Extracting Cryptocurrency Price Movements From the Reddit Network Sentiment

Stephen Wooley*, Andrew Edmonds*, Arunkumar Bagavathi†, Siddharth Krishnan*
Department of Computer Science
*University of North Carolina at Charlotte**, *Oklahoma State University*†
Email: swooley2@uncc.edu, aedmon16@uncc.edu, abagava@okstate.edu, skrishnan@uncc.edu

***Abstract*—Explosive growth in the value of cryptocurrencies like Bitcoin and Ethereum in recent years has attracted the attention of many speculators. Unlike traditional currencies, cryptocurrencies are not backed by any government agencies resulting in prices being strongly influenced by public opinion. Understanding the relationship between cryptocurrency prices and the public sentiment can lead to improved predictions of price movement. In this paper, we give an exploratory analysis of a network of 24 Reddit communities related to Bitcoin, Ethereum, or other cryptocurrencies to analyze Bitcoin and Ethereum price movements. We engineer a set of 112 time series features from submissions and comments made on the selected subreddits, run Granger causality tests on engineered time series against cryptocurrency price movements, and use these time series to forecast the cryptocurrency price movements using classification models. Results from these models support the Granger causality test results showing that with only lagged price values and lagged values from a single Reddit data derived feature, the direction of Bitcoin and Ethereum price movements can be predicted with 74.2% and 73.1% accuracy respectively.**

***Index Terms*—Cryptocurrency, PageRank, time series analysis, Granger causality, sentiment analysis, Gradient Boosting Machines**

I. Introduction

In 2008, the Bitcoin whitepaper introduced the idea of a peer-to-peer electronic currency which can be used for secure digital transactions without requiring the backing of a third-party institution [1]. The primary purpose of most cryptocurrencies is to enable a secure peer-to-peer market payment service. Using a Blockchain public ledger and proof-of-work concept, the network is safeguarded against fraud by requiring a network-wide agreement to any such fraud [2]. The promise of future widespread adoption and dramatic price fluctuations has attracted both investors and speculators, rapidly driving up prices in recent years. The market capitalization of the largest cryptocurrency Bitcoin, for example, peaked to approximately 325 billion USD in December 2017. Several cryptocurrencies have also exhibited similar rapid growth [3].

User discussions available in social media and other online forums like Wikipedia and Google trends have been tied to study the dynamics of cryptocurrency prices [4]–[9]. Existing works use methods like epidemic modeling in a social network [3] and sentiment analysis in user conversations [10], [11] to make price predictions. Existing work has also shown a medium-term positive correlation between price and online activity and argues that such relationship supports the validity of cryptocurrencies as speculative assets [4].

The website Reddit has been successfully used as a data source used to model user behavior. The forums' structure has been used to model public sentiment [12] and identify influencers [13]. The forums have also been analyzed using graph theory to model user's role and authority within the network [14]. We expand upon the previous work by extracting and engineering features from the network. Our feature set is a blend of temporal and structural properties of Reddit communities. We use machine learning models with these derived features to forecast future cryptocurrency price fluctuations. With exploratory analysis over a variety of features, we show that the models give better forecasting by combining the engineered features from Reddit communities along with features based on past price fluctuations.

We answer following research questions in this work:

RQ-1: Can we identify the top influencers within conversations discussing cryptocurrencies either as technology or a financial instrument? We extract shared user-base from a set of 24 Reddit communities then create a comment author to submission author directed graph. We use PageRank to rank author influence and weigh their sentiment.

RQ-2: Can we derive novel and significant features from the Reddit network model based on the sentiment of user's conversations and user activity? We derive a set of 112 time series conveying changes in the constructed graph structure, content-sentiment across time, and author importance.

RQ-3: Can we determine the significance of the derived features for modeling cryptocurrency price movements? We find statistical significance of derived features with Granger causality tests. Using classification models, we show the incremental accuracy when adding these novel features to models predicting cryptocurrency closing price movement.

