Data Min Knowl Disc (2009) 18:419–445
DOI 10.1007/s10618-008-0124-z

A link mining algorithm for earnings forecast and trading

Germán Creamer · Sal Stolfo

Received: 7 March 2007 / Accepted: 17 November 2008 / Published online: 18 December 2008

Abstract The objective of this paper is to present and discuss a link mining algorithm called CorpInterlock and its application to the financial domain. This algorithm selects the largest strongly connected component of a social network and ranks its vertices using several indicators of distance and centrality. These indicators are merged with other relevant indicators in order to forecast new variables using a boosting algorithm. We applied the algorithm CorpInterlock to integrate the metrics of an extended corporate interlock (social network of directors and financial analysts) with corporate fundamental variables and analysts' predictions (consensus). CorpInterlock used these metrics to forecast the trend of the cumulative abnormal return and earnings surprise of S&P 500 companies. The rationality behind this approach is that the corporate interlock has a direct effect on future earnings and returns because these variables affect directors and managers' compensation. The financial analysts engage in what the agency theory calls the "earnings game": Managers want to meet the financial forecasts of the analysts and analysts want to increase their compensation

Responsible editor: Eamonn Keogh.

A preliminary version of this paper was presented at the Link Analysis: Dynamics and Statics of Large Networks Workshop on the International Conference on Knowledge Discovery and Data Mining (KDD) 2006.

G. Creamer (✉) · S. Stolfo
Department of Computer Science, Columbia University, 500 W 120 St, New York, NY 10027, USA
e-mail: ggc14@columbia.edu

S. Stolfo
e-mail: sal@cs.columbia.edu

G. Creamer
Centrum Catolica, Pontificia Universidad Católica del Peru, Lima, Peru
e-mail: gcreamer@pucp.edu.pe

or business of the company that they follow. Following the CorpInterlock algorithm, we calculated a group of well-known social network metrics and integrated with economic variables using Logitboost. We used the results of the CorpInterlock algorithm to evaluate several trading strategies. We observed an improvement of the Sharpe ratio (risk-adjustment return) when we used "long only" trading strategies with the extended corporate interlock instead of the basic corporate interlock before the regulation Fair Disclosure (FD) was adopted (1998–2001). There was no major difference among the trading strategies after 2001. Additionally, the CorpInterlock algorithm implemented with Logitboost showed a significantly lower test error than when the CorpInterlock algorithm was implemented with logistic regression. We conclude that the CorpInterlock algorithm showed to be an effective forecasting algorithm and supported profitable trading strategies.

Keywords Link mining · Social network · Machine learning · Computational finance · Trading strategies · Data mining applications

1 Introduction

The application of networks to social science has a long tradition since the seminal works of Moreno (1932) and Milgram (1967) about the representation of group dynamics in a sociogram and the "small world" problem. In Milgram's experiment letters are passed from acquaintance to acquaintance. As a result, he showed how apparently distant people are connected by a very short chain of acquaintances. Most of the current literature in social networks is oriented to classify networks, to identify their properties, or to develop new cluster algorithms. Less attention has been devoted to use social networks as a forecasting tool. Recently, link mining has emerged as a new area of research that partially fills this gap.

Link mining[1] is a set of techniques that uses different types of networks and their indicators to forecast or to model a linked domain. Link mining has had several applications (Senator 2005) to different areas such as money laundering (Kirkland et al. 1999), telephone fraud detection (Fawcett and Provost 1999), crime detection (Sparrow 1991), and surveillance of the NASDAQ and other markets (Kirkland et al. 1999; Goldberg et al. 2003). One of the most important business applications of link mining is in the area of viral marketing or network-based marketing. Following this trend, Domingos and Richardson (2001) simulate markets as Markov random fields where customer value depends of the profitability of each customer according to its buying decisions and its capacity to influence other customers. Richardson and Domingos (2006) apply this latter work to knowledge-sharing sites. Leskovec et al. (2006) show, using a stochastic viral marketing model, that the effectiveness of recommendations of highly connected persons decline as the number of recommendations are larger than a certain threshold. However, very limited research has been done combining social network indicators with other relevant indicators. An innovative paper in the area of direct marketing is Hill et al. (2006) which has combined link mining

[1] For a recent survey see Getoor and Diehl (2005).

indicators with demographic and consumer-specific attributes to evaluate the response rate[2] of prospects associated to existent customers of a telecommunications company. In this paper we propose a link mining algorithm called CorpInterlock that merges social network indicators with any other relevant indicators to forecasting a variable that is mostly associated with the social network. We apply this algorithm for financial forecasting using social networks of corporate directors and financial analysts, however it can be applied to other related areas such as direct marketing.

We refer to the social network among directors as the basic corporate interlock, and the social network among directors and analysts as the extended corporate interlock. We use the definition of cumulative abnormal return (CAR) as the return of a specific asset less the average return of all assets in its risk-level portfolio for each trading date, and earnings surprise or forecast error (FE) as the difference between the forecast of financial analysts and the actual earnings at the end of the period of evaluation (see Appendix 1). The implementation of our algorithm specifically forecast CAR and FE using indicators of the basic and extended corporate interlock and a group of well-known investment variables presented in the appendix 1. From our perspective, we do not know of any previous research that has used social network indicators combined with economic determinants to forecast CAR and FE. We think that if the corporate interlock plays such an important role in corporate governance, it may also have an impact to forecast CAR and FE.

The reason that we study the extended social network of directors and analysts is because their relationship is part of what is called the principal agent problem in finance literature. The principal agent problem stems from the tension between the interests of the investors in increasing the value of the company (principals) and the personal interests of the managers (agents). This conflict of interest is evident in many of the recent bankruptcy scandals in publicly held US companies such as Enron and WorldCom, and has also led to the so-called “earnings game”. CEOs’ compensation depends on their stock options. So, top managers concentrate on the management of earnings and surprises. Wall Street companies want to keep selling stocks. Thus, analysts try to maintain positive reviews of the companies.[3] Once a prediction is published, CEOs do whatever is necessary to reach that prediction or boost the results above analysts’ prediction. CEOs play this game, even though a company may lose value in the long-term. Hence, the extended corporate interlock could help to transfer information between directors and analysts and also may bring more information to forecast earnings surprise than a basic corporate interlock. Additionally, we expect that statistics of an extended corporate interlock could be able to predict return or earnings surprises better than cumulative abnormal return because of the informal communications among directors and analysts that may explain earnings surprises. This methodology could also be applied to a larger class of measures as long as the social network used is relevant to the selected indicator. For instance, a labor economist

[2] Response rate is the number of solicitations that prospects respond in relation to the total number of solicitations.

[3] This situation is changing because of the regulations introduced by the regulation FD and the Sarbanes-Oxley Act of 2002.

may use a social network that includes board of directors members and workers leaders in order to evaluate labor productivity or quality of workers benefits.

The rest of the paper is organized as follows: Sect. 2 describes the "small world" model and the corporate interlock; Sect. 3 introduces the finance literature on earnings surprise; Sect. 4 presents the methods used to forecast the stock market: a link mining algorithm, and boosting; Sect. 5 explains in detail our forecasting and trading strategies; Sect. 6 presents the results of our forecast; Sect. 7 discusses the results, and Sect. 8 presents the conclusions. The appendix 1 introduces the main investment indicators used in this research.

2 Corporate interlock

Watts (1999), Watts and Strogatz (1998), Newman et al. (2001) and Newman et al. (2002) have formalized and extended the "small world" model. The relevant aspect of the "small world" model is that it is possible to characterize an undirected graph $G(V, E)$ by its structural indicators where $V = v_1, v_2, \ldots, v_n$ is the set of vertices, E is the set of edges, and e_{ij} is the edge between vertices v_i and v_j:

- Clustering coefficient: $C \doteq \frac{1}{n}\sum_{i=1}^{n} CC_i$, where:
  - $CC_i \doteq \frac{2|\{e_{ij}\}|}{deg(v_i)(deg(v_i)-1)} : v_j \in N_i,\ e_{ij} \in E$. Each vertex v_i has a neighborhood N defined by its immediately connected neighbors: $N_i = \{v_j\} :\ e_{ij} \in E$.
  - $deg(v_i)$ is the degree centrality or degree of a vertex $v_i : deg(v_i) \doteq \sum_j a_{ij}$
  - a_{ij} is an element of the adjacent matrix A of G
  - k is the average degree of the vertices
  - n is the number of vertices in G
- Mean of characteristic path lengths between its vertices: $L \doteq \frac{1}{n}\sum_j d_{ij}$, where $d_{ij} \in D$ and D is the geodesic distance matrix (matrix of all shortest path between every pair of vertices) of G.

In the case of a random network, these structural indicators are $L_{random} \approx \frac{ln(n)}{ln(k)}$ and $C_{random} \approx \frac{k}{n}$.

Using the above indicators, the four properties that characterize a "small world" network are:

I n is fixed and numerically large $n \gg 1$.

II k is fixed so that G is sparse ($k \ll n$), and with a minimum number of potential structures ($k \gg 1$).

III G is decentralized. So, there is not a single dominant vertex: $k_{max} \ll n$ where k_{max} is the maximal degree.

IV G must be strongly connected.

