Mining Influence in Evolving Entities: A Study on Stock Market

Chang Liao[*†], Yinfei Huang[‡], Xibin Shi[*†], Xin Jin[‡]
[*]School of Computer Science, Fudan University, Shanghai, China
[†]Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
[‡]Shanghai Stock Exchange, Shanghai, China
Email: cliao12@fudan.edu.cn, yfhuang@sse.com.cn, xbshi@fudan.edu.cn, xjin@sse.com.cn

***Abstract*—Mining influence in evolving entities is an important but challenging task, partly due to complex nature of it. In this paper, we mainly focus on the following problems on it with respect to stock market: (1) How to identify pairs of stocks that influence one another; (2) How to quantify the influence and capture group effects and dynamic nature of influence of each stock; (3) How to adopt approximate approaches so that we can improve the efficiency of the proposed model. To tackle these problems, a novel graph-based mining method, which utilizes time series and volume information collaboratively is proposed, and several optimized algorithms are presented. Besides, two extended metrics to capture the dynamic and group nature of influence based on the model are derived. Furthermore, we also suggest a potential application of the model to stock price prediction. The experimental results on both synthetic and real data sets verify the effectiveness and efficiency of our approach. Some insights on this paper can be the ideas of analyzing the influence of evolving entities using the social network analysis methods.**

I. Introduction

In an interconnected and dynamic changing world, it is well recognized that influence is a complex force that governs the dynamics and behaviors of involved ones. With the power of influence, the evolution of one entity can have a significant impact on the others. Usually, mining influence can be of great value to detect and control risks. It is especially applicable in the financial markets [1]. The bankruptcy of $Lehman\ Brothers$ caused by the US. $Subprime\ Mortgage\ Crisis$, and Portugal and Greeces sovereign debt risks due to the global financial market turmoil can be very good examples.

This paper aims at addressing the evolving entity influence data-mining problem, that is, how to effectively and efficiently discover and quantify influence of evolving entities based on their time series and volume. To achieve this, we first quantify the influence between them through time series and then analyze the relationship among entities by constructing a "entity social network". The entities and their relationships have been modeled as a weighted directed graph.

With the general problem setting, at least two challenges need to be addressed. One challenge lies in modeling influence in evolving entities. How to measure influence in entities is non-trivial, for influence strength between time series is hard to define and the dynamic and group nature is difficult to capture. The other is to adopt approximate approaches so that it can scale to real applications. Due to its computation intractability, finding the results efficiently is a challenge.

Recent works [2], [3], [4] have focused part of the problems addressed above, but the main drawbacks are the following:

1) These works ignore the popularity characteristics difference among each entity and assume it the same. However, entity with more volume often indicates higher influence status. What's more, influence is not just equal to similarity between entities. But these works have not distinguished them well.
2) These approaches do not support dynamically evolving entity influence quantification and have not taken group effect into account. But in real scenarios, group-orientation influence often exists, for instance, a movement in the stock markets can be always caused by some cohesive stocks, not individuals. And again, influences in entities are not stable and vary with the dynamic and volatile situations.
3) The process of influential entity discovery is always too costly and time consuming. When the proposed model is applied to real applications, like in stock market, consideration on how to make this process practical from running time perspective is a must.

In this paper we design several steps to address the above challenges and propose solutions to the problem of mining influence in evolving entities. In particular, we shall describe in detail the applications to stock market in this study. However, algorithms proposed can be generally applied to domains of traffic routing mechanism [5], climate change [6], etc.

In stock market, algorithms to combine and select various factors for measuring entity influence are devised [7], [8]. Nevertheless, these approaches ignore the causality among entities and still unable to obtain reasonable result [9]. With a different purpose from this paper, we are trying to identify leadership relation between evolving entities from time series data and formulate new indicators to represent the influence characteristics.

The main contributions are summarized as follows. We proposed a framework to measure stock's influence from a network perspectives that exploits both its intrinsic and causal property. Two extended types of influence are defined to capture its dynamic nature and group effect. Meanwhile, two optimized algorithms are introduced to improve the efficiency. The results of experiments on both synthetic and real data demonstrate our approach can outperform baseline approaches and its potential applications to price prediction.

II. Problem Definitions

In this section, we introduce several concepts and then formally formulate the problem.

*Definition 1 (**Influence Network**):* Define influence network $G = \{V, E, P\}$. V represents the set of evolving entities. $E \in R^{|V|\times|V|}$ is called the diffusion matrix such that $E(i, j) = e$ is the influence from node V_i to V_j and the value determines the tendency strength of adoption of influence. P is a popularity matrix, where P_{ii} is the relative influence strength that accounts for the ability to spread the influence.

