Signal or Noise in Social Media Discussions: The Role of Network Cohesion in Predicting the Bitcoin Market

Peng Xie, Hailiang Chen, and Jeffrey Hu

Journal of Management Information Systems, forthcoming

August 2020

Abstract

Prior studies have shown that social media discussions can be helpful in predicting price movements in financial markets. With the increasingly large amount of social media data, how to effectively distinguish value-relevant information from noise remains an important question. We study this question by investigating the role of network cohesion in the relationship between social media sentiment and price changes in the Bitcoin market. As network cohesion is associated with information correlation within the discussion network, we hypothesize that less cohesive social media discussion networks are better at predicting the next-day returns than more cohesive networks. Both regression analyses and trading simulations based on data collected from Bitcointalk.org confirm our hypothesis. Our findings enrich the literature on the role of social media in financial markets and provide actionable insights for investors to trade based on social media signals.

Keywords: social media analytics, network cohesion, financial technology, Bitcoin

Introduction

Financial market investors traditionally rely on conventional sources such as company disclosures, market news, analyst reports to obtain information for their investments. As social media becomes an important part of people's daily lives, investors now frequently use information from social media platforms such as Facebook and Twitter to make trading decisions [14, 56]. Empirical studies in the prior literature have also found evidence that information disseminated through social media can be value-relevant [7, 12, 21, 37]. However, extracting value-relevant information from the big data available on social media platforms is not a trivial task, because a lot of social media data can be just noises.

This study explores the possibility of distinguishing between value-relevant information and noise on social media from a network perspective and investigates the role of network cohesion in the relationship between social media sentiment and future market returns. Drawing from prior studies, we argue that directly connected nodes in a network tend to share similar information. Thus, a highly cohesive discussion network with many directly connected messages is likely to have a high level of information correlation, which results in decreased diversity and homogenous opinions and in turn, negatively affects the market return prediction accuracy. Building on these theoretical arguments, we hypothesize a negative relationship between social media discussion network' cohesion and its capability to predict future market returns accurately.

To empirically test our hypothesis we choose the Bitcoin market as our study context. Due to the excitement and enthusiasm for the new Blockchain technology, the price of Bitcoin is largely driven by pure speculation [4]. We conjecture that information from social media is likely to have a larger bearing on the Bitcoin market where there are very limited official information sources compared to the traditional stock market setting.

We collect social media data from Bitcointalk.org, the leading message board dedicated to Bitcoin-related discussions, from December 2012 to June 2017. Our sample consists of more than 500,000 messages posted on 12,441 different threads on the “Speculation” board. This message board offers a quoting feature allowing messages to quote others when they are posted. Most threads are active across multiple days and for each thread/day, we construct a discussion network of messages, in which two messages are linked together if one quotes the other.

We adopt two methods, regression analysis and trading simulation, to evaluate whether network cohesion can be used to help distinguish signal from noise in social media data. We first follow prior literature on the role of social media in financial markets and use regressions to test the predictive value of social media discussions by examining the relationship between next-day Bitcoin return and current-day social media sentiment. Next, we evaluate the moderating effect of discussion network cohesion on this relationship to explore the possibility of distinguishing signal from noise in social media from the perspective of network structure. Specifically, we measure the discussion network cohesion using both average degree and density and investigate whether highly cohesive discussion networks underperform in predicting the next-day Bitcoin return. We then run trading simulations to compare the trading strategies that leverage average degree and density of discussion networks with benchmark strategies such as momentum trading and trading based on aggregate social media signals.

To preview our main results, we find that the association between social media sentiment and next-day Bitcoin return increases significantly when the discussion network cohesion is lower. This result holds for both network cohesion measures (average degree and density) and is robust after controlling for various factors that could potentially affect Bitcoin returns including news sentiment in traditional media outlets, market indexes of other financial markets, investor attention and sentiment, and topic distributions in social media discussions. Our trading simulations also show that the trading strategies based on cohesion weighted sentiment outperform the benchmarks.

Our paper makes several important theoretical contributions. While prior research investigating the role of social media in financial markets has established that social media signals can be helpful in predicting the returns of financial assets, very limited effort is devoted to distinguishing signal from noise within the vast amount of social media data. Our paper bridges this gap by showing that network cohesion can be used to distinguish signal from noise in social media data, and more specifically, less cohesive networks are better at predicting future returns than more cohesive networks. The underlying mechanism is that posts that are more correlated with each other in terms of sentiment and topic coverage are more likely to quote each other, thereby increasing the network cohesion. This lack of opinion diversity in a cohesive discussion network contributes to its underperformance in predicting future returns.

Our paper also contributes to the stream of studies that investigate the effect of network cohesion on network performance. While many studies find high network cohesion to be desirable in various situations such as group performance [40] and knowledge transfer and building [9, 44, 53], others claim high network cohesion to be undesirable because it leads to a high level of redundancy [27, 47, 58]. Our paper shows that network cohesion is associated with information correlation in social media discussion networks and thus negatively affects their return predictability.

Lastly, our paper extends the study of social media's market return predictability from the traditional context of stock markets to the new context of emerging cryptocurrency markets. Due to the special characteristics (e.g., highly speculative) [4] of the cryptocurrency markets, a limited amount of fundamental information is available for investors to evaluate the cryptocurrency' market value. Our study shows that investor sentiments revealed in social media play an important role in determining the price movements of Bitcoins.

Apart from theoretical contributions, our research also offers important practical implications for social media analytics in financial markets. We demonstrate that network cohesion measures can be used to sort out the more valuable information from an enormous amount of social data generated each day. For users who would like to systematically use information on social media for decision-making, our insights can help them filter out low-quality information and allocate their limited resources to focus on the part of social media data that is more likely to be valuable.

Literature Review

It has been well documented in the literature that traditional financial reports and editorial media outlets can predict market price movements. Among these studies, many information channels have been examined. Examples include but are not limited to earnings press release [20], 10-Ks [36], Wall Street Journal columns [51], Dow Jones News Service [52], and IR firm news spinning [48]. With the rapid development of social media in recent years, there is a growing number of related studies examining sentiments extracted from social data (e.g., [7, 12, 21, 37]).

Unlike traditional financial advice sources, social media contents are usually loosely organized and informal. There is much debate about whether social media contents are valuable or just noises. Earlier studies do not find strong empirical support for its merit. Dewally [22] uses buy-and-sell recommendations from an online discussion group to predict stock market but fails to establish the relationship between the recommendations and the market's corresponding performance. Antweiler and Frank [2] study the effects of messages posted on Yahoo! Finance and find only mild influence. Tumarkin and Whitelaw [54] observe investor's message board activity on RagingBull.com but fail to find its association with industry-adjusted returns and abnormal trading volume. However, some other online communities have been shown to successfully predict market movements. Das and Chen [19] examine Yahoo's message board and document a significant relationship with 24 tech-sector stocks. Chen et al. [12] find that Seeking Alpha articles provide value-relevant information for long-term stock returns. Deng et al. [21] find that StockTwits messages significantly influence stock returns at the hourly level. In the

Bitcoin context, Mai et al. [37] study how social media impact Bitcoin values and find that the effects of social media are mainly driven by the silent majority (i.e., less active users).

The current literature suggests that social media discussions can be valuable for predicting returns in financial markets. One central question within this line of research is why social media offers value-relevant information at all given its informal and unregulated nature. Several explanations evolve from different perspectives. Tumarkin and Whitelaw [54] suggest three ways in which information shared on social media can influence readers: (1) the messages contain new information; (2) even if the messages do not contain new information, they at least provide an indication of general market sentiment; and (3) traders may recognize the trading momentum and follow the buy and sell recommendations to exaggerate the effect. Broadly speaking, many factors (e.g. reputation, enjoyment of helping, tenure in the field, and reciprocity) also motivate people to contribute knowledge in social networks [5, 55]. Besides the non-economic motivations mentioned above, economic reasons also exist. Message board viewers' trading decisions can have a price impact and expedite the convergence of market prices to what the message posters perceive to be fair. Because informed investors may not have the financial power to reap all the value conveyed in their private information, they have to stimulate other investors to move the price to the desired direction [24].

Despite that information transmitted through social media is generally valuable for making predictions, most prior studies assess the value relevance of social media data at the aggregate level without filtering, which could result in efficiency loss and unsatisfactory prediction accuracy. This study aims to evaluate the possibility of distinguishing signal from noise in social media from a network cohesion perspective. Network cohesion captures the general level of connectedness in a network [38]. It defines the characteristic as to how connected or fragmented a network is [11]. We use it to measure the connectedness of social media discussion networks in which messages are connected via the replying activities. In the next section, we develop our hypothesis regarding the moderating effect of discussion network cohesion in the relationship between social media sentiment and future Bitcoin return.