C works as a measure of order in G, where if $C \gg k/n$, then G is considered locally ordered, while random graphs are not ordered and therefore C_{random} is very small as the above property II ($k \ll n$) implies. If a graph is locally ordered or highly clustered, then it should have long characteristic path lengths in order to communicate its different clusters. Obviously, a random graph is not ordered, therefore $C_{random} \ll C$, and $L \approx L_{random}$. As a result, a simple way to evaluate the "small world" properties of a network is if the "small world" ratio ($SW \doteq \frac{C}{L} \cdot \frac{L_{random}}{C_{random}}$) is much larger than one.

Other additional indicators of social networks that we have used in this study are:

1. Closeness centrality (normalized): $C_c(v_i) \doteq \frac{n-1}{\sum_j d_{ij}}$, where d_{ij} is an element of the geodesic distance matrix D (Freeman 1979; Borgatti and Everett 2006).
2. Betweenness centrality $B_c(v_i) \doteq \sum_i \sum_j \frac{g_{kij}}{g_{kj}}$. This is the proportion of all geodesic distances of all other vertices that include vertex v_i where g_{kij} is the number of geodesic paths between vertices k and j that include vertex i, and g_{kj} is the number of geodesic paths between k and j (Freeman 1979).
3. Normalized clustering coefficient: $CC_i' \doteq \frac{deg(v_i)}{MaxDeg} CC_i$, where MaxDeg is the maximum degree of vertex in a network (de Nooy et al. 2005).

Several networks in the social and natural sciences have been identified to have the properties of a "small world" (Watts and Strogatz 1998; Barabasi 2002). We are particularly interested in those organizational studies about the corporate interlock or the social network of directors. Davis et al. (2003) have found that the basic corporate interlock of the major US corporations (those in the Fortune 500 list) between 1982 and 1999 has the characteristics of a "small world" as described above. A "small world" in the case of the corporate interlock implies that the average distance between firms, between directors, and (if applicable) between analysts is very short. Davis et al. (2003) also find that the basic corporate interlock is highly stable, even after major changes in corporate governance. Mintz and Schwartz (1985), following Mills (1956), study how commercial banks have a central position in the corporate interlock because of the participation of the major leaders of US non-financial corporations on the banks' boards. The original thesis of Mills (1956) is that a small group of business leaders, interconnected by being part of the same boards of directors, is able to coordinate policies, share practices, and finally control the major corporations. One of the contributions of the "small world" literature in this area is to understand that this connection in the corporate elite is based on the direct link among different actors such as directors, and is not necessarily based on the banking sector or does not require a high level of ownership concentration.[4] As we demonstrate in this paper, the "small world" phenomenon is observed in the basic and extended corporate interlock. Hence, the strength of these corporate interlocks might be the result of the interaction of different individuals that interact among several firms and boards, and not the result of a small central group that tries to control the society.

The use of the corporate interlock to understand and solve finance problems is just becoming more relevant. Larcker et al. (2005) have found that the distance between inside and outside directors, excluding the links when directors are part of the same board, affect CEO's compensation. The interesting aspect of this latter paper is that the authors control for standard economic determinants besides the organizational variables. Cohen et al. (2008) have observed that sell-side equity analysts outperform on their recommendations when there is an educational link to senior officers

[4] For a dynamic demonstration of the network of directors of the largest American companies see ⟨http://www.theyRule.net/⟩.

of companies that they follow. Very few previous papers have studied the economic effects of corporate interlocks such as their effect on the decision process of: 1. making political contributions (Mizruchi 1992), 2. poison pills (Davis 1991), and 3. switching from NASDAQ to NYSE (Rao et al. 2000).

We expect that the literature of finance may significantly enrich if the corporate interlock dimension is included into the analysis. Most of the current studies in corporate finance treat the board of directors of every firm independent of the rest. However, the above evidence as well as the one presented in this paper indicates that board of directors are highly linked among themselves and among other networks such as the network of financial analysts. An interesting line of research for board of directors is the valuation of their members according to their connectiveness with other companies. Members highly connected may bring additional businesses or relevant information that may improve company performance. In the case of companies that depend of government contracts, financial analysts might be very interested to take into account the connections among directors and senior government officers that may facilitate the access to future contracts. Likewise, companies that desire to have a global presence, may have a different level of valuation according to the participation of its directors into international business.

3 Earnings surprise

A very well-known phenomenon studied in the accounting and behavioral finance literature is the earnings surprise effect. Earnings surprise or forecast error refers to the difference between financial analysts' predictions and the actual earnings reported by companies. The earnings surprise effect emphasizes how the market reacts more to negative surprises than to positive surprises. Therefore, investors and fund managers have developed many trading strategies around the earnings announcement period and invest significant resources trying to predict earnings surprises. An important source of information for investors are the predictions of more than 3,000 analysts collated in huge databases created by several companies such as IBES International Inc., Zacks Investment Research, and First Call Corporation. These provide investors with a "consensus", or simple average of the market analysts' predictions, which they use to estimate what the market will do.

Other researchers use analysts' predictions for such forecasts, allowing them to make early investment decisions before quarterly announcements. The method they use is linear regression analysis using variables such as the characteristics of companies, and analysts. These studies suggested that analysts' forecasts may have predictive value (Ou and Penman 1989; Stober 1992; Bernard and Thomas 1990; Mendenhall 1991; Abarnabell and Bernard 1992; Peters 1993a,b). Brown et al. (1996) standardized a method to calculate the earnings surprise with an indicator that they call "earnings surprise predictor". This "earnings surprise predictor" outperforms the market using a portfolio of S&P 500 companies during the period 1985–1994.[5] We believe that recent developments in the area of machine learning and link mining can contribute to

[5] For a detailed list of references about the academic use of analysts' predictions see Brown (2000).

this debate, and especially formalize the study of patterns of behavior for trading and financial forecasting as proposed by the behavioral finance approach. This approach sustains that markets are inefficients and move on individual biases or behavioral patterns (Thaler 2005). In this paper we propose a link mining algorithm that improves the earnings and return predictions combining well-known corporate variables with metrics of a social network of directors and analysts. The association among directors and financial analysts may allow companies to adjust earnings to the forecast of financial analysts. However, this relationship is not easily captured by linear regression analysis. Link mining algorithms may explain the relationship among organizational and economic variables, and therefore improve stock price prediction.

Earlier studies on analysts and earnings surprise show at least two types of major variables that are typical of these studies. First, researchers have quantified companies' characteristics or actions, since companies' changes have been shown to relate to analysts' recommendations (Stickel 1995).[6] Secondly, there are variables which quantify analysts' predictions, such as the quality of their recommendation (Womack 1996; Elton et al. 1986; Barber et al. 2001); the accuracy of their past predictions (Brown 2001); the revisions they make (Peterson and Peterson 1995); the company variables they use (Finger and Landsman 1999; Stickel 1995; Krische and Lee 2000); the career moves of analysts (Hong and Kubik 2003); the timing of analyst's predictions;[7] the herding behavior of analysts (Clement and Tse 2005); and the information content of analysts' reports (Asquith et al. 2005).

Several studies have evaluated investment strategies that follow consensus recommendations of analysts. A particularly sophisticated model was developed by the company Starmine, which ranks analysts and makes its predictions "Smart estimate" using the forecasts of the most highly ranked analysts. Barber et al. (2001) find that after taking transaction costs into account, the high-trading level of strategies that follow consensus recommendations of analysts do not give a consistent return greater than zero. A similar result is obtained by Mikhail et al. (2002) even after taking into account analysts' prior performance. They recommend that those investors that still want to follow analysts' recommendations may benefit if they use the forecasts of highly ranked analysts with at least 5 years of superior performance in rankings surveys such as those collected by The Wall Street Journal. Jegadeesh et al. (2004) reported that analysts from sell-side firms recommend mostly "glamour stocks" (characterized by positive momentum, high growth, high volume, and relatively high prices); however, investors that blindly follow a strategy that invests in these recommended stocks may not obtain positive returns because investment in these stocks also requires favorable quantitative indicators (i.e. high value and positive momentum).[8]

[6] Beckers et al. (2004) find that after the European integration in 1992, country differences is not a relevant factor to explain earnings forecasts differences between analysts, however sector is still an important factor.

[7] Ivkovic and Jegadeesh (2004) find that the information content of upward earnings forecast revisions and recommendation upgrades increase near the earnings announcement date, while they are less informative in the week that follows this date. This situation is not observed for recommendation downgrades and downward revisions.

[8] Abarbanell (1991) finds that analyst's forecasts do not completely integrate the information of past prices changes; additionally, Abarnabell and Bernard (1992) find that the under-reaction of analysts to recent earnings is only a partial explanation for the under-reaction of stock prices to earnings.

4 Methods

4.1 Boosting

Adaboost is a machine learning algorithm invented by Freund and Schapire (1997) that classify its outputs applying a simple learning algorithm (weak learner) to several iterations of the training set where the misclassified observations receive more weight. Freund and Mason (1999) proposed a decision tree learning algorithm called an *alternating decision tree* (ADT). In this algorithm, boosting is used to obtain the decision rules and to combine them using a weighted majority vote.

