Twitter financial community sentiment and its predictive relationship to stock market movement

Steve Y. Yang, Sheung Yin Kevin Mo, and Anqi Liu

Twitter, one of the several major social media platforms, has been identified as an influential factor to financial markets by multiple academic and professional publications in recent years. The motivation of this study hinges on the growing popularity of the use of Twitter and the increasing prevalence of its influence among the financial investment community. This paper presents an empirical evidence of the existence of a financial community on Twitter in which users' interests align with the financial market related topics. We establish a methodology to identify relevant Twitter users who form the financial community, and we also present the empirical findings of network characteristics of the financial community. We observe that this financial community behaves similarly to a small-world network, and we further identify groups of critical nodes and analyze their influence within the financial community based on several network centrality measures. Using a novel sentiment analysis algorithm, we construct a weighted sentiment measure using tweet messages from these critical nodes, and we discover that it is significantly correlated with the returns of the major financial market indices. By forming a financial community within the Twitter universe, we argue that the influential Twitter users within the financial community provide a better proxy between social sentiment and financial market movement. Hence, we conclude that the weighted sentiment constructed from these critical nodes within the financial community provides a more robust predictor of financial markets than the general social sentiment.

Keywords: Twitter; Sentiment Analysis; Network Analysis; Financial Community; Network Centrality; Volatility; Regression Analysis

1.   INTRODUCTION

Behavioral finance has presented empirical evidence that financial decisions are largely driven by emotions and mood (Hilton 2001, Nofsinger 2005, Cohen and Kudryavtsev 2012, Brown and Cliff 2004). The motivation of this paper rests on the belief that social mood or sentiment in social media predicates upward or downward movement of the financial markets. Twitter, one of the main social media platforms, has been identified as an influential factor to the financial market with multiple sources of empirical evidence among academic and professional publications (Bollen *et al.* 2011, Zhang *et al.* 2011). The emergence of top influencers[1] and the occurrence of the disruptive events such as the 2013 Associated Press Hoax[2] have raised alarm on the potential threat of social media influence to the stability of the financial markets. As a potential solution to understand

---

[1] As the Securities Exchange Commission (SEC) permits tweets as a form of news release, several Twitter accounts have been identified as influential users related to the financial market. For instance, Carl Icahn, a hedge fund manager, has frequently disclosed information on Twitter with nearly 26,000 followers. (Available at http://blogs.wsj.com/moneybeat/2013/08/12/carl-icahn-warns-investors-to-watch-his-twitter-account)

[2] On April 23, 2013, a faked tweet from Associated Press Twitter account (since deleted) claiming there had been an explosion at the White House that injured President Obama caused the market reacted accordingly: From 1:08 p.m. to 1:10 p.m., the Dow Jones Industrial average plunged more than 100 points, from 14697.15 to 14548.58. Just as quickly though, it rebounded. By 1:13 p.m., it was back above 14690.

how Twitter influences financial market sentiment, this study defines a financial community with Twitter users whose interests align with the financial market. This subset of the Twitter users, an active investment community, contains more refined and precise information related to the financial market, therefore its influence to market sentiment can be pronounced.

Our hypothesis is that Twitter sentiment reflects the market participants' beliefs and behaviors toward future outcomes and the aggregate of the societal mood can present itself as a reliable predictor of financial market movement. However, not all users are equally influential in the social media, and those influential social media users will certainly have higher impact to the societal mood or sentiment (Cha *et al.* 2010). Reported evidence shown that there exist a community on Twitter whose primary concern is about financial investment. Those users who are harvesting information from these influential sources on the social media for their daily trading decisions forms the robust linkage between the social mood and financial market asset price movement. Hence this community would be more representative to market participant's beliefs, and consequently the sentiment extracted from this financial community would serve as a better predictor to the market movement.

In this paper, we seek to identify the corresponding investment community and pinpoint its major influencers in the social networks context. The primary research question is whether the beliefs and behaviors of major key players in such community reveals better signals to financial market movement. From a large-scale data crawling effort, we define a financial community as a group of relevant Twitter users with interests aligned with the financial market. We first identify 50 well-recognized investment experts' accounts in Twitter and use their common keywords to create the interests of the financial investment community. By constructing the two layers of the experts' followers, we apply a multitude of rigorous filtering criteria to establish a financial community boundary based on their persistent interests in the topic of financial investment.

After settling on a definition, we examine how messages from key influencers in the community interact with social mood or sentiment that tend to signal an impending upward or downward swing in the market price movement. We use key network metrics such as out-degree centrality and betweenness centrality to identify the financial community influencers and we conjecture that these key influencers along with their weight of their influence in the financial community will provide better predictors of financial market movement measures.

2.   BACKGROUND

This section provides a synopsis of the current understanding of the relationship between the investment financial community and social networks. We first describe the social media platforms, with a particular focus on Twitter, and then explore its influential role on the financial market. Given the large scale of the Twitter network, it will be important to further discuss the current techniques that are utilized for community detection and sentiment analysis.

2.1.   *Literature of Twitter and Social Media*

Social media has become a major channel of information worldwide. According to Kaplan and Haenlein (2010), social media is defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content" . These user-generated contents serve valuable source of information to business executives and decision-makers due to their profitable applications (Kaplan and Haenlein 2010). For example, Dell listed its continual sales and marketing effort in engaging customers through "external social media channels" in its 2012 annual report and subsequently reported \$1

million additional revenue due to sales alerts from Twitter[1].

Twitter is a micro-blogging social media platform that allows its users to send and read short messages of up to 140 characters, commonly known as "tweets". It has evolved with 517 million registered users as of 2012, generating over 340 million tweets daily[2]. The statistics is astonishing in recent web development, and Twitter has been consistently ranked as one of the top 15 websites with the highest volume. Twitter also demonstrates strong growing presence in the social media platforms among the top 15 websites[3].

2.2.   *Literature of Twitter Influence to Financial Market*

With Twitter becoming one of the largest social media platforms in the world, its influence to the financial market has been documented. A survey conducted by Thomson Reuters indicates that majority of financial market professionals and customers use social media for professional reasons[4]. This is a revealing trend that social media carries a substantial weight in influencing markets. Empirical evidences also demonstrate strong correlation between the two domains. Chen *et al.* (2013) finds that the articles and commentaries available on social media platforms contributes to the price discovery process and that investor opinions revealed on these platforms "strongly predict future stock returns and earnings surprises". In a similar study, Ruiz *et al.* (2012) showed the correlation between micro-blogging activity and stock-market events, and how it could be applied for developing profitable strategies. Bollen *et al.* (2011) further suggested that including public mood dimensions enhance the accuracy of market prediction. A similar study conducted by Zhang *et al.* (2011) investigated a set of emotional text Twitter data of six public opinion time series (dollar, $, gold, oil, job, and economy) and found their correlated effects with the financial market movement.

2.3.   *Literature of Social Network and Community Detection*

A social network is "a set of nodes representing people that are connected by links showing relations or flows between them" (González-Arangüena *et al.* 2010). A social network can be categorized into different network types, and the identification process depends on network metrics such as average path length and average clustering coefficient (Zhi-Yun *et al.* 2009). Large-scale networks across different domains were observed to follow similar patterns, such as scale-free distributions, the small-world effect and strong community structures (Tang and Liu 2010). The small-world phenomenon was first proposed by Travers and Milgram (1969) that "social networks are in some sense tightly woven, full of unexpected strands linking individuals seemingly far removed from one another in physical or social space". The small-world networks are often composed of "short chains of acquaintances" among social participants (Kleinberg 2000), highlighting key characteristics of small diameters and small average path lengths (Watts and Strogatz 1998). High clustering coefficient is also noted relative to "a purely random graph with the same number of links" (Jackson 2010). In summary, small-world networks are graphs with clustering coefficients much larger than random networks and with diameters that increase logarithmically with the number of nodes (Watts 1999). Small-world networks are "globally and locally efficient" in exchanging information among the network and examples were found in the neural networks, communication, and transport networks (Latora and Marchiori 2001).

---

[1]Internet-News article: What Keeps Twitter Chirping Along. This article describes Dell's attempt to use social media for expanding its market effort starting in 2007. Available at: http://www.internetnews.com/webcontent/article.php/3790161/What+Keeps+Twitter+Chirping+Along.htm

[2]Tech Crunch article: Twitter Passed 500M Users in June 2012, 140M of Them in US. Available at http://blog.twitter.com/2012/twitter-turns-six

[3]Statistics-Brain article: Top Websites By Traffic. Twitter is ranked #4 with more than 88 million visitors per month in 2012. Available at: http://www.statisticbrain.com/top-us-websites-by-traffic/

[4]Thomson Reuters Presentation: Social Networking for Market Professionals: How is it Changing the Face of Financial Markets.

Community is defined as "the set of nodes such that they interact with each other more frequently than with those nodes outside the group" (Tang and Liu 2010). Once a community is formed, its aggregate behavior can be insightful in representing specific interest groups. There are four common approaches to form a community: node-centric, group-centric, network-centric and hierarchy-centric (Tang and Liu 2010). Among the most common, the node-centric method requires each node in a group to satisfy a set of specified properties (Tang and Liu 2010). The selection of the optimal approach rests on the trade-off between stability and utility in how well the community reflects the common interest. Moreover, it facilitates the discovery of clustered group that achieves similarity among members and dissimilarity among different communities.

