2015 IEEE International Conference on Data Mining

Modeling Social Attention for Stock Analysis: An Influence Propagation Perspective

Li Zhang[1], Keli Xiao[1*], Qi Liu[2], Yefan Tao[2], Yuefan Deng[1,3]

[1]Stony Brook University, NY, USA

[2]University of Science and Technology of China, Anhui, China

[3]National Supercomputer Center in Jinan, Shandong, China

li.zhang.2@stonybrook.edu, keli.xiao@stonybrook.edu*, qiliuql@ustc.edu.cn,

yft@mail.ustc.edu.cn, yuefan.deng@stonybrook.edu

***Abstract* - With the rapid growth of usage of social network, the patterns, the scales, and the rate of information exchange have brought profound impacts on research and practice in finance. One important topic is the stock market efficiency analysis. Traditional schemes in finance focus on identifying significant abnormal returns triggered by important events. However, those events are merely identified by regular financial announcements such as mergers, equity issuances, and financial reports. Related data-driven approaches mainly focus on developing trading strategies using social media data, while the results are usually lack of theoretical explanations. In this paper, we fill the gap between the usage of social media data and financial theories. We propose a Degree of Social Attention (DSA) framework for stock analysis based on influence propagation model. Specifically, we define the self-influence for users in a social network and the DSA for stocks. A recursive process is also designed for dynamic value updating. Furthermore, we provide two modified approaches to reduce the computational cost. Our testing results from the Chinese stock market suggest that the proposed framework effectively captures stock abnormal returns based on the related social media data; and DSA is verified to be a key factor to link social media activities to the stock market.**

Keywords - Social Network; Influence Propagation; Stock Social Attention; Market Efficiency

I. INTRODUCTION

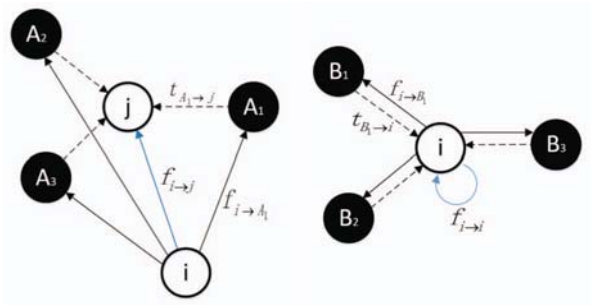
In this paper, we aim to explore the underlying relationship between social media and the stock market based on social influence models and the efficient market hypothesis. Since the efficient market hypothesis was formally introduced by [1], it was widely accepted and became one of the fundamental research topics in finance. Within the three forms of market efficiency (weak-form, semi-strong form, and strong form), we focus on the second one. While few studies are conducted for the strong form market efficiency because of its strict assumption that prices must reflect all public and private information, the other forms of market efficiency were widely studied in finance. In weak-form market efficiency, historical prices do not affect the future; technical analysis does not help in obtaining abnormal returns. Evidence against the weak form market efficiency can be found in many studies on momentum effect [2-4]. Semi-strong form efficient market hypothesis suggests that market prices should be fully reflected by public information, or abnormal returns will occur. This form of market efficiency is usually investigated based on event studies in order to identify the association between excess returns and different types of events, such as merger announcement [5] financial reports [6], analyst reports [7], and equity issuances [8]. However, studies on related topics are limited by the traditional event study database in which only standard types of events are included. To have a better understanding about financial market reactions on human activities, we focus on establishing a social influence based framework to 1) build a social influence system with a focus on stock-related activities, and 2) dynamically measure the social attention for specific stocks.

An appropriate influence propagation model would result good estimation of market influence for each post in a social network, and lead to effective studies on finance. The Independent Cascade (IC) model and the Linear Threshold (LT) model are considered as two of the most famous models in estimating social influence spread [9]. In both models, the influence spread is simply defined as the expected number of activated nodes. With the purpose of stock analysis, however, they may not be able to reflect the true market influence because every node is considered to be equally weighted during the spreading process. To associate social media activities with the stock market, we must model different market influence for social media users.

Reference [10] introduced a model that considers self-influence of individuals when computing social influence. **Fig 1(a)** shows the influence between two nodes in a social network. Solid lines represent the influence connections, and the arrows indicate directions of the influence propagation; dash lines with arrows represent the probability of a successful information propagation from one node to another. The main ideas of this work can be summarized as follows. First, the influence of node i on node j is determined by i's influence on

*Contact Author

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.117



(a) (b)

Fig 1. Social Networks and Influence Modeling

j's direct neighbors and their influence propagation to j. Second, the self-influence, which can be considered as the confidence, may not be always one (full confidence). However, the estimating methods for confidence level were insufficiently discussed in this work. To this end, we discuss more about the measurement of confidence with the thoughts of stock market analysis. On the other hand, although a lot of studies can be found in relevant topics on social network, few of them consider dynamic social influence. Therefore, while stock analysis based on social media activities becomes a new direction of research, we come up with improved methods for market influence modeling with the following perspectives.

- **Self-Influence (Confidence).** Most of the existing influence models do not consider the effect of self-influence or confidence during the influence spread estimation. However, in stock analysis, we believe that the market influence of individuals in a social network is determined by their social relations as well as their confidence in the articles they write, comment, and repost. If we assume people's confidence in specific topics mainly depends on the knowledge and expertise they have in handling related information, then different values must be assigned to describe the differences among people.
- **Dynamic Influence Updating.** Both the stock market and social media are dynamic systems. To explore the stock market dynamics by social media data, we propose an influence updating procedures to capture the real-time changes in a social media system. Based on the updated influence, we define the Degree of Social Attention for each stock. Hence, the two dynamic systems are able to connect together.
- **Computational Efficiency.** The computational complexity of influence modeling is extremely high because it contains high-dimensional matrix computation. A bigger challenge of computation is created with the purpose of dynamic stock analysis. To this end, we provide several alternative approaches as well as algorithm designs for computational efficiency.

In this paper, we mathematically define the confidence, and implement a social influence updating process based on the data of Weibo (the biggest social media platform in China). We further define the Degree of Social Attention (DSA) for stocks based on this dynamic influence modeling process. To evaluate our work, we study the market efficiency based on the DSA. The effects of DSA on market prices, volume, and abnormal returns are investigated. Our results confirm the effectiveness of our work by verifying the hypotheses we formulized based on the efficient market theory.

II. METHODOLOGY

In this section, we introduce our Degree of Social Attention (DSA) framework for stock analysis. First, we introduce the typical influence model and our improvement ideas in II.A - II.C. Then we define the DSA and discuss the theoretical ideas of its connection to the stock market based on the efficient market theory (II.D – II.E).