Hypothesis Development

Network cohesion has been recognized as a key factor that influences the characteristics of information exchange within a network. Although prior studies have shown that rich connections in social networks convey many benefits, such as reduced information acquisition cost [15], improved information accuracy [23, 32, 46], and boosted efficiency in communication [9], it is also documented that excessive connections in a network could lead to high information correlation [10] because the same information tends to circulate repeatedly within a well-connected network [41, 57]. Such a high level of information correlation will, in turn, result in increased redundancy and reduce the value of the information being transmitted in the network.

Abundant empirical evidence is available in the literature to support this information correlation view. Singh et al. [47] investigate the impact of open source project team network structures on project success. They find that knowledge creation is hampered when the team's external network becomes too cohesive. Information acquired from external resources become relatively redundant when the connections in the network are too dense. Similarly, Mizruchi and Stearn [41] find that a banker's probability to close a deal decreases if his/her personal network is densely connected since a densely connected personal network brings in a more limited range of views and expertise compared to a sparse personal network. Hansen et al. [27] study the performance of product development teams and show that teams with loosely connected networks complete novel tasks more quickly because each team member is more likely to contribute diverse views and knowledge. Lazer and Friedman [34] use an agent-based computer simulation to show that loosely connected groups achieve a higher performance in the long run when dealing with a complex problem because such a network guarantees diversity and is thus better for exploration for solutions.

The positive association between network cohesion and information correlation can be attributed to the inherent preference for similarity in the formation of social networks. It has been well-documented that connections between similar individuals are more probable than connections between dissimilar ones [39, 50] because similarity promotes affiliation, ease of communication, liking, and support [29, 33] while dissimilarity causes cognitive dissonance and uncomfortable psychological states when exposing to inconsistent beliefs and attitudes [45]. Empirical evidence on social media networks also exists to support this point. Adamic and Glance [1] study the links among political blogs during a period of two months before the 2004 presidential election and find that 85% of the links are between blogs of the same political ideology (e.g., Liberal to Liberal, Conservatives to Conservatives). Hargittai et al. [28] provide further evidence by analyzing the links among the blogs of top conservative and liberal bloggers and find more extreme polarization patterns that only 8.7% of these blogs contain links to opposite political communities. Similarly, Conover et al. [16] examine more than 250,000 tweets within six weeks of the 2010 U.S. congressional midterm elections and find that only 13% of retweets occur between different political communities.

The positive association between network cohesion and information correlation implies that a well-connected discussion network tends to embody highly correlated messages that bring in limited expertise and perspectives [30]. In contrast, a loosely connected discussion network facilitates information diversity [47] since each message in such a network is more likely to serve as a bridge to a different information source [27]. Information diversity improves the accuracy of the overall opinion of a discussion network because exposure to a diverse set of views encourages error correction and discrepancy resolution [17, 18, 26].

The collective wisdom of social media discussion networks generally improves with every additional message posted under the discussion thread. But beyond volume, non-redundant information enhances the collective wisdom further because it ensures information diversity. In our research context, high cohesion social media discussion networks encompass densely connected messages due to frequent replying activities. Such discussion networks are expected to contain highly correlated information, which reduces the overall informational value of the discussion networks and negatively affects their predictive accuracy for future market returns. In contrast, low cohesion social media discussion networks encompass sparsely connected messages from infrequent replying activities. Such discussion networks are expected to contain more divergent views, which enhances the informational value and prediction accuracy of the networks. Given two networks of an equal size (as measured by the number of messages), the one with low cohesion contains more heterogenous opinions and thus provides more informational benefits. In light of these arguments, we propose our hypothesis below.

Hypothesis: The cohesion of a social media discussion network negatively affects the discussion network's return predictability.

Data

Bitcoin

Our analyses are based on the prediction of Bitcoin market movement. Bitcoin is a decentralized peer-to-peer electronic payment platform. It is a web-based system that enables users to transfer value across the globe quickly and anonymously without the need for third-party verification [49]. The exchange rates between Bitcoin and fiat currencies are decided at specialized exchanges.

To track Bitcoin price movements, we adopt the Bitcoin Price Index (XBP) created by CoinDesk that is an average of Bitcoin prices across leading global exchanges. The XBP has been used by many companies such as New York Times, Wall Street Journal, Nasdaq, and so on. Similar to foreign exchange markets, the Bitcoin market is open 24 hours a day and seven days a week. We use Greenwich Mean Time (GMT) to define trading days and analyze the end-of-day closing XBP index (i.e., the price index at 24:00 GMT each day). The day *t* Bitcoin return is calculated as $(P_t - P_{t-1})/P_{t-1}$, where P_t is the close Bitcoin Price Index on day *t*.

Our data spans the period from December 1, 2012 to June 30, 2017, including 1,667 trading days (six days with no social media activity are removed). We choose December 1, 2012 as the start date because the Bitcoin price remained very low in its early years and the market was too small to attract enough public attention. The turning point occurred at the end of the year 2012 when Bitcoin quickly increased in value. Table 1 presents the descriptive statistics of Bitcoin-related variables. The radical expansion of the Bitcoin market is evident in Panel A of Table 1. The market capitalization of Bitcoin grew by more than three hundred times during the study period, averaging an increment of 28 million USD per day. The growth rate of Blockchain wallet users tells a similar story: Bitcoin market is gaining popularity rapidly in recent years.

[Table 1]

Panel B of Table 1 presents descriptive statistics for XBP and returns. Bitcoin market is very volatile, especially in the earlier years. However, with the development of the Bitcoin ecosystem, price volatility has decreased in recent years. The standard deviation of the XBP return is 7.28% in 2013, 3.91% in 2014, 3.56% in 2015, 2.53% in 2016, and 3.92% in the first half of 2017.

Panel C of Table 1 presents the descriptive statistics for trading volumes on the Bitcoin market. Within the study period, the value of Bitcoins being traded increases by almost half a million dollars ($448,104) per day. Over 1 billion dollars' worth of Bitcoins are traded during the most active day.

Bitcointalk.org

Bitcointalk.org is our primary source of social media data. It is a leading message board for Bitcoiners to share thoughts on various Bitcoin-related topics. This type of online communities is becoming an increasingly important information source for investors as the number of participating users continues to grow [26]. As of October 2019, Bitcointalk.org has accumulated 2,686,933 registered users and an average daily page view of 1,320,603 times; it receives on average 7,223 posts each day.[1] The timestamps of all posts are also based on GMT, so we have a consistent time frame for Bitcoin Price Index and social media data.

There are 239 boards on Bitcointalk.org at the time of writing, and each is dedicated to a particular topic such as technical issues, Bitcoin minings, and so on. However, many of these discussion boards are neither popular nor directly related to the Bitcoin market performance. In this research, we focus on the "Speculation" board, which is one of the most frequently posted boards and also dedicated to discussions on Bitcoin price movements and speculations. Our dataset contains more than 500,000 messages posted under 12,441 different threads in the Speculation board from December 1, 2012 to June 30, 2017.

[Figure 1]

Every registered user can start new threads. After the creation of a new thread, other users can join the discussion by sharing their views. There is a communication-enabling feature called

[1] For details, please refer to the statistics center of Bitcointalk.org, accessed October 26, 2019, https://bitcointalk.org/index.php?action=stats. In this paper, we use "posts" and "messages" interchangeably.

"quote" on Bitcointalk.org. An author may quote one or multiple existing messages when writing a new message, and then a link to each quoted message is added to the new message. Figure 1 shows an example of the discussion facilitated by the quoting function on this social media platform. Our discussion networks are constructed based on the quoting activities observed between messages under the same thread. Note that a new message can only quote previously posted messages, so the network we constructed is a directed network.

We follow the literature and assess the sentiment expressed in texts by calculating the fraction of negative words in the messages [12, 13, 36, 51, 52]. We adopt the negative word list developed by Loughran and McDonald [36] in our study. Also consistent with the text analysis literature in accounting and finance, we report our results based on the fraction of negative words instead of the fraction of positive words. When calculating the sentiment of a quoting message, the quoted message is not counted, even though the quoted message is embedded in the quoting message (see Figure 1 for an example).

We collect all messages under the 12,441 threads from the Speculation board. Panel A of Table 2 presents descriptive statistics for all the threads. Thread Duration is the time span between the first post and the last post within a thread. The distribution of Thread Duration is highly skewed. Except for a few threads that remain active for a long time, investors quickly lose interest in a thread and the discussion desists. "% 1st Day Post" is the percentage of first-day posts. Our data indicate that on average 66.68% of all discussions in a thread are posted on the first day.