An example is shown in Fig. 1: the number in each bracket () means popularity and the edge weight acts as the diffusion probability. $(E \rightarrow B \rightarrow A)$ act as an influence propagation path from E to A within its length is 2.

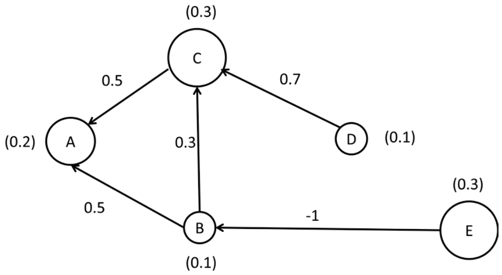


Fig. 1: Influence Network

*Definition 2 (**Global Influence**):* Given influence network, $f_{i\rightarrow V}$ is defined as the overall influence of evolving entity S_i, which represents the global influence strength on entity S_i over the whole network.

Along this line, two characteristics are focused to measure influence of entity. Firstly, it should be popular, which means it should be generally of good liquidity to spread influence. Secondly, it should be strongly correlated so as to be adopted by others. To be specific, it should have a high ability to spread the influence and ensure influence to be adopted. One thing to note is that, influence definition here is different from other types of networked influence, such as that in social network [10], which is quantified by combinations of various metrics.

*Definition 3 (**Evolving Entity Influence Mining**):* Given a set of entity series and volumes, the objective is to effectively and efficiently (1)quantify each entity's influence $f_{i\rightarrow V}$, (2)find top influential entity S_i, influential entity groups g and dynamic influential entity S_d.

This is the formal problem definition of this paper. Since previous models are not directly applicable, a framework that exploits both time series and fluctuation volume information is proposed. In particular, we focus influence mining problem on stock domain, and the so-called evolving entities is another saying of stocks in the rest of paper.

III. Stock Influence Mining Model

To address the stock influence mining task, we first have the following two assumptions to capture the influence.

Assumption 1: Stocks with high volume are expected to have high probability to spread influence to others. That is to say, the stocks with more occurrence are generally of good liquidity so as to have a potential high rate to spread influence.

Assumption 2: Stocks with temporal correlations are expected to have certain causal influence relations. In other words, some correlation can lead us to induce, under some conditions, the existence of causal relation. This assumption is widely used in economic time series analysis [9].

Based on these two assumptions, we construct the entity influence network based on the time series and volume. And then we propose a novel graph-based framework to calculate the overall influence of each stock by utilizing an influence propagation process. After that, two extended types of influence are derived.

A. Network Construction

To construct influence network, popularity matrix and diffusion matrix that model intrinsic and causal influence properties among stocks should both be formed. To be specific, popularity matrix models the ability to spread the influence while diffusion matrix imitates the propagation probability.

Popularity matrix Volume is taken into account in intrinsic influence property modeling. The influential stocks should be of good liquidity, which also implies large fluctuation volume on the stock. Given the volume $\{w_i(t)\}$ of stock S_i transactions, intrinsic influence strength of it at given time interval L is defined in Eq. 1 and the popularity matrix P can be defined as Eq. 2.

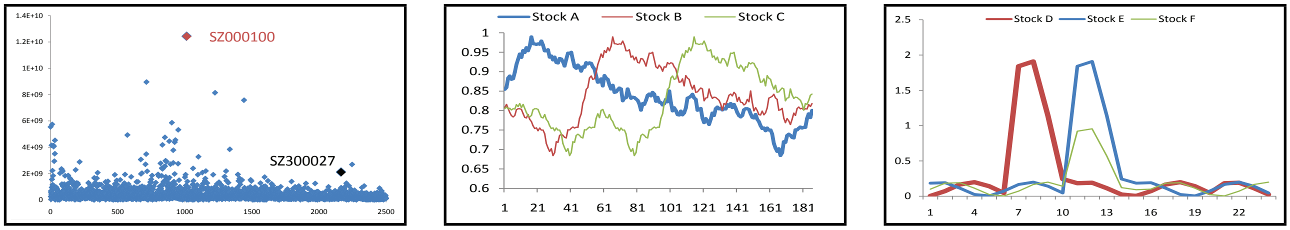
$$w_i = \sum\nolimits_{t=0}^{L} w_i(t) \tag{1}$$

$$P(i,j)=\begin{cases} \frac{w_i}{\sum_V w_i}, & j = i \\ 0, & otherwise \end{cases} \tag{2}$$

An illustration can be seen in Fig. 2(a), which describes the distribution of trading volumes among the 2516 stocks. TCL Group has the highest trading volume, while $Huayi$ $shares$(300027) has a related small one. In this case, TCL $Group$(000100) generally has a higher probability to spread the influence, which should be put more weight on. And our model tries to capture this feature. One thing to note here is that we just consider stocks whose trading volume are nearly the same magnitude , while letting alone extreme ones.

