University of Southampton
Faculty of Physical Sciences and Engineering
Electronics and Computer Science

Cryptocurrency Network Analysis on Twitter

by

Thanadon Fuengworatham
(tf2n17)

September 2018

Supervisor: Dr.Markus Brede
Second Examiner: Professor Timothy J Norman

A dissertation submitted in partial fulfilment of the degree of

MSc in Data Science

University of Southampton
Faculty of Physical Sciences and Engineering
Electronics and Computer Science

Cryptocurrency Network Analysis on Twitter

by

Thanadon Fuengworatham
(tf2n17)

September 2018

Supervisor: Dr.Markus Brede
Second Examiner: Professor Timothy J Norman

A dissertation submitted in partial fulfilment of the degree of

MSc in Data Science

Abstract

Incredible rise of cryptocurrency value leads to emerging of many actors in the market including new cryptocurrencies and investors. Social network, particularly Twitter, has been widely used by those actors to communicate, discuss, and express opinions about cryptocurrency. Despite this fact, no research has ever studied interactions between them in the social network. This lead to motivation that: Can we perform network analysis to extract insights from the trails of social interaction in Twitter?

In this dissertation, firstly, hashtag co-occurrence network was analysed to understand relationship between cryptocurrencies and social topics. This reveals that market ranking is related to how a cryptocurrency is discussed in the social network where higher-ranking cryptocurrencies tend to be mentioned in broad topics, while lower-ranking ones are specific only cryptocurrency-related topic. Secondly, network of active cryptocurrency users was analysed to study how users are connected based on their preference in cryptocurrencies. The finding shows that users who like Bitcoin and Ethereum are less likely to be interested in other cryptocurrencies, while users who like altcoins are more likely to be interested in various cryptocurrencies. Moreover, the user network exhibits homophily property where users who share the same preference are likely to be connected. Lastly, linear regression was performed to study influential factors in the social network. It reveals that users who tweets with focus on specific cryptocurrency topics are more likely to gain more social influence than users who tweet in diverse topics.

This dissertation presents a new dimension of cryptocurrency analysis which has never been done before in any related work. These results have proven that network analysis technique is capable for extracting insights from cryptocurrency social networks which could motivate further study on this direction.

Acknowledgements

I would like to express my appreciation to following people who make significant involvement on this dissertation.

Dr.Markus Brede, my supervisor, who has been giving me good advice through the period of this dissertation.

Professor Timothy J Norman, my second examiner, who helps comment and give suggestions to this dissertation.

Twitter who provide data used in this dissertation.

Researchers and literature referenced in this dissertation.

I am grateful to receive helps from all of them. This dissertation would not be possible without their good supports.

Statement of Originality

- I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.
- I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.
- I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

You must change the statements in the boxes if you do not agree with them.

We expect you to acknowledge all sources of information (e.g. ideas, algorithms, data) using citations. You must also put quotation marks around any sections of text that you have copied without paraphrasing. If any figures or tables have been taken or modified from another source, you must explain this in the caption and cite the original source.

I have acknowledged all sources, and identified any content taken from elsewhere.

If you have used any code (e.g. open-source code), reference designs, or similar resources that have been produced by anyone else, you must list them in the box below. In the report, you must explain what was used and how it relates to the work you have done.

I have used code and tools in following list.

- Tweepy API
- Python igraph package
- Gephi visualisation tool

You can consult with module teaching staff/demonstrators, but you should not show anyone else your work (this includes uploading your work to publicly-accessible repositories e.g. Github, unless expressly permitted by the module leader), or help them to do theirs. For individual

assignments, we expect you to work on your own. For group assignments, we expect that you work only with your allocated group. You must get permission in writing from the module teaching staff before you seek outside assistance, e.g. a proofreading service, and declare it here.

I did all the work myself, or with my allocated group, and have not helped anyone else.

We expect that you have not fabricated, modified or distorted any data, evidence, references, experimental results, or other material used or presented in the report. You must clearly describe your experiments and how the results were obtained, and include all data, source code and/or designs (either in the report, or submitted as a separate file) so that your results could be reproduced.

The material in the report is genuine, and I have included all my data/code/designs.

We expect that you have not previously submitted any part of this work for another assessment. You must get permission in writing from the module teaching staff before re-using any of your previously submitted work for this assessment.

I have not submitted any part of this work for another assessment.

If your work involved research/studies (including surveys) on human participants, their cells or data, or on animals, you must have been granted ethical approval before the work was carried out, and any experiments must have followed these requirements. You must give details of this in the report, and list the ethical approval reference number(s) in the box below.

My work did not involve human participants, their cells or data, or animals.

ECS Statement of Originality Template, updated August 2018, Alex Weddell aiofficer@ecs.soton.ac.uk

Contents

Abstract **i**

Acknowledgements **iii**

Statement of Originality **v**

1 Introduction **1**

1.1 Background and motivation 1

1.2 Aims and contributions 2

2 Related Work **4**

2.1 Overview of Social Network Analysis 4

2.2 Cryptocurrency Market Analysis 5

2.3 Bitcoin Transaction Analysis 6

2.4 Cryptocurrencies and Social Media 7

2.4.1 Sentimental Analysis 7

2.4.2 Text Mining and Machine Learning 8

2.4.3 Price and Topic Modelling 8

2.5 Research Gap Discussion 9

3 Dissertation Plan **11**

3.1 Contingency Plan 11

3.2 Time Management 11

4 Data Preparation **13**

4.1 Co-occurring Hashtag Data 13

4.1.1 Collecting Method 13

4.1.2 Cleaning Method 14

4.2 User Data 16

4.2.1 Collecting Method 16

4.2.2 Cleaning Method 17

5 Network Construction and Analysis Method **18**

5.1 Hashtag Co-occurrence Network 18

5.1.1 Network Explained 18

5.1.2 Edge Filtering 19

5.1.3 Community Detection 20

5.1.4 Cryptocurrency Hashtag Matching and Node Merging 21

5.1.5 Robustness Test for Hashtag Filtering 22

5.2 User Network 23

5.2.1 Network Explained 23

5.2.2 User Scoring 24

5.2.3 User Filtering 25

6 Experiments and Results **26**

6.1 Hashtag Co-occurrence Network Analysis 26

6.1.1 Relationship between Cryptocurrencies and Social Topics 26

6.1.2 Relationship between each Cryptocurrency 30

6.2 User Analysis 33

6.2.1 Correlation between User Preferences 34

6.2.2 Homophily Analysis 37

6.2.3 Influential Factor Analysis 41

6.3 Outliers Explained 44

7 Result Discussion and Future work **45**

7.1 Future Work and Limitation 46

8 Conclusion **47**

Bibliography **48**

List of Figures

1.1 Market capitalisation between September 2016 to July 2018 taken from coinmarketcap website [1]. 1

3.1 Contingency plan for handling unexpected events 11

3.2 Project plan. All objectives have been successfully achieved including the optional task which is the *Influential Factor Analysis* section in Experiment and Result chapter. 12

3.3 Project tracking dashboard for logging daily work. 12

4.1 Hashtags of names and abbreviations of top-30 cryptocurrencies which were used for searching tweets 14

4.2 ROC curve with 85% accuracy taken from Botometer literature [2]. 15

4.3 ROC curve with 79% accuracy from the test on collected data 15

4.4 Log-log histogram plot of hashtag frequency from the cleaned data. Dashed lines represent 3 thresholds that filter top 10000, 4000, and 1000 hashtags respectively 16

4.5 Example of bot active time taken from Botometer website [2]. Tweets were posted every hour for whole day. 17

5.1 Sample of hashtag co-occurrence network. Nodes and edges represent hashtags co-occurring in the same tweets. Colour distinguishes community 19

5.2 Example of hashtags in the detected communities. The community names represent topics that are reflected from the hashtags they contain. 20

5.3 Example of hashtags resulted from cryptocurrency hashtag matching method. . 21

5.4 Example of Unmerged network with Node A, B, C and D with edge weight equal to 1 22

5.5 Merging node A and C. The node degree and edge weight of node A and C are summed up. 22

5.6 Percentage that hashtags from communities in top-1000 are proportionate to communities in top-4000 dataset. 23

5.7 Percentage of hashtags from communities in top-4000 are proportionate to communities in top-10000 dataset. 23

5.8 User network structure. Node represents user, edge represents *'following'* relationship. Each user has score (0-1) indicating how much a user is active in each cryptocurrency and topic 24

5.9 Structure of user vectors consisting of top-30 cryptocurrencies, 5 topics and 1 other topics 25

6.1 Network of merged hashtags representing relationship between cryptocurrencies and topics. Five biggest nodes on the corners represent topics, while the orange and yellow nodes represent top-10 and top-11-to-30 cryptocurrencies respectively. 27

6.2 Heat map of log of edge weight in Figure 6.1. Dark green colours on the left shows the topics mostly co-occurred with the top cryptocurrencies and, generally, less co-occurred with lower-ranking ones 29

6.3 Heat map of normalised log of edge weight in Figure 6.1. The weights are normalised by the sum of all weights in their row. Dark green on the bottom left shows Bitcoin and Ethereum are strongly related to topics about buzzwords and advertisement 29

6.4 Cryptocurrencies hashtags network with 3 communities detected of which Modularity value is equal to 0.078. 31

6.5 Heat map of log of edge weight in Figure 6.4. Dark green around the corner shows strong relationship between higher-ranking cryptocurrencies. 32

6.6 Heat map of normalised log of edge weight in Figure 6.4. Green columns on the left show high-ranking cryptocurrencies strongly related to most cryptocurrencies. 32

6.7 Heat map of Spearman correlation coefficients between cryptocurrencies and topics. 34

6.8 Heat map of Kendall correlation coefficients between cryptocurrencies and topics. 34

6.9 Network of positive correlation coefficients from Spearman method. Three detected communities show strong correlation between the members. 36

6.10 Assortativity coefficients of user network based on each individual score. Colour presents ranking with green indicating rank 1-10 and topics, yellow indicating rank 11-20 and red indicating rank 21-30. Overall, top-10 cryptocurrencies and some crypto-related topics (in green) show higher values than the others meaning that they are more likely to be connected to others with similar behaviour. . . 38

6.11 Assortativity of each score with *p-value*. Most scores show homophily property with statistical significance at p-value <0.01. While only 27-AE and 28-ONT are not significant meaning that the way users tweeting about both cryptocurrencies are not correlated to the way they connect to each other. 39

6.12 Average euclidean similarity of user network with mean and standard deviation of results from null models. The test result shows statistical significance with p-value less than 0.01 40

6.14 Comparison of adjusted R-squared, P-value and AIC between 4 models in Figure 6.13. The combined model is indicated as the best model from highest adjusted R-squared and lowest AIC. 43

Chapter 1

Introduction

1.1 Background and motivation

Cryptocurrency is a digital currency, firstly introduced as Bitcoin in January 2009 by the unknown person with pseudonymous name Satoshi Nakamoto [3]. Blockchain is the core technology that applies cryptographic methodology to develop decentralised cash system. It allows users to create a reliable transaction without a need of a centralised system such as central bank. This capability led to some believes that cryptocurrencies would potentially disrupt the bank system in the future. Thanks to this potential, cryptocurrencies became mainstream and played important role in financial sector. Consequently, cryptocurrency market capitalization has been increasing and reached their peak at around $700 billion on 9 January 2018[4].

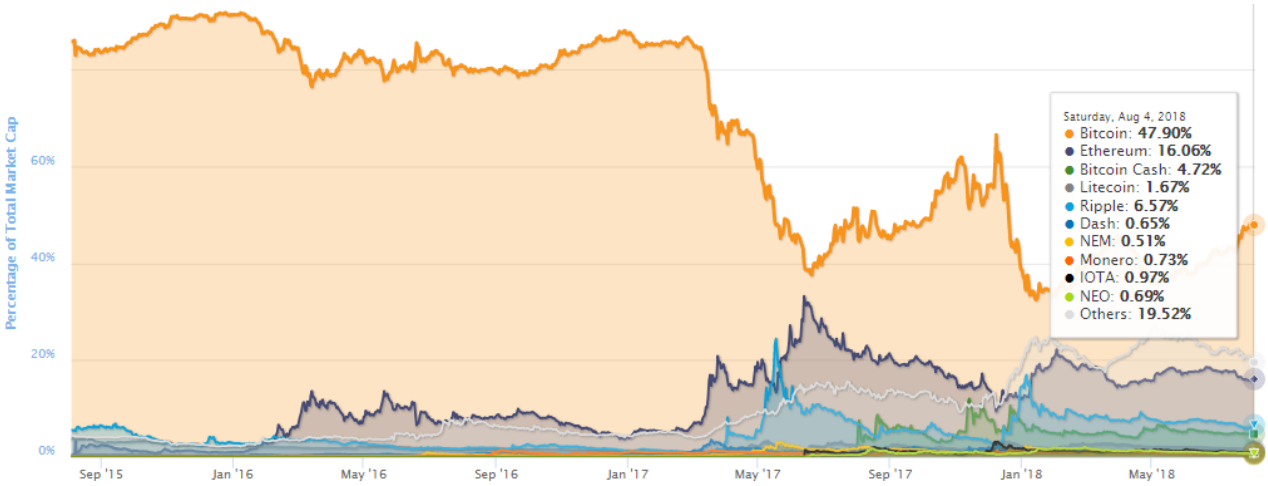


Figure 1.1: Market capitalisation between September 2016 to July 2018 taken from coinmarketcap website [1].

There have been many cryptocurrencies emerging since the first announcement of Bitcoin. As of 04 August 2018, there were around 1,700 cryptocurrencies currently in the market [1]. Most of them are generally cloned from the prominent cryptocurrency such as Bitcoin with minor modification on some features; however, some of them were invented with novel technologies which managed to take some market shares from Bitcoin. These innovative cryptocurrencies generally stayed on the top of the market and have their own community in cryptocurrency sector. According to Figure 1.1, it can be seen that although Bitcoin has been dominating more than 90% of the market shares, in recent years, other cryptocurrencies increasingly have more contributions to the market, with the reduction of Bitcoin market share to below 50%. This change shows that cryptocurrency sector is becoming more diverse. Thus, cryptocurrency is not only about Bitcoin any more; many new cryptocurrencies are emerging and playing more important role in the cryptocurrency market. Together with the fact that social network is widely used as a main media to distribute news and communication about cryptocurrencies. This leads to intriguing question: Can we use network analysis techniques to extract insights from users interaction in social network?.

To approach the question, social network analysis is used for this study. Although Twitter widely used as a main channel in cryptocurrency industry either for news distribution or discussion between users and communities, no research has ever studied interactions between those actors with the aim to understand their behaviour in the social network. Most research focused on price prediction from sentimental and topical analysis of Twitter contents. This leads to the motivation that the underlying social behaviour in cryptocurrency social network on Twitter would present a new aspect of cryptocurrency analysis and some insightful information could be revealed from social network analysis.

1.2 Aims and contributions

This dissertation aims to prove a hypothesis that network analysis can be used to extract insightful information related to cryptocurrency from interaction in the social network. We focus on understanding relationships between cryptocurrencies and social topics, user interactions, and social behaviours toward cryptocurrencies.

In this study, we selected 30 cryptocurrencies with top market capitalisation for analysis with the reason that the less prominent cryptocurrencies show less interactions in social network. Thus, they produce less data which might not be sufficient for this study with the limited time frame. Nevertheless, top-30 cryptocurrencies already gained 90% from all market capitalisations, as of 13 May 2018[1]. Therefore, the top-30 cryptocurrencies should be qualified to represent overall picture of social interaction related cryptocurrencies.

