correlation between Google search volumes and price was negative. Two studies established that higher Google search volumes tended to occur with lower bitcoin returns [113, 177]. Garcia et al [113] established this using a linear model, and also identified that three of the four largest daily drops in price were preceded by large increases in Google search volume. This study was corroborated by Büşra et al [177], who examined US Google search data. Subramaniam and Chakraborty [272] applied quantile regression methodology to Ethereum and Bitcoin data which supported Google searches leading to lower price returns when price returns were low whilst Google search volumes preceded higher prices returns when price returns were high. Smuts [265] corroborated, with strong, positive correlations observed when prices rose in 2017 and negative correlations reported when prices fell in 2018.

In five studies [1, 25, 114, 164, 287] there was no association between Google searches and price returns. Figá-Talamanca and Patacca [104] found Google searches were unpredictive of the mean bitcoin price return, but Google search data were predictive of the variance in bitcoin price returns and so may help in improving the accuracy of forecast models. In a later study, Figá-Talamanca and Patacca [103] found that the predictiveness of Google search data towards price volatility was not robust to splitting the dataset into sub-samples. Including dummy variables to reflect 'important' events led the effect to 'almost vanish' [103].

Bouri and Gupta [32] switched examining Google searches for the cryptoasset to examining Google searches relevant to measures of economic policy uncertainty in the US. Bouri and Gupta found such Google search volumes to rise with higher bitcoin prices. This variable was also found to be more predictive of price than a similar measure based on newspaper articles [32].

2.3.1.2 Wikipedia

Quantifying Wikipedia usage has been advocated as an alternative source of data [206] for anticipating stock market moves. Whilst Wikipedia data can be examined in terms of the number of views or edits [206], the cryptoasset literature focussed on views of the 'Bitcoin' page. Studies with data before 2016 found evidence for an association with price [30, 31, 60, 113, 114, 173, 174], with the exception of Glaser et al [116], and later studies typically found Wikipedia views not to be of predictive value [59, 221, 227]. ElBahrawy et al [93] considered 17 cryptoassets (timespan was 1 July 2015 to 23 January 2019) and found that with only five did Granger-causality tests support Wikipedia views as predictive of price; this included Bitcoin but excluded Ethereum. Dickerson examined contemporaneous associations, rather than whether Wikipedia data were predictive, and analysed a comparatively recent dataset (1 July 2015 to 3 March 2018). Positive SR correlations between views of Wikipedia pages on 'Bitcoin', 'Cryptocurrency' and 'Blockchain' and price were found [83].

Wikipedia page views have been found to be of less predictive value than both Google search volumes (when examining LASSO regression results [220,222]) and Reddit data (comparing wavelet analysis results for Bitcoin and Ethereum [227]).

A trading strategy based on Wikipedia views was less profitable than buying and holding the cryptoasset from February 2017 to January 2019 [93].

2.3.1.3 Social Media

An alternative measure of internet activity volume was activity on social media. Some studies included social media with Google search or Wikipedia views measures. Abraham et al. [2] input both Google search volumes and tweet volumes in a multiple linear regression model to predict the bitcoin price, having found that both variables had a significant, positive correlation with the bitcoin price. Garcia et al. [113] found that both more bitcoin-related tweets and more Facebook page reshares preceded higher prices. Ciaian et al. [59, 60] included number of new members and posts on `bitcointalk.org` with Wikipedia views. Only the number of new posts was found to be statistically significant when splitting the dataset into two different time periods [59]. Other studies examined just the volume of social media activity. A trading strategy informed by just Reddit activity outperformed buying and holding a cryptoasset and became less profitable when the trading volume was included [226]. Multifractal Detrended Fluctuation Analysis provided evidence for an association between Facebook likes of Bitcoin-related communities and price [188]. Laskowski and Kim [181] found Tweets to have a weak, negative correlation with price; this compared with a positive correlation for the Internet Relay Chat channel '#bitcoin-pricetalk'.

2.3.1.4 Other Sources

Only two studies examined the price-predictiveness of traffic to the cryptoasset's website. Mai et al [192] found website traffic not to be predictive of price [192]. Wang and Vergne [290] combined traffic with Bing search volumes to develop a 'public interest' metric that tended to rise before falls in price.

Different internet activity measures have been combined into a single metric designed to capture a specific characteristic. Wang and Vergne [290] combined Reddit, Facebook and Twitter data into a 'community interest' metric negatively associated with prices across a panel of Bitcoin, Litecoin, Peercoin, Ripple and Stellar. Goczek and Skliarov [118] combined Google search volumes with new posts, topics and members on bitcointalk.org with Bitcoin client downloads to create an 'attractiveness' metric positively associated with the bitcoin price. The limitation of such an approach is that it obfuscated which components of the combined metric contributed to or detracted from the observed correlation between the overall metric and price. Such an approach thus provided only limited information as to what was associated with the cryptoasset price.

2.3.1.5 Limitations

The results in the literature conflict regarding how the volume of internet activity and price are correlated. Even focussing on Google search volumes, in response to increases in internet activity, prices have been found to rise [2, 79, 83, 86, 167, 168, 173, 183, 198, 230], to fall [113, 177] and not to be associated [1, 25, 114, 164, 287].

This could be because different types of news event predominated in the different datasets examined which influenced the observed correlation between internet activity and price. Suppose mainly positive news events occurred across a dataset. A positive correlation between higher Google search volumes and higher increases in price might then be observed because when positive news events occur people both search on the internet to find out more and buy the cryptoasset [173, 183]. Using a different range of dates when negative news events were more common, news events might also drive people to search on the internet but instead sell the cryptoasset – resulting in a negative correlation being observed between Google search volumes and changes in price.

This explanation is consistent with the correlation between Google search volumes and price returns changing from positive to negative when moving from high price returns to low [265,272]. Positive news was likely to have been more common during higher price returns and negative news more common with lower returns. Liu and Tsyvinski established a positive association between volumes of Google search for 'bitcoin' and bitcoin price, but this correlation became negative when examining Google searches for 'bitcoin hack', and so upon altering the motivation for Google searched to be a negative news event [183].

Internet activity volume measures (Google searches, Wikipedia page views and social media activity) are typically analysed as predictors of price because they reflect the amount of interest in the cryptoasset [29, 31, 113, 114, 118, 168, 173, 174, 183]. Such metrics provide limited information on the positive or negative news events that caused this interest in the cryptoasset and so are the root causes of price variation.

The limited significance of finding an association between price and the volume of internet activity extends to other studies where the volume of internet activity across time is replaced by some other variable. This variable could be energy commodity prices [33, 153], transaction data [175], the cryptoasset's trading volume [12], macroeconomic variables and proxies for bitcoin demand and supply [59, 60]. Just as with Google search volumes, we may find that the value of a given metric moves closely with price within a certain range of dates, but, without knowing why the metric changed, we cannot be certain whether this association will persist with future data. For this reason, this thesis focusses on finding the root causes of price variation.

2.3.2 Sentiment

To answer why internet activity varied across time, a popular approach has been to consider the emotions ('sentiment') behind this internet activity.

2.3.2.1 Twitter

Twitter has been a popular social media source for sentiment analysis [2, 3, 9, 112, 114, 156, 164, 192, 198, 225, 231, 268, 269]. This text is converted into one or more sentiment metrics to measure the different aspects of the emotion in tweets. This typically involved the use of the Valence Aware Dictionary for sEntiment Reasoning (VADER) [142] (Tables 2.3 to 2.17) which provides scores on the proportion of text that is positive (expressing pleasure), negative (expressing displeasure) or neutral, or which can be used to generate an overall weighted average score.

Literature analysing tweets across earlier time periods typically determined that there was an association between Twitter sentiment and cryptoasset prices. This was found for data covering periods before 2015 [112,114,198] or within 2017 [268, 269]. Tweets expressing a positive sentiment were found to be predictive of the bitcoin price in the short-run [114, 198], whilst Garcia and Schweitzer found that more pleasure being expressed in tweets preceded greater polarisation which, in turn, occurred before rises in price [112].

Stenqvist and Lönnö [269] focussed on creating a predictive model using VADER sentiment metrics. The 79% accuracy of the model was used to justify the value of using sentiment metrics. The data was limited to a date range between 11 May to 11 June 2017. Rather than training a model for prediction, thresholds were applied to link aggregated Twitter sentiment changes over periods ranging from 5 minutes to 4 hours to bitcoin price fluctuations.

Steinert and Herff extended Twitter sentiment analysis to all cryptoassets where data were available except Bitcoin [268]. Using linear regression, Twitter activity and VADER sentiment were found to be predictive of cryptoasset returns, with Ethereum being in the top five cryptoassets ranked according to the mean coefficient of determination on the train data across all time lags. Evidence for an association remained upon applying the model to test data. The relative predictive value of the number of tweets compared to the value of positive, neutral and negative sentiment metrics was not evaluated. The results were also unstable with Twitter predictive of ether returns three hours after using the train data and 24 hours after on the test data. The coefficient of determination, although statistically significant, suggested that Twitter activity could explain only 2.5% of the variation in price returns, with test data. Data were limited to 45 days for the train dataset, extending from 21 March to 5 May 2017, and 26 days for the test dataset, from 9 May to 4 June 2017. This dataset occurred during a period of predominantly rising prices and terminated just before a 61% reduction in price from 12 June 2017 to 16 July 2017 (Figure 1.1).

Not all studies supported an association between Twitter sentiment and price. Two studies, using data mostly within 2014, found no association: Kaminski [156] (23 November 2013 to 7 March 2014); and Mai et al [192] (16 September 2014 to 16 December 2014). Perry-Carrera found a positive, statistically significant relationship between VADER sentiment metrics and price, but this existed only at the 5% level and was dependent on optimal lag selection (examining data across December 2017) [225]. Mai et al [191] recognised that hourly, but not daily, tweet sentiment was predictive of price (with data from 18 April 2014 to 18 August 2014) [191]. A common limitation of these studies was the small size of the dataset, comprising of one [225] or 3-4 months [156, 191, 192] of data. Furthermore, instead of VADER, Kaminski [156] used a more limited list of words to select emotional tweets.

Abraham et al [2] claimed that these previous studies were flawed in being conducted across earlier time periods when prices were continually going up. Abraham et al analysed a more recent dataset (4 March to 3 June 2018) examining Bitcoin and Ethereum, and found that information gained from sentiment analysis of tweets was of limited predictive value – particularly when prices were falling – because sentiment remained positive overall regardless of the direction of price. For Bitcoin, only one day saw tweet sentiment drop below zero, despite 11 out of 19 days showing price decreases. There was not a single day when tweet sentiment about Ethereum dropped below zero despite price fluctuations. Abraham et al [2] proposed that people who tweet about cryptoassets in a falling market are predominantly those who have a special interest in cryptoasset attributes and technology rather than their monetary value. The limited predictive value of the more recent Twitter sentiment data was also supported by Kim and Lee [164] (examining Korean tweets from November 2017 to April 2018) and Valencia et al [288] (who found forecasts based on Twitter data performed worse than a random classifier on 2018 data). Powell [231] was an exception, but this study examined the overall cryptoasset market capitalisation and was based on only 11 days of data.

In summary, the literature suggests two issues with using Twitter data. The first is the practical difficulty of extracting sufficient tweets from the Twitter API for reliable analysis, with the limited size of the Twitter dataset being a common problem across studies [2, 114, 156, 191, 192, 198, 225, 268, 269]. The second problem is that there does not seem to be sufficient evidence to support an association between Twitter sentiment and price when examining more recent data when prices fell.

Mai et al [192] attributed the lack of an association between Twitter sentiment and price to inherent flaws with Twitter data that limited the information provided by tweets. Tweets face length restrictions that both directly reduce the amount of content that can be placed in a tweet and encourage abbreviations that make tweets less interpretable. The information propagation model in Twitter discourages detailed discussions; tweets are sent from senders to followers with followers receiving little publicity in replying to these tweets. Tweets may also have a much shorter-term influence on price because finding older posts requires greater effort on Twitter than on discussion forums [192]. This may explain the finding of Abraham et al that only about half the tweets collected on any given day had an objective VADER score, the rest were strictly neutral [2].

Another limitation of Twitter is how to differentiate reliably between relevant and irrelevant tweets. Zheludev et al [305] found that Twitter sentiment was predictive in only a ‘narrow range of assets’ among UK and US foreign exchange and stock market markets. The main problem identified was how to isolate those tweets that captured an opinion on the financial asset’s future performance from all tweets that simply mentioned that financial asset [305]. Resolving this problem is complicated by the presence of fake accounts tweeting fake opinions. Ten million likely fake accounts have been created per week to tweet artificial opinions [210]. These Twitter bots have been found to skew measures of the popularity of different types of content [115]. The risk in cryptoassets is that traders ‘pump and dump’ [145], set up bots to post positive tweets to raise prices before selling.

2.3.2.2 Non-Linear Analysis

Proponents of transfer entropy [9, 84, 161] have criticised the tendency in the above literature to look for specifically linear associations between measures of internet activity and price. The problem has been illustrated with simulated, synthetic data that showed how linear models were less capable of capturing non-linear relationships than transfer entropy [84, 161] and more sensitive to extreme outliers [84].

Applying a non-parametric approach to evaluating the association between sentiment and price led to mixed results [9, 161], when measuring the sentiment of combined Twitter and StockTwits text. Aste [9] found a statistically significant positive KT correlation between the daily price and sentiment for both bitcoin (data from 1 September 2014 to 14 June 2018) and ether (data from 7 August 2015 to 14 June 2018). However, the transfer entropy values were statistically insignificant, perhaps due to difficulties in reliably measuring the transfer entropy value [9]. Keskin and Aste [161] applied transfer entropy to examine the association between the hourly price and sentiment for each cryptoasset across different 24-month windows of data. In the case of bitcoin, this supported a link that held regardless of the window examined (across August 2016 – 2018). By comparison, the evidence for a link between ether price and sentiment was weak when examining windows ending around January 2018.

In light of this criticism, this thesis selects non-parametric statistics that minimise the assumptions required. Hence, SR and KT are used to measure correlation rather than PMCC, in Chapter 3, and Wilcoxon Rank-Sum Tests are applied instead of t-tests to compare word frequencies across different time periods, in Chapters 5 to 7.

2.3.2.3 Discussion Forums

Discussion forums have been advocated as an alternative to Twitter as they are not subject to the same issues that constrain the informativeness of posts. Discussion forums lack tight length restrictions, encourage detailed discussion and make it easier to access older posts [192]. Furthermore, it is easier to facilitate the relevance of discussion forum data by selecting forums that are dedicated to the cryptoasset analysed and that have moderators to enforce this focus.

Choosing discussion forum data for analysis is supported empirically by studies finding that the sentiment of text on the discussion forum website `bitcointalk.org` is predictive of future changes in the bitcoin price. This has been established both by examining the linear associations between changes to sentiment and price within a dataset [160, 191, 192] and by examining the contribution from sentiment metrics to out-of-sample forecasting performance [192].

Kim et al [165] extended analysis to cryptoassets other than Bitcoin. This compared `forum.ethereum.org` activity with the ether price and `bitcointalk.org` activity with the bitcoin price, with VADER used to derive sentiment metrics from the text. The number of positive and very positive comments and positive replies were found to be predictive of the bitcoin price, and negative and very negative comments and positive user replies were found to be predictive of the ether price.

Xie et al [297] used more recent data from after 2016 (from 1 December 2012 to 30 June 2017). They applied sentiment analysis to messages from the ‘Speculation’ board of `bitcointalk.org`, because of its focus on bitcoin price movements. Xie et al [297] collected 12,441 threads and used thresholds to remove thread networks receiving few messages since these were less likely to be value-relevant. They quantified sentiment by the percentage of negative words in the messages, using a published list of words with negative implications in a financial context; topic modelling was only used as a control variable. They found that ‘broadcasts’ (standalone messages which were posted without quoting other existing messages) had a stronger association than ‘discussion’ messages. They found that the predictive power for bitcoin price movement was improved when the discussion network was less cohesive – meaning fewer authors quoted each other in the discussion network.

A common limitation of these analyses of discussion forum sentiment was a lack of recent data. Even Xie et al [297] ceased data collection on 30 June 2017, just before a period of unprecedented growth and decline (Figure 1.1). Although `bitcointalk.org` may have been a particularly large discussion forum before 2018, larger discussion forums may have arisen since. This issue is returned to in selecting the final dataset in Section 2.3.6.

2.3.3 Topic Modelling

Kim et al [166] criticised previous studies that used sentiment metrics (Section 2.3.2) for considering only one aspect of the information provided by a corpus of text – the conveyed emotion. Topic modelling was seen as an improvement on sentiment analysis as topics measure a wider variety of themes. Examining topic occurrence also removes the reliance on measures of sentiment to accurately capture the emotions behind social media text posted. Abraham et al [2] identified that a post's subject matter may appear neutral, when it may not be; for example, the current US Dollar price of a single bitcoin is a fact which does not itself carry sentiment. Two studies [166, 228] applied topic modelling to measure the popularity of different themes across time, with the association between topic occurrence and changes in price then evaluated.

The topics of Kim et al [166] were centred on 'keywords'. These 'keywords' were found by combining the most representative words from each topic generated by non-negative matrix factorisation with those found through the application of k-means to word embedding vectors, and by applying judgement to expand this set of keywords. The relevance of a social media post was to a topic defined by keywords was calculated by Kernel Density Estimation. Granger-causality testing provided the means of evaluating if a topic was predictive of the bitcoin price. Kim et al used data from `bitcointalk.org` [166]. This study was limited in that, of the topics created, only the concept 'China' was found to be predictive of the bitcoin price.

Phillips and Gorse [228] applied dynamic topic modelling. This provided a topic distribution for each social media post and a word distribution for each topic. The topic distribution for each post was used to track a topic's popularity across time. What a topic represented was determined by examining what the most probable words were in the word distribution for that topic. As the topic modelling was dynamic, the word distribution for each topic varied across time. This meant, to 'manually' label each topic, Phillips and Gorse relied on the 'gist of the topic' that was perceived to vary little across time [228]. Applying dynamic topic modelling also required assuming a set process by which documents were generated [228]. Having found the popularity of each topic across time, the association between topic prevalence and price was evaluated through a Hawkes model. This detects processes whereby a past event increases the probability of future events in an effect that is both additive over past events and that exponentially decays with time [202].

Other differences between Phillips and Gorse [228] and Kim et al [166] were that Phillips and Gorse analysed more recent data (from 30 August 2016 to 30 August 2017 compared with from 1 December 2013 to 21 September 2016), replaced `bitcointalk.org` data with Reddit submissions and examined both Bitcoin and Ethereum data. The results did not match. Phillips and Gorse specified 'China / announcements' as one of the Bitcoin topics, but this was not found to be predictive of price. Instead, price-relevant discussions centred on price (Bitcoin and Ethereum), trading (Bitcoin) and app development (Ethereum) [228].

2.3.4 Limitations to Examining Sentiment and Topics

A common limitation of studies that used metrics based on sentiment or topic occurrence was a reliance on subjective judgements to decide how to construct these metrics. This extended to measures of both 'perceived sentiment' [179] and topic occurrence [166, 228]. Kim et al [166] relied on subjectivity in expanding the list of words within each topic whilst Phillips and Gorse [228] had to manually apply labels based on a 'gist' of what words were most probable across time.

In engineering these metrics, price information was not used and so further analyses were required in order to determine whether these constructed metrics were price-relevant. This involved evaluating the association between each metric and price across a single dataset. Although an association might be found in one dataset across one time period, this conveyed no information as to whether this might persist upon examining other datasets covering other periods in time – an issue previously discussed in Section 2.3.1.5. Not all studies supported Twitter sentiment as predictive of price (see Section 2.3.2) whilst the topics identified as predictive of price differed when comparing Kim et al [166] with Phillips and Gorse [228].

Any identified associations also provided only limited information regarding what specific events or concerns caused price variation. Knowing that more positive (or more negative) sentiment posts preceded price rises (or falls) conveyed little as to why posts were positive or negative and so as to the specific events that happened with the price volatility.

This problem with ambiguity extended to the topic modelling approaches. Kim et al [166] found ‘China’ to be important. This topic was based on the keywords ‘china’, ‘chinese’ and ‘baidu’, and so it was ambiguous whether ‘China’ related to the importance of Chinese regulation, the Chinese economy, Chinese speculation or the Chinese adoption of cryptoassets. Phillips and Gorse [228] found ‘mainstream adoption/app development’ to predict higher ether prices. The most probable words found for this topic provided limited detail on what this topic represented: ‘hope’, ‘private’, ‘key’, ‘site’, ‘Google’, ‘Amazon’, ‘bittrex’, ‘code’, ‘trust’, ‘app’.

In summary, studies evaluating measures of sentiment or topic occurrence were flawed in requiring judgement to decide how best to extract information from the text and in leaving unclear whether an identified association might persist across time. They tended to generate results that lacked interpretability.

2.3.5 Words

Lamon et al [179] alternatively analysed individual words. They developed a predictive modelling approach that converted individual words into continuous, numeric vectors (using word2vec), passed these through a bi-direction Long Short-Term Memory neural network (LSTM) and then used a linear activation function to predict the change in bitcoin price 1 to 24 hours after. The highest prediction accuracy attained was 54.5%, which was obtained through using Reddit data, which was an improvement over news headlines and tweets. They did not seek to find potential causes of cryptoasset price variation, providing limited information on which events or concerns were likely causes of price movement.

2.3.6 Selection of Reddit Data

This thesis selects a dataset of Reddit submissions because of the relatively high number and activity of its users compared with alternatives [169]. Prior analyses of discussion forum text extracted data from `bitcointalk.org`, but such studies were typically conducted on data before the 2017-18 period examined by this thesis (see Section 2.3.2.3). The number of online users at 17:40 and 18:08 on 25 September 2018 (GMT) was, respectively, 1758 and 1643 on `bitcointalk.org` [23] and 8100 and 8300 on the Reddit subreddit ‘r/Bitcoin’ [240]. This supports that the activity on the Reddit subreddit had overtaken `bitcointalk.org`.

Examining Reddit data also meant that separate subreddits from the same website (`reddit.com`) could be selected for Bitcoin and Ethereum. This meant that any differences in the results for Bitcoin and Ethereum could not be attributed exclusively to differences in the website used. Just as Bitcoin had a large dedicated subreddit, so did Ethereum. The subreddit 'r/Bitcoin' had over 1.1 million subscribers as of 18:54 (GMT) on 23 August 2019 [240], whilst 'r/ethereum' had 436,000 subscribers on 14 May 2019 [245]. This was an advantage over other social media platforms where the data available on Ethereum was limited (such as with Telegram [265]).

Prior studies support Reddit as a price-predictive source of information. Reddit data have been found to outperform Wikipedia page views [227], news headlines and tweets [179] in predicting price. Phillips and Gorse [228] also used Reddit submissions to find topic occurrences associated with price.

Reddit forums (or 'subreddits') detail what the purpose of each post in the subreddit should be focussed on by providing a subreddit description, rules and guidelines. This purpose is enforced through the existence of moderators who can remove off-topic posts and ban spammers [243]. Moderators thus have significant power to block fake accounts and to ensure that the text remains relevant to the purpose of the subreddit. For example, in the 'Bitcoin' subreddit, moderators act to ensure that the 'primary topic is Bitcoin' [240].

Reddit text was publicly accessible through the Pushshift API [15]. For this thesis, Bitcoin analyses used text from 'r/Bitcoin' whilst Ethereum analyses combined several subreddits. The largest [72] Ethereum subreddit was 'r/ethereum', which was also moderated by Vitalik Buterin, the 'Creator of Ethereum' [7]. Following this forum's guidelines [245], its text was combined with that from 'r/ethtrader' and 'r/EtherMining'. Together, these had the most submissions containing the term 'ether' or 'eth' among Ethereum-specific subreddits [15] and have collectively been described as the most important subreddits [72]. Submissions data were selected over comments because the latter were prone to deviate onto arguments on bitcoin-irrelevant topics, such as religion, non-specific insults and different date formats (`https://www.reddit.com/r/Bitcoin/comments/9svjcp/10_years_ago_today_2008_oct_31/`).

Section 2.3.4 details how previously analysed sentiment-based or topic-based metrics were created without price information and generated results that lacked interpretability. In this thesis, price data are used to inform the criteria that are applied to extract words or topics associated with price from the Reddit submissions text, with the exact criteria varying across the different quantitative analyses (Chapters 5 to 7). To aid interpretation, where topics are analysed, these are specified such that the constituent and non-constituent words are clearly defined. The context in which the delineated words and topics are used is also determined to facilitate connecting the delineated words and topics with specific events and concerns that could be associated with price.

Table 2.3: Internet Activity Literature Review 2013 – 2014.

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Kristoufek 2013 [173]	Google Trends (weekly), Wikipedia Views	1/5/11–30/6/13	Google: 113 weeks; Wikipedia: 788 days	Linear (VAR/VECM)	Google: when price above trend price, increasing search queries correlates with higher price; but no association when below. Wikipedia views: no overall correlate unless separated into positive and negative phases, then higher interest leads to higher prices in positive phase and lower prices in negative phase	No	Yes
Glaser et al 2014 [116]	Wikipedia Views	1/1/11–8/10/13		Linear (ARCH)	Wikipedia page views did not influence bitcoin returns	No	No
Garcia et al 2014 [113]	Twitter, Wikipedia Views, Facebook, Google Trends and Bitcoin blockchain	9/1/09–31/10/13	6,827,894 tweets; 6,330,676 Wikipedia Views; 2,461 Facebook Reshares; 4,717,713 Bitcoin Users	Linear (VAR)	Four-tiered dataset composed of exchange data, social media activity, search trends and user adoption. Spikes in Google search volume and Wikipedia page views precede price declines. Three of the four largest daily drops in price preceded by first, fourth and eighth largest increases in Google Search volume. More bitcoin-related tweets or Facebook page reshares precede higher prices.	No	No
Kaminski 2014 [156]	Twitter	23/11/13–7/3/14	Twitter 1,612 tweets/day on average	Linear regression; PMCC; emotional tweets selected with list of words	Sum of total emotions and uncertainty sentiment had a strong positive correlate with trading volume on Bitstamp. Drop in price correlated with negative sentiment. Granger-causality testing did not confirm sentiment predictive.	Yes	No

Table 2.4: Internet Activity Literature Review 2015 part 1 (datasets ending in 2014).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Kristoufek 2015 [174]	Google Trends, Wikipedia Views	14/9/11–28/2/14		Wavelet analysis	Prices led Google searches up to June 2012 whilst from January to April 2013 the relationship was reversed. Wikipedia similar. China had an influence.	No	Yes
Polasik et al 2015 [230]	Google Trends (monthly), Nexis (monthly)	1/4/11–1/3/14		Linear regression	1% increase in Nexis articles with higher returns of 31-36 basis points. 1% rise in Google searches with 53-62 basis point increase. Articles undermining Bitcoin's reputation associated with lower price.	Yes	No
Bouoiyour et al 2014,2015 [30, 31]	Wikipedia Views	December 2010– June 2014		Frequency domain Granger causality tests on linear models	Evidence supports Wikipedia views as causing price at short- and long-run frequencies.	No	No
Bouoiyour and Selmi 2015 [29]	Google Trends	5/12/10–14/6/14		Linear ARDL bounds testing method, VEC Granger causality test and an innovative accounting approach	Google search has significant, positive impact in short-run but not long-run	No	No
Mai et al 2015 [191]	Bitcointalk.org (BT); Twitter (TW) (daily and hourly)	BT: 22/11/09–18/8/14; TW: 18/4/14–18/8/14	BT: 119,847 posts and 51,269 topics; TW: 3,348,965 unique tweets from 339,295 users	Linear (VAR/VECM); finance sentiment dictionary	Daily data: more bullish forum posts precedes higher bitcoin returns; more bearish forum posts precedes lower returns. Daily Twitter variables have negligible predictive power. Hourly data: more bullish tweets precedes higher returns with effect stronger if limited to opinion leaders' tweets. More bearish opinion leader's tweets precede lower returns. Hourly forum sentiment negligible predictive power.	Yes	No
Garcia and Schweitzer 2015 [112]	Google Trends, Twitter	1/2/11–1/12/14	19,578,671 Bitcoin-related tweets	Linear (VAR); lexicon technique for sentiment	More pleasure being expressed in tweets preceded greater polarisation and exchange volumes which, in turn, occurred before rises in price. These results informed a profitable trading strategy.	Yes	No

Table 2.5: Internet Activity Literature Review 2015 part 2 (datasets ending in 2015).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Georgoula et al 2015 [114]	Google Trends, Twitter, Wikipedia Views	27/10/14–12/1/15	2,125,243 tweets	Linear regression; Twitter sentiment by Support Vector Machines	Higher bitcoin prices preceded by more Wikipedia search queries and higher Twitter sentiment ratio in short-run. A significant association was not found for Google Trends and Tweet volumes data.	Yes	No
Matta et al 2015 [198]	Google Trends, Twitter	1/1/15–31/3/15	1,924,891 tweets	Cross-correlation; SentiStrength for sentiment	Positive mood predicts Bitcoin's price rise in 3-4 days. Higher Google Trends predicts higher price with zero lag.	Yes	No
Dokic et al 2015 [86]	Google Trends (monthly)	January 2011–April 2015		SR to test associations between price and search volume for different South-East Europe region countries	Strong correlation (0.85– 0.95) except Albania, Montenegro and Kosovo perhaps due to lack of data	No	No

Table 2.6: Internet Activity Literature Review 2016.

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Ciaian et al 2016 [60]	Wikipedia Views, Bitcointalk.org (new members and posts)	2009–2014		Linear (VAR/VECM)	Wikipedia views and new posts rise before higher bitcoin prices; new members rise before falling prices.	No	No
Ciaian et al 2016 [59]	Wikipedia Views, Bitcointalk.org (new members and posts)	1/11/09–31/5/15 (1/11/09–30/9/13 & 1/10/13–31/5/15)		Linear (VAR/VECM)	First period: investment attractiveness (new posts, new members on bitcointalk.org; Wikipedia views) had statistically significant, short-run predictive association with bitcoin price. Second period: new members and views became statistically insignificant; new posts remained significant.	No	Yes
Laskowski and Kim 2016 [181]	Twitter and Internet Relay Chat (IRC)	Twitter: 1/6/15–31/12/15; IRC: 1/6/15–12/12/15	Twitter: 12,105,833; IRC 64,712 (#bitcoin-pricetalk) to 1,113,243 (#dogecoin) messages	PMCC	Price has weak, negative correlation with tweets (-0.0191), positive correlation with IRC channel #bitcoin-pricetalk (0.5715) and negative correlations with #dogecoin (-0.3333) and #bitcoin-assets (-0.2991)	No	No
Kim et al 2016 [165]	bitcointalk.org and forum.ethereum.org	Bitcoin: 1/12/13–1/2/16; Ethereum: 7/8/15–8/2/16	Bitcoin: 13360 threads; Ethereum: 1449 threads	Linear (Granger-causality); VADER sentiment; Forecasts based on AODE	Bitcoin price associated with number of topics posted, positive/very positive comments and positive replies; 79.57% highest prediction accuracy with 6-day lag. Ether price associated with negative/very negative comments and positive replies; 71.823% highest accuracy with 6-day lag. Trading based on predictions raises investment by 35.09% compared with 19.29% price rise.	Yes	No

Table 2.7: Internet Activity Literature Review 2017 part 1 (datasets ending before 2017).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Li and Wang 2017 [182]	Google Trends, Twitter	1/1/11–31/12/13 & 1/7/13–31/12/14		Linear (VECM and ARDL)	Short-term positive impact from Google search and negative impact from tweets seen in earlier but not later dataset. Google search appears to have long-term, positive impact on bitcoin price while tweets do not.	No	Yes
Wang and Vergne 2017 [290]	Weekly 'public interest' combines: Bing searches, Alexa website traffic; 'community interest': Reddit, Facebook, Twitter	September 2014– August 2015	250 observations	Linear (fixed effects panel regression and GARCH-in-mean)	Both public interest and community interest have negative association with price, examining a panel of Bitcoin, Litecoin, Peercoin, Ripple and Stellar. GARCH-in-mean model supports investor aversion to volatility.	No	No
Kim et al 2017 [166]	Bitcointalk.org (discussion section), Google Trends, Wikipedia Views	1/12/13–21/9/16	Bitcointalk.org: 17,381 articles, 627,122 user comments; Google and Wikipedia: 1,026 observations	Linear Granger-causality tests on topics used on Bitcointalk.org; FFN prediction model used 10 topics, Google and Wikipedia data.	Ten topics identified of which 'China' statistically significant (p-value $< 5\%$) predictor of bitcoin price. Most accurate (80.39%) predictive model was three-layer FFN with a previous 12-day learning period.	No	No
Phillips and Gorse 2017 [226]	Number of new Reddit posts, subscribers and authors that post.	April 2015–September 2016 except Ethereum: August 2015–September 2016		Back-tested trading strategy involving Bitcoin, Litecoin, Ethereum and Monero and informed by epidemic-detecting HMMs	Trading strategy based on Reddit activity resulted in higher returns, Sharpe ratio and Sortino ratio and lower percentage drawdown over shorter period than buy and hold. Excluding trading volume and basing trades on unanimous HMM agreement raises profit.	No	No
Büşra et al 2017 [177]	Google Trends (weekly)	2011-16		Linear (ARIMA and least squares regression)	When US Google search volumes increase, bitcoin prices fall in a statistically significant association compared with a statistically insignificant association for Turkey Google search volumes.	No	No

Table 2.8: Internet Activity Literature Review 2017 part 2 (datasets ending in 2017).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Stenqvist and Lönnö 2017 [269]	Twitter	11/5/17–11/6/17	2,271,815 tweets	Thresholding aggregated Twitter sentiment change over periods ranging from 5 minutes to 4 hours to predict bitcoin price fluctuations; VADER sentiment	Aggregating tweet sentiments over a 30 minutes period with 4 shifts forward and applying a threshold of 2.2% to the sentiment change yielded a 79% accuracy. Number of predictions very low.	Yes	No
Lyudmyla et al 2017 [188]	Facebook	October 2016–July 2017		Multifractal Detrended Fluctuation Analysis	Likes of community ‘Bitcoin Product/service’ strongly correlated with price whilst number of likes in community ‘Blockchain’ weakly correlated with a lag.	No	No

Table 2.9: Internet Activity Literature Review 2018 part 1 (datasets ending before May 2017).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Mai et al 2018 [192]	Bitcointalk.org discussion board (BT), Twitter (TW), Google Trends, Bitcoin.org web traffic, TRNA Bitcoin news sentiment	BT: 1/1/12–31/12/14; TW: 16/9/14–16/12/14	BT: 343,769 posts and 15,420 topics from 17,215 users; TW: 3,348,965 tweets from 339,295 users	Linear (VECM, Granger-causality); finance sentiment dictionary; FEVD	More positive sentiment before higher bitcoin prices and more negative sentiment before lower prices. Predictive: sentiment of Bitcointalk.org messages and Google search volumes. Not predictive: tweet sentiment, TRNA Bitcoin news sentiment and Bitcoin.org web traffic. Social media improves out-of-sample forecasts across 3-month test period. FEVD supports social media sentiment explanatory power.	Yes	No
Kennis 2018 [160]	Bitcointalk.org; Reddit; Internet Relay Chat (IRC)	Data over year 2015		Granger-causality tests; sentiment found using classifiers trained on a random sample labelled by crowd-sourcing (Mechanical Turk)	Negative bitcointalk.org and news sentiment predictive of price whilst price predictive of Reddit sentiment.	Yes	No
Zhang et al 2018 [304]	Google Trends	1/6/11–1/2/17		Multifractal Detrended Cross-correlation Analysis	Google Trends and bitcoin price returns are significantly cross-correlated. Rolling window analysis suggests this cross-correlation diminished over time.	No	Yes
Kjærland et al 2018 [168]	Google Trends (weekly)	1/9/11–5/2/17	Google 279 observations	Linear (ARDL)	Significant positive relationship between Google searches and price. Negative news drives price down significantly.	No	No
Lamon et al 2018 [179]	Coindesk news, Twitter, Reddit	1/1/17–6/3/17	Approximately 4,000 news items, 100,000 tweets and 100,000 posts	Word2vec with LSTM predictive model	Overall accuracy of prediction was highest (54.5%) when predicting 12 hours in the future using a model trained on Reddit posts.	No	No

Table 2.10: Internet Activity Literature Review 2018 part 2 (datasets ending in May or June 2017).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Phillips and Gorse 2018 [227]	Reddit (posts per day, subscribers, authors), Google Trends (daily and weekly), Wikipedia Views	Bitcoin: 10/9/10–31/5/17; Ether: 8/8/15–31/5/17		Wavelet analysis	Short term: erratic and sparse periods where price falls lead higher online activity but not the reverse. Medium term: patches of strong relationship separated by long periods of no relationship. Long term: Reddit derived factors more predictive than Wikipedia. Google Trends has periods where unclear whether Google Trends or price leads. Bitcoin and Litecoin positively correlated in the medium and long term.	No	Yes
Steinert and Herff 2018 [268]	Twitter (three-hour intervals)	Train: 21/3/17–5/5/17; Test: 9/5/17–4/6/17	426,520 tweets for 181 cryptoassets (excluding Bitcoin)	Bonferroni-corrected tests of R^2 based on linear regression; VADER sentiment	Number and sentiment of Tweets can predict altcoin returns (including Ethereum and excluding Bitcoin from analysis).	Yes	No
Panagiotidis et al 2018 [220]	Google Trends, Wikipedia Views	17/6/10–23/6/17	2,533 daily observations	Linear (LASSO ‘glmnet’ and ‘lars’ packages); ADF Breakpoint unit root test	In predicting the bitcoin price, positive coefficients for above-trend internet volumes (Google and Wikipedia) and negative coefficients for below trend. Compared with Wikipedia: above-trend Google search has a larger positive coefficient; below-trend Google search has a larger negative coefficient.	No	Yes
Xie et al 2018 [297]	Bitcointalk.org	1/12/12–30/6/17	12,441 threads	Linear regression; sentiment measured by % of negative words; trading simulation	Demonstrated importance of broadcasts; predictive power improved when network less cohesive; simulation supports profitability of trading based on cohesion-weighted sentiment.	Yes	No

Table 2.11: Internet Activity Literature Review 2018 part 3 (datasets ending after June 2017 and before 2018).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Urquhart 2018 [287]	Google Trends	1/8/10 – 31/7/17		Linear (VAR)	Google search volumes not predictive of returns, volatility and volume but these features are predictive of Google searches. This holds for data after 27 October 2013 but not before.	No	Yes
Phillips and Gorse 2018 [228]	Technical and trading Reddit submissions	30/8/16–30/8/17		Hawkes model applied to occurrence of topics found by dynamic topic modelling	Manually-labelled topics ‘downward price movement’ and ‘risk/investment vs trading’ indicates bitcoin price falls; ‘substantial price movements’ indicate bitcoin volatility; and ‘fundamental cryptocurrency value’ and ‘mainstream adoption/app development’ indicate ether price rises.	No	No
Perry-Carrera 2018 [225]	Twitter	1/12/17–31/12/17	Greater than 500,000 tweets	Linear (VAR); VADER sentiment	Difficult to show relationship with sentiment except at 5% confidence with maximum lag criterion and optimal lag selection.	Yes	No