A. Social Influence Modeling

As shown in **Fig 1**, the probability of influence is propagated from node A to j can be denoted by $t_{A\to j}$, or t_{Aj} for simplification. The weighted-influence of i on j with the assessment of i's neighbor k, can be measured by $t_{kj}f_{i\to k}$. Assume $A = \{A_1, A_2, \dots A_n\}$ is the set of neighbors of node j, i's influence on j should be measured by the sum of the weighted-influence of i with the assessment of A. To solve this recursive function, an initial value should be assign to $f_{i\to i}$, the self-influence of i.

Based on influence model developed by [10], the influence of node i on j, can be defined as follows:

$$f_{i\to j} = \frac{1}{1+\lambda_j}\sum_{k\epsilon N_j} t_{kj}f_{i\to k}\,, \text{for } j \neq i$$
$$s.t.\ \ f_{i\to i} = \alpha_i \tag{1}$$

where N_j is j's trust-friend set. If $k\epsilon N_j$ then j and k are connected. And t_{kj} is the propagation probability form k to j. It can be measured by the probability that j will take actions on an article posted by k. For a special case, the propagation probability from i to itself, $t_{ii} = 1$. Parameter λ_j is the discount factor of j that measures the influence diminishing during propagation. We follow the setup in the original work that is to choose the same λ for each node i. $\alpha_i \in [0,1]$ is the prior constraint value being assigned to each node i. If i has a full confidence to the information, this value is assigned as one; if i has no confidence at all, it would be zero.

Then, the social influence of i can be defined as follows:

$$f_{i\to N} = \sum_{j\epsilon N} f_{i\to j.} \tag{2}$$

To solve the problem, rewrite (1) as:

$$f_{i\to j} = \frac{1}{1+\lambda}\sum_{k\epsilon N} t_{kj}f_{i\to k} + v_{ij}, \tag{3}$$

where N represents the entire network which contains all nodes, and we could use it instead of N_j because $t_{kj} = 0$ if $k \notin N_j$, so it would not change the value of final result. v_{ij} is the j-th entry in a vector $v_i = [0,0,\dots,v_{ii},\dots,0]'$, in which only the i-th entry v_{ii} is nonzero and it ensures $f_{i\to i} = \alpha_i$. Then we have the influence spread vector $\mathbf{f_i} = [f_{i\to 1}, f_{i\to 2}, \dots, f_{i\to n}]'$.

Based on (3), we get:

$$\begin{aligned}\mathbf{f_i} &= (\mathbf{I} + \lambda\mathbf{I})^{-1}(\mathbf{\Gamma}'\mathbf{f_i} + \mathbf{v_i}) \\ &= (\mathbf{I} + \lambda\mathbf{I} - \mathbf{\Gamma}')^{-1}\mathbf{v_i} \\ &= \mathbf{P}v_{ii}\end{aligned} \tag{4}$$

where $\mathbf{I}$ is the element matrix and $\mathbf{\Gamma}$ is the N by N influence transition matrix of t_{kj}; $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{\Gamma}')^{-1}$. In (4), $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{\Gamma}')$ is invertible because it is strictly diagonally dominant as $t_{ii} = 1$. As $f_{i\to i} = \alpha_i = v_{ii}p_{ii}$ and v_i only has the i-th entry v_{ii} nonzero, we get:

$$f_{i\to j} = \frac{\alpha_i}{p_{ii}}p_{ji}. \tag{5}$$

Therefore, (2) can be rewritten as:

$$f_{i\to N} = \sum_{j\epsilon N} f_{i\to j} = \frac{\alpha_i}{p_{ii}}\sum_{j=1}^{N} p_{ji} \tag{6}$$

In summary, given the influence transition matrix $\mathbf{\Gamma}$, the damping factor and the prior constraint α_i, we can compute the influence $f_{i\to j}$ and the social influence $f_{i\to N}$ based on (5) and (6).

B. Self-Influence: Confidence

Based on (1), $f_{i\to i}$ has to be given before the influence $f_{i\to j}$ can be computed. Reference [10] mentioned that $f_{i\to i}$ can be viewed as the confidence level of i, but no method has been provided for the estimation. We believe that the confidence levels are determined during the interactions among social media users. In terms of stock-related posts, the more feedback a node receive from its friends, the more confidence it gains. Here we do not consider whether the feedback is positive or negative, because either way would lead an increasing social influence.

As shown in **Fig 1(b)**, the self-influence (confidence) of node i, $f_{i\to i}$ depends on 1) how much influence i has on his neighbors, and 2) the probability of the neighbors would react on i's actions. The value of someone's confidence is high when his neighbors give more reactions to the articles posted by him or he has large influence on them. Therefore, we define the confidence of node i as follows.

***Definition 1*:** The confidence of user i in a social network is the sum of the product of the user's influence on its direct neighbors and the probability of feedback it receives from those neighbors.

Algorithm 1: Influence Updating

Input: $\lambda,\ \alpha_i, T_{now}$
Output: $f_{i\to N}^{T}$ and $\mathbf{f_i}^{T} = [f_{i\to 1}^{T}, f_{i\to 2}^{T}, \dots, f_{i\to n}^{T}]'$, for $T = T_{now}$

$\alpha_i = 1$;
for ($T = 0; T < T_{now}; T$++) **do**
  if ($T > 0$)
    $\mathbf{\Gamma}_{T-1} = \mathbf{\Gamma}_T$; //save the probability matrix for T-1
    $\mathbf{f_i}^{T-1} = \mathbf{f_i}^{T}$; //save the influence matrix for T-1
    //update α_i based on $\mathbf{\Gamma}_{T-1}$ and $\mathbf{f_i}^{T-1}$.
    $\alpha_i = \frac{1}{1+\lambda}\sum_{k\epsilon N} t_{ki}^{T-1} f_{i\to k}^{T-1}$;
  end if
  Compute $\mathbf{\Gamma}_T = [t_{ij}]_{n\times n}$ based on the past one-month data;
  for ($i = 0; i < n; i$++) **do**
    for ($j = 0; j < n; j$++) **do**
      $f_{i\to j} = \frac{\alpha_i}{p_{ii}}p_{ij}$;
      $\mathbf{f_i}^{T}$.pushback($f_{i\to j}$);
    end for
  end for
  $f_{i\to N}^{T} = \mathrm{sum}(\mathbf{f_i}^{T})$;
  return $f_{i\to N}^{T}, \mathbf{f_i}^{T}$;
end for

It can be mathematically expressed as:

$$f_{i\to i} = \frac{1}{1+\lambda}\sum_{k\epsilon N_i} t_{ki} f_{i\to k}, \tag{7}$$

where N_i is i's trust-friend set, t_{ki} is the propagation probability from k to i, and $f_{i\to k}$ is the influence of i on k. Then (1) is modified as follows:

$$f_{i\to j} = \frac{1}{1+\lambda}\sum_{k\epsilon N_j} t_{kj} f_{i\to k}\,, \text{for } \forall i, j \in N. \tag{8}$$

This recursive equation cannot be solved if $f_{i\to i}$ is not given in a static system. However, when it comes to a dynamic system, it can be solved with an initial value of $f_{i\to i}$.