[Table 2]

To test our hypothesis, we construct the thread-day networks (i.e., separate networks are created for each thread-day pair). Thread-day networks are used for two considerations: (1) different threads are likely to focus on different issues, and each has a unique impact on future price movements, and (2) information can be time-sensitive, the relevance of even the same topic may vary significantly over different days. Within each thread-day network, if message A quotes message B, an edge is created directing from A to B. A message may quote other messages or get quoted at the same time. Figure 2 shows examples of two real thread-day discussion networks with the same number of messages, but one network is more cohesive than the other. It is noteworthy that in the thread-day discussion networks, each message, rather than each author, is treated as a node since an author can post multiple messages in the same thread-day network and message is at a more granular level that enables us to capture more details.

[Figure 2]

A total of 60,202 thread-day networks are constructed. However, many of these networks contain only one post, which is insufficient to form a network. Therefore, we only focus on thread-day networks with more than one post, and there are 48,201 of them. Panel B of Table 2 presents descriptive statistics for the thread-day networks with more than one post. *PostCount* is the total number of messages in a thread-day network. The 25% percentile of *PostCount* is 3, meaning that a quarter of the observations are very small discussion networks. Because content quality (such as relevance, being informative and persuasive, humor, emotional appeal, and creativity) is important in attracting user attention and encourage user engagements in social media platforms [3, 35], we expect that discussion networks containing fewer messages are less likely to be value-relevant. Receiving a small number of messages is a signal that this thread is not appealing to investors, or it contains redundant topics already discussed elsewhere. Furthermore, calculating cohesion measures for networks with two nodes can create outliers that are extremely cohesive. For these reasons, we primarily focus on thread-day discussion networks with *PostCount* greater than two in our analyses.

We measure the discussion network cohesion with both average degree and density. *Average Degree* is operationalized as the total number of quotations divided by *PostCount* in a thread-day network. *Density* is operationalized as the total number of quotations divided by $PostCount \times (PostCount-1)$, because the discussion network is defined as a direct network. Note that in an undirected network, the denominator would be $NodeCount \times (NodeCount-1)/2$. One drawback of density as a measure of network cohesion is that it tends to be small for larger networks because the denominator grows with a polynomial rate when the number of nodes increases. In this regard, average degree is more "immune" to the network size than density [9].

Dow Jones Newswires

To account for traditional media news and mainstream discussions on Bitcoin, we collect news articles from all sources in Dow Jones Newswires (DJNS) through Factiva. We search for the "Bitcoin" keyword and manually downloaded all English articles published in our study period. Similar as social media sentiment, we assess the content and sentiment of news articles using the fraction of negative words. For each day, *DJNSCoverage* is an indicator variable denoting whether there is any news article on Bitcoin; *DJNSSentiment* is the average fraction of negative words across all news articles on Bitcoin when there is news coverage and zero otherwise.[2] Panel D of Table 1 presents the descriptive statistics for DJNS media coverage on Bitcoin.

Other Control Variables

We also control for a series of variables identified in related studies on the Bitcoin market (e.g., [37]). We include the returns on S&P 500 index and COMEX gold price in our regression model to capture the general investing trend, the Alexa website traffic rank for "Bitcointalk.org" and the Google Trends index for the keyword "Bitcoin" to capture the popularity of Bitcoin over time, the AAII Investor Sentiment to capture the general sentiment of individual investors, and the VIX index for the market's expectation of volatility. Because some of the control variables are not at the daily level, we assign the latest available values prior to day *t* to them in the model. Panel E of Table 1 presents the descriptive statistics for all these control variables.

Empirical Analysis

Social Media Sentiment and Bitcoin Return

Our task is to investigate the role of network cohesion in distinguishing signal from noise in social media discussions. Before introducing network cohesion into our analysis, we first evaluate whether and how social media sentiment offers value-relevant insights into the next-day Bitcoin returns. Using the thread-day panel data, we regress the next-day return on social media

[2] Our results do not change if *DJNSSentiment* is set to be a different value (e.g., its mean) when there is no news coverage.

sentiment and all control variables. The baseline analysis is conducted using the following panel data model:

$$R_{t+1} = \alpha + \beta_1 Sentiment_{it} + \delta X + \eta_{it}. \quad (1)$$

The dependent variable is the next-day Bitcoin return R_{t+1}.[3] On the right-hand side, $Sentiment_{it}$ is the average sentiment across all posts under thread *i* on day *t*. The coefficient estimate for $Sentiment_{it}$ reflects the effect of social media sentiment on the next-day return. *X* contains the control variables, which include the same-day return R_t on day *t*, the one-day lagged return R_{t-1}, the two-day lagged return R_{t-2}, the cumulative return over the past month $R_{t-30,t-3}$, the logarithmic transformation of the same-day trading volume $Log(TradingVolume_t)$, the logarithmic transformation of the number of posts in the thread-day network $Log(PostCount_{it})$, the logarithmic transformation of the number of page views of the thread $Log(PageView_i)$ at the time of data collection[4], the traditional media sentiment $DJNSSentiment_t$, the traditional media coverage $DJNSCoverage_t$, the return on S&P 500 index $S\&P500_t$, the return on COMEX gold price $COMEXGold_t$, the logarithmic transformation of Alexa web traffic rank for "Bitcointalk.org" $Log(AlexaRank_t)$, the Google Trends index for keyword "Bitcoin" $GoogleTrends_t$, AAII investor sentiment $AAIISentiment_t$, the VIX index VIX_t, and weekly dummies that control for time effects.

[3] Note that our data is a thread-day panel with *i* denoting a thread and *t* denoting a day. The dependent variable should be written as $R_{i,t+1}$, but since we are using different threads to predict the same next-day return, we simply write the dependent variable as R_{t+1}. The same notation applies to some of our control variables too.
[4] $Log(PageView_i)$ is mainly to control for the popularity of different threads. We do not have historical data on the number of page views for a thread on a specific day. We assume most page views occur within the days when there are new posts made under a thread.

There are 12,441 different threads in the data and for each thread the duration is relatively short (a thread lasts for 5 days on average). During our study period, about 300 new threads were created every day on this social media platform. This rapid creation of new threads resembles random draws from a large population of different thread topics, thus a random effect model would be more appropriate for our analysis [25]. In particular, we estimate a random effects model with standard errors clustered by thread.

[Table 3]

The estimation result of Equation (1) is shown in Table 3. The coefficient estimates for $Sentiment_{it}$ in Columns (1) and (2) are not statistically significant when we include small discussion networks in the estimation. The same coefficient estimate in Column (3) becomes negative and statistically significant at the 5% level when we focus on relatively larger networks (i.e., when the *PostCount* is greater than 6). The coefficient estimate for $Sentiment_{it}$ in Column (3) indicates that if the fraction of negative words in a thread-day discussion network is 1% higher, the next-day Bitcoin return is 0.091% lower. In short, the association between social media sentiment and next-day return becomes stronger and statistically significant when the discussion network is larger.[5] This is expected because when a thread receives very few posts, the discussions are very likely not value-relevant or interesting to other investors. On a related

[5] In untabulated analyses, the significance of the predictability remains at least at the 5% level when the *PostCount* threshold is raised to 7, 8, and 9 but diminishes to the 10% level as the *PostCount* threshold reaches 10, which severely cuts down the sample size.

note, this result also validates the notion that many messages posted on social media may be just noises and filtering out them could potentially lead to an improvement in prediction accuracy.

The results for the control variables in Table 3 are consistent with expectations. The sentiment revealed from news articles in Dow Jones Newswires has a strong association with the next-day Bitcoin return, and the effects are all statistically significant at the 1% level in all columns. Specifically, if the fraction of negative words in DJNS news articles is 1% higher, the next-day Bitcoin return is 0.148% to 0.158% lower. Media coverage in DJNS is negatively associated with the next-day Bitcoin return, implying that when there is media coverage on Bitcoin, it is on average more likely to be negative and drives down the Bitcoin price. Both past returns and trading volume are also negatively associated with future returns. We also observe a statistically significant negative relationship with the next-day Bitcoin return for both the S&P 500 return and COMEX gold price return.

The Role of Network Cohesion

After validating the value of social media sentiment in predicting returns in the Bitcoin market, we proceed to the main analysis of this study: testing the moderating effect of network cohesion in the relationship between social media sentiment and future Bitcoin return. To test our hypothesis, we compare the return predictability of discussion networks with different levels of network cohesion and test whether less cohesive networks are more accurate in predicting future Bitcoin returns than more cohesive networks. For this purpose, we add network cohesion and an interaction term between social media sentiment and network cohesion to Equation (1):

$$R_{t+1}=\alpha+\beta_1 Sentiment_{it}+\beta_2 Cohesion_{it}+\beta_3 Sentiment_{it}\times Cohesion_{it}+\delta X+\eta_{it}. \quad (2)$$

Results are reported in Table 4. Two different measures are used to capture the level of network cohesion: average degree and density. Columns (1) to (3) of Table 4 present the results when cohesion is measured by average degree, and Columns (4) to (6) of Table 4 present the results when cohesion is measured by density.