II. Background

Social media comments have demonstrated the ability to capture major trends seen in public opinion polls with sentiment analysis models [15]. Sentiment analysis applied to tweets has been succesfully shown to predict upward and downward price movements of the Dow Jones Industrial Average(DJIA) [16] and Bitcoin [17]. Existing works that tie online activity or sentiment to cryptocurrency prices have focused on online forums dedicated to a specific currency [10],

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00093

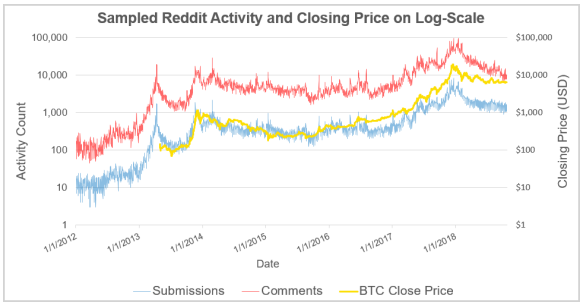

Fig. 1: Two-axis line chart detailing daily counts for Reddit comments and submissions on the left axis versus closing price on the right axis.

[18], tweets containing a specific currency name or hashtag such as "bitcoin" or "#bitcoin" [8], [17], [19], [20], or only included the primary subreddit for a given currency [4].

The Valence Aware Dictionary for sEntiment Reasoning(VADER) [21], was developed to address challenges posed by performing sentiment analysis on microblog-like content. The goal was to develop an algorithm for analyzing social media which was still generalizable, would not need large amounts of training data, and would be computationally fast enough for analysis of streaming data. VADER has been successfully applied to tweets and online forum content with F1 classification accuracy of 96% compared to 84% of human raters [21] and has been used for the purpose of predicting cryptocurrency prices [10], [19].

III. Dataset and Representation

Several researchers have modeled cryptocurrency price movements by using social media forums [8], [9] despite critics claiming that cryptocurrency-users are less active on Twitter [22]. A recent trend is to use the site Reddit as a data-source to study cryptocurrencies [4], [16] due to the longer character limit for posts and availability of the data.

We collected the data for this work from two different sources. The first data source is pushshift.io [1], which contains all Reddit submissions and comments data from January 2012 until October 2018. We use 24 subreddit communities that focus on discussing either Bitcoin, Ethereum, or cryptocurrencies in general. After aggregating the Reddit data, the resulting set includes over 1.7 million unique submissions and 20.6 million comments from 660,000 unique user accounts.

Finally, we use Coinmarketcap [2] for cryptocurrency price data. For each day, the data contain the Open-High-Low-Close prices for each cryptocurrency. The Bitcoin price series begins in April of 2013, while the Ethereum price series begins in August of 2015. The final date of data in both series is October 24, 2018. Figure 1 shows the daily comments and submissions from the sampled 24 Reddit communities along with closing prices from Coinmarketcap.

Changes in price and changes in levels of activity in these subreddits often accompany one another which suggests that the sampled Reddit data used here may exhibit a similar connection between closing price and activity to what has been found in previous works [4]–[7].

[1]http://files.pushshift.io/reddit/

[2]https://coinmarketcap.com/

IV. Graph Analysis

A. Bipartate Graph: Authors and Subreddits

Unlike existing works [4], [10], [18], [20] which focus on a single cryptocurrency, we explore similarity between the primary subreddits associated with the two most popular cryptocurrencies, 'r/bitcoin' and 'r/ethereum'. With this analysis we seek to understand how users are connected to these communities and look for signs of a shared user base among subreddit communities. Our assumption is that if subreddit pages share authors, subsequent analysis can view each subreddit as a part of the larger network without the need to focus only on a single subreddit for each currency.

1) Construction and Characteristics: We represent the Author-Subreddit relationship in a bipartite graph $B = (U, V, E)$ where U is the set of authors, V is the set of subreddits, and E is the set of directed edges connecting nodes in U and V. An edge $E_i \in E$ exists from $U_i \in U$ to $V_j \in V$ when author U_i made a submission or comment in subreddit V_j. The resulting graph has 669,759 authors, 24 subreddits, and 1,073,110 edges. The average degree of the node set U in $B(U, V, E)$ is 1.602, which shows a significant presence of users who are active in multiple Reddit communities within the 24 sampled subreddits.