C. Dynamic Influence Updating

Now we consider the time-varying influence in a dynamic system. We design the following influence updating process to ensure that the influence values reflect the up-to-date performance of social media users. As shown in **Algorithm 1**, the confidence for node $i \in N$ at time T is denoted by:

$$\begin{aligned} f_{i\to i}^{T} &= \frac{1}{1+\lambda}\sum_{k\epsilon N} t_{ki}^{T-1} f_{i\to k}^{T-1}, T = 0,1,2,\dots \\ &\text{s.t.} \quad f_{i\to i}^{0} = 1. \end{aligned} \tag{9}$$

We set $f_{i\to i}^{0} = 1$ for each node as the initial value by assuming people initially have full confidence when they join a social network. The influence of i on j, $f_{i\to j}$, and i's total social influence, $f_{i\to N}$, are further updated over time based on (5) and (6). The whole social influence system is updated with α_i based on the information in the previous time window.

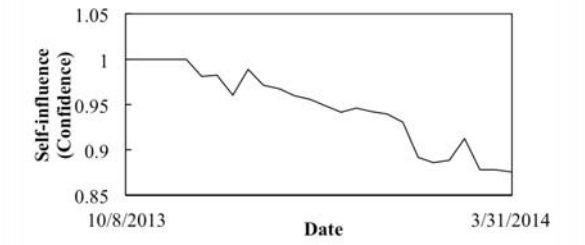


Fig 2. An Example of Dynamic Confidence

Fig 2 shows an example of the dynamic confidence for a social media user. As can be seen, the confidence is initialized to 1 and then fluctuates over time. Based on ***Definition 1***, these changes depend on the user's current social influence as well as the reactions from its direct neighbors.

An important issue of the algorithm is the determination of the length of time window. Large time windows may result the existence of accumulative outdated information or noise data; small ones may result biased samples because important information just recorded can be easily discarded. In this paper, we update the influence weekly based on a time window which covers the past four weeks' social media data. We provide detailed discussion on the window size in section V.

D. Degree of Social Attention

Given a social network $G = (N, E)$, where N is a set of nodes (social media users) and E is a set of edges, let $O_q|_{T_s}^{T_e}$ be the set of information flows regarding stock q, which appear in G from time T_s to T_e. The Degree of Social Attention (DSA) to stock q within the time frame can be defined as:

$$DSA_q|_l^m = \sum_{o \in O_{lm}} \mathcal{L}_o, \tag{10}$$

where $l, m > 0$ and $\mathcal{L}_o$ is the influence of information flow o. Assume the information flow o is posted by individual $i \in N$, we define $\mathcal{L}_o$ as the social influence of i, denoted by $f_{i \to N}$. For simplification, we use DSA_q to represent $DSA_q|_{T_s}^{T_e}$. Then, the DSA function is rewritten as:

$$DSA_q = \sum_{i \in W_q^i} \varphi_i d_i f_{i \to N}, \tag{11}$$

where W_q^i is the whole set of participators who discussed stock q between time T_s and T_e; d_i is the number of articles posted by i during that time; φ_i is the discount factor determined by i's personal characteristics such as age, title, education, number of followers, and frequency of article posting. To simplify the estimation process, we set φ_i to 1 for all individuals.

E. Stock Abnormal Returns

Here, we theoretically analyze the relationship between the stock market and social media activities. As a typical setting in finance, stock traders are separated into two types: informed traders and uninformed traders [11]. While the informed traders have significant advantages in terms of specialized information, technical skills, and capital power during trading, uninformed traders are considered as noise traders who trade on what they think is information but in fact is merely noise. Theoretically, when assets are mispriced, the activities of informed traders would pull prices back to fundamental values hence abnormal returns are reduced; the activities of uninformed traders would generate noise, and increase stock abnormal returns. Therefore, to explore the relationship between social media and the stock market, we need to figure out whether the social media users are informed traders or uninformed traders. Furthermore, stock abnormal returns must be the key to identify such relationship.

In finance, an abnormal return (excess return) is defined as the difference between the actual return and the expected return. For a stock q, its abnormal return at time T can be represented as follows:

$$AR_{q,T} = R_{q,T} - E(R_{q,T}), \tag{12}$$

where $AR_{q,T}$ is the abnormal return of stock q at time T; $R_{q,T}$ is the actual return of the stock; and $E(R_{q,T})$ is the expected return. To get the expected return, we first estimate the beta based on an OLS regression as follows:

$$R_{q,T} = \alpha_q + \beta_q(R_{m,T}) + \varepsilon_{q,T}, \quad \text{for } T = 0,1,2,\ldots, \tag{13}$$

where $R_{m,T}$ is the market return[1]; β_q is coefficient between the stock and market returns. Assume the risk-free rate at time T is $R_{f,T}$, the expected return for stock q can be estimated based on the Capital Asset Pricing Model (CAPM):

$$E(R_{q,T}) = R_{f,T} + \hat{\beta}_q(R_{m,T} - R_{f,T}). \tag{14}$$

Regarding the relationship between social media and the stock market, our first question is whether DSA can be directly used in stock forecasting. We believe that the answer also depends on whether the social media users are informed traders or uninformed traders. Thus, we propose the first testable hypothesis as follows.

***Hypothesis 1*:** There is no significant relationship between the Degree of Social Attention (DSA) and stock returns if most of the active social media users are uninformed traders. Some relationships may be found if most of the active users are informed traders.

Now we discuss the relationship between the DSA and stock abnormal returns. It is a common agreement that the Chinese market is empirically inefficient in the Semi-Strong form [12]. This indicates that public information is not fully reflected by the market price. During the trading period, uninformed traders usually cannot process information efficiently, so they simply follow the market, and cause over

[1] Market returns are computed based on market index, such as S&P 500 and Nasdaq composite. In our paper, we study the Chinese market and use CSI 300 as the market index.

trading and an increasing amount of abnormal returns. On the other hand, the trading activities of informed traders would pull the market price back to the intrinsic value, and reduce the abnormal return. Thus, our second hypothesis is formulized as:

***Hypothesis 2*:** If the majority of active social media users are uninformed traders, there should be a positive relationship between the Degree of Social Attention and the absolute value of abnormal returns; if informed traders are in the majority, a negative relationship should be found.

If the two hypotheses can be verified, then the effectiveness of DSA as the key factor to link social media activities to the stock market is confirmed. The result would further serve as a new evidence of semi-strong form inefficiency in the Chinese stock market. We extend the discussion in section V.