2.4. *Literature of Sentiment Analysis Techniques*

Sentiment analysis on text has been conducted by both academia and industry. By utilizing text mining techniques, sentiment analysis models have been applied to analyze mood in blogs. For instance, Dodds and Danforth (2010) measured the sentiment level of song lyrics and the content of blogs. They used the ANEW dictionary (Bradley and Lang 1999) to quantitatively assess words and evaluate the sentiment associated. This methodology can be applied large textual information. The recent growth of social media has increased interest in sentiment analysis of blogs and tweet messages. However, performing sentiment analysis on tweet messages is difficult due to the use of informal language expressions, misspellings and slangs (Nakov *et al.* 2013). A large number of Twitter sentiment analysis involves intensive data cleaning and filtering. Khan *et al.* (2014) proposed a three-step pre-processing procedure. First, the algorithm checks for spelling and grammatical mistakes. The following step involves word stemming, and removes stop words. Twitter-specific terms, such as URLs, hashtags, re-tweets (RTs) and usernames can be removed (Montejo-Ráez *et al.* 2012) or replaced by special tags (Agarwal *et al.* 2011). Unique attributes in tweets such as emoticons, repeating letters and punctuations, help to express the sentiment of the message. However, using such features in sentiment models is uncommon because they negatively impact the accuracy (Go *et al.* 2009, Agarwal *et al.* 2011). Mohammad *et al.* (2013) remove these features from their model and discovered that it improved the model performance.

The sentiment analysis techniques can be categorized into two main groups. The first group involves the use of a dictionary of positive and negative keywords. The sentiment is quantified by the number of word matches (Ortega *et al.* 2013, Dodds and Danforth 2010). Khan *et al.* (2014) applied this algorithm for opinion mining based on hybrid classifier. The second group adopts machine learning classifiers such as Nave Bayes, Support Vector Machine (SVM), and linear regression to identify the polarity of the text contents. This method requires extracting features from texts and training by existing polarity labels (Agarwal *et al.* 2011, Mohammad *et al.* 2013).

3.   DATA COLLECTION

Collecting Twitter data is an important milestone to gather empirical supporting evidence to construct the financial community and to identify behavioral patterns among its participants. Using Twitter API, we implemented a large-scale data collection system to collect three major types of data. First, the system gathers the friend-following relationship information between two Twitter users. Given the unique identification number of a specific user, the system can generate the identification number for all followers. This information can be useful in forming the structure of the social network. Second, the system collects user profile information including its location, language preference, account creation date, time zone. It also specifies the number of followers, the number of friends, and the number of total messages sent. The profile information has allowed our study understand the demographics of the financial community and their message propagation pattern. Lastly, the system tracks tweet and retweet messages by a majority of the members in the financial community. These messages can be translated into sentiment and therefore be used

for in-depth study related to financial phenomenon. By tracking the structural relation among participants of the community, their profile information and messages, the dataset has led to this empirical study related to the proposed financial community examining its network characteristics and significance in correlating the messages' sentiment with market movements.

4. METHODOLOGY

We use a degree-centric community detection method to form the financial community, which requires each community member to satisfy a set of specific requirements. We seek to understand the structure and network characteristics of the financial community in the Twitter universe through empirical evidence. The objective is to understand the static and dynamic properties of the network and how their message sentiment can be translated into measurable financial phenomenon.

4.1. *Definition of a Financial Community*

The financial community is defined as a subset of the Twitter users with similar interests in the financial market. The objective of establishing such community allows us to extract the most relevant users from the Twitter universe related to the financial market.

4.2. *Construction of the Financial Community*

- Identification of 50 well-recognized investment experts
  The proposed financial community begins with a collection of 50 user accounts which represents well-recognized investment experts and financial news providers (see Appendix A for the detailed account lists). The first list includes the top 25 traders as the most influential users on Twitter related to the financial market[1]. The selection of these traders was based on the frequency of posting, the number of followers and the usefulness of their tweets. The second list derives from the top 7 financial news providers who feature financial news in a timely and objective manner: Bloomberg, Forbes, Reuters, BusinessWeek, Financial Times, CNNMoney and CNBC[2]. Their associated Twitter accounts form the remaining 25 seeds. The assumption is that the 50 seed accounts specialize in delivering financial-related messages and their followers share similar interest, reflecting a strong likelihood that their followers belong to the active investment community.
- Forming the Financial Community Network
  It is the hypothesis of this study that the community can be built based on a set of influential seed accounts with their associated followers. The community network is formed by unique nodes and linkages. We obtain the pairwise data of the friend-follower relationships through Twitter API by extracting the followers of the 50 seed accounts. Once these linkages are identified, an algorithm sorts through the linkage table by extracting the unique nodes and their secondary followers are captured. The two-layer of the followers of the 50 seed accounts forms the basis of the financial community (see Figure 1).

---

[1]Option Trading IQ article: Top 25 Traders on Twitter. Available at http://www.optionstradingiq.com/top-25-traders-on-twitter/

[2]Top FOREX News article: Top 7 News Sources for Financial Trader, Available at http://www.topforexnews.com/top-7-news-sources-for-financial-trader/
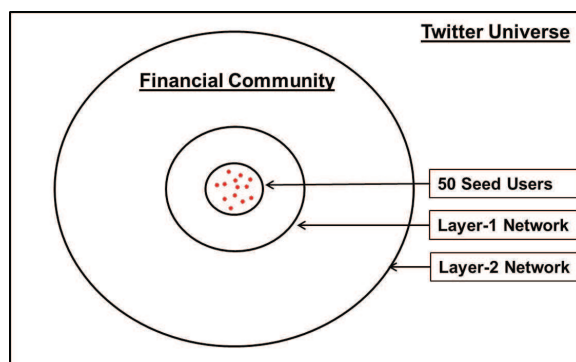
Figure 1. Financial Community from Twitter Universe

- Filtering Mechanism
    1. By Language Preference, Location and Messages
        With the goal of extracting the most relevant nodes in the community, we conduct a filtering procedure to eliminate nodes that might have relations with existing actors but have limited significance to the network. First, we restrict the domain to English language although we acknowledge that this methodology has the capability to apply across different languages. Second, we restrict the location accounts to all U.S. and U.K. domain. Any missing values from the profiles are disregarded. The rationale behind these two criteria is to ensure the quality of each node's contribution to the network. There exists a trade-off in encompassing a large network with as many related nodes as possible and in retaining the core quality of nodes with the best resemblance of a financial community. We aim to find an optimal balance between yielding a sufficiently large sample size and revealing significant information about the financial market.
    2. By Language Preference, Location and Messages
        One of the key objectives in filtering a relevant financial community is to ensure the messages posted by its members contain relevant financial-related keywords. Our approach includes extracting the top common keywords appeared in the tweet messages collected from the top 50 seed influencers from August 3 to September 3, 2013. We group these keywords into financial-related topics and apply them to screen out users whose most recent 20 messages do not belong to the relevant financial topics.

5. EMPIRICAL FINDINGS

5.1. *Financial Community Summary*

We identified a total number of 154,203 Twitter users in the financial community. We believe that our financial community consists of a robust set of Twitter users whose interests are related to the financial market. While incorporating additional layers of the relationships, we did not find significant expansion of the financial community.

5.2. *Financial Community User Classification*

It is essential to decompose the community into distinct user groups by their friend-following relationship. We adopt the Twitter user classification defined by Krishnamurthy *et al.* (2008) with the three distinctive user types: Broadcaster, Acquaintance and Odd users.

1. *Broadcaster:* The Twitter users who have a higher proportion of followers than people they follow.

2. *Acquaintances:* The Twitter users who demonstrate reciprocity in their relationships.
3. *Odd User/Small Cluster:* The Twitter users who have a higher proportion of people they follow than their followers.

The key metrics for classification is the ratio of the number of followers over the number of users who they follow (i.e. friend-following ratio). Broadcasters are defined as the users who have a friend-following ratio of over 2.00 and a minimum of 300 followers. Users with friend-following ratio (ff_ratio) between 0.50 and 2.00 are labeled as acquaintances and the rest are labeled as odd users. Our results indicate that there are 10,760 broadcasters, 131,105 acquaintances and 12,338 odd users (see Table 1). Majority of the community consists of acquaintances and the remaining splits about evenly between broadcasters and odd users.