Friedman et al. (2000), followed by Collins et al. (2004) suggested a modification of Adaboost, called Logitboost. Logitboost can be interpreted as an algorithm for step-wise logistic regression. This modified version of Adaboost—known as Logitboost—assumes that the labels $y_{i'}s$ were stochastically generated as a function of the $x_i's$. Then it includes $F_{t-1}(x_i)$ in the logistic function to calculate the probability of y_i, and the exponent of the logistic function becomes the weight of the training examples. Figure 1 describes Logitboost.

4.2 CorpInterlock: a link mining algorithm

In this paper we propose a link mining algorithm called CorpInterlock (see Fig. 2). This algorithm transforms a bipartite graph into a one-mode graph, selects the largest strongly connected component of a social network and ranks its vertices using several indicators of distance and centrality. These indicators are merged with other relevant indicators in order to forecast new variables using a machine learning algorithm.

The algorithm also calculates the "small world" ratio. Even though this ratio is not used as an input in the forecast and there is not a recommended level required for the CorpInterlock algorithm, this indicator is very useful to understand the nature of the corporate interlock. In our current application to finance problems, the "small world" property of a network may explain how information is transmitted. Step 3 of the CorpInterlock algorithm (Fig. 2) satisfies the requirement that a "small world" network must be strongly connected, and the weakly connected requirement of closeness centrality. Step 4 calculates the adjacency matrix A and geodesic distance matrix D used as inputs of the social network indicators and the "small world" ratio calculated in step 6.

Fig. 1 The Logitboost algorithm (Friedman et al. 2000). y_i is the binary label to be predicted, x_i corresponds to the features of an instance i, w_i^t is the weight of instance i at time t, h_t and $F_t(x)$ are the prediction rule and the prediction score at time t respectively

```
$F_0(x) \equiv 0$
for $t = 1 \ldots T$
    $w_i^t = \frac{1}{1+e^{-y_i F_{t-1}(x_i)}}$
    Get $h_t$ from weak learner
    $\alpha_t = \frac{1}{2} \ln \left( \frac{\sum_{i:h_t(x_i)=1, y_i=1} w_i^t}{\sum_{i:h_t(x_i)=1, y_i=-1} w_i^t} \right)$
    $F_{t+1} = F_t + \alpha_t h_t$
```

Input:
Two disjoint nonempty sets V_{11} and V_{12}.

1. Build a bipartite graph $G_1(V_1, E_1)$ where its vertex set V_1 is partitioned into two disjoint sets V_{11} and V_{12} such that every edge in E_1 links a vertex in V_{11} and a vertex in V_{12}.
2. Build a one-mode graph $G_2(V_2, E_2)$ in which there exist an edge between v_i and v_j: $v_i,\ v_j \in\ V_2$ if and only v_i and v_j share at least a vertex $u_i\ \in\ V_{12}$. The value of the edge is equal to the total number of objects in V_{12} that they have in common.
3. Calculate the largest strongly connected component of G_2 and call it $G_3(V_3, E_3)$.
4. Calculate the adjacency matrix A and geodesic distance matrix D for G_3. a_{ij} and d_{ij} are the elements of A and D respectively. The mean of all the distances d_{ij} is L.
5. For each vertex $v_i \in\ V_3$ calculate the following social network indicators:
   - Degree centrality: $deg(v_i) = \sum_j a_{ij}$
   - Closeness centrality (normalized): $C_c(v_i) \doteq \frac{n-1}{\sum_j d_{ij}}$
   - Betweenness centrality: $B_c(v_i) = \sum_i \sum_j \frac{g_{kij}}{g_{kj}}$, where g_{kij} is the number of geodesic paths between vertices k and j that include vertex i, and g_{kj} is the number of geodesic paths between k and j.
   - Clustering coefficient: $CC_i = \frac{2|\{e_{ij}\}|}{deg(v_i)(deg(v_i)-1)} : v_j\ \in\ N_i,\ e_{ij} \in\ E$
   - Normalized clustering coefficient: $CC'_i = \frac{deg(v_i)}{MaxDeg} CC_i$, where MaxDeg is the maximum degree of vertex in a network

6. For the complete network calculate the “small world” ratio: $SW\ =\ \frac{C}{L}\ \cdot\ \frac{L_{random}}{C_{random}}$, where $C\ =\ \frac{1}{n}\sum_{i=1}^{n} CC_i$, $L_{random} \approx \frac{ln(n)}{ln(k)}$ and $C_{random} \approx \frac{k}{n}$.
7. Merge social network indicators with any other relevant set of variables for the population under study such as analysts’ forecasts, labels and economic variables and generate test and training samples.
8. Run machine learning algorithm with above test and training samples to predict Y variable.

Output:
Prediction of Y and “small world” ratio.

Fig. 2 The CorpInterlock algorithm

4.2.1 *Application to forecasting earnings surprise*

We used the CorpInterlock link mining algorithm (see Fig. 2) to build a bipartite social network where the nodes of the partition V_{12} representing the directors and analysts are connected to nodes of the partition V_{11} representing companies that they direct or cover. This social network is converted into a one-mode network where the vertices are the companies and the edges are the number of directors and analysts that every pair of companies have in common. This is the extended corporate interlock. The basic corporate interlock is calculated in the same way using only directors. The algorithm merges a group of investment variables presented in the appendix 1 and a group of social network statistics obtained from the basic or extended corporate interlock. Finally, the algorithm predicts FE and CAR using a machine learning algorithm such as boosting.

We consider that this financial application of the CorpInterlock algorithm is appropriate because the increasing importance of organizational and corporate governance issues in the stock market requires the extraction of indicators from the extended and basic corporate interlock and integration with more traditional economic indicators in order to forecast CAR and FE. The indicators calculated by the CorpInterlock algorithm captures the power relationship among directors and financial analysts as follows:

1. *Degree centrality* directors and analysts of a company characterized by a high degree or degree centrality coefficient are connected among them through several companies.

2. *Closeness centrality* directors and analysts of a company characterized by a high closeness centrality coefficient are connected among them through several companies that are linked through short paths.
3. *Betweenness centrality* directors and analysts of a reference company characterized by a high betweenness centrality coefficient are connected among them through several companies. Additionally, the reference company mentioned above has a central role because it lies between several other companies, and no other company lies between this reference company and the rest of the companies.
4. *Clustering coefficient* directors and analysts of a company characterized by a high clustering coefficient are probably as connected among them as it is possible through several companies.

Each of the above measures show a different perspective of the connection between directors and analysts as described in the "earnings game" where the earnings forecast of analysts are aligned with management's expectations. Hence, we can include them in a decision system to forecast FE and CAR.

CorpInterlock can be implemented with any efficient machine learning algorithm that is appropriate for the problem under study. However, the importance of features used to predict earnings surprises, and cumulative abnormal returns may change significantly in different periods of time. As we do not know in advance what the most important features are and because of its feature selection capability, its error bound proofs (Freund and Schapire 1997), its interpretability, and its capacity to combine economic and organizational variables to optimize the earnings surprise and cumulative abnormal return prediction we decided to use boosting, specifically Logitboost, as our learning algorithm. Additionally, Creamer and Freund (2004, 2005, 2007) have already applied boosting to forecast equity prices and corporate performance and our tests showed that Logitboost performs significantly better than logistic regression, our baseline algorithm.

Dhar and Chou (2001) have already compared the predictive accuracy of tree-induction algorithms, neural networks, naive Bayesian learning, and genetic algorithms to classify the earnings surprise before announcement. They used a definition of earnings surprise or forecast error that we have also adopted in this research:

$$FE \doteq \frac{\mathrm{CONSENSUS}_q - \mathrm{EPS}_q}{|\mathrm{CONSENSUS}_q| + |\mathrm{EPS}_q|}$$

where $\mathrm{CONSENSUS}_q$ is the mean of earnings estimate by financial analysts for quarter q, and EPS_q is the actual earnings per share for quarter q. *FE* is a normalized variable with values between -1 and 1. Additionally, when $\mathrm{CONSENSUS}_q$ is close to zero and EPS_q is not, then the denominator will not be close to zero.

4.2.2 Other applications: viral marketing

The CorpInterlock algorithm could also be used in other domains where social network indicators are part of the inputs used in the prediction. A good example explored by Hill et al. (2006) comes from direct marketing. Corporations that use direct marketing intensively have large databases of their current customers, and they spend a significant amount of time and money trying to reach new customers. However, the response rate of prospects is extremely low considering that the best prospects receive a significant number of weekly offers from many different companies. There is the possibility that this response rate may increase when prospects are related to existent loyal customers. In our future research, we could use the CorpInterlock algorithm to test the following hypotheses: demographic, and marketing indicators combined with the indicators of a social network of existent customers and prospects that live in the same household may improve the selection of prospects that show a higher rate of response in relation to a network that does not include existent customers.

5 Experiments

We restricted our experiments to companies that are part of the US stock market. We obtained the price and return series from the Center for Research in Security Prices (CRSP), the accounting variables from COMPUSTAT,[9] the list of financial analysts and earnings forecast or consensus from IBES, and the list of directors from the Investor Responsibility Research Center. The list of directors exists only on an annual basis for the period 1996–2005. This restricts our analysis to this period. The number of companies under study changes every year. The minimum and maximum number of companies included in our study are 3,043 for 2005 and 4,215 for 1998.