III. Computation of Influence Matrix

One challenge of implementing the influence updating process is the problem of high computational complexity. The algorithm contains a process of N by N matrix inversion, $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{\Gamma}')^{-1}$ for each time T. (The propagation probability t_{ij} which forms the propagation matrix $\mathbf{\Gamma}$ can be calculated as the probability that node j reacts on node i's posts - comment, repost, like - during each time period) The complexity of matrix inversion with preferred methods, such as Gaussian Elimination, is $O(N^3)$. While it is difficult to reduce the computational complexity of inverting matrix, we provide two approaches to reduce matrix dimension with the purpose of stock analysis.

A. Efficient Approaches

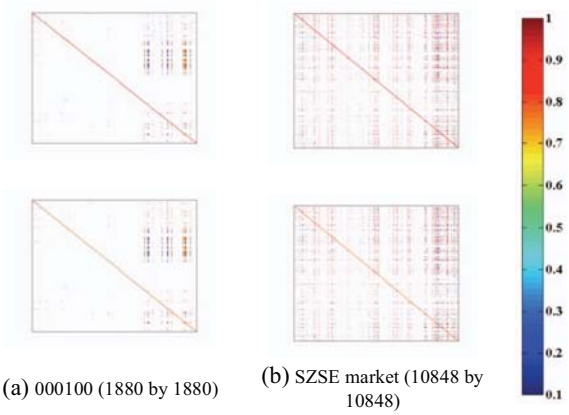
1) Market-based Approach

Although there are a huge number of users in the whole social network, it is not necessary to include all of them for stock analysis. We believe that only those who participate in the discussions of stock-related topics can influence the market. Furthermore, it can be true that people are only interested in stocks listed on the same market, say, NYSE or Nasdaq. In this case, if we separately consider the users who discuss stocks from different markets, we would only need to handle a matrix with smaller size for each market. Therefore, we modify the influence modeling as follows:

$$f_{i\to j,\theta} = \frac{1}{1+\lambda}\sum_{k\in N_m} t_{kj,\theta} f_{i\to k,\theta}\,, \text{for } j \neq i,$$
$$s.t.\ \ f_{i\to i,\theta} = \alpha_{i,\theta} \tag{15}$$
$$\theta = selected\ market$$

where the size of θ is the number of stock markets.

While this approach significantly reduces the size of matrix $\mathbf{P}$ by separately considering each market, the limitations are still obvious. First, it only works when the overlapping among these markets is small. If we find that most of people are interested in all markets, the size would not be reduced much.



(a) 000100 (1880 by 1880) (b) SZSE market (10848 by 10848)

Fig 3. Potential diagram of the influence matrix for (a) 000100 and (b) SZSE market, at 11/4/2013 (upper) and 1/13/2014 (lower)

Second, there are usually no more than three stock markets within a country. For instance, there are only two major stock markets in the U.S. (NYSE and Nasdaq). Therefore, the optimal case is to reduce the matrix size to 1/3 with which the complexity is still too high.

2) Stock-Based Approach

To further reduce the computational cost, we illustrate a stock-based approach which considers each single stock as an independent system. Similar to the market-based approach, we assume people are interested in a small group of particular stocks in a certain time but focusing on all of them. So that we can separate implement the modeling process based on data of each stock. We follow the same process of (15) where θ links to a selected stock but market.

Fig 3 shows an example of how influence matrix $\mathbf{F} = [f_{i\to j}]_{n\times n}$ is updated over time for both approaches. We choose the $\lambda = 0.176$ as suggested in [10]. As shown in the figure, $\mathbf{F}$ is an asymmetry matrix which indicates that $f_{i\to j} \neq f_{j\to i}$. The values on the diagonal denote self-influence or confidence. Blank areas denote zero-influence among users. **Fig 3(a)** and **(b)** plots the influence matrix for two examples, stocks 000100 and Shenzhen Stock Exchange (SZSE) market.

B. Algorithm Parrellization

The market-based and stock-based approach are designed to reduce the size of matrix $\mathbf{P}$, hence reduces the computational cost. We use the generalized minimal residual method (GMRES) [13] with QR factorization method to solve this matrix inversion problem, as shown in **Algorithm 2**. Moreover, since the divided tasks from the two approaches are independent with each other, parallel computing techniques can be easily applied to enhance the computational efficiency. We show the speedup results in Section V.

IV. Data Processing

We collected stock-related data from two major sources: the social media information and the stock market data. Our work focuses on the Chinese stock market - Shenzhen Stock

TABLE 1. WEIBO DATA PROFILING

Data	Feature Descriptions
Basic Identifications	Weibo ID Account ID Post Content Date and Time
Influence-Related	Number of "like" Number of Reposts Number of Comments
Reaction Tracking	Reaction ID

Exchange (SZSE) and Shanghai Stock Exchange (SSE) - and the social media activities in Weibo, the largest mobile social network in China.

A. Social Media Data

The API provided by Weibo has many restrictions for data collection, so an alternative way is implemented in our paper. We program based on an open source tool HtmlUnit, to model HTML documents of Weibo, then identify and retrieve information we need.

TABLE 1 summarizes the collected features of social media data. We collected three types of features for stock analysis. The first one contains the basic attributes of each posted article including the article ID, author account ID, the content, and the date and time for posting. The second one contains the features measuring social reactions, such as the number of times of the article being 'like', 'repost', and 'comment'. The last type of features includes the IDs of reactions, which we use to track the characteristics of each participator.

TABLE 2 reports the key statistics of the experimental data. The full sample contains 6-month data from October 2013 to March 2014. It includes stock-related articles of selected stocks listed on SZSE and SSE. Influence modeling is implemented based on the data of the users who have stock-related activities. The information of other users is discarded because we do not consider them as stock market participators. Even if some of them have large social influence, their influence on the stock market is considered to be zero.