Table 1.   Community Composition

Community Composition	Empirical Findings
Broadcasters	10,760 (7%)
Acquaintances	131,105 (85%)
Odd Users	12,338 (8%)

5.3.   *Financial Community Demographics*

The empirical scatter plot of the friend-following relationship is among the three distinctive user types in the financial community. It displays the number of followers and friends of each member in the community (see Figure 2). First, we construct two lines that represent the thresholds of each user type definition, i.e. slope of the follower-friend ratio. The first line defines the boundary between broadcasters and acquaintances (ff_ratio = 2). The nodes above this line are broadcasters. On the other hand, the second line defines boundary between acquaintances and odd users that all nodes below that line are odds (ff_ratio = 0.5). Thus, the majority nodes which stay between these two lines are acquaintances.
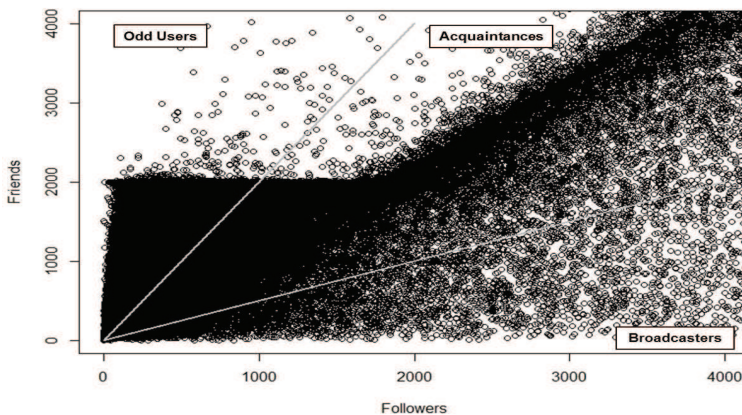

Figure 2. Scatter plot of Community by friend-follower relationship

The second observation is that most acquaintances are located at the diagonal line of the scatter plot (ff_ratio = 1). This demonstrates that the reciprocal relationship is most common among the community population, which is consistent with empirical finding in the Twitter universe (Krishnamurthy *et al.* 2008). The linear reciprocal relationship can be rationalized in the Twitter universe that a user tends to have the same number of followers and friends. In addition, it is noted that in the scatter plot there is a cut-off line at 2,000 friends. It can be explained by a restriction

imposed by Twitter which states that a user cannot follow more than 2,000 accounts if he does not have a sufficiently larger number of followers.

5.4. *Financial Community Population Growth*

We also examined the evolution of the financial community over time. We were able to extract the creation date of all the community members' Twitter accounts and consequently describe the community population growth curve. Since the initiation of Twitter in June 2006, the network has grown stagnantly for the first 2.5 years and rapidly after September 2008 reaching from 20,000 users to the current 154,327 users (see Figure 3).
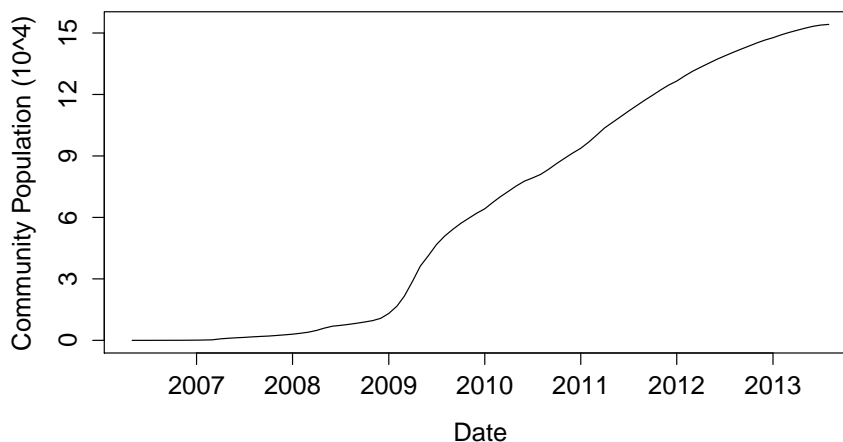

Figure 3. Financial Community Growth

5.5. *Network Characteristics of the Financial Community*

There are two elements in forming a network: Nodes and Edges. In this study, the 154,327 user accounts constitute the nodes for the financial community network and their pairwise friend-following relationships are used as forming the edges among the nodes. Due to the friend-follower relationship of the community structure, all nodes are connected with at least one edge and therefore the financial community is a large connected network without any isolated nodes.

Table 2.   Financial Community Network Summary Statistics

Number of Nodes	154,327
Number of Links	4,846,805
Average Out-Degree Centrality	35.71
Average Betweenness Centrality	420.2
Clustering Coefficient	0.15
Network Diameter	6
Average Path Length	2.72
Connected Component	1

We compare different types of network structure, and we find that the financial community best

8

resembles a small-world network. Our community exhibits some of the key characteristics of a small-world network such as its small diameter, and higher clustering coefficient relative to a random network. The descriptive characteristics are consistent with our empirical summary statistics of the financial community network (see Table 2). The low diameter at 6.00 illustrates that the longest shortest path is only 6 layers from a node to any other nodes in the network, which is a defining small-world network property. The diameter value is consistent with the result of comparable small-world network study at 4.67 and 3.43 by Cheng [1] and Bakhshandeh *et al.* (2011) respectively. In addition, the network displays a strong decaying power law of its degree distribution. This result is consistent with the findings of literature characterizing social media network as a small-world network.

5.6. *Critical Node Analysis in the Financial Community*

In the financial community, there exist users who play a central role in the connectedness of the network. These users, known as critical nodes, situate at the most critical locations of the community network and therefore bear a large weight in the network dynamic properties such as connectedness and message propagation pattern. Analyzing these nodes is essential for understanding the financial community because they represent the most influential users in the community in terms of facilitating the message propagation process and stabilizing the network structure.

Through social network analysis, we identify these critical nodes by applying centrality measures: out-degree centrality, betweenness centrality and closeness centrality. Our data captures the direction of the friend-follower relationship which contributes to the formation of a directed network. The three centrality measures incorporate direction as information propagates from the sender to their followers. This study tracks the profile information and tweet messages of the top 2,500 users ranked by each of the three centrality measures. The description of each centrality measure is defined as:

1. **Out-degree centrality** measures the number of followers a node have in the network.

$$C_D(v_i) = \sum_{j=1}^{n} a_{ij} \tag{1}$$

   where $\boldsymbol{A}$ denotes the adjacency matrix, a_{ij} is a binary term that values 1 if node i out-ties with j and values 0 otherwise, and n is the number of nodes in network (Borgatti and Everett 2006).

2. **Betweenness centrality** captures the number of shortest paths from all vertices to other nodes in the network that passes through the specific node.

$$C_B(v_i) = \sum_{j \neq i} \sum_{k \neq i} \frac{g_{jik}}{g_{jk}} \tag{2}$$

   where g_{jk} denotes the number of geodesic paths from node j to node k and g_{jik} denotes the number of geodesic paths from node j to node k that pass through node i (Borgatti and Everett 2006).

3. **Closeness centrality** captures the number of shortest paths from all vertices to other nodes in the network that passes through the specific node.

$$C_C(v_i) = \sum_{j=1}^{n} d_{ij} \tag{3}$$

---

[1] Sysomos article: Six Degree of Separation, Twitter Style. Available at: http://www.sysomos.com/insidetwitter/sixdegrees/

where $\boldsymbol{D}$ is the geodesic distance matrix, d_{ij} is the geodesic distance between node i and node j, and n is the number of nodes in network (Borgatti and Everett 2006).

Each group of these critical nodes shares certain common attributes. The profile data consists of the nodes' location, username, description, the number of messages they have posted, and the number of followers and friends. In studying of these critical nodes, we first examine the three important attributes: the number of tweeted messages, the number of followers and the number of friends. Samples of these critical nodes are provided in Table 3.

Table 3.   Critical Node Sample Users

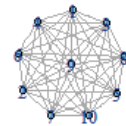| Critical Nodes	Sample Users
Out-Degree Centrality	@TheEconomist, @BreakingNews, @FinancialTimes, @FortuneMagazine, @CMEGroup
Betweenness Centrality	@themotleyfool, @Vanguard_Group, @ReformedBroker, @TheStreet, @NYSEEuronext
Closeness Centrality	@YESBANK, @currency4trades, @QNBGroup, @Bizzun, @FFinancialGroup, @sobertrader

a). avg_dc = 0.36 avg_bc = 0.20 avg_cc = 0.70

b). avg_dc = 0.26 avg_bc = 0.15 avg_cc = 0.62

c). avg_dc = 0.40 avg_bc = 0.24 avg_cc = 0.71

d). avg_dc = 1.00 avg_bc = 0.00 avg_cc = 1.00

Figure 4. Financial Community Network Topology

When we analyzed critical nodes with the highest out-degree centrality, we observed that they were followed by a large portion of the users in the community. A tweet message initiated by this group can be spread to a large domain in the network. Furthermore, nodes with the highest out-degree centrality can be much more influential than other nodes as their large number of followers can gain significant interest from the financial community and therefore attract more users who follow them. From the profile dataset, these top nodes with the highest out-degree centrality are broadcasters who have posted more than 10,000 tweets over the lifetime of the account, a much higher tweeting rate than that of normal broadcasters and the community. They also appeared to have a longer account history compared to the other two groups of critical nodes.

First, we look at critical nodes with the highest out-degree centrality, and we find that these critical nodes are followed by a large portion of the users in the community. A tweet message initiated by this group can be spread to a large domain in the network. Furthermore, nodes with the highest out-degree centrality can be much more influential than other nodes as their large

number of followers can gain significant interest from the financial community and therefore attract more users who follow them. From the profile dataset, these top nodes with the highest out-degree centrality are broadcasters who normally have tweeted more than 10,000 tweets over the lifetime of the account, a much higher tweeting rate than normal broadcasters and the community. They also have a longer account history compared to the other two groups of critical nodes since 2007.