Diffusion matrix Time series is considered in causal influence modeling and the polynomial function $f_i(t) = \sum_{t=0}^{L} k_r t^r$ is denoted to fit the series of stock S_i. The shape dissimilarity $\xi_{i,j}$, strength attenuation h' and time decay t' between series are used to described the influence between trigger and feedback of the entity pair, corresponding to the operations of amplitude transformation, amplitude scaling and time shifting of polynomial fitting. The relationships between them can be depicted in Eq. 3, and conjugate gradient descent methods [11] are adopted for optimal solutions. Upon these three factors, the diffusion matrix E is defined as Eq. 4, where α, β are weight parameters and σ is normalization term. One constraint is that $-\varrho < t' < \varrho$, where ϱ is predefined according to different requirements in real applications.

$$\min_{h',u',t',s'} \xi_{i,j} = \sum\nolimits_{t=0}^{L} (f_i(t) - (h' f_j(u't + t') + s'))^2 \tag{3}$$



(a) Trade Volumes (b) Time Decay (c) Strength Attenuation

Fig. 2: Illustration Examples

$$E(i,j)=\begin{cases} \frac{h'}{|h'|}e^{-\frac{\xi_{i,j}+\alpha h'+\beta t'}{\sigma}}, & t'>0 \\ 0, \quad otherwise & \end{cases} \tag{4}$$

The rationale behind Eq. 4 can be seen in Fig. 2(b) and Fig. 2(c). We infer causal relation exist among these stocks, because they exhibit similar trends. In Fig. 2(b), stock S_A has stronger influence on S_B than on S_C, while in Fig. 2(c), stock S_D has higher impact on S_E than on S_F, which is the result of time decay and strength attenuation accordingly. To conclude, our model aims at precisely modeling the causal influence by taking these factors into account.

B. Influence Quantification

Given two node S_i and S_j in influence network, we denote the indirect influence from S_i to S_j as the impact of S_i propagates to S_j within steps .

At first, we model the influence network as a markov chain, where indirect influence can be computed by an aggregate result of the direct mutual influences between direct connected ones. Then the global influence of the stock S_i is to be measured by firstly enumerating all the influence paths from S_i to others, secondly calculating the influence value of each path, and thirdly combing influence strength of all paths. Considering the decreasing property along the propagation process, we put $\lambda(k) = \frac{1}{(1+\lambda)^k}$ to describe the retained ratio of influence when the the length of propagation path is k.

Given a scalar λ and a network G, we define the matrix M as follows where K can be infinite. But in reality, the maximum value of K is set as a small value for simplicity.

$$M(E,\lambda) = [I - \frac{1}{1+\lambda}E]^{-1} = \sum_{k=0}^{K} \frac{1}{(1+\lambda)^k}E^k \tag{5}$$

Consider a network G with adjacency n-square matrix E and a scala λ, The vector of influence of stock S_i to others within k-distance is:

$$f_i = (M(E,\lambda))^T P_{,i} \tag{6}$$

Finally, the overall influence of stock S_i can be computed by summing influence value from its popularity strength by self-loops and diffusion strength by out-paths:

$$f_{i->V} = f_{ii} + \sum_{j \neq i} f_{ij} \tag{7}$$

C. Extended types of influence

Considering that the influence between stocks can be group oriented and dynamic, two extended types are derived.

For instance, group influence often exists in stock market, for instance, a movement in the stock markets can be always caused by some cohesive stocks, instead of individuals. A good example is sector rotation. For instance, the prosperity of steel or related infrastructure sectors leads the rise of the consumption and culture-related industries. To begin with, a set of nodes is defined as groups if they are similar via some metrics or domain knowledge.

*Definition 4 (**Group Influence**):* For group influence, we view the group as a big node, and thus we can measure the influence of group g as follows, where V^* means the nodes that do not exist in group g.

$$f_{g \to V} = \sum_{i \in g} f_{i \to V^*} \tag{8}$$

Another consideration is about dynamic. Previous assumptions of the influence of stock is static, however, this is not always true, especially when we study volatile or long-term time series. Influence strength can be amplified or weakened as time evolves.