B. Stock Data

Our stock data includes 10- minute price and volume of 10 highly active Chinese stocks from October 8th 2013 to March 31st 2014. The Shanghai Shenzhen CSI 300 Index is also collected as the market index. All prices are adjusted for dividends and splits. TABLE 3 shows the 10 stocks in our

TABLE 2. STATISTICS OF THE EXPERIMENTAL DATA

Data Sources	Properties	Statistics
Common	Time Scale	10/08/2013 – 03/31/2014
	Number of Days	174
Social Media	Number of Posts	139,855
	Number of Accounts	20,410
	Number of Posts per Day	803.76
Stock Market	Number of Stocks	10
	Number of Business Clusters	10
	Number of Trading days	119
	Data Frequency	10-minute
	Number of Time Points	3,094 per stock

Algorithm 2: GMRES with QR factorization method

Input: λ, $\boldsymbol{\Gamma}$
Output: $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \boldsymbol{\Gamma}')^{-1}$

$\mathbf{A} = \mathbf{I} + \lambda\mathbf{I} - \boldsymbol{\Gamma}'$
$\rightarrow \mathbf{A} * \mathbf{P} = \mathbf{I}$

//QR factorization
$\mathbf{A} = \mathbf{Q} * \mathbf{R}$
// where $\mathbf{Q} * \mathbf{Q}' = \mathbf{I}$ and $\mathbf{R}$ is a upper triangular matrix
$\rightarrow \mathbf{R} * \mathbf{P} = \mathbf{Q}'$
for $(i = 1; i \leq N; i{+}{+})$ **do**
//GMRES solve $\mathrm{P}_{\cdot i}$ for $\mathbf{R} * \mathrm{P}_{\cdot i} = \mathrm{Q}'_{\cdot i}$
$\mathrm{P}_{\cdot i} = 0$
$r = \mathrm{Q}'_{\cdot i} - \mathbf{R} * \mathrm{P}_{\cdot i}$
do while $\| r \| > 1e^{-6}$
$v_1 = r / \| r \|$
for $j = 1$ **step** 1 **until** k **do**
for $i = 1$ **step** 1 **until** j **do**
$h_{i,j} = (\mathbf{R}v_j)'v_i$
end for
$\tilde{v}_{j+1} = \mathbf{R}v_j - \sum_{i=1}^{j} h_{i,j}v_i$
$h_{j+1,j} = \| \tilde{v}_{j+1} \|$
$v_{j+1} = \tilde{v}_{j+1} / h_{j+1,j}$
end for
$\mathrm{P}_{\cdot i} = \mathrm{P}_{\cdot i} + V_j y_j$
$r = \mathrm{Q}'_{\cdot i} - \mathbf{R} * \mathrm{P}_{\cdot i}$
end do
end for
return P

sample being considered as the leader companies in different business clusters, they are discussed more often in social media than other companies. The Chinese stock market is open very Monday to Friday with two separated sessions. The morning session begins from 9:30 to 11:30; the afternoon session starts from 13:00 to 15:00. To avoid the noise information overnight and during the lunch break from 11:30 to 13:00, we remove the first 30 minutes of each session from our sample. Based on the sample, we compute stock returns as the current change of price divided by the previous price. Abnormal returns are computed based on the approach we introduced in section II.

C. Sample Selection

Based on the efficient market theory, sufficient trading activities would reduce the abnormal return. If significant relationship between social media activities and abnormal returns can be found in top trading stocks, more evidence must be found in less active ones. Therefore, we only focus on

TABLE 3. SELECTED STOCKS IN SAMPLE

Stock Code	000100	000157	000709	000783	002024
Company Name	TCL	Zoomlion Heavy Industry Sci. and Tech.	Hebei Steel	Changjiang Securities	Suning Commerce Group
Business Sectors	Consumer Goods	Industrial Goods	Basic Materials	Financial	Financial
Stock Code	600048	600221	600688	601018	601989
Company Name	Baoli Real Estate	Hainan Airlines	Sinopec Shanghai Petrochemical Corp	Ningbo Port	China Shipbuilding Industry Corp
Business Sectors	Financial	Services	Basic Materials	Services	Services

highly trading stocks which simultaneously have sufficient discussions in social media. Our sample includes 10 highly active Chinese stocks from different business sectors. They are selected based on the daily average trading volume and social media attention from October 2013 to March 2014.

V. Experimental Results

We evaluate our model based on two major considerations as follows. On one hand, we investigate whether the Degree of Social Attention (DSA), as the major index we propose in this paper, has significant association with abnormal returns. On the other, we check whether the computational cost is affordable during the influence updating process.

A. Sample Analysis and Determination of Time Window Size

Fig 4 shows the distribution of the sample social influence $f_{i\to N}$. Among the 20,410 accounts in our sample, we compute the distribution of the average social influence of i. As can be seen, 86.4% social media users are with social influence less than one ($ln(f_{i\to N}) < 1$). They are considered as small social influence participants. The rest 13.6% who have log social influence greater than one are considered as big and medium influence participants. Thus, we can assume most of the social media users are uninformed traders because of their small social influence.

To have a better understanding about the characteristics of social media activities, we define the weighted average influence (WAI) of user i on its direct neighbors as follows:

$$WAI_i = \overline{f_i} = \frac{1}{\|N_i\|}\sum_{i\in N_i} f_{i\to N.} \qquad (16)$$

The value of WAI is between 0 and 1. The higher WAI, the larger influence i has on each of its neighbor. If we set 0.5 as the baseline to separate "high WAI" and "low WAI", **Fig 5** shows the percentages of two groups of users for all 10 stocks in sample. As can be seen, the number of high WAI users is higher than the number of low WAI users for most of stocks, while stock 000709 and 600221 are two exceptions.

To determine appropriate moving window size for influence updating, we look at the frequency of new social media posts. As talked before, the high WAI accounts are considered as they have more impact for the influence updating. For statistical data, the frequency of posts for these

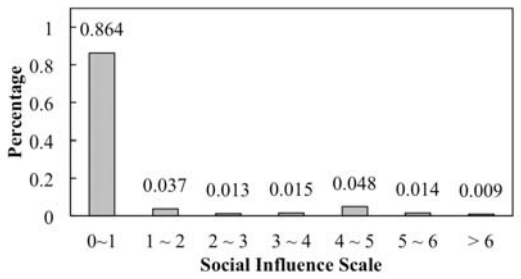


Fig 4. Distribution of the Social Influence $f_{i\to N}$

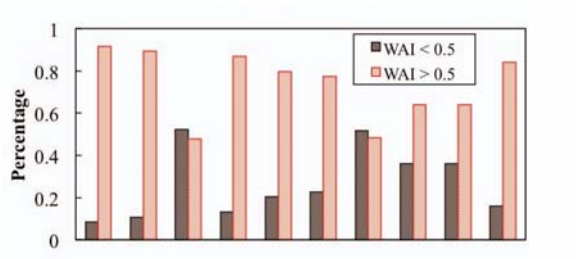


Fig 5. Percentage of the Weighted Average Influence $\overline{f_i}$

*From left to right: Stock 000100, 000157, **000709**, 000783, 002024, 600048, **600221**, 600688, 601018, 601989

accounts is about 6.45 days and the average time for all these accounts with at least one post is about 26.71 days. In our work, for simplified the process, we update the influence for every 7 days / 1 week based on the information from social media in the past 28 days / 4 weeks. When time moves forward, our influence system will keep updating based on dynamic information.

B. Experimental Design

Considering that uninformed traders are in the majority, we validate the effectiveness of our framework by testing the ***Hypotheses 1*** and ***2*** formulized in section II.