Table 2.12: Internet Activity Literature Review 2018 part 4 (datasets ending in 2018 and before May 2018).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Kjærland et al 2018 [167]	Google Trends (weekly)	1/1/13–20/2/18		Linear (ARDL and GARCH)	Higher publicity measured in Google Trends associated with higher price	No	No
Dickerson 2018 [83]	Google Trends, Wikipedia Views	1/7/15–3/3/18		SR to test associations. Trading strategy based on change in median search volume.	SR between bitcoin price and different Wikipedia keywords and between price and different Google keywords all over 0.85 indicating very strong relationship. The search term ‘Bitcoin’ had lowest correlation with price but resulted in trading strategy with highest returns	No	No
Cai et al 2018 [49]	Google Trends	4/9/17–3/3/18	181 trading days	Linear fixed-effects panel regression across 268 cryptoassets	Google search volumes for all three days prior positively associated with the subsequent cryptoasset price with the highest coefficient being for one day prior.	No	No
Rebane et al 2018 [242]	Google Trends	25/8/15–4/4/18 (Train data up to 23/1/17)		Comparison of different ARIMA and seq2seq RNN models across various datasets of predictors.	All seq2seq RNNs outperform ARIMA in forecasting the bitcoin price. Including Google search volumes for ‘Bitcoin’ and ‘Ethereum’ provided the best performance for long-term forecasts but reduced performance of short-term forecasts. Seq2seq RNNs showed poor performance in predicting after 16 December bitcoin price crash.	No	No

Table 2.13: Internet Activity Literature Review 2018 part 5 (datasets ending in May 2018 and after).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Liu and Tsyvinski 2018 [183]	Google Trends; Twitter (weekly)	Bitcoin: 1/1/11–31/5/18; Ethereum: 7/8/15–31/5/18		Linear regression	When weekly searches rose by a standard deviation the weekly price return increased by 1.84% (at 1 week) and 2.5% (at 2 weeks) for Bitcoin; and 4.36% (at 1 week), 3.45% (at 3 weeks) and 3.65% (at 6 weeks) for Ethereum. Tweet counts also have positive association with bitcoin returns. The ratio of ‘Bitcoin hack’ to ‘Bitcoin’ (proxy to negative sentiment) negatively and significantly predicted 1-5-week bitcoin price returns.	Yes	No
Abraham et al 2018 [2]	Google Trends, Twitter	Sentiment data: 4/3/18–3/6/18; PMCC tests: April 2014– June 2018	30,420,063 tweets	VADER sentiment; PMCC; linear regression	Sentiment of tweets not a reliable indicator of falling price. Google Trend and Twitter volume were highly, positively correlated with price. Regression model with Google and Twitter data accurately reflected price fluctuations.	Yes	No
Sovbetov 2018 [266]	Google Trends (weekly)	2010-2018	126 – 390 weeks (depending on analysis)	Linear (ARDL/ECM)	Google search has long-term but not short-term association with bitcoin and ether price at 1% significance level.	No	No

Table 2.14: Internet Activity Literature Review 2019-20 part 1 (datasets ending in 2017).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Aalborg et al 2019 [1]	Google Trends (Weekly and Daily)	1/3/12– 19/3/17		Univariate and multivariate linear regression with robust standard errors	Google trends statistically insignificant examining: weekly and daily data; univariate and multivariate regressions; and contemporaneous Google trends and Google trends in the past day or week.	No	No
Poyser 2019 [232]	Google Trends (weekly)	January 2013– May 2017		Bayesian Structural Time Series with Spike and Slab	Google search volume correlation with price varied in strength and sign across countries and time. 13/44 countries with correlation supported. US overall positive association 2013-17 but negative in 2014. Nigeria negative correlation, fading since 2013.	No	Yes
Goczek and Skliarov 2019 [118]	‘Attractiveness’ combines: Google Trends; new posts, topics and members on Bitcointalk.org; Bitcoin client downloads (monthly and daily)	July 2010– December 2017		PCA and Factor augmented VECM	Impulse response to increase in ‘attractiveness’ is positive, a result robust to using daily frequency and transforming variables into logarithmic form.	No	No
Figá-Talamanca and Patacca 2019 [104]	Google Trends	1/1/12– 31/12/17 (1/1/12– 31/12/14 & 1/1/15– 31/12/17)		ARMA with GARCH and EGARCH. Forecast model compared with observed values using out-of-sample data.	Google search affects variance but not the mean bitcoin return; trading volume affect both variance and mean. Forecast models enhanced by including Google search or trading volume measure.	No	Yes
Dastgir et al 2019 [79]	Google Trends (weekly)	1/1/13– 31/12/17 (1/1/13– 11/8/13 & 12/8/13– 31/12/17)		Copula-based Granger Causality in Distribution test	Positive association between Google search volumes and price. Overall and in second sample: causal relationship in extreme quantiles; first sample: Google search predictive of price in central quantiles.	No	Yes

Table 2.15: Internet Activity Literature Review 2019-20 part 2 (datasets ending in 2018 and before August).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Subramaniam and Chakraborty 2020 [272]	Google Trends	January 2013–March 2018		Granger causality testing applied to linear regression and quantile regression.	For both Ethereum and Bitcoin: when price returns high, higher Google search volume before higher price returns; when price returns low, higher Google search volume before lower price returns.	No	No
Valencia et al 2019 [288]	Twitter	16/2/18–21/4/18	20,789,572 tweets	VADER sentiment; performance of FNN, SVM and RF compared in predicting direction of price movement	Bitcoin: using just Twitter data led to lowest accuracy for FNN, SVM and RF that was worse than random. Including Twitter data deteriorated SVM and RF performance and improved FNN slightly. Ethereum: no model performed significantly better than random using Twitter and/or market data.	Yes	No
Kim and Lee 2019 [164]	Google Trend; Twitter (Korean data)	November 2017–April 2018	154,783 tweets (100,120 after removing noise)	Kernel Regularized Least Squares (KRLS); HMM	In forecasting the price with HMM, using trading volume and historic price results in higher accuracy and Area Under Curve than using Google search and/or Twitter sentiment data. KRLS finds no significant association between price and either Google search or Twitter sentiment metrics.	Yes	No
Aste 2019 [9]	Twitter and Stock-Twits	Bitcoin: 1/9/14–14/6/18; Ether: 7/8/15–14/6/18		Permutation tests applied to KT correlation and transfer entropy; PsychSignal measured sentiment.	Examining 1,944 cryptoassets, sentiment causes price and price causes sentiment across cryptoassets. The bitcoin price was significantly associated with 894 other cryptoasset prices; ether with 902. For bitcoin and ether, there was a significant, positive correlation between positive sentiment and price but transfer entropy was not significant.	Yes	No
Smuts 2019 [265]	Google Trends and Telegram (hourly)	1/12/17–30/6/18; validation data: May 2018; test data: June 2018	5,088 hours	VADER sentiment; PMCC testing; comparison of LSTMs in predicting direction of price change	PMCC tests: Google search volume and bitcoin price correlation positive in first two weeks of December 2017 but strongly negative by June 2018; 2018 inversion also observed for Ethereum; combined Telegram message sentiment weak, positive PMCC with bitcoin price, unstable association for Ethereum. Comparing LSTMs: Telegram data resulted in highest accuracy for bitcoin and Google Trends had marginal impact; Google Trends most accurate for ether and Telegram had marginal impact.	Yes	Yes

Table 2.16: Internet Activity Literature Review 2019-20 part 3 (datasets ending in or after August 2018 and before 2019).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
Panagiotidis et al 2019 [221]	Google Trends; Wikipedia Views	18/7/10–30/9/16 & 18/7/10–31/8/18		Linear VAR model and Factor-Augmented VAR model	VAR and Factor-augmented VAR suggest higher Google searches impact higher bitcoin prices but this occurs when Google searches above 7-day moving average in smaller dataset and below the moving average in the extended dataset. Shocks to Wikipedia do not seem to affect bitcoin.	No	Yes
Keskin and Aste 2019 [161]	Twitter and StockTwits (hourly)	August 2016 – August 2018		Linear (Granger-causality) and Transfer Entropy with results found over 24-month windows with 2-week stride; PsychSignal measured sentiment.	Simulation data supports transfer entropy as capturing non-linear relationships not detected by linear Granger-causality tests. Bitcoin: strong causative signal from sentiment to price and from price to sentiment. Ethereum: causative signal weak in windows ending around January 2018.	Yes	Yes
Bleher and Dimpfl 2019 [25]	Google Trends (hourly/ daily/ weekly)	Bitcoin: 14/09/13–30/9/18; Ethereum: 8/8/15–30/9/18		Ordinary Least Squares Linear Regression; Clark-West Test if RMSE reduced by Google data; Wald Test for fit; and Granger-Causality Tests for Google predictiveness	Wald and Granger-causality tests support Google data as predictive of volatility but not returns. Google reduces RMSE by statistically significant but economically insignificant amount. Examining rolling windows finds no extended period where Google search Granger-causes returns. Google search predictive with weekly data but RMSE ‘huge’.	No	Yes
Figá-Talamanca and Patacca 2019 [103]	Google Trends	January 2012–December 2018		VAR and EGARCH model with dummy variables included to reflect ‘important’ events.	Google search volume intensity does not affect mean bitcoin returns. The effect of Google search on price volatility was not found in one of four sub-samples and ‘almost vanish[ed]’ when dummy variables for important events included.	No	Yes
Bai et al 2019 [11]	Reddit	1/7/18–31/12/18	225,869 submissions and 1,624,674 comments from 101,564 authors	Google Cloud Natural Language API and LIWC for sentiment analysis; Student’s t-test to check Collective Cryptocurrency Price Prediction (C2P2) significant improvement in predicting 21 cryptoasset price series.	C2P2 outperforms ensemble of neural networks by 16% and an LSTM RNN by 5.1%-44.1%. These improvements are statistically significant.	Yes	No

Table 2.17: Internet Activity Literature Review 2019-20 part 4 (datasets ending in 2019 or dates unstated).

Author	Data Source (Daily)	Timescale	Number of Data Points (where stated)	Analytic Tools	Results	Sentiment Analysis	Cycle Analysis
ElBahrawy et al 2019 [93]	Wikipedia edit history and page views	1/7/15–23/1/19		SR correlation, Granger-causality test and trading strategy comparison	Positive correlation with price, share of exchange volumes and market share found for Wikipedia views and share of edits; Granger-causality supports predictive association in only 5 of 17 cryptoassets. Trading strategy based on page views outperforms price-based and random baselines (July 2015 to January 2018) but buy and hold only up to January 2017.	No	No
Chen et al 2020 [57]	Google Trends	2/2/17–1/2/19	740 observations	Boruta combined with forwards and backwards stepwise selection to select features. Prediction performance compared statistical (LR, LDA) with machine learning (RF, XGT, QDA, SVM and LSTM) approaches.	Feature selection approaches suggest Google search volumes predictive of sign change in bitcoin price. With daily data statistical outperforms machine learning but opposite true with 5-minute interval data. Daily data: LR highest accuracy of 66% compared with XGT 48.3%. 5-minute interval data (which excludes Google search volumes): LSTM highest accuracy of 67.2% compared with lowest LDA (51.5%).	No	No
Powell 2019 [231]	Twitter	1/4/19–15/4/19		Linear regression; sentiment calculated using ‘syuzhet’ package	Twitter sentiment has significant positive association with changes in cryptocurrency market capitalisation at 1% level while Twitter volume insignificant (p-value of 29.3%).	Yes	No
Bouri and Gupta 2019 [32]	Google Trends (monthly)	July 2010–May 2019	107 observations	EGARCH	Google search volumes related to economic uncertainty in US positively and strongly significantly associated with bitcoin returns; this was more important statistically and economically than a newspaper-based measure	No	No
Aggarwal et al 2019 [3]	Twitter	Unstated		CNN, LSTM and GRU performance compared using RMSE; VADER sentiment	LSTM lowest RMSE. LSTM suggests that positive sentiment tweet precedes price hike and negative sentiment precedes price fall	Yes	No
Pavlyshenko 2019 [222]	Google Trends; Wikipedia Views	Unstated		LASSO linear regression; Bayesian regression	LASSO regression selects both Google search volumes for ‘Bitcoin’ and views of Wikipedia page ‘cryptocurrency’ but coefficient for Google Trends exceeds that for Wikipedia. These results are robust to including an ‘expert correction term’ and to switching to Bayesian regression.	No	No

2.4 Methodologies

This section reviews the literature behind the analytic strategy that will be applied to understand the associations between Reddit data and price. This starts with a consideration of forecast models and characterises the limitations of such models in understanding the interrelationships in the dataset. This motivates a review of the causal inference methodologies that guides the analytic approach underpinning the quantitative analyses in this thesis (specified in Section 2.4.3).

2.4.1 Forecasting the Cryptoasset Price

The literature is dominated by comparisons of the accuracy of different forecasting models that leaves unclear the practical value of any 'optimal' model determined [123, 127, 143, 151, 158, 162, 200, 201, 212, 229, 283, 298]. Where practical use was considered, this was in the context of maximising the profitability of a trading strategy [5, 6, 8, 93, 132, 165, 189, 259, 262].

Twenty-one of the papers examining the predictiveness of internet activity metrics (Section 2.3) developed a forecast model [2, 3, 11, 25, 57, 83, 93, 104, 112, 164–166, 179, 192, 226, 242, 265, 268, 269, 288, 297]. Tables 2.18 to 2.21 (see end of Section 2.4) summarise 16 further studies that determined how to optimally create a forecast model to predict the cryptoasset price. This literature focussed on predicting the bitcoin price but some studies did examine cryptoassets other than Bitcoin, such Alessandretti et al [5], who analysed 1,681 such cryptoassets. Alessandretti et al [5] found that the most profitable investment portfolio was based on training separate Long Short-Term Memory neural networks for each cryptoasset and that this was profitable even with transaction fees of 1%.

Price was usually measured on a daily basis but data were examined with increased granularity at hourly [127], 15-minute [212], 10-minute and 10-second intervals [189,259]. This may in part reflect the source of the data available rather than any consensus on the appropriate time interval. Hourly data were collected directly from OKCoin [127] but an automated real time web scraper had to be developed for obtaining the 10 minute and 10 sec data from OKCoin and Coinbase [189]. Price was normally priced in US Dollars although the Chinese Yuan has also been used [8]. When past price data were used to forecast future prices, this past price data was stated as either price *per se* or subdivided into opening, maximum and minimum prices [143].

The literature disagreed on what variables should be considered in forecasting the cryptoasset price. Different data preparation techniques were also considered such as Principal Component Analysis [298] and Exponential Weighted Moving Averages [127]. Studies typically used data on historic price but varied regarding what other predictors should also be considered. Predictors used included:

- Trading volume [6, 104, 112, 127, 151, 262, 265, 283, 298]

- Bitcoin network metrics (such as miner revenue and block size) [123, 151, 262]

- Measures of internet search – particularly Google search volumes [25,57,104, 112, 164, 192, 242, 265, 283]. Dickerson [83] based a trading strategy exclusively on Google search volumes and Wikipedia views whilst ElBahrawy et al [93] analysed solely Wikipedia page views.

- Social media metrics covering cohesion [297], sentiment [11, 112, 164, 192, 265,288,297] and volume of activity [112,226,265]. Sometimes social media data were exclusively used to predict price [165, 179, 226, 268, 269]. Social media data were also supplemented by Google search volumes [2, 164], and Google search and Wikipedia page views [166].

- Data on other assets such as stock markets, commodities and other cryptoassets [151, 283].

Having identified the predictors to be used, the literature varied in the choice of methodology to be applied in converting these predictors into a cryptoasset price prediction. Methodologies can be subdivided into those based on:

- Linear Models [2, 8, 25, 57, 123, 127, 151, 189, 192, 259, 268, 288, 298] that included Auto-Regressive Moving Average (ARIMA) and Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH) [8, 104, 127, 158, 162, 200, 201, 283]
- Neural Networks that included Feedforward Neural Networks [6, 123, 166, 212, 229, 262, 288, 298] and Recurrent Neural Networks [3, 5, 11, 57, 151, 162, 179, 200, 201, 242, 265, 283].
- Recurrent Reinforcement Learning [132]
- Decision Trees and Random Forests [8, 57, 127, 189, 288]
- Gradient Boosting Machines [5, 57, 127, 132]
- Hidden Markov Model (HMM) [164, 226]
- Applying a threshold to changes in sentiment [269] or search volumes of a keyword [83]
- Averaged One-Dependence Estimators [165]
- Quadratic Discriminant Analysis [57]
- Learning the empirical conditional probability distribution [8]
- Collective Cryptocurrency Price Prediction [11]

Hence, in finding a forecasting approach that performed optimally in predicting price, the literature disagreed on how the price data should be handled, what predictors should be used and what methodology should be applied.

The problem with analysing a historical dataset to find an ‘optimal’ forecast model is that the data encountered in practice may substantially differ from the dataset used to train the model. This meant that a model trained to have a low error on the dataset could be highly inaccurate upon deployment. For example, a forecasting model might be trained to predict price with data prior to June 2017 [6, 8, 123, 127, 132, 143, 189, 200, 201, 259, 262]. However, in the extreme market conditions after June 2017, the variables driving price and/or the nature of causal relationships may have changed. This could lead the 'optimal' model to underperform in practice, meaning that any trading strategy based on this model would be unprofitable.

This suggests a need for a better understanding of what features have a robust association with price and the nature of such associations. For Bengio et al [20], this meant finding what are the true cause-effect relationships. This could potentially guide the creation of forecasting models that are more accurate when presented with new data or, at least, provide information as to the limitations forecasting models face when applied to cryptoassets.

The accuracy or profitability of a model provides limited information on what is associated with or causes changes in price. This is particularly true when predicting the future price with exclusively historic price data [8, 132, 143, 158, 162, 189, 212, 229, 259], which leaves open whether other predictors could be included to enhance model performance further. Even when using a range of predictors (such as [151, 283]), there is the issue of which of these predictors improved, were irrelevant or even reduced model performance.

In summary, there is not one optimal forecasting approach in the literature that can be applied to the data with the results interpreted. The problem with finding the optimal forecasting approach is that this may vary according to the time period examined. Understanding what variables cause price fluctuations facilitates developing a more robust forecasting model [20] but simply examining the results of a comparison of forecast models provides limited insight into such causal relationships. Hence, this thesis moves the debate from how best to forecast the price to what variables are best supported as causing movements in price.

2.4.2 Causal Inference

In the absence of experimental data, quantitative analyses have to rely on observational data to find plausible causes of changes in the cryptoasset price series. This dataset consists of a cryptoasset price series and time-stamped social media posts from forums dedicated to that cryptoasset. Such observational data are analysed to support a candidate event as causing a change in price. This is analogous to healthcare epidemiologists relying on observational data to understand what causes a disease when the available experimental data are limited. Observational data were instrumental in determining the link between smoking and lung cancer [74, 251].

Observational data cannot, however, prove what caused price to change. An observed association between an event and price can be the result of 'confounding bias' [223] rather than a causal connection. Changes in a third variable may have caused the event to occur and the price series to alter, resulting in the observed association [223]. Alternatively, the impact of the event may depend on a catalyst that was unique to the dataset [252].

The following reviews previous methodologies that have been proposed relevant to understanding what variables are supported as causing a time series to change. This literature is divided into three types: Directed Graphical Causal Models that represent causal relationships as a graph (Section 2.4.2.1); Functional Causal Models that instead use mathematical functions (Section 2.4.2.2); and those that avoid the use of models (Section 2.4.2.3). The identified limitations of these methodologies inform the overall analytic approach discussed in Section 2.4.3 (answering research question 4), which guides the quantitative analyses conducted in Chapters 5 to 7.

2.4.2.1 Directed Graphical Causal Models

The Directed Graphical Causal Model (DGCM) represents causal relationships with directed graphs. The directed graph consists of nodes, each representing a variable, and arrows (termed 'directed edges') linking pairs of variables [117, 223]. If the variable at the tail of an arrow is changed, holding all other variables fixed, then the variable the arrow points to should be affected [117].

Algorithms that use data to search for the DGCM either start with an entirely unconnected graph and iteratively add edges [58] or with an entirely connected graph and iteratively remove edges (PC algorithm [267]), with some approaches capable of accounting for unobserved confounding bias (Fast Causal Inference [303]).

DGCM search algorithms assume both the Markov condition and Faithfulness assumption [254] and depend on reliable conditional independence testing, which may require large datasets [117] particularly when allowing for non-linear relationships [253]. Applying DGCM to time series data further requires that the causal relationships between variables do not vary across time [253]. The Markov condition stipulates that for each variable X in the graph, X is independent of all variables not affected by X, conditional on the direct causes of X [117,254]. The faithfulness assumption requires that all conditional independence relations between variables are a consequence of this Markov condition [253].

DGCMs typically represent causal links as occurring between only pairs of variables. In practice, multiple events may have to occur for the overall effect to be sufficient in changing another variable [252]. For example, the measles virus is insufficient in causing measles; the individual must also lack immunity to the virus [252]. Similarly, the discussion of bans on social media may decrease price only in the context of concern regarding government regulation. Ban discussion may even increase price if this is in the context of repealing bans and deregulation. Applications of DGCMs typically do not account for such synergistic effects [150]. Hence, when synergistic effects are present, the faithfulness assumption does not hold [253]. Remedying this issue through more complicated graphical descriptions [150] leads to the issue of DGCM search algorithms becoming computationally unfeasible [253].

2.4.2.2 Functional Causal Models

A subgroup of DGCMs is the Functional Causal Model (FCM). Under an FCM, mathematical functions are additionally assumed that link the values of variables being affected to the values of exogenous variables, whose causes are left unspecified, and noise terms representing variables that cannot be measured [117, 223]. Data can then be used to both tune the parameters in these functions and to test if the proposed FCM is consistent with the data. Linear regression is the typical approach applied in cryptoassets (Tables 2.3 to 2.17). This is despite the methodology being unreliable in the presence of extreme outliers [271], which was an observed feature of cryptoasset price series. The median change in bitcoin prices over 2 years (1 January 2017 to 3 December 2018) was only 0.3247%, but the largest rise was 27.97% on 20 July 2017 and greatest fall was 20.21% on 16 January 2018. Although the strict assumptions specific to linear regression can be relaxed within an FCM framework, because the FCM depends on functions being specified, there remains a need to make quantitative assumptions regarding the relationships between variables [117,223,254].

2.4.2.3 Model-Free Approaches

The difficulty in justifying which functions should be used to describe a causal relationship has motivated research into a variety of 'model-free' [195] approaches that the authors advocate particularly compared with linear regression [12, 85].

In cryptoasset research, wavelet analysis, causality-in-quantiles tests, KT correlation and transfer entropy have been proposed. Wavelet analysis has been applied in cryptoasset research to provide insight into short-term, medium-term and long-term associations between variables [174,227]. This, however, assumes that the different time series compared are normally distributed [126], an assumption not found to hold in cryptoasset price series [55, 219]. The causality-in-quantiles test [12] replaces examination of what drives the conditional mean of the variable affected with examination of the different quantiles. Applying to bitcoin returns [12] found an association between trading volume and bitcoin returns but only when bitcoin returns were near the median value. A linear regression found no association [12]. Aste [9] applied KT correlation and found associations across 1944 cryptoassets between sentiment and price.

Transfer entropy has been advocated for its flexibility [253], being capable of detecting all orders of correlation [131]. Of particular interest has been the ability of transfer entropy to capture non-linear relationships between variables [84, 161]. Transfer entropy has been applied to evaluate the non-linear interdependence between: the prices of different cryptoassets [84]; the prices and sentiment of different cryptoassets [9]; and the price and sentiment for a given cryptoasset [161].

In transfer entropy [256], Kullback-Leibler divergence is applied to determine if the entropy rate for the true joint distribution is the same as that assuming independence between the variables. In theory, a non-zero transfer entropy should indicate that the variables are associated. In practice, transfer entropy typically relies on probabilities estimated based on limited sample data, which can lead to non-zero values even between independent variables. Hence, transfer entropy estimates are often compared against a control. This control can be transfer entropy calculated using the shuffled values of the hypothesised cause [195] or calculated based on assuming the hypothesised cause and effect are unrelated and simulating values of the effect [85]. A further issue is that transfer entropy assumes that the values compared are discrete, taking only a few different possible values [256] and so there is a need to determine how best to discretise the data [85]. Although transfer entropy can specify that variables are related, it does not provide detail on how they are related; for instance, whether there exists an increasing or decreasing association.

Random forests have been used in equity market research as an alternative to transfer entropy in evaluating feature predictiveness non-parametrically. Random forests are ensembles of decision trees [36], which do not assume a linear predictor-price relationship [208]. Booth et al [27, 28] demonstrated that random forests were more accurate in predicting future price than linear regression, support vector regression and neural networks. Random forests can also be used to compare the predictiveness of different features [36]. Booth et al advocated combining evaluations of feature importance with backwards elimination to reduce the features considered to a predictive subset before training a prediction model [27, 28]. More sophisticated feature selection algorithms are also available such as ‘Boruta’ [176], which has been applied in selecting a subset of features predictive of the direction of bitcoin price movement [57]. Boruta incorporates statistical comparisons with introduced shuffled noise variables [176]. Similar to transfer entropy, random forests provide limited information on whether variables have an increasing or decreasing association.

2.4.2.4 Limitations

A common theme among causal inference methodologies (be it DGCM, FCM or a ‘model-free’ approach) is that judgement is required regarding what the likely causes of price variation are before testing these features against the data. This could mean potentially critical causes of price variation may be missed if they are not considered before testing. In such cases, if the cause drives variation in both the price and a feature being tested, spurious correlations could result, leading to misinterpretations regarding which features caused price variations [223].

Current adaptations of causal inference techniques to time series analysis [253, 254] assess whether a variable at each unit of time examined (such as daily, weekly or monthly price) depends on other features. Examining the day-to-day, week-to-week, month-to-month or some other unit of price series variation ignores the phasic nature of price series movement. Visualising the price series (Figure 1.1), reveals distinct phases of movement whereby prices display rising or falling patterns over periods lasting a few months. From 1 January to 12 June 2017, ether prices rose 4,748%, whilst, from 13 January to 6 April 2018, prices fell 73%. The issue is that an association found when prices are in a rising pattern may not persist when prices revert to falling – a problem that was identified with Google search data (Section 2.3.1.5). This thesis thus takes a different approach, examining what is associated with and potentially causes the movement between different price phases.

2.4.3 Selection of Analytic Approach

This thesis examines phases in price movement rather than day-to-day variation, and applies a non-parametric approach to minimise assumptions made in analysing the data. Words and topics are directly extracted from social media text using criteria that imply an association with price. The criteria applied varies across the different studies (Chapters 5 to 7) but a common theme is that judgement is not required in deciding which features might be associated with price before analysing the data. The exact methodology evolves from discovering words associated with a single phase of volatile but overall falling bitcoin prices (Chapter 5) to extracting topics associated with shifts in the phase of price movement (Chapter 6) to delineating potential causes of phasic shifts in price (Chapter 7).

Two approaches are developed to reduce the risk that the discovered associations between an event discussed and price are due to some third, unknown variable. The 'mono-phase' analysis finds events supported as having a major effect in causing a single phasic shift in price. The 'multi-phase' analysis finds events supported as having a recurring influence on rising or falling prices. These analyses are detailed further in Chapter 7. The ideas behind these analyses have been derived from studies in healthcare epidemiology where again only observational data were available [35, 146, 251].

Table 2.18: Forecast Models Literature Review 2014 – 2016.

Author	Input Features	Time scale	Measure of success	Models	Result	Comments
Shah and Zhang 2014 [259]	Price 10 sec	6/5/14– 24/6/14	Accumulated profit	Bayesian regression, Latent source model	89% return in 50 days with Sharpe ratio of 4.1.	Retrospective model, bitcoin price relatively stable during time period. Doubling investment in around 60 days.
Madan et al 2014 [189]	Price daily, 10 min, 10 sec. Daily Bitcoin network and market data	Unstated	The ability to predict price	Linear (GLM), SVM, RF (16 features daily or 10 sec/10 min intervals).	Excellent daily price results for GLM, poor results for SVM and high accuracy but lower precision for RF.	Ten-minute data for GLM and RF had 50-55% accuracy at predicting price.
Greaves and Au 2015 [123]	Price and transaction data	1/2/12– 1/4/13	Improved performance of neural network model	Linear regression, SVM, FFN 2 hidden layers	Neural network only slightly better than regression at predicting bitcoin price change.	No data on return on investment.
Almeida et al 2015 [6]	Price and Volume	1/7/13– 1/5/15	Accumulated profit versus Trend Follower (price change day before dictates sell or invest next day)	FFN using MATLAB	Best performing network has 30 hidden neurons and 6 input delays and is slightly better than the Trend Follower. Volume improves result.	Error in code dramatically changed results. Best returns when bitcoin price rose in last quarter 2013.
McNally 2016, 2018 [200, 201]	Price, difficulty and hash rate	19/8/13– 19/7/16	Improved performance of neural network model	RNN, ARIMA, LSTM	RMSE of LSTM, RNN and ARIMA models were 6.87%, 5.45% and 53.74% respectively.	RNN best at bitcoin forecasting. No data on return on investment.
Hegazy and Mumford 2016 [132]	Five left derivatives of price	Unstated	Accuracy and Profit	Weighted Linear and logistic regression, GBT, GDA, RRL	GBT most accurate but RRL most profitable.	RRL had accuracy similar to logistic regression. GBT accrued low profit.

Table 2.19: Forecast Models Literature Review 2017.

Author	Input Features	Time scale	Measure of success	Models	Result	Comments
Katsiampa 2017 [158]	Volatility	1/12/10–1/12/14	Model parameters for GARCH	GARCH Model and variants	Measure volatility. Demonstrated AR-CGARCH modification best.	No data on return on investment.
Amjad and Shah 2017 [8]	Price in Chinese Yuan	2014, 2015 and 2016	Model comparison in terms profit, return, Sharpe ratio and accuracy. Variable training, validation and testing periods	EC, LR, LDA, RF, Baseline: ARIMA	All models outperformed the baseline ARIMA.	High return bitcoin investment (e.g. 6-7x, 4-6x and 3-6x return on investments for tests in 2014, 2015 and 2016), while maintaining prediction accuracy (> 60-70%) and Sharpe Ratio (> 2.0).
Indera et al 2017 [143]	Opening, low, high and closing price	12/3/12–11/3/17	The ability to predict price	Non-Linear Autoregressive with Exogenous Inputs based on FFN	Fit seemed strong.	No data on return on investment.
Sin and Wang 2017 [262]	Bitcoin features, Price and Volume	Train: 2/5/15–30/4/17; Test: 1/5/17–20/6/17	Accuracy and Profit	GASEN with 5 FFNs, Baseline: price change day before dictates sell or invest next day	GASEN accuracy of 64% over 50 days of predictions.	$10,000 produced $13,800.64 (baseline) and $18,484.43 (GASEN).
Pichl and Kaizoji 2017 [229]	Price	1/2/12– 1/8/17	The ability to predict price	FFN (10 day moving window for daily log return sampling)	Model captured the logarithmic return density distribution.	No data on return on investment.
Xu and Medarametla 2017 [298]	Price and Volume	1/1/17–11/10/17	The ability to predict gains	WLR, PCA, FFN, Baseline: buy and hold	All three models had gains better than baseline with FFN highest.	No financial data on return on investment.

Table 2.20: Forecast Models Literature Review 2018 (datasets ending in 2017).

Author	Input Features	Time scale	Measure of success	Models	Result	Comments
Guo and Antulov-Fantulin 2018 [127]	Hourly price and order book data	1/9/15–30/4/17	Model comparison based on RMSE and MAE	EWMA, GARCH, BEGARCH, STR, ARIMA, ARIMAX, STRX, GBT, RF, XGT, ENET, GP (70% training, 10% validation, 20% testing)	Ensemble method XGT and regularized regression ENET outperform other methods in most cases.	Compared models. No financial data on return on investment.
Jang and Lee 2018 [151]	Blockchain information (e.g. trading volume and miner revenue), macroeconomic (e.g. stock indices, gold, crude oil), currency ratios	13/9/11–21/7/17	Model comparison based on RMSE and MAPE	SVR, Linear regression, BNN (16 features lowest VIF, 200 days training), BNN applied with rollover framework	BNN better RMSE (for log price and log volatility (0.0069, 0.2325) and MAPE (0.0180, 0.5222) than linear regression (0.0935, 0.4823, 0.0712, 0.6263) and SVR (0.2742, 0.0404, 0.5297, 0.8629) (16 feature test data).	Emphasised importance of selecting features and applying a rollover framework with BNN.
Torres and Qiu 2018 [283]	Price and related from Cryptoassets (108), Currencies (30), Stocks (138) and Commodities (13). Includes Google Trends data.	1/5/13–30/9/17; Test: September 2017	Model comparison based on prediction accuracy.	RNN Models (GRU and LSTM), ARIMA, ARIMA without dynamic regression	Best RMSE LSTM 272.96 BTC/USD, GRU 274.02 BTC/USD, ARIMA 255.9 BTC/USD deteriorating to 1196.19 BTC/USD without dynamic regression.	Compared models. No financial data on return on investment.

Table 2.21: Forecast Models Literature Review 2018 (datasets ending in 2018).

Author	Input Features	Time scale	Measure of success	Models	Result	Comments
Khaldi et al 2018 [162]	Closing price	18/7/10–17/1/18	The ability to predict price	AR-GARCH, ELMAN (an RNN), EEMD-ELMAN Model (70% training, 15% validation, 15% testing)	EEMD-ELMAN (RMSE: 0.03; MAE: 0.02) outperformed ELMAN (RMSE: 0.19; MAE: 0.12) and AR-GARCH (RMSE: 0.04 and MAE: 0.03)	Compared models. No data on return on investment.
Nakano et al 2018 [212]	Price every 15 minutes. Training data 38,000 observations.	1/8/17–24/1/18	The ability to predict price return.	FNN 7 layers, Baseline: buy and hold	FNN better than buy and hold. Three different FNN strategies demonstrate: as number of classes less, the FNN performs better.	Compares models. No financial data on return on investment.
Alessandretti et al 2018 [5]	Daily price; market capitalisation value, normalised and rank; trading volume and age	11/11/15–24/4/18	Model comparison based on return on investment from portfolio based on model	LSTM, GBT, Simple Moving Average	LSTM highest returns; moving average lowest. LSTM profitable supposing transaction fees of up to 1%.	Examines how to design a profitable portfolio based on 1,681 cryptoassets excluding Bitcoin thus model accuracy not compared.

Chapter 3

Cryptoasset Price Co-movement

3.1 Introduction

Cryptoasset prices could be analysed as individual entities with the results found for different cryptoassets compared. Alternatively, all cryptoassets could be combined into a single entity, with analyses examining the causes of fluctuations in the combined price across time. This chapter applies a quantitative perspective to whether cryptoassets are sufficiently different to merit analysing cryptoassets individually rather than as a group (research question 1).

Whilst the literature has tested the statistical significance of associations between different cryptoasset price series [9, 85], the aim of this chapter is to assess and compare the strength of the associations between different price series.

If two cryptoasset price series were very strongly correlated, this would be consistent with very similar factors influencing the values of both cryptoassets. Hence, this would provide evidence supporting analysing both cryptoassets together as a single entity. Otherwise, the correlation value would suggest that the factors influencing the price differed across the cryptoassets. Here, treating the two cryptoassets as a single entity might obfuscate causes of price variation that were unique to one or other cryptoasset.

The prices of different cryptoassets are compared through measuring the correlations between the prices of different cryptoassets. Particularly popular correlation metrics are: Pearson's Product Moment Correlation Coefficient (PMCC); Spearman's Rho (SR); and Kendall's Tau (KT) [194]. Applying the PMCC assumes that cryptoasset returns follow normal distributions [299], which previous research has suggested to be an unreasonable assumption [55, 219]. The PMCC is further restricted in measuring linear relationships [299]. Hence, this chapter examines two non-parametric correlation measures: SR and KT. This paper primarily uses the SR methodology, with KT also being applied as a check on the robustness of the results. A cut-off of SR being more than 0.9 is used to indicate a very strong correlation and a cut-off of less than 0.1 a negligible correlation [255].

These results include the correlations between the bitcoin and ether price, which will be used to assess the extent there is a weak or strong correlation between these two cryptoassets' prices. The less correlated two different cryptoasset prices are, the stronger the case for examining the two cryptoassets separately and comparing results to better understand the causes of price variation unique to each cryptoasset. This chapter is based on research presented at the Cryptocurrency Research Conference 2018 [39].

3.2 Data Preparation

3.2.1 Choice of Cryptoasset

The focus is on the top ten cryptoassets by market capitalisation or liquidity. Reasons for focussing on larger cryptoassets were given in Section 2.2.1.

3.2.2 Choice of Dataset

All data were sourced from coingecko.com on 6 March 2018. This source covered thousands of cryptoassets and enabled downloading data in a CSV format.

As different cryptoassets were launched in different years, data availability varied. Two datasets were thus created. In the first dataset, all the cryptoassets were considered, which required beginning the time series from 9 November 2017. In the second dataset, a subset of cryptoassets where there was more data was considered. This enabled beginning the time series from 9 September 2016. Considering different time periods ensured greater robustness to the instability in correlation values over time [111, 219].

3.2.3 Preparing the Price Data

The daily US Dollar price of each cryptoasset selected was gathered from coingecko.com on 6 March 2018. This source had missing data for: 22 February 2018 (11 cryptoassets) and 8-10 August 2017 (NEO). Dates where a cryptoasset lacked data were removed from the dataset.

Rather than comparing the raw price series, the daily percentage change in price for each cryptoasset was calculated. This provided a closer proxy to the returns an investor would have received if they had held a particular cryptoasset on a certain day. As this calculation involved first differencing, it was also more robust should there be nonstationarity problems in the dataset [271].

3.3 Methodology

The approach was to measure the correlations between each pair of cryptoassets' daily percentage change in price and then to depict these results through a correlation network diagram. Section 3.3.1 provides the formulae for the SR and KT measures of correlation whilst Section 3.3.2 explains how the correlation network diagram was constructed.

3.3.1 SR and KT Formulae

Equation 3.1 is the formula [302] for the SR between a series x and y, where R_a is the variable a ranked by magnitude and $\bar{R_a}$ is the arithmetic mean value for the variable R_a.

$$SR(x,y) = \frac{\sum(R_{xi} - \bar{R_x})(R_{yi} - \bar{R_y})}{\sqrt{\sum(R_{xi} - \bar{R_x})^2 \sum(R_{yi} - \bar{R_y})^2}} \tag{3.1}$$

In interpreting SR, a cut-off of more than 0.9 is used to indicate a very strong correlation and a cut-off of less than 0.1 a negligible correlation, which are popular thresholds in the literature [255].

KT was used to check for robustness. Equation 3.2 is the formula [159] for the KT between a series x and y, where:

- n_c is a count of the number of pairs of values where the ordering in series x matches that in y
- n_d is a count of the number of pairs where the ordering does not match
- pairs of tied values are ignored in the above counts
- n_a is the total number of pairs of datapoints that are not tied in series a

$$KT(x,y) = \frac{n_c - n_d}{\sqrt{n_x} \times \sqrt{n_y}} \tag{3.2}$$

3.3.2 Correlation Networks

The top ten correlation values were depicted in a correlation network. A network consists of circular nodes connected by lines called edges. Here, the nodes represent the daily returns for different cryptoassets whilst each edge has a weight that is the correlation between the linked cryptoassets' returns. Diagrammatically, the stronger the association between two cryptoassets' returns, the wider the line connecting their nodes [96]. The networks were initially created using SR. The networks were then redrawn to evaluate the impact of switching the correlation measure to KT. To aid interpretability, the nodes were arranged such that more correlated cryptoassets were placed closer together. This cannot always be perfectly achieved in a two-dimensional space [96], so instead an approximate force-embedded algorithm approach was applied [110].

3.3.3 Software

Correlation networks were applied using the programming language R. The correlation matrices were implemented using base R function 'cor', which did not require the installation of additional packages. The correlation network was implemented using the package qgraph [96], which was specifically designed for this process.

3.4 Results

3.4.1 Cryptoassets Selected

The cryptoassets selected were: Bitcoin, Litecoin, Ethereum, Ethereum Classic, Monero, NEO, Bitcoin Cash, Tron, Cardano, Qtum, Ripple, EOS, Stellar and USD Tether.