*Definition 5 (**Dynamic Influence**):* Considering the fact that usually the recent influence relations are more important, and they should receive higher score. The dynamic influence in stock S_i can be derived as Eq. 9, where $\tau = 1, 2, ..., T$, representing the time interval over time T.

$$f_{i \to V} = \frac{1}{\sum_{\tau=1}^{T} \exp(\tau - T)} \sum_{\tau=1}^{T} \exp(\tau - T) \cdot f_{i \to V}{}^{(\tau)} \tag{9}$$

D. Algorithm Complexity Analysis

The complete algorithm is presented in Algorithm 1, which contains three main parts: influence network construction, influence quantification and influence ranking. The algorithm complexity of network construction is $O(|V|^2)$, where the most expensive time consuming process is to calculate the matrix in Stage 2, with the minimum complexity $O(|V|^{2.4})$, while the most optimal complexity of the third part is $O(|V|log|V|)$, with the divide-and-conquer strategy. One thing to mention is that in dynamic ranking process, we should first split the time interval into segments according to the cycle of time series, which is usually predefined, and loop the algorithm for each segment.

Algorithm 1 Stock Influence Mining Framework.

```
Input:
    $Dataset$ is entity time series,$\lambda$ is damping parameter,$K$ is propagation path length
Output:
    Influence Value and Rank in each node $f_{i\to V}$, $R_{i\to V}$;
//Stage 1: Construct Influence Network
for every entity pair $(i,j)$ do
    Determine $P(i,j)$ using Eq. 2
    Determine $E(i,j)$ using Eq. 4
end for
//Stage 2: Quantify the influence of each stock
Calculate the stock influence $f_{i\to V}$ using Eq. 7
Determine the influential groups and etc. using Eq. 8 and Eq. 9
//Stage 3: Rank the stocks According to the influence value
```

IV. Optimizations Strategies In Stock Market

The high computational complexity $O(|V|^{2.4})$ makes the design of efficient solutions of the model difficult. However, there are some heuristics we can make use of when applied to stock market. In this section, two optimization strategies based on characteristics on the network structure of stock domain are proposed.

A. Strategy 1: Stock Influence Approximation

In real applications, how to quickly estimate influence of entities without compromising much accuracy means more than compute exact influence in them in a relative much longer time. The problem then turns to how to design an efficient algorithm that can quickly estimate influence of each entity. In stock market, the influence between most of stocks is not strong, which can be ignored by a threshold. In other words, a common stock always has limited direct influence relation. Inspired by this, we no longer use the adjacency matrix but the adjacency list to store the graph and use lemma 1 to reduce the influence computation cost of stocks.

*Lemma 1 (**Shared segments**):* Random walk paths generated by different nodes contain the shared segments, which can be reused to save computational cost. For example, influence propagation along path $\{a,b,c\}$ and $\{d,b,c\}$ are both computed on influence quantifications for node a and d, but they contain the same segment $\{b,c\}$.

Algorithm 2 Stock Influence Approximation.

```
Input:
    $G=(V,E,P)$ is stock influence network, $\lambda$ is damping parameter, $Q$ is adjacent list, $K$ is propagation length
Output:
    Influence in each node $f_{i\to V}$;
initialize $f_{i\to V} = P_{ii}$
for every node $i \in$ V do
    PathRecursion($i$,0,$\lambda$)
end for
return  Influence in each node $f_{i\to V}$;
Procedure PathRecursion($i$,$k$,$\lambda$)
$k = k+1$
for every node u in in-neighbor set $Q_i$
    $f_{i,\cup k} = f_{i,\cup k-1} + P_{ii} \cdot Q_{i,u} \cdot \frac{1}{1+\lambda} \cdot f_{u,\cup k-1}$
    if $k < K$ then
        PathRecursion($i$,$k$,$\lambda$)
    end if
end for
```

With Lemma 1 in hand, we adopt a memory-based search method to optimize the algorithm. For each vertex, we set an initial disturbance on it and calculate the influence strength through depth-first search method. We store each path-based influence value and iterate afterwards. Regarding sparsity in influence among stocks, we can reduce the time cost to $O(K\cdot|E|)$, where $|E|$ is relatively small. The main process is described as Algorithm 2.

B. Strategy 2: Top-Z Influential Stocks Discovery

In practice, we are usually more interested in finding Top-Z influential entities while ignoring others. For instance, in terms of regulatory oversight, top few influential stocks can mean a lot, also known as $Pareto\ rule$. In real stock market, there is a dramatic distinction between the influence of stocks. That is, some are of key status, while others are of no impacts. Thus in this case, instead of computing exact influence value of each entity and then ranking it, we turn to design an algorithm that can quickly estimate the upper bound of each entity's influence and filter out insignificant ones conversely. To tackle this problem, we first find out the upper bound of influence of each stock according to lemma 2 and then use it to speed up the algorithm.