2) Personalized PageRank: Personalized PageRank is a modification to the PageRank algorithm adjusted to capture the importance of one topic with respect to another topic [23]. The Personalized PageRank algorithm differs from standard PageRank algorithm in such a way that the teleport always moves to a node in set S. From the bipartite graph B, we calculate Personalized PageRank of 'r/bitcoin' and 'r/ethereum' across all members of V. For a stochastic adjacency matrix M of size $|U|$ + $|V|$ x $|U|$ + $|V|$ of a bipartite graph B and probability of teleporting 1 – α, the Personalized PageRank matrix M' contains values given by the formula:

$$M'_{ij} = \begin{cases} \alpha M_{ij} + \dfrac{(1-\alpha)}{|S|} & \text{if } i \in S, \\ \alpha M_{ij} & \text{otherwise.} \end{cases} \quad (1)$$

We use the Personalized PageRank formula to asses the similarity of other subreddit communities to the subreddit pages 'r/bitcoin' (Figure 2a) and 'r/ethereum' (Figure 2b). It is notable from Figure 2a that 'r/bitcoin' is highly similar to the general cryptocurrency subreddit ('r/cryptocurrency') followed by bitcoin focused page ('r/btc') and Ethereum focused subreddits such as 'r/ethtrader' and 'r/ethereum'. Additionally, from Figure 2b we note that the subreddit 'r/ethereum' has the highest similarity to 'r/bitcoin' followed by 'r/cryptocurrency'. Interestingly, Figure 2b shows that only one out of the top six subreddits related to 'r/ethereum' are related specifically to Ethereum while half of these are related to Bitcoin.

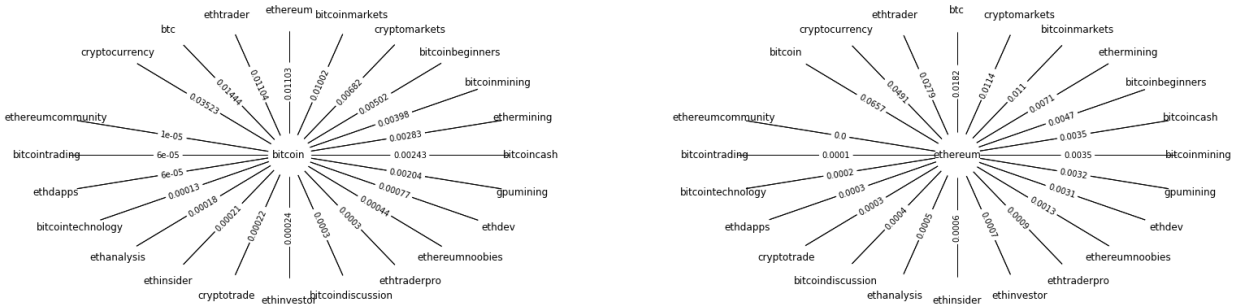

Fig. 2: Personalized PageRank scores relative to 'r/bitcoin' on the left (Fig. 2a) and 'r/ethereum' on the right (Fig. 2b). Figures are ordered in a clockwise direction in a decreasing fashion mirroring the PageRank scores.

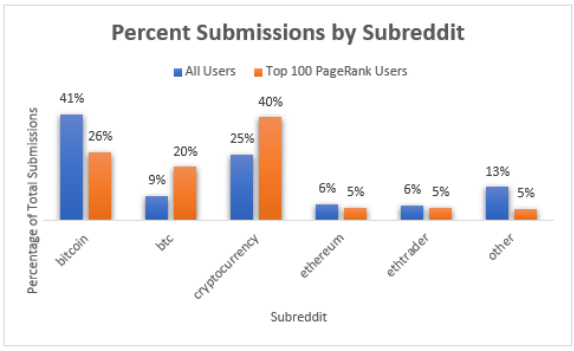

Fig. 3: Bar chart showing distribution of submissions for all users compared to only the top 100 PageRank users by subreddit.

B. Directed Graph: Authors

1) Construction and Characteristics: From the Reddit data we also construct a directed graph $G = (V, E)$ based on author conversations, where V is a set of users who authored a submission, comment, or both and $E_i \in E$ exists from V_i to $V_j \in V$ if V_i comments to a submission authored by V_j, regardless of how many comments connect the two. This edge is intended to capture the interaction of one contributing member of the Reddit community (V_i) exposed to and expressing interest in the submission content of an author (V_j). The constructed graph G has 599,489 nodes and 7,278,390 edges. The average node degree of nodes in the graph $G(V, E)$ is 12.1, showing the average contributing user interacts with about 12 unique author's submissions.

2) Identifying the Most Influential Authors with PageRank: In this work, we use standard PageRank algorithm with probability α = 0.85 of traveling on an edge and probability 0.15 of teleporting to a random node in G to identify influential authors. We examine the top 100 most influential users determined by PageRank versus the general network to understand where important contextual information might be found. Our authority set of 100 users accounts drive disproportionately higher volume of conversations consisting of 51,314 submissions out of 1.7 million total submissions.