The analysis in this dissertation is divided into two main parts, which aims to answer the question in different aspects. First part is to study interaction between different cryptocurrencies in the social network. To achieve this, network of co-occurrence hashtag was built from the hashtags related to the cryptocurrencies. Then network analysis techniques were performed to extract relationship between cryptocurrencies and the underlying communities. Note that communities are represented by groups of hashtags that often co-occurred in the same tweet which will be elaborate in Network construction method chapter. The analysis from this part will answer following questions:

- What is the relationship between cryptocurrencies in social network and how is it related to social topics?
- How is cryptocurrency ranking related to interaction between cryptocurrencies and topics in the social network?

Second part aims to analyse interactions and relationships between users who are interested in different cryptocurrencies. Firstly, user vectors were built from their hashtags to present how much they are interested in each cryptocurrency and then the correlation between user preferences in specific cryptocurrencies are calculated to show how likely a user who likes a specific cryptocurrency, will also like another cryptocurrency. Secondly, homophily analysis was performed on the user network to show how likely that users who have the same cryptocurrency preference tend to connect to each other. Finally, user influence is analysed to study which factors or user behaviour are correlated to influence in the social network. At the end of this part, following questions will be answered:

- Which cryptocurrencies are strongly related in terms of attracting the similar group of users?
- If users interested in a specific cryptocurrency, which other cryptocurrencies they are likely to like?
- If users interested in a specific cryptocurrency, how likely their friends will also like it?
- How likely that users who share the same preference are connected together?
- What are the factors or tweeting behaviour that are correlated to user influence in the social network?

The results from this study would provide more understanding about cryptocurrency relationship and users behaviour in the social network. Consequently, this should make the market more transparent to the public and facilitate investors to efficiently analyse the market and create a better strategy for their investment.

Chapter 2

Related Work

This chapter is to present literature related to social network analysis and cryptocurrencies which currently exist in this research fields. With this knowledge, we aim to present to gaps that can be closed in this study.

2.1 Overview of Social Network Analysis

As discussed in the previous chapter, social network analysis was used in this dissertation to extract insights from interactions between the actors in the social network of cryptocurrencies. Social network analysis is an analysis technique that uses graph theory [5] to explain relationship and interaction between actors in the social network. To perform the analysis, a network is, firstly, constructed with nodes and edges representing actors and their relationships respectively. Then various network measures such as nodes degree, centrality and community detections techniques are used to describe relationship and interactions between actors in the network. Note that nodes can be any entities of interest such as users or hashtags.

Due to an emerge of online social network platforms, valuable information from the trails of users interaction in the platforms enables to the various application from social network analysis [6]. Amongst those important platforms, Twitter is the most popular one used as a data source for social network analysis. This is due to its functional design which is not limited to only a group of friends, but more focuses on public interactions from users across the social network. This leads to diverse research from general purposes such as a study of user influence from Twitter activity [7] or topical analysis on Twitter hashtags [8, 9], to very specific purposes such as sentimental analysis for stock market prediction [10] in which Dow Jones price movement was associated with moods reflected in related tweets. These examples demonstrate that Twitter data can be used to extract insights from user interactions in social

network for various purposes.

2.2 Cryptocurrency Market Analysis

Resent research from City, University of London[11] studied all cryptocurrencies in the market focusing on the evolution of market shares. The surprising result showed that the neutral model of evolution can be used to explain the behaviour of market shares movement over time. Firstly, market shares of top-5 cryptocurrencies were compared over the 4-year period (2013-2017) in line with the Spearman's correlation measures. This experiment exhibited a decrease trend of Bitcoin market share in the next 10 years with more 50% fluctuation over the period. Whilst the top-5 currencies are gaining more market share increasing to about 20% recently. Secondly, the plot between market shares and rankings exhibited power-law distribution shape. Thus, only a small group of top-ranking cryptocurrencies occupy a majority of the market shares. Next, the neutral model of evolution could be used to explain many underlying properties in the market such as the distribution of market share over years, the evolution of market share in Bitcoin and turnover in ranking. Finally, they also analysed whether underlying technologies (Proof-of-work and Proof-of-stake) affect the way people decide to invest in a cryptocurrency. The result showed no correlation between technology and the investment force. In conclusion, this research showed a significant contribution to all cryptocurrencies in the market and revealed that power laws and neutral model are capable to explain the evolution of cryptocurrency in the market.

Correlation of price movement between cryptocurrencies has been of interest in few studies. Research from Gandal et al. [12] studied competition between 7 important cryptocurrencies in the market, based on the movement of price returns correlation. The research started with a motivation that whether network effects in Bitcoin will continue and lead to a situation that Bitcoin, as the first cryptocurrency becomes more popular and finally take over the whole market. To test this hypothesis, correlation coefficients of price returns between pairs of the cryptocurrencies were calculated and analysed over 3 periods. The result illustrated that most of the periods, Bitcoin exhibited a winner-take-all effect on the market. However, in the end, the study cannot have a solid conclusion on this due to an uncertainty of the correlation over the period. It should be noted that, in this study, Ethereum, the second most influential cryptocurrency at the time of writing, was not invented yet. The result would have been more interesting if Ethereum was included.

Similarly, research from Burnie, A. [13] used correlation coefficients of price returns between cryptocurrencies to construct a correlation network, where nodes represent cryptocurrencies and edges represent correlation coefficients of connected nodes. The network was used to analyse how shared characteristics between nodes affect their correlation strength. The characteristics

taken into account included the fork status of cryptocurrencies, core technology, a supply mechanism, functionality etc. This research studied only top-10 important cryptocurrencies in terms of market cap and liquidity, an ability to be easily converted to cash. Both Spearman and Kendall correlation were calculated and compared to test the robustness of the result. In summary, external features, such as functionality and supply mechanism, were weakly correlated to the price changes. While some cryptocurrency pairs where one is a fork from the other showed a positive correlation. However, the significant level of the correlation was not sufficient to conclude that a specific factor led to co-movement of cryptocurrency prices.

2.3 Bitcoin Transaction Analysis

Another aspect related to user privacy has been studied by Fleder et al.[14]. They questioned whether Bitcoin account, with the anonymous property, can be de-anonymised by performing network analysis on Bitcoin transaction networks. To approach this question, they imitated criminal behaviour by crawling account number in online forums published by users who were not aware of their privacy. Usernames and account numbers were mapped and then used to build Bitcoin transaction network with nodes representing user accounts and edges representing transactions between users. To identify important nodes in the network, Page Rank algorithm [15] was applied which resulted in a criminal event on Bitcoin investigated by FBI being spotted as an outstanding activity. In line with the usernames crawled from the online forums, the study could identify that one of the users was closely related to this event. These results led to the conclusion that network analysis has a potential for spotting important event from Bitcoin transaction and, thus, Bitcoin account is not fully anonymous.

Not only user privacy, transaction network can be associated with the trading price. Two studies from John Mern et al.[16] and Kondor et al.[17] shared the similar motivation that singular values extracted by PCA, Principal component analysis, could be used to predict price changes in Bitcoin. The first research[16] approach this problem by using a machine learning model. They compared result from 3 neural network models separately trained from:

1) Network structural properties e.g., degree distribution, network diameters

2) Singular values produced from PCA

3) Embedded singular values generated by CNN auto-encoder, a deep learning model that compresses data to intense feature representation.

The result revealed that embedded singular values performed better than the ordinary singular values; however, it still could not beat the model from network properties. The second research[17] approached the problem in a different way. Firstly, they focused only the most

active users which were used to build the transaction networks. Then, for prediction, rather than using a machine learning model, they analyse the correlation between the change over time of the singular values and Bitcoin price. As a result, the singular value from the first principal component showed a positive correlation to the Bitcoin price. From both studies, it showed that singular values of transaction network could be used as features to predict Bitcoin price.

2.4 Cryptocurrencies and Social Media

2.4.1 Sentimental Analysis

Sentiments expressed in social media are often used to associate with the price movement. Research from Kaminski, J. [18] studied how sentiment toward Bitcoin in Twitter is related to price and volume changes in the market. They firstly defined 4 sets of keywords that represent 4 types of sentiment, positive, negative, both positive&negative, and uncertainty. Then they searched for the tweets that contained these keywords together with "Bitcoin" keyword. The number of bitcoin-related tweets in each sentiment type was associated with Bitcoin price and volume and then used to calculate Pearson correlation and test with Grange causality method. The study found that sentiment in the social network did not cause a change in price although it was significantly correlated. On the other hand, the causality test showed that the increase in uncertainty sentiment caused a decrease in Bitcoin volume. This study concluded that sentiment in the Twitter can reflect the situation in the market but it was still not clear that whether it caused price changes.

More practical work from Garcia, D. et al. [19] has been researched with the aim to create practical algorithmic trading strategies for Bitcoin by using wide range features from both economic factors and social signal. Many factors from social media related to Bitcoin term were used as signal consisting of Google search count, tweet count, emotion and opinion polarization from twitter. For prediction, Vector Auto-Regression (VAR), a time series model for multidimensional features which is rearranged in vector form, was applied to predict changes in price returns from both financial and social signal which vary over time. The model aimed to detect signal from real-time social activities and economic factors that precede the changes in Bitcoin returns. The study found that opinion polarization in tweets preceded the increase in Bitcoin price and volume. Thus, using social signals for algorithmic trading has the potential to make a profit in the Bitcoin market.

Although the results from the studies above could not strongly conclude which specific social signal caused changes in price, they clearly showed a strong connection between cryptocurrencies market and user interaction in social media which is worth noting.

2.4.2 Text Mining and Machine Learning

Text mining and machine learning techniques have been extensively used to predict cryptocurrencies price from social network data. A study from Colianni, S. et al. [20] performed text mining process on tweets containing Bitcoin keyword to extract sentiment vectors and bag of words features [20]. The features then were labeled with price trend, up or down, at the time of tweet posted in hourly and daily basis. Three different algorithms, Logistic Regression, SVM and Naive Bayes were used to train models and compared to find the most accurate model for Bitcoin price prediction. As a result, a bag of words features with Naive Bayes model outperformed the other models in daily basis prediction with 95.00% accuracy. While sentiment vectors with logistic regression performed best in hourly basis with 98.58% accuracy.

More extensive use of data sources was studied in the research from Lamon, C. et al. [21]. Not only tweets but also headlines from news site were scrapped to use to predict three cryptocurrencies including Bitcoin, Ethereum and Litecoin. They proposed an interesting way to label the text data. Rather than labeling text by sentiment polar, positive and negative, they directly labeled the text by price changes in one-day ahead. This method allowed the machine learning model to learn price trend directly from the text vector, rather than transforming to sentiment then price. As a result, words that strongly contributed to the price changes could be extracted from the models. For example, break, continue and spike, showed a strong correlation to the price increase of the next day, while begin, bitcoin hard showed a high correlation to the opposite trend. Thus, not only this method was able to predict the price fluctuation from the social expressions but also presented words that influenced the price changes.

2.4.3 Price and Topic Modelling

Apart from machine learning related approaches, some analysis methods from other fields have been used to study the relationship between social interaction and cryptocurrencies price. Very recent research, in 2018, from Phillips, R. C. et al. [22] applied wavelet coherence analysis technique to study why the correlation between the market price and the number of post in online forums varies over time. From the study, coherence analysis can well explain the variation of the correlation. It revealed an interesting finding that when the market becomes a bubble-like state, where the price changes rapidly, the positive correlation between forum posts and price became significantly stronger. In contrast, some special events or news, that impact the market price, can cause just inconsistently change of correlation which makes the market hard to be predicted.

Another work from Phillips, R. C. et al. [23] applied Hidden Markov Model, which was designed for influenza detection, to predict the price changes. With the motivation that the shape of

the evolution of epidemic spread and price bubbles are similar, they use number of Reddit [24] posts in a group of a specific cryptocurrency and some other features to predict *boom phase*, where the price starts rising, of financial bubble event in the market. The entry and exit of epidemic state, presented by hump shape in the graph, were used to represent the boom and panic phase of the bubble event. These transition phases were then used as buy and sell signals and tested with a traditional trading agent. As a result, the epidemic model can outperform the traditional algorithm as it only traded at strong signals and minimise transaction fees, while traditional algorithm traded straight away based on the up and down signal, which may not be effective, although correctly predicted when the price increased the only small amount.

Apart from the social factors studied above, topics discussed in social media also have potential to reflect the market situation and predict price movement. As studied by Linton, M. et al. [25], Dynamic topic modeling technique was used to detect an important event in the cryptocurrency industry. The research found that topics discussed in online forums were relevant to the important events in the same period. This demonstrates there is a strong connection between social conversation and events in the real world. Another work from Phillips, R. C. et al. [26] intensified this finding from the successful application of topic modeling to identify price change in the market. The research started by using Dynamic topic modeling to extract topics from online forums. Then, the volume of the topics discussed over time was associated with the price fluctuation. To analyse the co-movement between price and topics, Hawkes model was used to study the relationship. The result of this analysis indicated that the rise of discussion in topic *Risk and investment* preceded price fall event, while topic about *Fundamental of cryptocurrency value* preceded the increase in price return. These findings clearly showed that topics discussed in the social media could be used as a signal to predict trends in the market.

2.5 Research Gap Discussion

As reviewed in the previous sections, cryptocurrencies have been researched in many aspects both economic and technical fields, while most of the studies focused on price prediction. Social interaction has been widely used to associate with price movement in the cryptocurrencies market. As proven by the literature discussed in the last sections, social media have a strong relationship to the market. Most of the works used sentiment analysis techniques to extract social sentiment for price prediction. The techniques were successfully used to analyse sentiment in Twitter [18, 20, 27], online forums [28, 19] and new headlines [21] together with predictive model to predict price fluctuation. Apart from sentimental analysis, research from [25, 26] proved that topics discussed in the social network are important features that can well explain the market situation and price movement.

Some literature [21, 22, 23, 26] focused on broader analysis by analysing multiple cryptocur-

rencies. However, only few prominent cryptocurrencies were analysed in most research. Only literature [11], studied all existing cryptocurrencies in the market but it was limited to only market share analysis. The literature [13] has studied the relationship between important cryptocurrencies in the market, but only 10 cryptocurrencies were limited in this research and also the study only focused on the correlation of price movement.

Taking all of these into account, there is still a big gap for further research to understand cryptocurrency relationship and social interaction toward each cryptocurrency as well as user behaviour in the market. This dissertation aims to fill this gap by analysing top-30 cryptocurrencies with 90% of the whole market capitalization as of 14 May 2018 [1]. Social network analysis technique is used to study interactions between the cryptocurrencies in Twitter, as proven in [19, 27, 28] that Twitter has a strong connection to cryptocurrency market.

Chapter 3

Dissertation Plan

3.1 Contingency Plan

To handle unexpected events that might happen during the dissertation period, a contingency plan was prepared as shown in Figure 3.1.