*Lemma 2 (**Upper bound**):* Given an influence network, node V_i's total influence value should not be larger than $(1+\frac{1}{\lambda})p_i e_i$, where e_i and p_i is the sum of $E(,i)$ and $P(i,)$.

Proof. Given the decay function $\lambda(k) = \frac{1}{(1+\lambda)^k}$, it can be proved in the following way: The value of all the entries in matrix $(E^T)^k$ is smaller than 1 and the sum value of each column is 1, so we can transform Eq. 5 and conclude $f_{i,k} \le \frac{1}{(1+\lambda)^k} E^T P_{i,}$, followed by $f_{i->V} \le (1+\frac{1}{1+\lambda}+...+\frac{1}{(1+\lambda)^K})p_i e_i$ of Eq. 6. While $1+\frac{1}{1+\lambda}+...+\frac{1}{(1+\lambda)^K} \le (1+\frac{1}{\lambda})$, it comes to be true that $f_{i->V} \le (1+\frac{1}{\lambda})p_i e_i$.

Algorithm 3 Top-Z Influential Stocks Discovery.

```
Input:
    $G=(V,E,P)$ is stock influence network, $\lambda$ is damping parameter, $Z$ is defined number, $K$ is propagation length
Output:
    Top-Z influential stocks
for every node i do
    $U_i = (1+\frac{1}{\lambda})p_i e_i$
    $IsBound_i = True$
end for
while $|S| < Z$ do
    //Find node d with biggest $U_d$ in $U$
    if $IsBound_d == True$ then
        Compute $U_d = f_{d\to V}$
        $IsBound_d = False$
    else
        $S = S \cup d$
    end
end
return S
```

For finding the Top-Z influential stocks, traditional method first computes $|V|$ stocks' influence and then ranks them by their values. By contrast, we first compute each stock influence upper bound, and then use only a small $|N|$ stocks to capture the top influential ones. Thus, the cost of this task changes from $|V|$ times influence computation to ($|N|$+1) times. The main process is described as Algorithm 3.

V. Experiments

In this section, we present various experiments to evaluate the efficiency and effectiveness of the proposed approach in both simulated and real data derived from stock market .

A. Experimental Setup

Three data sets are experimented, the first two were crawled from [12], and the third is generated synthetically:

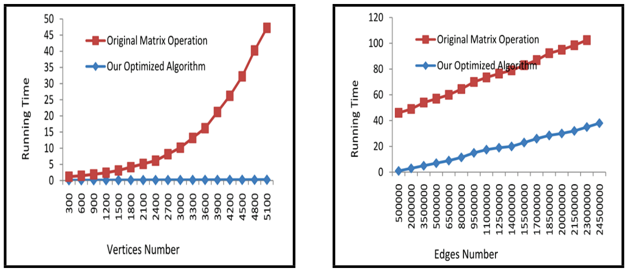
- **Stock historical price data:** It contains reinstated closing prices and trading volumes ranging in the period of year 2008 to 2012 (1219 business days) of 1346 stocks in Chinese stock market. Others are ignored for the sake of incomplete information, such as new entrants and etc.
- **Stock financial report data:** It includes reports on the corresponding 1346 stocks quarterly and annual financial statistics ranging from year 2008 to 2012.
- **Synthetic data:** The data set includes edge weights and vertices. The value of links are picked from a beta distribution, given the vertices and edge number.

We evaluate our method in the following four aspects:

- **Accuracy on the model:** We compare the proposed model against the pagerank algorithm [2], which treats the initial importance of stocks the same, for the task of stock's influence quantification and top influential ones discovery. The xiaoxiao's algorithm [4], which infers causality and detects shaker from evolving entities, acts as another baseline that we used for evaluating the performance of the model.
- **Running time on the model:** we compare the time cost of two optimized algorithms with that of algorithm 1 using both real and synthetic data.
- **Extended influence Types:** We combine domain knowledge to demonstrate how effectively our method can identify the group-oriented and dynamic influence.
- **Price Prediction Case Study:** We suggest one potential application case of our proposed model. In particular, we apply it to price prediction of $CSI800$.