From Figure 3, we see that 'r/bitcoin' accounts for the most significant portion of user submissions across the entire set of authors. However, we see that the top PageRank authors make most of their submissions in 'r/cryptocurrency' and 'r/btc'. So while 'r/bitcoin' may contain the largest portion of posts in the greater network, there is likely more highly-regarded content that can be found in some of the other subreddits, specifically 'r/cryptocurrency' and 'r/btc'. Thus, focusing solely on 'r/bitcoin' would exclude key conversations and so aggregating the data across subreddits allow us to capture all of this information.

V. Time Series Analysis

A. Time Series Creation

The directed authors graph G is supplemented with submission titles, submission body, and comments to engineer a total of 112 different time series for use in subsequent causality testing and predictive models. We first create two time series for raw submission and comment counts. Table I shows the number of features created and tested broken down by the feature engineering method. The following subsections detail the method for engineering each of the feature types seen in Table I.

1) Sentiment Time Series: We use VADER [21] to analyze sentiment of the subreddit submission titles, submission body, and comments. A compound score lower than –0.5 is interpreted as negative sentiment and a score above 0.5 is taken as positive sentiment. Furthermore, we define very-negative and very-positive as compound scores less than –0.7 and greater than 0.7, respectively. We construct distinct time series for counts of each sentiment label for submission title, submission body, and comments. We also construct the respective time series for percentage of occurrences of a sentiment out of total possible occurrences. In total we create 12 time series for sentiment counts and 12 time series for percentage of sentiment occurrence. The numbers in Table I reflect these four sentiment categories across the three sources consisting of submission title, submission body, and comments.

TABLE I: Number of Time Series Tested with Both Bitcoin and Ethereum Grouped by Engineering Method and Number Found with 5% Significance for Bitcoin and Ethereum

Time Series Feature Engineering Method	Number Tested	Significant for Bitcoin	Significant for Ethereum
Counts All Authors with Equal Weight	14	1	6
Percentage of Sentiment	12	2	1
Network Characteristics	4	3	2
Nonzero PageRank Counts	14	0	3
Percent Sentiment Nonzero PageRank	12	4	1
Top 50% PageRank Counts	14	1	4
Top 25% PageRank Counts	14	3	2
Counts by Normalized PageRank	14	1	1
Counts by Binned PageRank	14	2	8
Total	**112**	**17**	**28**

2) Network Characteristic Time Series: We also construct time series to represent the network dynamics of the directed authors graph G. Unlike previous research, which uses time series of daily submissions and comments [6], [7], [10], [20], we use time series of number of nodes, number of edges, average degree, and average clustering coefficient of the graph G. This feature set reflects the levels of conversation and discussion for the active Reddit contributors.

3) Time Series Factoring Author Importance Into Daily Values: We use sentiment scores combined with influential-author data based on PageRank scores from section 4.2 to generate the next 3 subcategories of time series features. We build the following time series for the number of submissions, the number of comments, and sentiment time series based on PageRank scores.

a) Time Series Using PageRank to Determine Cutoffs: We generate time series based on three PageRank criterion (non-zero, top 25%, and top 50%). In other words, we derive time series features using only authors who have a PageRank scores satisfying one of these criterion.

b) Time Series Using Normalized PageRank as a Weight: The second category of features is a set of time series features based on normalized PageRank values.

c) Time Series Using Binned PageRank as a Weight: The third category of features is a set of time series features weighing user content by binning the author's PageRank scores into four bins: [0, 0.25], (0.25, 0.5], (0.5, 0.75], and (0.75, 1].

B. Rolling Window Z-Score Transformations

Similar to other works [10], [18], we apply rolling z-score transformations to all of our time series. We use this transformation for two reasons. First, this allows a balanced comparison between time series on different scales such as the cryptocurrency price time series and our other time series like average degree. Second, this transforms each time series to be stationary with constant variance to follow Granger causality hypothesis testing [24].

$$Z_{X_t} = \frac{X_t - \bar{X}(X_{t-k})}{\sigma(X_{t-k})} \tag{2}$$

The rolling z-score uses a rolling window of 'k' days. Here, we take the difference between each daily value and the mean value over the past 'k' days then divide by the standard deviation over the past 'k' days. The result is a stationary time series with mean of 0 and standard deviation of 1.