Risk	Impact	Mitigation Plan
Data lost or file corrupted	Could not work on project	• Regularly backup data on One Drive provided by the university • Regularly commit codes to GITHUB • Use overleaf for writing report which provide back-up feature and online accessible
University closed	Could not contact professor face-to-face	• Commit project scope at the early stage of dissertation • Contact via email or Skype if necessary
Twitter API out of service	Could not use Twitter data for analysis	• Put data collection process as first priority and start at early stage. • Change data source to other social media such as Reddit or other online forums

Figure 3.1: Contingency plan for handling unexpected events

3.2 Time Management

This dissertation was planned as Figure 3.2 before the dissertation period. It has been used as a guideline to keep work on the track. First of all, we started the project almost a month earlier as we were aware of Twitter limitation that can collect data only back to a week period. We planned beforehand that we needed data around at least a month period and, thus, we needed that early starts to collect the amount of data as planned. After the data collection period, we have been working by following this plan. In general, all steps have been successfully achieved on plan with, in total, 8 regular meetings I have with my supervisor through the dissertation period. Additionally, the optional work, No. 8. on Figure 3.2, has also been successfully

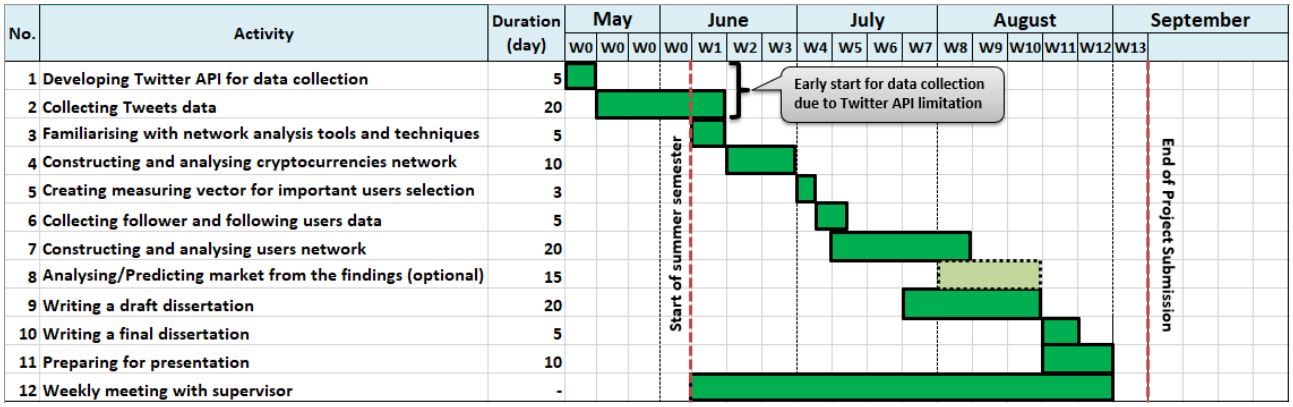
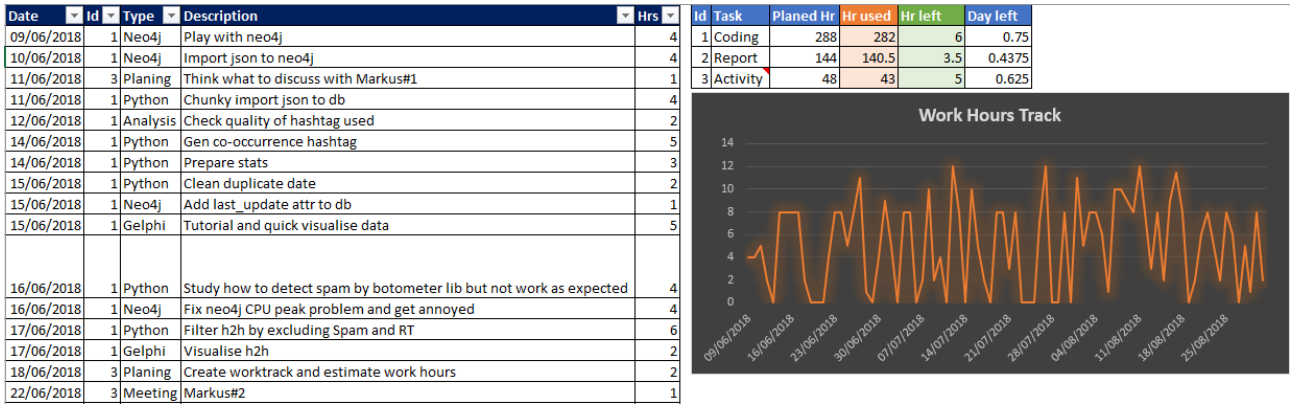


Figure 3.2: Project plan. All objectives have been successfully achieved including the optional task which is the *Influential Factor Analysis* section in Experiment and Result chapter.



Date	Id	Type	Description	Hrs
09/06/2018	1	Neo4j	Play with neo4j	4
10/06/2018	1	Neo4j	Import json to neo4j	4
11/06/2018	3	Planing	Think what to discuss with Markus#1	1
11/06/2018	1	Python	Chunky import json to db	4
12/06/2018	1	Analysis	Check quality of hashtag used	2
14/06/2018	1	Python	Gen co-occurrence hashtag	5
14/06/2018	1	Python	Prepare stats	3
15/06/2018	1	Python	Clean duplicate date	2
15/06/2018	1	Neo4j	Add last_update attr to db	1
15/06/2018	1	Gelphi	Tutorial and quick visualise data	5
16/06/2018	1	Python	Study how to detect spam by botometer lib but not work as expected	4
16/06/2018	1	Neo4j	Fix neo4j CPU peak problem and get annoyed	4
17/06/2018	1	Python	Filter h2h by excluding Spam and RT	6
17/06/2018	1	Gelphi	Visualise h2h	2
18/06/2018	3	Planing	Create worktrack and estimate work hours	2
22/06/2018	3	Meeting	Markus#2	1

Id	Task	Planed Hr	Hr used	Hr left	Day left
1	Coding	288	282	6	0.75
2	Report	144	140.5	3.5	0.4375
3	Activity	48	43	5	0.625

Figure 3.3: Project tracking dashboard for logging daily work.

achieved which is the *influential factor analysis* experiment in the last section of *experiment and result* chapter.

During the working period, project tracking dashboard in Figure 3.3 was created in MS Excel for tracking work hours and summary of daily work. Firstly, 3 types of work, coding, reporting and project activities such as meeting, were designed with efforts estimated as shown on the upper-right of Figure 3.3. Next, detail and progress of everyday work have been logged on the left side. Lastly, the graph on the right shows working hours which have been tracked for monitoring work consistency through the period of this dissertation. As a result, this dashboard has been used as a reminder and work guideline for us to stay on track as planned.

Chapter 4

Data Preparation

In this dissertation, there are 2 datasets, co-occurring hashtag, and user data, prepared separately based on two main analyses of hashtag co-occurrence network and user network. Firstly, tweets related to cryptocurrencies were collected of which containing hashtags were used to build a hashtag co-occurrence network. Secondly, top-1000 users with most tweets from the first dataset were used as initial users of which followings and followers were collected to construct a user network for the user analysis. The details of collecting and cleaning steps are explained in each section.

4.1 Co-occurring Hashtag Data

4.1.1 Collecting Method

As discussed in the last chapter, there are 30 top-ranking cryptocurrencies studied in this dissertation. Firstly, to collect the data from Twitter, hashtags that represent each of the cryptocurrencies were defined in Figure 4.1. These hashtags were taken from the names and abbreviations that were used in the market as of 13 May 2018. After that Tweepy [29], a wrapper of Twitter API written in Python, was used to search for the tweets that contain the cryptocurrencies hashtags. The collection period lasted for 45 days, from 13 May to 26 June 2018. As a result, around 6,000,000 tweets and 600,000 users related to the cryptocurrencies were collected. After the cleaning process, around 4,000 hashtags with 310,000 links between them were extracted from the cleaned tweets and then used to build the hashtag co-occurrence network which is detailed in the next chapter.

#	Name	Abbr	#	Name	Abbr	#	Name	Abbr
1	bitcoin	btc	11	neo	-	21	zcash	zec
2	ethereum	eth, ether	12	monero	xmr	22	icx	-
3	ripple	xrp	13	dash	-	23	omisego	omg
4	bitcoincash	bch	14	nem	xem	24	lisk	lsk
5	eosio	eos	15	tether	usdt	25	zilliqa	zil
6	litecoin	ltc	16	vechain	ven	26	bitcoingold	btg
7	cardano	ada	17	ethereumclassic	etc	27	aeternity	ae
8	stellar	xlm	18	bytecoin	bcn	28	ontology	ont
9	iota	miota	19	binancecoin	bnb	29	verge	xvg
10	tron	trx	20	qtum	-	30	steem	-

Figure 4.1: Hashtags of names and abbreviations of top-30 cryptocurrencies which were used for searching tweets

4.1.2 Cleaning Method

Data cleaning process is an essential part of this dissertation. To ensure the quality of the research, irrelevant data and noises must be reduced to keep only data that suits the aims of the study. Follows are the steps that were used to clean this dataset.

1) Lowercase and remove duplicate hashtags from each tweet

Hashtags that have the same meaning might be in different cases. To unify such hashtags, the lower case transformation is applied to all hashtags in the dataset. In addition, duplicate hashtags that exist in the same tweet were consolidated into one. This is to reduce noises as duplicate hashtags will become duplicate pair when building co-occurring hashtag pair.

2) Excluding tweets posted by bot users

The main problem caused by bot users is that they generally post tons of tweets for advertising purposes. These tweets are mostly repetitive and meaningless which could be considered as noises for this study. With this reason, tweets from bot users were excluded from the data.

Botomenter [2], a Python library for bot detection in Twitter, was utilised to detect bots in this dataset. The library uses a pre-trained machine learning model to classify whether a user is bot. More than a thousand features from 6 feature types were used to train the model i.e., user profile, user friends, network pattern, content, sentiment and timing activity.

As specified in the literature [2], the model could 85% correctly classify bot from unseen data. To verify the result, we followed the evaluation process in the literature and tested it on this dataset. Firstly, 100 user samples were selected and manually labeled based on user content

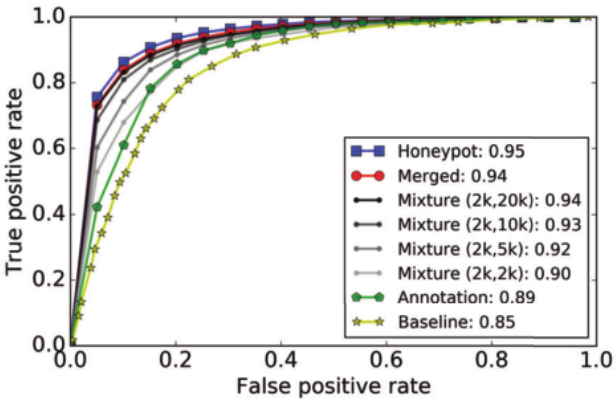


Figure 4.2: ROC curve with 85% accuracy taken from Botometer literature [2].

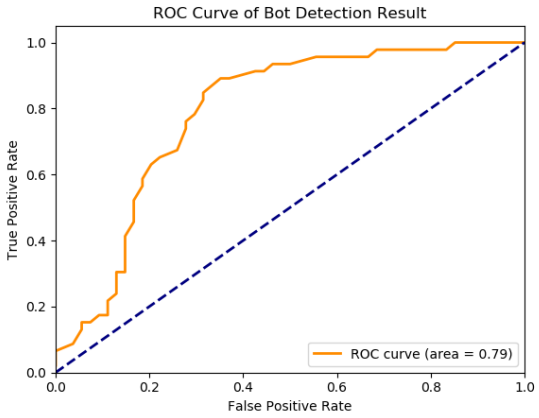


Figure 4.3: ROC curve with 79% accuracy from the test on collected data

and active time. That is the users who repeatedly tweeted the same content or have been active more than 20 hours were labeled as a bot. The test result was then used to plot the ROC curve and compared with the result from the literature.

According to Figure 4.2 and Figure 4.3, the result tested on this data showed 79% accuracy which is fairly consistent with 85% accuracy specified in the literature. With this small error compared to plenty of time that can be saved from this utilisation, the library was used to detect bot users in this data. Nevertheless, there is a potential that misclassified bots that produced high tweets volume might generate significant noise to the data. To ensure the quality of tweets in this data, users with more than 300 tweets, were manually checked. As a result of this cleaning, the non-sense bot tweets which do not truly reflect social interaction should mostly be eliminated.

3) Excluding retweet and tweets with non-English hashtags

In Twitter, a retweet is a tweet which is a clone of the other. Specifically, all retweets from the same original tweet are identical. Hence, this step is to simply remove duplicate tweets from the data. In addition, to make the analysis interpretable, tweets which contained non-English hashtags were excluded.

4) Excluding less frequent hashtags

In this step, the less frequent hashtags as presented on the left part of the histogram plot in Figure 4.4, were removed. With regards to the right-skewed shape of the distribution plot, it can be interpreted that most of the hashtags have been used only few times, while some small amount was widely used in the social network. In other words, the high frequent hashtags on

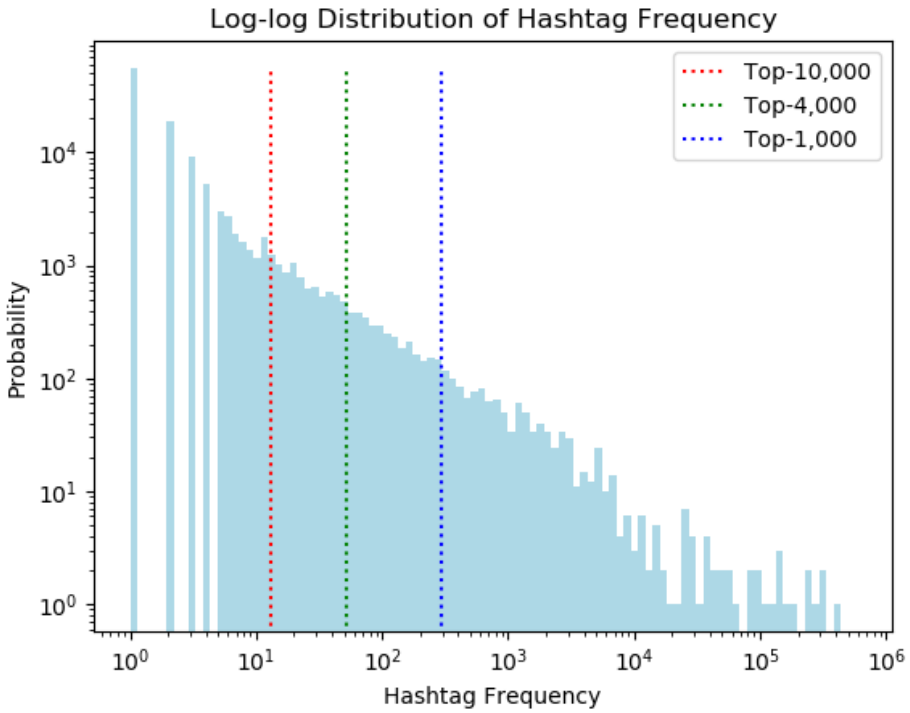


Figure 4.4: Log-log histogram plot of hashtag frequency from the cleaned data. Dashed lines represent 3 thresholds that filter top 10000, 4000, and 1000 hashtags respectively

the long tail could be used to represent the majority of social interactions in cryptocurrencies social network.

To choose an optimal threshold for hashtag filtering, we performed a robustness test on three datasets filtered by three thresholds in Figure 4.4. As a result, top-4000 is selected as it produced the most robust result. The detail of robustness test is described in method section.

4.2 User Data

4.2.1 Collecting Method

The objective of this dataset is for constructing a network of users who are active about cryptocurrencies. Based on assumption that users who are connected to the base users are willing to receive news and updates related to the top-30 cryptocurrencies, top-1000 users with the highest number of tweets were selected from the last dataset and used as base users to collect other users who are connected to them. In other words, these connected users are presumed to be interested in the top-30 cryptocurrencies. Thus, in the first step, users who were following or followed by the base users were collected together with their relationship and used as nodes and edges in the user network. Next, tweets posted by each user were retrieved to use for user

scoring which will be specified in the next chapter.

In summary, three types of data were collected. First, followings and followers of the base users. Second, the relationship between the users, namely, whether a user is following or followed by the other. Third, tweets of the users.

4.2.2 Cleaning Method

For an efficient study of user interaction, bot users are required to be excluded from this dataset, however, due to the usage limit of Botometer API [2] and time constraint of this dissertation, the library is not practical for this dataset. To achieve this task in this limited time frame, we programmed bot detecting constraints based on the bot statistics in Figure 4.5, acquired from the last section. The constraints are defined based on active time and tweet volume. Follows are the rule for bot classification.