[Table 4]

Because social media sentiment is measured using the fraction of negative words, return predictability is reflected in a negative coefficient estimate on the sentiment measure. If high network cohesion weakens the prediction accuracy of a social media discussion network, we expect a positive coefficient estimate on the interaction term $Sentiment_{it}\times Cohesion_{it}$.

The coefficient estimates for the interaction term $Sentiment_{it}\times Cohesion_{it}$ in all columns are positive and statistically significant at least at the 5% level. Column (3) indicates that for large discussion networks with size above the median, when the fraction of negative words is 1% higher in discussion networks with an average degree of 0.534 (sample mean), the next-day Bitcoin return is 0.059% lower (-0.059%=(-0.297+0.445×0.534)×100%). In contrast, when the discussion networks' average degree is one standard deviation (0.207) lower, the same 1% increase in the fraction of negative words is associated with a 0.151% decrease (-0.151%=(-0.297+0.445×(0.534-0.207))×1%) in magnitude in the next-day Bitcoin return. In other words, the association between social media sentiment and future returns becomes much stronger as the network becomes less cohesive. We obtain very similar results when cohesion is measured by density as shown in Column (6). The association between a 1% increase in the fraction of negative words and the next-day Bitcoin return is -0.096% (-0.102% = (-0.245+3.109×0.048)×1%) when the network density is at the sample mean of 0.048, but this association becomes -0.198% (-0.219% = (-0.245+3.109×(0.048-0.033))×1%) when the network density is one standard deviation (0.154) lower. The average improvement in return predictability at the daily level seems mild even though it doubles as the network cohesion is one standard deviation lower, but if we consider the compounded returns over a long time, the difference can be huge. In sum, these results support our hypothesis that the cohesion of a social media discussion network negatively affects its accuracy in predicting future Bitcoin returns.

It is important to note that the coefficient estimates for $Sentiment_{it}$ and $Sentiment_{it} \times Cohesion_{it}$ together in Table 4 also imply that the effect of sentiment (measured by fraction of negative words) on future return can become positive when the discussion networks have high cohesion. In other words, when the underlying sentiment is more negative, the return is however more positive. Put it differently, the prediction made based on information extracted from highly cohesive networks can be completely wrong or even opposite. This result is consistent with our claim that less cohesive networks are better at predicting future returns than more cohesive networks and that there are a lot of noises in social media data so we need to be selective in order to identify the valuable information. In Table OA1 of Online Appendix, we report a subsample analysis to further validate our result.

Trading Simulation

In this section, we verify our findings with trading simulations using the same dataset. Kastens and Schroeder [31] suggest a set of rules to be followed in trading simulations. These rules include (1) the trades should be frequent enough during the trading period; (2) if some parameters of the trading strategy are generated using historical data in a certain period, the trading simulation should avoid that period; (3) the trading strategy cannot be too complex, otherwise it would be less practical in reality; (4) the trading strategy has to be historically realistic. It should be reasonable and possible for a trader to react in a specific fashion given the conditions of that trading moment; and (5) the trading profits should be compared to a benchmark. We design five trading strategies based on the above-mentioned guidelines: random, past returns, equally weighted sentiment, average degree weighted sentiment, and density weighted sentiment. The first three strategies serve as the benchmark for comparison. The second rule does not apply in our situation since we do not train any model to generate parameters for our trading strategies.

Suppose an investor starts with $1 and makes investment from January 1, 2013 to June 30, 2017. The performance of each trading strategy is measured by the average end-of-period balance across 10,000 trading repetitions. We first consider the case in which short selling is not allowed. Under random investing, the investor has a 50% chance to invest the entire balance during each day, and avoid the market (i.e., no position) otherwise. Under investing based on past returns, we adopt the "momentum-style" trading whereby positive (negative) returns are presumed to be followed by continued positive (negative) returns [42]. Specifically, the investor invests the entire balance during the current day with a probability equal to the percent of days with positive returns in the past (starting from December 1, 2012, the beginning of our data collection period, to the current day), and holds no position otherwise. No future information is used in this trading strategy, which conforms to the historical realism rule [31]. Under investing based on equally/average degree/density weighted sentiment, the investor makes trades based on social media signals. The investor first observes the percentile of the current day's social media sentiment in all available history, and then invests in the next day with a probability associated with that percentile. For example, if the current day's sentiment (measured by the fraction of negative words) is at the 10th percentile of the entire history (i.e., the sentiment is quite positive), the investor will invest the entire balance in the next day with a 90% chance, and avoid the market with a 10% chance. Again, the rule of historical realism is upheld and the behavior of the investor is reasonable and justifiable. When calculating the average degree (density) weighted sentiment, the discussion networks with a higher average degree (density) are given a lower weight because they are expected to be less informative based on our findings. Specifically, if the average degree (density) of a discussion network is at the *x*th percentile among all discussion networks in a certain day, the weight assigned to the sentiment of this particular discussion network is 1-*x*%.

As Bitcoin made astronomical gains during recent years, many investors started taking a bearish view. Shortly after the beginning of our study period, Bitcoin short-selling becomes possible (e.g., margin trading, option trading, prediction markets) through Bitfinex and ICBit [43]. Therefore, we also consider the possibility of Bitcoin shorting. In this set of simulations, instead of the choice to avoid the market by holding no position, the investor can short a number of Bitcoins equivalent to the worth of the current balance whenever needed. Since the highest actual return is still less than 100% in the study period, the investor will never go bankrupt.

To summarize, the end-of-period balance for each strategy is calculated as follows. Based on the probability to invest (p_t) during each day, we generate a random trading signal S_t. The trading signal S_t is either 1 (trade with a long position) or 0 (not trade when shorting is not allowed or trade with a short position when shorting is allowed), and the expected value of S_t is p_t across all 10,000 repetitions. When shorting is not allowed, the end-of-period balance of the investment given the initial investment of \$1 is $Balance_T = \prod_{t=1}^{T}(1 + S_t R_t)$, where the R_t is the Bitcoin return on day t. When shorting is allowed, the investor will take a long position if $S_t = 1$ and a short position if $S_t = 0$. The end-of-period balance of the investment given the initial investment of \$1 is then $Balance_T = \prod_{t=1}^{T}[(1 + R_t)\, S_t + (1 - R_t)(1 - S_t)]$.

The results of the trading simulation are shown in Table 5. Panel A presents the results when shorting is not allowed, while Panel B presents the results when shorting is allowed. Among the five strategies, both average degree and density weighted sentiment strategies consistently outperform the other three benchmark strategies regardless of whether shorting is allowed or not. The average end-of-period balance for the cohesion weighted sentiment strategy can reach \$50 or more when shorting is not allowed, although the end-of-period balance is much lower when shorting is allowed. Since Bitcoin price is mostly increasing during the study period, shorting Bitcoin is usually not a wise action retrospectively. Despite this, we can still achieve an average return of 300% to 400% with a trading strategy based on the cohesion weighted sentiment strategy. The density weighted sentiment strategy also achieves the largest maximum ending balances of $1,067 without shorting and $830 with shorting. Among the three benchmark strategies, the random strategy always leads to the lowest average end-of-period balance in both panels. The median of the equally weighted sentiment strategy is larger than the median of the past returns strategy, but the relationship between the two in terms of the average is reversed. In the rightmost column of Table 5, we also present the percentage gains of each strategy against the equally weighted sentiment strategy. Even compare with the best benchmark strategy, the cohesion weighted sentiment strategy is able to achieve a gain of 35% to over 100% in different scenarios. Therefore, these results demonstrate that trading based on network cohesion weighted sentiment yields much better profits than existing trading strategies.

[Table 5]

Topic Modeling

To distinguish value-relevant information from noise, this study focuses on the role of network cohesion in social media discussion networks. To test our hypothesis, we have investigated how network cohesion moderates the association between social media sentiment and future Bitcoin return. One concern is that the topics of social media discussions may influence the Bitcoin returns and the discussion network cohesion at the same time. For instance, certain topics are more value-relevant and may have a larger impact on returns, and these topics are naturally associated with certain level of network cohesion. Without controlling for the discussion topics, our analyses may suffer from omitted variable bias and identify a spurious relationship.