B. Evaluation on the accuracy of the model

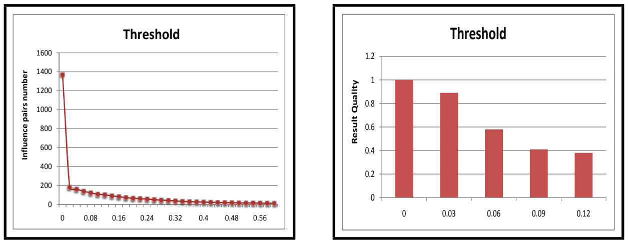
For accuracy analysis, we take the ranking result of the stocks according to their financial statements as the ground truth. That is, we rank the stock via the average ranking result of various criteria, such as assets, profit, share, momentum, etc. For instance, in the first quarter of 2012, $ICBC$ (601398) ranked first according to its total profits and also first in the number of total share; thus the overall ranking of $ICBC$ (601398) is $(1+1)/2 = 1$ considering these two factors. We compare our proposed model with Xiaoxiao's method [4] and Pagerank method [2]. Accuracy metric $\frac{|\overrightarrow{r} \cap \overrightarrow{t}|}{5}$ is defined, where the top 5 of the result from the model is taken as $\overrightarrow{r}$, and top 10% stocks from the ground truth as $\overrightarrow{t}$.



(a) Evaluation on vertices

(b) Evaluation on edges

Fig. 4: Comparison on Running Time Cost



(a) Influence Pairs Number

(b) Result Quality

Fig. 5: Evaluation on Coal Industry Over Different Threshold

The data set between April 2nd and June 29th in 2012 in Chinese stock market are used to test the effectiveness of the model. The accuracy across different industries is displayed in Fig. 3. Clearly, the proposed model outperforms baseline methods. To explain, Xiaoxiao's method performs quite well when just considering causality relationship between stocks, while popularity characteristics counts when considering influential stocks among all.

C. Evaluation on the time cost of the algorithms

For time cost analysis, we evaluate the scalability of the two proposed algorithms compared with algorithm 1.

To test the performance of Algorithm 2, we evaluate the performance of the scalability of it compared with the original algorithm 1 by using the synthetic data. The test results are shown in Fig. 4. From Fig. 4(a), the running time of algorithm 1 quickly exhibits sharp growth curve while algorithm 2 grows little. The same result is in Fig. 4(b). Let P_1 be the approximate result of stock influence ranking list by algorithm 2 and P_2 be the result of that by original algorithm 1. Then we define $sim(P_1, P_2) = kendall(P_1, P_2)$ to measure the quality of approximation algorithms under different threshold value. Clearly, we have $0 \leq sim(P_1, P_2) \leq 1$. We ran the algorithm in coal industry data between April 2nd and June 29th in 2012 in Chinese Stock market. The results are shown in figure 5. From figure 5(a), we can see most influence strength between stocks are not strong, which can thus be pruned to reduce computation cost. In figure 5(b), we can see the result is of high quality when the threshold value is smaller than 0.03. And thus, in many applications, when the accuracy is not a strict requirement, our approximate algorithm is quite suitable.

To evaluate the performance of Algorithm 3, we test the search number for top Z influential ones discovery of the 1346 real stock data. The result is presented in Fig. 6. From it, we can observe that the search number is quite small with respect to the entire search space(1346). Obviously, the computation

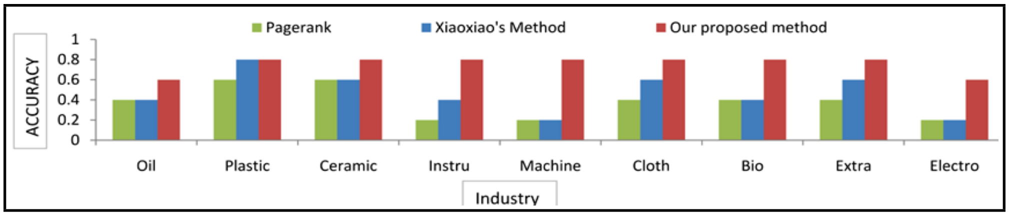


Fig. 3: Rank Accuracies according to industries

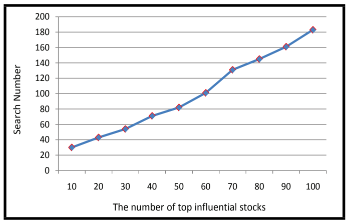


Fig. 6: Evaluation on search number

TABLE I: Top 5 dynamic influential stocks

ID	Stock Name	Industry
600000	Pudong Development	Finance
600036	China Merchants Bank	Finance
600019	Baosteel	Steel
600028	China Petrochemical	Oil
600030	CITIC Bank	Finance

cost is reduced a great deal. What's more, this algorithm is an exact algorithm, which reduces the time cost without compromising the quality.