Of these, the following had data available from 9 September 2016: Bitcoin, Ethereum, Ethereum Classic, Litecoin, Monero, NEO, Ripple, Stellar and USD Tether. Table 3.1 lists the abbreviations used for the different cryptoassets.

Table 3.1: Cryptoassets, abbreviations and data availability

Cryptoasset	Abbreviation	Data Start
Bitcoin	btc	9 September 2016
Ethereum	eth	9 September 2016
Ethereum Classic	etc	9 September 2016
Litecoin	ltc	9 September 2016
Monero	xmr	9 September 2016
NEO	neo	9 September 2016
Ripple	xrp	9 September 2016
Stellar	xlm	9 September 2016
USD Tether	usdt	9 September 2016
Bitcoin Cash	bch	9 November 2017
Cardano	ada	9 November 2017
EOS	eos	9 November 2017
Tron	trx	9 November 2017
Qtum	qtm	9 November 2017

3.4.2 Correlation Values

3.4.2.1 Spearman's Rho

Figure 3.1 shows the top ten correlations for each time period considered, and displays these results as correlation network diagrams. Table 3.2 then provides all the SR values for all cryptoassets in the smaller dataset (from 9 November 2017 to 6 March 2018) whilst Table 3.3 provides the SR values for the subset of cryptoassets where data were available for the longer dataset (from 9 September 2016 to 6 March 2018).

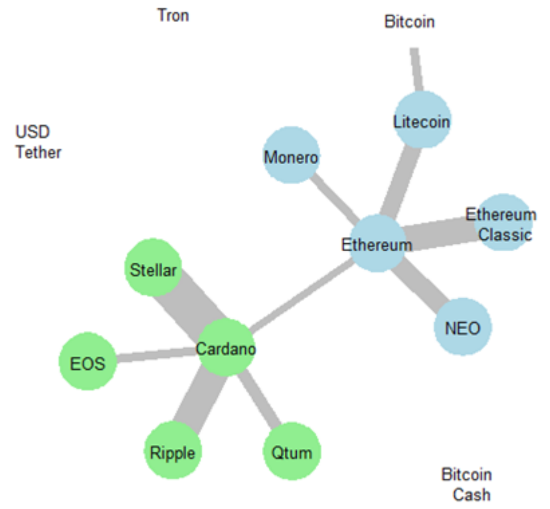
3.4.2.2 Kendall's Tau

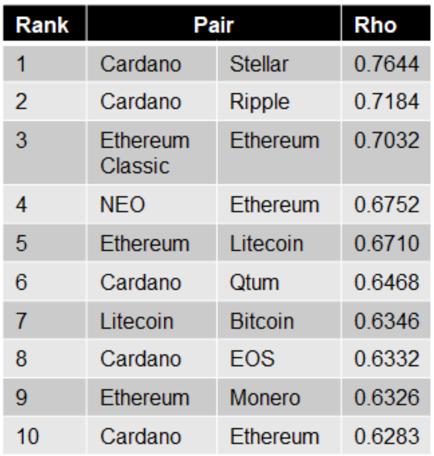
Using KT led to results similar to SR. There were the following exceptions in the smaller dataset (from 9 November 2017 to 6 March 2018):

- The KT correlation between Qtum and Ethereum and between Qtum and Cardano are in the top ten
- The link between Ethereum and Monero is lost

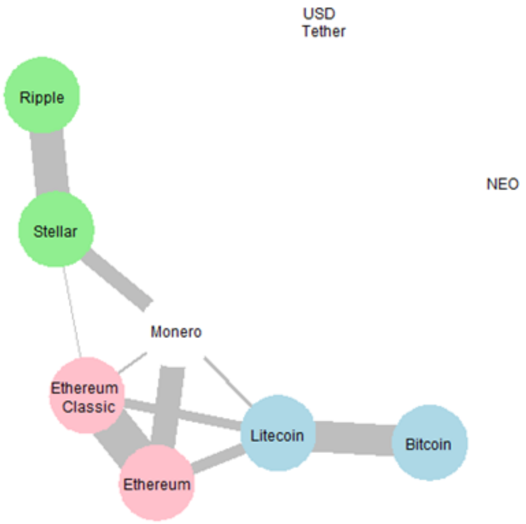
Examining KT in the larger dataset (from 9 September 2016 to 6 March 2018), Bitcoin's correlation with Monero is in top ten rather than Ethereum Classic's association with Monero.

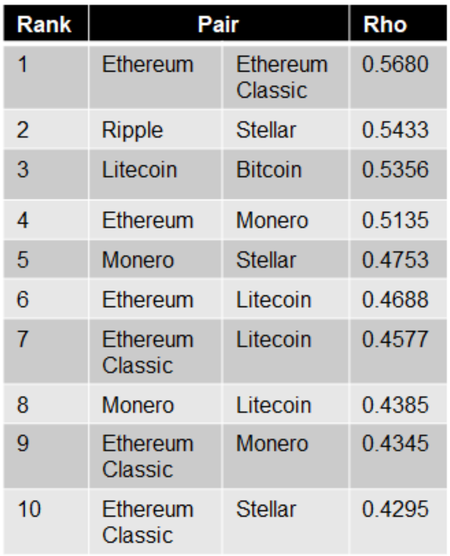
Figure 3.1: Correlation network diagrams depicting the Spearman's rho correlations between different cryptoasset returns. Each node represents a cryptoasset and each edge represents a correlation between returns. Links in the top 10 of correlation values are displayed to improve interpretability. The first diagram relates to data from 9 November 2017 to 6 March 2018, whilst the second relates to data from 9 September 2016 to 6 March 2018 (which constrained the list of cryptoassets considered). This was presented at the Anglia Ruskin Cryptocurrency Research Conference 2018 [39]





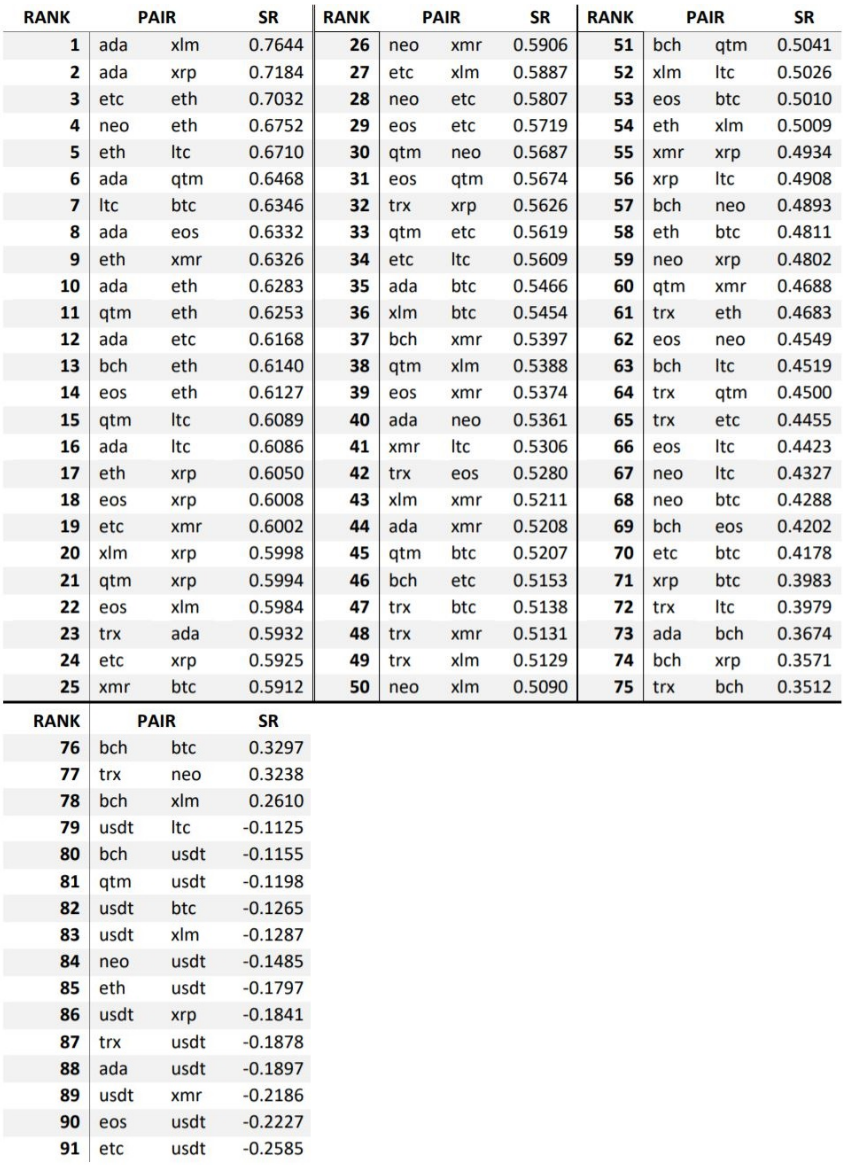
Rank	Pair		Rho
1	Cardano	Stellar	0.7644
2	Cardano	Ripple	0.7184
3	Ethereum Classic	Ethereum	0.7032
4	NEO	Ethereum	0.6752
5	Ethereum	Litecoin	0.6710
6	Cardano	Qtum	0.6468
7	Litecoin	Bitcoin	0.6346
8	Cardano	EOS	0.6332
9	Ethereum	Monero	0.6326
10	Cardano	Ethereum	0.6283





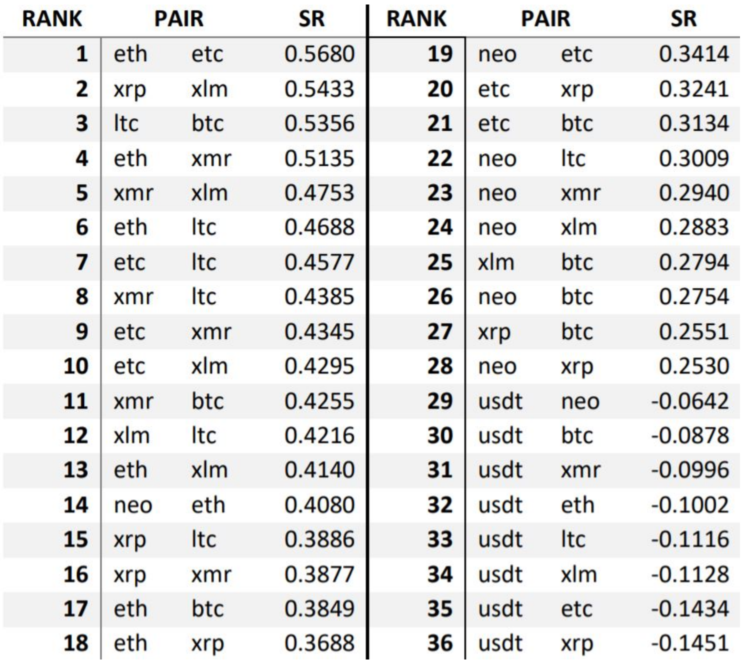
Rank	Pair		Rho
1	Ethereum	Ethereum Classic	0.5680
2	Ripple	Stellar	0.5433
3	Litecoin	Bitcoin	0.5356
4	Ethereum	Monero	0.5135
5	Monero	Stellar	0.4753
6	Ethereum	Litecoin	0.4688
7	Ethereum Classic	Litecoin	0.4577
8	Monero	Litecoin	0.4385
9	Ethereum Classic	Monero	0.4345
10	Ethereum Classic	Stellar	0.4295

Table 3.2: Spearman's Rho (SR) correlations between the daily percentage change in prices from 9 November 2017 to 6 March 2018. This considers all cryptoassets listed in Table 3.1; see Table 3.1 for meaning of abbreviations.



RANK	PAIR		SR
1	ada	xlm	0.7644
2	ada	xrp	0.7184
3	etc	eth	0.7032
4	neo	eth	0.6752
5	eth	ltc	0.6710
6	ada	qtm	0.6468
7	ltc	btc	0.6346
8	ada	eos	0.6332
9	eth	xmr	0.6326
10	ada	eth	0.6283
11	qtm	eth	0.6253
12	ada	etc	0.6168
13	bch	eth	0.6140
14	eos	eth	0.6127
15	qtm	ltc	0.6089
16	ada	ltc	0.6086
17	eth	xrp	0.6050
18	eos	xrp	0.6008
19	etc	xmr	0.6002
20	xlm	xrp	0.5998
21	qtm	xrp	0.5994
22	eos	xlm	0.5984
23	trx	ada	0.5932
24	etc	xrp	0.5925
25	xmr	btc	0.5912
26	neo	xmr	0.5906
27	etc	xlm	0.5887
28	neo	etc	0.5807
29	eos	etc	0.5719
30	qtm	neo	0.5687
31	eos	qtm	0.5674
32	trx	xrp	0.5626
33	qtm	etc	0.5619
34	etc	ltc	0.5609
35	ada	btc	0.5466
36	xlm	btc	0.5454
37	bch	xmr	0.5397
38	qtm	xlm	0.5388
39	eos	xmr	0.5374
40	ada	neo	0.5361
41	xmr	ltc	0.5306
42	trx	eos	0.5280
43	xlm	xmr	0.5211
44	ada	xmr	0.5208
45	qtm	btc	0.5207
46	bch	etc	0.5153
47	trx	btc	0.5138
48	trx	xmr	0.5131
49	trx	xlm	0.5129
50	neo	xlm	0.5090
51	bch	qtm	0.5041
52	xlm	ltc	0.5026
53	eos	btc	0.5010
54	eth	xlm	0.5009
55	xmr	xrp	0.4934
56	xrp	ltc	0.4908
57	bch	neo	0.4893
58	eth	btc	0.4811
59	neo	xrp	0.4802
60	qtm	xmr	0.4688
61	trx	eth	0.4683
62	eos	neo	0.4549
63	bch	ltc	0.4519
64	trx	qtm	0.4500
65	trx	etc	0.4455
66	eos	ltc	0.4423
67	neo	ltc	0.4327
68	neo	btc	0.4288
69	bch	eos	0.4202
70	etc	btc	0.4178
71	xrp	btc	0.3983
72	trx	ltc	0.3979
73	ada	bch	0.3674
74	bch	xrp	0.3571
75	trx	bch	0.3512
76	bch	btc	0.3297
77	trx	neo	0.3238
78	bch	xlm	0.2610
79	usdt	ltc	-0.1125
80	bch	usdt	-0.1155
81	qtm	usdt	-0.1198
82	usdt	btc	-0.1265
83	usdt	xlm	-0.1287
84	neo	usdt	-0.1485
85	eth	usdt	-0.1797
86	usdt	xrp	-0.1841
87	trx	usdt	-0.1878
88	ada	usdt	-0.1897
89	usdt	xmr	-0.2186
90	eos	usdt	-0.2227
91	etc	usdt	-0.2585

Table 3.3: Spearman's Rho (SR) correlations between the daily percentage change in prices from 9 September 2016 to 6 March 2018. This considers the subset of cryptoassets with data from 9 September 2016 as specified in Table 3.1; see Table 3.1 for meaning of abbreviations.



RANK	PAIR		SR	RANK	PAIR		SR
1	eth	etc	0.5680	19	neo	etc	0.3414
2	xrp	xlm	0.5433	20	etc	xrp	0.3241
3	ltc	btc	0.5356	21	etc	btc	0.3134
4	eth	xmr	0.5135	22	neo	ltc	0.3009
5	xmr	xlm	0.4753	23	neo	xmr	0.2940
6	eth	ltc	0.4688	24	neo	xlm	0.2883
7	etc	ltc	0.4577	25	xlm	btc	0.2794
8	xmr	ltc	0.4385	26	neo	btc	0.2754
9	etc	xmr	0.4345	27	xrp	btc	0.2551
10	etc	xlm	0.4295	28	neo	xrp	0.2530
11	xmr	btc	0.4255	29	usdt	neo	-0.0642
12	xlm	ltc	0.4216	30	usdt	btc	-0.0878
13	eth	xlm	0.4140	31	usdt	xmr	-0.0996
14	neo	eth	0.4080	32	usdt	eth	-0.1002
15	xrp	ltc	0.3886	33	usdt	ltc	-0.1116
16	xrp	xmr	0.3877	34	usdt	xlm	-0.1128
17	eth	btc	0.3849	35	usdt	etc	-0.1434
18	eth	xrp	0.3688	36	usdt	xrp	-0.1451

3.5 Discussion

The absence of very strong correlations reflected that the prices of different cryptoassets did not move in perfect synchrony. In both datasets examined, the SR values were all below the popular 0.9 cut-off for a very strong correlation [255]. All SR values in the longer dataset (9 September 2016 to 6 March 2018) were found to be less than 0.6, which some conventions would consider only a 'moderate' correlation [255]. This supports analysing each cryptoasset individually and then comparing the results.

The results further reveal the distinctiveness of USD Tether as a cryptoasset. This reflects the distinct functionality of USD Tether tokens, with its status as a 'crypto-voucher' token discussed further in Chapter 4. Whilst the correlation values between most cryptoasset returns were positive, USD Tether was negatively (albeit weakly [255]) correlated with the other cryptoassets.

The lack of association was as expected because, unlike the other cryptoassets, one USD Tether is exchangeable for one US Dollar and so price volatility is minimal across time [187]. If the USD Tether price were less than one US Dollar, there is an incentive to buy USD Tether and exchange for US Dollars to make a profit. This higher demand would likely raise the price of USD Tether until the profit opportunity is no more. If the USD Tether price exceeded one US Dollar, there is instead an incentive to swap US Dollars for USD Tether and sell the USD Tether for a profit. This higher supply would likely diminish the USD Tether price until the profit opportunity is removed. Hence, whilst the values of most cryptoassets fluctuate across time, the price of USD Tether is unlikely to move far from one US Dollar, consistent with the weak, negative correlation observed between USD Tether and other cryptoassets.

Why this correlation was consistently negative required further investigation. Examining trading exchange data from coingecko.com provides a possible explanation. Cryptoassets are often bought using USD Tether, and so USD Tether is often sold whilst a cryptoasset is being bought. This suggests that sudden increases in the demand for cryptoassets (raising their prices) is likely to coincide with sudden increases in the supply of USD Tether (decreasing the USD Tether price), which could explain the observed negative correlation.

In terms of which crypto-fuel and crypto-transaction systems should be analysed, the results support the validity of comparing Bitcoin with Ethereum. The correlation between these prices was not in the top ten in either dataset. Examining the longer dataset (Table 3.3), the SR value was only 0.3849 compared with a correlation value of 0.5356 between Bitcoin and Litecoin. This is consistent with distinct variables existing that affect either Bitcoin or Ethereum but not both. This motivates comparing Bitcoin and Ethereum to understand these distinctive factors.

Chapter 4

Cryptoasset Classification and Analysis of Non-Conventional Fundamentals

4.1 Introduction

This chapter first develops a cryptoasset classification (results in Section 4.3.2) to analyse, from a qualitative perspective, research question 1 on whether cryptoassets vary sufficiently to merit analysing each cryptoasset individually. Then, fundamentals are derived from the classification that, in theory, may underpin the valuations of different cryptoassets. Hence, these fundamentals are used to inform the elucidation of theoretical causes of price variation for each type of cryptoasset (see Section 4.4), in response to research question 5.

Following on from limitations found in existing cryptoasset classifications (detailed in Section 2.1.2), criteria are developed and applied that consider the characteristics of the tokens. This results in a new cryptoasset classification (see Section 4.3.2) that defines different types of cryptoasset according to certain facets of the functionality of the token that are shared among constituent cryptoassets.

These facets could potentially provide reasons why a participant might buy or sell a token other than to profit from an increase in price. Over time, these reasons for buying or selling a token may become more or less valid, which might cause the popularity of that cryptoasset to vary, causing fluctuations in price. For instance, a major retailer may decide to accept a cryptoasset as a form of money which would make that cryptoasset more popular, raising the price. This suggests that the cryptoasset classification might help to explain cryptoasset price variation.

A analysis of the fundamentals is performed on each type of cryptoasset identified in the cryptoasset classification. These analyses are displayed as an assessment framework (see Section 4.4). The framework translates the characteristics of the different token types into the risks and benefits relevant to an owner holding each category of token. This goes beyond the profit that may be received from buying the cryptoasset at a low price and selling at a higher price. This provides considerations, other than speculation, a potential buyer of a cryptoasset within a certain category might evaluate before making a purchase decision or a holder of a token might consider before deciding whether to sell. Such considerations provide insights into the theoretical fundamentals underpinning price. These will be supplemented by the quantitative analyses of social media in Chapters 5 to 7, as part of the overall ‘quantamental’ analytic approach of this thesis.

This chapter is based on research published in the journal *Ledger* [40]. The cryptoasset classification developed was also included with written evidence provided by Eversheds Sutherland (International) LLP and published by the UK Parliament Digital Currencies Inquiry (reference: DGC0020) [100].

4.2 Methodology

4.2.1 Scope

The cryptoassets considered are those where the token is:

1. an entirely digital store of value
2. publicly available
3. supported by a blockchain

Publicly available cryptoassets are likely to have the most available data, whilst the support of the blockchain has been seen as a differentiating characteristic of cryptoassets [137, 235]. Using this scope, the most financially significant cryptoassets are selected for the dataset.

4.2.2 Determining Financial Significance

The most financially significant cryptoassets are found using market capitalisation or liquidity, as justified in Section 2.2.1. For robustness, two metrics from three websites are examined at multiple timepoints [63, 64, 67–69, 105, 106]. Lists of the top ten cryptoassets by market capitalisation and liquidity were collected from coinmarketcap.com at 14:27 on 4 October 2017, 15:48 on 30 October 2017, and 10:27 on 18 December 2017. Examining coincap.io at 15:58 on 30 October 2017 and 10:28 on 18 December 2017 corroborated with coinmarketcap.com, whilst examining onchainfx.com led to similar results, except that this website did not include Tether in its rankings. Where two lists disagree, cryptoassets from both rankings are included.

The top five ICOs by amount raised as of 18 December 2017 are also included [264]. To mitigate against the risk of cryptoassets failing to launch, coinmarketcap.com was used to restrict the list to where either the tokens or futures exchangeable for the tokens could be bought.

4.2.3 Criteria Applied to Each Cryptoasset

1. What is the purpose, the functionality, and the rights associated with the token?

2. How is the supply of tokens determined over time?

3. How is the cryptoasset related to other cryptoassets?

These questions characterise the fundamentals of each cryptoasset, the characteristics that bring value to owning a token other than anticipation of a price increase. These criteria were applied using information on each cryptoasset sourced from whitepapers, official websites, and third-party commentary.

4.3 Results

4.3.1 Dataset

The financially most significant cryptoassets were found to be: Bitcoin, Ethereum, Ripple, Bitcoin Cash, Litecoin, Dash, NEM, NEO, Monero, Ethereum Classic, Tether, Qtum, Zcash, Cardano, Bitcoin Gold, EOS, AirSwap, Filecoin, the Bancor Protocol, Qash, and Kin. Financial information associated with these cryptoassets is provided in Tables 4.1 and 4.2. The veracity of the information released by BitConnect has been questioned [105, 279] and, similarly, Tezos was involved in accusations of dishonesty [147, 148]. Both are excluded.

Table 4.1: Market capitalisation and liquidity (transaction volume over last 24 hours) measured in USD for the cryptoassets selected based on these metrics, as of 10:27 on 18 December 2017.

Cryptoasset	Market Capitalisation	Liquidity
Bitcoin	318,567,613,388	13,070,000,000
Bitcoin Cash	31,514,053,090	877,377,000
Bitcoin Gold	5,190,273,036	199,003,000
Cardano	12,661,355,262	349,895,000
Dash	8,456,546,893	250,788,000
EOS	4,593,527,046	387,014,000
Ethereum	69,594,352,659	2,062,100,000
Ethereum Classic	3,528,696,852	493,391,000
Litecoin	17,294,853,905	1,198,410,000
Monero	5,411,241,508	182,633,000
Ripple	28,770,594,399	1,072,940,000
NEM	7,150,157,999	93,046,400
NEO	4,820,946,000	532,062,000
Qtum	3,163,793,147	1,147,910,000
Tether	1,128,439,474	2,070,980,000
Zcash	1,531,318,793	331,762,000

Table 4.2: Amount raised in US Dollars at ICO for the top five cryptoassets selected on this basis, with data sourced from `smithandcrown.com/icos` at 14:46 on 18 December 2017 [264].

Cryptoasset	Amount Raised
AirSwap	365,000,000
Filecoin	262,000,000
Bancor Protocol	153,000,000
Qash	108,170,000
Kin	97,500,000

4.3.2 Classification

The analysis identifies three groups ('crypto-transaction', 'crypto-fuel' and 'crypto-voucher') as well as 'hybrids' and the potential overlap between categories. How cryptocurrencies were allocated to the different groups is shown in Table 4.3 after the description of the different categories.

4.3.2.1 Crypto-transaction

A crypto-transaction cryptoasset is defined as a cryptoasset that is designed primarily for transacting value and therefore to be a form of 'electronic cash' [211].

Crypto-transaction tokens are usually designed to be easily transferrable, with minimal barriers to acquisition. Value is not derived from some underlying asset, but rather it is determined by a network of users (see Section 4.4). Among the cryptoassets examined (except Monero), this value was further supported by fixing the total amount of tokens that will ever be created. Examining the websites of crypto-transaction systems suggests that the availability of exchanges and/or merchants who will accept the tokens is an important consideration. Electronic cash is only useful if it can be exchanged directly for goods or services, or if exchange can occur easily through some other currency.

Crypto-transaction tokens were the first form of cryptoasset, beginning with Bitcoin in 2009. Despite this, new systems are still being created, such as Bitcoin Cash, Bitcoin Gold, Qash, and Kin in 2017. The development of a new codebase usually focuses on resolving perceived limitations in a previous attempt to create electronic cash (typically Bitcoin). The underlying code is often an amended copy of that of an older token, except for Qash and Kin. Even Bitcoin was developed to remove a perceived limitation, specifically the dependence of previous electronic cash systems on a central governing entity [211].

Improvements focus on speeding transactions [65, 91, 207, 249]; changing the mining algorithm to prevent centralisation [34, 65, 119, 207]; improving scalability [53, 207]; and enhancing liquidity [186]. There is a distinct subgroup that is concerned with privacy (Dash [91], Monero [207] and Zcash [19,34]), a finding corroborated by other researchers [76,214,296]. There is a second subgroup where the crypto-transaction token was developed to support a specific platform that can provide a suite of financial (Ripple [248] and Qash [186]) or social media (Kin [144]) services. Dash is unusual in seeking to change the governance structure through enabling network participants to vote on governance and budgeting proposals [91]. How improvements are prioritised and the strategies pursued to implement a given enhancement is system-specific.

4.3.2.2 Crypto-fuel

A crypto-fuel cryptoasset is defined as a cryptoasset intended to enable developers to create blockchain-supported applications. They are typically launched with a blockchain platform that is designed to enable the token to be used as a fuel for the created applications to operate. It is a term sourced from the Ethereum whitepaper [97].

The blockchain platform often has *smart contract* functionality, which enables the creation of accounts that behave in a pre-programmed, rule-based way in response to changes in the network, and so forms the basis of decentralised applications [62,78,95,97,140,236].

The blockchain platform can be used to facilitate ICOs, explaining the popularity of basing ICOs on crypto-fuel systems, such as with Etherparty, the Bancor Protocol, and CoinDash all based on Ethereum, and Ecobit on NEM. The blockchain platform can, however, also be more broadly applied to create a new crypto-voucher system (examples include the Bancor Protocol discussed in the next section), or some other type of network that runs independently of a central authority.

Crypto-fuel development usually starts as a fresh project (as with Ethereum and NEM) or as a fork from some other crypto-fuel's codebase (as with Ethereum Classic). They rarely evolve just from a crypto-transaction system. The new codebase typically focuses on improving the process for creating blockchain applications over a prior cryptoasset. This can mean simplifying the creation of applications [78,87,95,97,108,140,236]; raising flexibility [87,140]; improving scalability [78, 95, 140]; easing regulatory compliance [140]; preventing subsequent changes to the code [62]; or reducing the costs of usage [95].

The underlying architecture behind crypto-fuels varies significantly both from the perspective of the experience of the developer in creating an application to how the cryptoasset is created and distributed. Developers may have to learn a new programming language [62, 78, 97, 140], or be able to use a preferred language [87, 95, 236], whilst cryptoasset supply might be fixed [87], increase indefinitely [97], or increase up to a fixed cap [62].

4.3.2.3 Crypto-voucher

A crypto-voucher cryptoasset is defined to be a cryptoasset whose tokens carry the right to a predefined asset.

The asset to which the token-holder has rights varies. For example, USD Tether is exchangeable one-to-one with the US Dollar (or equivalent spot value in Bitcoin) [187]; tokens on the Bancor Protocol are exchangeable at fixed ratios with other cryptoassets [135]; and Filecoin tokens will be transferrable for data storage space [178]. In AirSwap, the token is temporarily locked up to register signals to peers of an intention to buy or sell Ethereum-based tokens [276].

As well as depending on the demand for an underlying asset, crypto-voucher tokens are also often dependent on one or more external blockchains. In the case of Filecoin, this dependence means the existence of bridges that enable participants to exploit the functionality of multiple other blockchains [178]. In contrast, the Bancor Protocol and AirSwap are run on top of the Ethereum blockchain [135, 277], whilst Tether uses the Omni Layer protocol, which runs on the Bitcoin blockchain [187].

Crypto-voucher systems are usually not the most dominant cryptoassets from the perspective of liquidity or market capitalisation (except for Tether), but are more prevalent among recent ICOs (AirSwap, Filecoin, and the Bancor Protocol).

4.3.2.4 Hybridisation

The distinction between crypto-fuel and crypto-transaction cryptoassets can be complicated by market forces turning crypto-fuel tokens into a store of value, in this respect taking on the properties of a crypto-transaction token; conversely, in some cases, the creation of new protocols is used to give additional crypto-fuel functionality to a crypto-transaction cryptoasset.

The extent to which such 'hybridised' cryptoassets fulfil an alternative role determines the extent to which the considerations associated with that other role are relevant (Figure 4.1). For example, Bitcoin was designed for transacting value and thus put in the crypto-transaction group [211]. Subsequently, the Omni Layer was developed so that Bitcoin could acquire crypto-fuel functionality [278]. However, the primary function for the Bitcoin token continues to be in transacting value and so it remains in the crypto-transaction group. Ethereum is in the crypto-fuel group but market forces have sometimes used it to purchase goods and services from merchants, although, in practice, this is very difficult [152]. The Ethereum whitepaper continues to describe ether as a 'crypto-fuel' [97].

4.3.2.5 Overlap

Linked to hybridisation is the issue of overlap, in particular between crypto-fuel and crypto-transaction tokens. Determining which cryptoasset falls within each of these categories will therefore require a determination of the primary function of the relevant cryptoasset. The starting point for forming this judgement was how the functionality of the token was explained within its whitepaper, as this is the best evidence of the original design over the token. As tokens evolve, a value judgement may be required to determine the primary function of the cryptoasset in question. This involves a consideration of how market participants are actually using the cryptoasset, and the effect of changes to the design, for example as a result of votes on its use, or changes shown by later whitepapers. This capacity for evolution may be part of the original code used for a cryptoasset, for example operation codes (opcodes) were baked into the original Bitcoin design which, although not part of the original function, were later reactivated in Bitcoin Cash, making it both spendable and compatible with smart contracts [244]. Activation events such as these may alter the classification of the cryptoasset and reinforces the fact that different market participants may legitimately come to different conclusions of how a cryptoasset is to be categorised, based on a different judgement of the primary function of a given cryptoasset. These differences in view point will reflect the fact that, for some cryptoassets, the primary function changes depending on the scenario in which it is used and, therefore, a cryptoasset may have multiple concurrent uses. However, in this case the distinction between crypto-fuel and crypto-transaction cryptoassets is still important to analysing a particular use.

Table 4.3: Allocation of cryptoassets across the different groups. The development of crypto-fuel functionality for Qash is discussed in the associated whitepaper [186].

Crypto-Transaction	**Crypto-Fuel**	**Crypto-Voucher**
Bitcoin	Ethereum	AirSwap
Bitcoin Cash	Ethereum Classic	Bancor Protocol
Dash	NEM	Filecoin
Litecoin	NEO	Tether
Monero	Qtum	
Ripple	Cardano	
Zcash	EOS	
Qash (currently)	Qash (planned)	
Bitcoin Gold		
Kin		

4.4 Implications for the Analysis of Fundamentals

Figure 4.1 provides a framework for questions which might be relevant for buyers and sellers assessing a given cryptoasset. This is not intended to cover all the potential risks and opportunities that may be associated with a cryptoasset. Instead, this framework translates the characteristic that defined each type of cryptoasset into implications of that characteristic for buyers and sellers of cryptoassets. The focus is on the cryptoasset token as this is the entity being bought or sold and so the part of the cryptoasset system that is likely to be relevant for the buyers or sellers.

The implied theoretical fundamentals that underpin cryptoasset value are to be compared with the results of the subsequent quantitative analyses (in Chapter 8). Any discrepancies between the quantitative analyses results and these identified fundamentals would be suggestive of the price-relevance of factors outside the characteristics of the tokens – of which the importance of speculation is discussed below (Section 4.5).

The application of questions highlighted in Figure 4.1 in assessing a cryptoasset is mostly self-evident. However, some of the questions raised require further elucidation. These issues are discussed below.

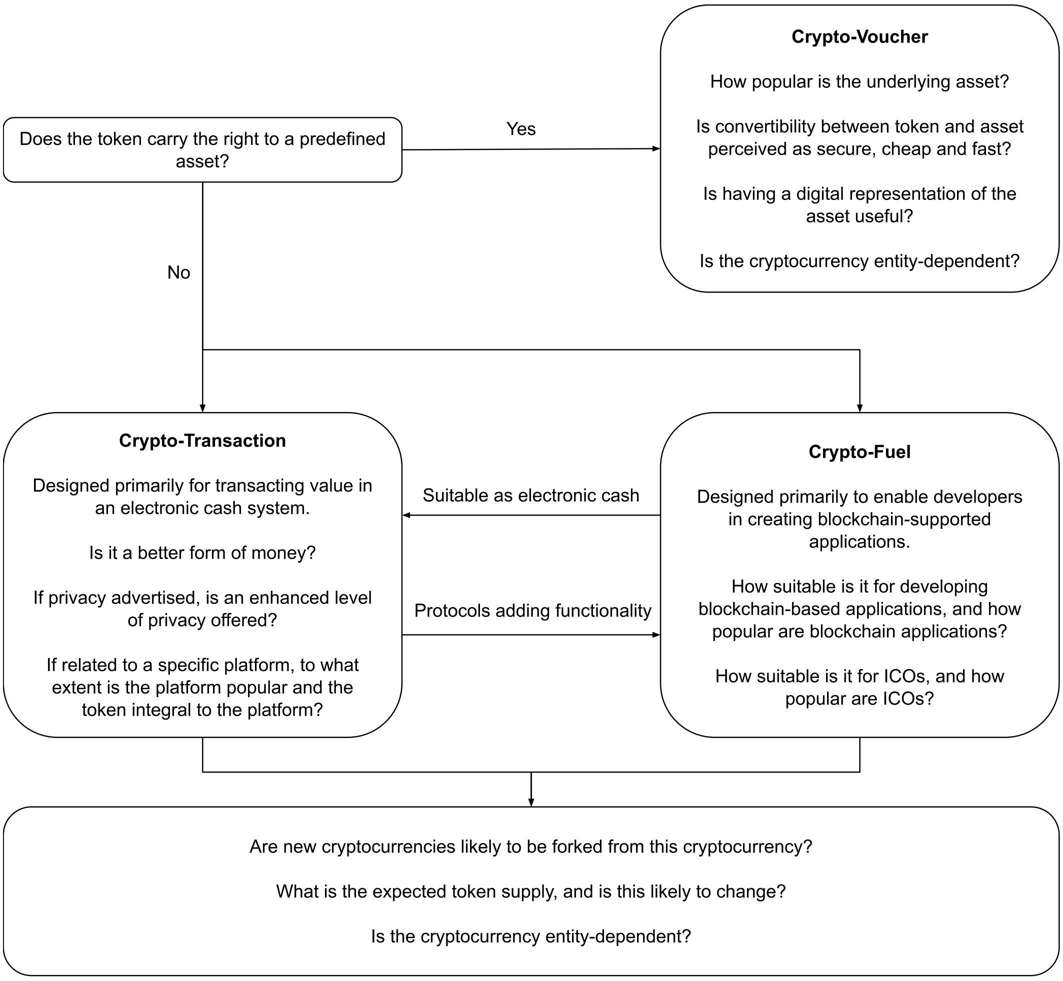

Does the token carry the right to a predefined asset?
Yes
No
Crypto-Voucher
How popular is the underlying asset?
Is convertibility between token and asset perceived as secure, cheap and fast?
Is having a digital representation of the asset useful?
Is the cryptocurrency entity-dependent?
Crypto-Transaction
Designed primarily for transacting value in an electronic cash system.
Is it a better form of money?
If privacy advertised, is an enhanced level of privacy offered?
If related to a specific platform, to what extent is the platform popular and the token integral to the platform?
Suitable as electronic cash
Protocols adding functionality
Crypto-Fuel
Designed primarily to enable developers in creating blockchain-supported applications.
How suitable is it for developing blockchain-based applications, and how popular are blockchain applications?
How suitable is it for ICOs, and how popular are ICOs?
Are new cryptocurrencies likely to be forked from this cryptocurrency?
What is the expected token supply, and is this likely to change?
Is the cryptocurrency entity-dependent?

Figure 4.1: A fundamental analysis assessment framework for cryptoassets.

4.4.1 Determining a 'better' form of money

Two-thirds of cryptoasset payment companies' transactions were found to be between national currency and cryptoasset [138], underlining the importance of national currencies as a competing form of money. Hence, following Hileman and Bank of England Governor Mark Carney [52, 137], cryptoasset and national currencies are compared regarding each of the economic functions of money, to determine whether a given cryptoasset has the potential to truly represent 'The Best Money in the World' [53].

1. As a long-term store of value:

    - The paper notes underpinning the value of bank accounts deteriorate and must be replaced. The Federal Reserve spent USD 726.6 billion on new paper notes in 2017 [216], about 85% of which replaced deteriorated paper notes [215] which typically last about 6-7 years [217]. Cryptoassets' digital form does not deteriorate over time.
    - Investors cannot be sure to recover the value invested in highly volatile cryptoassets; the continuous creation of new systems of cryptoasset means there is a risk of previous systems becoming obsolete and so losing value; flaws in the underlying code may suddenly render the cryptoasset valueless. For some cryptoasset systems, such as Bitcoin, the process for verifying and recording transactions (*mining*) could in theory become dominated by a single entity, who could then spend the same token many times and/or block all transaction validation (though the risk of such a so-called '51% Attack' is demonstrably smaller in established proof-of-work-based cryptoassets due to their increased size [292]). Hence, crypto-transaction systems often seek to prevent miner centralisation.

2. As a unit of account:

    - A paper currency cannot measure value in fractions of a coin, whereas a digital currency is infinitely divisible, suggesting cryptoassets could be particularly valuable for micropayments.
    - The high volatility of cryptoassets undermines its use in the consistent measurement of the value of goods or services [137].

3. As a medium of exchange:

    - Cryptoassets facilitate global transactions without an intermediary, potentially offering faster and more private transactions.
    - Paper currencies are more valuable as a medium of exchange because they have a much larger userbase than cryptoassets. This could explain why scalability is an issue for some crypto-transaction systems: lack of scalability constrains the potential userbase. This also suggests the importance of liquidity: the easier it is to enter and exit a cryptoasset, the more useful it is as a medium of exchange [296].