We applied the CorpInterlock algorithm described in Fig. 2 using the softwares EMT (Stolfo et al. 2006) and Pajek (de Nooy et al. 2005) to obtain the basic and extended corporate interlock. We computed the investment signals and a group of the social network statistics introduced in Sect. 1 [average distance, betweenness centrality, closeness centrality, degree centralization, degree, and clustering coefficient (normalized and unnormalized)] of the basic and extended corporate interlock. We merged our accounting information, analysts' predictions (consensus) and social networks statistics using quarterly data, and selected the last quarter available for every year. Most of the fundamental and accounting variables used are well-known in finance literature and Jegadeesh et al. (2004) demonstrated that these variables are good predictors of cross-sectional returns (see the appendix 1 for an explanation of the variables used). We forecasted two different trends: FE and CAR. In both cases, we labeled an instance as 1 if the trend was positive and -1 otherwise. We calculated the label of CAR using the cumulative abnormal return of the month following the earnings announcement. CAR is calculated as the return of a specific asset less the value weighted average return of all assets in its risk-level portfolio according to CRSP. We computed FE using the predictions of the analysts available 20 days before the earnings announcement as fund managers may suggest (Dhar and Chou 2001). Fund managers take a position, short or long,[10] a certain number of days before the earnings announcement and, according to their strategy, they will liquidate the position a given number of days after the earnings announcement. If fund managers know the trend of

[9] COMPUSTAT is an accounting database managed by Standard & Poor's.

[10] Long or short positions refer to buy a specific asset or to sell a borrowed asset based on the expectation that price of the asset will increase or decrease respectively.

FE or CAR, they make take a position according to their expectations; however they do not need to know exactly what the future stock price is going to be. They profit when the market moves in the direction expected, and above a certain threshold, even though the market movement might not be in the exact amount forecasted. For this reason, the emphasis of this paper is in the improvement of the prediction of the trend of FE and CAR–and not in their value–with the inclusion of the extended corporate interlock information.

We implemented the CorpInterlock algorithm using Logitboost. To evaluate the difficulty of the classification task, we compared our method with random forests (Breiman 2001), and logistic regression (Cessie and Houwelingen 1992). The latter algorithm was our baseline method. We implemented ADTs and Logitboost with 50 iterations, and random forests with 100 trees and five features[11] using the Weka package (Witten and Frank 2005). We generated eight training models for each learning algorithm on a growing window, each one for every year from 1997 to 2004. Our first data set is from 1997 so that it has the accumulated data of 1996 and 1997. Every year we tested our training model with two test samples of the following year from 1998 to 2005. As a result we have 16 test samples that we can use to evaluate how our algorithm and our trading strategies perform with samples of different time periods. The test errors that we obtained were the result of averaging the results of running our models with all variables over the 16 sets.

As we are including all the companies that are part of the US stock market for every year, if a company is listed during our period of evaluation it becomes part of our sample. Likewise, if a company is delisted during our period of evaluation, then this company is not anymore part of our sample. Therefore, we avoided the very common survivorship bias. We eliminated companies that did not have earnings or CAR information.

We ran linear regressions using FE and CAR as dependent variables, and evaluated the importance of the variables listed in the appendix 1 for the model. We tested our model for heteroscedasticity and multicollinearity using the test of Breusch and Pagan (1979), and the variance inflation factor (VIF) (Davis et al. 1986) respectively. We did not find heteroscedasticity or multicollinearity in our sample. In any case, if there was any multicollinearity, it was overcome by boosting's feature selection capability.[12]

We also tested several trading strategies assuming that traders will take a position based on the results of our forecast and will liquidate their positions a month after each earnings announcement. These trading positions are taken only in the last quarter of the year because the data of our social network indicators is annual. We did not take into account transaction costs considering that they are very low because there is only a major buy–sell transaction per year, and all our strategies are affected by the same costs. So, there is no difference in relative terms. The trading positions that we simulate have the opposite sign of the Logitboost prediction for FE. In cases that we predict a positive FE, we take a short position and vice versa. The reason for this policy is

[11] We implemented random forests with five features in order to optimize its performance.

[12] The regression is heteroscedastic if the variance of the residuals is not constant across observations. Multicollinearity is the presence of correlation among dependent variables. For a more detailed presentation see Greene (2007).

that the numerator of FE is the difference between the consensus of financial analysts and actual EPS for each quarter. Hence, a positive FE indicates that financial analysts overestimate EPS. As a result, a short position might be profitable and vice versa. We restricted our analysis to trading strategies using FE because our tests presented below showed that the prediction of FE by the CorpInterlock algorithm outperformed the prediction of CAR. The trading strategies that we test are the following:

I Take only long positions for negative FE.
II Take long and short positions for negative and positive FE respectively.
III Take long and short positions for negative and positive FE respectively only for the most precise decile.
IV Take only long positions for negative FE when analysts predict that earnings will be larger than consensus.

We evaluated the results of the trading strategies using the Sharpe ratio. The Sharpe ratio is a risk adjusted return indicator calculated as the mean of cumulative abnormal return divided by its standard deviation.

We expect that the "long only" strategies (I and IV) perform better than the "long/ short" strategies (II and III) because the former strategies are based on the direction of the social networks indicators. If these indicators are not very strong, it does not necessarily mean that the stock will not perform well. Trading strategy III assumes that to take a trading position, signals should be above a certain threshold or actions should be taken only based on the most precise forecasts regardless of the sign of the prediction. Trading strategy IV limits the "long only" strategy to those cases where analysts anticipate that earnings will be larger than consensus. This last strategy reinforces the positive selection of analysts with the forecast capability of our algorithm. The shortcoming with these last two strategies is that the number of observations might be reduced substantially affecting the Sharpe ratio.

We split the presentation of our results before and after 2001 because in October 23, 2000, the SEC issued regulation Fair Disclosure (FD). This regulation requires that companies disseminate material information evenly, without giving any preferences to any investor or analyst. Critics of this regulation indicated that market volatility may increase and the volume of information disseminated in the market will be reduced. However, Lee et al. (2004) neither find any significant increase in volatility, nor an increase in certain components of the bid-ask spread around new releases as a result of regulation FD. During the year 2001 there were also a significant numbers of IPOs, mergers, and acquisitions that were affected by the presence of analysts; it was also the last "bullish" year of the internet "bubble", and also after this year the market became more regulated.

6 Results

The "small world" ratio for the basic and extended corporate interlock is much larger than one according to Table 1. Hence, both corporate interlocks are clearly considered to be of the "small world" type as Davis et al. (2003) found for the Fortune 500 companies. Even though we did not use the "small world" ratio as a predictor, the above results confirm that directors and analysts belong to a “small world” network characterized by the connection among its members by a very short chain of acquaintances. As the “small world” properties of the extended corporate interlock are stronger than those observed in the basic corporate interlock, we think that the strength of these corporate interlocks might be the result of the relationships of different individuals that interact among several firms and boards, and not the result of a small central group that tries to control the society as was proposed by Mills (1956).

We also find that the average distance among boards is stable during a 10 years period, although the increase from 4.1° to 4.85° is slightly larger than the one observed by Davis et al. (2003) during the eighties and nineties. The extended corporate interlock shows a similar increase in the average distance from 3.2° to 3.75°. The indicator that shows a major change is the average degree of the extended corporate interlock which decreases about 50% between the periods 1996–2000 and 2002–2005. During the year 2001, there is a major jump of this indicator which might be explained by the importance of analysts in the last period of the internet bubble. In May 10, 2002 the Securities and Exchange Commission (SEC) approved the rule 2711 “Research Analysts and Research Reports” issued by the National Association of Securities Dealers (NASD), and the rule 472 “Communications with the Public” issued by the New York Stock Exchange (NYSE). These rules establish that no research analyst might be controlled by a firm’s investment banking department. It also shows that the company that is subject of the report can review the report only for factual accuracy checks. This additional regulations may explain the significant reduction of average degree in the period 2002–2005.

The implementation of CorpInterlock using Logitboost with all variables shows a significantly lower test error than its implementation using logistic regression, our baseline algorithm, and Adaboost. The Logitboost implementation shows similar test errors than the implementation of CorpInterlock using random forests. Additionally, the test errors for the prediction of FE are much lower than those observed for the prediction of CAR (see Table 2). Based on these results, we decided to limit our analysis to the implementation of the CorpInterlock algorithm with Logitboost to predict FE.

The regression analysis for the prediction of FE (Table 5) shows a higher adjusted R-square (0.43) for the extended corporate interlock in relation to the basic corporate interlock (0.37) during the period 1996–2001. This advantage disappear or is reversed during the period 2002–2005. In all cases, the *p*-value of the F-statistics is highly significant indicating that the model has explanatory power. Additionally, Fig. 3 shows that the plot of the majority of residuals against forecasts for the extended corporate interlock using all variables during the period 1996–2001 follows a horizontal line in the graph indicating equality of variance or homoscedasticity. A similar behavior is observed in the rest of the cases.