Hypothesis 1 suggests that the degree of social attention (DSA) cannot directly explain the price movement of each stock in terms of capturing the stock return. We test this hypothesis by model (i) as follow:

$$R_{q,T} = \mu_{0,q} + \mu_{1,q} dsa_{q,T} + \mu_{2,q} R_{q,T-1} + err, \qquad (17)$$

where $R_{q,T}$ is the return of stock q at time T; $dsa_{q,T} = \log(DSA_{q,T})$; $\mu_{0,q}$ is a constant term and $\mu_{1,q}$ is the coefficient between $R_{q,T}$ and $dsa_{q,T}$; $\mu_{2,q}$ is the coefficient between $R_{q,T}$ and $R_{q,T-1}$; $err \sim i.i.d$. The DSA values are non-negative because social influence is positive all the time. However, returns can be positive or negative. While DSA cannot separate optimistic and pessimistic discussions in social media, we include the return with one lag as an independent variable to capture momentum of the stock price movement. Following ***Hypothesis 1***, we do not expect to find any significant relationship between the return and DSA.

Based on our ***Hypothesis 2***, we expect to find positive relationship between the absolute value of abnormal return and DSA. Model (ii) is used for this test:

$$|AR_{q,T}| = \mu'_{0,q} + \mu'_{1,q} dsa_{q,T} + err' \qquad (18)$$

where $|AR_{q,T}|$ is the absolute value of abnormal return; $\mu'_{0,q}$ is a constant term and $\mu'_{1,q}$ is the coefficient between $|AR_{q,T}|$ and $dsa_{q,T}$; $err' \sim i.i.d$.

As an additional check, we test the relationship between the trading volume and DSA by model (iii) as follows:

TABLE 4. SIGNIFICANCE TEST BASED ON MARKET-BASED APPROACH

Stock Code		(i) Return ($R_{q,T}$)			(ii) Abnormal Return ($\lvert AR_{q,T}\rvert$)		(iii) Logarithmic Volume ($vol_{q,T}$)	
		(Intercept)	$dsa_{q,T}$	$R_{q,T-1}$	(Intercept)	$dsa_{q,T}$	(Intercept)	$dsa_{q,T}$
000100	Coefficient	-1.110e-04	-6.546e-07	-0.2388	0.0024	**4.582e-05**	15.5344	**0.1571**
	p-value	(0.1398)	(0.8456)	(0.0000)	(0.0000)	**(0.0961)**	(0.0000)	**(0.0000)**
000157	Coefficient	-2.013e-04	-9.167e-07	-0.1931	0.0012	**6.568e-05**	13.8218	**0.1208**
	p-value	(0.0000)	(0.5100)	(0.0000)	(0.0000)	**(0.0116)**	(0.0000)	**(0.0000)**
000709	Coefficient	-8442e-05	3.309e-08	-0.3350	0.0028	-3.901e-05	13.6853	**0.4416**
	p-value	(0.2702)	(0.9903)	(0.0000)	(0.0000)	(0.8177)	(0.0000)	**(0.0000)**
000783	Coefficient	-3.324e-04	2.269e-06	-0.1127	0.0014	**1.550e-04**	13.8036	**0.2395**
	p-value	(0.0000)	(0.1820)	(0.0000)	(0.0000)	**(0.0011)**	(0.0000)	**(0.0000)**
002024	Coefficient	-2.400e-04	-9.497e-07	-0.1414	0.0021	**3.935e-04**	15.1372	**0.3618**
	p-value	(0.0176)	(0.6861)	(0.0000)	(0.0000)	**(0.0000)**	(0.0000)	**(0.0000)**
600048	Coefficient	-3.076e-04	-1.128e-07	-0.1182	0.0012	**9.242e-05**	14.3525	**0.1091**
	p-value	(0.0000)	(0.8182)	(0.0000)	(0.0000)	**(0.0021)**	(0.0000)	**(0.0000)**
600221	Coefficient	-4.484e-05	3.089e-08	-0.3872	0.0025	1.140e-04	13.9949	**0.0860**
	p-value	(0.5454)	(0.9250)	(0.0000)	(0.0000)	(0.1503)	(0.0000)	**(0.0000)**
600688	Coefficient	-3.206e-04	1.108e-06	-0.1982	0.0020	**2.051e-04**	14.0461	**0.8334**
	p-value	(0.0000)	(0.7493)	(0.0000)	(0.0000)	**(0.0082)**	(0.0000)	**(0.0000)**
601018	Coefficient	-1.523e-04	-2.339e-06	-0.2892	0.0022	**5.422e-05**	13.3157	**0.4356**
	p-value	(0.0285)	(0.1158)	(0.0000)	(0.0000)	**(0.0000)**	(0.0000)	**(0.0000)**
601989	Coefficient	-2.962e-04	8.288e-07	-0.1616	0.0015	**2.320e-04**	14.8642	**0.3992**
	p-value	(0.0000)	(0.9278)	(0.0000)	(0.0000)	**(0.0021)**	(0.0000)	**(0.0000)**

$$vol_{q,T} = \mu''_{0,q} + \mu''_{1,q} dsa_{q,T} + err'' \quad (19)$$

where $vol_{q,T}$ is the volume of stock q at time T; $\mu''_{0,q}$ is a constant term and $\mu''_{1,q}$ is the coefficient between $vol_{q,T}$ and $dsa_{q,T}$; $err'' \sim i.i.d$. We expect a positive relationship to be found between the volume and DSA.

The DSA is computed based on both the market-based approach and the stock-based approach. All series are confirmed to be stationary based on the KPSS test. For the estimators in above three models, we conduct student's t test to verify the significance. A p-value of 0.05 indicates a 95 percent confidence level.

C. Result Anlaysis

TABLE 4 reports the results of three testing models based on the DSA computed by the market-based approach. Several empirical findings can be concluded from the results. First, among the 10 stocks in sample, no relationship between the stock return and DSA can be found based on the results of model (i). None of the p-values of DSA coefficients is less than 0.1. The result is consistent with our ***Hypothesis 1***. Second, positive relationship between the absolute value of abnormal return and DSA is identified for 8 stocks in 10 based on model (ii). The two exceptions are stock 000709 and stock 600221. The first one has a negative DSA coefficient which is meaningless in terms of economic implications, so the insignificant result (p-value = 0.8177) is still as expected; the second one has a positive coefficient with p-value = 0.1503, which indicates an 85% confidence level. Another interesting finding is that the two stocks happen to be the two anomalies in terms of the distributions of WAI (see **Fig 5)**. Therefore, the results are in favor of ***Hypothesis 2***. Third, the results of model (iii) show significant relationship between the trading volume and DSA for all stocks in sample.