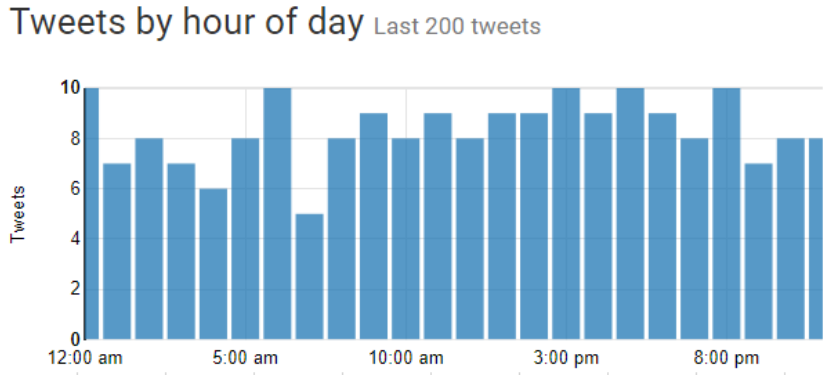


Figure 4.5: Example of bot active time taken from Botometer website [2]. Tweets were posted every hour for whole day.

1) Users who are continuously active more than 20 hours a day are classified as bot. From Figure 4.5, bot users are continuously active for the whole day without sleeping. The minimum rate of sleeping hours is 4 hours. Thus, users who are active more than 20 hours are considered as bots.

2) Users who tweet more than 78 tweets a day are classified as bots. This number is taken from the statistic of bot cleaning in the last section where the maximum number of the tweet posted by human users is around 3,500 within 45 days, which is around 78 tweets per day.

Chapter 5

Network Construction and Analysis Method

This chapter is divided into two sections according to the networks studied in this dissertation. Overall, methods that were used to construct, analyse and test the networks are described in detail.

5.1 Hashtag Co-occurrence Network

5.1.1 Network Explained

This network presents how cryptocurrencies are mentioned in various topics represented by a group of related hashtags. The network is constructed from hashtags that co-occurred with top-30 cryptocurrencies hashtags specified in Figure 4.1. As presented by the sample network in Figure 5.1, nodes and edges represent hashtags and co-occurrence relationship respectively. While The thickness or weight of the edges show how often they co-occurred in the same tweets. That is the thicker a link is, the more frequent they co-occurred. Node size presents weighted degree, which is a measure of how many connections it has to other nodes weighted by the edge weight. The bigger a node is, the more connections it has to other nodes with weight taken into account. Lastly, all network visualisations in this dissertation were created by Gephi [30], a graph visualisation, and network analysis tool.

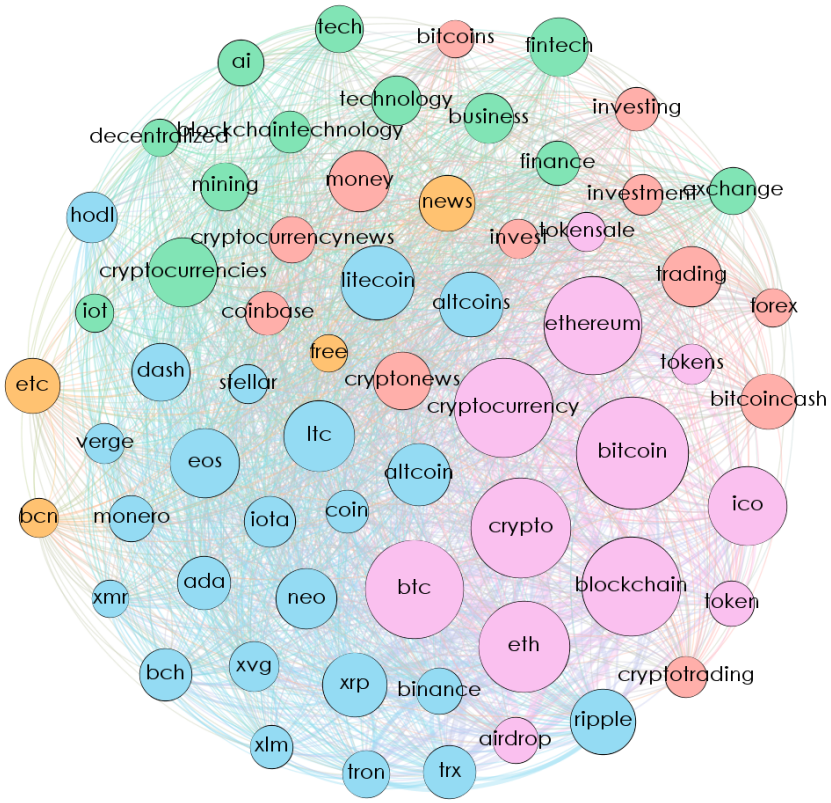


Figure 5.1: Sample of hashtag co-occurrence network. Nodes and edges represent hashtags co-occurring in the same tweets. Colour distinguishes community

5.1.2 Edge Filtering

This process is to remove edges of hashtags that are considered as co-occurring by chance. Hashtags that are widely used in Twitter are more likely to co-occur with others than the less popular ones. Such a case can generate bias in which the popular hashtags always gain more connection to others. These would affect results of network analysis. To avoid such a noise, the edges of the hashtags that co-occurred by chance were removed before performing network analysis.

$$P(A \cap B) = P(A)P(B) \tag{5.1}$$

To detect such edges, *Independence events rule* from probability theory [31] was applied. According to the theory, the probability that two independent events occur together is equal to the product of the probabilities for each event as presented in equation (5.1). This can be applied to hashtag co-occurrence events where the probability that two independent hashtags co-occur is equal to the product of the probabilities that each hashtag occurs. In other words, if the

Crypto News	Crypto Buzzword&Ads	Business&Tech	Trading	General
altcoins	crypto	cryptocurrencies	money	news
altcoin	cryptocurrency	fintech	trading	usa
cryptonews	blockchain	technology	investing	love
hodl	ico	business	cryptotrading	entrepreneur
cryptocurrencynews	token	tech	invest	worldcup
exchange	airdrop	finance	investment	follow
coinbase	coin	mining	forex	rt
consensus2018	tokens	blockchaintechnology	market	success
bittrex	free	ai	trade	music
cryptoexchange	tokensale	iot	stocks	art

Figure 5.2: Example of hashtags in the detected communities. The community names represent topics that are reflected from the hashtags they contain.

probability of the hashtag co-occurrence, $P(A \cap B)$, is less than the product of each individual probability, $P(A)P(B)$, then the hashtags are considered to co-occur by chance.

From the equation, $P(A \cap B)$ is calculated from the frequency of hashtag co-occurrence divided by the total number of tweets. While $P(A)$ and $P(B)$, the probability that hashtag A or B occurs in a tweet are calculated from the number of hashtag A or B divided by the total number of tweets. As a result, approximately 10,000 from 320,000 edges were removed, which resulted in around 310,000 edges remaining.

5.1.3 Community Detection

Community detection is a network analysis technique that is used to cluster nodes into groups where nodes that are strongly related belong to the same group, so-called *Community*. There are a few algorithms invented for community detection. In this dissertation, Fast unfolding algorithm from Blondel, V. D. et al. [32] was used due to its efficiency. As evaluated in the literature, it is faster than other methods and also produces a decent *Modularity* value, a measure of connection strength between nodes in the same group.

In hashtag co-occurrence network, related hashtags are formed into a community which is considered as a topic. As a result of community detection, there were five communities detected. Each of them is named based on topics reflected from the hashtags it contains. Figure 5.2 shows the example of hashtags that belongs to each community. Follows are the description of each community.

- **Crypto News**: contains hashtags related to news about cryptocurrency market.
- **Crypto Buzzword&Ads**: contains buzzwords about cryptocurrency such as blockchain, token and ico. Also, some words convey advertisement meaning such as free, tokensale and airdrop - an activity to give away cryptocurrency.

- **Business&Tech**: contains hashtags related to business and technology such as business, finance, ai and iot.

- **Trading**: contains hashtags related to trading such as trading, invest, forex and stocks.

- **General**: contains hashtags related to news about the cryptocurrency market.

5.1.4 Cryptocurrency Hashtag Matching and Node Merging

Once communities were detected, all hashtags under a community were merged into a supernode representing a topic. Similarly, hashtags that represent the same cryptocurrency should be merged into a supernode that represents specific cryptocurrencies. For example, bitcoin, bitcoins, btc and btc_news nodes should be merged into one bitcoin node.

To identify the representative hashtags, we used cryptocurrency names and abbreviations defined in Figure 4.1 as keywords to search for hashtags that contain these keywords. The matching hashtags were manually verified again to exclude some irrelevant hashtags before being used to represent the cryptocurrencies. Figure 5.3 shows an example of the listed hashtags.

Once all hashtags were identified, the hashtags in the community or cryptocurrency were then merged into one supernode. When nodes are merged the node degree and edge weight are summed up. As shown in Figure 5.4 and 5.5, node A and C are merged into the supernode AC. As a result, the node size which represents the weighted degree become double as seen from the bigger node, while the edge weight is summed up to 2.

Bitcoin	**Ethereum**	**Ripple**	**Bitcoincash**	**Eosio**	**Litecoin**
bitcoin	ethereum	ripple	bitcoincash	eosio	litecoin
btc	eth	ripple	bchusd	eos	litecoin
btcusd	ether	ripplenet	bitcoin_cash	eosio	litecoinfam
hitbtc	ethereum	xrp	bch	eosdac	paywithlitecoin
btcnews	myetherwallet	xrpusd	bchpls	eosico	litecoinnation
freebtc	etherum	xrpthestandard	bchforeveryone	eosusd	litecoincash
...	...	...	...	...	...

Figure 5.3: Example of hashtags resulted from cryptocurrency hashtag matching method.

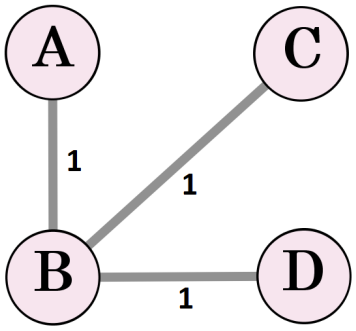


Figure 5.4: Example of Unmerged network with Node A, B, C and D with edge weight equal to 1

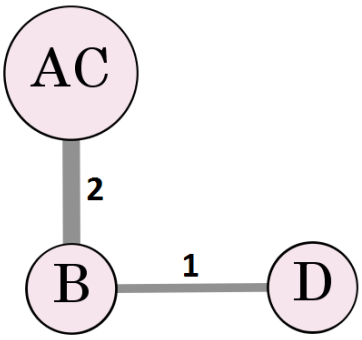


Figure 5.5: Merging node A and C. The node degree and edge weight of node A and C are summed up.

5.1.5 Robustness Test for Hashtag Filtering

As discussed in the data preparation chapter, a robustness test is performed on data generated from 3 thresholds - lower mean(13), mean(52) and upper mean(292). As shown in Figure 4.4, the thresholds result in top-10,000, top-4,000 and top-1,000 most frequent hashtags respectively. To test the robustness of the datasets, we performed community detection and compared the results how hashtags from communities in smaller dataset distribute to the communities in the larger dataset. The dataset that gives the most robust result is selected for analysis. Figure 5.6 and 5.7 show results from the comparison between the proportion of hashtags from smaller to a larger dataset.

Top-1000 vs Top-4000

According to Figure 5.6, there are two findings indicating that top-4000 dataset is a good representative of the hashtag population. First, all communities in the top-1000 dataset exist in top-4000 dataset and also most hashtags are highly proportionate to the corresponding communities around 80-90%. This result shows that although 3000 additional hashtags are included in top-4000 dataset, the communities can still maintain most hashtags found in the smaller dataset.

Second, General&Trading that is ambiguous in top-1000 are clearly detected as two separate communities in top-4000 dataset. This demonstrates that top-4000 can produce more precise community. With respect to two findings above, it can be concluded that top-4000 hashtags can produce clearer topics found in tweets than top-1000 hashtags.

Top-4000 vs Top-10000

According to Figure 5.7, the number of community in top-10000 dataset increases from 5 to 8; however, the new communities are very small and cannot be identified from their presented

Top-1,000 \ Top-4,000	Crypto News	Crypto Buzzword & Ads	General	Trading	Business & Tech
Crypto News	91%	0%	3%	4%	3%
Crypto Buzzword & Ads	3%	87%	4%	2%	5%
General & Trading	0%	4%	39%	43%	14%
Business & Tech	1%	3%	4%	4%	89%

Figure 5.6: Percentage that hashtags from communities in top-1000 are proportionate to communities in top-4000 dataset.

Top-4,000 \ Top-10,000	Crypto News	Crypto Buzzword & Ads	General	Trading	Business & Tech	N/A	N/A	N/A
Crypto News	87%	2%	1%	11%	1%	0%	0%	0%
Crypto Buzzword & Ads	4%	84%	3%	4%	4%	1%	0%	0%
General	1%	14%	80%	2%	3%	0%	0%	0%
Trading	1%	4%	3%	91%	1%	0%	0%	0%
Business & Tech	1%	4%	2%	18%	74%	0%	0%	0%

Figure 5.7: Percentage of hashtags from communities in top-4000 are proportionate to communities in top-10000 dataset.

hashtag. This can be implied that some noises are added in the larger set. Next, considering the rest communities in top-10000, they present the same communities as existed in the smaller set with approximately 80-90% of corresponding hashtags belonging to the same communities. This shows the robustness of the communities detected in top-4000 dataset where most of belonging hashtags stay in the same communities even though the sample size is as twice smaller as top-10000 dataset.

Summary

According to the comparisons above, it can be concluded that top-4000 dataset is the most suitable dataset. Not only does it present sensible and precise communities, but also show robustness where the high proportion of hashtags are maintained in the corresponding communities in the larger datasets with less noise presented.

5.2 User Network

5.2.1 Network Explained

User network is a group of users on Twitter who are very active about cryptocurrencies. In this study, top-1000 users who most frequently tweeted about cryptocurrencies and their connections were used to construct the network. This is due to the motivation that users who frequently tweeted about cryptocurrencies act like distributors of cryptocurrency-related contents, while users who are connected to this group are interested to receive the contents from them. For this reason, the top-1000 users and their connections are used to represent of cryptocurrency social network.

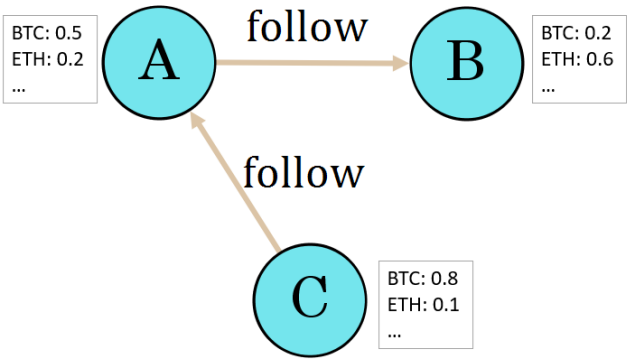


Figure 5.8: User network structure. Node represents user, edge represents *'following'* relationship. Each user has score (0-1) indicating how much a user is active in each cryptocurrency and topic

According to the Figure 5.8, nodes and edges in this network represent users and *following* relationship respectively. Each user has a score (0-1) indicating that how much they are active about each cryptocurrency and topic. All nodes have the same size because this network is for Homophily analysis where node measure is not required. Similarly, all edge weights are equal to 1 as *following* relationship has no quantity.