The most important variables in the prediction of FE using the extended corporate interlock are lagged cumulative abnormal return for the preceding 6 months (CAR1) and for the second preceding 6 months (CAR2), total accruals to total assets (TA), size, the lagged of the number of analysts predicting that earnings surprise increase (ANFORLAG), the lagged value of FE (FELAG), consensus, betweenness centrality,

Table 1 Social network indicators for the corporate interlock of total US stock market. C is average of clustering coefficient

Year	All companies	Number of companies in strong component (n)	C	Degree centrality	Betweenness centrality	Closeness centrality	Degree (k)	Distance (L)	SW
(a) Extended Corporate Interlock for US stock market									
1996	3815	3702	0.578	0.009	0.001	0.317	46.103	3.202	31.060
1997	4123	3997	0.570	0.008	0.001	0.317	45.237	3.197	34.249
1998	4215	4089	0.544	0.008	0.001	0.313	42.180	3.240	36.207
1999	4132	4022	0.587	0.016	0.000	0.347	80.614	2.940	18.835
2000	3875	3746	0.561	0.013	0.001	0.335	61.184	3.042	22.586
2001	3521	3383	0.651	0.051	0.000	0.392	189.706	2.647	6.798
2002	2863	2635	0.505	0.007	0.001	0.271	21.986	3.781	40.767
2003	2835	2646	0.539	0.008	0.001	0.277	23.153	3.689	41.868
2004	3042	2882	0.549	0.007	0.001	0.274	23.854	3.695	45.101
2005	3043	2878	0.566	0.007	0.001	0.271	23.887	3.748	45.670
(b) Basic Corporate Interlock for US stock market									
1996	3815	1202	0.241	0.007	0.003	0.250	8.899	4.115	25.699
1997	4123	1303	0.234	0.006	0.003	0.241	8.428	4.282	28.384
1998	4215	1477	0.227	0.005	0.002	0.237	8.367	4.338	31.765
1999	4132	1505	0.226	0.005	0.002	0.235	8.412	4.370	31.815
2000	3875	1430	0.229	0.005	0.002	0.230	7.734	4.469	33.630
2001	3521	1460	0.230	0.004	0.002	0.222	7.364	4.636	35.900
2002	2863	1133	0.220	0.006	0.003	0.229	7.080	4.502	28.064
2003	2835	1154	0.203	0.005	0.003	0.227	6.865	4.531	27.497
2004	3042	1150	0.198	0.005	0.003	0.221	6.346	4.637	29.479
2005	3043	1123	0.197	0.005	0.003	0.212	5.726	4.846	32.046

Last column is the "small world" ratio for the US stock market calculated as $SW = \frac{C}{L} \cdot \frac{L_{random}}{C_{random}}$, where $L_{random} \approx \frac{ln(n)}{ln(k)}$ and $C_{random} \approx \frac{k}{n}$

Table 2 Mean of test errors for learning algorithms by CAR and FE

	Total		CAR		FE	
	Mean (%)	SD (%)	Mean (%)	SD(%)	Mean (%)	SD (%)
Logistic regression	39.44**	9.59	48.33	2.92	30.54**	3.88
Random forests	33.11	14.66	47.49	2.81	18.73	1.59
Adaboost	33.87*	13.92	47.47	2.74	20.27**	2.25
Logitboost	33.33	14.58	47.56	2.97	19.09	2.35

*, ** Represent significance levels of 1% and 5% respectively for the paired t-test of the difference between test errors among each algorithm and Logitboost

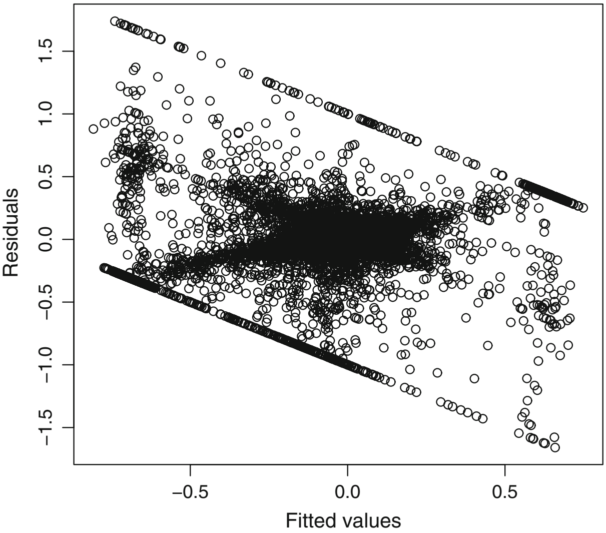

Fig. 3 Residuals plotted against FE fitted values for extended corporate interlock using all variables, 1996–2001

and closeness centrality.[13] Most of the economic variables mentioned above that are associated with FE are either lagged variables (CAR1, CAR2, and FELAG) or reflect the peer effect of financial analysts (ANFORLAG and consensus). This is not surprising if we take into account that FE is based on analysts' expectations. Additionally, according to the "earnings game" companies that have shown earnings surprises in the past or analysts that have predicted earnings surprise in the past may also have similar trends in the future as long as there are still players participating in this game. So, variables that are able to capture expectations of analysts or the peer effect among analysts have a higher predictive power than the rest.

The trading strategies based on FE show that the predictions using the extended corporate interlock lead to a higher risk-adjusted return (Sharpe ratio) than those using the basic corporate interlock or only economic variables (Table 3) during the period 1998–2001. This advantage is not maintained during the period 2002–2005 as was

[13] These last two variables are only relevant during the period 1996–2001.

Table 3 Sharpe ratio using Logitboost forecast of FE for US stock market for the following trading strategies: long and short portfolio when FE expected is −1 and 1 respectively (panel a); long only when FE expected is −1 (panel b); long and short portfolio when FE expected is −1 and 1 respectively only for the most exact decile (panel c), and long only when FE expected is −1 and analysts predict earnings larger than consensus (panel d). Sharpe ratio is calculated as the mean of abnormal returns during a complete year divided by its standard deviation. Abnormal returns used for calculation are the mean of monthly abnormal returns for each year. "Econ. only" stands for Economic variables only; "Extended net" stands for Extended Corporate Interlock, and "Basic net" for Basic Corporate Interlock

	Econ. only	Extended net	Basic net
(a) Long/short strategy			
1998–2001	1.51	1.72	1.60
2002–2005	2.08	2.07	2.06
1998–2005	1.74	1.84	1.78
(b) Long only strategy			
1998–2001	1.86	1.87	1.74
2002–2005	2.05	2.05	2.05
1998–2005	1.99	1.99	1.91
(c) Long/short strategy. Top decile			
1998–2001	1.01	1.10	0.80
2002–2005	1.09	0.89	0.95
1998–2005	1.08	1.01	0.90
(d) Long only strategy when analysts predict that earnings are larger than consensus			
1998–2001	1.22	1.20	1.19
2002–2005	1.50	1.49	1.48
1998–2005	1.39	1.37	1.38

also observed in the regression analysis. The "long only" strategy is the most risk-adjusted profitable strategy (Tables 3, 4; panel b). In all cases, the trading strategies generate significant abnormal returns according to the t-statistic (Table 4) and the Sharpe ratio is larger during the period 2002–2005. We think that the accumulation of additional years of training improve the trading and forecasting capability of the algorithm.

When we only use the cases where analysts predict that earnings are larger than consensus (Tables 3, 4; panel d), the difference between the Sharpe ratios of different sets is very small. However, the abnormal return (Table 4) is larger than the return of the other trading strategies.

7 Discussion

Our results indicate that the CorpInterlock algorithm leads to profitable trading strategies forecasting the FE during all the years under study. This finding can be explained if we consider that many fund managers or their representatives have influence or even have a seat or more in the board of the corporations where they invest. Hence, they can use their knowledge about the financial health of the companies where they have some presence to optimize their portfolios. Additionally, institutional investors have access to their own research team and could maintain certain independence of the analysts' influence. They are able to deeply evaluate the companies in which they are interested

Table 4 Abnormal return using Logitboost forecast of FE for US stock market for the following trading strategies: long and short portfolio when FE expected is −1 and 1 respectively (panel a); long only when FE expected is −1 (panel b); long and short portfolio when FE expected is −1 and 1 respectively only for the most exact decile (panel c), and long only when FE expected is −1 and analysts predict earnings larger than consensus (panel d)

	Econ. only (%)	Extended net (%)	Basic net (%)
(a) Long/short strategy			
1998–2001	1.17**	1.22**	1.19**
2002–2005	1.78**	1.77**	1.78**
1998–2005	1.47**	1.49**	1.48**
(b) Long only strategy			
1998–2001	2.90**	2.94**	2.91**
2002–2005	2.60**	2.59**	2.60**
1998–2005	2.75**	2.77**	2.75**
(c) Long/short strategy. Top decile			
1998–2001	2.79**	2.52*	2.14*
2002–2005	2.61*	2.44*	2.45*
1998–2005	2.70**	2.48**	2.30**
(d) Long only strategy when analysts predict that earnings are larger than consensus			
1998–2001	3.46**	3.45*	3.21*
2002–2005	3.40**	3.36**	3.39**
1998–2005	3.43**	3.40**	3.31**

"Econ. only" stands for Economic variables only; "Extended net" stands for Extended Corporate Interlock, and "Basic net" for Basic Corporate Interlock. Returns are monthly

*,** Represent significance levels of 1% and 5% respectively for the t statistic

in investing. Therefore, they have an understanding of the fundamental valuation of the companies where they invest regardless of the day to day market speculation. This fact explains that even though our algorithm is able to improve the forecast of the trend of FE in relation to logistic regression, our baseline algorithm, the inclusion of the social network information improves the prediction of FE only in the period 1998–2001. The main explanation is that the period 1998–2001 corresponds to the last part of the internet "bubble". During this period, stock prices increased very quickly and the valuation multiples such as price-to-earnings ratio of technology companies like YAHOO were much higher than what a fundamental analysis would indicate. Many individual investors were participating in the market, and even small investors left their regular jobs to become full-time day traders. An important source of information for these investors was the forecast of the analysts (consensus). Suddenly, technology analysts became stars and were interviewed in popular shows. Their opinions were able to influence the market and therefore the returns, while fundamental or value investors had less importance. Additionally, analysts were also hired by investment banks that were participating in new deals such as IPOs, mergers, and acquisitions. Analysts had a strong pressure from the investment bankers to favorably cover companies where they expected to have a new deal or already had one. Also, if an analyst was covering a company that was merged or acquired another company, suddenly she expanded her coverage to a new company or even a new industry, if the company was trying to diversify itself. For example, Microsoft has grown through acquisitions and has significantly expanded its initial area of economic activity as "software developer".