4.4.2 Forks

When the codebase of a cryptoasset forks, it effectively splits into two versions: the original and a new version that implements perceived improvements. Unless all users and miners then switch to one version, the result is two distinct cryptoassets [54]. If the original transaction data is copied across, the owners of the original cryptoasset may receive free tokens of the new cryptoasset. This occurred when Ethereum Classic forked from Ethereum, and when Bitcoin Cash forked from Bitcoin. A tendency for investors to purchase cryptoassets intending to benefit from such events has been observed [246].

4.4.3 Token Supply

There is likely to be an inverse relationship between the price and the expected supply of tokens in circulation. Potential participants should therefore consider how new tokens will be created over time and their distribution mechanism. For many cryptoassets [62, 65, 91, 97, 207, 211] the supply over time is determined formulaically by the codebase.

4.4.4 Entity Dependence

Entity-dependent cryptoassets are characterised as when the system becomes dependent on a small number of operators (see Section 2.2.2 and Burnie et al [41]). This can be by design, such as with Tether controlling the creation and destruction of tokens [187], or by evolution. For example, a few market participants could potentially hoard a significant proportion of a cryptoasset in circulation, giving them power over its price. The importance of who controls the verification and recording of transactions has been particularly emphasised [41]. A widely held concern with Bitcoin is whether a miner could have sufficient computing power to instigate a 51% attack [292], enabling them to block all transactions and to spend the same tokens repeatedly [34, 65, 207]. Participants should thus consider the implications of entity-dependence.

4.5 Limitations to the Analysis of Fundamentals

The problem with using the identified fundamentals (Section 4.4) to price cryptoassets is that speculation can act as an important factor that obscures the effect of these fundamentals [31, 173, 260]. Speculation can mean that fluctuations in price may occur even if the fundamentals suggested by this article remain unchanged.

A similar issue was observed with Internet companies during the ‘Dotcom Bubble’ where valuations were often based on speculation rather than profitability. However, Demers and Lev found that when these valuations fell during the ‘Dotcom Bubble’, those Internet companies with the strongest fundamentals were the most resilient [81]. This suggests that the fundamentals highlighted by this article may be particularly important for investors in identifying cryptoassets with midto long-term value.

Another issue is the nascent nature of the trading infrastructure, with exchanges facing difficulties in handling surges in demand, denial-of-service attacks, and theft [75]. The threat of losing access to cryptoasset holdings may trigger investors to sell even if the cryptoasset's fundamentals are strong, contributing to the high price variation. Infrastructure difficulties may explain why prices can differ across exchanges, for example, with Bitcoin prices varying by USD 4000 between different exchanges on 8 December 2017 [75].

4.6 Discussion

4.6.1 Classification

A classification is established based on the intended functionality of the different tokens of the most financially significant cryptoassets. This basis distinguishes the classification from previous systems that considered a regulatory [73, 92, 102, 107, 136, 257], technology [275] and ontological [133] perspective.

This supports there being substantial, qualitative variation in the characteristics of the tokens being bought and sold across different cryptoasset systems. Hence, cryptoassets are to be analysed individually rather than treating all cryptoassets, regardless of token type, as a single entity. Specifically, Bitcoin is found to be of a distinct type (crypto-transaction) compared with Ethereum (crypto-fuel), which supports analysing Bitcoin and Ethereum separately and comparing the results.

Burniske and Tatar also created a tripartite classification, but Burniske and Tatar examined the digital resource provisioned [48] not the token functionality. Burniske and Tatar restricted the term 'cryptocurrencies' exclusively to crypto-transaction systems as only these systems intend to provide a new form of currency. Their 'crypto-commodity' category was similar to this chapter's 'crypto-fuel' type in that both refer to systems where the intention is to support other applications. Burniske and Tatar also corroborate in distinguishing Bitcoin (a crypto-transaction or true 'cryptocurrency') from Ethereum (a crypto-fuel or 'cryptocommodity') [48].

The Burniske and Tatar system differs from this chapter's classification in how the third group is specified. Burniske and Tatar referred to 'cryptotokens' where finished goods and services are provided [48], whilst here the third group is centred on 'crypto-voucher' systems where the token carries a right to a predefined asset.

A cryptoasset being of the 'crypto-voucher' type is relevant to the price dynamics of the cryptoasset, and so, by excluding this category, the Burniske and Tatar system is less suitable in understanding the heterogeneity across cryptoassets. This can be seen by analysing the example of USD Tether. Under Burniske and Tatar, USD Tether is a true cryptocurrency [48] because it is intended to provide a form of currency [187]. However, holders of USD Tether also have a right to exchange one-to-one with US Dollars and so, under this chapter's classification, USD Tether is primarily a crypto-voucher system (Figure 4.1). In terms of understanding future price dynamics, knowing that USD Tether is a crypto-voucher is more informative than knowing it is being used as a currency. This is because USD Tether's status as a crypto-voucher system means that the price of USD Tether is unlikely to move far from one US Dollar, assuming participants believe in the exchangeability between USD Tether and US Dollars. This is consistent with the finding, in the previous Chapter 3, that USD Tether had a negative (albeit weak [255]) correlation with the other cryptoasset prices.

How to divide between different cryptoasset types is also more ambiguous in the Burniske and Tatar taxonomy. For instance, the token Kin could be seen as a true cryptocurrency as it is a currency intended for 'payments' [144], but Kin also supports Kik, a messaging platform, and so it supports provision of a finished product, making it a cryptotoken. By comparison, the classification in this chapter sees Kin as a crypto-transaction token because of its primary intended use in payments.

CryptoCompare created an alternative tripartite classification [196] that considered 'natural grouping[s]', combining perspectives from: regulation, industry classification, rationales for holding tokens and the 'economic value drivers.' 'Economic value drivers' related to whether price was driven mainly by changes to demand or supply from a network or by changes in the value of an underlying asset.

CryptoCompare identified that most cryptoassets were fungible, that is to say each token is interchangeable with any other token. Fungible tokens were then classified, with terminology adopted from terms used by regulators (see Section 2.1.2), into: 'payment', 'utility' and 'asset-security' tokens. The CryptoCompare classification corroborates with Burniske and Tatar [48] and this chapter in dividing Bitcoin from Ethereum: bitcoin is a 'payment' token and ether is a 'utility' token [196].

'Payment' tokens are similar to crypto-transaction tokens in that both are intended for transacting value, but payment tokens must also be used across 'all networks' [196]. In this chapter the need for use across 'all networks' was removed. Regardless of whether a crypto-transaction token is intended for global use or to support a specific platform, the token still provides a form of money and so the functions of money are still fundamentally applicable in assessing its value. The term 'all networks' is also ambiguous as, technically, bitcoin tokens (specified as 'payment' tokens) can only be used on the one network – Bitcoin.

The classification in this chapter does not contain 'utility' tokens that 'offer digital access to an application or to some service' [196] as this could capture a variety of different types of access and a plurality of services. Kin provides access to the Kik messaging platform (rather than being used across 'all networks' [196]) and so could be seen as a utility rather than payment token. However, it is intended as a form of money [144] and so, in its fundamentals, it is more similar to other crypto-transaction tokens such as bitcoin than other 'utility' tokens such as ether. Ether, unlike Kin, is intended to support application development.

The classification in this chapter advances on comparable alternatives because it analyses the characteristics of the token being bought or sold. This means that the issues considered are more pertinent to cryptoasset buyers and sellers and so are likely to be more relevant to the price dynamics.

Following this classification, analysis will examine crypto-transaction and crypto-fuel systems, and not crypto-voucher systems. In crypto-voucher systems, the value of the token is likely to be dominated by changes in the value of the underlying asset, and so modelling the token price may best be met through modelling the underlying asset. This means that insights on what drives the valuation of a specific crypto-voucher token are least likely to be generalisable to cryptoassets in general. The focus will be on Bitcoin and Ethereum, as the largest crypto-transaction and crypto-fuel systems respectively (see Table 2.2).

4.6.2 Analysis of the Fundamentals for Bitcoin and Ether

Bitcoin and Ethereum are selected for analysis because of their distinct functionality and for the reasons detailed in Section 2.2. The fundamentals for Bitcoin and Ethereum are next determined by applying the questions detailed in Figure 4.1.

Bitcoin is a crypto-transaction system. This suggests that changes in the actual and expected benefit Bitcoin provides as a form of money over time may influence the bitcoin price. This utility can be evaluated by considering bitcoin as a long-term store of value, unit of account and medium of exchange (from Section 4.4.1).

The other considerations specific to crypto-transaction systems are less relevant. Bitcoin is not related to a specific platform and buyers interested in privacy are likely to be drawn more to cryptoassets that advertise this as a major feature (such as Dash [91], Monero [207] and Zcash [19, 34]).

Ethereum is a crypto-fuel system. This suggests that the suitability of Ethereum in developing blockchain-based applications and in launching ICOs may affect the ether price, along with the overall popularity of developing blockchain-based applications and launching ICOs.

There may be cross-over in the fundamentals affecting the bitcoin and ether price. Forks could occur in either cryptoassets' code-base, which may increase the price (see Section 4.4.2). Following from Section 4.3.2.4, the fundamentals described for bitcoin may have some relevance to the ether price, as ether can be used as a form of money; whilst the fundamentals described for Ethereum may also be relevant to bitcoin, as the Omni Layer provides Bitcoin with limited crypto-fuel functionality [278]. The extent such hybridisation is supported by the data as having an influence on price dynamics is revisited in the comparison of fundamentals with quantitative analysis results in Chapter 8.

The other shared fundamentals identified in the assessment framework are of less relevance to the price. For both Bitcoin and Ethereum, the supply of new tokens follows a fixed, predictable schedule across time, reducing the potential influence on price from unexpected shocks to tokens supply [97,211]. The size of Bitcoin and Ethereum (see Table 2.2) also reduces the ability of a few operators to seize control of these cryptoassets, rendering them entity-dependent.

For both bitcoin and ether, the price may vary because of a change in the current fundamentals for that cryptoasset or because of a change in the anticipated future fundamentals for that cryptoasset. For instance, demand for bitcoin may rise because a major retailer decides to accept bitcoin for goods and services, improving the current value of bitcoin as a medium of exchange. Alternatively, the expected future userbase for Bitcoin may expand because of current changes in the Bitcoin codebase that improve scalability and so increase the potential number of users that Bitcoin could handle at a future date. This improvement in the potential of Bitcoin may induce speculators to buy bitcoin now in anticipation of a larger future demand for bitcoin. Hence, changes in actual, present fundamentals and expected future fundamentals should both be considered in understanding why the price of a cryptoasset has changed.

Chapter 5

Words Associated with Bitcoin Price Phases

5.1 Introduction

Figure 5.1 illustrates how the bitcoin price followed three distinct phases of movement across 2017-18:

- **Stage 1 (from 1 January to before 16 December 2017)**: Prices rose to 1954.30% of the initial value, from 997.73 to a peak, all time high price of 19498.68 US Dollars.

- **Stage 2 (from 16 December 2017 to before 29 June 2018)**: Prices fell overall, in a cyclical pattern, to 5908.70 US Dollars (30.30% of the December peak).

- **Stage 3 (from 29 June 2018 to before 15 November 2018)**: Prices traded within a band of 30.30% - 42.32% of the highest value in the series (19498.68 US Dollars). The median price, across stage 3, was 6499.06 US Dollars (9.99% above 29 June 2018). Throughout prices remained above the 29 June 2018 value, and so the prices did not fall overall.

After 15 November 2018, the price fell below the 29 June 2018 value and, by the end of the dataset, the price was 3967.52 US Dollars.

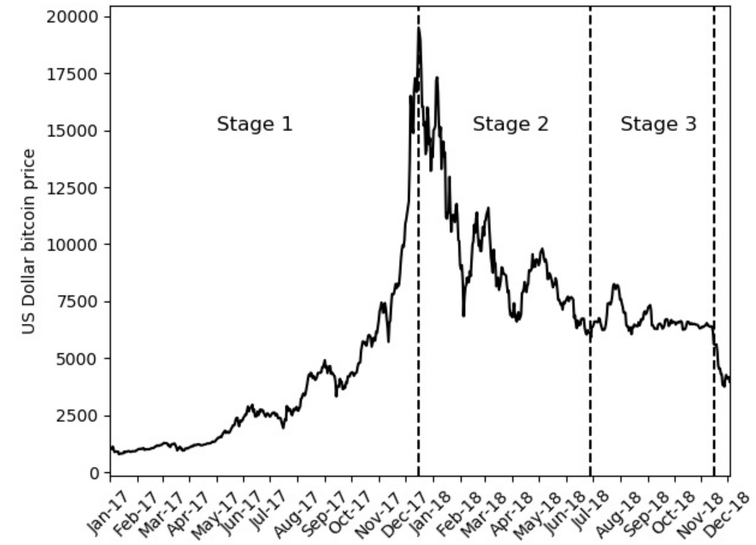


Figure 5.1: The daily bitcoin price in US Dollars from 1 January 2017 to 3 December 2018. The horizontal axis is formatted such that each tick corresponds to the first day of the labelled month. Data sourced from the Charts API of Blockchain Luxembourg S.A. [26].

This thesis examines how word use on Reddit varied across these three different bitcoin price phases, comparing the middle phase of falling prices with the rising phase before and a relatively stable phase after. A similar, three-phase pattern was not evident in the ether price series where the rising prices across 2017 was interrupted by a five-month period of volatility where prices only rose 3% overall (see Figure 1.1).

As published in Royal Society Open Science [45], word use is analysed from two perspectives. Firstly, the most frequent words are compared across phases. Secondly, the three-phase pattern is exploited, with words associated with the stage 2 identified through comparison of word frequencies with the stages before (stage 1) and after (stage 3). This involves developing a new Data-Driven Phasic Word Identification (DDPWI) approach that identifies which words have daily frequencies that are statistically significantly higher or lower in stage 2 compared with both the time period before and after. The resulting 'price dynamic words' are interpreted using approaches developed to elucidate the context in which these words are used across the different phases.

5.2 Data Preparation

5.2.1 Data Sources

The dataset extended from 1 January 2017 to 3 December 2018. US Dollar Bitcoin Price was sourced from Blockchain Luxembourg S.A. through their 'Charts API' [26], and the text for each submission to the 'Bitcoin' subreddit was extracted using the Pushshift API [15]. Submissions data were selected over comments because the latter were prone to deviate onto arguments on bitcoin-irrelevant topics, such as religion, non-specific insults and different date formats (`https://www.reddit.com/r/Bitcoin/comments/9svjcp/10_years_ago_today_2008_oct_31/`). The Reddit submissions were processed as follows.

5.2.2 Engineering Word Frequency Data from Reddit Submissions Text

The submissions were filtered and the text processed and tokenised to produce word lists.

Submission Filtering

The following submissions were filtered out: those authored by ‘rBitcoinMod’, as these consisted primarily of automated text stating forum guidelines for the ‘Daily Discussion’ and ‘Mentor Monday’; those authored by ‘crypto_bot’, as these consisted mainly of automated, daily data updates on the bitcoin network; submissions with identical text to another submission; blank submissions; and submissions that had been entirely removed, thus whose text consisted of only ‘[deleted]’ or ‘[removed]’ [300].

Text Pre-Processing

1. All text was put into the lower case.

2. The accepted currency codes [172] ‘btc’ and ‘xbt’ were converted into the synonymous ‘bitcoin’.

3. The following were removed respectively: strings of 50 or more consecutive word characters (as this is too long to represent a word); URLs; HTML tags (e.g. ‘&amp’); the new line character (‘\n’); Twitter (e.g. ‘@john’) and Reddit handles (e.g. ‘/u/john’ and ‘/r/john’); references to deleted text (‘[removed]’ and ‘[deleted]’); and non-ASCII text (e.g. Cyrillic alphabet or emoticons).

4. The US Dollar was referred to in 11.30% submissions as: ‘$’, ‘usd’, ‘dollar(s)’ and ‘us dollar(s)’. These were treated as synonymous and were all replaced by ‘dollar_marker_symbol’.

5. Punctuation and apostrophes were removed unless these were inside words to indicate abbreviations (e.g. ‘o’clock’).

6. ‘tx’ was used to abbreviate the word transaction [301] and thus was replaced by the word ‘transaction’. Both ‘ln’ and ‘lightning network’ were replaced with ‘ln’. The terms ‘telephone number’ and ‘phone number’ were replaced with ‘phone_number’.

Daily Word Frequencies

Text was converted into word lists using Python package NLTK version 3.3 and its associated download ‘punkt’. NLTK removed ‘stopwords’ which were high frequency words unrelated to a particular topic (such as ‘me’, ‘we’, ‘a’ or ‘the’). The term “n’t” was included as an abbreviation for the stopword ‘not’.

Words with the same meaning but different grammatical case were combined. Each word was lemmatised using NLTK’s ‘WordNetLemmatizer’. The context of a word was determined by looking it up in a dictionary and mapping different cases of the same word to a base form. This failed for unusual words (e.g. ‘bitcoins’ and ‘bitcoin’, and ‘ICO’ and ‘ICOs’) that were not in the dictionary. Hence, stemming was subsequently applied using ‘SnowballStemmer’. This applied to all words a set of rules that ignore the context of the word, and so was extendable to rare words. The ‘snowball’ stemmer was chosen as it is the least likely to treat words of the same concept differently or words of a different concept the same [154]. For cryptoasset mining, two abbreviations ‘miner’ and ‘mine’ were merged into ‘mine’.

To prevent skewing by a few longer submissions, each word was counted once if present in a given submission. Words in 100 or less submissions were removed. There were 326,945 submissions with 131,656 words of which 3,900 were found in more than 100 submissions. A ‘day’ was specified to be from 00:00 on a given day to before 00:00 on the next date (GMT).

Daily counts of the 3,900 words were normalised by dividing the count by the daily total number of submissions to ensure that word frequency measured the proportion of submissions containing a term. The number of submissions per day were the number that remained after text processing.

5.3 Methodology

5.3.1 Identifying Words by Absolute Frequency

The words that were in at least 5% submissions in any one of the identified stages of the bitcoin price series were identified. This was to determine the extent certain words dominated discussions across time.

5.3.2 Identifying Words by Relative Frequency

Comparing Word Frequencies Across Stages

A methodology was required to statistically evaluate for which words the daily frequencies were typically higher or lower in one stage of the price series compared with the previous stage.

Extreme outliers were present, even for a popular word such as 'bitcoin', which never fell below 35% submissions on a given day. Across three days, the popularity of 'bitcoin' fell from 46.75% submissions (19 July 2017) to 38.58% (20 July 2017) to recover to 48.71% the next day (21 July 2017). This precluded using the t-test in comparing daily word frequencies across price phases, as this is sensitive to extreme outliers [37, 293].

Instead, the non-parametric equivalent, the two-sided Wilcoxon Rank-Sum Test, was used to delineate which words had daily frequencies that had changed significantly across different phases in the price series. An additional Bonferroni correction, such that the p-value cut-off (1%) was divided by the number of tests (3,900), ensured that the identification of significant words was robust [199].

Applying DDPWI to Identify Price Dynamic Words

The DDPWI approach identifies those words where the change in frequency from phase 1 to 2 (rising prices shifting to falling) and from phase 2 to 3 (falling prices ceasing to fall further) are opposite and both statistically significant. The words that changed statistically significantly were identified using the two-sided Wilcoxon Rank-Sum Test and restricted to those with above 1% frequency in phase 2. We define the words resulting from applying DDPWI as the 'price dynamic' words.

5.3.3 Context of Price Dynamic Words

Identifying the Context

An iterative procedure for generating the theme of a typical sentence that contained one of the price dynamic words was developed:

1. Let W represent a chain of words. Initially, $W = [w_1]$, where w_1 was the specific word of interest.
2. Extract only submissions that contain all words in W.
3. Find the most frequent word in these submissions and append to W.
4. Repeat (ii) - (iii) until W is of length 5, excluding the word of interest, or there exists at least two words of the same highest frequency in step (ii).
5. The result was a chain of related words, $W = [w_1, w_2, ...]$.

Every iteration reduced the number of submissions considered. Generic words (e.g. 'bitcoin' and 'would') that provided little thematic content and synonyms were censored.

Sentiment of the Context

Sentiment was measured, using the VADER [142] algorithm, for submissions that contained the price dynamic words, using 'bitcoin' as a control. VADER was designed for social media text and so is able to handle both emoticons and slang [142,165]. Text processing was thus minimised to converting 'tx' (a bitcoin-specific abbreviation [301]) into 'transaction' and removing tokens that should not have a sentiment (e.g. URLs, HTML tags and '[deleted]'). Individual submissions with a compound sentiment score of less than -0.2 were labelled as 'negative' and those with a score of at least 0.2 as 'positive' [165]. The number of positive sentiment submissions was divided by the number of positive and negative sentiment submissions to derive the positive sentiment metric. The negative sentiment metric was similarly normalised. The sentiment metrics were calculated for submissions across the past 90 days to prevent noise in the metric from obscuring the identification of underlying trends in the sentiment over time.

5.4 Results

5.4.1 Reddit Submissions Descriptive Statistics

Table 5.1 presents descriptive statistics on Reddit submissions, showing a decline in Reddit activity as prices stabilised. On average, over 500 submissions were posted per day when prices were most volatile in stages 1 and 2; this fell 46% with stage 3.

Table 5.1: Descriptive statistics for Reddit Submissions (1 January 2017 to 3 December 2018).

Stage	Days	Submissions	Submissions per Day
All Data	702	326945	465.73
1	349	181327	519.56
2	195	101110	518.51
3	139	38706	278.46

5.4.2 Most Frequent Words by Absolute Frequency

Figure 5.2 lists those words in at least 5% (one in twenty) submissions in all stages. The term ‘bitcoin’ was the commonest, in about half of submissions. The other terms conveyed the persistent popularity of discussion around the bitcoin price (‘dollar_marker_symbol’ and ‘price’), acquiring bitcoin (‘get’, ‘buy’, ‘make’), opinions (‘like’), innovation (‘new’) and exchanges (‘exchang’). All these terms had a statistically significant fall from stages 1 to 2 except ‘exchang’ (p-value 1.95e-01). The ‘dollar_marker_symbol’ term rose significantly in popularity from stages 2 to 3 (p-value 4.94e-12).

Figure 5.3 lists those words that were in at least 5% submissions in an incomplete number of stages. Twenty fell significantly from stage 1 to 2 with 14 falling to below the 5% threshold. ‘Blockchain’ became popular in stage 3 (46.12% rise on stage 2, p-value of 3.39e-10) and so did ‘market’ (48.14% rise on stage 2, p-value of 1.72e-17). Cryptocurrency discussions more than doubled in frequency from phase 1 to 2, an upward trend that continued to phase 3. The term ‘coinbas[e]’ (frequency of 5.83% in stage 1) referred to the cryptoasset exchange Coinbase (`https://www.coinbase.com/`).

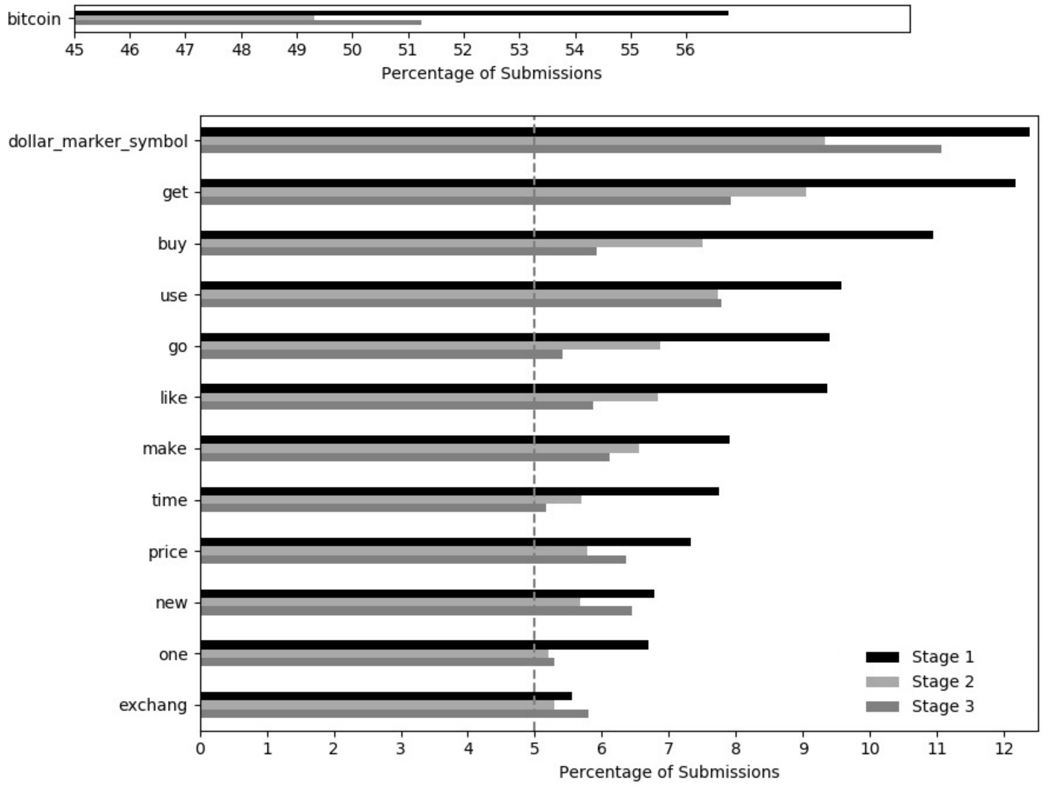


Figure 5.2: Words in at least 5% submissions in all stages, and the percentage of submissions they were in for each stage. Bitcoin is graphed separately because it was more than twice as frequent as the next word. The dashed, vertical grey line represents the 5% cut-off. The top bar represents the percentage of submissions containing the term in stage 1; the middle bar is the percentage in stage 2; and the bottom bar is the percentage in stage 3. Each 'word' is a lemmatised and then stemmed version of the original word. For example, 'exchang' represents exchange, exchanges, exchanged and exchanging, and 'use' represents 'use', 'uses', 'used' and 'using'. The term 'dollar_marker_symbol' represents different synonyms for the US Dollar (see Section 5.2.2).

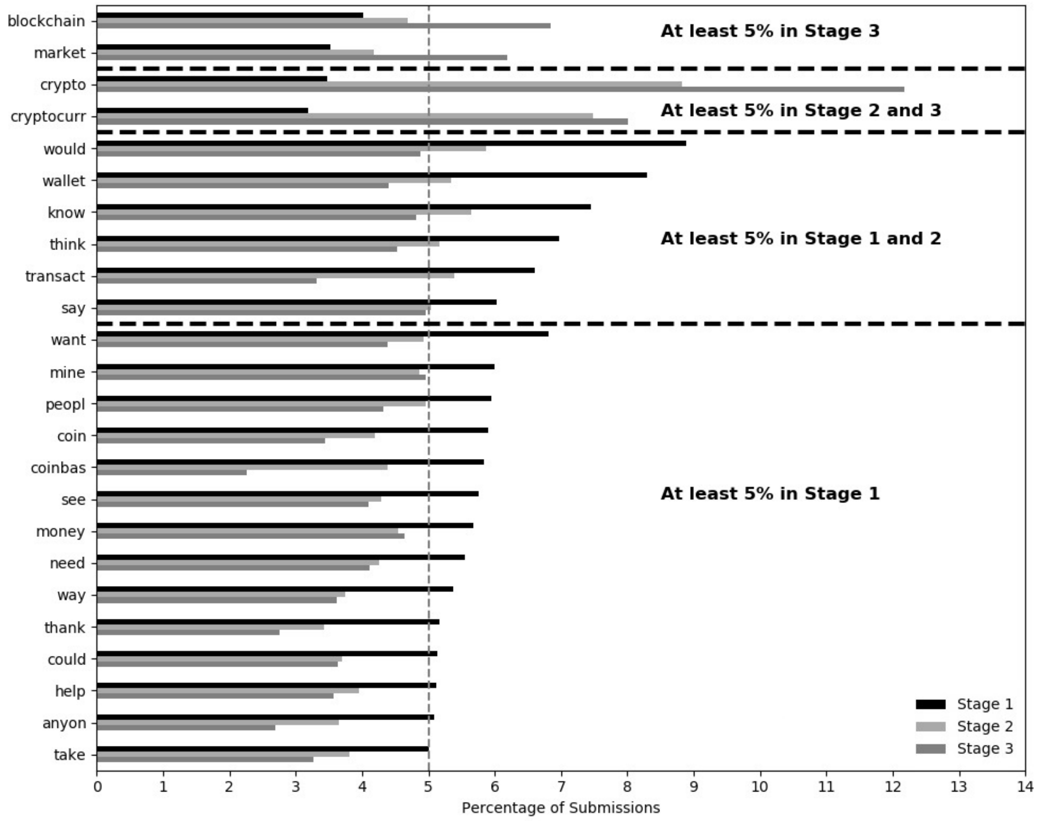


Figure 5.3: Words in at least 5% submissions in at least one stage but not all, and the percentage of submissions they were in for each stage. These words consisted of four groups demarcated by the black, horizontal lines: those words in at least 5% submissions in stage 1 alone (bottom words); in stages 1 and 2 (penultimate from bottom); in stages 2 and 3 (penultimate from top); and in stage 3 alone (top). The dashed, vertical grey line represents the 5% cut-off. The top bar represents the percentage of submissions containing the term in stage 1; the middle bar is the percentage in stage 2; and the bottom bar is the percentage in stage 3. Each 'word' is a lemmatised and then stemmed version of the original word. For example, 'exchang' represents exchange, exchanges, exchanged and exchanging, and 'use' represents 'use', 'uses', 'used' and 'using'.

5.4.3 Comparing Word Frequencies Across Stages and Identifying Price Dynamic Words

Eleven words demonstrated a statistically significant change in frequency when moving from both phase 1 to 2 and phase 2 to 3.

Six words rose across both phasic shifts: ‘investor’, ‘market’, ‘million’, ‘crypto’, ‘launch’ and ‘platform’. Two words fell across both phasic shifts: ‘segwit’ and ‘fee’.

Three words fulfilled the definition of a price dynamic word (Section 5.3.2) in that the change in frequency was opposite and statistically significant from phase 1 to 2 and from phase 2 to 3: ‘tax’ and ‘ban’ rose from stage 1 to 2 and fell from stage 2 to 3; whilst ‘dollar_marker_symbol’ fell from stage 1 to 2 and rose from stage 2 to 3.

5.4.4 Context of Price Dynamic Words

Identifying the Context

The word ‘ban’ occurred most with ‘china’ and ‘exchang[es]’ in stage 1 (see Table 5.2) but these associated words did not continue into stages 2 or 3. In stage 2, bans were mentioned in the context of ‘central’ ‘bank’ ‘cryptocurr[ency]’ regulation. A subanalysis of the ten most frequent words associated with ‘ban’ demonstrated ‘trade’ (11.98%) and ‘ad’ (11.20%) were specific to stage 2, and ‘googl[e]’ (12.75%) was specific to stage 3. When ‘ban’ and ‘trade’ were run together (stage 2), ‘korea’ was the most frequent word (42.21%). When ‘ban’ and ‘ad’ were run together, the chain of associations were: ‘facebook’ (42.36%), ‘googl[e]’ (22.95%) and then ‘twitter’ (78.57%). In stage 3, ‘ban[s]’ by the ‘india[n]’ ‘reserv[e]’ ‘bank’ became a topic. When ‘ban’ was paired with ‘googl’, ‘ad’ had the highest frequency (46.15%). In submissions with these three words, ‘end’ (50.00%) was the most frequent.

US Dollars were discussed the most with the word ‘buy’ in stages 1 and 2. This pair was mentioned more with ‘price’ in stage 1 and ‘sell’ in phase 2. Phase 3 was distinct - ‘price’ (20.63%) was mentioned more than ‘buy’ (13.70%) and dollars and price were mentioned most frequently with the word ‘market’ (26.02%).

The word ‘tax’ occurred most frequently in association with the word ‘pay’ throughout all three stages. Associated with ‘pay[ing]’ ‘tax’, was ‘capit[al]’ ‘gain[s]’ (stage 1 and 2) and ‘buy[ing]’ (stage 1) and/or ‘sell[ing]’ (stage 1 and 2) bitcoin.

Table 5.2: Chain of most frequent words associated with price dynamic words: ‘tax’, ‘ban’ and ‘dollar_marker_symbol’. At each step, submissions were reduced to those containing all previous words in the chain and then the most frequent word in these submissions was found and expressed as a percentage of submissions. For example, starting with stage 1 submissions containing the word ‘tax’, the most frequent word was ‘pay’ (34.17% of those submissions). The word that was most frequent in submissions with the words ‘tax’ and ‘pay’ was ‘buy’, in 35.12% of these submissions. In submissions that contained ‘tax’, ‘pay’ and ‘buy’, 48.73% contained the word ‘sell’.

Stage	Chain
	‘ban’
1	‘china’ (38.50%) - ‘exchang[e]’ (30.20%) - ‘price’ (31.69%) - ‘peopl[e]’/‘would’/‘time’/‘trade’ (55.17%)
2	‘cryptocurr[ency]’ (29.63%) - ‘bank’ (23.10%) - ‘central’ (37.50%) - ‘govern’ (36.36%) - ‘time’/‘technolog[y]’ (75.00%)
	Starting with ‘ban’ and ‘trade’, censoring ‘cryptocurr[ency]’ and ‘crypto’
	‘korea’ (42.21%) - ‘south’ (83.08%) - ‘say’/‘plan’ (20.37%)
	Starting with ‘ban’ and ‘ad’, censoring ‘cryptocurr[ency]’ and ‘crypto’
	‘facebook’ (42.36%) - ‘googl[e]’ (22.95%) - ‘twitter’ (78.57%) - ‘plan’ (36.36%)
3	‘cryptocurr[ency]’ (27.12%) - ‘india’ (31.33%) - ‘bank’ (34.62%) - ‘reserv[e]’/‘court’ (44.44%)
	Starting with ‘ban’ and ‘googl[e]’, censoring ‘cryptocurr[ency]’ and ‘crypto’
	‘ad’ (46.15%) - ‘end’ (50.00%) - ‘month’ (44.44%) - ‘next’/‘news’ (75.00%)
	‘dollar_marker_symbol’
1	‘buy’ (24.05%) - ‘price’ (27.75%) - ‘time’ (39.43%) - ‘one’ (47.38%) - ‘peopl[e]’ (63.57%)
2	‘buy’ (18.66%) - ‘sell’ (30.66%) - ‘price’ (47.22%) - ‘peopl[e]’/‘time’ (49.80%)
3	‘price’ (20.63%) - ‘market’ (26.02%) - ‘time’ (50.00%) - ‘exchang[e]’/‘trade’ (54.78%)
	‘tax’
1	‘pay’ (34.17%) - ‘buy’ (35.12%) - ‘sell’ (48.73%) - ‘gain’ (55.21%) - ‘capit[al]’ (75.47%)
2	‘pay’ (28.95%) - ‘gain’ (36.13%) - ‘capit[al]’ (61.29%) - ‘year’/‘sell’ (45.26%)
3	‘pay’ (25.57%) - ‘one’/‘would’ (40.30%)

Sentiment of the Context

Examining bitcoin mentions (see Figure 5.4) demonstrated that positive sentiments were more than twice as frequent than negative across all three stages. Sentiment initially became more negative during the phase of falling prices, but from March 2018 this trend reversed. Overall bitcoin mentions fell during phase 2, from over 58% to below half of submissions, but from April 2018 onwards this reversed.

Figure 5.5 shows that twice as many ‘ban’ submissions were negative than positive, and there was a drift towards more negative sentiment over time. There were periods of particularly high interest where frequency was above 1.6%: the 90 days up to October-November 2017 (phase 1), and in April 2018 and June 2018 (both phase 2).

Similar to bitcoin, the frequency of US Dollar mentions fell at the start of phase 2 with this trend reversing from April 2018 (Figure 5.6). Sentiment was twice as positive than negative across the three phases. Sentiment became more negative during phase 2 and, unlike with bitcoin, this trend reversed only with the shift from phase 2 to 3.

Interest in ‘tax’ (Figure 5.7) began to rise just before phase 2, more than doubling in frequency from less than 0.8% (90 days to November 2017) to fluctuating around 1.6% submissions (March-May 2018, phase 2). Frequency subsequently fell to about 0.6% by August 2018 (phase 3). There were more than 2.5 times as many positive than negative submissions mentioning ‘tax’ across the dataset.

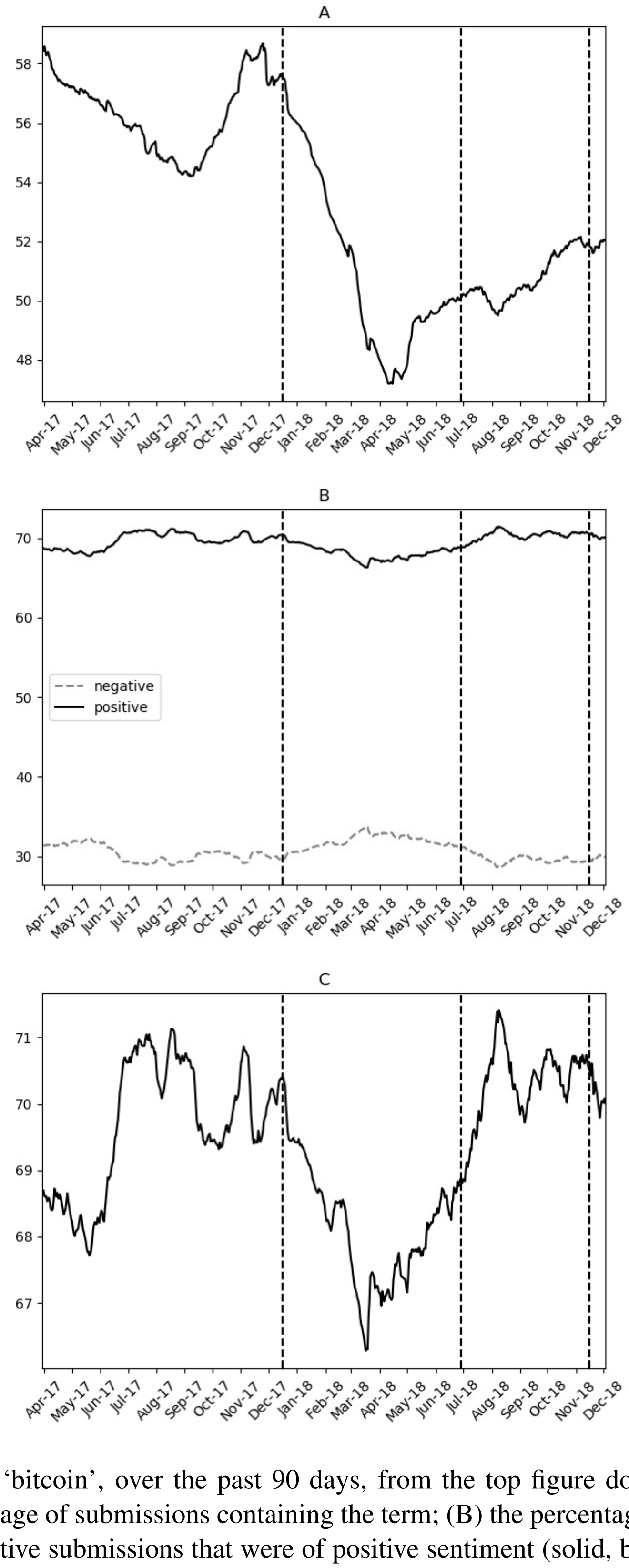


Figure 5.4: For ‘bitcoin’, over the past 90 days, from the top figure down: (A) the percentage of submissions containing the term; (B) the percentage of negative and positive submissions that were of positive sentiment (solid, black line) or negative sentiment (dashed, grey line); (C) the percentage that were of positive sentiment.