To address this concern, we adopt topic modeling techniques to measure the topic distributions of each thread-day discussion network and then add all topic weights within each network as control variables in the regression model. The basic idea of topic modeling in textual analysis is that documents are represented as a distribution over a certain number of latent topics, and each topic is characterized by a distribution of words. We employ Latent Dirichlet Allocation (LDA) proposed by Blei et al. [6] to identify the latent topics from the social media messages and use them to infer the topic weights of the thread-day networks.

For this purpose, we group all messages under a thread-day discussion network together as one document and generate a set of documents as the input for LDA. To concentrate on the words that are meaningful for identifying topics, we remove stop words and high-frequency context-specific words such as “bitcoin” and “btc”. To improve the interpretability of the topics, we also remove numbers and dates in the text. We also drop the thread-day discussion networks that have too few words (i.e., documents that contain less than 25 words).

To find the optimal number of topics, we fit the LDA model with different numbers of topics and calculate the perplexity score on a held-out test dataset using 10-fold cross-validation. A lower perplexity score indicates a better language model performance and a good rule of thumb is to pick a number of topics that produces a reasonable output and after the perplexity has started to decrease at a slower rate.[6] As shown in Figure OA1 of the Online Appendix, the decrease in the perplexity score becomes much slower after the number of topics reaches 40. Thus, we choose 40 as the number of topics in order to fit a parsimonious model.[7] To illustrate the topics uncovered by the LDA model, we provide the top ten most frequent words for each topic in Table OA2 of the Online Appendix.

The LDA model also produces a topic distribution over the 40 topics for each thread-day discussion network. To control for the topic coverage of social media discussions, we add these 40 topic weights as additional control variables to Equation (2) and re-estimate our main model as in Table 4. The new results are presented in Table OA3 of the Online Appendix. In short, our main findings remain robust after controlling for the topic coverage of social media discussions.

Information Correlation and Network Formation

Through regression analysis and trading simulation, we have shown that less cohesive social media discussion networks are more accurate in predicting future Bitcoin returns than more cohesive networks. Our key argument for this empirical result is that information correlation plays an important role in the moderating effect of network cohesion. In this section, we provide further evidence to support our argument by studying how information correlation affects network formation, and more specifically, the quoting behavior between two messages. A network becomes more cohesive when messages within this network quote others more often.

[6] For details, please refer to "Stanford Topic Modeling Toolbox", *The Stanford Natural Language Processing Group*, accessed September 27, 2019, https://nlp.stanford.edu/software/tmt/tmt-0.4/.
[7] Our results remain largely the same if a different number of topics is chosen and are available upon request.

For this purpose, we run a logistic regression to test if information correlation increases the probability of quotation between two messages after controlling for other factors that may also affect the dependent variable.

To capture information correlation between a pair of messages, we define the following two measures. $OpionionDistance_{i,j}$ is the absolute difference between the sentiments of message i and j. $TopicSimilarity_{i,j}$ is the cosine similarity between the topic distribution vectors of message *i* and message *j*. A low opinion distance and a high topic similarity indicate high information correlation between two messages.

The dependent variable of the logistic regression is $Quotation_{i,j}$, which is 1 if message j quotes message *i* and 0 otherwise. Quotation only occurs between messages under the same thread, and a message can only quote messages posted prior to itself. For a thread with *N* messages, the number of pairs is $N\times(N\text{-}1)/2$. In our sample of 12,441 threads, the total number of possible pairs is 653,215,410. To reduce the sample size, we utilize the fact that a quoting message is more likely to quote those recent messages before itself but unlikely to quote those messages that are very far ahead. For instance, 98.6% of all actual quotations occur within a 100-mesasage range (i.e., there are no more than 98 messages between a quoting message and a quoted message in the timeline of a thread). Even if we lower the threshold from 100 to either 20 or 30 messages, the proportion of such quotations remains as high as 91.8% and 94.6%, respectively. Therefore, for our test we construct three datasets in which we include all pairs of messages that are no farther than 20, 30, or 100 messages away, respectively.

In the logistic regression, we also control for other factors that affect the chance of message *j* quoting message *i*. These factors can be divided into three groups: variables concerning both messages *i* and *j*, variables about message *i* only, and variables about message *j* only. The first group includes two variables. $log(TimeLag_{i,j} + 1)$ is the logarithm of the timestamp difference (measured in seconds) between message *i* and message j. $log(MessagesInBetween_{i,j})$ is the logarithm of the number of messages between message *i* and message *j* when all messages under the thread are sorted by posting time. The second group of control variables about message *i* includes: (1) $Sentiment_i$, which is the fraction of negative words in message *i*; (2) $log(Length_i)$, which is the logarithm of the number of words in message *i*; (3) $log(AuthorActivitiy_i + 1)$, which is the logarithm of the activity score[8] of message *i*'s author at the time when message *i* was posted; (4) $TopicConcentrationHHI_i$, which is the Herfindahl–Hirschman Index calculated based on message *i*'s topic distribution and used to capture the concentration of the topic distribution (i.e., the larger $TopicConcentrationHHI_i$ is, the more concentrated in a few topics message *i* is; (5) $CitingOther_i$, which is a binary variable set to 1 if message *i* itself quotes some other messages and 0 otherwise; and (6) $Log(CitedCount_i+1)$, which is the logarithm of the number of times message *i* is quoted by messages posted before message *j* along the time line of the thread. The third group of control variables about message *j* is defined similarly as the group for message i, except that $Log(CitedCount_j+1)$ is not included because

[8] Activity score is defined by the forum site as min(*time*×14, *posts*), where *time* is the number of two-week periods in which the user has posts since registration and *posts* is the total number of posts made by the user.

message j cannot be quoted before it is posted. When any of the above control variables has zero obervations, we add 1 to it before the log transformation. The descriptive statistics of the variables are presented in Table OA4 of the Online Appendix.

The model estimation results are presented in Table 6. In all three columns, the coefficient estimates for $OpionionDistance_{i,j}$ are negative and statistically significant at the 1% level, and the coefficient estimates for $TopicSimilarity_{i,j}$ are postiive and statistically significant at the 1% level. The economic significance is also meaningful for both information correlation measures. For example, in the sample of $MessagesInBetween_{i,j} \leq 100$ (Column 3), when everything else being equal, a 1% increase in the magnitude of opinion distance between two messages leads to a 3.07% decrease (=exp(-3.025/100)-1) in the odds of quotation between them, and a 10% increase in the magnitude in topic similarity between two messgaes leads to a 9.05% increase (=exp(0.866/10)-1) in the odds of quotation between them.[9] These results show that quotation between messages is more likely if they express similar opinions (in terms of sentiment) and contain similar topics.[10] A dense discussion network with more quotations is thus expected to have a higher level of information redundancy and lack opinion diversity. This finding serves as strong evidence that information correlation indeed plays an important role in

[9] The standard deviation of opinion distance and topic similarity is 4.4% and 30.7%, respectively.
[10] In addition to our key variables of interest, we have the following interesting findings from the control variables: (1) negative messages are more likely to be quoted; (2) longer messages tend to be quoted; (3) messages covering a broader range of topics are more likely to be quoted; (4) messages posted by users with less posts or shorter membership tenure are more likely to be quoted; (5) messages quoting other prior messages or already quoted many times by other messages are more likely to be quoted; (6) quoting messages tend to be negative in sentiment; (7) quoting messages tend to be longer and also cover a broader range of topics; (8) quoting messages are more likely to be written by users with more posts or longer membership tenure; and (9) any message becomes less likely to be quoted as time passes by and as more messages are posted within the same thread.

explaining why less cohesive discussion networks are more accurate in predicting future returns than more cohesive networks.

*[*Table *6]*

Conclusion

This study investigates the role of network cohesion in distinguishing signal from noise in social media discussions. By analyzing the emerging and highly speculative Bitcoin market, we empirically document a negative relationship between social media discussion network cohesion and its prediction accuracy for future Bitcoin price movement. Information correlation plays an important role in explaining this result. Social media posts that are similar in opinion sentiment and topic coverage tend to form a tie in discussion networks, which results in increased information redundancy and decreased information diversity. Our findings can help investors extract valuable signals from social media information based on network structure and make more reliable investment decisions. Future research can explore other methods to sort out valuable information from the increasingly large amount of social media data generated every day to support decision making in various contexts.