D. Evaluation on extended type of influence

We show how effectively the model can capture group and dynamic influence in real stock market. Since the model only works on a relatively small stock influence network with limited information, the results may not ideally reflect the real facts. But to some extent, it demonstrates that the group and dynamic influence discovered by our model is consistent to real stock market.

The result of influential group discovery in $SSE50$ is presented in Fig. 7. From the results, we can get some meaningful conclusions. The groups consisting of $Zhongjing$ $Gold$ and $China$ $Unicom$ are both in the rank 1st group, while in reality, these stocks are of high influence in Chinese stock market [12].

Table 1 shows an example of dynamic ranking of top 5 influential ones among the 1346 stocks, given time interval ranging from 2008 to 2012, and time segment is set as one year round. The recent achievements of these five stocks confirms the effectiveness of our results [12]. An interesting finding is that the finance industry has the largest portions of the top influential stocks, which implies the prosperity of the finance sector in China these years.

E. Case Study

The proposed model can be helpful in many cases, and here we focus on its application to price prediction. Intuitively, some stocks are so active that they will impact the overall trend. Given a stock pool, we can estimate its overall trend based on influential stocks obtained from our approach. Detailed information lies in Eq. 10, where $d(t,j)$ is $(p(t)-p(t-1))/p(t-1)$, representing the volatility of stock S_j, while $p(t)$ is its price value at time t and Δ is the smallest time interval.

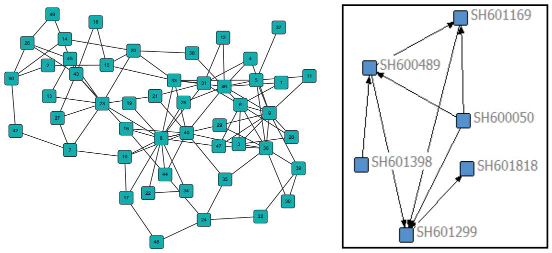


Fig. 7: Group Influence results

$$d(t+\Delta) = \frac{1}{\sum_{j=1}^{Z} f_{j\to V}} \sum\nolimits_{j=1}^{Z} f_{j\to V} \cdot d(t,j) \tag{10}$$

To describe the price prediction task of the stock index clearly, we use the price prediction on $CSI800$ on the first 60 days of 2012 by the model as an example. $CSI800$ is compiled by the pool of 800 stocks and top influential stocks are chosen to predict the overall trend. Actually the more number of stocks selected, the better performance is; and here we picked the top 15 ones. The result is displayed in Fig. 8. Compared with real data, the predicted price curve matches the trend of real price changing nicely, somewhat leading, except for small deviations. The result confirms the usefulness of our model in price predictions. One thing to mention is that the proposed method in addition to other factors can offer a better guide to investors, for the price fluctuation is a result of many reasons in the market. When applied to real trading strategies, a lot of other things should be taken into account. However, it's not the focus of this paper.

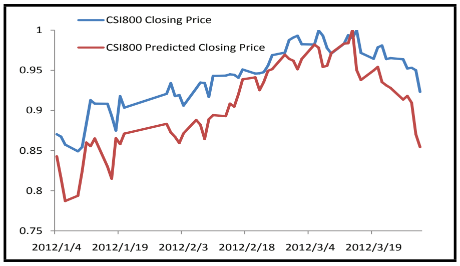


Fig. 8: CSI800 Index vs. Predicted CSI800 Index

VI. Related Work

There are several areas of work that we build upon: time series analysis, network modeling and financial data mining.

Time Series Analysis The similarity between the sequences in a fixed time intervals has been studied in [13], [14]. [15] addressed the view that influence is not equal to similarity. Causal relations investigated by Granger Causality tests have been shown in [16], [17]. [18] focus on analyzing meaningful events in the time series rather than the time series numerical values directly.

Network Modeling [19] proposed the first network model in stock market. [1] illustrate an international financial network, where the nodes represent major financial institutions and the links are the strongest existing relations among them. [20] proposed an approach to measure financial contagion. Influence diffusion has attracted considerable attentions in social networks [21], [22], [23]. [10] proposed the method of extracting influential nodes based on the propagation models.

Financial Data Mining Financial data mining [24], [25] is another area of the related work. An empirical analysis of clusters of interest rates in money and capital markets is performed [26]. [27] use support vector machine with a hybrid feature selection method to stock trend prediction, while [28] apply neural networks for forecasting stock market returns. However, this paper tries to discover influence in stock markets by applying social network analysis methods.