C. Bivariate Granger Causality Tests

Bivariate Granger causality tests check whether one time series is useful in predicting another time series. For two time series x and y, x Granger-causes y if lagged values of x are statistically significant in predicting the current value of y. We use data from July 1, 2016 to October 24, 2018 for these tests. Let x_t and y_t be the values of time series x and y at time t. Let α_i be the fitted coefficient of y_{t-i} and let β_i be the fitted coefficient of x_{t-i}. Let α_0 be the intercept and ϵ_t be the error at time t. To test whether time series x Granger-causes time series y, we look at the following equation for time series y at time t which is the two-variable model for Granger causality [24] and test the following null hypothesis of non-causality, H_0, with a Wald test [25]:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + ... + \alpha_s y_{t-s} + \beta_1 x_{t-1} + ... + \beta_s x_{t-s} + \epsilon_t \tag{3}$$

$$H_0 : \beta_1 = ... = \beta_s = 0 \tag{4}$$

If a p-value for this test is significant, we can reject the null hypothesis and claim that time series x Granger-causes time series y. As seen in the equation above, the test designed to identify a linear relationship, which means that non-linear relationships could still exist in cases where the null hypothesis fails to be rejected. In our experiments we use Granger causality testing on a total of 112 feature-derived time series against closing prices for Bitcoin and Ethereum, checking both directions of Granger causality.

TABLE II: P-values for a Sample of 5 out of 112 Features Tested for Granger Causality with Bitcoin Closing Price

Lags	*Comment Count*	*Submission Title Negative Top 25% PageRank*	*Average Degree*	*Number of Edges*	*Number of Nodes*
1 day	0.053*	0.158	0.053*	0.029**	0.034**
2 day	0.262	0.234	0.115	0.154	0.205
3 day	0.479	0.090*	0.013**	0.184	0.314
4 day	0.634	0.096*	0.025**	0.299	0.467
5 day	0.682	0.112	0.024**	0.443	0.438
6 day	0.803	0.051*	0.036**	0.573	0.580
7 day	0.868	0.047**	0.052*	0.591	0.566

*Significant at 10% level.
**Significant at 5% level.

D. Granger Causing Cryptocurrency Prices

In this method we use z-score time series, one at a time, in a bivariate Granger causality test against the z-score time series for cryptocurrency closing prices using lags ranging from 1 day to 14 days. This means that each z-score time series is included in 14 different Granger causality tests checking which range of lags up to a maximum of 14 lags are significant. Table II contains p-values from a sample set of Granger causality

tests. The p-values correspond to the Wald statistic, and p-values under 0.05 allow us to reject the null hypothesis that all coefficients are zero with 95% confidence. Table I summarizes the number of time series showing significance at the 5% level based on the feature engineering method and cryptocurrency.

There are 17 time series for Bitcoin and 28 for Ethereum found significant at the 5% level. For Bitcoin, we find average degree, number of nodes, and number of edges to be significant. The latter two are noteworthy because the raw submission and comment counts are only significant at the 10% level. The only network characteristic features which show significance at the 5% level are number of nodes and number of edges. All sentiment categories show significance in at least one engineering method for each currency.

E. Granger Causing Network and Sentiment Time Series

Out of our 112 feature time series, Granger causaility tests show that Bitcoin price Granger-causes 75 time series and Ethereum price Granger-causes 45 at the 5% significance level. Both currencies Granger-cause the number of nodes and number of edges at the 5% significance level. Bitcoin closing price also Granger-causes the average clustering coefficient and the average degree at the 10% significance level. Analysis in other works have shown an inter-individual influence in Bitcoin purchasing due to the interactions of search volume, social media activity, and price [5]. Our Granger causality testing shows that Bitcoin price is not only correlated with measures of activity levels but also network measures such as the clustering and average degree of authors

VI. Predicting Daily Cryptocurrency Price Movements

The Granger causality tests on our engineered feature set shows that there exist significant links between our engineered time series from Reddit and the cryptocurrency price movements. As a next step, we test how well these time series can help predict the movements of Bitcoin and Ethereum closing prices using classification models.

A. Dataset and Processing

We use data from July 1, 2016 through July 24, 2018 as training data. The most recent 3 months of the collected data, from July 25, 2018 to October 24, 2018, is used as test data. Figure 4 shows the training and testing split of the Bitcoin and Ethereum time series. The testing data contains an almost even split of 48 downward price movements and 45 upward price movements for Bitcoin. Ethereum has a greater skew with 56 downward price movements and 37 upward price movements.