5.2.2 User Scoring

User vectors are required to understand the relationship between users and cryptocurrencies. The vectors indicating how much a user actively tweeted about specific cryptocurrencies or topics. As presented in Figure 5.9, in total, there are 36 scores contained in a vector. First 30 scores are from the top-30 cryptocurrencies, 5 scores from the detected and 1 additional score from other topics apart from cryptocurrencies and detected topics.

$$UserScore = P(Hashtag_{relevant}) = \frac{\#Hashtag_{relevant}}{\#Hashtag_{total}} \tag{5.2}$$

A user score is calculated from a probability that hashtags related to a cryptocurrency or topic are used by this user. According to equation (5.2), a user score is equal to the number of the relevant hashtags used by the user divided by the number of total hashtags from the user. For example, 5 out of 10 hashtags used by a user are on the list of hashtags representing Bitcoin. As a result, the user's Bitcoin score is equal to 5 divided by 10 which is 0.5. Thus, to calculate a user vector, this process is repeated 36 times for each score in a vector.

User Vector (Length: 36)		
Top-30 Cryptocurrencies (30)	**5 Detected Topics (5)**	**Other topics (1)**
1-BTC, 2-ETH, ... , 30-STEEM	Crypto-News, Crypto-Buzzword&Ads, Business&Tech, Trading, General	Others

Figure 5.9: Structure of user vectors consisting of top-30 cryptocurrencies, 5 topics and 1 other topics

5.2.3 User Filtering

Due to the aim to study the network of users who are active about cryptocurrencies together with limitation from Twitter API, a filter was applied to select only the most suitable data for this study under the time constraint of this dissertation. Follows condition was applied to filter users.

1) When collecting followers and friends of the initial top-1000 users, only users with at least 10 links to the initial users are selected. This is to select only users who are strongly connected to the initial users.

2) Keep only users with sum of all top-30 cryptocurrency scores, Crypto-News, and Crypto-Buzz&Ads above average. This is to filter only users who are active in cryptocurrencies more than average.

3) Keep only users with sum of General and Other scores below average. This is to filter out users who are mostly active in general topics.

Chapter 6

Experiments and Results

This chapter is to describe experiments in this dissertation. By design, the experiments are divided into two main parts. First, hashtag co-occurrence network is analysed to understand how cryptocurrencies are related to each other and to topics discussed in Twitter. Second, the network of users who are active about cryptocurrencies are analysed to understand the relationship between their connections based on their preferences, represented by the user vectors. Lastly, user influence, measured by the number of retweets, is analysed to understand which factors are correlated to user influential in cryptocurrency social network. Note that the relationship and preference presented in these experiments are measured based on the number of hashtags related to cryptocurrencies or topics as discussed in the previous chapter.

6.1 Hashtag Co-occurrence Network Analysis

6.1.1 Relationship between Cryptocurrencies and Social Topics

In this experiment, this hashtag co-occurrence network is used to study the relationship between cryptocurrencies and the social topics in the sense that how each cryptocurrency is related to each topic and why. We aim to answer the following questions:

- Which cryptocurrencies are strongly related to which topics?
- How is cryptocurrency ranking related to the topics they were mentioned on?

To set up the experiment, the hashtag co-occurrence network is constructed and then community detection is applied to identify topics related to the cryptocurrencies. Then hashtags in

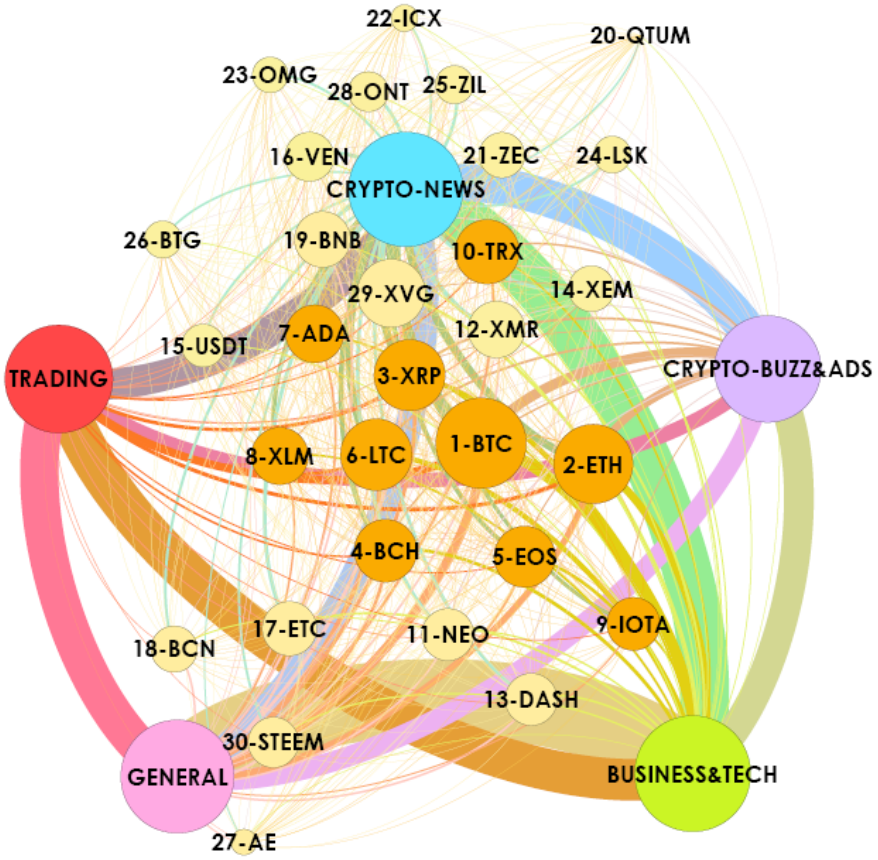


Figure 6.1: Network of merged hashtags representing relationship between cryptocurrencies and topics. Five biggest nodes on the corners represent topics, while the orange and yellow nodes represent top-10 and top-11-to-30 cryptocurrencies respectively.

the same communities and hashtags that convey similar cryptocurrencies are merged into supernodes, each of which represents individual topic and cryptocurrency as presented in Figure 6.1 which is created by Gephi [30], a network visualisation tool.

As shown in Figure 6.1, the topic nodes are fixed on the five corners and the layout algorithm *Force Atlas 2* [33] was applied to arrange the cryptocurrency nodes. The algorithm balances distance between nodes where nodes that are strongly related, presented by high weights, are placed close to each other, while nodes with weak connections are far apart. Topic nodes are coloured differently to clearly present connections between nodes, while the cryptocurrencies nodes are coloured in oranges for top-10 cryptocurrencies and yellow for the lower-ranking ones.

Overview Analysis

From the graph, in general, it can be seen that most of the low-ranking cryptocurrencies are distributed around Crypto-News topic, while the top-10 cryptocurrencies stay in the middle of the network which is surrounded by all topics. This implies that **low-ranking cryptocur-**

rencies were only mentioned in topics closely related to cryptocurrencies market, while the top-10 cryptocurrencies are widely mentioned in more general topics such as Trading, Business&Tech, and General. Additionally, some cryptocurrencies that have special characteristics are spotted on the corresponding topic. For example, 9-IOTA, which uses a novel technology rather than blockchain, is closely related to technology topic. There are some other distinct relationships which are noteworthy for further investigation. To conclude, **this visualisation can be used for overview analysis of how social talked about cryptocurrencies in which topics and how each cryptocurrency is related to each other**.

Weight Analysis

To analyse the strength of the relationship between topics and cryptocurrencies, two heat maps are plotted. Figure 6.2 presents weights in absolute view, while Figure 6.3 present weights normalised by the sum of all weights in the same row, which represent the same topic. The pixel colour represents edge weights showing how often a cryptocurrency appear in a topic. Green means high weight and red means low weight. The prefix of cryptocurrencies in the column name represents their ranking.

According to Figure 6.2, there are three highlights presented. First, the **higher-ranking cryptocurrencies are more often mentioned in various topics than the lower-ranking ones**. This can be seen from more green pixels on higher-ranking (on the left) and more red pixels on lower-ranking (on the right) cryptocurrencies.

Second, in the aspect of topic popularity, **cryptocurrency news/updates topic is the most popular one that mostly used in the social network to associate with cryptocurrencies.** This can be seen from all green pixels in the first row. Following that business&technology, general, trading and buzzword&advertisement topics show less popular in respective order. This can be seen from the decreasing number of green pixels on each row accordingly.

Third, it is interesting to note that only around top-6 cryptocurrencies are strongly related to buzzwords and advertisement topics. This implies that **advertisement is more widespread on prominent cryptocurrencies than the less prominent ones.**

Normalised Weight Analysis

Next, in Figure 6.3, the connections are analysed in normalised form where each weight in a row is divided by the sum of all weights in that row. This presents how often each cryptocurrency is associated with the topic compared to other cryptocurrencies in the same topic. There are three findings should be highlighted.

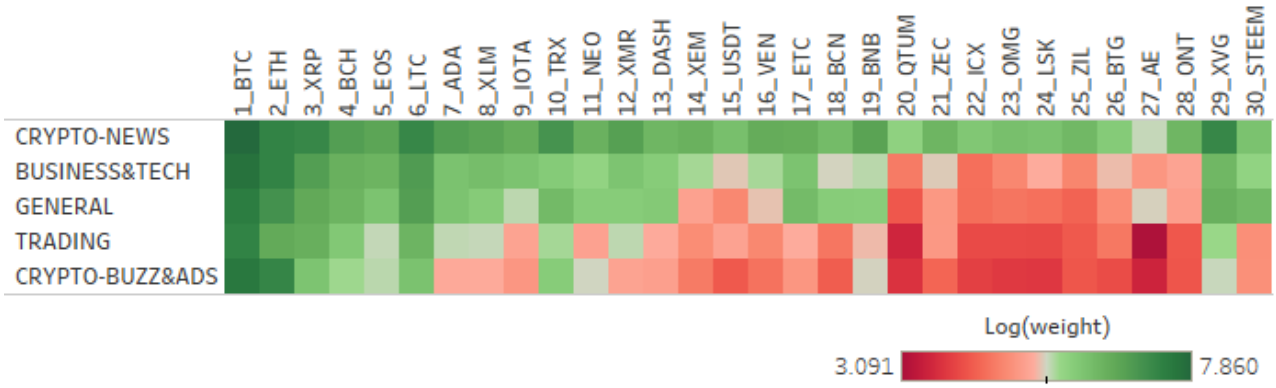


Figure 6.2: Heat map of log of edge weight in Figure 6.1. Dark green colours on the left shows the topics mostly co-occurred with the top cryptocurrencies and, generally, less co-occurred with lower-ranking ones

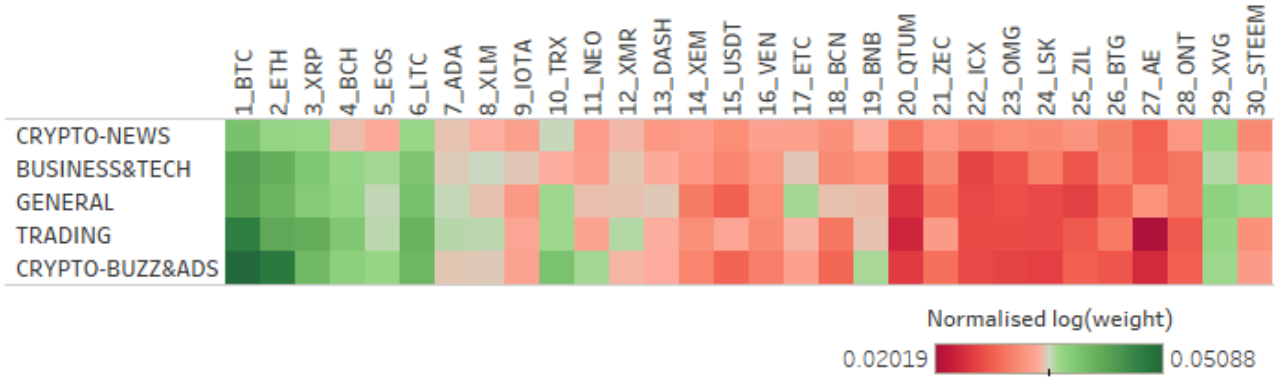


Figure 6.3: Heat map of normalised log of edge weight in Figure 6.1. The weights are normalised by the sum of all weights in their row. Dark green on the bottom left shows Bitcoin and Ethereum are strongly related to topics about buzzwords and advertisement

First, according to the gather of the green pixels on the left, it can be interpreted that top-6 cryptocurrencies are most relevant to all topics, especially, 1-BTC and 2-ETH with topic Crypto-Buzz&Ads as seen from the darkest green pixels. This means the **top-6 cryptocurrencies are not only of interest in the cryptocurrency related topics but also of interest in general**. Similarly, 29-XVG seems to be an outlier that is also strongly related to all topics but it is in rank 29 while other in rank 1 to 6. This will be discussed in *outlier explained* section

Second, some specific cryptocurrencies stand out from other cryptocurrencies in some topics as seen from green among red pixels i.e., 17-ETC and 30-STEEM in general topic, 12-XMR in trading topics, and 11-NEO and 19-BNB in cryptocurrency buzzwords and advertisement topic. These specific relationships are interesting for further investigation. However, this is beyond the scope of this dissertation and could be noted as a future work.

Lastly, from the darkest green pixels on the left, it is interesting that 1-BTC and 2-ETH, are highly associated with advertisements which are by far more frequently compared to other cryptocurrencies as seen from contrast colour. This reflects the same insights found in third findings in the last section, but clearer that **1-BTC and 2-ETH are outstandingly often**

associated with advertisement and buzzwords.

Summary

This experiment shows how the cryptocurrencies are related to topics discussed in the social network. First, as shown in the overview analysis, cryptocurrency and topics network in Figure 6.1 can spot an outstanding relationship which is able to reflect some facts in the cryptocurrency market. These relationships can be used as initial clues for further investigation which might lead to actionable insights. Thus, **the network can be used for cryptocurrency analysis and extract some insights from social interaction**.

Second, from all analyses above, the main finding shows that the high-ranking cryptocurrencies are widely discussed in various topics while the lower-ranking ones mostly appeared in cryptocurrencies related topics. This can be concluded that **the bigger a cryptocurrency is the more prominent it seems to be in social network and likely to be mentioned in various topic including general ones**

6.1.2 Relationship between each Cryptocurrency

This experiment is to analyse the relationship between cryptocurrencies and aims to answer following questions:

- Which cryptocurrencies are strongly related to each other in terms of that they were often mentioned together in the social network?
- How their relationships are related to their rankings?

Overview Analysis

To analyse the relationship between cryptocurrencies in general, we used the same network in Figure 6.1 but without the topic nodes. After that community detection was performed to detect groups of cryptocurrencies that have strong connections together. The result in Figure 6.4 shows that there are three groups detected with *Modularity 0.078*. The groups are differentiated by colour.