TABLE 5 reports the results of same models based on the DSA computed by the stock-based approach. The results are consistent with the conclusions we make for the market-based approach. First, ***Hypothesis 1*** is supported by the results of model (i), in which there is no evidence to show the relationship between the stock return and DSA. Second, we find significant evidence to support the positive relationship between magnitude of the abnormal return and DSA for 8 of 10 stocks in the sample. Also, stock 000709 and stock 600221 are still the two exceptions. Hence, ***Hypothesis 2*** is supported.

In summary, the effectiveness of our DSA framework is verified for both of the market-based approach and the stock-based approach. We conclude that the DSA is significantly correlated to the absolute abnormal return and trading volume, while it does not directly affect the stock return. The testing results suggest that DSA serves as an important factor to link social media activities and the stock market. It will contribute on research and practice in finance, such as price forecasting, risk management, and other asset pricing problems.

D. Computational Efficiency

Now we study the computational efficiency of our DSA framework. The major computational cost of the framework is the dynamic influence modeling process, so we compare the performance of the market-based approach and the stock-based approach with a benchmark algorithm that considered

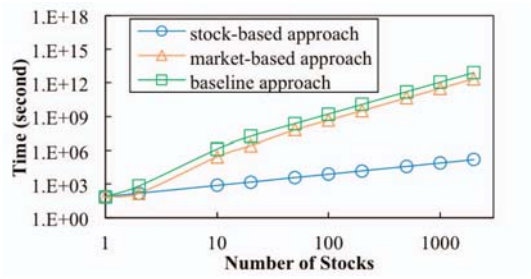


Fig 6. Computational Costs of Three Approaches: Solving One Influence Matrix with Different Stock Number

TABLE 5. SIGNIFICANCE TEST BASED ON STOCK-BASED APPROACH

Stock Code		(i) Return ($R_{q,T}$)			(ii) Abnormal Return ($\|AR_{q,T}\|$)		(iii) Logarithmic Volume ($vol_{q,T}$)	
		(Intercept)	$dsa_{q,T}$	$R_{q,T-1}$	(Intercept)	$dsa_{q,T}$	(Intercept)	$dsa_{q,T}$
000100	Coefficient	-1.015e-04	-9.082e-06	-0.2388	0.0024	**4.666e-05**	15.5338	**0.1647**
	p-value	(0.1781)	(0.2534)	(0.0000)	(0.0000)	**(0.0902)**	(0.0000)	**(0.0000)**
000157	Coefficient	-2.038e-04	-9.048e-07	-0.1930	0.0012	**6.252e-05**	13.8214	**0.1201**
	p-value	(0.0000)	(0.3960)	(0.0000)	(0.0000)	**(0.0157)**	(0.0000)	**(0.0000)**
000709	Coefficient	-9.334e-05	1.804e-05	-0.3352	0.0028	-1.102e-05	13.6856	**0.4359**
	p-value	(0.2434)	(0.6999)	(0.0000)	(0.0000)	(0.9478)	(0.0000)	**(0.0000)**
000783	Coefficient	-3.065e-04	-1.535e-05	-0.1124	0.0014	**1.541e-04**	13.8029	**0.2365**
	p-value	(0.0000)	(0.1407)	(0.0000)	(0.0000)	**(0.0012)**	(0.0000)	**(0.0000)**
002024	Coefficient	-2.636e-04	3.426e-06	-0.1417	0.0021	**3.897e-04**	15.1363	**0.3612**
	p-value	(0.0148)	(0.7118)	(0.0000)	(0.0000)	**(0.0000)**	(0.0000)	**(0.0000)**
600048	Coefficient	-3.155e-04	1.326e-07	-0.1183	0.0012	**9.623e-05**	14.3507	**0.1137**
	p-value	(0.0000)	(0.8556)	(0.0000)	(0.0000)	**(0.0013)**	(0.0000)	**(0.0000)**
600221	Coefficient	-3.950e-05	-1.010e-07	-0.3872	0.0025	9.273e-05	13.9930	**0.0715**
	p-value	(0.5872)	(0.8567)	(0.0000)	(0.0000)	(0.2179)	(0.0000)	**(0.0036)**
600688	Coefficient	-3.180	4.848e-07	-0.1983	0.0020	**2.222e-04**	14.0445	**0.8258**
	p-value	(0.0000)	(0.9015)	(0.0000)	(0.0000)	**(0.0039)**	(0.0000)	**(0.0000)**
601018	Coefficient	-1.511e-04	-2.475e-06	-0.2894	0.0022	**5.576e-05**	13.3156	**0.4406**
	p-value	(0.0298)	(0.9770)	(0.0000)	(0.0000)	**(0.0000)**	(0.0000)	**(0.0000)**
601989	Coefficient	-2.756e-04	-3.875e-05	-0.1602	0.0015	**2.457e-04**	14.8640	**0.4055**
	p-value	(0.0000)	(0.2574)	(0.0000)	(0.0000)	**(0.0011)**	(0.0000)	**(0.0000)**

all social media users as a single group. The experiments are based on synthetic data for larger sample size. According to the statistics of our 6-month social media data that covers 100% user information of the stocks in sample, the average number of users for one stock is 2,177.31. We assign 2,000 social media users for each stock to run the tests. The propagation probabilities for each pair of users are randomly assigned. Currently there are 2,314 stocks trading in the Chinese stock market, and we set the maximum stock number as 2,000 for the experiments. **Fig 6** plots the average computational time for the three approaches based on 10 fair experiments. As we can see, the stock-based approach performs the best among the three approaches. A slight improvement is also found for the market-based approach. Furthermore, the stock-based approach creates opportunities of parallel computing. **Fig 7** plots the speedup for different matrix dimension with parallel computing. With more cores, larger speedup and better computational efficiency can be obtained. For instance, for 4 million nodes, we can get 5.52 times speedup by using 8 cores, and the speedup achieve 8.14 times by using 16 cores. For our algorithm, the speedup is very efficient due to small communication cost among cores. One issue is the reduction of speedup performance when increasing the number of nodes. For example, when we handle 400k nodes with 16 cores, we get 10.29 times speedup, while handling 4 million nodes reduces the speedup to 8.14 times. However, the total time consuming in our analysis is controlled in a tolerable level with good scalability using high-performance computing methods.