We frame the daily price movement prediction as a binary classification model where the goal is to predict a price increase or decrease in the future based on observations from the past. Each data sample here is one day. The independent variables are lagged values from the derived feature time series and lagged values from the closing price time series.

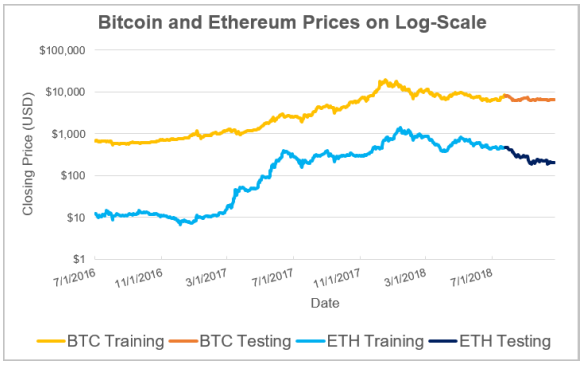

Fig. 4: Line chart displaying log of Bitcoin and Ethereum closing prices during training and testing period.

TABLE III: Top Five Models by Additional Reddit Feature for Bitcoin

Bitcoin Model Features	Accuracy	Sensitivity	Specificity
Submission Counts Top 25% PageRank	74.19%	60.00%	87.50%
Submission Title Top 25% PageRank^{-}	74.19%	88.89%	60.42%
Submission Body Non-Zero PageRank^{-}	70.97%	73.33%	68.75%
Comments Normalized PageRank^{-}	70.97%	66.67%	75.00%
Submission Title Percent^{-}	70.97%	77.78%	64.58%

TABLE IV: Top Five Models by Additional Reddit Feature for Ethereum

Ethereum Model Features	Accuracy	Sensitivity	Specificity
Submission Title Binned PageRank^{--}	73.12%	67.57%	76.79%
Submission Body Non-Zero PageRank^{++}	73.12%	62.16%	80.36%
Comment Counts Non-Zero PageRank	72.04%	64.86%	76.79%
Submission Body Counts^{-}	70.97%	59.46%	78.57%
Submission Title Top 25% PageRank^{++}	70.97%	70.27%	71.43%

$^{-}$denotes time series using only negative sentiment.
$^{--}$denotes time series using only very negative sentiment.
$^{++}$denotes time series using only very positive sentiment.

B. Preliminary Exploration of Classification Methods

We compare the performance of baseline models using only lagged price data against models using lagged price data supplemented by the Reddit-derived features. We make this comparison across four models: *Logistic Regression*, *Gaussian Naive Bayes*, *Support Vector Machines*, and *Gradient Boosting Machines* (GBM). GBM models are an ensemble model of boosted trees. Subsequent decision trees are built by applying weights to incorrectly classified data samples, allowing these subsequent trees to correctly classify a portion of the previously misclassified data samples. We observe gains in accuracy when including the Reddit-derived features. Additionally, GBM classifier outperforms the other models by more than 10% in accuracy. Therefore, we choose GBM classifiers for an in-depth feature performance comparison.

C. Measuring per Feature Accuracy Gains

We build GBM classifiers to mirror the bivariate Granger causality tests. This similar bivariate construction allows the comparison of a baseline model using only past closing price data and one with a single additional feature derived from

the Reddit data to see the incremental prediction accuracy attributed to the additional feature. We follow this strategy for all of 112 engineered time series features. Of the 112 time series used in the Granger causality testing, 39 result in an increase in prediction accuracy when compared to baseline model for Bitcoin, and 30 for Ethereum. Table III includes summary statistics for the five best performing models to predict Bitcoin price movements. Table IV contains results for Ethereum models in the same format.

For both Bitcoin and Ethereum, the feature with the largest incremental gain in accuracy comes from an engineering method which incorporated author-importance through their PageRank score. While the approach in use here does not show that factoring in author-importance is essential to receive equivalent or superior results, the results do show strong gains when using the author-importance derived from proposed directed author graph G. Among the Bitcoin price movement prediction models, the highest accuracy gains are from features which use node-importance from our directed graph to completely exclude content created by less-important authors. These results suggest that the most useful information conveyed through the user content in these Reddit communities originates from a small portion of the community.