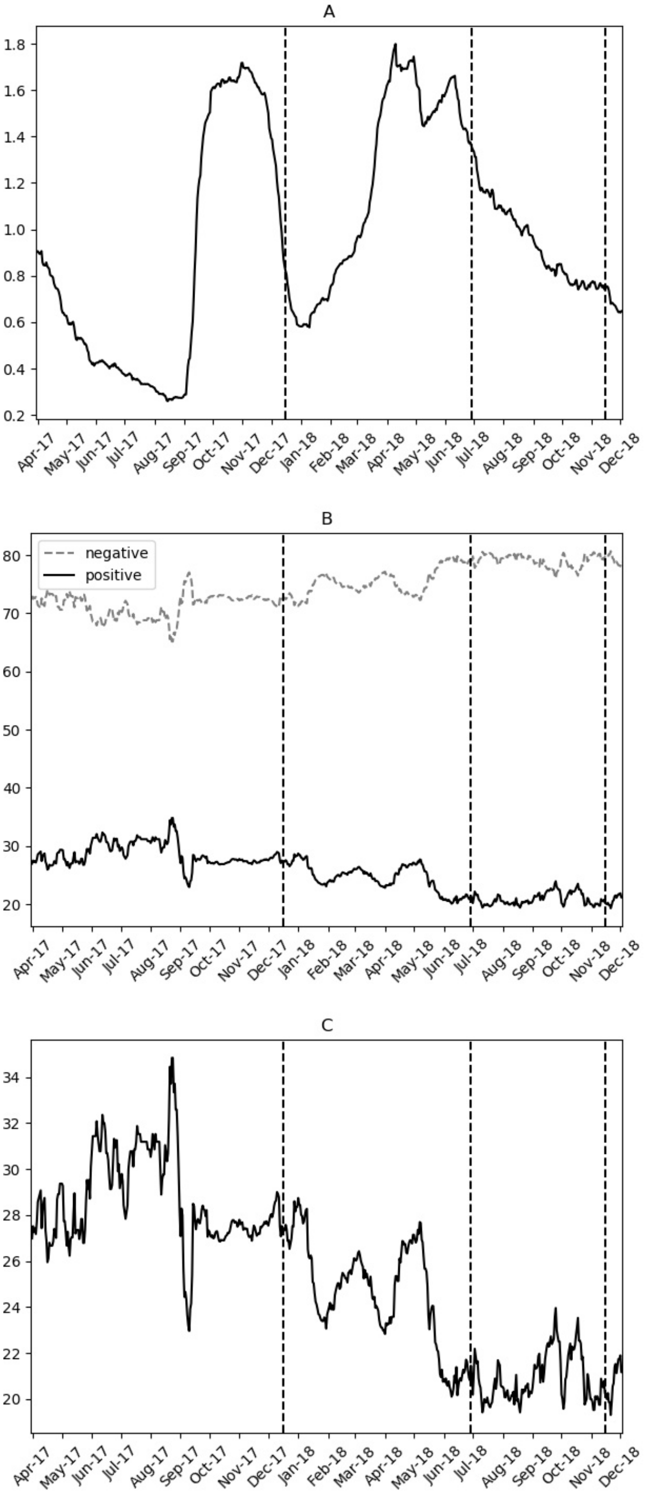


Figure 5.5: For 'ban', over the past 90 days, from the top figure down: (A) the percentage of submissions containing the term; (B) the percentage of negative and positive submissions that were of positive sentiment (solid, black line) or negative sentiment (dashed, grey line); (C) the percentage that were of positive sentiment.

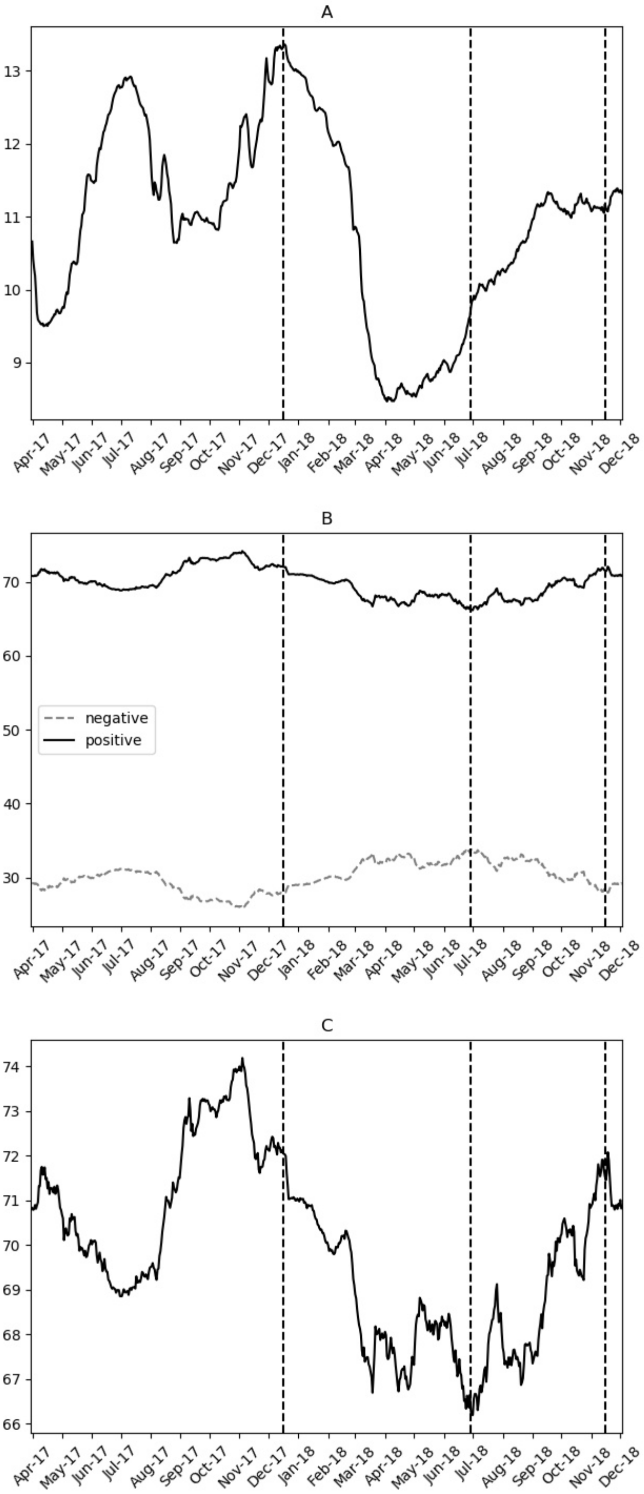


Figure 5.6: For 'dollar_marker_symbol', over the past 90 days, from the top figure down: (A) the percentage of submissions containing the term; (B) the percentage of negative and positive submissions that were of positive sentiment (solid, black line) or negative sentiment (dashed, grey line); (C) the percentage that were of positive sentiment.

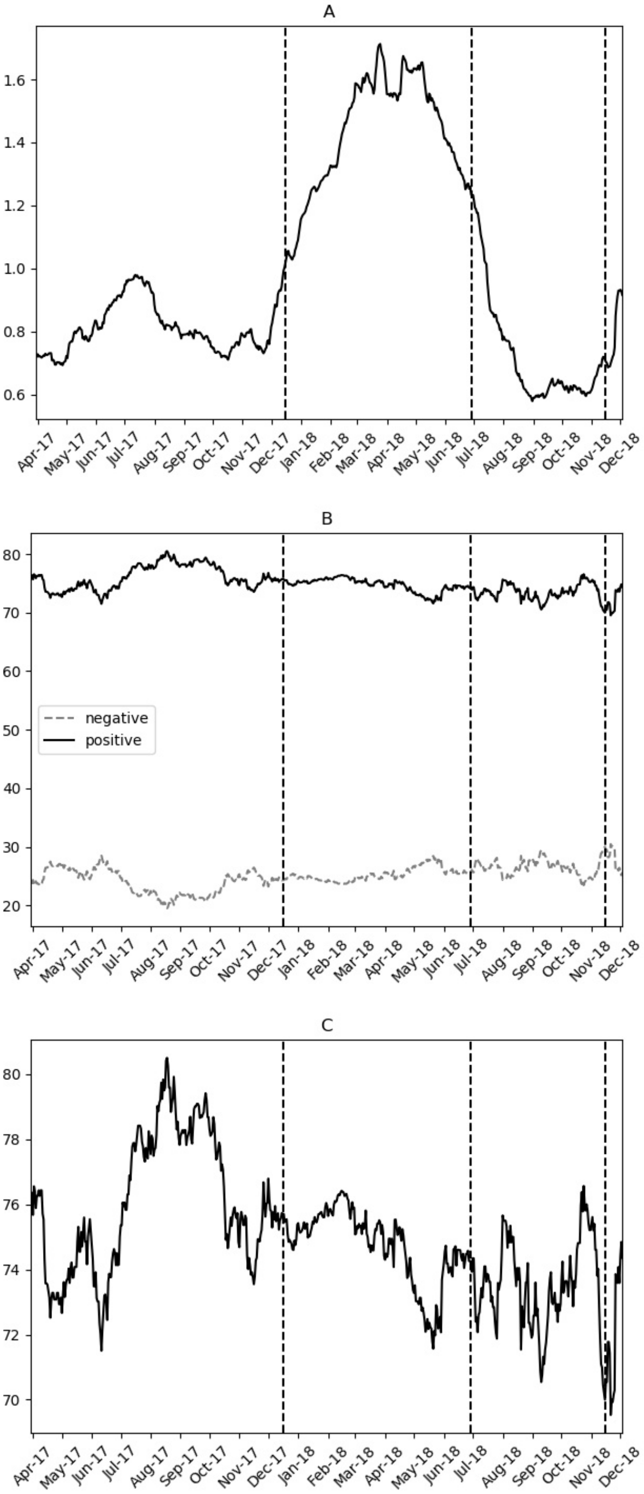


Figure 5.7: For 'tax', over the past 90 days, from the top figure down: (A) the percentage of submissions containing the term; (B) the percentage of negative and positive submissions that were of positive sentiment (solid, black line) or negative sentiment (dashed, grey line); (C) the percentage that were of positive sentiment.

5.5 Discussion

Examining the most frequent words established the evolving nature of Reddit discussions across the three phases (Figures 5.2 and 5.3). During stage 1, discussions were more orientated towards people considering entering the bitcoin network, thus the particularly high popularity of ‘get’, ‘buy’, ‘want’, ‘wallet’ and ‘mine’, and the exchange Coinbase was more frequently considered than in subsequent periods. These words became less popular during stages 2 and 3. During stages 2 and 3, the frequency of submissions discussing crypto and cryptocurrencies more than doubled that in stage 1; likewise, there was an uplift in discussion of blockchain. This would be consistent with interest broadening from bitcoin to other forms of cryptoasset and their associated blockchain technology. Evolving Reddit discussions were further reflected in certain words changing statistically significantly in frequency between one phase and the next (Section 5.4.3). For example, there was a decline in the debate concerning the ‘segwit’ bitcoin fork, whereas there was rising popularity in trading (‘investor’, ‘market’) and cryptoasset innovation (‘crypto’, ‘launch’).

Applying DDPWI identified three ‘price dynamic’ words whose frequencies were associated with the volatile, falling prices of phase 2. The words ban and tax were statistically significantly higher and US Dollars lower during phase 2 compared with both before (phase 1) and after (phase 3).

The word ‘ban’ occurred in a shifting context (Table 5.2) of consistently negative sentiment. This context changed from regulation in China (phase 1) to South Korea (phase 2) to India (phase 3), whilst discussions about internet company bans on adverts became evident only in phases 2 and 3. Discussions of ‘bans’ became particularly frequent from September - November 2017 (just before phase 2) and rose in frequency from January 2018 to a peak in April 2018 (in phase 2) (Figure 5.5). Higher concern over bans coincided with speculation of or actual bans being implemented. For example, the phase 1 activity occurred with China announcing a ban on exchanges in September 2017 with the last exchange closing in November [239]. In previous studies covering earlier time periods, the effect of ‘China’ on the bitcoin price has been suggested [31, 174] and, using topic modelling, the concept ‘China’ was predictive towards the bitcoin price [166].

During phase 2, there was speculation as to the extent to which cryptoasset activities would be banned in South Korea [238]. Facebook banned cryptoasset adverts from January - June 2018 [88], followed by announcements of bans by Twitter [250] and Google [89] in March 2018. The chain of word frequencies in Table 5.2, stage 3, could be explained by a court decision, in India, to uphold the cryptoasset ban, made in July 2018 [285], and Google’s ban on cryptoasset adverts being partially ended in October 2018 [89].

The identified significance of US Dollar discussions was consistent with the importance of speculation in the 2017-18 pricing cycle, an issue raised by the House of Commons Treasury Committee [141]. In stages 1 and 2, US Dollars were most mentioned in the context of buying bitcoin (Table 5.2), with a fall in US Dollar mentions in phase 2 consistent with declining buying enthusiasm (Figure 5.6). In the period of relative price stability (phase 3), ‘buy’ no longer most commonly occurred with US Dollars, nor was it in the chain of popular words (Table 5.2). With more stable prices, there was thus less evidence for speculation.

The price dynamic word ‘tax’ showed a statistically significant increase in frequency from phase 1 to phase 2 and fall from phase 2 to phase 3. ‘Tax’ most frequently occurred with ‘pay’ across all phases. The words ‘capit[al]’ and ‘gains’ were other close associates in stages 1 and 2. Gains on bitcoin trading have been deemed liable to Capital Gains Tax in the US, UK, Japan and Australia [180]. The price gains in phase 1 would have generated a tax liability for traders who sold bitcoin. In order to meet this, they might have sold further bitcoin in stage 2 when tax was due at the end of the financial year, thus driving a downwards trend in prices. The positive sentiment (Figure 5.7) across all stages may reflect that the need to pay tax is associated with making a financial gain.

Chapter 6

Topics associated with Phasic Shifts in Price

6.1 Introduction

This chapter applies machine learning to delineate topics associated with phasic shifts in the price series. Pairs of contiguous phases are compared, identifying words that rose or fell in frequency across phases. This removes the inherent constraint of DDPWI where, after the second phase, the word must revert in frequency. The resulting increase in words extracted is addressed by a neural network methodology (word2vec) which consolidates the resulting words into fewer topics. Grouping similar words together may help to identify events or concerns where discussion is better indicated through the use of a group of words rather than one specific word, such as with market sentiment ('bear', 'bearish' and 'bull').

This chapter aims to optimise the word2vec-based topic modelling approach using Bitcoin data. The optimal model is applied to analyse the shift in bitcoin prices from rising in Stage 1 to falling in Stage 2 (see Section 5.1). The material in this chapter was presented at ACM SIGIR [43]. The methodology used for text preparation and the identification of words that changed statistically significantly in frequency is identical to that described in Chapter 5. The bitcoin price phases are described in Section 5.1, with the addition of a Stage 4 to the price series from 15 November 2018 to 22 January 2019 where prices fell to 55% of the previous low

(29 June 2018) and then recovered to 60%.

6.2 Framework

Word2vec models were trained ('gensim' Version 3.5.0 [289]) using text from all submissions from 00:00 1 January 2017 to before 00:00 23 January 2019 (GMT). This used the default hyperparameter values suggested by gensim [289], except that the number of noise words drawn (in the case of negative sampling) and iterations were increased to 20 to reflect the limited dataset size [204,205]. Words with a total frequency below 100 were excluded.

Two-sided Wilcoxon Rank-Sum Tests were applied to delineate words that changed statistically significantly in frequency across stages in the price series. The trained word2vec models assigned to each word a vector of 100 continuous-scaled numbers. The risers and fallers were placed on separate undirected graphs ('NetworkX' Version 2.2 [213]) where each edge had a weight corresponding to the cosine similarity between the words' vectors [157]. The weight thus measured how similar the context was in which the two words were used [204]. A threshold was applied to remove the edges with the lowest cosine similarities. Topics were identified as groups of more than one word that were connected with each other and not connected with words outside the group.

Compared with applying k-means [163, 166], this graphical approach to clustering word2vec-represented words obviated the need to select the number of topics and allowed for polysemy.

6.3 Experiments in Topic Modelling Optimisation

6.3.1 Datasets

Table 6.1 shows a decline in the number of words that statistically significantly rose and fell over time as successive stages had fewer associated days and submissions, and so less data were available. This was exacerbated by a decline in Reddit activity. Over 500 submissions per day were being posted on average in Stages 1 and 2, the periods when prices were most volatile. This fell 46% as prices stabilised (Stage 3) and by a further 10% in Stage 4.

Table 6.1: Descriptive statistics for Reddit Submissions (1 January 2017 to 22 January 2019).

Stage	Days	Submissions	Submissions per Day	Risers	Fallers
All	752	338415	450.02	N/A	N/A
1	349	181327	519.56	N/A	N/A
2	195	101110	518.51	129	586
3	139	38706	278.46	83	40
4	69	17272	250.32	63	8

6.3.2 Model Variants

Experiments compared four different word2vec architectures in deploying this framework. Using a neural network trained to predict the current word using its context, the Continuous Bag-of-Words model (CBOW), was evaluated against training to predict the context using the current word, the continuous Skip-gram (SG) model [204]. Computational complexity being mitigated through the original approach of Hierarchical Softmax (HS) [204] was assessed against the alternative Negative Sampling (NEG) [205]. The following percentile thresholds applied to the graph were compared: 90, 95, 99, 99.90, 99.95 and 99.99. A pre-trained model was not used in comparison as these were developed for words without stemming and lemmatisation [120].

6.3.3 Evaluation Metrics

A ‘group’ here refers to two or more words that are connected by edges. The words within each group generated should be similar to each other and dissimilar with words outside the group. The median cosine similarity between words within the same group (‘INTRA’) and between words in a group and words outside (‘INTER’) were calculated. These are of the same scale and so INTER was deducted from INTRA to provide a measure of the quality of the groups generated. Using just this quality metric resulted in only one or two groups being generated with the exception of the words that fell from Stages 1 to 2.

The more groups, the more potential, distinct topics that can be interpreted from them, and so the quality metric was multiplied by the number of groups generated (Equation 6.1), resulting in an evaluation score (‘EVAL’). Models with a negative INTRA or INTER score were excluded. This meant that the same percentage increase in either quality or number of groups had the same impact on the evaluation metric.

$$EVAL = (INTRA - INTER) \times \text{Number of Groups} \tag{6.1}$$

6.3.4 Evaluating Model Variants

Tables 6.2 to 6.6 compare the results of applying different word2vec architectures to each dataset of rising or falling words. For each word2vec architecture (‘Model’), a threshold was selected (‘Threshold’) that maximised the value of the evaluation score with each table showing the resulting optimal number of groups (‘Groups’) and evaluation score (‘EVAL’).

Examining Stages 1 to 2 (see Tables 6.2 and 6.3), the evaluation score was similar across different word2vec architectures applied to the same dataset. The exception was that, for the risers from Stages 1 to 2 (Table 6.3), the evaluation score for SG with HS was 41% higher (11.45) compared with the second highest value (8.12). The architecture SG with HS had the highest evaluation score for words falling from Stages 2 to 3 (Table 6.4) and rising from Stages 3 to 4 (Table 6.6). There was, however, no consistent tendency for one word2vec architecture to outperform all others: the architecture SG with HS had the third highest evaluation score in Table 6.5 and second highest in Table 6.2. The results for the words falling from Stages 3 to 4 are not shown as there were only eight words and only one group was generated across the models compared.

A further result was that the more words available for extracting topics, the more groups were generated and the higher the optimal threshold.

Table 6.2: Grouping 586 Words Falling from Stages 1 to 2

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.90	0.6863	0.1939	73	35.95
SG, HS	99.90	0.5946	0.0360	67	37.43
CBOW, NEG	99.90	0.6626	0.0206	62	39.81
CBOW, HS	99.90	0.5742	0.002178	62	35.46

Table 6.3: Grouping 129 Words Rising from Stages 1 to 2

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.00	0.6139	0.1865	19	8.12
SG, HS	99.00	0.6815	0.0790	19	11.45
CBOW, NEG	99.00	0.5068	0.0573	15	6.74
CBOW, HS	99.00	0.4806	0.0297	17	7.67

Table 6.4: Grouping 40 Words Falling from Stages 2 to 3

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	95.00	0.5893	0.1696	7	2.94
SG, HS	95.00	0.6013	0.0530	8	4.39
CBOW, NEG	95.00	0.5879	0.0351	6	3.32
CBOW, HS	95.00	0.5476	0.0120	7	3.75

Table 6.5: Grouping 83 Words Rising from Stages 2 to 3

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.00	0.5514	0.2154	8	2.69
SG, HS	99.00	0.5881	0.0895	10	4.99
CBOW, NEG	99.00	0.5517	0.0509	12	6.01
CBOW, HS	99.00	0.4932	0.0286	12	5.58

Table 6.6: Grouping 63 Words Rising from Stages 3 to 4

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.00	0.6918	0.2246	8	3.74
SG, HS	99.00	0.7712	0.0607	9	6.39
CBOW, NEG	99.00	0.8036	0.0317	8	6.18
CBOW, HS	99.00	0.7005	0.0060	8	5.56

6.3.4.1 Topic Modelling

The shift from Stage 1 (rising) to Stage 2 (falling) prices had the most associated data (Table 6.1). The optimal model for falling words was CBOW with NEG (Table 6.2) and for rising was SG with HS (Table 6.3). The largest groups (with more than three words) are displayed in Figures 6.2 and 6.1 using a force-embedded algorithm (following Fruchterman and Reingold [110]) to display the graph for each group.

The topics generated by the different approaches were similar. The optimal model identified eight topics in the fallers which when the other three model variants were examined were constant. For risers, the results were again constant for the SG with NEG and with HS, but no 'ICO' topic could be found for CBOW with NEG and no 'Startup' topic could be found for CBOW with HS.

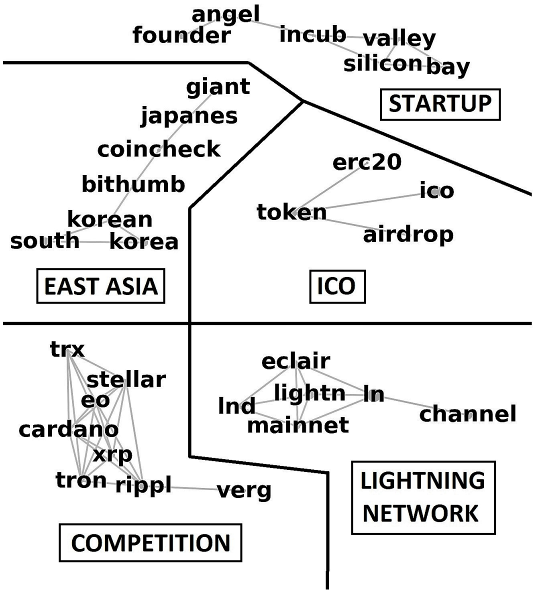


Figure 6.1: Groups Rising in Frequency from Stages 1 to 2.

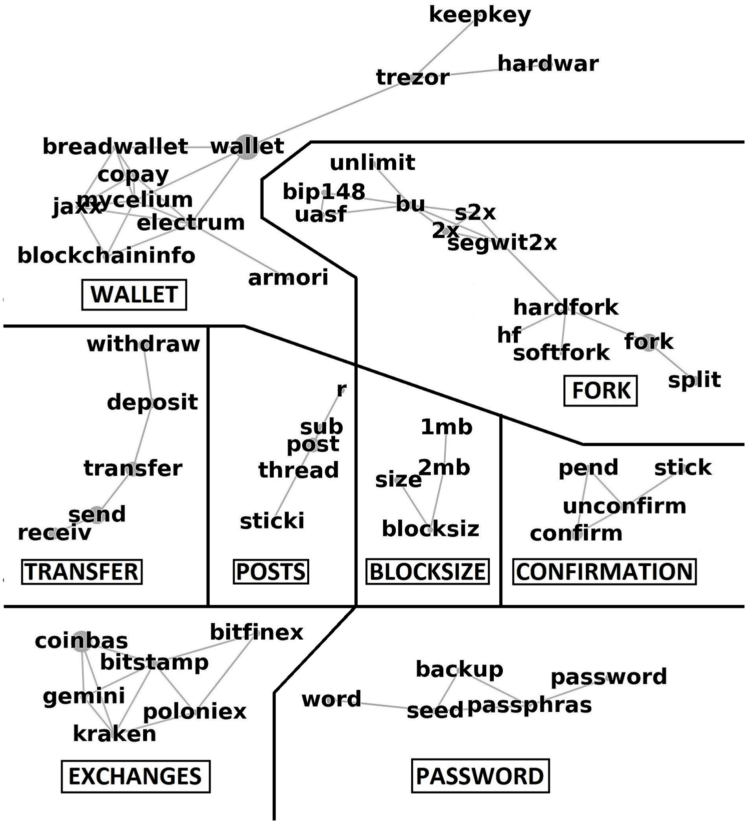


Figure 6.2: Groups Falling in Frequency from Stages 1 to 2.

6.3.4.2 Smaller Groups Identified

Smaller groups (with three or fewer words) identified for words **rising** in frequency from Stages 1 to 2:

- Regulation: ('cftc', 'chairman'), ('regul', 'g20')
- Sentiment: ('bull', 'bear'), ('rebound', 'slump'), ('bullish', 'bearish', 'sentiment')
- Manipulation: ('whale', 'manipul')
- Influencers: ('mcafe', 'john'), ('buffett', 'warren', 'buffet')
- Social Media: ('discord', 'telegram')

Smaller groups identified for words **falling** in frequency from Stages 1 to 2:

- Price: ('cad', 'dollar_marker_symbol', 'worth')
- Acquiring: ('sell', 'purchas', 'buy'), ('ach', 'wire')
- Understanding: ('best', 'safest', 'safe'), ('explain', 'eli5')
- Hash Rate: ('difficulti', 'hashpow', 'hashrat')
- Related to 'Fork': ('btu', 'bch', 'bcc'), ('agreement', 'nya'), ('btc1', 'core')
- Influencers: ('silbert', 'barri'), ('garzik', 'jeff')
- Firms: ('okcoin', 'cni', 'btcc')
- Mining pools: ('antpool', 'pool')

6.4 Interpretation

The largest groups of rising words clustered around five topics (Figure 6.1) centred on ‘East Asia’, ‘Competition’, ‘Startup’, ‘ICO’ and the ‘Lightning Network’. Regarding East Asia, Japanese Coincheck and South Korean Bithumb were both subject to investigations and hacks [218], whilst ‘giant’ could refer to large Japanese firms entering partnerships with exchanges to accept bitcoin [273]. Bitcoin competitors that became more discussed included Tron (‘trx’), Stellar, EOS (‘eo’), Cardano, Ripple (‘rippl’, ‘xrp’) and Verge (‘verg’). The ‘Startup’ topic focussed on incubators (‘incub’), the Silicon Valley (‘silicon’, ‘valley’ and ‘bay’), investment (‘angel’) and founders. There was also growing interest in ICOs (‘ico’) and the Lightning Network. Examining the smaller groups revealed interest in regulation that reflected reported fears of global regulation from the G20 [51] and actions from the US CFTC [203]. There was further rising interest in sentiment and market manipulation, a possible response to falling prices, as well as in social media websites Telegram and Discord.

The largest groups of falling words clustered around eight topics (Figure 6.2) which reflected a notable fall in discussion around how Bitcoin works. This involved topics covering ‘Wallet’, ‘Transfer’, ‘Exchanges’, ‘Password’ and ‘Posts’. In response to escalating confirmation times (topic ‘Confirmation’), a split (topic ‘Fork’) emerged between Bitcoin Unlimited (‘bu’ and ‘unlimit’), for a larger blocksize limit (topic ‘Blocksize’), and Segregated Witness (SegWit), for moving information off network [71]. This involved protests such as the User-Activated Soft Fork (‘uasf’) Bitcoin Improvement Proposal 148 (‘bip148’) [134] and the abortive compromise SegWit2x [18]. Results show how interest in this debate declined after SegWit was implemented and Bitcoin Cash was forked, both on 1 August 2017 [71]. The smaller groups corroborated with the themes identified by the larger groups and further showed a declining interest in the price of bitcoin and in acquiring bitcoin.

The importance of the views of specific influencers was suggested by the inclusion of their names among rising and falling words (see Section 6.3.4.2). With rising prices (Stage 1) becoming falling (Stage 2), Barry Silbert and Jeff Garzik became less mentioned, whilst John McAfee and Warren Buffett became more popular.

6.5 Discussion

Removing the requirement of DDPWI (Chapter 5) that the word frequency change must revert in the next phase resulted in more words being delineated (see Table 6.1). This study optimised word2vec-based topic modelling to determine topics among the words that changed in frequency. Comparing different word2vec architectures showed that no single architecture consistently provided optimal results showing the need to compare all the architectures when applying this approach. The optimal models were applied to the most significant shift in bitcoin prices from across 2017 to 2018 and this led to the emergence of intuitive groups.

Topics that rose in frequency from phase 1 to 2 could be linked with events or concerns occurring within phase 2 (see Section 6.4). These results suggest a higher interest in regulation (as previously observed in Chapter 5), developments in East Asian markets and the views of influencers such as John McAfee and Warren Buffett. These topics provide only possible, not proven, explanations for why bitcoin prices shifted from rising (phase 1) to overall falling and volatile (phase 2).

Topics that fell in frequency from phase 1 to 2 could be intuitively linked with subjects that had been of interest in phase 1 but were no longer as relevant by phase 2 (see Section 6.4). These topics could be split between those capturing an interest in Bitcoin forks and those suggestive of an overall enthusiasm for Bitcoin, such as discussing how to acquire bitcoin, its price and how it works. As discussed in Section 6.4, the discussion of forks in phase 1 coincided with a debate over forks in the Bitcoin codebase. The higher apparent enthusiasm for Bitcoin in phase 1 was probably symptomatic of the rising prices in phase 1 that may have encouraged a desire to hold Bitcoin.

6.6 A Comparison of DDPWI with Word2vec-based Topic Modelling

Whilst DDPWI resulted in three price dynamic words ('ban', 'tax' and US Dollars) associated with phase 2, the word2vec-based approach identified 129 words which rose in frequency and 586 words which fell in frequency as bitcoin prices shifted from phase 1 to phase 2 (see Table 6.1).

The word2vec-based topic modelling approach consolidated these words into topics. This involved black-box neural networks that meant the DDPWI methodology was more transparent. The word2vec-based topic modelling approach also had greater computational requirements. There was a need to train the word2vec model on cryptoasset-related text. As no one word2vec architecture was consistently optimal, further computation was also required in comparing all architectures before applying the optimal approach.

Word2vec-based topic modelling can be deployed to compare any two datasets of word frequencies. Whereas, DDPWI requires specifically a consecutive, triphasic dataset where the word frequencies in a single phase in the price series can be compared with the phases chronologically before and after.

The next chapter extracts plausible causes of phasic shifts in price from social media text using the word2vec-based topic modelling approach. Both the 'mono-phase' and 'multi-phase' analyses require a tool that can extract concepts that changed in frequency across two datasets. Unlike DDPWI, word2vec-based topic modelling is sufficiently flexible to provide such a tool and so was used as part of the methodology (see Section 7.4).

Chapter 7

Cryptoasset Phasic Shifts and Causality

7.1 Introduction

Previous chapters examined the association between word use and a particular phase or topic use and phasic shifts in the bitcoin price (Chapters 5 and 6). The current chapter evolves the debate from what events and concerns are associated with different phases in price to what are the plausible causes of these phasic shifts in price. A standardised pipeline is developed and applied to both the bitcoin and ether price series that applies a common approach to preparing the text and finding phases in the cryptoasset price series. Examining both Bitcoin and Ethereum enables a comparison of the results which is used to determine if the discovered causes are specific to a cryptoasset or shared between them.

In considering causality, if an event occurs as price changes, that event could be driving the change in price, but a reasonable alternative explanation is that the event is in response to the change in price. To exclude the latter possibility, cause must come before effect as the future cannot affect the past [35, 122, 146]. Hence, the event must precede the price change, and such events, therefore, may be predictive.

Previous literature has focussed on models that assess if certain features are predictive of the cryptoasset price (see Sections 2.3 and 2.4.1). However, establishing a predictive relationship does not prove a causal link because of ‘confounding bias' [223]. That is to say if one event occurs before another, both may be the symptoms of a third factor changing [223] or there may have been a catalyst unique to that dataset without which the causal link ceases [252].

Ideally, experiments would be carried out to reduce the risk of confounding bias [223, 251], but for cryptoassets we have only observational data. Although observational data cannot prove that a candidate caused a change it can provide evidence that favours this explanation over confounding bias [223,251]. It is in this context that healthcare epidemiologists often operate to find the underlying causes of disease, as, for instance, with the link between smoking and lung cancer [74,251].

7.2 The Causality Framework

The approach (see Figure 7.1) is to filter words from social media text, group words of similar meaning to identify the underlying concepts, and then to apply quantitative causality criteria. There is then an examination of the context of the delineated concepts and evaluation of the coherence of suggested causal links with known facts [35]. Healthcare epidemiology literature suggests two distinct approaches to constructing the quantitative causality criteria.

The first approach uses the strength of the association to support a causal link [35, 251]. The larger the increase in the candidate cause and the greater the effect, the more any third, unconsidered, 'confounding' variable would have to affect both for the association to be spurious and not indicative of a causal relationship [74, 125, 251]. This is applicable to identifying rare, unpredictable black swan events that have a one-off influence on a single, major phasic shift in the price series. The 'mono-phase' analysis (see Figure 7.1) focuses on the major change in the price series which is the shift in movement from the phase of rising prices before to the phase of falling prices after the all time high price. This analysis filters for words that were statistically significantly higher in frequency in the latter phase of falling values. The causality criteria used are: frequency is more than three-fold higher [125] in the phase of falling prices than the phase of rising prices, and frequency is higher within the 24 hours before the maximum price. A cut-off is used that the concept must be more than three-fold higher in frequency to reduce the risk that the detected association is spurious. This is consistent with recommendations in the epidemiology literature regarding the definition of what constitutes ‘strong support for causation’ [125].

The alternative approach places value in relationships that consistently recur despite a changing context [35, 146]. The more an observed association recurs across different contexts, the more likely any unobserved variables would have changed in value and impact, and so the less likely that the observed association is due to some unobserved variable driving both candidate cause and effect. This approach can detect potential causes with a recurring effect on the price series. In the ‘multi-phase’ analysis (see Figure 7.1), words are filtered for where daily frequency was statistically significantly different comparing all phases of rising values with all phases of falling values. A concept captured a potential recurring cause of rising values if its frequency was higher in every phase of rising values compared with the previous phase and higher within the 24 hours before each phase of rising values. Concepts reflecting potential causes of falling values have a higher frequency in every phase of falling values compared with the previous phase and a higher frequency within the 24 hours before each phase of falling values.

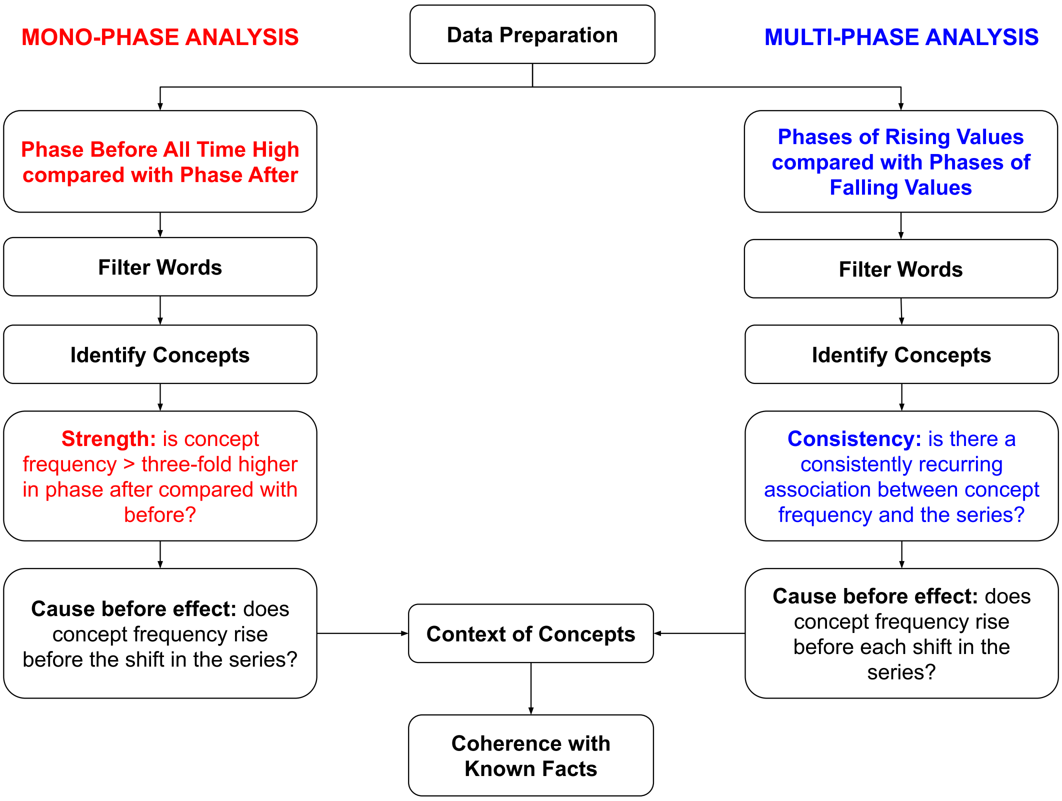

MONO-PHASE ANALYSIS
Data Preparation
MULTI-PHASE ANALYSIS
Phase Before All Time High compared with Phase After
Phases of Rising Values compared with Phases of Falling Values
Filter Words
Filter Words
Identify Concepts
Identify Concepts
Strength: is concept frequency > three-fold higher in phase after compared with before?
Consistency: is there a consistently recurring association between concept frequency and the series?
Cause before effect: does concept frequency rise before the shift in the series?
Context of Concepts
Cause before effect: does concept frequency rise before each shift in the series?
Coherence with Known Facts

Figure 7.1: The causality framework. This evaluates evidence for or against an event and/or concern on social media having an impact on price. The framework begins in the box labelled 'Data Preparation'. The mono-phase analysis follows the route on the left and the multi-phase analysis follows the route on the right; differences in approach are indicated by coloured text. The process terminates in the box labelled 'Coherence with Known Facts'.

7.3 Data preparation

7.3.1 Dataset

The dataset extended from 1 January 2017 to 14 May 2019 and included: Reddit submissions text sourced using the Pushshift API [15], the US Dollar bitcoin price from the Charts API of [26] and the US Dollar ether price from [99]. Text from subreddit 'r/Bitcoin' was used for Bitcoin analyses and combined text from 'r/ethereum', 'r/ethtrader' and 'r/EtherMining' was used for Ethereum analyses (see Section 2.3.6 in Chapter 2).

7.3.2 Dividing the price series into phases

The price data were divided into phases using local maxima and minima to define the boundaries (see Figure 7.2). A date represented a local maximum if the price was higher than on any other date 28 days (4 weeks) before and after. That date was a local minimum if the price was instead lower than on any other date 28 days before and after. Phases terminating just before a local maximum were rising price phases, those ending just before a local minimum were falling price phases. Sometimes there were several consecutive minima with the last value being the lowest; all such minima except the last, lowest value were ignored.

The length of the window was specified at 28 days before and after because a longer window risked merging rising and falling price phases. For example, examining bitcoin, the 28-day window delineated a phase where bitcoin prices fell 65% from the all time high price on 16 December 2017 to 5 February 2018 (see Figures 1.1 and 7.2). Doubling the length of this window to 56 days would have enlarged this phase of price movement to include the subsequent 70% increase in prices from 5 February 2018 to 5 March 2018. Using shorter time windows would have reduced the size of the price phases, limiting the amount of data available when applying Wilcoxon Rank-Sum Tests to filter words in the mono-phase analysis (described in Section 7.4.1.1). This would have reduced the power of such tests [37].