References

1. Adamic, L. A., and Glance, N. The political blogosphere and the 2004 US election: divided they blog. *3rd International Workshop on Link Discovery* (*LinkKDD*), 2005, pp. 36-43.
2. Antweiler W. and Frank M.Z. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59, 3 (2004), 1259-1294.
3. Ashley C., and Tuten T. Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. *Psychology & Marketing*, 32, 1 (2015), 15-27.
4. Baek C. and Elbeck M. Bitcoins as an investment or speculative vehicle? A first look. *Applied Economics Letters*, 22, 1 (2015), 30-34.
5. Bhattacharyya, S., Banerjee, S., Bose, I., and Kankanhalli, A. Temporal Effects of Repeated Recognition and Lack of Recognition on Online Community Contributions. *Journal of Management Information Systems*, 37, 2 (2020), 536-562.
6. Blei D.M., Ng A.Y., and Jordan M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, (2003), 993-1022.
7. Bollen J., Mao H., and Zeng X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2, 1 (2011), 1-8.
8. Borgatti S.P., Everett M.G., and Johnson J.C. *Analyzing Social Networks*. London: Sage, 2018.
9. Brown J. S., and Duguid P. Organizational learning and communities-of-practice**:** toward a unified view of working, learning, and innovation. *Organization Science*, 2, 1 (1991), 40-57.
10. Burt R.S. *Structural Holes: the Social Structure of Competition*. Cambridge: Harvard University Press, 2009.
11. Cavalcanti T.V., Giannitsarou C., and Johnson C.R. Network cohesion. *Economic Theory*, 64, 1 (2017), 1-21.
12. Chen H., De P., Hu Y.J., and Hwang B.H. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27, 5 (2014), 1367-1403.
13. Chen H., Hu Y.J., and Huang, S. Monetary incentive and stock opinions on social media. *Journal of Management Information Systems*, 36, 2 (2019), 391-417.
14. CNBC. How investors are using social media to make money. 2016. https://www.cnbc.com/2016/06/09/how-investors-are-using-social-media-to-make-money.html.

15. Coleman J.S. Social capital in the creation of human capital. *American Journal of Sociology*, 94 (1988), S95-S120.
16. Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. Political polarization on twitter. *5th International AAAI Conference on Weblogs and Social Media* (*ICWSM*), 2011, pp. 89-96.
17. Corfman, K. P., and Kahn, B. E. The influence of member heterogeneity on dyad judgment: Are two heads better than one? *Marketing Letters*, 6, 1 1995, 23-32.
18. Da, Z., and Huang, X. Harnessing the wisdom of crowds. *Management Science*, 66, 5 (2020), 1847-1867.
19. Das S.R. and Chen M.Y. Yahoo! for Amazon: sentiment extraction from small talk on the web. *Management Science*, 53, 9 (2007), 1375-1388.
20. Davis A.K., Piger J.M., and Sedor L.M. Beyond the numbers: measuring the information content of earnings press release language. *Contemporary Accounting Research*, 29, 3 (2012), 845-868.
21. Deng S., Huang Z.J., Sinha A.P., and Zhao H. The interaction between microblog sentiment and stock return: An empirical examination. *MIS Quarterly*, 42, 3 (2018), 895-918.
22. Dewally M. Internet investment advice: investing with a rock of salt. *Financial Analysts Journal*, 59, 4 (2003), 65-77.
23. Echols A., and Tsai W. Niche and performance: the moderating role of network embeddedness. *Strategic Management Journal*, 26, 3 (2005), 219-238.
24. Gray W., and Kern A. Talking your book: Social networks and price discovery. 2011. https://ssrn.com/abstract=1767452.
25. Greene W.H. *Econometric Analysis*. New York: Pearson Education, 2003.
26. Gu B., Konana P., Raghunathan R., and Chen H.M. Research note—The allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, 25, 3 (2014), 604-617.
27. Hansen M.T., Podolny J. M., and Pfeffer J. So many ties, so little time: a task contingency perspective on corporate social capital in organizations. In K. S. Cook, and J. Hagan (eds.) *Research in the Sociology of Organizations.* Greenwich: JAI Press, 2001, pp. 21-57.
28. Hargittai, E., Gallo, J., and Kane, M. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*, 134, 1-2 (2008), 67-86.
29. Huston, T. L., and Levinger, G. Interpersonal attraction and relationships. *Annual Review of Psychology*, 29, 1 (1978), 115-156.

30. Horwitz, S. K., and Horwitz, I. B. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, 33, 6 2007, 987-1015.
31. Kastens T.L., and Schroeder T.C. A trading simulation test for weak-form efficiency in live cattle futures. *Journal of Futures Markets*, 15, 6 (1995), 649-676.
32. Koka B.R., and Prescott J.E. Strategic alliances as social capital: A multidimensional view. *Strategic Management Journal*, 23, 9 (2002), 795-816.
33. Lakin, J. L., and Chartrand, T. L. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14, 4 (2003), 334-339.
34. Lazer D., and Friedman A. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52, 4 (2007), 667-694.
35. Lee D., Hosanagar K., and Nair H.S. Advertising content and consumer engagement on social media: Evidence from Facebook. *Management Science*, 64, 11 (2018), 5105-5131.
36. Loughran T., and McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The *Journal of Finance*, 66, 1 (2011), 35-65.
37. Mai F., Shan, Z., Bai Q., Wang X., and Chiang R. H. L. How does social media impact Bitcoin value? A test of the silent majority hypothesis. *Journal of Management Information Systems*, 35, 1 (2018), 19-52.
38. Martí J., Bolíbar M., and Lozares C. Network cohesion and social support. *Social Networks*, 48 (2017), 192-201.
39. McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 1 (2001), 415-444.
40. Milanov H., and Shepherd D.A. The importance of the first relationship: The ongoing influence of initial network on future status. *Strategic Management Journal*, 34, 6 (2013), 727-750.
41. Mizruchi M. S., and Stearns L. B. Getting deals done: the use of social networks in bank decision-making. *American Sociological Review*, 66, 5 (2001), 647-671.
42. Motiwalla L., and Wahab M. Predictable variation and profitable trading of US equities: a trading simulation using neural networks. *Computers & Operations Research*, 27, 11-12 (2000), 1111-1129.
43. Quartz. How to short bitcoins (if you really must). 2013. https://qz.com/69630/how-to-short-bitcoins-if-you-really-must/.
44. Reagans R., and McEvily, B. Network structure and knowledge transfer: the effects of cohesion and range. *Administrative Science Quarterly*, 48, 2 (2003), 240-267.

45. Rogers, E. M., and Bhowmik, D. K Homophily-heterophily: Relational concepts for communication research. *Public Opinion Quarterly*, 34, 4 (1970), 523-538.
46. Rowley T., Behrens D., and Krackhardt D. Redundant governance structures: An analysis of structural and relational embeddedness in the steel and semiconductor industries. *Strategic Management Journal*, 21, 3 (2000), 369-386.
47. Singh P.V., Tan Y., and Mookerjee V. Network effects: The influence of structural capital on open source project success. *MIS Quarterly*, 35, 4 (2011), 813-829.
48. Solomon D.H. Selective publicity and stock prices. *The Journal of Finance*, 67, 2 (2012), 599-638.
49. Sun Yin, H. H., Langenheldt, K., Harlev, M., Mukkamala, R. R., and Vatrapu, R. Regulating cryptocurrencies: a supervised machine learning approach to de-anonymizing the bitcoin blockchain. *Journal of Management Information Systems*, 36, 1 (2019), 37-73.
50. Sunstein, C. R. *Republic.com 2.0.* Princeton: Princeton University Press, 2007.
51. Tetlock P.C. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62, 3 (2007), 1139-1168.
52. Tetlock P.C., Saar-Tsechansky M., and Macskassy S. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63, 3 (2008), 1437-1467.
53. Tortoriello M., Reagans R., and McEvily B. Bridging the knowledge gap: The influence of strong ties, network cohesion, and network range on the transfer of knowledge between organizational units. *Organization Science*, 23, 4 (2012), 1024-1039.
54. Tumarkin R., and Whitelaw R.F. News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57, 3 (2001), 41-51.
55. Wasko M.M., and Faraj S. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly*, 29, 1 (2005), 35-57.
56. WSJ. Everyone’s a Day Trader Now. 2020. https://www.wsj.com/articles/everyones-a-day-trader-now-11595649609.
57. Yoo, E., Gu, B., and Rabinovich, E. Diffusion on Social Media Platforms: A Point Process Model for Interaction among Similar Content. *Journal of Management Information Systems*, 36, 4 (2019), 1105-1141.
58. Yu J., Gilbert B.A., and Oviatt B. M. Effects of alliances, time, and network cohesion on the initiation of foreign sales by new ventures. *Strategic Management Journal*, 32, 4 (2011), 424-446.