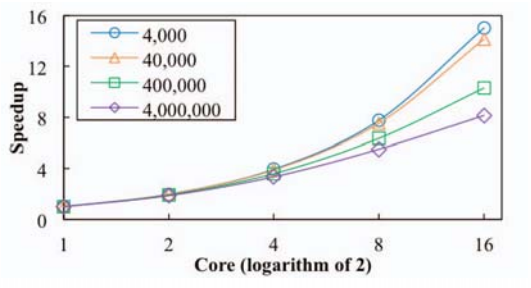


Fig 7. Speedup Results for Four Matrix Dimension (4k, 40k, 400k, 4m Nodes) with Different Computing Cores

VI. RELATED WORK

Related work can be generally classified into two groups. The first group of studies focuses on identifying the relationship between social media and the stock market. Reference [14] investigated the relationship between tweet sentiment and stock returns, message volume and trading volume, and disagreement and volatility. Reference [15] discovered a positive relationship has been discovered between disagreement on stock-related articles and the trading volume. Online forums reflect the major activities of uninformed traders, but not informed traders such as institutional investors [16]. Reference [17] claimed that, in terms of the online attention about thinly traded microcap stocks, positive abnormal returns are most likely to be associated with the stocks with the most discussions. Reference [18] showed that noise trading is highly correlated to stock price volatility, while the effect of reverse causation is even stronger. Reference [19] suggested that the actions of retweets and mentions is not solely triggered by followership. Hence, social influence cannot simply measured by popularity. Reference [20] claimed that it is a common case that users keep retweeting valuable messages in order to validate nice contents or friend users. Reference [21] believed that there is a tendency that social media users would be attracted by the Internet stock messages with less noise and well processed contents.

The second group of papers builds forecasting models based on social media data. Reference [22] proposed a forecasting model which investigates how other financial markets would affect Russian market. An important variable

has been defined in this model: positive or negative news in the past. Although few studies consider the social influence model in stock analysis, social media data has already be used in market stock market prediction. Reference [23] presented a model to predict market behavior based on public event related to target companies. Reference [24] proposed a framework to learn association between news and the stock reactions. Reference [25] established a stock prediction model using blog content.

Our study is in the first group. While most of the previous work merely considers the frequency of discussion, our study is based on dynamic social influence that better explains the real impact of a discussion on the stock market. Also, our hypotheses and experiments are made for high-frequency financial data which is seldom studied before.

VII. Conclusion

In this paper, we propose a Degree of Social Attention (DSA) framework for stock analysis. We improve the existing influence models by (1) mathematically defining the self-influence (confidence) and (2) designing algorithms to capture dynamic influence in a social network with the consideration of computational efficiency. We further define the DSA based on the proposed dynamic influence process. By testing our two hypotheses formulized under the assumptions of semi-strong inefficiency of the Chinese stock market, we verify the effectiveness of our framework with both the market-based approach and the stock-based approach. We find positive relationship between the absolute value of abnormal return and DSA. We also verify the market volume is associated with DSA at most of time. On the other hand, the results can serve as new evidence to support the semi-strong market inefficiency in the Chinese stock market, and also show the significance of DSA as an important factor to link social media activities to the stock market. In addition, we confirm that the stock-based approach for dynamic influence modeling has the best performance for its computational efficiency.

Acknowledgement

This research was partially supported by the Funds for the joint innovation center of finance and regional development (a key project of social science for Shandong, Grant No. 14AWTJ01-14), the Natural Science Foundation of China (Grant No. 61403358) and the Fundamental Research Funds for the Central Universities of China (Grant No. WK0110000042).

References

[1] E. F. Fama, "Efficient capital markets: A review of theory and empirical work*," *The journal of Finance,* vol. 25, pp. 383-417, 1970.

[2] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of Finance,* vol. 48, pp. 65-91, 1993.

[3] N. Jegadeesh and S. Titman, "Profitability of momentum strategies: An evaluation of alternative explanations," *The Journal of Finance,* vol. 56, pp. 699-720, 2001.

[4] E. F. Fama and K. R. French, "Dissecting anomalies," *The Journal of Finance,* vol. 63, pp. 1653-1678, 2008.

[5] A. Agrawal, J. F. Jaffe, and G. N. Mandelker, "The post-merger performance of acquiring firms: a re-examination of an anomaly," *The Journal of Finance,* vol. 47, pp. 1605-1621, 1992.

[6] R. Ball and P. Brown, "An empirical evaluation of accounting income numbers," *Journal of accounting research,* pp. 159-178, 1968.

[7] P. L. Davies and M. Canes, "Stock prices and the publication of second-hand information," *Journal of Business,* pp. 43-56, 1978.

[8] A. Brav, C. Geczy, and P. A. Gompers, "Is the abnormal return following equity issuances anomalous?," *Journal of Financial Economics,* vol. 56, pp. 209-249, 2000.

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137-146.

[10] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, "Pagerank with priors: An influence propagation perspective," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 2740-2746.

[11] A. S. Kyle, "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society,* pp. 1315-1335, 1985.

[12] S. Ma, *The efficiency of China's stock market*: Ashgate Aldershot, 2004.

[13] M. Sosonkina, L. T. Watson, and R. K. Kapania, "A new adaptive GMRES algorithm for achieving high accuracy," 1996.

[14] T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welpe, "Tweets and Trades: the Information Content of Stock Microblogs," *European Financial Management,* vol. 20, pp. 926-957, 2014.

[15] W. a. F. Antweiler, M.Z., "Is all that talk just noise? The information content of internet stock message boards," *Journal of Finance,* vol. 59, pp. 1259-94, 2004.

[16] S. Das, Martinez-Jerez, A. and Tufano, P., "eInformation: a clinical study of investor discussion and sentiment," *Financial Management,* vol. 34, pp. 103-37, 2005.

[17] S. Sabherwal, Sarkar, S.K. and Zhang, Y., "Online talk: does it matter?," *Managerial Finance,* vol. 34, pp. 423-36, 2008.

[18] J. L. Koski, Rice, E.M. and Tarhouni, A., "Noise trading and volatility: Evidence from day trading and message boards," 2004.

[19] M. Cha, Haddadi, H., Benevenuto F. and Gummadi K.P., "Measuring user influence in Twitter: The million follower fallacy," in *the International Conference on Weblogs and Social Media*, 2010, pp. 10-17.

[20] D. Boyd, Golder, S. and Lotan, G., "Tweet, tweet, retweet: conversational aspects of retweeting on Twitter," in *the 43rd Hawaii International Conference on System Sciences*, 2010, pp. 1-10.

[21] B. Gu, Konana, P. and Chen, H-W., "Melting-pot or homophily? An empirical investigation of user interactions in virtual investment-related communities," 2008.

[22] B. H. a. A. M. Kutan., "The impact of news, oil prices, and global market developments on russian financial markets.," *The Economics of Transition,* vol. 13, pp. 373-393, 2005.

[23] M. S. V. Lavrenko, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan., "Mining of concurrent text and time series.," in *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, 2000, pp. 37-44.

[24] R. P. S. a. H. Chen., "Textual analysis of stock market prediction using breaking financial news: The azfin text system. ," *ACM Transaction Information Systems,* vol. 27:12:1-12:19, 2009.

[25] M. H. M. J. Ginsberg, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant., "Detecting influenza epidemics using search engine query data.," *Nature,* vol. 457, pp. 1012-1014, 2009.