The second group of critical nodes, users with the highest betweenness centrality, is equally split between broadcasters and acquaintances with few odd users. This illustrates that this group of critical nodes is not composed of mainly broadcasters, but nodes that are situated between the most followed broadcasters and least followed odd users. Their connectedness is important as they can transmit information most effectively within the network. An interesting observation is that these critical nodes are more specialized in the financial industry with majority being writers or publishing editors who are highly interested in financial market. Most accounts were created from 2007 to 2009.

The last group of critical nodes, users with the lowest closeness centrality measure, consists of community members who are within short distance to the most central location within the network. These nodes have many followers compared to those critical nodes measured in out-degree centrality and betweenness centrality. In the financial community, their messages can take a shorter distance to spread among the network and therefore the underlying message propagation is more direct. Many nodes are observed to have less number of users they follow and less number of tweeted messages, compared to the previous two groups. Most of these accounts were created from 2008 to 2011.

As an illustration of the importance of centrality measures related to the community structure, we are interested in exploring how the current financial community compares against sample community structures. We select four sample communities to showcase extreme scenarios with distinct characteristics. We then calculate the normalized centrality score for every community node and compute the average among all the node scores to gauge the connectedness of the community (see Figure 4). Figure 4(a) shows a symmetric structure with two nodes as the central hub. The peripheral nodes are directly attached to the center of the community and therefore the overall structure has a relatively high average closeness centrality score. Figure 4(b) features a similar structure with one central node with more connections. This biased phenomenon results in a lower average centrality scores across all three measures. Figure 4(c) is another symmetric structure featuring three nodes as the central hub and higher average centrality scores. In contrast, Figure 4(d) is a fully connected community with each node linked to all other nodes. This structure is an ideal framework for message propagation as the distance between any two nodes in the community is 1. Therefore, the betweenness centrality and the closeness centrality for all nodes is 0 and 1, respectively. In summary, these sample communities feature distinct network structures in terms of key centrality measures. For the financial community, the average scores of out-degree centrality, betweenness centrality and closeness centrality are 0.720, 0.005 and 0.360, respectively. These scores are normalized from 0 to 1 to facilitate the comparison with the sample structures. The high average out-degree centrality score illustrates that the financial community contains more direct linkages among nodes, while the low average betweenness centrality score shows that the central hub among the network is widely spread among many nodes. In addition, the average closeness centrality score reflects that a substantial portion of the community users are close to the center of the network. These observations show that the financial community is more densely connected at the local level. In addition, the connectivity is not strong at the global level but there is evidence of clustering around the center of the network.

Based on the location data, we extract the top 2,500 users with the highest betweenness centrality from the financial community in the U.S. and map their population density (see Figure 5). New York and California are identified as the top two states containing the largest number of critical nodes, a fact reflective of their wealth distribution and status as financial and media centers. The next level consists of the states of Massachusetts, Texas, Illinois and Florida. These are among the U.S. states with the largest and wealthiest population. Moreover, some of their cities such as Boston

and Chicago, serve as major financial hubs. Lastly, the final level comprises mostly states on the east coast like Pennsylvania, Virginia, Georgia, New Jersey, Maryland and District of Columbia. They tend to be situated with close proximity to New York City and have significant business ties with regard to the number of financial corporation headquarters. The population map of the critical nodes has two significant interpretations: First, it reflects to a certain degree where the most influential nodes in the financial community are most likely located. Second, their tweeting activities might have direct implications on the location of the events. Knowing the location of the source provides a competitive advantage in tracking the scope of the events.
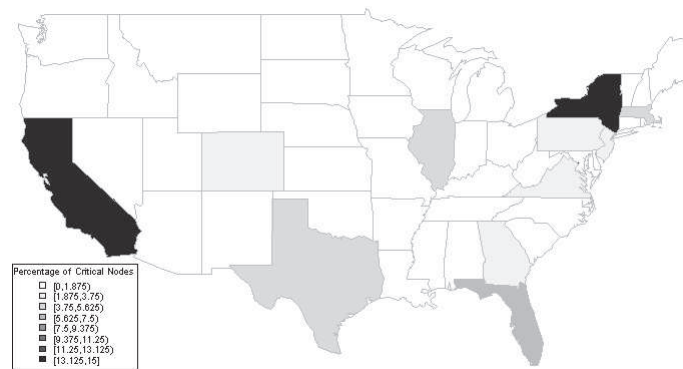

Figure 5. Critical Nodes Location in Financial Community

6.   SENTIMENT ANALYSIS

6.1.   *Motivation*

The tweet message is a unique form of textual information with key distinguishable characteristics. First, it is limited to 140 characters and therefore the message is often conveyed in a condensed and concise form. Second, it contains common jargons that do not conform to traditional English grammar. Third, different users have their unique style of composing tweet messages and therefore results in the heterogeneous form of expression.

Sentiment analysis is applied to determine the sentiment level of tweet messages. A major challenge is to find the optimal algorithm that measures sentiment with highest accuracy. Despite using different types of algorithms for sentiment analysis, the selection process often rests on the unique nature of the textual information and its specific topic. In this study, we propose a new and innovative sentiment analysis algorithm. The strength of our proposed algorithm lies in its ability to analyze tweet messages and extract their sentiments based on the relevance to financial market topics. In addition, credibility of the message source plays an important role in our sentiment analysis algorithm as we believe that not all messages carry equivalent weight of information. Thus, messages from more credible sources should have a much larger influence on the financial community. This is manifested in the credibility measure and here, we propose using different centrality measures to benchmark the associated credibility. For instance, a similar positive message tweeted by a broadcaster will have a larger sentiment weight than an acquaintance. Our algorithm resulted in substantial improvements in accuracy using the dictionary of sentimental keywords proposed by Hu and Liu (2004).

6.2.   *Methodology*

To overcome the challenges of performing sentiment analysis on Twitter messages, our algorithm was developed to objectively evaluate the sentiment from the tweet message related to the financial

topics. It consists of four main components: data pre-processing, entity matching, sentiment word matching, and score computation (see Algorithm 1).

1. Data Pre-processing

    To properly evaluate the tweet data, it is essential to convert the raw data into a machine readable format for the algorithm to process. This data pre-processing procedure involves *filtering, stemming, splitting and n-gram recognition.*

    Word filtering: Unique attributes in tweet messages such as URLs and emoticons have empirically shown to cause minimal impact to the message sentiment. The following procedure was applied to the set of tweet messages:
    - Removal of URLs, User ID '@' and Hashtag '#'
    - Removal of emoticons and punctuations
    - Removal of stop words such as 'is', 'the' and 'at'

    Word stemming: Each word in the tweet message is reduced to its root form. This stemming process allows individual word to be uniformly treated and therefore aligned to its universal meaning. For example, words such as 'disappointing', 'disappointed', 'disappoints' are reduced to its stem word 'disappoint'.

    Word splitting and n-gram recognition: The procedure allows the algorithm to recognize units of individual words and phrases that consists of two or more words. The algorithm has the capability to process uni-grams (single word), bi-grams (two-word phrase) and tri-grams (three-word phrase). This feature is especially valuable for identifying phrases such as 'give up' so that the actual sentimental meaning is not lost.

2. Financial Entity Matching

    After the data pre-processing step, the next procedure is to pinpoint those messages that are related to the financial market topics. The rationale behind this critical step is to ensure that the level of noise is minimized in the study. We propose an approach to first form a list of financial entities by extracting commonly used languages from the top financial news broadcasters and traders' Twitter accounts, and then matching them against the individual words and phrases in the tweet message. Furthermore, each financial entity is quantitatively assigned a score to reflect its proximity to the financial related topics.

    Entity Extraction from top financial Twitter users: Sentiment analysis does not perform well if members of the financial community tweet messages that are unrelated to financial market. We conducted an entity matching process to extract messages with financial interests. A financial keyword corpus was then created using a sample dataset from 10/05/2013 to 02/05/2014 from the top financial news broadcasters and traders' Twitter accounts. The sample dataset was categorized according to the type of message initiator. Messages from the top news agencies and traders were labeled as "key messages" while messages from less important community users were "noise messages". By running text parsing and text filtering processes for the two groups, two respective keyword lists are obtained with frequency of occurrence reported. Through text parsing, stop words were dropped and only nouns and noun phrases were extracted from tweet messages. Entities that demonstrated higher occurrence in the "key messages" group were preserved to form the entity corpus while others are discarded. This step effectively screens for financial entities that are more commonly mentioned through established financial Twitter accounts. In addition, the corpus pairs each entity with a weight based on the frequency of occurrence. Higher weight can be interpreted as closer relevance to financial market topics.

    Company Name and Ticker: Another source of financial entity identification is the name and ticker symbol of major U.S. domestic companies. If a specific company or its ticker is mentioned, there is a high likelihood that the message contains financial-related information. As a result, the financial entity dictionary contains the names and their ticker symbols for over 6,000 companies listed on the three largest U.S. exchanges:

the New York Stock Exchange (NYSE), the NASDAQ stock market (NASDAQ) and the American Stock Exchange (AMEX).