The analysts of Microsoft have to understand the new business operations. This latter idea also explains why in 2001 there is such an unusual increase in the degree of the extended corporate interlock of the US market as Table 1 shows.

The relationship between analysts and directors is partially explained by the "earnings game" that we introduced in Sect. 1. The value of the stock options of CEO's and senior managers depends on the earnings surprises. Managers try to reach or improve the analysts' predictions. At the same time, analysts need the investment banking business because their compensation might be based on it. As a result there are incentives on both sides to find a mutually satisfying prediction and selective disclosure of material information to analysts. If the same game is played in several companies with various common directors, then the inclusion of analysts in the social network of directors may increase the profitability of trading strategies formulated around FE during the period 1998–2001. However, the degree of the extended corporate interlock and the profitability of its associated trading strategies are reduced in the period 2002–2005. This contraction might be explained because of the regulations introduced by the Sarbanes-Oxley Act and the regulation FD that prohibits selective disclosure of material information.

The results of our trading strategies is consistent with the results of Cohen et al. (2008) who show the profitability of a trading strategy that takes a long position with "buy" recommendations and school ties among analysts and directors, and a short position with "buy" recommendations without school ties only until regulation FD was established. Regulation FD enforces the fair disclosure of material information to all interested parties. Hence, the direct link that Cohen et al. (2008) show when directors and analysts share the same "Alma mater" or—according to our research—when directors and analysts are connected through several companies indicate that these connections allowed analysts to receive privileged information that may have improved their predictions. In our simulations of trading strategies, the Sharpe ratio is smaller when we use either the top decile of predictions or only those cases where analysts predict that earnings are larger than consensus, however their abnormal returns are higher than those of other strategies. The main explanation for this apparent contradiction is that the selection of the above sample significantly reduce the number of observations. Hence, their standard deviations increase and their Sharpe ratios decrease. The trading strategy where analysts predict that earnings are larger than consensus show the highest abnormal return of all the trading strategies. In this case, the importance of social network indicators is less relevant because there is already a combination of the selection of those companies with the best analysts' forecast and the restriction that analysts are part of the major strongly connected network similar to the results of Cohen et al. (2008). A possible explanation is that the best companies to invest attract a large group of analysts which help to strength the social network of directors and analysts and to transmit information not only about earnings, but also about competitors' plans or strategies. This additional information may help the board to take more informed decisions and as a consequence, improve corporate performance.

The problem explored in this paper has some similarities with the direct marketing problem that Hill et al. (2006) approached. In both cases, the algorithm must make a decision to invest in a prospect (mail an offer) or in a stock (long or short position), and the prediction is improved when very well-known marketing or investment variables are combined with social network properties. Cohen et al. (2008), Hill et al. (2006) and Creamer and Stolfo (2006) show that social networks have a significant effect in how people take investment decisions. Acquiring a new credit card, expanding phone services, recommending a firm or taking a long position in a stock are decisions that might be affected by the social network affiliation of the economic agent. The details of how the information transmission process happens might be different. In the direct marketing case, the continuous interaction between neighbors or members of the same household may explain that it is easier to expand services or acquire prospects when they are affiliated to loyal customers because loyal customers may act as diffusion agents. In the extended corporate interlock case, the school ties (Cohen et al. 2008) might explain the relevance of the social network variables as predictors, however the social network effect might be due to more recent associations such as professional networks or the attraction of similar people to same firms. Many corporations choose directors that are highly connected in the industry because of the additional information or business that these connections may bring. Hence, top 500 Fortune companies may have very well-known directors such as Vernon Jordan who is in the top ten list of the most connected directors in the study of Davis et al. (2003). This bias of companies to select highly connected directors is confirmed by our regression analysis (see Table 5) which shows that the most important social network variables for the period 1996–2001 are betweenness centrality and closeness centrality. The most successful analysts of these companies might also be very well connected either by education or social status. If this is the case, the likelihood that analysts and directors know each other, interact and exchange information in professional or social events or clusters increase. This interaction among directors and analysts may determine that some companies are highly connected with other companies by very few degrees (closeness centrality) or that they become "bridges" to facilitate the connection among many other firms (betweenness centrality).

8 Conclusions

The link mining algorithm, CorpInterlock, demonstrated to be a flexible mechanism to increase the profitability of trading strategies using social network indicators. The capacity to improve the forecast of earnings surprises and abnormal return using a mixture of well-known economic indicators with organizational and behavioral variables also enriches the debate between the modern finance theory and behavioral finance to show how behavioral patterns can be recognized under a rigorous method of analysis and forecast.

The basic and extended corporate interlocks have the properties of a "small world" network. However, the expansion of the original corporate interlock to include new actors, such as financial analysts, bring additional information especially during a "bull" market that leads to profitable trading strategies.[14]

[14] A "bull" market is a market where prices are increasing or there is the expectation that they will increase. "Bear" market is the opposite.

Table 5 Results of regression model for the periods 1996–2001 (panel a) and 2002–2005 (panel b) using FE as the dependent variable and the following independent variables: 1. only economic variables, and 2. economic and social network variables (All)

Variables	Extended Corp. Interlock		Basic Corp. Intelock	
	Economic var.	All	Economic var.	All
1996–2001				
CAR1	0.035	0.035	0.024	0.024
	(6.854)***	(6.967)***	(2.658)**	(2.564)*
CAR2	−0.025	−0.026	0.008	0.007
	(−4.991)***	(−5.354)***	(0.886)	(0.871)
SIZE	0.012	0.011	0.008	0.009
	(9.526)***	(−6.848)***	(4.146)***	(4.219)***
FREV	−0.045	−0.041	−0.115	−0.116
	(−1.809)	(−1.643)	(−2.276)*	(−2.304)*
LTG	0.000	0.000	0.000	0.000
	(−0.363)	(−0.087)	(0.255)	(−0.073)
SUE	0.000	0.000	0.000	0.000
	(−0.103)	(−0.139)	(0.119)	(0.128)
SG	0.000	0.000	0.000	0.000
	(−0.215)	(−0.206)	(−0.380)	(−0.371)
TA	−0.062	−0.065	−0.031	−0.031
	(−3.687)***	(−3.840)***	(−1.246)	(−1.245)
CAPEX	0.069	0.066	0.005	0.007
	(1.942)	(1.849)	(0.109)	(0.145)
BP	0.000	0.000	0.000	0.000
	(−0.083)	(0.009)	(1.402)	(1.399)
EP	0.000	0.000	0.000	0.000
	(0.355)	(0.293)	(−0.604)	(−0.591)
ANFOR	0.001	0.001	0.002	0.002
	(1.257)	(1.296)	(1.435)	(1.377)
ANFORLAG	0.005	0.006	0.004	0.004
	(5.975)***	(6.219)***	(4.339)***	(4.337)***
FELAG	0.646	0.645	0.594	0.594
	(84.612)***	(84.481)***	(50.159)***	(50.109)***
CONSENSUS	−0.001	−0.001	−0.001	−0.001
	(−3.423)***	(−3.321)***	(−0.874)	(−0.866)
CC'		−0.011		−0.001
		(−1.108)		(−0.058)
deg		0.007		−0.918
		(0.062)		(0.534)
B_c		−5.417		0.388
		(−2.768)**		(0.221)
C_c		0.253		−0.042
		(2.731)**		(−0.277)
Adj. R square	0.432	0.433	0.371	0.371
p-value (F-stat.)	0.000	0.000	0.000	0.000
2002–2005				
CAR1	0.036	0.036	0.036	0.036
	(3.490)***	(3.505)***	(2.514)*	(2.517)*
CAR2	−0.004	−0.004	0	0.000
	(−0.473)	(−0.476)	(−0.018)	(0.036)
SIZE	0.004	0.003	0.003	0.004
	(2.133)*	−1.545	−1.327	(1.324)