VII. Conclusion

In this paper, we devise models and algorithms for the problem of mining influence in evolving entities, with applications to stock market. The superiority of the model is demonstrated by extensive experiments. Meanwhile, evaluation with large real data set and applications to other domains, such as climate science and intelligent transportation, are the future work of our research.

Acknowledgment This work was supported in part by National Science & Technology Support Program No.2012BAH13F04 and No.2012BAH13F02.

References

[1] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White, "Economic networks: The new challenges," *science*, vol. 325, no. 5939, p. 422, 2009.

[2] D. Wu, Y. Ke, J. X. Yu, S. Y. Philip, and L. Chen, "Detecting leaders from correlated time series," in *Database Systems for Advanced Applications*. Springer, 2010, pp. 352–367.

[3] B.-K. Yi, N. D. Sidiropoulos, T. Johnson, H. Jagadish, C. Faloutsos, and A. Biliris, "Online data mining for co-evolving time sequences," in *Data Engineering, 2000. Proceedings. 16th International Conference on*. IEEE, 2000, pp. 13–22.

[4] X. Shi, W. Fan, J. Zhang, and P. Yu, "Discovering shakers from evolving entities via cascading graph inference," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1001–1009.

[5] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie, "Discovering spatio-temporal causal interactions in traffic data streams," in *SIGKDD 2011*. Association for Computing Machinery, Inc., August 2011.

[6] A. N.-M. Y. L. C. P. J. H. A. C. Lozano, H. Li and N. Abe, "Spatial-temporal causal modeling for climate change attribution," in *SIGKDD 2009*. Proceedings of the 15th SIGKDD Conference on Knowledge Discovery and Data Mining, 2009, pp. 587–596.

[7] Y. L. Becker, H. Fox, and P. Fei, "An empirical study of multi-objective algorithms for stock ranking," in *Genetic Programming Theory and Practice V*. Springer, 2008, pp. 239–259.

[8] M. Zhu, D. Philpotts, R. Sparks, and M. J. Stevenson, "A hybrid approach to combining cart and logistic regression for stock ranking," *The Journal of Portfolio Management*, vol. 38, no. 1, pp. 100–109, 2011.

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

[10] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *AAAI*, vol. 7, 2007, pp. 1371–1376.

[11] T. Abatzoglou and B. O'Donnell, "Minimization by coordinate descent," *Journal of Optimization Theory and Applications*, vol. 36, no. 2, pp. 163–174, 1982.

[12] Sina!Finance, http://finance.sina.com.cn/realstock/, 2013.

[13] J. Jiang, K. Ma, and X. Cai, "Non-linear characteristics and long-range correlations in asian stock markets," *Physica A: Statistical Mechanics and its Applications*, vol. 378, no. 2, pp. 399–407, 2007.

[14] D. Pohl and A. Bouchachia, "Financial time series processing: A roadmap of online and offline methods," in *Business Intelligence and Performance Management*. Springer, 2013, pp. 145–162.

[15] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 7–15.

[16] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 66–75.

[17] M. T. Bahadori and Y. Liu, "On causality inference in time series," in *2012 AAAI Fall Symposium Series*, 2012.

[18] G. Ganeshapillai, J. Guttag, and A. Lo, "Learning connections in financial time series," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 109–117.

[19] R. N. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 11, no. 1, pp. 193–197, 1999.

[20] K.-H. Bae, G. A. Karolyi, and R. M. Stulz, "A new approach to measuring financial contagion," *Review of Financial studies*, vol. 16, no. 3, pp. 717–763, 2003.

[21] 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*. ACM, 2003, pp. 137–146.

[22] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 88–97.

[23] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2008, pp. 67–75.

[24] D. Zhang and L. Zhou, "Discovering golden nuggets: data mining in financial application," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 34, no. 4, pp. 513–522, 2004.

[25] B. Kovalerchuk and E. Vityaev, "Data mining in finance," *Advances in relational and hybrid methods*, 2000.

[26] T. Di Matteo, T. Aste, and R. N. Mantegna, "An interest rates cluster analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 339, no. 1, pp. 181–188, 2004.

[27] M.-C. Lee, "Using support vector machine with a hybrid feature selection method to the stock trend prediction," *Expert Systems with Applications*, vol. 36, no. 8, pp. 10 896–10 904, 2009.

[28] D. Enke and S. Thawornwong, "The use of data mining and neural networks for forecasting stock market returns," *Expert Systems with applications*, vol. 29, no. 4, pp. 927–940, 2005.