Financial Entity Score Computation: All words and phrases from the tweet message are matched against the financial entity dictionary. Each message contains a score based on the degree of relevance to financial topics. The higher the financial entity score is, the heavier its weight counting towards its sentiment is. For instance, a tweet message containing the ticker symbol for General Electric 'GE' has a financial entity score of +1. If multiple financial entities are matched in one Twitter message, the financial entity score weight is equal to the highest weight among the matched financial entities (see Eq.(4)).

$$
\begin{aligned}
S^i_{entity} &= max(\omega(\boldsymbol{W^i_{message}} \cap \boldsymbol{W_{fe}})) \\
i &= 1, 2, \ldots, N
\end{aligned}
\tag{4}
$$

where S^i_{entity} is the financial entity score of message i, $\boldsymbol{W^i_{message}}$ is the word set split from message i, $\boldsymbol{W_{fe}}$ is the financial entity word set, $\omega(\boldsymbol{W})$ is the financial entity weights set of word set $\boldsymbol{W}$, N is the number of message and $n_{matched}$ is the number of matched financial entity.

3. Sentiment Word Matching

We use a comprehensive dictionary (SentiWordNet) that consists of over 8,000 sentimental words with respective positive and negative scores ranging from 0 to 1. For instance, the word 'happy' consists of a positive score with 0.125 and negative score with 0. We adopted this dictionary because of the wide scope of words in its collection and its track record in conducting accurate sentiment analysis (Khan *et al.* 2014, Bravo-Marquez *et al.* 2013). Moreover, the quantitative scoring of individual words facilitates our objective of computing sentiment score.

Word Matching: Every word and phrase in the tweet message is matched against the dictionary. The frequency of individual words and phrases within a given tweet message serves as an input to the sentiment scoring equation.

Detection of Words with Negative Connotation: With the message sentiment computed, it is important to handle words with negative connotation such as 'not' that reverse the actual sentiment meaning of the message. The algorithm uses a binary flag for detecting whether such words exist inside the tweet message in question.

Message Sentiment Computation: With the number of word occurrence and negative flag detection, the sentiment of a single tweet message can be generated. The message sentiment score ranges from -1 to 1, with -1 expressing the most negative sentiment, 0 being neutral, and +1 the most positive sentiment. The equation for computing sentiment score of message i can be found in Eq.(5). Among all the messages initiated by the top 2,500 critical nodes, the sentiment distribution approximately follows a normal distribution (see Figure 6).

$$
\begin{aligned}
S^i_{sentiment} &= \frac{\sum_j n^j_i \times s(j)}{\sum_j n^j_i} \times S^i_{entity} \times \mathrm{sgn}(i) \\
\mathrm{sgn}(i) &= \begin{cases} -1 & \text{if } \boldsymbol{W^i_{message}} \cap \boldsymbol{W_{neg}} \neq \emptyset \\ 1 & \text{others} \end{cases}
\end{aligned}
\tag{5}
$$

where $S^i_{sentiment}$ is the sentiment score of message i, S^i_{entity} is the financial entity score of message i computed by Eq.(4), $\boldsymbol{W^i_{message}}$ is the word set split from message i, $\boldsymbol{W_{neg}}$

is the negative connotation word set, n_i^j is the number of occurrence of SentiWordNet word j in message i, $s(j)$ if the sentiment score of word j
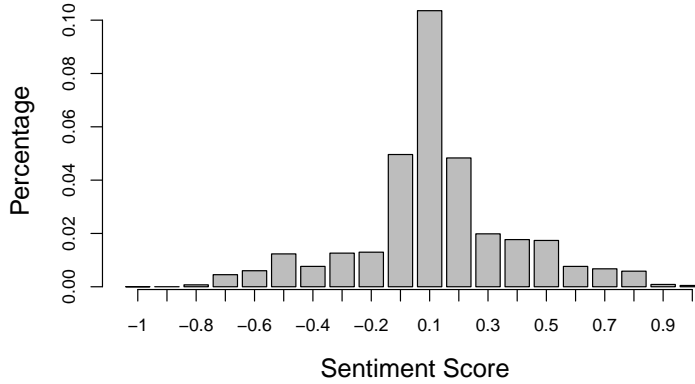

Figure 6. Histogram of Message Sentiment Distribution

4. Sentiment Daily Score Computation

The last procedure in the sentiment analysis algorithm is to compute the sentiment score for the day of observation for a given user. Three components are factored into the computation process: the score of the financial entity, the message sentiment score and the centrality score for the message initiator (see Eq.(6)). The algorithm then takes the average of the active user sentiment generated for each day to generate the daily sentiment score for regression studies.

$$S(t) = \frac{1}{N} \sum_{j=1}^{N} \omega_c^j \frac{\sum_{k=1}^{n_j(t)} S_{sentiment}^k(t)}{\sum_{k=1}^{n_j(t)} S_{entity}^k(t)} \tag{6}$$

where $S(t)$ is the daily sentiment score of day t, ω_c^j is the centrality weight of user j, $n_j(t)$ is the number of message by user j on day t, $S_{entity}^k(t)$ and $S_{sentiment}^k(t)$ are entity score and sentiment score of message k computed by Eq.(4) and Eq.(5) respectively.

7.   MARKET REGRESSION OF SENTIMENT

This section demonstrates the value of extracting sentiment based on the social structure of the financial community. A key hypothesis of this study is that Twitter sentiment extracted from the network's critical nodes serves as a reliable predictor of financial market movement. We believe that not all messages carry equivalent weight of information and therefore the message initiated from a more credible source should have larger influence on the financial community. Through regression analysis, we test whether specific critical nodes in the financial community have predictive power to key financial market measures. From the previous section, we identified three groups of 2,500 critical nodes based on key centrality measures from the financial community: betweenness centrality, out-degree centrality and closeness centrality. Through their tweet messages, we extracted their sentiments using the sentiment analysis algorithm and determined the statistical relationship with the historical daily return of major market returns and volatility indices.

Data: Twitter Messages
Result: Message Sentiment Score

while *not at the end of document* **do**

    $Message_{raw} \leftarrow$ read current message;

    Step 1: Data-Preprocessing;

    $Message_{splitted} \leftarrow$ Word-splitting(Message);

    $Message_{splitted} \leftarrow$ Remove-stop-words($Message_{splitted}$);

    $Message_{splitted} \leftarrow$ Word-stemming($Message_{splitted}$);

    Step 2: Financial Entity List Processing;

    $WordList_{FinancialEntity} \leftarrow$ Financial Entity List created from SAS;

    $MatchedWords_{FinancialEntity} \leftarrow Message_{splitted} \cap WordList_{FinancialEntity}$;

    $Score_{FinancialEntity} \leftarrow \max\{MatchedWords_{FinancialEntity}\}$;

    if $Score_{FinancialEntity} > 0$ **then**

        Step 3: Sentiment Word List Processing;

        $WordList_{Sentiment} \leftarrow SentiWordNetList$;

        $MatchedWords_{Sentiment} \leftarrow Message_{splitted} \cap WordList_{Sentiment}$;

        $Score_{SentiWordNet} \leftarrow \frac{1}{N} \sum\{MatchedWords_{Sentiment}\}$;

        Step 4: Inverse Factor;

        if *negative words/phrases (not n't) detected* **then**

            | $InverseFactor \leftarrow -1$

        else

            | $InverseFactor \leftarrow 1$

        end

        Step 5: Final Message Sentiment Score [-1,1];

        $Score_{Sentiment} \leftarrow Score_{FinancialEntity} \times Score_{SentiWordNet} \times InverseFactor$

    else

        Drop $Message_{raw}$;

        continue;

    end

end

Algorithm 1: Sentiment Analysis Algorithm Flowchart

7.1. *Data Sources*

We applied the regression analysis on 1,606,104 tweet messages for the period between 02/15/2014 and 06/15/2014. The daily sentiment series for the three major critical node groups were generated from their messages and then weighed with respect to the normalized centrality measures. In addition, we adopted the returns series for 6 exchange-traded funds which served as proxies for the historical daily market returns and volatility. The daily return is computed by taking the log return of market price (see Eq.(7)) from 02/15/2014 to 06/15/2014.

$$r_t = \log(\frac{p_t}{p_{t-1}}) \tag{7}$$

where r_t denotes the return of day t and p_t is the price of day t.

7.2. *Comparison among Number of Lags in Sentiment*

The objective is to examine whether the result is significant across different lags of the sentiment time series in relation to the financial market returns and volatility. A single lag refers to the sentiment of the previous day and the lag-1 regression model therefore captures the statistical

Table 5.   Exchange-traded Funds Ticker and Corresponding Index

Market Return	
SPY	S&P 500
DIA	Dow Jones Industrial Average
QQQ	NASDAQ
IWV	Russell 2000
Market Volatility	
VIX	CBOE Volatility Index

relationship of how the aggregate sentiment series of the previous day affects the current market movement. We ran multiple experiments (up to 10 lags of the sentiment series) for all centrality groups and found that the lag-1 regression model yields the most consistent and significant result (see Table 4). The models from lag-2 to lag-10 are not significant, suggesting that the sentiment from longer duration has minimal influence to the financial market. The result shows improvement compared with Bollen's finding that the lag-1 Twitter sentiment (generated from the OpinionFinder tool) is significant with p-value = 0.08[*] (Bollen *et al.* 2011). Therefore, we chose to use lag-1 sentiment for our modeling purpose.