Table 5 continued

Variables	Extended Corp. Interlock		Basic Corp. Intelock	
	Economic var.	All	Economic var.	All
FREV	−0.097	−0.097	−0.114	−0.115
	(−3.360)***	(−3.354)***	(−1.394)	(−1.413)
LTG	0	0	0.001	0.001
	(−0.463)	(−0.391)	(2.420)*	(2.337)*
SUE	0.002	0.002	0.004	0.004
	−0.587	−0.601	−1.316	(1.310)
SG	0	0	0	0.000
	(−0.158)	(−0.155)	(−0.142)	(−0.170)
TA	0.009	0.01	0.037	0.036
	−0.498	−0.516	−1.379	(1.356)
CAPEX	−0.047	−0.047	−0.059	−0.059
	(−0.831)	(−0.829)	(−0.860)	(−0.862)
BP	0	0	−0.001	−0.001
	(−0.493)	(−0.491)	(−1.315)	(−1.244)
EP	0	0	0.063	0.063
	−0.24	−0.244	(4.418)***	(4.369)***
ANFOR	0	0	0	0.000
	(−0.143)	(−0.120)	(−0.127)	(−0.077)
ANFORLAG	0.003	0.003	0.002	0.002
	(3.855)***	(3.896)***	(3.549)***	(3.572)***
FELAG	0.661	0.661	0.69	0.690
	(63.097)***	(63.071)***	(49.831)***	(49.823)***
CONSENSUS	−0.004	−0.004	−0.004	−0.004
	(−0.911)	(−0.951)	(−1.009)	(−0.947)
CC'		−0.014		0.006
		(−1.316)		(0.451)
deg		0.009		2.317
		−0.015		(1.133)
B_c		−0.14		−1.822
		(−0.077)		(−1.189)
C_c		−0.017		−0.144
		(−0.134)		(−0.833)
Adj. R square	0.444	0.444	0.474	0.473
p-value (F-stat.)	0	0	0	0.000

Models include intercept and dummy variables to control for economic sector of activity which are not included in the table, the rest of variables are included. Economic variables are cumulative abnormal return for the preceding 6 months (CAR1) and for the second preceding 6 months (CAR2) since the earnings announcement day; natural logarithm of market capitalization (SIZE); analysts earnings forecast revisions to price (FREV); mean of analysts' long-term growth forecast (LTG); standardized unexpected earnings (SUE); sales growth (SG); total accruals to total assets (TA); rolling sum of capital expenditures to total assets (CAPEX); book to price ratio (BP); earnings to price ratio (EP); number of analysts predicting that earnings surprise increase (ANFOR) and its lagged value (ANFORLAG); and lagged forecast error (FELAG). Social network variables are clustering coefficient (CC'), degree centrality (deg), betweenness centrality (B_c), and closeness centrality (C_c). Numbers in parentheses are t-statistics

*,**,*** Represent significance levels of 0.1%, 1%, and 5% respectively

The application of link mining algorithms to problems of finance or social sciences may enrich the discussion in two ways: on one hand, a link mining algorithm can contribute to the understanding of social phenomena with the integration of different domains and especially quantifying the network perspective. On the other hand, the complex social problems offer scenarios to tests the adequacy or the development of new algorithms to solve interdisciplinary problems. For example, the oil supply is controlled by rich-oil countries with authoritarian or autocratic governments. A link mining algorithm may help to integrate the different domains in play: political, social, economical and cultural, and to find links that may bring new solutions to old problems.

A future line of research is the extension of the CorpInterlock algorithm to problems of direct marketing that showed to be very similar to the investment questions explored in this paper. The key problem is to quantify the main social network that offer additional information about how agents take economic decisions.

Acknowledgments This research was supported by the National Science Foundation grant “Email Mining Toolkit Supporting Law Enforcement Forensic Analysis“ from the Digital Government research program, No. 0429323. GC would also like to thank Patrick Jardine for proof-reading the article; David Waltz, Hubert Delany, Shlomo Hershkop, Zeph Grunschlag, Tony Jebara, Vasant Dhar, and the anonymous referees of this journal and of the “Link Analysis: Dynamics and Statics of Large Networks“ Workshop on SIGKDD 2006 for their valuable comments. The opinions presented are the exclusive responsibility of the authors.

Appendix 1 Investment signals used for prediction

We do not include firm-specific subscripts in order to clarify the presentation. Subscript q refers to the most recent quarter for which an earnings announcement was made. The fundamental variables are calculated using the information of the previous quarter (SUE,SG,TA,and CAPEX) and our notation is similar to the notation used by Jegadeesh et al. (2004).

Variable	Description	Calculation detail
SECTOR	Two-digit sector classification according to the Global Industrial Classification Standards (GICS) code	Energy 10, Materials 15, Industrials 20, Consumer Discretionary 25, Consumer Staples 30, Health Care 35, Financials 40, Information Technology 45 Telecommunication Services 50, Utilities 55
Price momentum		
CAR1	Cumulative abnormal return for the preceding 6 months since the earnings announcement day	$[\Pi_{t=m-6}^{m-1}(1+R_t)-1]$ $-[\Pi_{t=m-6}^{m-1}(1+R_{tw})-1]$, where R_t is return in month t, R_{tw} is value weighted market return in month t, and m is last month of quarter
CAR2	Cumulative abnormal return for the second preceding 6 months since the earnings announcement day	$[\Pi_{t=m-12}^{m-7}(1+R_t)-1]$ $-[\Pi_{t=m-6}^{m-1}(1+R_{tw})-1]$
Analysts variables		
ANFOR (ANFORLAG)	Number of analysts predicting that earnings surprise increase (lagged value)	
CONSENSUS	Mean of earnings estimate by financial analysts	
FELAG	Lagged forecast error	$\frac{\text{CONSENSUS}_q-\text{EPS}_q}{\lvert\text{CONSENSUS}_q\rvert+\lvert\text{EPS}_q\rvert}$ (Dhar and Chou 2001) where *EPS* is earnings per share

Earnings momentum		
FREV	Analysts earnings forecast revisions to price	$\sum_{i=0}^{5} \frac{\text{CONSENSUS}_{m-i} - \text{CONSENSUS}_{m-i-1}}{P_{m-i-1}}$ where P_{m-1} is price at end of month $m-1$, and i refers to the previous earnings revisions
SUE	Standardized unexpected earnings	$\frac{(EPS_q - EPS_{q-4})}{\sigma_t}$ where EPS is earnings per share, and σ_t is standard deviation of EPS for previous seven quarters
Growth indicators		
LTG	Mean of analysts' long-term growth forecast	
SG	Sales growth	$\frac{\sum_{t=0}^{3} \text{Sales}_{q-t}}{\sum_{t=0}^{3} \text{Sales}_{q-4-t}}$
Firm size		
SIZE	Market cap (natural log)	$ln(P_q\ shares_q)$ where $shares_q$ are outstanding shares at end of quarter q
Fundamentals		
TA	Total accruals to total assets	$\frac{\triangle\text{C.As.}_q - \triangle Cash_q - (\triangle\text{C.Lb.}_q - \triangle\text{C.Lb.D}_q) - \triangle\text{T}_q - \text{D\&A}_q}{\frac{(\text{T.As.}_q - \text{T.As.}_{q-4})}{2}}$ where $\triangle X_q = X_q - X_{q-1}$ and C.As., C.Lb., C.Lb.D., T,D&A, and T.As. stands for current assets, current liabilities, debt in current liabilities, deferred taxes, depreciation and amortization, and total assets respectively.
CAPEX	Rolling sum of capital expenditures to total assets	$\frac{\sum_{t=0}^{3} \text{capital expenditures}_{q-t}}{(T.As._q - T.As._{q-4})/2}$
Valuation multiples		
BP	Book to price ratio	$\frac{\text{book value of common equity}_q}{\text{market cap}_q}$, where market $\text{cap}_q = P_q\ \text{shares}_q$
EP	Earnings to price ratio (rolling sum of EPS of the previous four quarters deflated by prices)	$\frac{\sum_{t=0}^{3} EPS_{q-t}}{P_q}$
Social networks		
$deg(v_i)$	*Degree centrality or degree:* number of edges incidents in vertex v_i	$\sum_j a_{ij}$, where a_{ij} is an element of the adjacent matrix A
$C_c(v_i)$	*Closeness centrality (normalized):* inverse of the average geodesic distance from vertex v_i to all other vertices	$\frac{n-1}{\sum_j d_{ij}}$, where d_{ij} is an element of the geodesic distance matrix D (Freeman 1979; Borgatti and Everett 2006)

$B_c(v_i)$	*Betweenness centrality:* proportion of all geodesic distances of all other vertices that include vertex v_i	$\sum_i \sum_j \frac{g_{kij}}{g_{kj}}$, where g_{kij} is the number of geodesic paths between vertices k and j that include vertex i, and g_{kj} is the number of geodesic paths between k and j (Freeman 1979)
CC_i	*Clustering coefficient:* cliquishness of a particular neighborhood or the proportion of edges between vertices in the neighborhood of v_i divided by the number of edges that could exist between them (Watts and Strogatz 1998)	$\frac{2\lvert\{e_{ij}\}\rvert}{deg(v_i)(deg(v_i)-1)} : v_j \in N_i,\ e_{ij} \in E$, where each vertex v_i has a neighborhood N defined by its immediately connected neighbors: $N_i = \{v_j\}: e_{ij} \in E$
CC_i'	Normalized clustering coefficient	$\frac{deg(v_i)}{MaxDeg} CC_i$, where MaxDeg is the maximum degree of vertex in a network (de Nooy et al. 2005)
C (not used for forecasting)	Mean of all the clustering coefficients	$\frac{1}{n} \sum_{i=1}^{n} CC_i$
SW (not used for forecasting)	"Small world" ratio (Watts and Strogatz 1998).	$\frac{C}{L} \frac{L_{random}}{C_{random}}$, where $L_{random} \approx \frac{ln(n)}{ln(k)}$ and $C_{random} \approx \frac{k}{n}$
Labels		
LABELFE	Label of forecast error (FE)	1 if $CONSENSUS \geq EPS$ (current quarter) , -1 otherwise
LABELCAR	Label of cumulative abnormal return (CAR)	1 if $CAR_{m+1} \geq 0$, -1 otherwise, where CAR_{m+1} refers to the CAR of the month that follows the earnings announcement