As bitcoin prices rose across 2017, there were brief phases where bitcoin prices reversed upon reaching round values. This occurred at 1000 US Dollars (1285.14 to 941.92 from 3-24 March 2017); 3000 US Dollars (2961.83 to 1931.21 from 11 June to 16 July 2017); and 5000 US Dollars (4911.74 to 3319.63 from 1-14 September 2017). Traders sell at round values that represent a large return on their investment to prevent losing this return to subsequent volatility, even if their view of the cryptoasset is unchanged [56]. Therefore, these phases were incorporated into the overall rising price phase.

When technical traders believe that a certain price level is a support or resistance level, they will buy (pushing prices up) as prices fall to that support level and sell (pushing prices down) as prices rise to that resistance level [209]. When prices approach a round-valued price this can drive reversals in trend even if opinion of the cryptoasset is otherwise unchanged [4, 90, 261, 291]. These phases where the connect between price and non-price events and concerns is weak were excluded.

In 2017, the ether price rose to 394.66 US Dollars (12 June), fell to near 150 US Dollars (155.42 US Dollars, 16 July 2017), then rose again to 391.42 US Dollars (1 September 2017) (Figure 1.1). This supports a 400 US Dollar price resistance level identified by the media at the time [14, 295]. Hence, the phase from 12 June (where the barrier was first neared) to before 23 November 2017 (when the barrier was exceeded) is removed from analysis.

In 2018, the bitcoin price fell to 5908.70 US Dollars (29 June 2018), recovered and tested the barrier again at 6050.94 (14 August 2018). Hence the 6000 US Dollar support level has been described as a ‘crucial test’ [77]. The phase from 29 June 2018 to before 15 November 2018 (when prices finally fell below the barrier) is removed from analysis.

After attaining a local minimum in mid-December 2018, neither the bitcoin nor ether price fell further. This point thus marks the end of the 2017-18 price cycle this thesis focusses on, and so the last phase of data analysed ends mid-December 2018 for both cryptoassets (14 December for Ethereum and 15 December for Bitcoin).

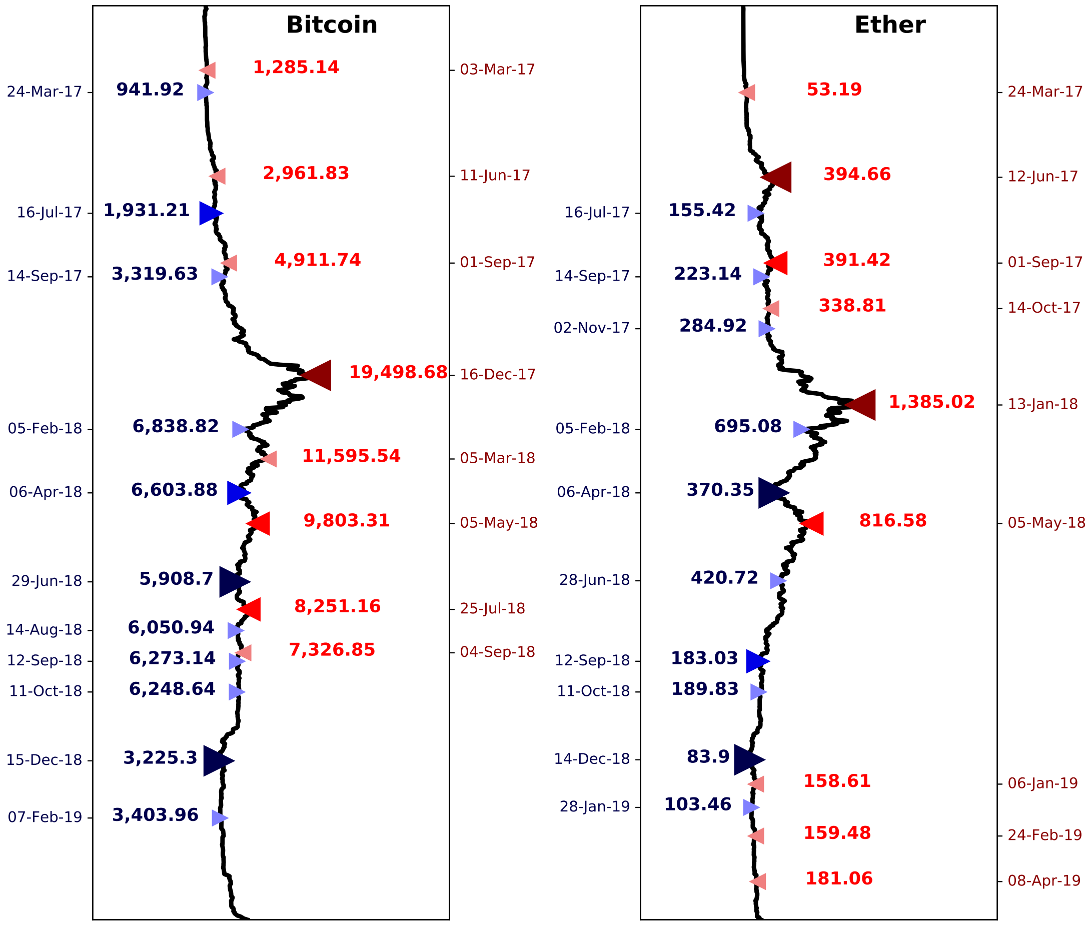


Figure 7.2: Comparison of ether and bitcoin US Dollar Price Local Extrema (1 January 2017 to 14 May 2019). Local minima indicated by blue '▷' and local maxima by red '◁'. Smallest, lightest-coloured symbols indicate most extreme (highest or lowest) price for 28 days (4 weeks or about 1 month) before and after; next size 56 days (8 weeks or about 2 months); and largest, darkest-coloured 84 days (12 weeks or about 3 months) before and after. Dates of minima on left and maxima on right.

7.3.3 Text preparation

Reddit submission processing involved: removing blank, duplicate and automated submissions, standardising text of synonymous meaning and deleting text not relating to words. Each submission was converted from a string of text into a list of distinct words.

7.3.3.1 Submissions filtered out

The following submissions were removed:

- Automated submissions authored by the following: ‘AutoModerator’, ‘CommunityPoints’, ‘rBitcoinMod’ and ‘crypto_bot’;
- Those consisting of duplicate text;
- Those containing just ‘[deleted]’ or ‘[removed]’; and
- Blank submissions.

7.3.3.2 Text processing

All text was placed into lower case and strings of 50 or more word characters (too long to represent a word) were removed. The following details the approach to standardising synonymous words (e.g. ‘BTC’ and ‘bitcoin’) and replacing terms of multiple words (‘smart contract’) with single words (‘smartcontract’). This accounted for words being separated by whitespace characters (‘smart contract’) and hyphens (‘smart-contract’); and for spelling variants (decentralised and decentralized).

Currency codes – Cryptoasset codes were replaced by the name of the associated cryptoasset (see Table 7.1). References to '1BTC' or '1 XBT' became '1 bitcoin'; 'ETH' and 'ether(s)' became 'ethereum'. This was applied to the top 10 cryptoassets by market capitalisation and/or liquidity (13:41 GMT; 21 May 2019): Bitcoin, Ethereum, Ripple, Bitcoin Cash, Litecoin, Binance Coin, Tether, Stellar, Cardano and Tron. The cryptoassets EOS, Matic Network and NEO did not have a distinct currency code. The abbreviation SAT was further replaced with satoshi [22]. Other cryptoassets were added to this list where highlighted by previous runs of the methodology: Golem, Verge, Ethereum Classic, Bitcoin Unlimited, Iconomi, Distributed Credit Chain, UChain, Bancor, Maker DAO, DIGIX and Auctus. Before this, references to US dollars were standardised.

Table 7.1: Replacing currency code with associated name. Conversion of ETC into 'ethereumclassic' was conducted prior to lower-case conversion to prevent confusion with *et cetera* ('etc.').

Replacing Term	**Terms Replaced**
dollarmarkersymbol	'(us/u.s.) dollar(s)'; 'usd'; '$'
ethereum	'eth'; 'ether(s)'
ethereumclassic	'ETC'; 'ethereum classic'
bitcoincash	'bch'; 'bitcoin cash'; 'bcash'
bitcoin	'btc'; 'xbt'; 'bitcoins'
satoshi	'sat(s)'; 'satoshis'
tron	'trx'
ripple	'xrp'
stellar	'xlm'
cardano	'ada'
litecoin	'ltc', 'litecoins'
golem	'gnt'
tether	'usdt'
binancecoin	'bnb'; 'binance coin(s)'
verge	'xvg'; 'verge currency'
bitcoinunlimited	'bu'; 'btu'; 'bitcoin unlimited'
iconomi	'icn'
distributedcreditchain	'distributed credit chain'; 'dcc'
uchain	'ucn'
bancor	"bancor(')(s) network token"; 'bnt'
makerdao	'maker dao'; 'dai'
digix	'dgx'; 'dgd'; 'digix dao'; 'digix gold token(s)'
auctus	'auc'

Improvement proposals – The following improvement proposal references were standardised: ‘bitcoin improvement proposal(s)’ and ‘bips’ were converted to ‘bip’; ‘ethereum improvement proposal(s)’ and ‘eips’ were converted to ‘eip’; and ‘ethereum request(s) for comment(s)’ was changed to ‘erc’. References to the same numbered proposal were standardised through removing the gap between the proposal type (‘erc’) and number of proposal (‘20’). Hence, ‘erc-20’, ‘erc 20’ and ‘erc20’ all became ‘erc20’.

Cryptoasset, financial, regulator and nationality words – Ethereum-related (Table 7.2); bitcoin-related (Table 7.3); cryptoasset-related (Table 7.4); and finance-related (Table 7.5) terminology were standardised.

Table 7.2: Ethereum concepts standardised.

New Term	**Words Replaced**
smartcontract	‘smart contract(s)’
evm	‘ethereum virtual machine’
dapp	‘decentralized application(s)’; ‘dapp(s)’; ‘dap(s)’
dao	‘decentralized autonomous organization(s)’; ‘dao(s)’
dac	‘decentralized autonomous corporation(s)’; ‘dac(s)’
ico	‘initial coin/token offering(s)’; ‘token generation event(s)’; ‘ico(s)’; ‘ito(s)’; ‘tge(s)’
eea	‘enterprise ethereum alliance’

Table 7.3: Bitcoin concepts standardised. The term ‘lightening’ is a common mistake in spelling ‘lightning’ [281].

New Term	**Words Replaced**
ln	‘light(e)ning network(s)’
segwit	‘segregated witness’; ‘sw’
segwit2x	‘b2x’; ‘s(w)2x’; ‘s(w)2mb’; ‘segwit 2mb’; ‘segwit2mb’; ‘segwit 2x’
nya	‘bitcoin scaling agreement at consensus 2017’; ‘new york agreement’

Table 7.4: Cryptoasset concepts standardised.

New Term	Words Replaced
cryptoasset	‘crypto currency/ies’; ‘crypto asset(s)’; ‘cryptocurrency/ies’
delegatedproofofstake	‘delegated proof of stake’; ‘dpos’
proofofstake	‘proof of stake’; ‘pos’
proofofwork	‘proof of work’; ‘pow’
proofofauthority	‘proof of authority’; ‘poa’
byzantinefaulttolerance	‘byzantine fault tolerance’; ‘bft’
directedacyclicgraph	‘directed acyclic graph(s)’; ‘dag’
storeofvalue	‘store of value’; ‘sov’
mediumofexchange	‘medium of exchange’; ‘moe’
unitofaccount	‘unit of account’; ‘uoa’
cpu	‘central processing unit(s)’; ‘cpus’
gpu	‘graphics processing unit(s)’; ‘gpus’
asic	‘application specific integrated circuit(s)’; ‘asics’
asicboost	‘asic boost’
uasf	‘user activated soft fork(s)’
hashrate	‘hash power’; ‘hash rate’
twofactorauthentication	‘two/2/multi factor authentication’; ‘2fa’
ddos	‘distributed denial of service’
ipfs	‘interplanetary file(s) system’
pki	‘public key infrastructure’
publickey	‘public key(s)’
privatekey	‘private key(s)’
nonce	‘number used only once’
hardfork	‘hard fork’; ‘hf’
softfork	‘soft fork’
hd	‘hierarchical deterministic’
explain	‘eli5’
fud	‘fear(,) uncertainty(,) (and) doubt’
ai	‘artificial intelligence’
transaction	‘tx’
txid	‘transaction id(entification)’
tpsec	‘transaction(s) per second’; ‘tps’

Table 7.5: Finance concepts and nationalities standardised.

New Term	**Words Replaced**
etf	‘exchange traded fund(s)’
etp	‘exchange traded product(s)’
otc	‘over the counter’
dex	‘decentralised exchange(s)’
cex	‘centralised exchange(s)’
pumpanddump	‘pump(s/ed/ing) and dump(s/ed/ing)’
marketcap	‘market cap(italisation)(s)’
larger	‘bigger’; ‘larger’
technicalanalysis	‘ta’; ‘technical analysis’
fundamentalanalysis	‘fa’; ‘fundamental analysis’
kyc	‘know your customer/client’
sec	‘securities and exchange commission’
ftc	‘federal trade commission’
cftc	‘commodity futures trading commission’
fdic	‘federal deposit insurance corporation’
doj	‘department of justice’
g20	‘group of twenty/20’
pboc	“people’s bank of china”; ‘pbc’
cboe	‘chicago board options exchange’
ice	‘intercontinental exchange’
p2p	‘peer to/2 peer’
korea	‘(south) korea(n)’
france	‘french’
china	‘chinese’

7.3.3.3 Text removed

The following were removed respectively: URLs; HTML tags (e.g. '&amp'); the new line character ('\n'); references to deleted text ('[removed]' and '[deleted]'); greetings ('hey', 'hi' and 'hello') and non-ASCII text (e.g. Cyrillic alphabet or emoticons). Punctuation and apostrophes were removed unless these were inside words to indicate abbreviations (e.g. 'o'clock').

7.3.3.4 Creating lists of words from strings of text

The processed text was tokenised into word lists using Python package NLTK version 3.3 and its associated download 'punkt'. 'Stopwords' were then removed using the list provided by NLTK, supplemented by abbreviations for 'not' ("n't"); 'I am' ('im', "i'm"); 'you are' ("you're", 'youre'); '(s)he is' ('(s)hes', "(s)he's"); 'they are' ('theyr', "they'r", "they're", 'theyre'); and 'we are' ('wer', "we'r", "we're"). Words were also removed that contained no letters, thus deleting any numbers, along with references to thousands ('5k' or '14k'), millions ('5m', '1m'), multiples ('10x'), ranks ('1st', '2nd' or '4th') and images ('img'). Words were lemmatised using NLTK's 'WordNetLemmatizer', and stemmed using 'SnowballStemmer'. The 'snowball' stemmer was selected in being least likely to treat words of the same concept differently or words of a different concept the same [154]. Table 7.6 lists lemmatised and stemmed words that were standardised as they referred to similar concepts.

Table 7.6: Lemmatised and stemmed words standardised.

New Word	Word Replaced
'mine'	'miner'
'newbi'	'noob'; 'n00b'; 'newb'
'buy'	'purchas'
'ad'	'advertis'; 'advert'
'mew'	'myetherwallet'; 'myetherwalletcom'; 'wwwmyetherwalletcom'
'verif'	'verifi'; 'verif'
'repli'	'respons'
'might'	'mayb'
'partner'	'partnership'

7.3.4 Measuring frequency

With each submission represented as a list of words, the number of submissions across a defined time period that contained each word could be counted. This was then divided by the total number of submissions such that the ‘frequency’ or ‘popularity’ of a word was the proportion of submissions across a defined time period that contained that word at least once. Extending to groups containing multiple words, frequency was the proportion of submissions containing at least one word from that group. Daily frequency referred to the proportion of submissions containing a word or a word from a group on each day. Following the sources on price data [26, 99], a ‘day’ was specified to be from 00:00 on a given day to before 00:00 the next date (GMT).

7.4 Methodology

An overview of the methodology is provided in Figure 7.1.

7.4.1 Mono-phase analysis

7.4.1.1 Filter words

One-tailed Wilcoxon Rank-Sum Tests (SciPy package version 1.1.0) and a Bonferroni-corrected p-value threshold of 1% were applied to filter for those words where the daily word frequency tended to be higher in the phase after the all time high price compared with before. Prior to this, extremely rare words in 100 or less submissions were removed.

7.4.1.2 Identify concepts

Word2vec-based topic modelling was applied following the methodology specified in Chapter 6. This produced groups of connected words, which were merged into single ‘concepts’ (such as ‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’), and words unconnected with any other word (‘korea’), which were treated as concepts consisting of only one word. Hence, the ‘concepts’ examined consisted of one or more words that shared a similar meaning. This required the use of word2vec models [204, 205], which were trained on the processed text from all submissions, as well as Python packages 'gensim' [289] version 3.5.0 and 'NetworkX' [213] version 2.2.

7.4.1.3 Apply causality criteria: strength and cause before effect

Mono-phase concepts were more than three-fold higher in popularity [125] across the phase after the all time high price compared with the phase before, and increased in frequency before the shift in phase. Determining whether frequency rose before the shift involved examining one hour, two hours, three hours, and so on, up to 24 hours before the shift and evaluating whether the proportion of submissions containing the concept within any of these windows was higher compared with all the submissions in the same phase but before that window.

7.4.2 Multi-phase analysis

7.4.2.1 Filter words

Two-tailed Wilcoxon Rank-Sum Tests (SciPy package version 1.1.0) and a Bonferroni-corrected p-value threshold of 1% were applied to extract those words where the daily word frequency tended to be higher or lower comparing all phases where prices rose with all phases where prices fell. Prior to this, extremely rare words in 100 or less submissions were removed.

7.4.2.2 Identify concepts

Words more frequent as prices rose were split from those more popular as prices fell. Word2vec topic modelling was applied, as in Section 7.4.1.2, to convert each set of words into a set of concepts: 'rising-price concepts' consisted of words higher in frequency as prices rose and 'falling-price concepts' consisted of words more frequent as prices fell.

7.4.2.3 Apply causality criteria: consistency and cause before effect

Rising-price, multi-phase concepts were rising-price concepts that rose in frequency with every shift to rising prices and within the 24 hours before every shift to rising prices. Falling-price, multi-phase concepts were falling-price concepts that rose in frequency with every shift to falling prices and within the 24 hours before every shift to falling prices. Removed from the analysis was any concept that consistently rose in popularity across every shift in price, independent of whether prices were rising or falling, as any rise in popularity could have been an artefact of the long-term trend.

7.4.3 Context of concepts

Establishing the context of a concept involved finding the top five most common words occurring in submissions containing at least one word from that concept. The context was established for each mono-phase and multi-phase concept. The following words from the text were removed before running the analysis as these did not aid in the interpretation of the concept: the name of the cryptoasset being analysed, 'account', 'actual', 'add', 'address', 'ago', 'alreadi', 'also', 'amount', 'anyon', 'appli', 'back', 'blockchain', 'come', 'communiti', 'could', 'crypto', 'cryptoasset', 'current', 'day', 'differ', 'drive', 'end', 'even', 'everi', 'exchang', 'extra', 'feel', 'find', 'first', 'get', 'give', 'go', 'group', 'happen', 'howev', 'includ', 'keep', 'know', 'let', 'like', 'look', 'lot', 'make', 'mani', 'may', 'money', 'much', 'multipl', 'need', 'next', 'one', 'peopl', 'pleas', 'put', 'rememb', 'right', 'run', 'say', 'see', 'similar', 'someth', 'start', 'still', 'take', 'talk', 'thing', 'think', 'time', 'two', 'use', 'user', 'want', 'way', 'whole', 'work', 'would', 'year' and 'yet'. If two or more words were in the same percentage of submissions (rounded to two decimal places), such words were treated as being ranked equally.

7.5 Results

7.5.1 Comparison of Bitcoin and Ethereum price phases

Both the bitcoin and the ether price rose to an all time high at the end of 2017 and beginning of 2018, to then oscillate with an overall decline in value until mid-December 2018 (see Figure 7.2). There was a disparity in the timing of the all time high price for bitcoin (16 December 2017) and ether (13 January 2018).

It appears that different price levels acted as barriers at different times. Whilst bitcoin prices rose across 2017, ether prices reverted upon nearing 400 US Dollars [14, 295] (12 June and 1 September 2017), only increasing above this level after five months. Whilst ether prices fell from 5 May to mid-December 2018, bitcoin prices recovered upon falling to 6000 US Dollars [77] (29 June and 14 August 2018) and only fell below this level after four months.

Based on local extrema (see Figure 7.2) and price barriers, six phases of price movement with ether and eight with bitcoin were demarcated (see Table 7.7). Table 7.7 further shows which of these phases were used in order to compare daily word frequencies so as to filter words (see Sections 7.4.1.1 and 7.4.2.1). Word2vec topic modelling was then applied to create concepts from these words, with the specifications of the word2vec topic modelling approach provided in Table 7.8. Descriptive statistics for the different phases are provided in Table 7.9.

Table 7.7: For each phase in the cryptocurrency price series: the date range, price movement, overall percentage increase and in which Wilcoxon Rank-Sum Test that phase was used.

(A) Bitcoin

Phase	Dates	Price Movement	Increase
1	1 January to before 16 December 2017	Rise	1,854%
2	16 December 2017 to before 5 February 2018	Fall	-65%
3	5 February 2018 to before 5 March 2018	Rise	70%
4	5 March to before 6 April 2018	Fall	-43%
5	6 April to before 5 May 2018	Rise	48%
6	5 May to before 29 June 2018	Fall	-40%
7	29 June 2018 to before 15 November 2018	Sideways	-5%
8	15 November 2018 to before 15 December 2018	Fall	-43%

(B) Ether

1	1 January to before 12 June 2017	Rise	4,748%
2	12 June to before 23 November 2017	Sideways	3%
3	23 November 2017 to before 13 January 2018	Rise	241%
4	13 January to before 6 April 2018	Fall	-73%
5	6 April to before 5 May 2018	Rise	120%
6	5 May to before 14 December 2018	Fall	-90%

(C) Phases compared in the Wilcoxon Rank-Sum Tests

Cryptocurrency	Analysis Type	Rising Price Dataset	Falling Price Dataset
Bitcoin	mono-phase	1	2
Bitcoin	multi-phase	1,3,5	2,4,6,8
Ethereum	mono-phase	3	4
Ethereum	multi-phase	1,3,5	4,6

Table 7.8: Word2vec-based topic modelling specifications. In the multi-phase analysis, words more frequent as prices rose were split from those more popular as prices fell before finding topics (see Section 7.4.2.2); the 'Higher Word Frequency Dataset' indicates in which dataset word frequency was higher. As in Chapter 6, in applying word2vec, the Continuous Bag-of-Words model (CBOW) was compared against the continuous Skip-gram (SG) model, and Hierarchical Softmax (HS) was assessed against Negative Sampling (NEG). The optimal model ('Model') and threshold ('Threshold') is provided.

MONO-PHASE ANALYSIS

Cryptoasset	Model	Threshold
Bitcoin	SG, HS	99.0
Ethereum	CBOW, HS	99.0

MULTI-PHASE ANALYSIS

Cryptoasset	Higher Word Frequency Dataset	Model	Threshold
Bitcoin	Falling Price	SG, HS	99.0
Bitcoin	Rising Price	CBOW, NEG	99.9
Ethereum	Falling Price	CBOW, NEG	99.0
Ethereum	Rising Price	CBOW, HS	99.9

Table 7.9: Descriptive statistics for phases in the bitcoin and ether price series: the number of days and submissions.

Bitcoin

Phase	Days	Submissions
Rise	406	204344
Fall	168	86290
1	349	180898
2	51	48048
3	28	13290
4	32	12302
5	29	10156
6	55	17213
7	139	38700
8	30	8727

Ether

Phase	Days	Submissions
Rise	242	61010
Fall	306	68034
1	162	30328
2	164	54372
3	51	24037
4	83	27552
5	29	6645
6	223	40482

7.5.2 Mono-phase concepts and their context

Ether prices rose 241% (phase 3) to an all time high price on 13 January 2018 before falling 73% (phase 4). Only ‘feb’ met the criteria for a mono-phase concept and was excluded as it reflected the timing of phase 4.

Bitcoin prices rose 1854% to an all time high price on 16 December 2017 during phase 1 and then fell 65% (phase 2). Ten mono-phase concepts rose more than three-fold with this shift to falling prices and increased within the 24 hour period before entering the falling price phase (see Figure 7.3). The words occurring with these concepts (see Table 7.10) suggested three themes: regulatory bans (‘korea’ and ‘minist’/‘ministri’); concerns over whether to sell bitcoin or switch to an altcoin (‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’; ‘airdrop’; ‘binanc’/‘hitbtc’; ‘hashflar’; and ‘discord’); and discussion of the practicalities of transacting bitcoin (‘batch’, ‘bech32’ and ‘changelli’). Two further concepts (‘merri’ and ‘christma’/‘holiday’/‘xmas’) also met the mono-phase criteria but were excluded because these were most likely due to the timing of phase 2, which began on 16 December 2017.

The context of the altcoin group (‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’) reflected the contexts of each cryptoasset named. Three of these six cryptoassets increased more than three-fold in the proportion of submissions from phase 1 to 2: Cardano rose 721.44%; Tron 562.63%; and Ripple (represented by ‘rippl’) 309.36%. Examining the top five words occurring with each of Cardano, Tron and Ripple and the altcoin group (‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’) revealed, in each case, that they were discussed with: ‘ethereum’, ‘buy’, price (‘price’ or US Dollars) and another cryptoasset (‘bitcoincash’ or ‘rippl’ and ‘verg’ in the case of Tron). Further details in Table 7.11.

The concept ‘binanc’/‘hitbtc’ combines two different cryptoasset exchanges: Binance and HitBTC. Interest in Binance rose 1327.89% in frequency compared with only 163.55% for HitBTC. The context in which ‘binanc’ was used was similar to the concept ‘binanc’/‘hitbtc’, with the top ten words being shared and the top three words having the same ranking (‘coinbas’, US Dollar mentions and send).

Further details in Table 7.12.

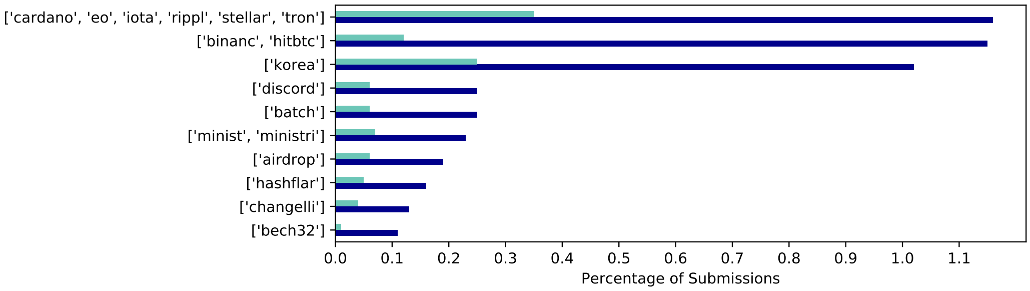


Figure 7.3: Frequency data for mono-phase concepts in the case of Bitcoin. This shows the percentage of all submissions containing the concept in phase 1 (light green) and phase 2 (blue).

Table 7.10: Top five words occurring with each Bitcoin mono-phase concept in phase 2. ‘Frequency’ is the percentage of submissions containing each word, providing the context of that concept. Concepts given in bold and grouped into themes (in capitals). ‘DMS’ is an abbreviation for ‘dollarmarkersymbol’, used to represent mentions of US Dollars.

REGULATORY BAN

‘korea’		**‘minist’/‘ministri’**	
Word	Frequency	Word	Frequency
ban	26.07	financ	60.00
trade	23.22	ban	32.73
regul	14.26	korea	29.09
market	13.85	trade	27.27
govern	12.83	india	23.64

SELL OR SWITCH TO ALTCOIN

‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’		**‘airdrop’**		**‘binanc’/‘hitbtc’**	
Word	Frequency	Word	Frequency	Word	Frequency
ethereum	15.65	free	30.11	coinbas	17.00
buy	14.21	token	20.43	DMS	15.73
DMS	13.13	coin	16.13	send	15.37
coin	11.69	new	13.98	transact	14.83
bitcoincash	8.63	fork	11.83	fee / transfer	14.47

‘hashflar’		**‘discord’**	
Word	Frequency	Word	Frequency
mine	55.70	join	24.79
cloud	29.11	pump	20.66
DMS	27.85	server / member	14.88
profit	11.39	pumpanddump	14.05
buy / sell	10.13	new	11.57

TRANSACTION PRACTICALITIES

‘batch’		**‘bech32’**		**‘changelli’**	
Word	Frequency	Word	Frequency	Word	Frequency
transact	65.00	segwit	69.23	transact	42.19
segwit	55.83	wallet	65.38	send	32.81
coinbas	44.17	support	48.08	DMS	28.12
fee	40.00	send/transact	40.38	help	20.31
implement	27.50	electrum	36.54	support	18.75

Table 7.11: Top five words occurring with each of Cardano, Tron and Ripple ('rippl') compared with the Bitcoin mono-phase concept 'cardano'/'eo'/'iota'/'rippl'/'stellar'/'tron' in phase 2 of the bitcoin price series. 'Frequency' is the percentage of submissions containing each word, providing the context of the specific altcoin or the group of altcoins. 'DMS' is an abbreviation for 'dollarmarkersymbol', used to represent mentions of US Dollars.

'cardano'/'eo'/'iota'/'rippl'/'stellar'/'tron'		**'cardano'**	
Word	Frequency	Word	Frequency
ethereum	15.65	rippl	45.83
buy	14.21	price/bitcoincash/ethereum/litecoin	37.50
DMS	13.13	buy	33.33
coin	11.69	analysi/nem	29.17
bitcoincash	8.63	wallet	25.00

'tron'		**'rippl'**	
Word	Frequency	Word	Frequency
coin	25.00	ethereum	17.43
DMS/buy	13.64	buy	14.22
ethereum	11.36	DMS	12.39
fee/binanc/help/rippl	9.09	bitcoincash	10.78
bring/verg/bite/futur/invest/new/week	6.82	coin	10.32

Table 7.12: Top ten words occurring with Binance ('binanc') compared with the Bitcoin mono-phase concept 'binanc'/'hitbtc' in phase 2 of the bitcoin price series. 'Frequency' is the percentage of submissions containing each word, providing the context of the word 'binanc' or concept 'binanc'/'hitbtc.' 'DMS' is an abbreviation for 'dollarmarkersymbol', used to represent mentions of US Dollars.

'binanc'/'hitbtc'		**'binanc'**	
Word	Frequency	Word	Frequency
coinbas	17.00	coinbas	17.58
DMS	15.73	DMS	16.21
send	15.37	send	15.62
transact	14.83	transfer	15.04
fee/transfer	14.47	buy	14.65
buy	14.29	transact	14.45
new	13.56	fee/new	13.67
trade	12.48	trade	12.30
help	12.12	wallet	12.11
wallet	11.93	help	11.52

7.5.3 Multi-phase concepts and their context

With Bitcoin, two multi-phase concepts were linked to falling prices: ‘market’ and ‘sale’. The top two words occurring with ‘market’ were ‘price’ and US Dollars across each phase of falling prices. The concept ‘sale’ was discussed in a varying context in different phases of falling prices: with ‘buy[ing]’ and ‘sell[ing]’ in phases 2 and 6, ‘token’ sales in phases 4 and 6 and ‘black’ ‘friday’ sales in phase 8 (see Table 7.13).

With Ethereum, ten multi-phase concepts were identified. Three of these were associated with rising prices: ‘tax’, US Dollars and ‘hit’. ‘Hit’ was discussed with US Dollars (over 40% submissions in each phase of rising prices) and US Dollars were frequently discussed with ‘bitcoin’(over 15%). The concept ‘tax’ was considered with ‘gain’ (over 30% submissions in each phase of rising prices); ‘pay’ (over 25%); US Dollars (over 24%) and ‘trade’ (over 23%). Further details in Table 7.14.

The remaining seven multi-phase concepts related to falling ether prices. With the exception of ‘game’, all these could be split into two themes: price (‘market’ and ‘bear’/‘bearish’/‘bull’) and innovation (‘featur’; ‘ceo’/‘cofound’; ‘project’/‘team’; and ‘makerdao’/‘stablecoin’). In each phase of falling prices, ‘bear’/‘bearish’/‘bull’ was discussed with ‘market’ (over 45% submissions) and ‘market’ was discussed with US Dollars (over 20%) and ‘price’ (over 18%). Price was discussed in the context of ‘bitcoin’, which was in over 16% ‘market’ submissions. The context of discussions around innovation varied but referred to new ‘token[s]’ in over 10% submissions across all concepts and across all phases of falling prices. The concept ‘game’ was discussed in the context of using gaming machines to mine ether in phase 4 (24.39% submissions) and ‘play[ing]’ games in phase 6 (14.62% submissions). Further details in Table 7.15.

After the tables showing the context of the multi-phase concepts, Tables 7.16 and 7.17 provide further detail on the percentage change in popularity for Bitcoin multi-phase concepts and Ethereum multi-phase concepts respectively.

Table 7.13: Top five words occurring with each Bitcoin falling-price, multi-phase concept in phases 2, 4, 6 and 8. Concepts given in bold. 'Frequency' is the percentage of submissions containing each word, providing the context of that concept. 'DMS' is an abbreviation for 'dollarmarkersymbol', used to represent mentions of US Dollars.

'market'

Phase 2		**Phase 4**		**Phase 6**		**Phase 8**	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
price	23.71	price	21.67	DMS	23.48	DMS	27.80
DMS	21.51	DMS	16.48	price	21.29	price	25.56
buy	20.23	buy	15.37	buy	17.15	bear	20.48
trade	16.58	sell	12.96	new	13.14	buy	19.28
new	16.33	new	10.56	trade	12.90	sell	13.30

'sale'

Phase 2		**Phase 4**		**Phase 6**		**Phase 8**	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
buy	26.30	token	20.45	buy	22.22	buy	30.77
sell	19.45	sell	19.32	DMS	19.26	price	28.21
DMS	17.53	DMS/price/market	17.05	sell	17.78	DMS	25.64
price	13.70	buy	15.91	token	17.04	friday	20.51
new	11.78	mt / gox	14.77	busi	14.81	black / market	19.23

Table 7.14: Top five words occurring with each Ethereum rising-price, multi-phase concept in phases 1, 3 and 5. Concepts given in bold. 'Frequency' is the percentage of submissions containing each word, providing the context of that concept. 'DMS' is an abbreviation for 'dollarmarkersymbol', used to represent mentions of US Dollars.

'hit'

Phase 1		**Phase 3**		**Phase 5**	
Word	Frequency	Word	Frequency	Word	Frequency
DMS	46.82	DMS	46.30	DMS	41.27
price	27.95	bitcoin	22.22	bitcoin/check/mean/price/wallet	17.46
buy	25.23	new	20.63	buy/help/never/hold/activ/transact	15.87
new	24.09	high	16.67	new/best/list/move/bite/secur	14.29
bitcoin	23.64	mine	16.14	mine/rate/worth/fund	12.70

US Dollar mentions

Phase 1		**Phase 3**		**Phase 5**	
Word	Frequency	Word	Frequency	Word	Frequency
buy	28.70	buy	22.47	price	17.97
price	24.85	price	19.72	bitcoin	15.67
bitcoin	21.13	bitcoin	16.18	token	12.67
invest	15.48	new	14.34	market	12.44
sell	14.00	mine	11.53	buy	11.75

'tax'

Phase 1		**Phase 3**		**Phase 5**	
Word	Frequency	Word	Frequency	Word	Frequency
buy	32.27	gain	31.14	gain	35.71
gain	31.47	DMS	27.19	trade	31.43
pay	30.68	pay	26.75	pay	28.57
DMS	25.50	buy	25.44	DMS	24.29
trade	23.90	trade	24.56	capit	22.86

Table 7.15: Top five words occurring with each Ethereum falling-price, multi-phase concept in phases 4 and 6. Concepts given in bold and grouped into themes (in capitals). ‘Frequency’ is the percentage of submissions containing each word, providing the context of that concept. ‘DMS’ is an abbreviation for ‘dollarmarkersymbol’, used to represent mentions of US Dollars.

PRICE

‘market’				**‘bear’/‘bearish’/‘bull’**			
Phase 4		**Phase 6**		**Phase 4**		**Phase 6**	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
DMS	22.22	DMS	21.43	market	45.57	market	49.69
price	21.53	price	18.50	bitcoin	32.07	DMS	24.64
buy	17.21	bitcoin	17.60	DMS	23.63	price	22.59
bitcoin	16.22	trade	16.81	price	23.21	bitcoin	21.97
new	15.44	new	16.36	buy	21.52	new	14.37

INNOVATION

‘project’/‘team’				**‘featur’**			
Phase 4		**Phase 6**		**Phase 4**		**Phase 6**	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
token	22.92	token	20.48	new	36.73	new	31.69
new	20.58	develop	19.16	help	26.12	platform	26.11
ico	17.96	new	17.79	token / develop	22.45	token	24.52
develop	17.68	ico	16.27	build	21.22	project	21.02
market	14.78	platform	16.00	check / price	20.82	develop	20.06

‘ceo’/‘cofound’				**‘makerdao’/‘stablecoin’**			
Phase 4		**Phase 6**		**Phase 4**		**Phase 6**	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
interview	16.42	DMS	12.55	DMS / token	24.19	DMS	19.45
token	12.77	platform	12.29	stabl	19.35	token	14.71
project	11.31	token	11.37	price	16.94	new	11.22
ico	10.58	new	11.24	maker / project	12.90	price	10.97
develop	10.22	project	11.11	decentr / market	12.10	coin	10.72

POLYSEMIC

‘game’			
Phase 4		**Phase 6**	
Word	Frequency	Word	Frequency
mine	25.39	new	21.11
new	24.08	play	14.62
card	19.69	token	12.31
gpu	16.73	launch	12.22
buy	16.37	buy	11.62

Table 7.16: For each Bitcoin multi-phase concept, the percentage change in frequency with each shift to falling prices. Frequency was measured as proportion of submissions containing that concept. Both concepts identified were associated with falling prices.

Concept	**Phase 1 to 2**	**Phase 3 to 4**	**Phase 5 to 6**	**Phase 7 to 8**
sale	42.55	25.09	15.44	32.02
market	12.09	22.81	10.73	30.29

Table 7.17: For each Ethereum multi-phase concept, the percentage change in frequency with each shift to rising (upper section) or falling (lower section) prices. The first three concepts were associated with rising prices, and the next seven concepts were associated with falling prices. Frequency was measured as proportion of submissions containing at least one word from that concept.

Concept	**Phase 2 to 3**	**Phase 3 to 4**	**Phase 4 to 5**	**Phase 5 to 6**
tax	61.67		21.95	
hit	37.91		11.63	
dollarmarkersymbol	12.53		0.36	
makerdao, stablecoin		209.09		4.48
bear, bearish, bull		75.22		35.49
ceo, cofound		63.73		25.57
market		26.21		32.04
project, team		24.52		23.01
game		23.61		23.11
featur		21.45		35.64

7.6 Coherence with known facts

Of the Bitcoin mono-phase themes (see Table 7.10), regulatory bans are the closest to capturing a specific external event. Discussion of ‘korea’ and ‘minist’/‘ministri’ occurred with the debate between the Ministry of Finance and Justice in South Korea as to whether a ban on cryptoasset trading activity should be implemented, with one proposal being that cryptoassets are a scam that should be subject to criminal charges [149]. On 16 December 2017, when prices changed to falling, South Korean news media reported how North Korea was using hacks of South Korean exchanges to fund its regime, encouraging South Korean support for a ban [130]. This could have triggered South Koreans to sell bitcoin holdings before this became illegal and possibly even criminal [149]. Since approximately a fifth of bitcoin transactions were in South Korean Won at the time [149], it is coherent with known events that this caused the shift from rising to falling prices. The presence of ‘india’ in 23.64% ‘minist’/‘ministri’ submissions may reflect concerns over bitcoin regulation, including rumours of a possible ban in India during phase 2 [185].