Table 4.   Statistical signicance (p-values) of lag-n sentiment to SPY

Number of Lags	BC Score[1]	DC Score[2]	CC Score[3]
1	0.01 [**]	0.14	0.70
2	0.07 [*]	0.42	0.39
3	0.48	0.39	0.26
4	0.57	0.91	0.66

[1] Sentiment score weighted by betweenness centrality

[2] Sentiment score weighted by degree centrality

[3] Sentiment score weighted by closeness centrality

[**]$p < 0.05$

[*]$p < 0.1$

7.3.   *Linear Regression Model of Sentiment*

A linear regression model is applied to examine the relationship between the daily index return and the lag-1 Twitter message sentiment from the critical nodes. In particular, we investigated whether the lagged series of message sentiment have significant statistical relations with market returns and volatility. The dependent variable in the regression model is the daily return of respective market index. In this study, we used SPY, DIA, QQQ, IWV as the market returns and VIX as the volatility returns. These market indices represent major distinct components of the financial market by the characteristics of the underlying stock such as market capitalization and industry sector (see Table 5). The independent variable is the lagged time series of the message sentiment from three critical node groups. In the experiment, we test the lag-1 sentiment for their respective significance to the returns series (see Eq.(8)). If a significant relationship is observed, it suggests that returns lag behind the sentiment movement and therefore sentiment has predictive property over returns.

$$r_t = \beta_0 + \beta_1 S_{t-1} + \epsilon \tag{8}$$

where r_t denotes the daily return of day t and S_{t-1} denotes the daily sentiment score of day $t-1$

7.4.   *Comparison among Centrality Groups*

It is important to examine whether there is a fundamental difference of the regression result across the three different centrality measures in relation to the financial market returns and volatility. Varying the number of critical nodes in each group, we found that the betweenness centrality (BC) group consistently outperformed the degree centrality (DC) and closeness centrality groups (CC) (see Appendix B). The sentiment regression model of the BC group has shown significance across all market returns at the level of 95% (see Table 6). In addition, the positive coefficients of the model demonstrate the predictive capability that the more positive the message sentiment is, the higher market returns it leads to. For volatility, the betweenness centrality group is also more significant than the two other groups in terms of its significance level (see Table 6). The result shows that more positive sentiment leads to lower volatility level, vice versa. It is consistent with the observation that negative sentiment can cause a higher volatility spike, suggesting that bad news on Twitter increases the volatility of price return in the stock market.

7.5.   *Comparison among Number of Top Critical Nodes*

Along the same intuition that not all messages carry equivalence of information, we investigated whether there is an optimal number of a critical node for each centrality group in explaining financial market movement. For the extreme scenarios, too many critical node users may introduce unnecessary noise but too few users may omit key contributing sentiment for the regression study. Varying the number of critical nodes for each centrality group might yield results that reveal the emerging critical point and, therefore, lead to an enhanced indicator for explaining market movements. In this study, we investigate all three centrality measures starting from the top 100 users to 2,500 users in each group at an increment of 100 additional users. For instances, we first examine the top 100 users in the betweenness centrality group and then the top 200 users in the same group. We observe that an optimal point exists when the number of critical nodes in the group is 200 (see Figure 7). The coefficient for the 200-user group with the highest centrality measure is the most significant and consistent among all market indices. With the incorporation of more critical nodes in the regression model, we find that the p-value remains stabilized under the 0.05 level (except VIX volatility measure) for the models against market returns. This illustrates that our regression result is robust across different number of top critical nodes.

Table 6.   Lag-1 Sentiment Linear Regression Statistics
(n=200)

Market Indices	BC Score[1]		DC Score[2]		CC Score[3]	
	coeff	p-value	coeff	p-value	coeff	p-value
SPY	0.15	0.01**	0.15	0.07*	0.11	0.35
DIA	0.12	0.03**	0.13	0.11	0.09	0.41
QQQ	0.17	0.05**	0.16	0.20	0.13	0.46
IWV	0.16	0.01**	0.17	0.07*	0.11	0.38
VIX	-0.94	0.10*	-0.81	0.29	-1.14	0.29

[1] Sentiment score weighted by betweenness centrality

[2] Sentiment score weighted by degree centrality

[3] Sentiment score weighted by closeness centrality

**$p < 0.05$

*$p < 0.10$

Table 7.   Lag-1 Sentiment Linear Regression Statistics
(n=500)

Market Indices	BC Score[1]		DC Score[2]		CC Score[3]	
	coeff	p-value	coeff	p-value	coeff	p-value
SPY	0.16	0.01**	0.18	0.12	-0.10	0.54
DIA	0.14	0.03**	0.15	0.18	-0.06	0.69
QQQ	0.19	0.05**	0.17	0.30	-0.22	0.35
IWV	0.18	0.01**	0.19	0.11	-0.10	0.55
VIX	-1.01	0.09*	-0.95	0.36	0.91	0.53

[1] Sentiment score weighted by betweenness centrality

[2] Sentiment score weighted by degree centrality

[3] Sentiment score weighted by closeness centrality

**$p < 0.05$

*$p < 0.10$

Table 8.   Lag-1 Sentiment Linear Regression Statistics
(n=1000)

Market Indices	BC Score[1]		DC Score[2]		CC Score[3]	
	coeff	p-value	coeff	p-value	coeff	p-value
SPY	0.18	0.01**	0.23	0.11	-0.05	0.81
DIA	0.15	0.03**	0.20	0.16	-0.05	0.82
QQQ	0.21	0.05**	0.22	0.31	-0.09	0.79
IWV	0.19	0.01**	0.25	0.10*	-0.03	0.91
VIX	-1.11	0.09*	-1.24	0.34	-0.15	0.94

[1] Sentiment score weighted by betweenness centrality

[2] Sentiment score weighted by degree centrality

[3] Sentiment score weighted by closeness centrality

**$p < 0.05$

*$p < 0.10$

Table 9.   Lag-1 Sentiment Linear Regression Statistics
(n=2000)

Market Indices	BC Score[1]		DC Score[2]		CC Score[3]	
	coeff	p-value	coeff	p-value	coeff	p-value
SPY	0.19	0.01**	0.25	0.13	0.02	0.94
DIA	0.16	0.03**	0.22	0.18	0.05	0.86
QQQ	0.22	0.05**	-0.00	0.23	-0.10	0.82
IWV	0.20	0.01**	0.28	0.12	0.06	0.85
VIX	-1.15	0.09*	-1.40	0.35	-0.79	0.76

[1] Sentiment score weighted by betweenness centrality

[2] Sentiment score weighted by degree centrality

[3] Sentiment score weighted by closeness centrality

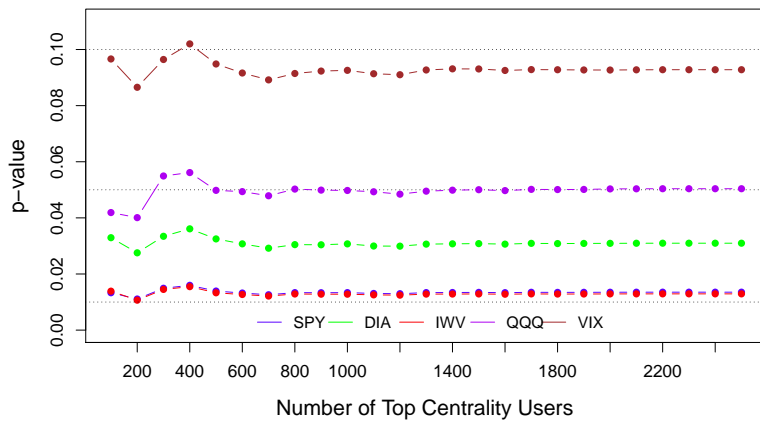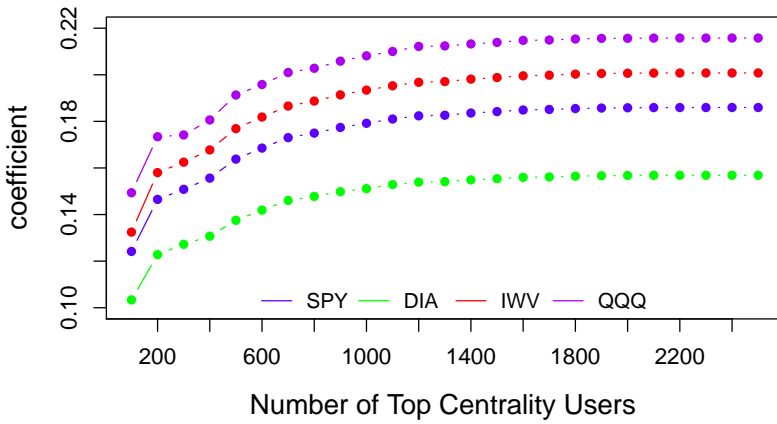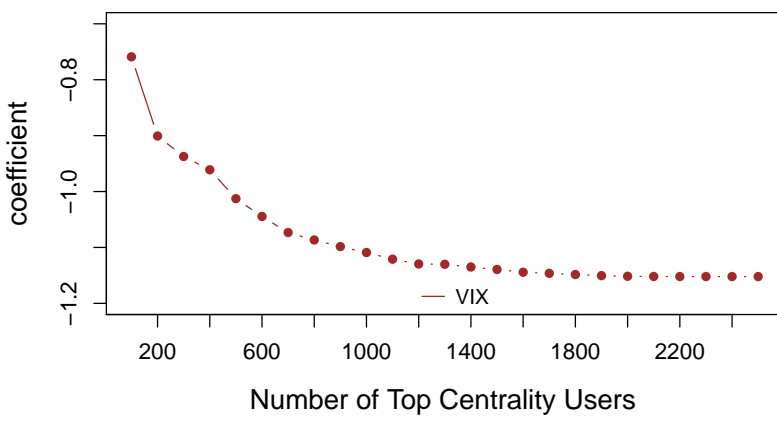**$p < 0.05$