References

Abarbanell J (1991) Do analysts earnings forecasts incorporate information in prior stock price changes? J Account Econ 14:147–165

Abarnabell J, Bernard V (1992) Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior. J Finance 47:1181–1207

Asquith P, Mikhail MB, Au AS (2005) Information content of equity analyst reports. J Financ Econ 75:245–282

Barabasi A (2002) Linked: the new science of networks. Perseus, Cambridge, MA

Barber B, Lehavy R, McNichols M, Trueman B (2001) Can investors profit from the prophets? Security analysts recommendations and stock returns. J Finance 56:531–563

Beckers S, Steliaros M, Thomson A (2004) Bias in European analysts' earnings forecasts. Financ Anal J 60:74–85

Bernard VL, Thomas JK (1990) Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. J Account Econ 13

Borgatti SP, Everett M (2006) A graph-theoretic perspective on centrality. Soc Netw 28:466–484

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breusch TS, Pagan A (1979) A simple test for heteroscedasticity and random coefficient variation. Econometrica 47:1287–1294

Brown LD (2000) I/B/E/S Research Bibliography, 6th edn. I/B/E/S International Incorporated. http://www2.gsu.edu/~wwwacc/Faculty/lbrown/Bibliography.pdf

Brown LD (2001) How important is past analyst forecast accuracy? Financ Anal J 57:44–49

Brown LD, Han JCY, Keon EF Jr, Quinn WH (1996) Predicting analysts' earnings surprise. J Invest 5:17–23

Cessie SL, Houwelingen JCV (1992) Ridge estimators in logistic regression. Appl Stat 41:191–201

Clement M, Tse S (2005) Financial analyst characteristics and herding behavior in forecasting. J Finance 40:307–341

Cohen L, Frazzini A, Malloy C (2008) Sell side school ties. Working paper, Harvard Business School

Collins M, Schapire RE, Singer Y (2004) Logistic regression, adaboost and Bregman distances. Mach Learn 48:253–285

Creamer G, Freund Y (2004) Predicting performance and quantifying corporate governance risk for latin american adrs and banks. In: I Proceedings of the financial engineering and applications conference, MIT-Cambridge

Creamer G, Freund Y (2005) Using adaboost for an equity investment/board balanced scorecard. In: Machine learning in finance workshop in NIPS 2005, Whistler, B.C

Creamer G, Freund Y (2007) A boosting approach for automated trading. J Trading (Summer 2007):84–95

Creamer G, Stolfo S (2006) A link mining algorithm for earnings forecast using boosting. In: Proceedings of the link analysis: dynamics and statics of large networks workshop on international conference on knowledge discovery and data mining (KDD), Philadelphia, PA

Davis CE, Hyde JE, Bangdiwala S, Nelson J (1986) Modern statistical methods in chronic disease epidemiology, chapter An example of dependencies among variables in a conditional logistic regression. Wiley, New York

Davis G (1991) Agents without principles? The spread of the poison pill through the intercorporate network. Adm Sci Q 36:586–613

Davis G, Yoo M, Baker W (2003) The small world of the american corporate elite, 1982–2001. Strateg Organ 1:301–326

de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York

Dhar V, Chou D (2001) A comparison of nonlinear methods for predicting earnings surprises and returns. IEEE Trans Neural Netw 12:907–921

Domingos P, Richardson M (2001) Mining the network value of customers. In: KDD '01: proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp 57–66

Elton JE, Gruber MJ, Grossman S (1986) Discrete expectational data and portfolio performance. J Finance 41:699–714

Fawcett T, Provost F (1999) Activity monitoring: noticing interesting changes in behavior. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99), pp 53–62

Finger CA, Landsman WR (1999) What do analysts' stock recommendations really mean? Working paper, University of Illinois and U.N.C., Chapel Hill

Freeman L (1979) Centrality in networks: I. conceptual clarification. Soc Netw 1:215–239

Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In: Machine learning: proceedings of the sixteenth international conference, pp 124–133

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comp Sys Sci 55:119–139

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 38:337–374

Getoor L, Diehl CP (2005) Link mining: a survey. SIGKDD Explorations 7:3–12

Goldberg HG, Kirkland JD, Lee D, Shyr P, Thakker D (2003) The NASD securities observation, news analysis and regulation system (sonar). In: IAAI 2003, Acapulco, Mexico

Greene W (2007) Econometric analysis, 6th edn. Prentice Hall, Upper Saddle River, NJ

Hill S, Provost F, Volinsky C (2006) Network-based marketing: identifying likely adopters via consumer networks. Stat Sci 21:256–276

Hong HG, Kubik JD (2003) Analyzing the analysts: career concerns and biased earnings forecasts. J Finance 58:313–351

Ivkovic Z, Jegadeesh N (2004) The timing and value of forecast and recommendation revisions. J Financ Econ 73:433–463

Jegadeesh N, Kim J, Krische SD, Lee CMC (2004) Analyzing the analysts: when do recommendations add value? J Finance 59:1083–1124

Kirkland JD, Senator TE, Hayden JJ, Dybala TG, Goldberg H, Shyr P (1999) The nasd regulation advanced detection system (ads). AI Mag 20:55–67

Krische SD, Lee CMC (2000) The information content of analyst stock recommendations. Working paper, Cornell University

Larcker DF, Richardson SA, Seary AJ, Tuna I (2005) Back door links between directors and executive compensation. Working paper

Lee CI, Rosenthal L, Gleason KC (2004) Effect of regulation FD on asymmetric information. Financ Anal J 60:79–89

Leskovec J, Adamic LA, Huberman BA (2006) The dynamics of viral marketing. In: EC ’06: proceedings of the 7th ACM conference on electronic commerce, pp 228–237, ACM, New York, NY, USA

Mendenhall RR (1991) Evidence on the possible underweighting of earnings information. J Account Res 29:170–179

Mikhail MB, Walther B, Willis R (2002) Do security analysts exhibit persistent differences in stock picking ability? J Financ Econ 74:67–91

Milgram S (1967) The small world problem. Psychol Today 2:60–67

Mills C (1956) The power elite. Oxford Press, New York

Mintz B, Schwartz M (1985) The power structure of American business. University of Chicago Press, Chicago

Mizruchi M (1992) The structure of corporate political action: interfirm relations and their consequences. Harvard University Press, Cambridge, MA

Moreno J (1932) Application of the group method to classification. National committee on prisons and prison labor, New York

Newman M, Strogatz S, Watts D (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64

Newman MEJ, Watts DJ, Strogatz SH (2002) Random graph models of social networks. Proc Natl Acad Sci USA 99(Suppl 1):2566–2572. doi:10.1073/pnas.012582999

Ou JA, Penman SH (1989) Accounting measurement, price-earnings ratios, and the information content of security prices. J Account Res 27

Peters D (1993a) Are earnings surprises predictable? J Invest 2:47–51

Peters D (1993b) The influences of size on earnings surprise predictability. J Invest 2:54–59

Peterson D, Peterson P (1995) Abnormal returns and analysts earnings forecast revisions associated with the publication of ’stock highlights’ by value line investment survey. J Financ Res 18:465–477

Rao H, Davis G, Ward A (2000) Embeddedness, social identity and mobility: why firms leave the NASDAQ and join the New York Stock Exchange. Adm Sci 45:268–292

Richardson M, Domingos P (2006) Markov logic networks. Mach Learn 62:107–136

Senator TE (2005) Link mining applications: progress and challenges. SIGKDD Explor 7:76–83

Sparrow M (1991) The application of network analysis to criminal intelligence: an assessment of the prospects. Soc Netw 13:251–274

Stickel SE (1995) The anatomy of the performance of buy and sell recommendations. Financ Anal J 51:25–39

Stober T (1992) Summary financial statements measures and analysts’ forecasts of earnings. J Account Econ 15:347–372

Stolfo S, Creamer G, Hershkop S (2006) A temporal based forensic discovery of electronic communication. In: Proceedings of the national conference on digital government research, San Diego, California

Thaler R (2005) Advances in behavioral finance II. Princeton University Press, Princeton, NJ

Watts D (1999) Networks, dynamics, and the small-world phenomenon. Am J Sociol 105:493–527

Watts D, Strogatz S (1998) Collective dynamics of small world networks. Nature 393:440–442

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco

Womack K (1996) Do brokerage analysts’ recommendations have investment value? J Finance 51:137–167