The remaining Bitcoin mono-phase concepts could be reflections of a change in mind-set among bitcoin-holders prior to selling. Before selling, holders of bitcoin are likely to become concerned as to the future of bitcoin (theme ‘Sell or Switch to Altcoin’ in Table 7.10) and to consider how to transact the bitcoin held (theme ‘Transaction Practicalities’). Concerned holders of bitcoin may consider: rival cryptoassets (‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’ and ‘airdrop’); Binance, an exchange selling more than 150 cryptoassets [21]; and whether to stop reinvesting mining ‘profit[s]’ from Hashflare (‘hashflar’) to generate more bitcoin [237]. Other bitcoin-holders may dismiss concerns raised on social media platforms (‘discord’) as price manipulation (‘pumpanddump’). Before selling bitcoin, holders may consider the practicalities of: reducing ‘fee[s]’ through batching transactions (‘batch’) [129]; seeking ‘support’ on exchanges (‘changelli’); and determining whether transferring bitcoin from a ‘bech32’ address is ‘support[ed]’ [258].

All the concepts delineated for Ethereum were multi-phase, having a recurring impact on price over time. Innovation (‘project’/‘team’, ‘featur’, ‘ceo’/‘cofound’ and ‘makerdao’/‘stablecoin’) was associated with falling prices (Table 7.15). This suggests that ether holders disposing of their ether to capitalise on new ‘token[s]’ from new cryptoassets was a cause of price falls. This included ‘project[s]’ or ‘team[s]’ ‘develop[ing]’ ($\geq$ 17.68% submissions) ‘new’ ($\geq$ 17.79%) ‘token[s]’ ($\geq$ 20.48%) through ICOs (‘ico’; $\geq$ 16.27%). Mentioned in relation to this was ‘ceo’/‘cofound’ (‘project’ $\geq$ 11.11% submissions) and ‘featur’ (‘project’ $\geq$ 15.51% submissions). A separate innovation theme related to interest in MakerDAO, which was launched in December 2017 enabling holders to exchange their ether for Dai, a decentralised ‘stablecoin’ designed to maintain its value in US Dollars [193].

For Ethereum, price discussed in the context of ‘hit’ was supported as causing prices to rise whilst ‘market’ price and sentiment (‘bear’/‘bearish’/‘bull’) discourse were associated with price falls (see Tables 7.14 and 7.15). These discussions happened in the context of ‘bitcoin’ which was a top five co-occurring word throughout. This suggests a source of ether price volatility was traders analysing the ether price and comparing it with bitcoin before buying or selling ether.

The multi-phase concept ‘market’ was identified as a consistent driver for both falling bitcoin prices and falling ether prices. This was discussed in the context of price as well as buying, trading and selling (see Tables 7.13 and 7.15). This supports the widespread influence of technical traders who use just price information to make trading decisions on cryptoasset price series and is consistent with evidence for price barriers at 400 US Dollars for ether and 6000 US Dollars for bitcoin (see Figure 1.1).

Including contextual analysis in the framework has shown that some multi-phase concepts were polysemic - being used in a different context in different price phases. In some cases, this could be because the concept is an artefact of distinct themes of discussion each happening to include the polysemic concept. For instance, in the case of Ethereum, ‘game’ was used in the context of using ‘gam[ing]’ machines to mine ether in phase 4 (‘mine’, ‘card’, ‘gpu’) and ‘play[ing]’ ‘game[s]’ in phase 6 (see Table 7.15). Both include the word ‘game’ but are otherwise distinct issues and so examining the context reveals that ‘game’ is probably a spurious result.

In contrast, with Bitcoin, the polysemic concept ‘sale’ became popular in all four phases of falling prices making coincidence less plausible (see Table 7.13). The concept ‘sale’ was mentioned in terms of ‘buy[ing]’ and ‘sell[ing]’ in phases 2 and 6, a ‘token’ sale in phases 4 and 6 and ‘black’ ‘friday’ sales in phase 8. For ‘sale’ to be irrelevant to price, distinct, irrelevant themes including ‘sale’ would have to arise at the correct time across four different phases (falling price phases 2, 4, 6 and 8) and within 24 hours before each phase to meet the multi-phase concept criteria. A tenable explanation is that ‘sale’ is a general term that captures concern regarding bitcoin before decisions to sell. If holders are concerned about bitcoin, they could be more sensitive to any ‘sale’ of bitcoin (phases 2 and 6); more interested in ‘token’ ‘sale[s]’ to exchange bitcoin for other tokens (phases 4 and 6); and more tempted by ‘black’ ‘friday’ ‘sale[s]’ where bitcoins are exchanged for discounted products or sold to generate cash to buy such products (phase 8). This suggests the concept ‘sale’ may have value as a negative sentiment indicator that warns of future falls in price.

The association of ‘tax’ with rising ether prices could be explained by the timing of phases 3 and 5, which coincided with the end of tax years when ‘pay[ment]’ of ‘capit[al]’ ‘gain[s]’ ‘tax’ becomes due (see Table 7.14). The end of the tax year in some countries, such as the USA [155], is on 31 December (phase 3 is from 23 November 2017 to 13 January 2018) but in the UK on 5 April (phase 5 was from 6 April to 5 May 2018) [109]. Tax returns are also due in the US by April [155].

7.7 Discussion

The developed framework is designed to capture two distinct types of potential cause of shifts in the cryptoasset price series: the ‘mono-phase’ with a one-off, strong impact and the ‘multi-phase’ that repeatedly causes shifts. Application to Bitcoin and Ethereum data supports both of these phenomena occurring.

The results suggest a one-off effect of regulatory bans on bitcoin, a repeated effect of rival innovations on ether and the influence of technical traders, captured through market price discourse, on both cryptoassets. Traders seem to be comparing the prices of different cryptoassets: the Ethereum multi-phase concepts discussed with price commonly referred to ‘bitcoin’, and the Bitcoin mono-phase concept covering altcoins (‘cardano’/‘eo’/‘iota’/‘rippl’/‘stellar’/‘tron’) was discussed with US Dollars.

The difference in results between bitcoin and ether is consistent with the difference in the timing of the price phases (Table 7.7) and the all time high price (see Figure 1.1 in Chapter 1). These cryptoassets were also of different token functionality (Chapter 4).

The concepts delineated by multi-phase analysis may have implications for forecast models, since these concepts have a predictive association with price that persists across time. Multi-phase concepts may provide an improvement on sentiment metrics such as VADER that have found social media posts to be positive even during falling prices [2]. This extends to polysemic concepts, if their context supports such concepts as acting as proxies for positive or, in the case of ‘sale’, negative sentiment. The concept ‘market’ was supported as a consistent driver of falling prices for both bitcoin and ether. The other multi-phase concepts differed, suggesting that different predictors may be suitable for different cryptoassets.

The use of forecast models would have to account for the possible presence of mono-phase concepts that have a one-off, major effect on price. A plausible example of such an event would be the one-off effect of rumours of regulatory bans in South Korea on the bitcoin price. These could be considered analogous to ‘black swan’ [274] events, being unexpected and having a major impact, but they can be rationalised with the benefit of hindsight. The rarity of such events means there are limited data available for understanding how they affect price and so for informing how the forecast model could adapt to their presence. The magnitude and unpredictability of these effects make them difficult to model and they may invalidate results derived from predictions based on the multi-phase concepts.

7.7.1 Limitations of Causality Analysis

The quantitative analytic methodologies developed rely on observational rather than experimental data. As presented here, observational data may support a causal link as plausible, particularly relative to alternative explanations [74, 251], but such data cannot prove a causal relationship [223, 251]. Let us suppose that a statistical association between event X and event Y has been established, and that X is being tested to see if it causes Y. Three reasons have been identified for why X might not be the cause of Y: X is a response to Y, X and Y are symptoms and X causing Y requires a catalyst.

7.7.2 X is a Response to Y

This was an issue when considering the DDPWI and word2vec-based topic modelling results. For instance, it is plausible that the declining discussion of US Dollars, mentioned with buying bitcoin, was a response to prices falling rather than a cause (see Section 5.5). This is addressed by placing X before Y in time, which was incorporated into the mono-phase and multi-phase analyses. These methodologies required that the popularity of a concept rose before the phasic shift in price.

7.7.3 X and Y are Symptoms

Event X may have occurred before Y because of a third event that caused X and then Y to occur [117, 223, 303].

Mono-phase analysis reduces this risk through examining the most significant phasic shift, where prices moved from rising to the all time high value to falling, and considering just concepts where frequency was more than three-fold higher after the shift in phase. This means that for a mono-phase result to be spurious the unmeasured variable would have to have a strong association with both the concept popularity and the price phase. The time of year could be such a variable. This could explain why ‘christma’/‘holiday’/‘xmas’ were extracted in the Bitcoin analysis, where the latter phase’s time-span included Christmas, and why the Ethereum analysis produced ‘feb’, because February was in the latter phase compared.

Multi-phase analysis reduces this risk through requiring that the association between the use of a concept and a move to rising or falling prices persists across different phasic shifts. Hence, a multi-phase result that a concept is associated with rising prices is compromised if there exists an unmeasured variable that happened to occur with each phase of rising prices. For instance, the association between ‘tax’ and phases of rising prices may be because these phases coincided with key dates in the tax calendar.

In both the mono-phase and multi-phase analysis, the use of concepts helped to reduce this risk. Cases where examining the word or words within the concept contributed to determining such a risk as likely have been identified (‘christma’/‘holiday’/‘xmas’, ‘feb’ and ‘tax’). Examining contextual words that occurred in submissions that contained the word or words in a concept provided further insight. The Bitcoin mono-phase analysis extracted the concepts ‘binanc’/‘hitbtc’ and ‘changelli’. These were discussed with ‘send’, ‘transact’ and US Dollar references (see Table 7.10). Hence, the discussion of exchanges may have reflected a desire to dispose of bitcoin rather than caused this need. This suggests that exchange discussions may not have been the root cause of falling prices but instead some other event drove the need to sell bitcoin and this was the true root cause of bitcoin prices falling.

7.7.4 X Causing Y Requires a Catalyst

Another risk is that *X* causing *Y* may depend on the presence of a catalyst – there needs to be other events that also occurred [252]. The detected causal relationship between *X* and *Y* may then fail to persist if the catalyst becomes absent. Despite being described as a common issue in epidemiology [252], this is not typically accounted for in some causal inference approaches such as DGCMs [150].

This is a risk in the mono-phase analysis that compares just two phases in time. Rumours of South Korean regulatory bans may have impacted the bitcoin price because of a context specific to the end of 2017 and beginning of 2018. Bitcoin-holders may have been particularly sensitive to rumours of a ban at this time because cryptoasset exchanges had just been banned in China (September – November 2017) [239]. Rumours of a ban in South Korea may have fuelled speculation that other countries would follow China’s lead. This may have caused panic selling as holders sold their bitcoin before being unable to trade bitcoin, which could have caused the period of falling prices.

This is less of a risk with multi-phase analysis because this requires the association between proposed cause and effect to persist across multiple phases. Hence, any catalyst required for the multi-phase concept to affect price would have to have been present across the various phases analysed. Even if there were a catalyst meeting this criterion, this suggests that the catalyst persisted across time, reducing the risk of the catalyst becoming absent with future data.

Chapter 8

Quantamental Analysis of Bitcoin and Ethereum

This chapter performs a quantamental analysis of the bitcoin and ether price that matches the results of the quantitative analyses with the fundamentals previously identified for Bitcoin and Ethereum in Chapter 4. Section 8.1 compares the results for the different quantitative analyses to determine what events and concerns these supported as being the causes of shifts in the bitcoin and ether price phases. Section 8.2 then applies the quantamental analyses. This is to address research question 9 in Section 1.3.5.

8.1 Comparison of Quantitative Analysis Results

This section compares the results of the quantitative analyses applied to Bitcoin, and then examines the extent Ethereum results differed.

8.1.1 Bans

The quantitative analyses results suggest that concerns over regulation banning cryptoasset trading in South Korea had a one-off effect which may have been responsible for bringing to a close the phase of rising prices across 2017.

DDPWI established that the word ‘ban’ was particularly frequent during the stage of falling, volatile bitcoin prices and that ‘ban’ was discussed in the context of South Korea and ‘trade’. A negative impact on price is also consistent with the persistently negative sentiment associated with submissions containing ‘ban’ (see

Section 5.4.4).

When considering causality, the mono-phase analysis corroborated DDPWI, returning ‘korea’ and the Ministry of Finance (‘financ’ in 60% ‘minist’/‘ministri’ submissions in Table 7.10) as plausible causes of the shift to falling prices at the end of 2017, concepts that were discussed with the word ‘ban’ ($\geq$ 26.07 submissions) and ‘trade’ ($\geq$ 23.22 submissions) (see Table 7.10). The multi-phase analyses did not return any results related to regulation or South Korea, suggesting that this influence of ban concerns on price was not a recurring effect.

In the DDPWI results, during the stage of falling bitcoin prices, ‘ban’ was also discussed with adverts and internet companies (Facebook, Google and Twitter). This may have captured discussion around internet companies banning cryptoasset adverts (see Section 5.5). When examining causality, neither the mono-phase nor multi-phase analyses supported such policies as influencing price. Internet company bans on cryptoasset adverts were unlikely to have caused the shift to falling prices because the bans were implemented after the shift took place. The first ban was announced by Facebook in January 2018 [88].

8.1.2 US Dollars

DDPWI found that, as prices fell, there was less discussion of US Dollars. US Dollars was discussed with ‘buy’ across stages when prices were rising or falling but, when prices stabilised, ‘buy’ was no longer captured by contextual analysis (see Section 5.4.4). Prices may have stabilised because fewer speculators were ‘buy[ing]’ bitcoin.

The multi-phase analysis, that considered causality, established that ‘market’ (discussed with ‘price’ and US Dollars) rose in the 24 hours before and with every shift to falling prices. This is consistent with holders of bitcoin discussing the bitcoin ‘market’ and concluding that they should sell. These results convey the difficulty in interpreting US Dollars because, combined with other words such as ‘buy’ and ‘market’, this had different implications for the price.

8.1.3 Tax

Although DDPWI found 'tax' to have been discussed more as prices were falling, neither the mono-phase nor multi-phase analyses supported tax as causing phasic shifts in price. Concerns over tax causing prices to fall is also inconsistent with the persistently positive sentiment of submissions containing the word 'tax' (see Section 5.4.4). DDPWI may have extracted 'tax' because the timing of the stage of falling prices included April, when tax returns are due in the US [155] and this is the end of the tax year in the UK [109]. This would explain why interest in 'tax' rose up to April 2018 and then declined (see Figure 5.7). Hence, concerns over tax were likely to have reflected the time of year.

8.1.4 Is Ethereum Different?

None of the Ethereum results suggested that government regulation had a direct influence on the ether price, whilst innovation was found to have a repeated effect on the ether price but not on the bitcoin price.

Government regulation may, however, have had an indirect influence on the ether price. There is evidence to suggest that the ether and bitcoin price were being compared (see Tables 7.14 and 7.15). Also, the ether price peaked on 13 January 2018, one month after bitcoin (16 December 2017). It is plausible that the holders of ether, having seen bitcoin prices fall across a month, started to question the valuation of ether, triggering them to sell, reducing the ether price.

Both the Ethereum and Bitcoin results suggest the recurring influence of technical trading, with historic price information used to determine whether to buy or sell the cryptoasset. Multi-phase analyses identified 'market' (discussed with US Dollars and 'price') as a recurring cause of both falling ether and bitcoin prices. Price levels could also be identified that acted as price barriers for both Bitcoin and Ethereum (discussed in Section 7.5.1).

The multi-phase analyses identified further price-related concepts for Ethereum. This included discussion of market sentiment ('bear'/'bearish'/'bull'), associated with falling ether prices, and price-related concepts associated with rising prices (US Dollars and 'hit'). These concepts were used with the word 'bitcoin', and so they may capture observations regarding the bitcoin price influencing ether buy or sell decisions.

8.2 Quantamental Analysis

8.2.1 Bitcoin as Money

Concerns over regulatory bans on bitcoin trading causing a shift to falling prices is consistent with the fundamentals identified in Section 4.6.2 as underpinning the bitcoin price.

Bitcoin, as a crypto-transaction system, is primarily designed for transacting value. As Bitcoin provides a form of money, the functions of money should be relevant to understanding why the price of bitcoin has changed. These functions are to provide: a store of value, a unit of account and a medium of exchange. Price may be affected by not just an alteration to the actual, present functionality of Bitcoin as a form of money, but also by any changes in the expected future functionality of Bitcoin as a form of money (as explained in Section 4.6.2).

In the extreme case, if a ban on the trading of bitcoin in a country were entirely effective, the holders of bitcoin in that country would become unable to sell their bitcoin. This would mean the value of the bitcoin they hold would be, in practice, worthless, removing the functionality of Bitcoin as a store of value.

A complete ban on all bitcoin trading is unlikely to be achievable. Holders may transport their bitcoin abroad and sell in foreign exchanges, a practice facilitated by the digital nature of bitcoin [211]. Alternatively, holders may find, perhaps illegal, channels by which bitcoin could be sold to other buyers, such as transacting bitcoin from wallet to wallet after meeting the other party in person. These provide possible means of circumventing government regulation, but they are less practical and involve greater risk than being able to sell bitcoin in a local exchange. Hence, concerns over government bans on bitcoin trading are likely to lead to concerns over whether bitcoin will remain a viable store of value.

Such concerns over the future viability of bitcoin as a store of value may impact on the current value of bitcoin as a medium of exchange. A merchant would need to invest in new infrastructure to begin accepting bitcoin in exchange for goods and services. If the bitcoin received can then not be exchanged for national currency, and so is essentially valueless, this infrastructure investment would have generated a loss. Concerns over bitcoin trading bans may thus cause merchants to hesitate over adapting to accept bitcoin for goods and services, reducing the current viability of bitcoin as a medium of exchange.

Bans on bitcoin trading in different jurisdictions may also damage the reputation of Bitcoin. A ban on bitcoin trading suggests that the government views the trading of bitcoin as inappropriate for society. This could fuel a negative opinion of bitcoin among potential participants, which may discourage the use of bitcoin. With fewer users, bitcoin would be a less valuable medium of exchange.

Hence, concerns over government regulatory bans could have caused concerns over the viability of Bitcoin as a store of value and medium of exchange, which would have reduced the case for holding bitcoin. This may have reduced the demand for bitcoin and caused more holders to sell, causing the shift to a phase of falling prices. This suggests that this result from quantitative analysis of concerns over regulatory bans causing a shift to falling prices is plausible from the perspective of the fundamentals identified.

8.2.2 Ethereum for Application Development

As discussed in Section 4.6.2, Ethereum, a crypto-fuel system, was primarily designed to enable the development of blockchain-supported applications. Ether can be acquired to benefit from the suitability of Ethereum in developing applications and launching ICOs. Bitcoin, designed as a crypto-transaction system, does not similarly support application development.

If a new innovation occurred that offered improvements in application development or in raising funds through an ICO, examining the fundamentals suggests that developers are more likely to switch from ether to this new, rival technology than from bitcoin. Hence, the price of ether is more likely to be influenced by concerns over new innovations than bitcoin. This supports the plausibility of the quantitative analysis results that suggest that innovation influenced the ether price.

8.3 Conclusion

Overall, this chapter has demonstrated the value of a quantamental analysis. In the case of Bitcoin, the fundamentals characterised it as a form of money and thus it was vulnerable to regulatory bans. For Ethereum, the fundamentals characterised it as a platform for developing blockchain applications and thus it was more vulnerable to new, competitive technologies.

Chapter 9

Conclusion

9.1 Addressing the Research Questions

The recent trend has been to move from describing Bitcoin as a cryptocurrency to making it part of a wider universe of ‘cryptoassets’ [48]. This reflects the fact that tokens such as ether offer more than cryptocurrency [48] and tokens such as bitcoin are perceived as being too volatile to be a viable currency [52]. This raises the question that if they are assets then there should be ‘fundamentals’ (see Section 1.1) underlying their value. A quantamental analysis was performed that identified what these fundamentals might be and then examined whether these were consistent with with the results of quantitative analyses of social media data. This research was conducted by considering a series of questions (see Section 1.3), which are addressed here.

1. **Should cryptoasset price series be analysed individually or in aggregate?**

It was found that cryptoassets are a heterogenous universe of assets that should be analysed individually. This was shown by the lack of any very strong correlations between the prices of different cryptoassets (see Chapter 3) and by three distinct types of token functionality being identifiable (see Chapter 4). Hence, cryptoassets were analysed individually.

2. **Which cryptoassets are to be analysed?**

Four criteria were applied to decide which cryptoassets to analyse (specified in Section 2.2), namely: consistently in the top ten by market capitalisation and liquidity, entity-independence, sufficiently large, publicly available, social media database and tokens are of a different type. Hence, Bitcoin was selected as the largest cryptoasset with Ethereum chosen as a comparator, being the second largest cryptoasset and having a functionally distinct token from Bitcoin (see Chapter 4). Also, the prices of the bitcoin and ether tokens were only weakly correlated (see Chapter 3).

3. **What social media data are to be used?**

A critical review of the literature was conducted to guide what data would be analysed in performing the quantitative analyses (see Section 2.3). This supported selecting a dataset of discussion forum posts, with Reddit subreddits analysed due to the existence of subreddits dedicated to Bitcoin and Ethereum with large userbases. Measures of the emotional content of posts or the occurrence of generic topics (such as 'China' [166]) were replaced by an examination of the frequency of the use of specific words (e.g. 'ban') and delimited groups of words (e.g. 'cardano'/'eo'/'iota'/'rippl'/'stellar'/'tron'). Examining a word or a group of words and the context in which they were being used facilitated linking their use with specific events or concerns captured by their discussion.

4. **What analytic approach is to be applied?**

The literature review also guided the analytic approach that underpinned the methodologies used throughout the quantitative analyses (Section 2.4). Previous literature decided on a generic measure of social media activity (such as the volume or sentiment of posts) and then tested for an association with, typically, the daily change in price. The current thesis instead used the phases observed in cryptoasset price series to inform the extraction of events and concerns from social media text in a non-parametric approach.

This removed the need to pre-select the possible cause of price variation before testing whether the data were supportive, facilitating finding new, plausible causes of phasic shifts in price that may not have otherwise been considered. For example, the data supported a higher concern over regulatory bans as causing a phasic shift to falling prices, an issue that had not been previously discovered in empirical studies (Kim et al [166] and Phillips and Gorse [228]).

Reviewing the literature from healthcare epidemiology facilitated evolving the methodology from considering what events and concerns were associated with changes in the bitcoin price movement (see Chapters 5 and 6) to what could have caused observed phasic shifts in the bitcoin and ether price (see mono-phase and multi-phase analyses in Chapter 7).

5. **What benefit does a participant receive from holding a cryptoasset token and how might this influence the value of the token?**

Examining whitepapers, official websites, and third-party commentary showed three distinct reasons, other than to profit from speculation, for why a token might be held: to be used as money ('crypto-transaction' tokens); to use a platform for developing blockchain-supported applications ('crypto-fuel'); and to acquire rights to a pre-defined asset ('crypto-voucher') (see Chapter 4). Crypto-vouchers were not analysed further because price was likely to be dominated by changes in the value of the underlying asset.

The benefits from holding different types of token suggested 'fundamentals' that might underpin cryptoasset valuations. More bitcoin may be bought (resulting in higher prices) if the current or expected functionality of Bitcoin as money improves. More ether might be bought with improvements in the suitability of Ethereum in developing blockchain-based applications and launching ICOs.

6. **What words were associated with the phase of volatile but overall falling bitcoin prices 2017-18?**

The first quantitative analyses involved developing a triphasic methodology (DDPWI) that extracted three words that were more frequent ('ban' and 'tax') or less frequent (US Dollars) in the stage of falling prices compared with the phases both before and after (see Chapter 5).

7. **How can we evolve the results from words associated with phases to topics associated with phasic shifts in the bitcoin price?**

A word2vec-based topic modelling methodology was developed that extracted topics, rather than words, that rose or fell with phasic shifts in price (see Chapter 6). The word2vec-based topic modelling methodology was more flexible than DDPWI in being able to compare any two datasets of word frequencies.

8. **How can we evolve the analysis to find potential causes of phasic shifts in the bitcoin and ether price?**

The word2vec-based topic modelling methodology informed the mono-phase and multi-phase analyses that analysed causality (see Chapter 7). The mono-phase analysis used the strength of the association to support causes of a single phasic shift in price. The multi-phase analysis looked for relationships that consistently recurred despite a changing context across time.

9. **How do the results for Bitcoin and Ethereum compare? Are the insights for each cryptoasset shared or unique?**

The quantamental analyses in this thesis characterised Bitcoin and Ethereum as distinct entities with distinct events and concerns that influence price. Considering the fundamentals, Bitcoin presents a form of money while Ethereum provides a platform for developing applications. The quantitative analyses were consistent with these fundamentals (see Chapter 8) and suggested that regulatory ban concerns had a one-off, major negative effect on the bitcoin price while concerns over new innovations had a recurring negative influence on the ether price.

Where Bitcoin and Ethereum were found to be similar was in the evidence found for speculation having an influence on price. Technical traders use observed price data to decide whether to buy or sell, and their effect is suggested by the apparent presence of price barriers at 400 US Dollars for ether and 6000 US Dollars for bitcoin. This is further consistent with multi-phase analyses delineating price discussions as influencing price. The contextual analyses in Chapter 7 suggest that such traders may have been comparing the ether with the bitcoin price.

9.2 Future Work

In future work, the quantamental analytic strategy developed in this thesis could be applied to Bitcoin and Ethereum in a future time period, to other cryptoassets and to prices series in other asset classes to understand what events or concerns influence price across time. The specific methodologies developed in the quantitative analyses (DDPWI, word2vec-based topic modelling, mono-phase analysis and multi-phase analysis) could be applied to other fields of research.

Bibliography

[1] Halvor Aarhus Aalborg, Peter Molnr, and Jon Erik de Vries. What can explain the price, volatility and trading volume of Bitcoin? *Finance Research Letters*, 29:255–265, June 2019.

[2] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review*, 1(3), 2018. `https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview`.

[3] Apoorva Aggarwal, Isha Gupta, Novesh Garg, and Anurag Goel. Deep Learning Approach to Determine the Impact of Socio Economic Factors on Bitcoin Price Prediction. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–5, Noida, India, August 2019. IEEE.

[4] Raj Aggarwal and Brian M. Lucey. Psychological barriers in gold prices? *Review of Financial Economics*, 16(2):217–230, January 2007.

[5] Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, and Andrea Baronchelli. Anticipating Cryptocurrency Prices Using Machine Learning. *Complexity*, 2018:1–16, November 2018.

[6] Joo Almeida, Shravan Tata, Andreas Moser, and Vikko Smit. Bitcoin prediciton using ANN. June 2015. `https://www.academia.edu/17464066/Bitcoin_Stock_Prediction_Using_Artificial_Neural_Networks`.

[7] Jose Alvarez. Crypto Profiles: Vitalik Buterin, Creator of Ethereum. *Blockonomi*, January 2018. `https://blockonomi.com/vitalik-buterin-profile/`.

[8] Muhammad Amjad and Devavrat Shah. Trading bitcoin and online time series prediction. In Oren Anava, Azadeh Khaleghi, Marco Cuturi, Vitaly Kuznetsov, and Alexander Rakhlin, editors, *Proceedings of the Time Series Workshop at NIPS 2016*, volume 55 of *Proceedings of Machine Learning Research*, pages 1–15, Barcelona, Spain, 09 Dec 2017. PMLR.

[9] Tomaso Aste. Cryptocurrency market structure: connecting emotions and economics. *Digital Finance*, 1(1-4):5–21, November 2019.

[10] Edward N.W. Aw, Christopher R. Dornick, and John Q. Jiang. Combining Quantitative and Fundamental Analysis: A Quant-amental Approach. *Journal of Investing*, 23(2):28–43, 2014.

[11] Chongyang Bai, Tommy White, Linda Xiao, V. S. Subrahmanian, and Ziheng Zhou. C2p2: A Collective Cryptocurrency Up/Down Price Prediction Engine. June 2019. arXiv: 1906.00564.

[12] Mehmet Balcilar, Elie Bouri, Rangan Gupta, and David Roubaud. Can volume predict bitcoin returns and volatility? a quantiles-based approach. *Economic Modelling*, 64:74 – 81, 2017.

[13] Mark C. Ballandies, Marcus M. Dapp, and Evangelos Pournaras. Decrypting Distributed Ledger Design – Taxonomy, Classification and Blockchain Community Evaluation. November 2019. arXiv:1811.03419 [cs].

[14] Mihi Bamburic. Ethereum passes $400. *BetaNews*, November 2017. `https://betanews.com/2017/11/23/ethereum-price-november-23/`.

[15] Jason Michael Baumgartner. Pushshift API. October 2018. `https://github.com/pushshift/api`.

[16] Dirk G. Baur, KiHoon Hong, and Adrian Lee. Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money*, 54(C):177–189, 2018.

[17] Pedro Bao, Antnio Portugal Duarte, Helder Sebastio, and Srdjan Redzepagic. Information Transmission Between Cryptocurrencies: Does Bitcoin Rule the Cryptocurrency World? *Scientific Annals of Economics and Business*, 65(2):97–117, June 2018.

[18] Mike Belshe. [Bitcoin-segwit2x] Segwit2x Final Steps. November 2017. https://lists.linuxfoundation.org/pipermail/bitcoin-segwit2x/2017-November/000685.html.

[19] Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized Anonymous Payments from Bitcoin (extended version). May 2014.

[20] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *CoRR*, 2019. https://arxiv.org/abs/1901.10912.

[21] Binance. Buy and Sell Cryptocurrency. *Binance*, August 2019. https://www.binance.com/en/buy-sell-crypto.

[22] Bitcoin Wiki. Satoshi (unit). *Bitcoin Wiki*, February 2018. https://en.bitcoin.it/wiki/Satoshi_(unit).

[23] bitcointalk.org. Statistics Center. *bitcointalk.org*, September 2018. https://bitcointalk.org/index.php?action=stats.

[24] BitInfoCharts. New Cryptocurrencies 2017. *BitInfoCharts*, March 2018. https://bitinfocharts.com/new-cryptocurrencies-2017.html.

[25] Johannes Bleher and Thomas Dimpfl. Today I got a million, tomorrow, I don't know: On the predictability of cryptocurrencies by means of Google search volume. *International Review of Financial Analysis*, 63:147–159, May 2019.

[26] Blockchain Luxembourg S.A. Blockchain charts & statistics API, December 2018. `https://www.blockchain.com/api/charts_api`.

[27] Ash Booth, Enrico Gerding, and Frank McGroarty. Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications*, 41(8):3651–3661, June 2014.

[28] Ash Booth, Enrico Gerding, and Frank McGroarty. Predicting equity market price impact with performance weighted ensembles of random forests. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 286–293, London, UK, March 2014. IEEE.

[29] Jamal Bouoiyour and Refk Selmi. What does bitcoin look like? *Annals of Economics and Finance*, 16(2):449 – 492, November 2015.

[30] Jamal Bouoiyour, Refk Selmi, and Aviral Tiwari. Is Bitcoin Business Income or Speculative Bubble? Unconditional vs. Conditional frequency domain analysis. *Munich Personal RePEc Archive*, 59595, 2014. `https://mpra.ub.uni-muenchen.de/59595/`.

[31] Jamal Bouoiyour, Refk Selmi, and Aviral Kumar Tiwari. Is Bitcoin business income or speculative foolery? New ideas through an improved frequency domain analysis. *Annals of Financial Economics*, 10(1), June 2015.

[32] Elie Bouri and Rangan Gupta. Predicting Bitcoin returns: Comparing the roles of newspaper- and internet search-based measures of uncertainty. *Finance Research Letters*, 101398, December 2019. In Press, Corrected Proof. Available at: `https://linkinghub.elsevier.com/retrieve/pii/S1544612319307020`.

[33] Elie Bouri, Naji Jalkh, Peter Molnár, and David Roubaud. Bitcoin for energy commodities before and after the December 2013 crash: diversifier, hedge or safe haven? *Applied Economics*, 49:5063–5073, March 2017.

[34] Sean Bowe, Taylor Hornby, and Nathan Wilcox. Zcash Protocol Specification Version 2017.0-beta-2.7. July 2017. `https://github.com/zcash/zips/blob/master/protocol/protocol.pdf`.

[35] Austin Bradford Hill. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300, May 1965.

[36] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[37] Patrick D Bridge and Shlomo S Sawilowsky. Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. *Journal of Clinical Epidemiology*, (3):229–235, 1999.

[38] Paul Bukowski. *Basics of Quantitative Equity Investing*. John Wiley & Sons, Inc, Hoboken, New Jersey, 2013.

[39] Andrew Burnie. Exploring the interconnectedness of cryptocurrencies using correlation networks. In *Cryptocurrency Research Conference 2018*, Cambridge, UK, 2018. Anglia Ruskin University. `https://arxiv.org/abs/1806.06632`.

[40] Andrew Burnie, James Burnie, and Andrew Henderson. Developing a cryptocurrency assessment framework: function over form. *Ledger*, 3:ISSN 2379–5980, July 2018. `https://doi.org/10.5195/ledger.2018.121`.

[41] Andrew Burnie, Andrew Henderson, and James Burnie. Putting Names to Things: reconciling cryptocurrency heterogeneity and regulatory continuity.

Journal of International Banking and Financial Law, 33(2):83–86, February 2018.

[42] Andrew Burnie and Safwan Mchawrab. Pricing of internet companies: Financial and non-financial value drivers. *Bankers, Markets & Investors*, (146):21–37, January-February 2017.

[43] Andrew Burnie and Emine Yilmaz. An Analysis of the Change in Discussions on Social Media with Bitcoin Price. In *42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 889–892, Paris, France, July 2019. ACM SIGIR.

[44] Andrew Burnie and Emine Yilmaz. Data from: Social media and bitcoin metrics: which words matter. *Dryad, Dataset*, v2, September 2019. `https://doi.org/10.5061/dryad.8n6m564`.

[45] Andrew Burnie and Emine Yilmaz. Social media and bitcoin metrics: which words matter. *Royal Society Open Science*, 6, 2019.

[46] Andrew Burnie, Emine Yilmaz, and Tomaso Aste. Analysing social media forums to discover potential causes of phasic shifts in cryptocurrency price series. *Frontiers in Blockchain*, 3:1, 2020.

[47] Andrew Burnie, Emine Yilmaz, and Tomaso Aste. Data from: Analysing social media forums to discover potential causes of phasic shifts in cryptocurrency price series. *Dryad, Dataset*, v3, February 2020. `https://doi.org/10.5061/dryad.q2bvq83f6`.

[48] Chris Burniske and Jack Tatar. *Cryptoassets: the innovative investor's guide to bitcoin and beyond*. McGraw-Hill, New York, New York, 2018.

[49] Zhao Cai, Fei Liu, Eric Lim, Chee-Wee Tan, and Zhiqiang Zheng. Unraveling the Effects of Google Search on Volatility of Cryptocurrencies. In *Thirty ninth International Conference on Information Systems*, San Francisco, California, 2018. Association for Information Systems (AIS).

[50] David Canellis. Three types of cryptocurrency tokens explained as quickly as possible. *The Next Web*, November 2018. https://thenextweb.com/hardfork/2018/11/19/cryptocurrency-tokens-explained/.

[51] Francesco Canepa. G20 leaders to hold fire on cryptocurrencies amid discord: sources. *Reuters*, March 2018.

[52] Mark Carney. The Future of Money. In *The Inaugural Scottish Economics Conference*, Edinburgh, UK, March 2018. Bank of England.

[53] Bitcoin Cash. Bitcoin Cash - Peer-to-Peer Electronic Cash. *Bitcoin Cash - Peer-to-Peer Electronic Cash*, October 2017. https://www.bitcoincash.org/.

[54] Amy Castor. A Short Guide to Bitcoin Forks. *CoinDesk*, March 2017. https://www.coindesk.com/short-guide-bitcoin-forks-explained/.

[55] Stephen Chan, Jeffrey Chu, Saralees Nadarajah, and Joerg Osterrieder. A Statistical Analysis of Cryptocurrencies. *Journal of Risk and Financial Management*, 10(4):12, May 2017.

[56] James Chen. Profit Taking. *Investopedia*, March 2018. https://www.investopedia.com/terms/p/profittaking.asp.

[57] Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365:112395, February 2020.

[58] David Maxwell Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2:445–498, February 2002.

[59] Pavel Ciaian, Miroslava Rajcaniova, and d'Artis Kancs. The economics of bitcoin price formation. *Applied Economics*, 48:1799 – 1815, 2016.

[60] Pavel Ciaian, Miroslava Rajcaniova, and d'Artis Kancs. The digital agenda of virtual currencies: Can BitCoin become a global currency? *Information Systems and e-Business Management*, 14(4):883–919, November 2016.

[61] Pavel Ciaian, Miroslava Rajcaniova, and d'Artis Kancs. Virtual relationships: Short- and long-run evidence from BitCoin and altcoin markets. *Journal of International Financial Markets, Institutions and Money*, 52:173–195, January 2018.

[62] Ethereum Classic. Ethereum Classic. *Ethereum Classic*, 2017. `https://ethereumclassic.github.io/`.

[63] CoinCap. CoinCap Rankings. *CoinCap.io*, October 2017. Accessed `http://coincap.io/` at 15:58 on 30 October 2017.

[64] CoinCap. CoinCap Rankings. *CoinCap.io*, December 2017. Accessed `http://coincap.io/` at 10:28 on 18 December 2017.

[65] CoinDesk. What is the Difference Between Litecoin and Bitcoin? *CoinDesk*, February 2014. `https://www.coindesk.com/information/comparing-litecoin-bitcoin/`.

[66] CoinMarketCap. Cryptocurrency market capitalizations. *CoinMarketCap*, December 2017. `https://coinmarketcap.com/`.

[67] CoinMarketCap. Cryptocurrency market capitalizations, December 2017. Accessed `https://coinmarketcap.com/` at 10:27 on 18 December 2017.

[68] CoinMarketCap. Cryptocurrency market capitalizations, October 2017. Accessed `https://coinmarketcap.com/` at 14:27 on 4 October 2017.

[69] CoinMarketCap. Cryptocurrency market capitalizations, October 2017. Accessed `https://coinmarketcap.com/` at 15:48 on 30 October 2017.

[70] CoinMarketCap. Cryptocurrency Market Capitalizations. *coinmarketcap.com*, January 2019. Accessed `https://coinmarketcap.com/` at 20:40 on 17 January 2019.

[71] Cointelegraph. What is Bitcoin Cash? *Cointelegraph*, January 2018.

[72] Christina Comben and Coin Rivet. Three Ethereum Subreddits you should start following. *Yahoo Finance*, May 2019. `https://finance.yahoo.com/news/three-ethereum-subreddits-start-following-080019921.html`.

[73] United States Securities and Exchange Commission. Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934: The DAO. Technical Report 81207, July 2017.

[74] Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *JNCI: Journal of the National Cancer Institute*, 22(1):173–203, 1959.

[75] Chloe Cornish, Izabella Kaminska, Philip Stafford, Hannah Murphy, and Adam Samson. Chaotic trading marks new surge in bitcoin price. *Financial Times*, December 2017.

[76] crypto oracle. TOP 5 PRIVACY CRYPTOCURRENCIES - FUTURE OF CRYPTO SECURITY. *Steemit*, July 2017.