*$p < 0.10$

Figure 7. Different Market Indices Comparison (p-value)

(a)

(b)

Figure 8. Different Market Indices Comparison (Coefficient)

7.6.   *Comparison with Different Market Indices*

The regression model reveals that different market indices exhibits varying level of market reaction towards the sentiment series of key centrality groups. We adopt the message sentiment of the betweenness centrality group since it has a stable and consistent statistical relationship with the market movement. For different market comparison, SPY and IWV have the best performance in terms of the model significance, while VIX and QQQ have the worst among all other indices (see Figure 7). This result suggests that the sentiment by the BC group can predict the S&P 500 and the Russell 3000 index with higher model accuracy over other indices. In addition, all models appear to follow similar shape with $n = 200$ at the most significant critical point. In terms of the parameter coefficient, the result shows that QQQ has the highest positive coefficient, followed by IWV and SPY, and DIA has the lowest positive coefficient. This order is consistent with the historical volatility level of the 4 different markets: with QQQ as the most volatile market index, followed by IWV, SPY and DIA.

7.7.   *Multivariate Regression Model and Cross Validation*

This section expands the current framework by incorporating additional exploratory terms and seeks to validate the robustness of the regression model. The motivation behind the procedure is that the three centrality user groups are theoretically different subsets of the financial community and thus their aggregated opinions can be dissimilar. Centrality measures reflect the importance of nodes expressing their opinion, and the interaction among two distinct groups can be a significant factor in the regression model. Therefore, by trying combinations of the first-order, second-order and interaction terms, the model selection process becomes more robust in the search for the best predictive model to the financial market.

A total of 27 combinations of the model were examined based on the interaction terms, first-order and second-order sentiment series. We found that the first-order BC and DC models remain the most significant after comparing the AIC and p-value across all models. This illustrates that the interaction and second-order term do not have significant influence in improving the result. This result might be due to the consensus of sentiment among different centrality groups. The predictive capability of the model can significantly improve the accuracy in the case when two groups express polarized opinions towards their beliefs in the financial market. The model with the interaction term can capture both group' sentiment and yields result that is more reflective towards the actual market movement. However, in the case when investors' opinion is in sync with a uniform direction, the model will not show significant difference.

In addition, cross-validation is applied on these two models to assess how they perform in practice with accuracy estimation. We tested up to 3 folds on the dataset because of the small sample of observations for training and testing, and we found that the mean squared error is significantly small (see Table 10 and Table 11). This justifies that the sentiment provided by the top 200-user betweenness centrality group and degree centrality group yield reliable and consistent result.

Table 10.   Cross-Validation of Mean Squared Error
(Betweenness Centrality Group)

fold	Market Indices				
	SPY	DIA	QQQ	IWV	VIX
1	3.31e-05	3.48e-05	5.94e-05	3.66e-05	0.004
2	5.13e-05	4.04e-05	1.38e-04	5.90e-05	0.003
3	2.26e-05	2.30e-05	4.88e-05	2.72e-05	0.002
mean	3.57e-05	3.28e-05	8.21e-05	4.09e-05	3.07e-03

Table 11.   Cross-Validation of Mean Squared Error
(Degree Centrality Group)

fold	Market Indices				
	SPY	DIA	QQQ	IWV	VIX
1	3.51e-05	3.71e-05	6.04e-05	3.87e-05	0.004
2	5.41e-05	4.24e-05	1.42e-04	6.21e-05	0.003
3	2.09e-05	2.21e-05	4.44e-05	2.53e-05	0.002
mean	3.67e-05	3.39e-05	8.22e-05	4.20e-05	3.09e-03

8.   DISCUSSION

The market sentiment regression highlights the significant influence of critical nodes towards movements in the financial market. Through performance comparison, the sentiment expressed by the betweeness centrality group is more significant and impactful than those by other community members. Critical nodes ranked by betweeness centrality degree yield the highest parameter significance close to the 99% level. In addition, the BC group also yields the highest positive sentiment coefficients against market return at 0.15**, 0.16**, 0.18** and 0.19** for the top 200, 500, 1000 and 2000 user group respectively. The volatility regression exhibits similar observations of sentiment coefficients at -0.94*, -1.01*, -1.11*, -1.15* with better significance level by the critical nodes ranked by betweenness centrality (see Table 6). This is consistent with the definition of critical nodes because of their central contribution to network connectedness. Grouping users in financial community by centrality analysis transforms the regression result to be more precise and accurate in explaining market movements.

Another significant finding of this study is the effect of using a smaller subset of the critical node groups. The tradeoff analysis of varying the number of critical nodes reveals the optimal point in yielding significant signals to financial market returns and volatility. This study found that selecting the top 200 users in the betweenness centrality group provide the most significant signal. A group with more than 200 critical node users dilutes the significance of the model but the result remains robust at a high significance level. The systematic search for the optimal number among the key centrality groups reinforces the main principle that not all messages should be treated with equal weight. Lastly, our comparison among different market indices signifies the value of extracting message sentiment based on social structure of the investment financial community. The sentiment series has potential applications on pairs trading strategy across multiple market indices.

This study shows that the behavior of critical nodes can be used to yield a more reliable indicator for the financial asset's price movement. It is worth noting that the current critical node analysis does not factor in the effect of tweets being read by unregistered users who may be active investors. This channel of information propagation is achieved by searching the Twitter website for specific keywords and the associated tweets would appear regardless whether the users are followers or not. This would facilitate the propagation mechanism of tweets to a wider community, including those tweets broadcasted by the critical nodes. Measuring the impact of this unobservable user group to the financial market will be a challenging problem, and we plan to address this issue in future studies. This empirical study of the financial community has three major contributions to the current literature of financial market and Twitter sentiment. First, it addresses the hypothesis that Twitter sentiment reflects the market participants' beliefs and behaviors toward future outcomes and the aggregate of the societal mood can present itself as a reliable predictor of financial market movement. Second, the concepts of leveraging critical nodes in the financial community generate a robust linkage between the social mood and financial market asset price movement. Moreover, the findings of critical nodes serve as an important guidance for regulatory authorities in paying attention to avoid manipulative or malicious actions such as the 2013 Associated Press hacking incident. Lastly, the empirical study provides insightful observations about the demographics and network structure of the financial community. By decomposing the community into unique user

types, beliefs and behaviors of market participants can be better understood. With the continual growth of the Twitter universe, the dynamic characteristics of the financial community can be better understood in terms of its network structure and the message sentiment influence towards the financial market movement.

9. CONCLUSION AND FUTURE RESEARCH WORK

This paper documents that there is a robust correlation between the social mood and financial market asset price movement by establishing a financial community. A key finding is that Twitter sentiment generated by critical nodes in the financial community yields a reliable indicator to predict financial market movement. We find that Twitter sentiment of the critical nodes has predictive power over market returns, and it consistently predicates market volatility as well. In the sentiment study, we show that different groups of critical nodes exert different degree of impacts to financial asset price and volatility movement.

For future work, this research can be extended into two areas related to the use of the financial community: First, the behaviors and beliefs of the critical nodes can be modeled as surrogates for information diffusion in financial markets. Sentiment or mood of the financial community can be modeled as the aggregate of the individual sentiment with memories. Swing of community sentiment can then be modeled as a complex and interactive system. Second, an agent-based simulation can be applied to the network growth and message propagation of the financial community. By modeling the dynamic properties of the community, the agent-based simulation can replicate important events in the social media platform and capture their key factors on the financial market movement.

Appendix A:

Top 25 Traders' Twitter Accounts: @howardlindzon, @alphatrends, @pkedrosky, @grassosteve, @chessNwine, @allstarcharts, @FOCUS_ON_RISK, @reformedbroker, @harmongreg, @ritholtz, @irenealdridge, @AnneMarieTrades, @optionalpha, @jimcramer, @PeterLBrandt, @1nvestor, @vader7x, @downtowntrader, @smbcapital, @tlmontana, @KeeneOnMarket, @ppearlman, @bespokeinvest, @fibline.

Top 25 Financial News Providers: @BloombergMrkts, @BloombergNews, @Forbes, @Reuters, @BW, @ftfinancenews, @CNNMoney, @CNBCFastMoney, @WSJ, @YahooFinance, @CNBCWorld, @MarketWatch, @IBDinvestors, @MSN_Money, @FoxBusiness, @RealMarketNews, @NYSE, @TabbFORUM, @EconBizFin, @CNBCClosingBell, @dowbands, @FXstreetNews, @Street_Insider, @HedgeWorld, @cr_harper.