In general, it can be noticed that 1-BTC is the largest node and stay in the middle of the network. This means **1-BTC most frequently co-occurred with other cryptocurrencies**. Next, when looking at the community level, first, **higher ranking cryptocurrencies tend**

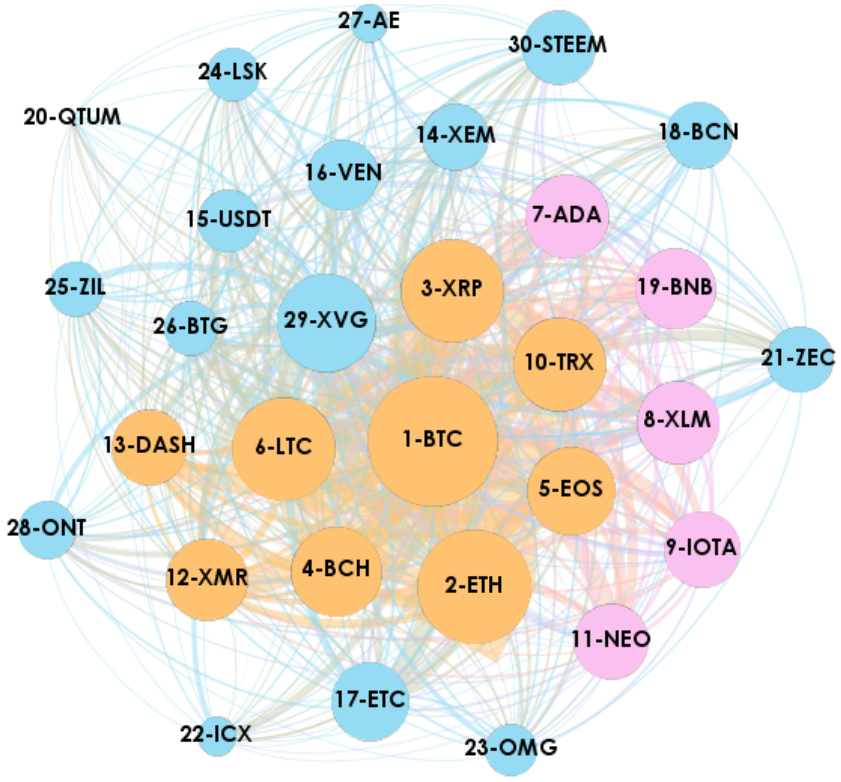


Figure 6.4: Cryptocurrencies hashtags network with 3 communities detected of which Modularity value is equal to 0.078.

to stay in the same community and distributed around 1-BTC as seen in the orange community.

Second, in the blue community, it seems that 29-XVG is the most influential cryptocurrencies in this group as presented by its largest node. However, this is still not clearly identified and, thus, will be further explored in the heat map analysis in the section below.

Third, in the pink community, there is one outlier, 19-BNB from rank 19 while others are from around rank 7 to 11. From further investigation, this cryptocurrency is a crypto-exchange platform which allows users to make an exchange between cryptocurrencies. This should make it related to other cryptocurrencies being traded on the platform, especially, the prominent ones which might more often co-occur due to their high volume. Therefore, this might be the reason why 19-BNB has the extraordinary stronger relationship to the top-ranking cryptocurrencies than the other low-ranking ones.

Weight Analysis

Next, to analyse the strength of the relationships, the edge weights of the graph in Figure 6.4 are visualised as the heat map in Figure 6.5. There are two highlights shown in this heat map. First, according to the green pixels around upper-left corner of Figure 6.5, cryptocurrencies around

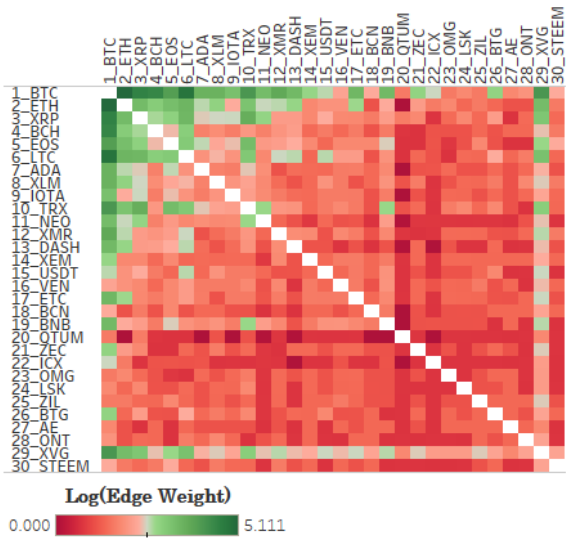


Figure 6.5: Heat map of log of edge weight in Figure 6.4. Dark green around the corner shows strong relationship between higher-ranking cryptocurrencies.

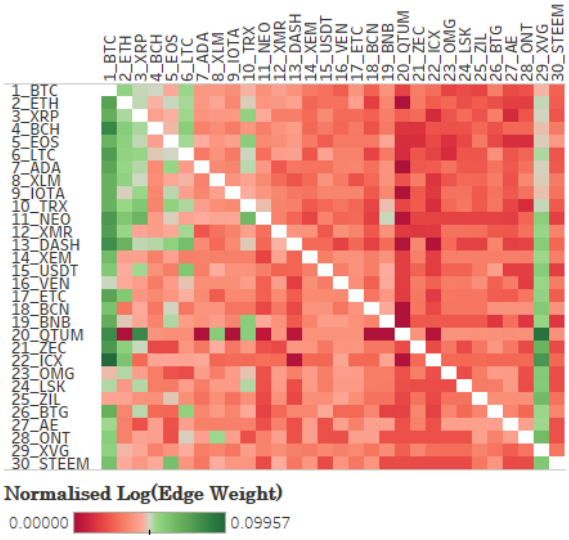


Figure 6.6: Heat map of normalised log of edge weight in Figure 6.4. Green columns on the left show high-ranking cryptocurrencies strongly related to most cryptocurrencies.

rank 1 to 10 are strongly related and more related compared to the relationship between the lower-ranking ones presented in red pixels. Second, it should be noted that 1-BTC and 29-XVG are only two cryptocurrencies that are strongly related to most of the other cryptocurrencies as seen from the mostly green stripe on their row. This can be explained more clearly in the next visualisation.

Normalised Weight Analysis

The findings above are more clearly presented in Figure 6.6 which presents heat map of the weight edges in normalised form where the value in each pixel is the fraction of total weights in that row. According to many green pixels around some first left columns in Figure 6.6, they show that most top cryptocurrencies, rank 1-6, are strongly related to other cryptocurrencies, particularly, 1-BTC and 2-ETH at the first 2 columns. This reflects the same results as Figure 6.5 but more clearly that 1-BTC and 2-ETH are significantly stronger than the others. Therefore, the first findings can be concluded that **high-ranking cryptocurrencies are more likely to be mentioned with other cryptocurrencies than the low-ranking ones, especially 1-BTC and 2-ETH**

Next, there are some outstanding relationships spotted on both heat maps. First, on Figure 6.6, 29-XVG in rank 29, show outstanding result where it occupies green pixels amongst the low-ranking cryptocurrencies with the most intense centred at rank 20. This implies that **29-XVG is the centre of low-ranking cryptocurrencies** of which related tweets often co-occurred with 29-XVG. Similarly, 10-TRX in rank 10 shows many green pixels connected to the top 10

in which stronger than the cryptocurrencies in rank 7, 8 and 9.

Second, some pairs stand out from the heat map and do not follow the trend. This can be noticed from green pixels in the middle of red as mostly seen in row 20-QTUM and some in row 19-BNB, 26-BTG, and 28-ONT. These outliers are worth noted for further investigation as it might imply some clue that can lead to some interesting insights. However, this is beyond the scope of this dissertation which aims to study the overall market, not on specific cryptocurrencies.

Summary

There are three key findings found in this experiment. First, this experiment starts from the overview analysis of cryptocurrencies network which was able to detect three groups of strongly related cryptocurrencies as separate communities in Figure 6.4. This result can answer the first question of this experiment that which cryptocurrencies are strongly related together.

Second, the heat maps are plotted to analyse the strength of the connections between cryptocurrencies. The analysis revealed that **higher-ranking cryptocurrencies, particularly 1-BTC and 2-ETH, show strong association to most cryptocurrencies, while the lower-ranking ones are less associated by other cryptocurrencies**.

Third, the analysis can identify some outliers that are interesting for further investigation. The highlight outlier is 29-XVG which is strongly related to the majority of lower-ranking cryptocurrencies around rank 20. These findings show that the **hashtag co-occurrence network can be used to analyse the relationship between cryptocurrencies in the market and identify outliers via the social interaction in Twitter**.

6.2 User Analysis

This section contains three analyses on users who are active in cryptocurrencies. First, correlation of user scores is calculated to understand the preferences of users over difference cryptocurrencies and how they are related. Second, homophily analysis is used to study how likely that users who have the same preferences are connected together. Third, user influence, measured by retweet count, is analysed using the linear regression method. This is to understand which factors are correlated to users influence in the social network of cryptocurrencies.

6.2.1 Correlation between User Preferences

This experiment is to understand the preferences that a user have over different cryptocurrencies and topics. This aims to answer the following questions:

- If a user like specific cryptocurrency, how likely that he will also like another cryptocurrency?
- How cryptocurrency ranking is related to user preferences in the social network?

To analyse the relationship between user preferences, the correlation of the user scores is calculated. We choose Spearman and Kendall methods and exclude Pearson which is restricted in measuring linear relationship while the other two methods can well explain the monotonic relationship that is exhibited in the relationship between user scores.

Correlation Analysis

Figure 6.7 and 6.8 respectively present Spearman and Kendall correlation coefficients. It can be seen that both correlation methods show the same pattern where most cryptocurrencies and topics are positively correlated together, as seen by green pixels, while only 1-BTC, 2-ETH, and buzzwords&advertisement are negatively and less correlated to others. This can be noticed from red and white pixels along their rows in Figure 6.7 and 6.8. This result implies that **users who are active about 1-BTC, 2-ETH, and Crypto-Buzz&Ads are unlikely to be interested in other cryptocurrencies, while users who are active about altcoins are likely to be interested in the other altcoins**.

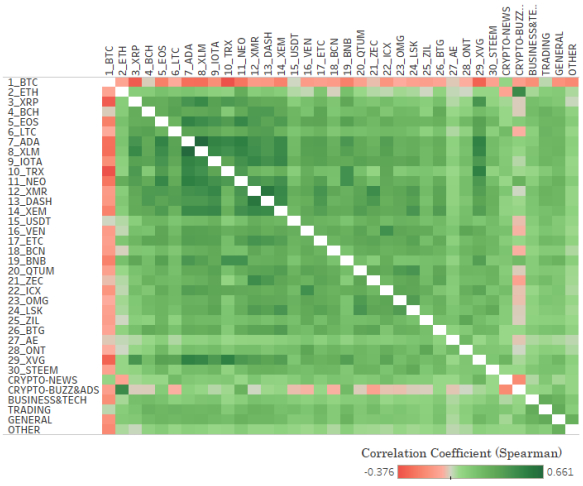


Figure 6.7: Heat map of Spearman correlation coefficients between cryptocurrencies and topics.

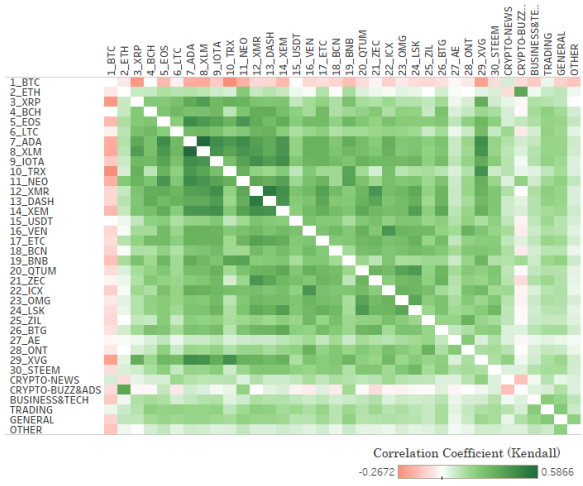


Figure 6.8: Heat map of Kendall correlation coefficients between cryptocurrencies and topics.

When taking a closer look on Figure 6.7, there are some pairs stand out from the others as shown by dark green and dark red pixels. For example, 2-ETH vs Crypto-Buzz&Ads and 7-ADA vs 8-XLM show strongly positive correlation around 0.6, which means users tend to be interested in the cryptocurrencies in the same pair. Another example is that 1-BTC vs 3-XRP, 10-TRX, and 29-XVG show negative correlation around 0.3, which means users who like 1-BTC are unlikely to be interested in these cryptocurrencies. Thus, these findings demonstrate that this correlation analysis method can extract some insights from the user behaviour in social network and these insights can be further analysed to understand the specific relationship between the cryptocurrencies.

Network Analysis on Positive Correlation

To find groups of cryptocurrencies or topics that are strongly correlated, the network of positive correlation (green pixels in Figure 6.7) was built. As shown in Figure 6.9, the nodes represent cryptocurrencies and topics, node size represents the weighted degree – the higher, the bigger. While edges represent a Spearman correlation coefficient between nodes scaled by the edge width. Spearman is used rather than Kendall as it can present clearer positive correlation as seen in Figure 6.7.

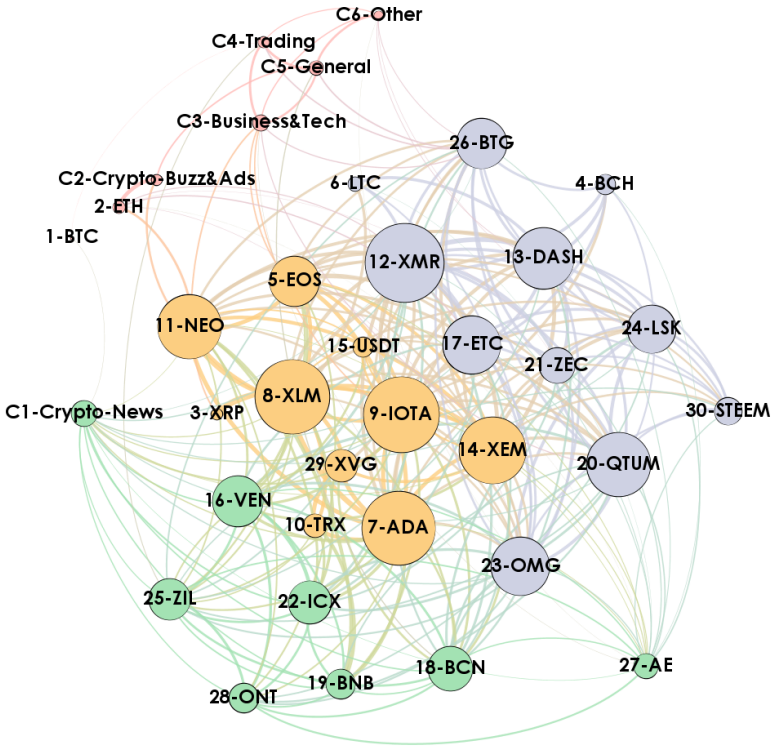


Figure 6.9: Network of positive correlation coefficients from Spearman method. Three detected communities show strong correlation between the members.

Secondly, the community detection method was performed to detect clusters of nodes that are strongly correlated. As a result, five clusters were detected and shaded in different colours. This can be interpreted that the members belonging in the same cluster are more correlated than others in different clusters. There are two key findings in this analysis.

First, from the correlation network, it can be seen that 1-BTC, 2-ETH and all topics except Crypto-News were grouped in the same cluster. This means users who are interested in prominent cryptocurrencies, which are 1-BTC and 2-ETH at the time of writing, are likely to talk more about general topics as presented by the topics nodes in the red group. In other words, **when users are very active about Bitcoin and Ethereum, they are unlikely to care much about altcoins**. This can also be spotted from the red and white first two columns on the heat map in Figure 6.7 and Figure 6.8.

Second, for other clusters, there is a slight trend that the clusters are mixed from cryptocurrencies in relative rankings. As seen in Figure 6.9, the orange cluster is the group of high-ranking altcoins, most of the members are from rank 3 to 10, and some members are from lower ranks. Blue cluster is built from around 10s to early 20s ranking, while the green cluster is built from 10s to almost 30 ranking. This shows that users in the social network tend to be interested

in cryptocurrencies in close ranking. For example, if they are interested in prominent altcoins such as 3-XRP or Ripple, they are likely to talk about cryptocurrencies around rank 3 and above rank 10 as shown in the orange group. However, 29-XVG in this orange group again can be marked as an outlier which will be discussed in *outlier explained* section.

Summary

Overall, the result can be concluded that **users who like altcoins are likely to like other altcoins around their ranks, while users who actively talked about the top ranks, Bitcoin and Ethereum, show less interest in altcoins but more engagement in general topics**. Moreover, the visualisation can illustrate some distinct relationships and detail between cryptocurrencies in pairs which could be further investigated for in-depth understanding.