[77] Anthony Cuthbertson. Bitcoin price passes $7,000 as remarkable recovery continues into seventh week. *The Independent*, May 2019. `https://www.independent.co.uk/life-style/gadgets-and-tech/news/bitcoin-price-latest-btc-cryptocurrency-usd-exchange-rate-a8911186.html`.

[78] Patrick Dai, Neil Mahi, Jordan Earls, and Alex Norta. Smart-Contract Value-Transfer Protocols on a Distributed Mobile Applica-

tion Platform. 2017. https://qtum.org/uploads/files/a2772efe4dc8ed1100319c6480195fb1.pdf.

[79] Shabbir Dastgir, Ender Demir, Gareth Downing, Giray Gozgor, and Chi Keung Marco Lau. The causal relationship between Bitcoin attention and Bitcoin returns: Evidence from the Copula-based Granger causality test. *Finance Research Letters*, 28:160–164, March 2019.

[80] Rafael Delfin. A General Taxonomy for Cryptographic Assets. 2018. https://assets.ctfassets.net/sdlntm3tthp6/6mqu1HTdBKG46Q6iqa26uE/df09eaf16935053c99c8fcdce658c7ae/General_Taxonomy_for_Cryptographic_Assets.pdf.

[81] Elizabeth Demers and Baruch Lev. A rude awakening: Internet shakeout in 2000. *Review of Accounting Studies*, 6(2):331–359, 2001.

[82] Amrita Dhillon, Grammateia Kotsialou, Peter McBurney, and Luke Riley. *Introduction to Voting and the Blockchain: some open questions for economists*. Number 416 in Working Paper Series. University of Warwick, Department of Economics, Centre for Competitive Advantage in the Global Economy, Coventry, UK, June 2019.

[83] Alexander Dickerson. Algorithmic Trading of Bitcoin Using Wikipedia and Google Search Volume. *SSRN Electronic Journal*, 2018. https://www.ssrn.com/abstract=3177738.

[84] Thomas Dimpfl and Franziska J. Peter. Group transfer entropy with an application to cryptocurrencies. *Physica A: Statistical Mechanics and its Applications*, 516:543–551, February 2019.

[85] Thomas Dimpfl and Franziska Julia Peter. Using transfer entropy to measure information flows between financial markets. *Studies in Nonlinear Dynamics and Econometrics*, 17(1), January 2013.

[86] Kristian Dokic, Mirjana Radman Funari, and Katarina Potnik Gali. The Relationship between the Cryptocurrency Value (Bitcoin) and Interest for It in the Region. *SSRN Electronic Journal*, 2015. `https://www.ssrn.com/abstract=3281915`.

[87] Andreu Rodriguez i Donaire. Why NEM is Your Production Level Blockchain Platform. *NEM*, July 2017. `https://blog.nem.io/production-level-blockchain/`.

[88] Jillian D'Onfro. Facebook is reversing its ban on cryptocurrency ads. *CNBC*, June 2018.

[89] Jillian D'Onfro. Google reverses ban on cryptocurrency exchange advertising in US Japan. *CNBC*, September 2018.

[90] Michael Dowling, Mark Cummins, and Brian M. Lucey. Psychological barriers in oil futures markets. *Energy Economics*, 53:293–304, January 2016.

[91] Evan Duffield and Daniel Diaz. Dash: A PrivacyCentric CryptoCurrency. 2015.

[92] European Banking Authority (EBA). Report with advice for the european commission: on crypto-assets. January 2019.

[93] Abeer ElBahrawy, Laura Alessandretti, and Andrea Baronchelli. Wikipedia and Cryptocurrencies: Interplay Between Collective Attention and Market Performance. *Frontiers in Blockchain*, 2:12, October 2019. `https://doi.org/10.3389/fbloc.2019.00012`.

[94] Hermann Elendner, Simon Trimborn, Bobby Ong, and Teik Ming Lee. The Cross-Section of Crypto-Currencies as Financial Assets: An Overview. *SFB 649 Discussion Paper*, (2016-038, SFB 649, Economic Risk), October 2016.

[95] EOS.IO. EOS.IO Technical White Paper. December 2017. `https://github.com/EOSIO/Documentation`.

[96] Sacha Epskamp, Anglique OJ Cramer, Lourens J. Waldorp, Verena D. Schmittmann, and Denny Borsboom. qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48(4):1–18, 2012.

[97] Ethereum. White Paper. January 2019. `https://github.com/ethereum/wiki/wiki/White-Paper`.

[98] Etherscan. Ethereum Charts And Statistics. *Ethereum (ETH) Blockchain Explorer*, January 2019. `https://etherscan.io/charts`.

[99] Etherscan. Ethereum (Ether) Historical Prices. *Ethereum (ETH) Blockchain Explorer*, May 2019. `https://etherscan.io/chart/etherprice`.

[100] Eversheds Sutherland (International) LLP. Eversheds Sutherland (International) LLP - written evidence. *Digital Currencies Inquiry*, DGC0020, May 2018. `http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/treasury-committee/digital-currencies/written/81375.pdf`.

[101] Financial Conduct Authority (FCA). Cryptoassets Ownership and attitudes in the UK: Consumer survey research report. March 2019.

[102] Financial Conduct Authority (FCA). Guidance on cryptoassets. *Consultation Paper*, CP19/3, January 2019.

[103] Gianna Figá-Talamanca and Marco Patacca. Disentangling the relationship between Bitcoin and market attention measures. *Journal of Industrial and Business Economics*, August 2019.

[104] Gianna Figá-Talamanca and Marco Patacca. Does market attention affect Bitcoin returns and volatility? *Decisions in Economics and Finance*, 42(1):135–155, June 2019.

[105] Onchain Finance. Cryptoasset rankings and metrics for investors. *Cryptoasset network value, market cap, rankings & metrics*, October 2017. Accessed `https://onchainfx.com/` at 15:58 on 30 October 2017.

[106] Onchain Finance. Cryptoasset rankings and metrics for investors. *Cryptoasset network value, market cap, rankings & metrics*, December 2017. Accessed `https://onchainfx.com/` at 10:28 on 18 December 2017.

[107] Swiss Financial Market Supervisory Authority (FINMA). FINMA publishes ICO guidelines. *FINMA*, February 2018.

[108] Qtum Foundation. Qtum Blockchain Economy Whitepaper. 2017. `https://qtum.org/uploads/files/ef2723f33deef1875ef17361f7c696ef.pdf`.

[109] Jane Frecknall-Hughes. Why the UK tax year begins on April 6 (it's a very strange tale). *The Independent*, April 2016. `http://www.independent.co.uk/money/why-the-uk-tax-year-begins-on-april-6-it-s-a-very-strange-tale-a6970801.html`.

[110] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[111] Neil Gandal and Hanna Halaburda. Can We Predict the Winner in a Market with Network Effects? Competition in Cryptocurrency Market. *Games*, 7(4):16, July 2016.

[112] David Garcia and Frank Schweitzer. Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science*, 2(9), 2015.

[113] David Garcia, Claudio J. Tessone, Pavlin Mavrodiev, and Nicolas Perony. The digital traces of bubbles: feedback cycles between socio-economic sig-

nals in the Bitcoin economy. *Journal of The Royal Society Interface*, 11(99), 2014. `https://doi.org/10.1098/rsif.2014.0623`.

[114] Ifigeneia Georgoula, Demitrios Pournarakis, Christos Bilanakos, Dionisios N. Sotiropoulos, and George M. Giaglis. Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices. *SSRN Electronic Journal*, 2015. `http://www.ssrn.com/abstract=2607167`.

[115] Zafar Gilani, Reza Farahbakhsh, and Jon Crowcroft. Do Bots impact Twitter activity? In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 781–782, Perth, Australia, 2017. ACM Press.

[116] Florian Glaser, Kai Zimmermann, Martin Haferkorn, Moritz Weber, and Michael Dowling. BITCOIN - ASSET OR CURRENCY? REVEALING USERS' HIDDEN INTENTIONS. In *Twenty Second European Conference on Information Systems (ECIS 2014)*, Tel Aviv, Israel, 2014. Association for Information Systems.

[117] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, June 2019.

[118] Łukasz Goczek and Ivan Skliarov. What drives the Bitcoin price? A factor augmented error correction mechanism investigation. *Applied Economics*, 51(59):6393–6410, December 2019.

[119] Bitcoin Gold. Bitcoin Gold - GPU Bitcoin Mining (Official Website). *Bitcoin Gold*, December 2017.

[120] Google. word2vec, July 2013. `https://code.google.com/archive/p/word2vec/`.

[121] Google. Google trends, December 2018. `https://trends.google.com/trends`.

[122] C. W. J. Granger. Testing for Causality. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[123] Alex Greaves and Benjamin Au. Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin. 2015.

[124] Robert J. Greer. What is an asset class, anyway? *Journal of Portfolio Management*, 23(2):86–91, 1997.

[125] David A Grimes and Kenneth F Schulz. Bias and causal associations in observational research. *The Lancet*, 359(9302):248–252, January 2002.

[126] 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*, 11(5/6):561–566, November 2004.

[127] Tian Guo and Nino Antulov-Fantulin. An experimental study of Bitcoin fluctuation using machine learning methods. February 2018. arXiv:1802.04065 [cs, stat].

[128] Hansel. A Better Taxonomy for Cryptocurrencies. *Medium*, September 2018. `https://medium.com/swlh/a-better-taxonomy-for-cryptocurrencies-cbffd2e1b58c`.

[129] David A. Harding. Saving up to 80% on Bitcoin transaction fees by batching payments. *Medium*, August 2017. `https://bitcointechtalk.com/saving-up-to-80-on-bitcoin-transaction-fees-by-batching-payments-4147ab7009fb`.

[130] Colin Harper. North Korean Hackers Attacked South Korean Exchanges, NIS Claims. *The Merkle Hash*, December 2017. `https://themerkle.com/north-korean-hackers-attacked-south-korean-exchanges-nis-claims/`.

[131] Saike He, Xiaolong Zheng, Daniel Zeng, Kainan Cui, Zhu Zhang, and Chuan Luo. Identifying Peer Influence in Online Social Networks Using Transfer

Entropy. In G. Alan Wang, Xiaolong Zheng, Michael Chau, and Hsinchun Chen, editors, *Intelligence and Security Informatics Pacific Asia Workshop, PAISI 2013*, volume LNCS 8039, Beijing, China, August 2013. Springer.

[132] Kareem Hegazy and Samuel Mumford. Comparitive Automated Bitcoin Trading Strategies. 2016. `https://pdfs.semanticscholar.org/50de/ca6799f76ca00dd3adfa1fd86c4be3f4c068.pdf`.

[133] Jeff Herbert and Martin Stabauer. Bitcoin & co: An ontology for categorising cryptocurrencies. In *2015 M-SPHERE: Book of Papers*, pages 45–55. Accent, Zagreb, Croatia, January 2016.

[134] Alyssa Hertig. Bitcoin UASF Proposal Quietly Activates - to Little Effect. *CoinDesk*, August 2017. `https://www.coindesk.com/bitcoin-uasf-proposal-quietly-activates-little-effect`.

[135] Eyal Hertzog, Guy Benartzi, and Galia Benartzi. Bancor Protocol: Decentralized networks for smart contract based tokens to enable continuous convertibility between them. October 2017. `https://www.bancor.network/static/bancor_protocol_whitepaper_en.pdf`.

[136] Stan Higgins. CFTC Aligns With SEC: ICO Tokens Can Be Commodities. *CoinDesk*, October 2017. `https://www.coindesk.com/cftc-no-inconsistency-sec-cryptocurrency-regulation/`.

[137] Garrick Hileman. Alternative Currencies: A Historical Survey and Taxonomy. *SSRN Electronic Journal*, March 2013. `https://doi.org/10.2139/ssrn.2747975`.

[138] Garrick Hileman and Michel Rauchs. 2017 Global Cryptocurrency Benchmarking Study. *SSRN Electronic Journal*, 2017. `https://ssrn.com/abstract=2965436`.

[139] William Hinman. Digital Asset Transactions: When Howey Met Gary (Plastic). *SEC.gov*, June 2018.

[140] Charles Hoskinson. Why we are building Cardano. June 2017. https://whycardano.com/.

[141] House of Commons Treasury Committee. Crypto-assets HC 910. *UK Parliament House of Commons*, 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmtreasy/910/91002.htm.

[142] C.J. Hutto and Eric Gilbert. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, Michigan, 2014. AAAI Press.

[143] N.I. Indera, I.M. Yassin, A. Zabidi, and Z.I. Rizman. Non-linear Autoregressive with Exogeneous input (NARX) bitcoin price prediction model using PSO-optimized parameters and moving average technical indicators. *Journal of Fundamental and Applied Sciences*, 9(3S):791–808, January 2017.

[144] Kik Interactive. Kin: a decentralized ecosystem of digital services for daily life. May 2017. http://www.kinecosystem.org/static/files/Kin_Whitepaper_V1_English.pdf.

[145] Investopedia. Pump And Dump. *Investopedia*, November 2003. https://www.investopedia.com/terms/p/pumpanddump.asp.

[146] John P. A. Ioannidis. Exposure-wide epidemiology: revisiting Bradford Hill. *Statistics in Medicine*, 35(11):1749–1762, May 2016.

[147] Anna Irrera and Steve Stecklow. Tezos organizers hit with second lawsuit over cryptocurrency fundraiser. *Reuters*, November 2017.

[148] Anna Irrera and Steve Stecklow. Tezos organizers sued in California over crypto currency project. *Reuters*, November 2017.

[149] Kim Jaewon. South Korea joins in Asia-wide bitcoin crackdown. *Nikkei Asian Review*, December 2017. https://asia.nikkei.com/

Economy/South-Korea-joins-in-Asia-wide-bitcoin-crackdown.

[150] Ryan G. James, Nix Barnett, and James P. Crutchfield. Information Flows? A Critique of Transfer Entropies. *Physical Review Letters*, 116(23):238701, June 2016.

[151] Huisu Jang and Jaewook Lee. An Empirical Study on Modeling and Prediction of Bitcoin Prices With Bayesian Neural Networks Based on Blockchain Information. *IEEE Access*, 6:5427–5437, 2018.

[152] Jean-Pierre Buntinx. What is MyEtherShop? *The Merkle*, August 2017. https://themerkle.com/what-is-myethershop/.

[153] Qiang Ji, Elie Bouri, David Roubaud, and Ladislav Kristoufek. Information interdependence among energy, cryptocurrency and major commodity markets. *Energy Economics*, 81:1042–1055, June 2019.

[154] Anjali Ganesh Jivani. A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Applications*, 2(6):1930–1938, 2011.

[155] Julia Kagan. Tax Year Definition. *Investopedia*, July 2019. https://www.investopedia.com/terms/t/taxyear.asp.

[156] Jermain Kaminski. Nowcasting the Bitcoin Market with Twitter Signals. June 2014. arXiv:1406.7577 [cs].

[157] Andrei Kashcha. Exploring word2vec embeddings as a graph of nearest neighbors: anvaka/word2vec-graph. January 2019. https://github.com/anvaka/word2vec-graph.

[158] Paraskevi Katsiampa. Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters*, 158:3–6, September 2017.

[159] M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3), November 1945.

[160] Marvin Aron Kennis. Multi-channel discourse as an indicator for Bitcoin price and volume movements. November 2018. arXiv:1811.03146 [cs, q-fin, stat].

[161] Z. Keskin and T. Aste. Information-theoretic measures for non-linear causality detection: application to social media sentiment and cryptocurrency prices. June 2019. arXiv:1906.05740 [physics, q-fin].

[162] Rohaifa Khaldi, Abdellatif El Afia, Raddouane Chiheb, and Rdouan Faizi. Forecasting of Bitcoin Daily Returns with EEMD-ELMAN based Model. In *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications - LOPAL '18*, pages 1–6, Rabat, Morocco, 2018. ACM Press.

[163] Han Kyul Kim, Hyunjoong Kim, and Sungzoon Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, November 2017.

[164] Kwansoo Kim, Sang-Yong Tom Lee, and Said Assar. Coin Market Behavior using Social Sentiment Markov Chains. Xi'an, China, 2019. Association for Information Systems (AIS). `http://www.pacis2019.org/wd/Submissions/PACIS2019_paper_196.pdf`.

[165] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLOS ONE*, 11(8):e0161197, August 2016.

[166] Young Bin Kim, Jurim Lee, Nuri Park, Jaegul Choo, Jong-Hyun Kim, and Chang Hun Kim. When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctua-

tion. *PLOS ONE*, 12(5):e0177630, May 2017. https://doi.org/10.1371/journal.pone.0177630.

[167] Frode Kjærland, Aras Khazal, Erlend Krogstad, Frans Nordstrøm, and Are Oust. An Analysis of Bitcoin's Price Dynamics. *Journal of Risk and Financial Management*, 11(4), October 2018.

[168] Frode Kjærland, Maria Meland, Are Oust, and Vilde Øyen. How can Bitcoin Price Fluctuations be Explained? *International Journal of Economics and Financial Issues*, 8(3):323–332, 2018. https://www.econjournals.com/index.php/ijefi/article/view/6446/pdf.

[169] Megan L. Knittel and Rick Wash. How "True Bitcoiners" Work on Reddit to Maintain Bitcoin. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*, Glasgow, UK, 2019. ACM Press.

[170] Grammateia Kotsialou, Luke Riley, Amrita Dhillon, Toktam Mahmoodi, Peter McBurney, Paul Massey, and Richard Pearce. Using Distributed Ledger Technology for Shareholder Rights Management. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden, 2018.

[171] Marius Kraemer. Breakdown of Cryptocurrency Market: 12 Major Cryptocurrency Categories. *Master The Crypto*, May 2018.

[172] Kraken. Bitcoin currency code: XBT vs BTC. *Kraken*. http://support.kraken.com/hc/en-us/articles/360001206766-Bitcoin-currency-code-XBT-vs-BTC.

[173] Ladislav Kristoufek. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3(3415), December 2013.

[174] Ladislav Kristoufek. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PLOS ONE*, 10(4):e0123923, April 2015.

[175] Ladislav Kristoufek. Is the Bitcoin price dynamics economically reasonable? Evidence from fundamental laws. *Physica A: Statistical Mechanics and its Applications*, 536, June 2019.

[176] Miron Bartosz Kursa. Package Boruta. July 2018. `https://cran.r-project.org/web/packages/Boruta/Boruta.pdf`.

[177] Büşra Kutlu, Durmuş Sezer, and Umut Tolga Gümüş. CAN BITCOINS' PRICES BE PREDICTED BY GOOGLE TRENDS DATA? AN EXAMPLE OF TURKEY WITH COMPARISION OF USA. *International Journal of Academic Value Studies*, 3(10):167–177, 2017.

[178] Protocol Labs. Filecoin: A Decentralized Storage Network. August 2017. `https://filecoin.io/filecoin.pdf`.

[179] Connor Lamon, Eric Nielsen, and Eric Redondo. Cryptocurrency Price Change Prediction Using News and Social Media Sentiment. 2018. `http://cs230.stanford.edu/files_winter_2018/projects/6929537.pdf`.

[180] Matthias Langer. Taxation of Cryptocurrencies in Europe. *Crypto Research Report*, December 2017. `https://cryptoresearch.report`.

[181] Marek Laskowski and Henry M. Kim. Rapid Prototyping of a Text Mining Application for Cryptocurrency Market Intelligence. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 448–453, Pittsburgh, Pennsylvania, July 2016. IEEE.

[182] Xin Li and Chong Alex Wang. The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. *Decision Support Systems*, 95:49–60, March 2017.

[183] Yukun Liu and Aleh Tsyvinski. Risks and Returns of Cryptocurrency. *NBER Working Paper Series*, (24877), August 2018. `http://www.nber.org/papers/w24877`.

[184] Manan Lohia. A Classification of Crypto-tokens. *Medium*, April 2019. `https://medium.com/towardsblockchain/a-classification-of-crypto-tokens-aa416af26c40`.

[185] Natasha Lomas. PSA: No India hasn't banned Bitcoin but its still talking tough on crypto. *TechCrunch*, February 2018. `http://social.techcrunch.com/2018/02/03/psa-no-india-hasnt-banned-bitcoin-but-its-still-talking-tough-on-crypto/`.

[186] QUOINE Pte. Ltd. Whitepaper version 1.9: Providing liquidity to the non-LIQUID Crypto Economy. October 2017. `https://liquid.plus/`.

[187] Tether Ltd. Tether: Fiat currencies on the Bitcoin blockchain. June 2016. `https://tether.to/wp-content/uploads/2016/06/TetherWhitePaper.pdf`.

[188] Kirichenko Lyudmyla, Bulakh Vitalii, and Radivilova Tamara. Fractal time series analysis of social network activities. In *2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, pages 456–459, Kharkov, Ukraine, October 2017. IEEE.

[189] Isaac Madan, Shaurya Saluja, and Aojia Zhao. Automated Bitcoin Trading via Machine Learning Algorithms. 2014.

[190] Daniele Magazzeni, Peter McBurney, and William Nash. Validation and Verification of Smart Contracts: A Research Agenda. *Computer*, 50(9):50–57, 2017.

[191] Feng Mai, Qing Bai, Zhe Shan, Xin (Shane) Wang, and Roger H.L. Chiang. From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance. *SSRN Electronic Journal*, 2015. `http://www.ssrn.com/abstract=2545957`.

[192] Feng Mai, Zhe Shan, Qing Bai, Xin (Shane) Wang, and Roger H.L. Chiang. How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis. *Journal of Management Information Systems*, 35(1):19–52, January 2018.

[193] MakerDAO. Dai is now live! *The Maker Blog*, August 2017. `https://blog.makerdao.com/dai-is-now-live/`.

[194] Dominique Drouet Mari and Samuel Kotz. *Correlation and dependence*. Imperial College Press, London, UK, 2001.

[195] R. Marschinski and H. Kantz. Analysing the information flow between financial time series: An improved estimator for transfer entropy. *The European Physical Journal B*, 30(2):275–281, November 2002.

[196] Leon Marshall, Andrey Repin, Constantine Tsavliris, and Charlie Humberstone. Cryptoasset Taxonomy Report 2018. October 2018. `https://www.cryptocompare.com/media/34478555/cryptocompare-cryptoasset-taxonomy-report-2018.pdf`.

[197] Annie Massa, Lily Katz, and Matthew Leising. Ripple Has Tried to Buy Its Way Onto Major Exchanges for Cryptocurrency. *Bloomberg*, April 2018.

[198] Martina Matta, Ilaria Lunesu, and Michele Marchesi. Bitcoin Spread Prediction Using Social And Web Search Media. *DeCAT 2015: Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services*, June 2015.

[199] John H. McDonald. *Handbook of Biological Statistics*. Sparky House, Baltimore, Maryland, 3 edition, 2014.

[200] Sean McNally. *Predicting the price of Bitcoin using Machine Learning*. PhD Thesis, Dublin, National College of Ireland, 2016.

[201] Sean McNally, Jason Roche, and Simon Caton. Predicting the Price of Bitcoin Using Machine Learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 339–343, Cambridge, UK, March 2018. IEEE.

[202] Hongyuan Mei and Jason M Eisner. The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, California, 2017.

[203] David Meyer. Cryptocurrencies Like Bitcoin Are Commodities, Federal Judge Says. Here's Why That Matters. *Fortune*, March 2018.

[204] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. January 2013. arXiv 1301.3781 [cs].

[205] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., Lake Tahoe, Nevada, 2013.

[206] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3(1), December 2013.

[207] Monero. The Merits of Monero: Why Monero vs Bitcoin. *Monero.How*, 2017. `https://www.monero.how/why-monero-vs-bitcoin`.

[208] James N. Morgan and John A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434, 1963.

[209] Casey Murphy. Support and Resistance Basics. *Investopedia*, April 2019. `https://www.investopedia.com/trading/support-and-resistance-basics/`.

[210] Steven Musil. Twitter ramps up effort to combat abusive bots, trolls. *CNET*, June 2018.

[211] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008. `https://bitcoin.org/bitcoin.pdf`.

[212] Masafumi Nakano, Akihiko Takahashi, and Soichiro Takahashi. Bitcoin technical trading with artificial neural network. *SSRN Electronic Journal*, February 2018. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3128726`.

[213] NetworkX. Software for complex networks. 2019. `https://networkx.github.io/`.

[214] Shen Noether, Adam Mackenzie, and the Monero Research Lab. Ring Confidential Transactions. *Ledger*, 1:1–18, December 2016.

[215] Board of Governors of the Federal Reserve System. FAQs: How does the Federal Reserve Board determine how much currency to order each year? *The Fed*, December 2014.

[216] Board of Governors of the Federal Reserve System. FAQs: How much does it cost to produce currency and coin? *The Fed*, March 2017.

[217] Federal Reserve Bank of New York. Currency Processing and Destruction. *Federal Reserve Bank of New York*, May 2014.

[218] Stephen O'Neal. From Coincheck to Bithumb: 2018's Largest Security Breaches So Far. *Cointelegraph*, June 2018.

[219] Jörg Osterrieder, Julian Lorenz, and Martin Strika. Bitcoin and Cryptocurrencies - Not for the Faint-Hearted. *International Finance and Banking*, 4(1):56–94, January 2017.

[220] Theodore Panagiotidis, Thanasis Stengos, and Orestis Vravosinos. On the determinants of bitcoin returns: A LASSO approach. *Finance Research Letters*, 27:235–240, December 2018.

[221] Theodore Panagiotidis, Thanasis Stengos, and Orestis Vravosinos. The effects of markets, uncertainty and search intensity on bitcoin returns. *International Review of Financial Analysis*, 63:220–242, May 2019.

[222] Bohdan M. Pavlyshenko. Bitcoin Price Predictive Modeling Using Expert Correction. In *2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT)*, pages 163–167, Lviv, Ukraine, September 2019. IEEE.

[223] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference In Statistics: A Primer*. Wiley, Chichester, UK, 2016.

[224] Daniel Pele and Miruna Mazurencu-Marinescu-Pele. Using High-Frequency Entropy to Forecast Bitcoins Daily Value at Risk. *Entropy*, 21(2), January 2019.

[225] Brian Perry-Carrera. Effect of sentiment on bitcoin price formation. 2018. `https://sites.duke.edu/djepapers/files/2018/06/brianperrycarrera-dje.pdf`.

[226] Ross C. Phillips and Denise Gorse. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, Honolulu, Hawaii, November 2017. IEEE.

[227] Ross C. Phillips and Denise Gorse. Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PLOS ONE*, 13(4):e0195200, April 2018.

[228] Ross C. Phillips and Denise Gorse. Mutual-Excitation of Cryptocurrency Market Returns and Social Media Topics. In *Proceedings of the 4th International Conference on Frontiers of Educational Technologies - ICFET '18*, pages 80–86, Moscow, Russian Federation, 2018. ACM Press.

[229] Lukáš Pichl and Taisei Kaizoji. Volatility Analysis of Bitcoin Price Time Series. *Quantitative Finance and Economics*, 1(4):474–485, 2017.

[230] Michal Polasik, Anna Iwona Piotrowska, Tomasz Piotr Wisniewski, Radoslaw Kotkowski, and Geoffrey Lightfoot. Price Fluctuations and the Use of Bitcoin: An Empirical Inquiry. *International Journal of Electronic Commerce*, 20(1):9–49, September 2015.

[231] Ross Powell. Market Manipulation and a Case for the Further Regulation of Social Media and the Finance Industry. *Skidmore College Creative Matter Economics Student Theses and Capstone Projects*, 125, 2019. `https://creativematter.skidmore.edu/econ_studt_schol/125`.

[232] Obryan Poyser. Exploring the dynamics of Bitcoins price: a Bayesian structural time series approach. *Eurasian Economic Review*, 9(1):29–60, March 2019.

[233] Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*, 3(1), December 2013.

[234] Tobias Preis, Daniel Reith, and H. Eugene Stanley. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719, December 2010.

[235] PricewaterhouseCoopers. Making sense of bitcoin, cryptocurrency, and blockchain. *PwC*, April 2018. https://www.pwc.com/us/en/industries/financial-services/fintech/bitcoin-blockchain-cryptocurrency.html.

[236] The Neo Project. NEO White Paper. October 2017. https://github.com/neo-project/docs.

[237] Pranav Ramarao. Hashflare Math: Should you Reinvest? *Medium*, December 2017. https://medium.com/@pranav.rr93/hashflare-making-60x-returns-in-1-year-9dda301a4943.

[238] Elaine Ramirez. Facts And Myths Surrounding Crypto In South Korea: Death, Taxes And Bans. *Forbes*, February 2018.

[239] Kenneth Rapoza. Cryptocurrency Exchanges Officially Dead In China. *Forbes*.

[240] r/Bitcoin. *Reddit*, September 2018. https://www.reddit.com/r/Bitcoin/.

[241] Revealing Reality. How and why consumers buy cryptoassets: A report for the FCA. 2019.

[242] Jonathan Rebane, Isak Karlsson, Stojan Denic, and Panagiotis Papapetrou. Seq2Seq RNNs and ARIMA models for Cryptocurrency Prediction: A Comparative Study. In *SIGKDD Workshop on Fintech (SIGKDD Fintech'18)*, London, UK, August 2018. ACM. https://www.semanticscholar.org/paper/Seq-2-Seq-RNNs-and-ARIMA-models-for-Cryptocurrency-Rebane/c1ec480f005244a2cfb93f1d3ad15c2d22b864d6.

[243] reddit. What is a moderator? *Reddit Help*, December 2019. https://www.reddithelp.com/en/categories/reddit-101/moderators/what-moderator.

[244] Jamie Redman. Op-Codes and Scripting Capabilities Coming to Bitcoin Cash. *Bitcoin News*, April 2018. https://news.bitcoin.com/op-codes-and-scripting-capabilities-coming-to-bitcoin-cash/.

[245] r/ethereum. *Reddit*, March 2019. https://www.reddit.com/r/ethereum/.

[246] Bailey Reutzel. Logical or Not, Bitcoin's Coming Fork Is Boosting Its Price. *CoinDesk*, October 2017. https://www.coindesk.com/logical-not-bitcoins-coming-fork-boosting-price/.

[247] Luke Riley, Grammateia Kotsialou, Amrita Dhillon, Toktam Mahmoodi, Peter McBurney, and Richard Pearce. Deploying a Shareholder Rights Management System onto a Distributed Ledger. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, pages 2381–2383, Montreal, Canada, May 2019.

[248] Ripple. Ripple - One Frictionless Experience To Send Money Globally. 2017. https://ripple.com/.

[249] Ripple. XRP The Digital Asset for Payments. 2017. https://ripple.com/xrp/.

[250] Kate Rooney. Twitter bans cryptocurrency advertising, joining other tech giants in crackdown. *CNBC*, March 2018.

[251] Paul R. Rosenbaum. *Observation and Experiment: An Introduction To Causal Inference*. Harvard University Press, Cambridge, Massachusetts, 2017.

[252] Kenneth J. Rothman. Causes. *American Journal of Epidemiology*, 185(11):1035–1040, June 2017.

[253] J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, July 2018.

[254] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1), December 2019. `https://doi.org/10.1038/s41467-019-10105-3`.

[255] Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, May 2018.

[256] Thomas Schreiber. Measuring Information Transfer. *Physical Review Letters*, 85(2):461–464, July 2000.

[257] European Securities and Markets Authority (ESMA). Advice: Initial coin offerings and crypto-assets. ESMA50-157-1391, January 2019.

[258] Kai Sedgwick. Everything You Should Know About Bitcoin Address Formats. *Bitcoin News*, February 2019. `https://news.bitcoin.com/everything-you-should-know-about-bitcoin-address-formats/`.

[259] Devavrat Shah and Kang Zhang. Bayesian regression and Bitcoin. In *52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 409–414, Monticello, Illinois, 30 September - 3 October 2014. IEEE.

[260] Robert Shiller. Diverse views on asset bubbles. In William C. Hunter, George G. Kaufman, and Michael Pomerleano, editors, *Asset Price Bubbles:*

The Implications for Monetary, Regulatory and International Policies, pages 35–39. MIT Press, Cambridge, Massachusetts, 2003.

[261] Robert J. Shiller. *Irrational Exuberance*. Princeton University Press, Princeton, New Jersey, 2000.

[262] Edwin Sin and Lipo Wang. Bitcoin price prediction using ensembles of neural networks. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 666–671, Guilin, China, July 2017. IEEE.

[263] Brianne Smith. The Life-Cycle and Character of Crypto-Assets: A Framework for Regulation and Investor Protection. *Journal of Accounting and Finance*, 19(1):156–168, 2019.

[264] Smith + Crown. ICOs and Crowdsales. *Cryptofinancial News, Research & Analysis*, January 2017. `https://www.smithandcrown.com/icos/`.

[265] Nico Smuts. What Drives Cryptocurrency Prices?: An Investigation of Google Trends and Telegram Sentiment. *ACM SIGMETRICS Performance Evaluation Review*, 46(3):131–134, January 2019.

[266] Yhlas Sovbetov. Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero. *Journal of Economics and Financial Analysis*, 2(2):1–27, 2018.

[267] Peter Spirtes and Clark Glymour. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, 9(1):62–72, April 1991.

[268] Lars Steinert and Christian Herff. Predicting altcoin returns using social media. *PLOS ONE*, 13(12):e0208119, December 2018.

[269] Evita Stenqvist and Jacob Lönnö. Predicting Bitcoin price fluctuation with Twitter sentiment analysis. 2017. `http://www.diva-portal.org/smash/get/diva2:1110776/FULLTEXT01.pdf`.

[270] Stefania Stimolo. Token classification: the differences between crypto, stable coin, security, utility and equity. *The Cryptonomist*, December 2018.

[271] James Stock and Mark Watson. *Introduction to Econometrics*. Pearson, Harlow, UK, 3 edition, 2012.

[272] Sowmya Subramaniam and Madhumita Chakraborty. Investor Attention and Cryptocurrency Returns: Evidence from Quantile Causality Approach. *Journal of Behavioral Finance*, 21(1):103–115, January 2020.

[273] Sujha Sundararajan. Japanese Electronics Retail Giant Launches Bitcoin Payments. *CoinDesk*, January 2018. https://www.coindesk.com/japanese-electronics-retailer-launches-bitcoin-payments.

[274] Nassim Nicholas Taleb. *The Black Swan: The Impact Of The Highly Improbable*. Incerto. Penguin Books, London, UK, revised edition, 2010.

[275] Paolo Tasca and Claudio J. Tessone. A Taxonomy of Blockchain Technologies: Principles of Identification and Classification. *Ledger*, 4, February 2019.

[276] AirSwap Team. The AirSwap Token. *AirSwap Blog*, September 2017. https://blog.airswap.io/the-airswap-token-42855fe5e120.

[277] AirSwap Team. Swap: A Peer-to-Peer Protocol for Trading Ethereum Tokens. June 2017. https://swap.tech/whitepaper/.

[278] Omni Team. Omni Layer: An open-source, fully-decentralized asset platform on the Bitcoin Blockchain. *Omni Layer*, 2017. http://www.omnilayer.org/.

[279] thegrinder. The BitConnect scam exposed. *Steemit*, April 2017. https://steemit.com/scam/@thegrinder/the-bitconnect-scam-exposed.

[280] Ramakrishna Thurimella and Yeturu Aahlad. The Hitchhikers Guide to Blockchains: A Trust Based Taxonomy. *WANdisco*, November 2018.

[281] Marko Ticak. Lightening vs. LightningWhats the Difference? *Grammarly*, November 2016. https://www.grammarly.com/blog/lightening-vs-lightning/.

[282] Kyle Torpey. Comparing Digital Currencies? Market Cap Doesn't Tell the Whole Story. *Bitcoin Magazine*, December 2016.

[283] Douglas Garcia Torres and Hongliang Qiu. Applying Recurrent Neural Networks for Multivariate Time Series Forecasting of Volatile Financial Data. https://www.researchgate.net/publication/322027012_Applying_Recurrent_Neural_Networks_for_Multivariate_Time_Series_Forecasting_of_Volatile_Financial_Data.

[284] Intelligent Trading. CryptoAsset Classifications. *Intelligent Trading*, September 2020. https://intelligenttrading.org/guides/cryptoasset-classifications/.

[285] Upmanyu Trivedi and Rahul Satija. Cryptocurrency Virtually Outlawed in India as Top Court Backs Ban. *Bloomberg*, July 2018.

[286] Unhashed. The Ultimate List of Bitcoin and Alt-Cryptocurrency Forks. *UNHASHED*, 2020. https://unhashed.com/bitcoin-cryptocurrency-forks-list/.

[287] Andrew Urquhart. What causes the attention of Bitcoin? *Economics Letters*, 166:40–44, May 2018.

[288] Franco Valencia, Alfonso Gmez-Espinosa, and Benjamn Valds-Aguirre. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy*, 21(6), June 2019.

[289] Radim Řehůřek. models.word2vec - Word2vec embeddings. *gensim: topic modelling for humans*, January 2019.

[290] Sha Wang and Jean-Philippe Vergne. Buzz Factor or Innovation Potential: What Explains Cryptocurrencies' Returns? *PLOS ONE*, 12(1):e0169556, January 2017.

[291] Frank Westerhoff. Anchoring and Psychological Barriers in Foreign Exchange Markets. *Journal of Behavioral Finance*, 4(2):65–70, June 2003.

[292] Bitcoin Wiki. Weaknesses. *Bitcoin Wiki*, July 2017. `https://en.bitcoin.it/wiki/Weaknesses#Attacker_has_a_lot_of_computing_power`.

[293] Christopher John Wild and George Arthur Frederick Seber. *Chance encounters: A First Course in Data Analysis and Inference*. Wiley, New York, New York, 2000.

[294] Oscar Williams-Grut. Everything you need to know about the complex relationship between Ripple and cryptocurrency XRP. *Business Insider*, March 2018.

[295] Josiah Wilmoth. Bitcoin Price Sets New All-Time High as Crypto Market Cap Nears $250 Billion. *CCN Markets*, November 2017. `https://www.ccn.com/bitcoin-price-sets-new-all-time-high-as-crypto-market-cap-nears-250-billion/`.

[296] Willy Woo. The network effects of volatility and liquidity, Bitcoin vs other payment coins. *Woobull*, December 2016. `http://woobull.com/the-network-effects-of-volatility-and-liquidity-bitcoin-vs-other-payment-coins/`.

[297] Peng Xie, Hailiang Chen, and Yu Jeffrey Hu. Network structure and predictive power of social media in the bitcoin market. *Georgia Georgia*

Tech Scheller College of Business Research Paper, No. 17-5, June 2018. http://dx.doi.org/10.2139/ssrn.2894089.

[298] Justin Xu and Dhruv Medarametla. Using Bitcoin Pricing Data to Create a Profitable Algorithmic Trading Strategy. 2017.

[299] Weichao Xu, Yunhe Hou, Y.S. Hung, and Yuexian Zou. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276, January 2013.

[300] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A Biterm Topic Model for Short Texts. In *22nd International World Wide Web Conference (WWW2013)*, pages 1445–1455, Rio de Janeiro, Brazil, May 2013. International World Wide Web Conference Committee (IW3C2).

[301] yogi. Terminology. *bitcointalk.org*, November 2012. https://bitcointalk.org/index.php?topic=126798.0.

[302] Jerrold H. Zar. Spearman rank correlation: Overview. *Wiley StatsRef: Statistics Reference Online*, 2014. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05964.

[303] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, November 2008.

[304] Wei Zhang, Pengfei Wang, Xiao Li, and Dehua Shen. Quantifying the cross-correlations between online searches and Bitcoin market. *Physica A: Statistical Mechanics and its Applications*, 509:657–672, November 2018.

[305] Ilya Zheludev, Robert Smith, and Tomaso Aste. When Can Social Media Lead Financial Markets? *Scientific Reports*, 4(1), May 2015.