VII. Conclusion and Discussion

In this paper, we graphically model Reddit communities related to cryptocurrencies and analyze the relationship between network structure and market fluctuations. Our results show that not only are the Reddit-derived features correlated with price changes, but they contain information that improve market predictions. From our experiments, sentiment time series, average node degree, and average clustering coefficient features are shown to be correlated with observed price changes.

Previous related works [16], [26] have shown results in similar predictions problems by using a Self-Organizing Fuzzy Neural Network (SOFNN) which we would like to apply in a future endeavor to explore more complex non-linear relationships. Additional future work could compare the performance of PageRank with other node importance measures and explore using the Personalized PageRank similarity to generate a dataset for predicting the movements of less popular cryptocurrency. Finally, we also would like to analyze the relationship of network activity to the volatility of cryptocurrency markets by using GARCH models, which have proven useful in modeling market volatility for a variety of instruments [27].

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[2] M. Crosby, P. Pattanayak, S. Verma, V. Kalyanaraman *et al.*, "Blockchain technology: Beyond bitcoin," *Applied Innovation*, vol. 2, no. 6-10, p. 71, 2016.

[3] A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, and A. Baronchelli, "Evolutionary dynamics of the cryptocurrency market," *Royal Society open science*, vol. 4, no. 11, p. 170623, 2017.

[4] R. C. Phillips and D. Gorse, "Cryptocurrency price drivers: Wavelet coherence analysis revisited," *PloS one*, vol. 13, no. 4, p. e0195200, 2018.

[5] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[6] M. A. Nasir, T. L. D. Huynh, S. P. Nguyen, and D. Duong, "Forecasting cryptocurrency returns and volume using search engines," *Financial Innovation*, vol. 5, no. 1, p. 2, 2019.

[7] L. Steinert and C. Herff, "Predicting altcoin returns using social media," *PloS one*, vol. 13, no. 12, p. e0208119, 2018.

[8] M. Matta, I. Lunesu, and M. Marchesi, "Bitcoin spread prediction using social and web search media." in *UMAP Workshops*, 2015, pp. 1–10.

[9] M. Linton, E. G. S. Teo, E. Bommes, C. Chen, and W. K. Härdle, "Dynamic topic modelling for cryptocurrency community forums," in *Applied Quantitative Finance*. Springer, 2017, pp. 355–372.

[10] 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, p. e0161197, 2016.

[11] 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," *arXiv preprint arXiv:1805.00558*, 2018.

[12] A. Park, M. Conway, and A. T. Chen, "Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach," *Computers in human behavior*, vol. 78, pp. 98–112, 2018.

[13] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, and T. T. Kwon, "Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors," in *Proceedings of the 2015 acm on conference on online social networks*. ACM, 2015, pp. 233–243.

[14] C. Buntain and J. Golbeck, "Identifying social roles in reddit using network structure," in *Proceedings of the 23rd international conference on world wide web*. ACM, 2014, pp. 615–620.

[15] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[16] 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.

[17] J. Kaminski, "Nowcasting the bitcoin market with twitter signals," *arXiv preprint arXiv:1406.7577*, 2014.

[18] Y. B. Kim, J. Lee, N. Park, J. Choo, J.-H. Kim, and C. H. Kim, "When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation," *PloS one*, vol. 12, no. 5, p. e0177630, 2017.

[19] L. Kristoufek, "Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era," *Scientific reports*, vol. 3, p. 3415, 2013.

[20] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," *SMU Data Science Review*, vol. 1, no. 3, p. 1, 2018.

[21] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[22] I. Hernandez, M. Bashir, G. Jeon, and J. Bohr, "Are bitcoin users less sociable? an analysis of users' language and social connections on twitter," in *International Conference on Human-Computer Interaction*. Springer, 2014, pp. 26–31.

[23] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the 12th international conference on World Wide Web*. Acm, 2003, pp. 271–279.

[24] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[25] A. Buse, "The likelihood ratio, wald, and lagrange multiplier tests: An expository note," *The American Statistician*, vol. 36, no. 3a, pp. 153–157, 1982.

[26] G. Leng, G. Prasad, and T. M. McGinnity, "An on-line algorithm for creating self-organizing fuzzy neural networks," *Neural Networks*, vol. 17, no. 10, pp. 1477–1493, 2004.

[27] J. Chu, S. Chan, S. Nadarajah, and J. Osterrieder, "Garch modelling of cryptocurrencies," *Journal of Risk and Financial Management*, vol. 10, no. 4, p. 17, 2017.