6.2.2 Homophily Analysis

Homophily analysis is applied in this experiment to study behaviour of users on Twitter, how they are connected to each other based on their preference. It uses a measure called *Assortativity coefficient* [34], a Pearson correlation of degree between two connected nodes, to determine how likely that nodes with the similar attribute are connected together. The value is ranged between -1 to 1 where 1 indicates perfect assortative network in which all connected nodes have the same attribute value, while -1 indicates perfect disassortativity meaning that attribute values of two connected nodes perfectly vary in the opposite direction i.e., perfect negative correlation between attribute values of all pairs of connected nodes in the network.

This experiment is separated into two parts. First part is for analysing the assortativity of individual scores, which presents how users are connected based on their score. The second part is for analysing the assortativity of the user vector which presents how users who have the same tweeting behaviour are connected. This experiment aims to answer:

- How likely that users who like a specific cryptocurrency or topic are connected to each other? For example, if a user like Bitcoin, how likely that his/her friends will also like it?

- How likely that users who have the same preference, in term of how they tweet about cryptocurrencies and topics, are connected to each other?

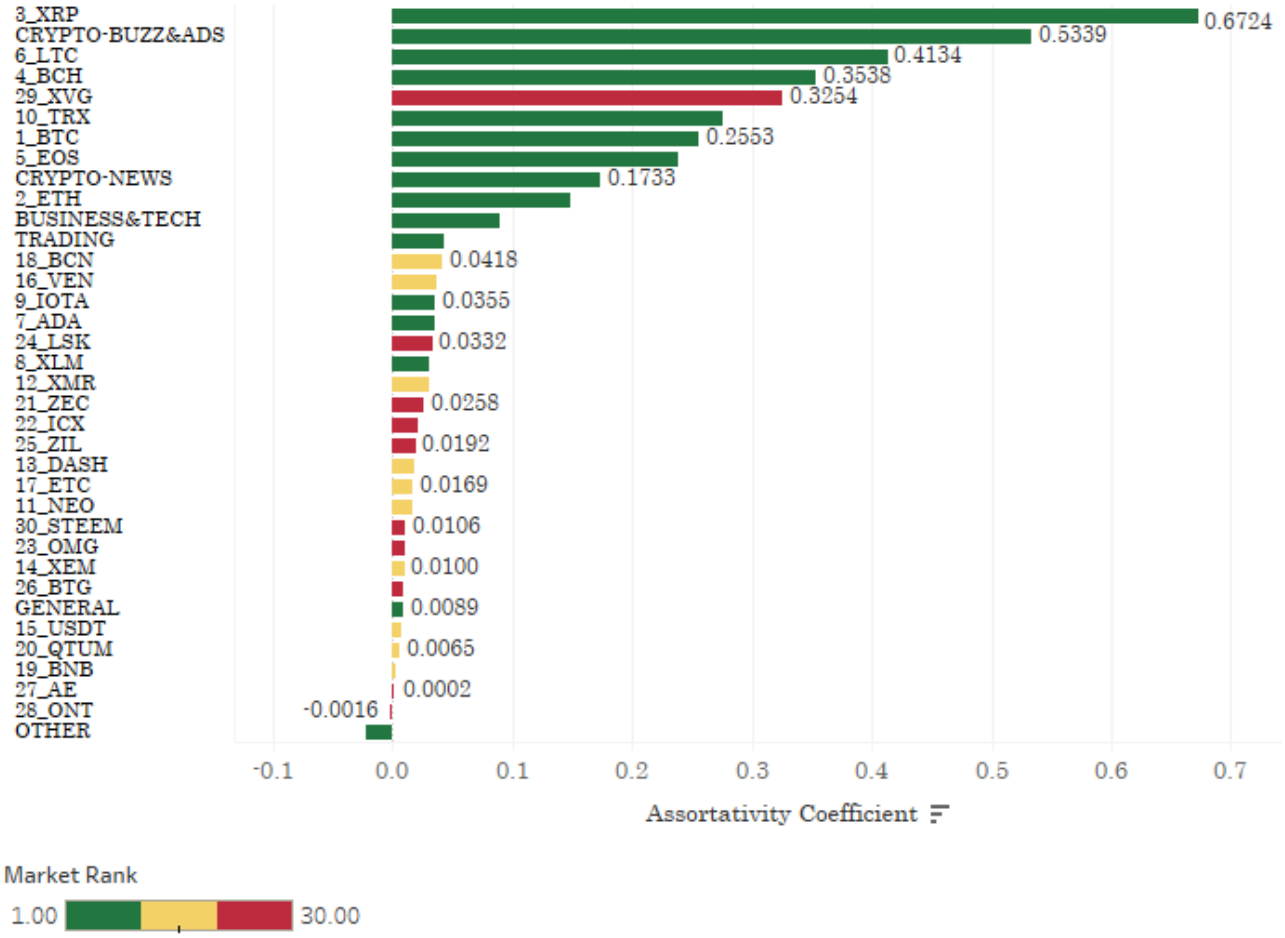


Figure 6.10: Assortativity coefficients of user network based on each individual score. Colour presents ranking with green indicating rank 1-10 and topics, yellow indicating rank 11-20 and red indicating rank 21-30. Overall, top-10 cryptocurrencies and some crypto-related topics (in green) show higher values than the others meaning that they are more likely to be connected to others with similar behaviour.

Assortativity for Individual Score

To set up this experiment, firstly, the network of the active users is constructed. After that igraph [35], a python package for network analysis, is used to calculate assortativity coefficients of the user scores based on the Newman MEJ literature [34]. This results in 36 coefficient values with 34 positive and 2 negative assortativity as shown in Figure 6.10. The bar chart is coloured according to cryptocurrencies ranking where topics and rank 1-10 are in green, rank 11-20 are in yellow and rank 20-30 are in red.

To ensure the significance of the results, we performed hypothesis testing of the assortativity results. We generated a set of 200 null models which are a randomly rewired form of the user network generated by Erdos Renyi method [36]. Then, we calculated assortativity coefficients of these null models which resulted in the distribution of the values from the random networks. With this distribution, we could calculate *z-score* and then *p-value* of the original results which are presented in Figure 6.11. The results from all scores are 0.01 statistically significant except

Score Name	Assort. Coef.	P-Value
1_BTC	0.25534	< 1.69E-299
2_ETH	0.14900	< 1.69E-299
3_XRP	0.67236	< 1.69E-299
4_BCH	0.35383	< 1.69E-299
5_EOS	0.23825	< 1.69E-299
6_LTC	0.41336	< 1.69E-299
7_ADA	0.03551	< 1.69E-299
8_XLM	0.03086	2.02E-270
9_IOTA	0.03555	< 1.69E-299
10_TRX	0.27625	< 1.69E-299
11_NEO	0.01642	5.40E-91
12_XMR	0.03000	1.17E-203
13_DASH	0.01829	1.54E-130
14_XEM	0.01003	6.13E-55
15_USDT	0.00807	1.21E-27
16_VEN	0.03678	< 1.69E-299
17_ETC	0.01687	1.28E-78
18_BCN	0.04176	< 1.69E-299
19_BNB	0.00346	0.00018
20_QTUM	0.00653	5.26E-17
21_ZEC	0.02580	3.64E-198
22_ICX	0.02100	4.96E-122
23_OMG	0.01006	4.87E-31
24_LSK	0.03317	1.69E-299
25_ZIL	0.01916	1.39E-105
26_BTG	0.00913	3.84E-25
27_AE	0.00023	**0.43363**
28_ONT	-0.00158	**0.12936**
29_XVG	0.32537	< 1.69E-299
30_STEEM	0.01060	1.13E-40
CRYPTO-NEWS	0.17329	< 1.69E-299
CRYPTO-BUZZ&ADS	0.53390	< 1.69E-299
BUSINESS&TECH	0.08974	< 1.69E-299
TRADING	0.04374	< 1.69E-299
GENERAL	0.00887	1.13E-29
OTHER	-0.02150	2.13E-118

Figure 6.11: Assortativity of each score with *p-value*. Most scores show homophily property with statistical significance at p-value <0.01. While only 27-AE and 28-ONT are not significant meaning that the way users tweeting about both cryptocurrencies are not correlated to the way they connect to each other.

27-AE and 28-ONT. This can be interpreted that **the way users tweeted about 27-AE and 28-ONT are not correlated to the way they follow or are followed by other users. On the other hand, the other scores are significantly correlated to the user behaviour**.

According to Figure 6.10, overall, it can be seen from the green bars that most of the top-10 cryptocurrencies, most topics, and 29-XVG, an outlier, are located on the top of the chart with assortativity coefficient more than 0.1. While cryptocurrencies below rank 10 and general topics show by far less assortativity, particularly, users who tweeted about other topics showing negative assortativity. Thus, this result can be interpreted that **users who are interested in top-10 cryptocurrencies (except 7-ADA, 8-XLM, and 9-IOTA) and topics related to cryptocurrencies, business and technology are more likely to be connected with other users who have the same behaviour than users who tweeted about the lower-ranking cryptocurrencies and general topics.** In contrast, the users who tweeted about other topics seem not to be connected to each other as seen from the negative assortativity.

When considering individual scores in the bar chart, it is interesting that 3-XRP comes as the first place with value 0.67 followed by Crypto-Buzz&Ads with value 0.53. This demonstrates that 3-XRP have very strong communities in Twitter. Likewise, users who tweeted about buzzwords and advertisement are likely to be connected to each other. The latter might reflect some behaviour of cryptocurrency advertisement in Twitter i.e., advertising users like to follow each other to gain more audiences. Further investigation of this clue might lead to better understanding of crypto-advertising behaviour in Twitter.

Assortativity for User Vector

To consider all scores as a representative of user behaviour, user vector is used for the calculation. As suggested in literature [37], we can measure the vector assortativity by firstly

User network	200 Null models			
	Mean	Std	Z-Score	P-Value
0.73873	0.67979	0.00008	712.03918	< 1.69E-299

Figure 6.12: Average euclidean similarity of user network with mean and standard deviation of results from null models. The test result shows statistical significance with p-value less than 0.01

calculating the average similarity of every connected user with Euclidean similarity as shown in equation (6.1), $\vec{v}_{src}$ is source user vector and $\vec{v}_{target}$ is target user vector.

$$Sim_{Euclidean} = \frac{1}{1 + ||\vec{v}_{src} - \vec{v}_{target}||_2} \tag{6.1}$$

Secondly, hypothesis testing is performed to test the significance of the result. To achieve this, 200 random networks with the same nodes and edges numbers as the original network are generated and used as null models for testing. As shown in Figure 6.12, mean and standard deviation are calculated and then used to find *z-score* then *p-value* respectively. As a result, it is statistically significant that similar users tend to be connected to each other in the user network as presented by similarity greater than 0 and *p-value* less than 0.01.

$$Assortativity_{vector} = \frac{Sim_{network} - Sim_{null}}{Sim_{null}} \tag{6.2}$$

To be precise, we can quantify the level of assortativity in this network by measuring how it differs from the result from random models as shown in equation (6.2). This results in positive assortativity of *0.0867*, which means the user network is around 9% more assortative than a random network. Therefore, it can be concluded that **users in the social network of cryptocurrency exhibit *homophily* property in which users tend to connect with others who have the same behaviour**.

Summary

The assortativity coefficients from individual scores reveal that **users who tweeted about top-10 cryptocurrencies and topics related to cryptocurrencies, business and technology are most likely to be connected to others with similar behaviour**. This is also true with the lower ranking cryptocurrencies, trading, and general topics but with weaker likelihood. On the other hand, **users who tweeted about other topics show negative assortativity meaning that they are unlikely to be connected to others with similar behaviour**.

For assortativity of user vector, it presents an overall view of how users connect to each other based on their tweeting behaviour. Similarly, the analysis shows that the user network has homophily property, where **users with the same behaviour are likely to connect to each other with the chance of around 9% more than random networks**.

6.2.3 Influential Factor Analysis

This experiment aims to understand which factors are correlated to users influence in the cryptocurrency social network. With motivation from the literature [38] about correlation between diversity of tweet topics and user influence, in this study, we focus on studying whether users who tweet about a specific cryptocurrency or users who tweet about various cryptocurrencies are more likely to be more influential in the network and which cryptocurrency or topic is more correlated to social influence. Follows are the questions answered in this experiment:

- Which factors are correlated to user influence in the cryptocurrency social network?
- Between tweeting about fewer cryptocurrencies(or topic) and more diverse cryptocurrencies(or topic), which one tends to gain more social influence?
- Tweeting about which cryptocurrency-related topic tends to gain more social influence?

In this study, user influence is measured by retweet count, which is the frequency that tweets of a user is retweeted. To answer the question above, we performed linear regression to predict retweet count from 2 types of feature. The first type is user metadata consisting of follower count, friend(followee) count, number of tweet and entropy of topic/cryptocurrency in user hashtags. The second type is user score consisting of 36 features from the user scores defined in the last chapter.

$$Entropy\ H(\#Hashtag_{topic}) = -\sum_{n=1}^{N_{topic}=36} p(\#Hashtag_{topic}) \log_2 p(\#Hashtag_{topic}) \tag{6.3}$$

Firstly, the entropy is calculated and individually used to predict retweet count. This aims to find whether tweeting by focusing on a few specific cryptocurrencies (small entropy) or diverse of them (large entropy) tend to gain more influence from the social network. Equation (6.3) shows the formulae used for the calculation. The number of hashtags related to cryptocurrencies or topics is used to calculate the entropy. Thus, more diverse of the hashtags results in larger entropy. As shown in Figure 6.13 (a), the regression result shows that entropy is statistically significant with the negative coefficient. This means lower entropy results in higher

Entropy Feature		
Features	**Coef.**	**P-value**
(Intercept)	2052.6	5.27E-08 ***
crypto&topic entropy	-453.7	0.0015 ***

(a) Entropy feature with significant negative coefficient meaning that tweeting by focusing on a smaller number of topics/cryptocurrencies gain more social influence

Best-AIC Metadata Features		
Features	**Coef.**	**P-value**
(Intercept)	516.172078	1.82E-07 ***
followers_count	0.494767	< 2e-16 ***
friends_count	-0.450888	< 2e-16 ***
twt_count	-0.020289	0.0556 .

(b) Best-AIC model from user-metadata feature. The entropy feature was included in the full model but was then removed by stepwise method to produce the best AIC

Best-AIC User Score Features		
Features	**Coef.**	**P-value**
(Intercept)	-146.7	0.72997
10-XRP	4412.7	2.88E-05 ***
Business&Techology	11417.7	0.000192 ***
2-ETH	4833.9	0.006361 **
Crypto-Buzz&Ads	1807.1	0.007341 **
1-BTC	1455	0.031357 *
Others	-3361.2	0.071277 .
General	-7841.1	0.124344

(c) Best-AIC model from user-score feature reduced from 36 scores.

Best-AIC Combined Features		
Features	**Coef.**	**P-value**
(Intercept)	-6.19E+02	0.0017 **
followers_count	4.95E-01	< 2e-16 ***
friends_count	-4.59E+00	< 2e-16 ***
10-XRP	3.78E+03	4.64E-06 ***
Business&Techology	9.89E+03	6.06E-05 ***
Crypto-Buzz&Ads	2.09E+03	7.54E-05 ***
twt_count	-1.89E-02	0.0741 .
2-ETH	2.07E+03	0.1495

(d) Best-AIC model from combination of the best features in a) and b).