Appendix B: LAG-1 SENTIMENT VS. MARKET INDEX RETURNS
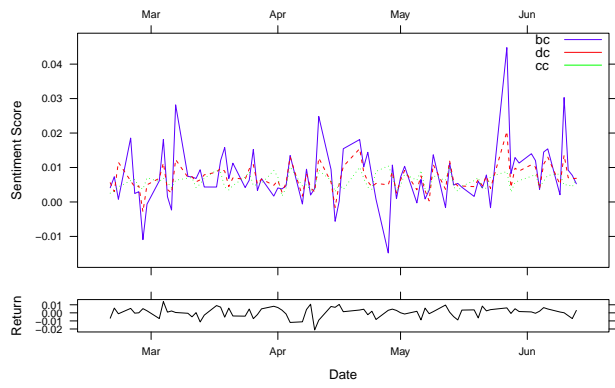

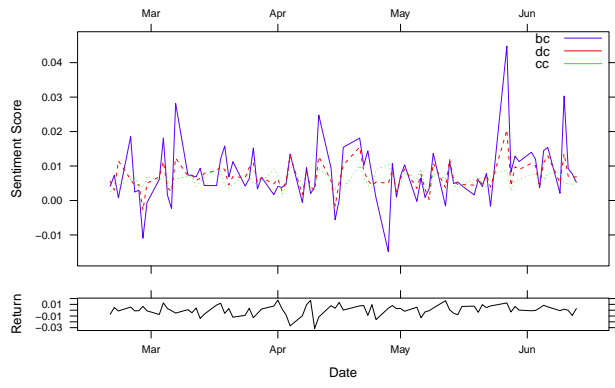
Figure B1. Lag-1 Sentiment Score vs. SPY Return
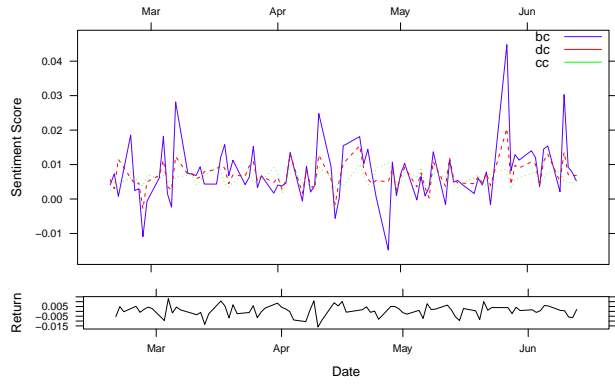

Figure B2. Lag-1 Sentiment Score vs. QQQ Return

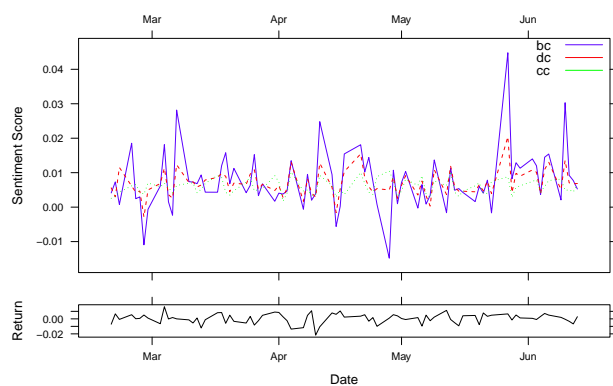Figure B3. Lag-1 Sentiment Score vs. DIA Return
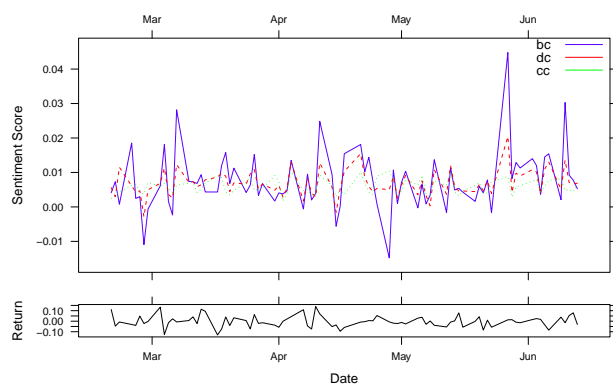
Figure B4. Lag-1 Sentiment Score vs. IWV Return

Figure B5. Lag-1 Sentiment Score vs. VIX Return

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., Sentiment Analysis of Twitter Data. In *Proceedings of the Proceedings of the Workshop on Languages in Social Media*, LSM '11, Portland, Oregon, pp. 30–38, 2011 (Association for Computational Linguistics: Stroudsburg, PA, USA).

Bakhshandeh, R., Samadi, M., Azimifar, Z. and Schaeffer, J., Degrees of separation in social networks. In *Proceedings of the Fourth Annual Symposium on Combinatorial Search*, 2011.

Bollen, J., Mao, H. and Zeng, X., Twitter mood predicts the stock market. *Journal of Computational Science*, 2011, **2**, 1 – 8.

Borgatti, S.P. and Everett, M.G., A Graph-theoretic perspective on centrality. *Social Networks*, 2006, **28**, 466 – 484.

Bradley, M.M. and Lang, P.J., Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

Bravo-Marquez, F., Mendoza, M. and Poblete, B., Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis. In *Proceedings of the Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, Chicago, Illinois, pp. 2:1–2:9, 2013 (ACM: New York, NY, USA).

Brown, G.W. and Cliff, M.T., Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 2004, **11**, 1 – 27.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P.K., Measuring User Influence in Twitter: The Million Follower Fallacy.. *ICWSM*, 2010, **10**, 10–17.

Chen, H., De, P., Hu, J. and Hwang, B.H., Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 2013.

Cohen, G. and Kudryavtsev, A., Investor Rationality and Financial Decisions. *Journal of Behavioral Finance*, 2012, **13**, 11–16.

Dodds, P. and Danforth, C., Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, 2010, **11**, 441–456.

Go, A., Bhayani, R. and Huang, L., Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009, pp. 1–12.

González-Arangüena, E., Manuel, C. and Del Pozo, M., Centrality in Directed Social Networks: A Game Theoretic Approach. In *Proceedings of the ICNAAM 2010: International Conference of Numerical Analysis and Applied Mathematics 2010*, Vol. 1281, pp. 815–818, 2010.

Hilton, D.J., The Psychology of Financial Decision-Making: Applications to Trading, Dealing, and Investment Analysis. *Journal of Psychology and Financial Markets*, 2001, **2**, 37–53.

Hu, M. and Liu, B., Mining and Summarizing Customer Reviews. In *Proceedings of the Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, Seattle, WA, USA, pp. 168–177, 2004 (ACM: New York, NY, USA).

Jackson, M.O., *Social and economic networks*, 2010, Princeton University Press.

Kaplan, A.M. and Haenlein, M., Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 2010, **53**, 59 – 68.

Khan, F.H., Bashir, S. and Qamar, U., TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 2014, **57**, 245 – 257.

Kleinberg, J., The Small-world Phenomenon: An Algorithmic Perspective. In *Proceedings of the Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, Portland, Oregon, USA, pp. 163–170, 2000 (ACM: New York, NY, USA).

Krishnamurthy, B., Gill, P. and Arlitt, M., A Few Chirps About Twitter. In *Proceedings of the Proceedings of the First Workshop on Online Social Networks*, WOSN '08, Seattle, WA, USA, pp. 19–24, 2008 (ACM: New York, NY, USA).

Latora, V. and Marchiori, M., Efficient Behavior of Small-World Networks. *Phys. Rev. Lett.*, 2001, **87**, 198701.

Mohammad, S.M., Kiritchenko, S. and Zhu, X., NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T. and Ureña López, L.A., Random Walk Weighting over Sentiwordnet for Sentiment Polarity Detection on Twitter. In *Proceedings of the Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, Jeju, Republic of Korea, pp. 3–10, 2012 (Association for Computational Linguistics: Stroudsburg, PA, USA).

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V. and Wilson, T., Semeval-2013 task 2: Sentiment analysis in twitter. , 2013.

Nofsinger, J.R., Social Mood and Financial Economics. *Journal of Behavioral Finance*, 2005, **6**, 144–160.

Ortega, R., Fonseca, A. and Montoyo, A., SSA-UO: Unsupervised Twitter Sentiment Analysis. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (* SEM)*, Vol. 2, pp. 501–507, 2013.

Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A., Correlating Financial Time Series with Micro-blogging Activity. In *Proceedings of the Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, Seattle, Washington, USA, pp. 513–522, 2012 (ACM: New York, NY, USA).

Tang, L. and Liu, H., Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2010, **2**, 1–137.

Travers, J. and Milgram, S., An experimental study of the small world problem. *Sociometry*, 1969, **32**, 425–443.

Watts, D.J., *Small worlds: the dynamics of networks between order and randomness*, 1999, Princeton university press.

Watts, D.J. and Strogatz, S.H., Collective dynamics of small-worldnetworks. *nature*, 1998, **393**, 440–442.

Zhang, X., Fuehres, H. and Gloor, P.A., Predicting Stock Market Indicators Through Twitter I hope it is not as bad as I fear. *Procedia - Social and Behavioral Sciences*, 2011, **26**, 55 – 62.

Zhi-Yun, Z., Bao-Hua, M., Hai-Ming, H., Jian-Zhi, G. and Jie-Jiao, Y., Regular Small-World Network. *Chinese Physics Letters*, 2009, **26**, 110502.