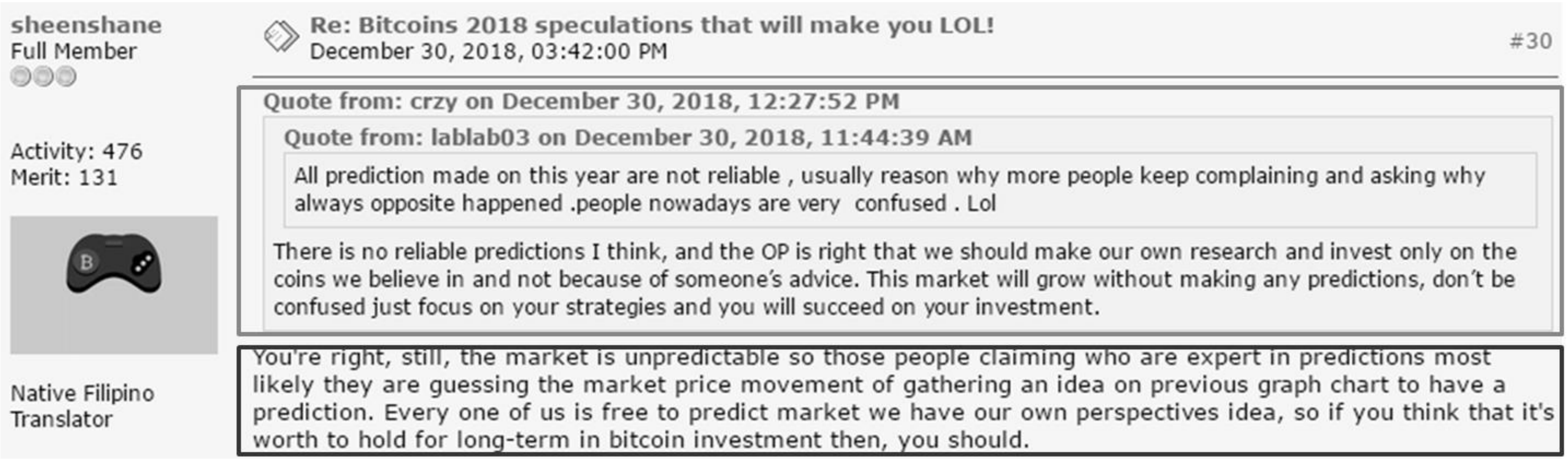

Figure 1 **Illustration of Quoted Message and Quoting Message on Bitcointalk.org**

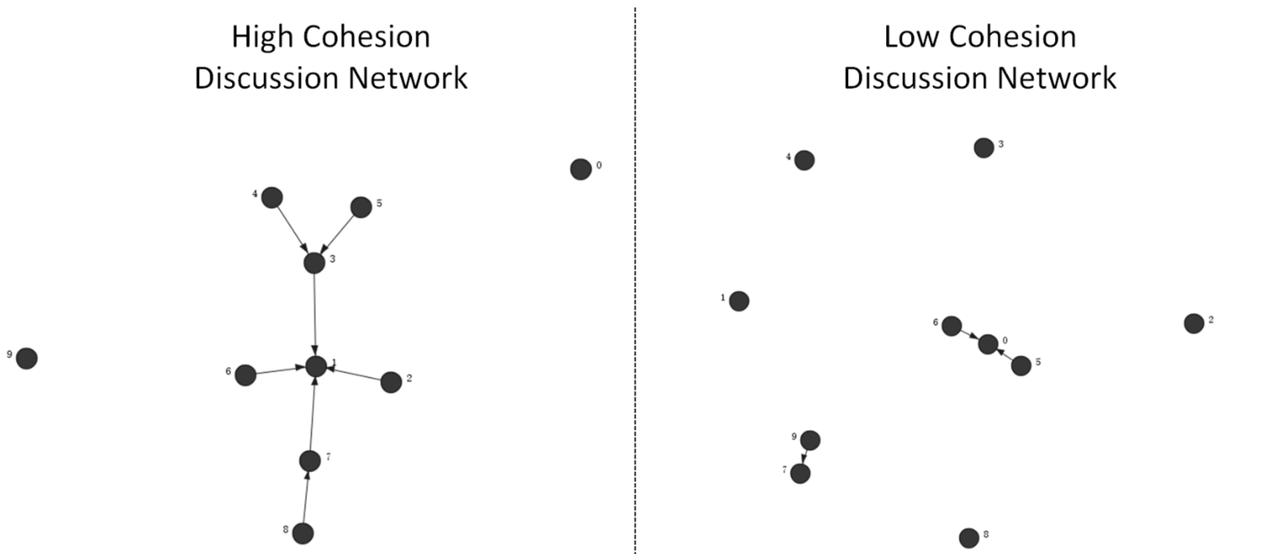

Figure 2 **Illustration of High vs. Low Network Cohesion for a 10-node Discussion Network**

Table 1 Descriptive Statistics – Bitcoin Market and Control Variables

	Mean	Min	Median	Max	Std. Dev	Avg. Daily Increment	Obs.
Panel A: Currency							
Units in Circulation		10,511,875		16,419,900		3,531.40	1,667
Market Capitalization		$133,372,357		$47,218,046,132		$28,143,858	1,667
Blockchain Wallet Users		46,429		14,968,009		8,919	1,667
Panel B: Bitcoin Price Index and Return							
Price Index	500.030	12.500	412.980	3,018.540	455.811		1,667
Return	0.42%	-28.26%	0.23%	41.59%	4.58%		1,667
Panel C: Trading Volume							
Trading Volume in Dollar Terms	$119,610,730	$1,371,188	$66,491,033	$1,209,302,552	$146,816,373	$448,104	1,667
Daily # Transactions	125,690	10,120	89,058	354,151	91,654	148.61	1,667
Panel D: Dow Jones Newswires (DJNS) on Bitcoin							
DJNS Sentiment	0.012	0	0.009	0.100	0.013		1,667
DJNS Coverage	0.594	0	1	1	0.491		1,667
Panel E: Other Control Variables							
Return on S&P 500 Index	0.0005	-0.039	0.0005	0.039	0.008		1,667
Return on COMEX Gold Price	0.0001	-0.035	0	0.054	0.007		1,667
Alexa Web Traffic Rank	6,331.551	1,736	5,679	18,990	3,714.808		1,667
Google Trends Index	3.893	1	3	15	2.696		1,667
AAII Investor Sentiment	0.063	-0.352	0.063	0.386	0.129		1,667
VIX Index	14.840	9.750	13.890	40.740	3.543		1,667

Table 2 Descriptive Statistics – Thread-day Networks

	Mean	Min	25th Percentile	Median	75th Percentile	Max	Std. Dev	Obs.
Panel A: Thread								
Thread Duration (Days)	4.839	1	1	2	4	739	17.741	12,441
% 1st Day Posts	66.68%	0.01%	37.5%	75%	100%	100%	33.35%	12,441
# Views	3,460.42	129	833	1,557	2,860	2,030,570	29,709.57	12,441
Panel B: Thread-day Networks (PostCount>1)								
Sentiment (% Negative Words)	1.40%	0	0.65%	1.34%	2.08%	50%	3.88%	48,201
Post Count	10.120	2	3	6	12	1,073	15.514	48,201
Total # Authors	7.463	1	3	5	9	200	7.136	48,201
Average Degree	0.487	0	0.333	0.500	0.667	1.182	0.266	48,201
Density	0.132	0	0.025	0.069	0.167	0.5	0.154	48,201

Table 3 Association between Sentiment and Next Day Bitcoin Return

	R_{t+1}	R_{t+1}	R_{t+1}
	(1)	(2)	(3)
PostCount Threshold	> 2 (10% percentile)	> 3 (25% percentile)	>6 (50% percentile)
$Sentiment_{it}$	0.022 (0.73)	-0.050* (-1.67)	-0.091** (-2.14)
R_t	-0.320*** (-25.91)	-0.323*** (-24.53)	-0.339*** (-20.08)
R_{t-1}	-0.352*** (-39.71)	-0.354*** (-36.92)	-0.355*** (-31.79)
R_{t-2}	-0.350*** (-33.36)	-0.348*** (-31.17)	-0.368*** (-26.34)
$R_{t-30,t-3}$	-0.144*** (-27.42)	-0.145*** (-26.01)	-0.155*** (-22.71)
$Log(TradingVolume_t)$	-0.005*** (-9.12)	-0.005*** (-8.72)	-0.005*** (-8.27)
$Log(PostCount_{it})$	0.0003 (0.77)	0.0004 (1.00)	0.0003 (0.35)
$Log(PageView_i)$	0.0002 (0.51)	0.001 (1.32)	0.001 (0.99)
$DJNSSentiment_t$	-0.158*** (-5.65)	-0.148*** (-4.95)	-0.158*** (-4.14)
$DJNSCoverage_t$	-0.002** (-2.46)	-0.002*** (-2.84)	-0.003*** (-2.92)
$S\&P500_t$	-0.214*** (-5.40)	-0.218*** (-5.07)	-0.244*** (-4.34)
$COMEXGold_t$	-0.302*** (-5.85)	-0.287*** (-5.02)	-0.223*** (-2.88)
$Log(AlexaRank_t)$	0.013* (1.85)	0.012 (1.53)	0.008 (0.86)
$GoogleTrends_t$	0.00001 (0.02)	-0.0002 (-0.39)	-0.0003 (-0.45)
$AAIISentiment_t$	-0.230* (-1.75)	-0.252* (-1.75)	-0.240 (-1.30)
VIX_t	-0.0004** (-2.11)	-0.0004* (-1.73)	-0.0004 (-1.33)
Week Dummies	√	√	√
# Obs.	40,621	34,894	22,878
R^2	0.304	0.311	0.314