Figure 6.13: Coefficients and p-values of four linear regression models with retweet count as the target. Features on each model are selected by stepwise method. * presents significant level

retweet count. In other words, **tweeting about specific topics tends to gain more social influence than tweeting about various topics.**

Secondly, the entropy feature is used together with the user metadata to predict the retweet count. Then stepwise method [39] , a model selection method, is used to select the best features that produce the lowest Akaike information criterion or AIC [40]. AIC is a measure of model quality with the trade-off between model simplicity and how well it can fit the target data - less is the better. In this experiment, both forward and backward stepwise are applied and then the model that produces the best result is selected to show in the Figure 6.13. According to the result in Figure 6.13 (b), entropy was excluded from the model, while follower and friend counts are the most significant features followed by tweet count. This means entropy is a trivial feature compared to the others. Positive coefficient of follower count means that users are likely to gain more retweet when they have more followers. On the other hand, negative coefficients in friend count and tweet count mean that users who are following more users and post more tweets tend to gain less social influence.

Thirdly, to find that which cryptocurrency or topic tends to gain influence in the social network, the user scores are used for the regression. Similarly, the stepwise methods are applied to select the best features. As a result, 7 from 36 scores are selected. As shown in Figure

6.13 (c), topics about 10-XRP and business&technology are the most significant features, followed by 2-ETH, buzzwords&advertisement, 1-BTC, and other topics. All of them comes with positive coefficients except other topics. While general topics, with the negative coefficient, is not significant but included in the model. These results can be interpreted that **tweeting about those positive-coefficient topics might gain more retweet, especially about 10-XRP and business&technology. On the other hand, tweeting about other topics might reduce chances to be retweeted in this cryptocurrency social network. While tweeting about general topics together with some of the topics above might also reduce the retweeted chance**.

Lastly, to find the best model to predict user influential in the network, we combine the best features from previous models and then performed stepwise feature selection again. As a result, 1-BTC, Other and General scores are removed from the models as demonstrated in Figure 6.13 (d). All features except tweet count and 2-ETH are highly significant with the p-value less than 0.001 while tweet count shows small significance at the p-value less than 0.1 and 2-ETH is not significant. When considering coefficients sign, the result shows the same as previous models where all features except friend count and tweet count are positively correlated.

By comparing the results between all models in Figure 6.14, we found that the combined model produces the best results followed by metadata, user score, and entropy models. This is indicated by R-squared, p-value, and AIC which are used to measure the quality of a linear model. According to Figure 6.14, from the p-values, we can interpret that the results from all models are statistically significant. For adjusted R-squared, around 0.33 from the best model means all features in the model can explain around 33% of the variance in retweet count. As close as that, 0.32 or 32% of the variance in retweet can be explained by only metadata features. In contrast, user scores and entropy features can explain just a tiny bit of variance in retweet count, around 0.5% and 0.1% respectively. When looking at the improvement from the metadata model to the combined model, R-squared increases just around 0.004 or 0.04%. This small improvement shows that user scores are trivial factors to associate with user influence compared to the metadata features.

Model	Adjusted R-squared	P-value	AIC
Best-AIC Combined Features	0.3261	< 2.2e-16	152076.1
Best-AIC Metadata Features	0.3225	< 2.2e-16	152117.1
Best-AIC User Score Features	0.004792	4.65E-08	155357.5
Entropy Feature	0.001078	0.001501	155382.9

Figure 6.14: Comparison of adjusted R-squared, P-value and AIC between 4 models in Figure 6.13. The combined model is indicated as the best model from highest adjusted R-squared and lowest AIC.

Although metadata features can produce the best indicators amongst other as seen from 33% of retweet count can be explained by it, there is still a big gap of around 67% variance of retweet count that cannot be explained in this experiment, which focuses on studying the relationship between cryptocurrency-related topics and user influence. Further study focusing on user influence prediction could be conducted to fulfill this gap.

Summary

From this experiment, first, entropy shows the negative correlation to retweet count. Thus, **focusing on tweeting about fewer topics are associated with a chance to gain more social influence**. Second, by using user scores for the regression, it shows that **tweeting about 10-XRP, business and technology topics are the most significant factor that are correlated to social influence**, while 2-ETH, buzzword, advertisement, and 1-BTC show less significant correlation. Finally, **follower and friend count are the most influential features to predict retweet count**, while user scores and entropy show very small influence to retweet count. Nevertheless, when combined the metadata and user scores features together, it can produce a better model with a bit of improvement from the metadata model.

6.3 Outliers Explained

From the analysis, it is interesting that there are 2 cryptocurrencies, 10-TRX and 29-XVG, that are often spotted as outliers. From further investigation, this might be due to strong communities they have in Twitter. Both 10-TRX and 29-XVG have a very high number of followers, around 300,000, which are much higher than other cryptocurrencies around their ranking, especially 29-XVG which is in rank 29 but its number of followers is almost the same as the cryptocurrencies in top six. This finding can roughly explain their strong relationship with various topics; however, deeper analysis on both cryptocurrencies would be interesting. This clue from social interaction might imply something about the evolution of both cryptocurrencies in the future.

Chapter 7

Result Discussion and Future work

From the experiment results, it can be seen that network analysis techniques are capable to extract insights from the social interaction in Twitter. The analyses on those results focus more on explaining the overall picture of cryptocurrency market and less on investigations of specific relationship of individual cryptocurrencies although some interesting relationship was spotted from the visualisations. In this chapter, the key findings are discussed more broadly together with other related work and some limitation.

Cryptocurrency ranking is related to how they are discussed in the social network where higher ranking cryptocurrencies tend to attract many cryptocurrencies and a wide range of topics to be associated with. On the other hand, lower ranking cryptocurrencies are not of interest to be associated with other cryptocurrencies and they are mostly mentioned in topics specifically related to cryptocurrency. Thus, prominent cryptocurrencies tend to gain more audiences and, thus, become more popular, while the less prominent ones attract less audience and become less popular. This finding implies preferential attachment or rich-get-richer property [41] of user behaviour having toward cryptocurrencies in the social network. This phenomenon can be linked to the power-law shape of market shares against ranking plot in literature [11]. Due to this rich-get-richer property, this finding could explain why only top 30 from around 1700 cryptocurrencies managed to gain 90% capitalisation of the whole market [1].

Similarly, cryptocurrency ranking also shows some influence to user behaviour in cryptocurrency network. This finding can be simply concluded that users who like altcoins seem to like other altcoins too, while users who like Bitcoin and Ethereum do not care much about other cryptocurrencies but are more interested in general topics. This result shows that social network analysis could reveal user behaviours related cryptocurrencies. This could be linked to investor behaviour and, with further study, it has a potential for predicting trend in the market. Lastly, it is worth noting that this finding has never been presented in any related works before. This proves that social network analysis is a novel method in this field which

might motivate future research focusing on this direction. Second, we could detect homophily property in this user network. This means users who have the same preference or behaviour in the social network are likely to be connected. The effect of homophily network in this literature [42] showed that ones are easier to be influenced by others if the network is homophily. Thus, amongst group of active users, ones are more likely to be easily influenced by other, especially, the users who are interested in Ripple (3-XRP), which show the highest assortativity measure. This insight would be useful for ones who plan to gain influence in the cryptocurrencies network i.e., strategies for spreading news or advertisement about cryptocurrency in the social network.

According to influential factor analysis, although the result shows that topics and diversity of topics are not efficient enough for prediction, we could identify their correlation with social influence. We found that rather than tweeting about diverse cryptocurrencies, focusing on fewer cryptocurrencies is more likely to gain more social influence, measured by retweet count. This findings could be used to design tweet strategies for gaining influence in the social network.

7.1 Future Work and Limitation

Not only could this study present the general key findings discussed above, it can also identify outlier and extract insights for individual cryptocurrencies. Further specific analysis on some of those insights has proven that they reflect some facts in the market. However, there are still more insights left out in the visualisation that have not been investigated and would be useful for ones with specific preference in individual cryptocurrencies. With this gap, further study could be conducted to understand market in-depth by associating outliers and insight from the visualisations with cryptocurrency fundamental such as technology, code base, features, etc. as studied in literature [13].

Lastly, it should be noted that due to the time constraints and limitation of Twitter API, the data in this study is limited as described in the data preparation chapter. The results found in this study should be remarked with the scope of this dataset.

Chapter 8

Conclusion

We started the study with the hypothesis that network analysis can be used to extract insightful information from the user interactions in the social network. We focus to answer questions that what is the relationship between cryptocurrencies and how they interact to each other with a focus on two aspects from social discussion and user behaviour in line with another additional study about user influence.

Hashtag co-occurrence network was analysed to understand how cryptocurrencies are related together and also to the topics discussed in the social network. Two key findings found from this analysis. First, the ranking of cryptocurrencies in the market is related to the way they are mentioned in the social network. Thus, the higher ranking a cryptocurrency is on the market, the more likely it is associated with other cryptocurrencies and also various topics including general ones. On the other hand, lower ranking cryptocurrencies are less attractive to other cryptocurrencies and also less likely to be associated with general topics. Second, prominent cryptocurrencies, especially Bitcoin(1-BTC) and Ethereum(2-ETH), are by far more likely to be associated with buzzwords and advertisements than the less prominent ones.

Network of users who are active about cryptocurrencies was analysed to understand behaviour of users in cryptocurrency social network. The result shows that users who are interested in altcoins are likely to be interested in other altcoins, while users who like prominent cryptocurrencies, Bitcoin(1-BTC) and Ethereum(2-ETH), show less interest in altcoins but more engagement in general topics. Furthermore, homophily analysis reveals that network of the active users exhibits assortative property where users who have the same tweeting behaviour are likely to connect to each other. Particularly, users who are interested in top-10 cryptocurrencies are more likely to be connected than the users who are interested in the lower-ranking ones.

User influence, measured by retweet count, was analysed by performing linear regression with focus on studying whether users who are tweeting about specific cryptocurrency topics or a wide

range of topics tend to gain more social influence. The results show that less diversity of topics in tweets is statistically significant to retweet. Thus, tweeting by focusing on specific topics is more correlated to gaining more influence in the social network. Similarly, tweeting about Ripple (3-XRP), business, technology, and Ethereum (2-ETH) are most significantly correlated with social influence when comparing to other topics. However, the best regression model shows that tweeting behaviour features are trivial compared to features from user metadata consisting of followers and friends (followees) count. We found that followers count is positively correlated with the number of retweets gained, while friends count is negatively correlated. Thus, the more follower and less friend a user has, the more influence he tends to gain from the social network.

Furthermore, it should be noted that visualisations in this dissertation, which were created from social interactions, can identify outliers and some strong relationships between cryptocurrencies and topics in both aspects of social discussion and user behaviour. This proves that network analysis techniques can extract insights from social interaction in the social network and further investigation on these relationships might reveal some actionable result that helps for investment in the cryptocurrency market.

Bibliography

[1] CoinMarketCap, "Cryptocurrency market capitalizations," 2018.

[2] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions : Detection , Estimation , and Characterization," no. Icwsm, pp. 280–289, 2017.

[3] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," *Www.Bitcoin.Org*, p. 9, 2008.

[4] W. Martin, "Cryptocurrency market passes $700 billion on january 3, bitcoin rises - business insider," 2018.

[5] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," *Cambridge University Press*, vol. 1, p. 116, 1994.

[6] M. Mincer and E. Niewiadomska-Szynkiewicz, "Application of social network analysis to the investigation of interpersonal connections," *Journal of Telecommunications and Information Technology*, vol. 2012, no. 2, pp. 83–91, 2012.

[7] M. Cha, H. Haddai, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter : The Million Follower Fallacy," *International AAAI Conference on Weblogs and Social Media*, pp. 10–17, 2010.

[8] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1031–1040, 2011.

[9] L. Weng and F. Menczer, "Topicality and impact in social media: Diverse messages, focused messengers," *PLoS ONE*, vol. 10, no. 2, pp. 1–18, 2015.

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

[11] A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, and A. Baronchelli, "Evolutionary dynamics of the cryptocurrency market," pp. 1–16, 2017.

[12] N. Gandal and H. Halaburda, "Can We Predict the Winner in a Market with Network Effects? Competition in Cryptocurrency Market," *Ssrn*, pp. 1–21, 2016.

[13] A. Burnie, "Exploring the Interconnectedness of Cryptocurrencies using Correlation Networks," pp. 1–29, 2018.

[14] M. Fleder, M. S. Kester, and S. Pillai, "Bitcoin Transaction Graph Analysis," pp. 1–8, 2015.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet And Web Information Systems*, vol. 54, no. 1999-66, pp. 1–17, 1998.

[16] J. Mern, "Structure and Evolution of Bitcoin Transaction Network," pp. 1–9.

[17] D. Kondor, I. Csabai, J. Szüle, M. Pósfai, and G. Vattay, "Inferring the interplay between network structure and market effects in Bitcoin," *New Journal of Physics*, vol. 16, pp. 0–10, 2014.

[18] J. Kaminski, "Nowcasting the Bitcoin Market with Twitter Signals," no. September 2014, pp. 1–16, 2014.

[19] D. Garcia and F. Schweitzer, "Social signals and algorithmic trading of Bitcoin," *Royal Society Open Science*, vol. 2, no. 9, pp. 1–19, 2015.

[20] S. Colianni, S. Rosales, and M. Signorotti, "Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis," pp. 1–5, 2015.

[21] C. Lamon, E. Nielsen, and E. Redondo, "Cryptocurrency Price Prediction Using News and Social Media Sentiment," 2017.

[22] R. C. Phillips and D. Gorse, "Cryptocurrency price drivers: Wavelet coherence analysis revisited," *PLoS ONE*, vol. 13, no. 4, pp. 1–21, 2018.

[23] "Predicting cryptocurrency price bubbles using social media data and epidemic modelling," *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, vol. 2018-January, pp. 1–7, 2018.

[24] Reddit, "Bitcoin - the currency of the internet," 2018.

[25] M. Linton, E. G. S. Teo, E. Bommes, C. Chen, and W. K. Härdle, "Dynamic Topic Modelling for Cryptocurrency Community Forums," *Ssrn*, 2016.

[26] R. C. Phillips and D. Gorse, "Mutual-Excitation of Cryptocurrency Market Returns and Social Media Topics," no. Ickea, 2018.

[27] T. R. Li, A. S. Chamrajnagar, X. R. Fong, N. R. Rizik, and F. Fu, "Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model," vol. 1, no. 603, pp. 1–9, 2018.

[28] Y. B. Kim, J. G. Kim, W. Kim, J. H. Im, T. H. Kim, S. J. Kang, and C. H. Kim, "Predicting fluctuations in cryptocurrency transactions based on user comments and replies," *PLoS ONE*, vol. 11, no. 8, pp. 1–17, 2016.

[29] Tweepy, "Tweepy api website," 2018.

[30] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," 2009.

[31] J. Schiller, R. A. Srinivasan, and M. Spiegel, "Schaum's outline of probability and statistics, 4th edition," pp. –1, 2012.

[32] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. 1–12, 2008.

[33] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PLOS ONE*, vol. 9, pp. 1–12, 06 2014.

[34] M. E. Newman, "Mixing patterns in networks," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 67, no. 2, p. 13, 2003.

[35] IGraph, "Python igraph manual," 2018.

[36] P. Erdös and A. Rényi, "On random graphs i," *Publicationes Mathematicae Debrecen*, vol. 6, p. 290, 1959.

[37] K. Pelechrinis and D. Wei, "Va-index: Quantifying assortativity patterns in networks with multidimensional nodal attributes," *PLOS ONE*, vol. 11, pp. 1–13, 01 2016.

[38] L. Weng and F. Menczer, "Topicality and impact in social media: Diverse messages, focused messengers," *PLoS ONE*, vol. 10, no. 2, pp. 1–18, 2015.

[39] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Critical Care*, vol. 9, p. 112, Jan 2005.

[40] H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 199–213. New York, NY: Springer New York, 1998.

[41] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, vol. 64, p. 025102, Jul 2001.

[42] M. Yavaş and G. Yücel, "Impact of homophily on diffusion dynamics over social networks," *Social Science Computer Review*, vol. 32, no. 3, pp. 354–372, 2014.