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 4 Moderating Effect of Network Cohesion

	Cohesion Measured by Average Degree			Cohesion Measured by Density		
	R_{t+1}	R_{t+1}	R_{t+1}	R_{t+1}	R_{t+1}	R_{t+1}
	(1)	(2)	(3)	(4)	(5)	(6)
PostCount Threshold	> 2	>3	>6	> 2	>3	>6
$Sentiment_{it}$	-0.053	-0.156***	-0.297***	-0.060**	-0.116***	-0.245***
	(-1.44)	(-3.04)	(-3.59)	(-1.96)	(-2.75)	(-3.97)
$Cohesion_{it}$	-0.002*	-0.002	-0.005	-0.010***	-0.005	-0.020
	(-1.68)	(-1.37)	(-1.56)	(-2.83)	(-0.82)	(-1.09)
$Sentiment_{it} \times Cohesion_{it}$	0.175***	0.229***	0.445***	0.632***	0.683**	3.109***
	(2.62)	(2.67)	(2.96)	(3.55)	(2.44)	(3.35)
R_t	-0.320***	-0.323***	-0.339***	-0.320***	-0.323***	-0.339***
	(-25.90)	(-24.53)	(-20.07)	(-25.93)	(-24.54)	(-20.09)
R_{t-1}	-0.352***	-0.354***	-0.355***	-0.352***	-0.354***	-0.355***
	(-39.70)	(-36.90)	(-31.78)	(-39.74)	(-36.94)	(-31.86)
R_{t-2}	-0.350***	-0.348***	-0.368***	-0.351***	-0.348***	-0.368***
	(-33.36)	(-31.18)	(-26.38)	(-33.36)	(-31.17)	(-26.40)
$R_{t-30,t-3}$	-0.144***	-0.145***	-0.155***	-0.144***	-0.145***	-0.155***
	(-27.43)	(-26.04)	(-22.75)	(-27.44)	(-26.04)	(-22.73)
$Log(TradingVolume_t)$	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***
	(-9.09)	(-8.75)	(-8.31)	(-9.05)	(-8.74)	(-8.28)
$Log(PostCount_{it})$	0.0003	0.0004	0.0002	0.0002	0.001	0.001
	(0.74)	(0.95)	(0.29)	(0.48)	(1.34)	(1.50)
$Log(PageView_i)$	0.0002	0.0005	0.0004	0.0003	0.001	0.0003
	(0.42)	(1.11)	(0.74)	(0.54)	(1.10)	(0.64)
$DJNSSentiment_t$	-0.158***	-0.148***	-0.158***	-0.158***	-0.147***	-0.158***
	(-5.64)	(-4.94)	(-4.13)	(-5.64)	(-4.94)	(-4.12)
$DJNSCoverage_t$	-0.002**	-0.002***	-0.003***	-0.002**	-0.002***	-0.003***
	(-2.48)	(-2.86)	(-2.91)	(-248)	(-2.85)	(-2.91)
$S\&P500_t$	-0.214***	-0.217***	-0.244***	-0.214***	-0.217***	-0.244***
	(-5.41)	(-5.07)	(-4.33)	(-5.39)	(-5.06)	(-4.33)
$COMEXGold_t$	-0.301***	-0.287***	-0.223***	-0.301***	-0.287***	-0.221***
	(-5.85)	(-5.02)	(-2.88)	(-5.84)	(-5.01)	(-2.86)
$Log(AlexaRank_t)$	0.013*	0.012	0.008	0.013*	0.012	0.008
	(1.83)	(1.51)	(0.83)	(1.83)	(1.52)	(0.84)
$GoogleTrends_t$	0.000005	-0.0002	-0.0003	0.00002	-0.0002	-0.0003
	(0.01)	(-0.41)	(-0.46)	(0.04)	(-0.39)	(-0.48)
$AAIISentiment_t$	-0.228*	-0.250*	-0.239	-0.230*	-0.249*	-0.237
	(-1.74)	(-1.74)	(-1.29)	(-1.75)	(-1.72)	(-1.27)
VIX_t	-0.0005**	-0.0004*	-0.0004	-0.0005**	-0.0004*	-0.0004
	(-2.13)	(-1.73)	(-1.32)	(-2.16)	(-1.73)	(-1.31)
Week Dummies	√	√	√	√	√	√
# Obs.	40,621	34,894	22,878	40,621	34,894	22,878
R^2	0.304	0.311	0.314	0.305	0.311	0.314

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 5. Trading Simulation (Starting with $1; 10,000 Repetitions)

	Mean	Median	Min	Max	Std. Dev.	Percentage Gains Against "Equally Weighted Sentiment"
Panel A: Short selling is not allowed						
Random	20.423	13.308	0.394	289.631	23.280	-44.1%
Past returns	32.641	21.673	0.835	885.429	36.510	-10.6%
Equally weighted sentiment	36.528	26.546	1.210	572.849	34.706	N/A
Average degree weighted sentiment	49.526	35.720	1.330	656.593	47.776	35.5%
Density weighted sentiment	53.388	37.949	1.548	1,066.641	51.206	46.2%
Panel B: Short selling is allowed						
Random	0.994	0.163	0.0002	138.464	4.211	-62.4%
Past returns	3.062	0.568	0.0003	402.672	12.165	15.7%
Equally weighted sentiment	2.647	0.631	0.002	754.723	10.947	N/A
Average degree weighted sentiment	4.456	1.217	0.003	320.038	12.167	68.3%
Density weighted sentiment	5.370	1.393	0.003	829.764	17.453	102.9%

Table 6 Information Correlation and Network Formation

	$Quotation_{i,j}$	$Quotation_{i,j}$	$Quotation_{i,j}$
	(1)	(2)	(3)
$MessagesInBetween_{i,j}$ Threshold	*≤20*	*≤30*	*≤100*
$OpinionDistance_{i,j}$	-2.920***	-2.948***	-3.025***
	(-26.18)	(-26.90)	-28.23)
$TopicSimilarity_{i,j}$	0.863***	0.866***	0.866***
	(120.98)	(123.75)	(126.89)
$Log(TimeLag_{i,j}+1)$	-0.092***	-0.098***	-0.104***
	(-75.14)	(-80.92)	-86.84)
$Log(MessagesInBetween_{i,j})$	-1.358***	-1.362***	-1.394***
	(-392.70)	(-424.15)	-487.75)
$Sentiment_i$	2.381***	2.360***	2.357***
	(23.69)	(23.93)	(24.45)
$Log(Length_i)$	0.122***	0.125***	0.126***
	(51.97)	(54.21)	(56.04)
$TopicConcentrationHHI_i$	-0.164***	-0.181***	-0.211***
	(-16.46)	(-18.51)	(-22.09)
$Log(AuthorActivity_i+1)$	-0.036***	-0.041***	-0.054***
	(-16.37)	(-19.36)	(-26.18)
$CitingOther_i$	0.138***	0.106***	0.050***
	(29.16)	(22.79)	(11.16)
$Log(CitedCount_i+1)$	0.657***	0.767***	1.028***
	(85.42)	(105.58)	(154.89)
$Sentiment_j$	1.940***	1.957***	1.971***
	(19.40)	(19.93)	(20.55)
$Log(Length_j)$	0.121***	0.126***	0.131***
	(51.86)	(54.77)	(58.83)
$TopicConcentrationHHI_j$	-0.371***	-0.374***	-0.375***
	(-37.24)	(-38.19-16.67)	(-39.30)
$Log(AuthorActivity_j+1)$	0.155***	0.155***	0.151***
	(65.97)	(67.43)	(67.39)
$CitingOther_j$	-3.414***	-3.411***	-3.413***
	(-302.98)	(-310.62)	(-321.51)
Week Dummies	√	√	√
# Obs.	7,216,601	10,089,372	23,702,216

Note: *** p<0.01, ** p<0.05, * p<0.1.