A link mining algorithm for earnings forecast using boosting

Germán Creamer  
Center for Computational Learning Systems  
Columbia University  
475 Riverside MC 7717  
New York, NY 10115  
gcreamer@cs.columbia.edu

Sal Stolfo  
Department of Computer Science  
Columbia University  
500 W 120 St  
New York, NY 10027  
sal@cs.columbia.edu

ABSTRACT
The objective of this paper is to present and discuss the results of a link mining algorithm called CorpInterlock that integrates the metrics of an extended corporate interlock (social network of directors and financial analysts) with corporate fundamental variables and analysts’ predictions (consensus) in order to forecast the trend of the cumulative abnormal return and earnings surprise using the boosting approach. 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 or business of the company that they follow.

We found that the basic and extended corporate interlock of the US stock market has the properties of a “small world” network. Based on this, we calculated a group of well-known social network metrics and integrated with economic variables using alternating decision trees (ADTs) implemented with Logitboost. We observed a significant reduction of the test error of the experiments when we used the extended corporate interlock instead of either the basic corporate interlock with fundamental variables and consensus or when only fundamental variables and consensus were included, during a “bull” market (1997-2001). The basic corporate interlock showed to be more effective during a “bear” market (2002-2003).

Categories and Subject Descriptors
H.4.2 [Information Systems]: Information Systems Applications Type of systems|Decision support

General Terms
Economics, Management, Algorithms

Keywords
Link mining, link analysis, social network, machine learning, computational finance, boosting, time series, pattern analysis, data mining applications

1. INTRODUCTION
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 ([32, 38, 4, 23, 1, 34, 33]). Brown et al [6] standardized a method to calculate the “earnings surprise” with an indicator that they call “earnings surprise predictor” (ESP). This “earnings surprise predictor” outperforms the market using a portfolio of S&P 500 companies during the period 1985-1994. We believe that recent developments in the area of machine learning and link mining can contribute to 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 inefficient and move on individual biases or behavioral patterns [40]. 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.

1.1 Earnings surprise

Several studies have also 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. [3] 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. [24] 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 five-years of superior performance in rankings surveys such as those collected by The Wall Street Journal. Jegadeesh et al. [20] 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 who 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).

1.2 Social networks and company interlock

The application of networks to social science has a long tradition since the seminal works of Moreno [29] and Milgram [25] 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.

Several networks in the social and natural sciences have been identified to have the properties of a “small world” (see [42, 2]). We are particularly interested in those organizational studies about the corporate interlock or the social network of directors of major corporations. Davis et al. [10] have found that the interlock network structure of the major US corporations (those in the Fortune 500 list) between 1982 and 1999 has the characteristics of a “small world” as described in section 2. These authors also find that the interlock network is highly stable, even after major changes in corporate governance. Mintz and Schwartz [27], following Mill [26]’s thesis, study how commercial banks have a central position in the interlock network because of the participation of the major leaders of US nonfinancial corporations on the banks’ boards. The original thesis of Mills 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.

Larcker et al. [22] 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. 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 [28], 2. poison pills [9], and 3. switching from NASDAQ to NYSE [35].

From our perspective, we do not know of any previous research that has used social network indicators combined with economic determinants to forecast cumulative abnormal returns (CAR) and earnings surprises or the forecast error (FE). We think that if the corporate interlock plays such an important role in corporate governance, it may also have an impact to forecast earnings surprises and CAR.

In this research we refer to the social network among directors as the basic corporate interlock, the social network among directors and analysts as the extended corporate interlock, and cumulative abnormal return as the return of a specific asset less the average return of all assets in its risk-level portfolio for each trading date.

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. 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 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 earnings surprises better than cumulative abnormal return because of the relationship among directors and analysts that may explain earnings surprises.

Considering the existence of the “earnings game”, our objectives in this paper are: a) evaluate whether the basic and extended corporate interlock (directors and analysts) of the US stock market has the properties of a “small world” network; b) evaluate the contribution of social network indicators of the basic and extended corporate interlock to predict the trend of PE and CAR, and c) present and test an algorithm to integrate analysts’ predictions, and economic and social network indicators using the boosting approach.

The rest of the paper is organized as follows: Section 2 describes the “small world” model; section 3 presents the methods used to forecast the stock market: a link mining algorithm, and boosting; section 4 explains in detail our forecasting strategy; section 5 presents the results of our fore-
2. SMALL WORLD

Watts [41] and Watts et al. [42, 30, 31] 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 $V = v_1, v_2, ..., v_n$ 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: $C = \frac{1}{n} \sum_{i=1}^{n} CC_i$, where:
  - $CC_i = \frac{2\delta(e_{ij})}{\delta(v_i)^2} : e_{ij} \in N_i$, $e_{ij} \in E$.
  - $\delta(e_{ij})$ is the degree centrality or degree of a vertex $v_i$: $\delta(v_i) = \sum_i a_{ij}$.
  - $a_{ij}$ 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: $L = \frac{1}{n} \sum_i d_{ij}$, where $d_{ij} \in D$ and $D$ is the geodesic distance matrix of $G$. In the case of a random network, these structural indicators are $L_{random} \approx \ln(n)$ and $C_{random} \approx \frac{\ln(\ln(n))}{\ln(n)}$.

Using the above indicators, the four properties that characterize a “small world” network are:
1. $n$ is fixed and numerically large $n \gg 1$.
2. $k$ is fixed so that $G$ is sparse ($k \ll n$), and with a minimum number of potential structures ($k \gg 1$).
3. $G$ is decentralized. So, there is not a single dominant vertex: $k_{max} \ll n$ where $k_{max}$ is the maximal degree.
4. $G$ must be strongly connected.

$C$ works as a measure of order in $G$, where if $C >> k/n$, then $G$ is considered locally ordered, while random graphs are not ordered and therefore $C_{random}$ is very small as the above property 2 ($k \ll n$) 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 ($SW = L_{random} / \ln(n)$) is much larger than one.

Other additional indicators of social networks that we have used in this study are:
1. Closeness centrality (normalized): $C_{c}(v_i) = \frac{1}{\sum_{ij} d_{ij}}$, where $d_{ij}$ is an element of the geodesic distance matrix $D$ [14, 5].
2. Betweenness centrality $B_{s}(v_i) = \sum_{ij} \frac{g_{kij}}{g_{ij}}$. 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 $g_{ij}$ is the number of geodesic paths between $k$ and $j$ [14].
3. Normalized clustering coefficient: $CC'_i = \frac{\delta(e_{ij})}{\delta(v_i)} CC_i$, where MaxDeg is the maximum degree of vertex in a network [11].

\[ C = \frac{1}{k} \sum_{i=1}^{k} CC_i, \quad CC_i = \frac{2\delta(e_{ij})}{\delta(v_i)^2} : e_{ij} \in N_i, e_{ij} \in E. \]

\[ L = \frac{1}{n} \sum_i d_{ij}, \quad d_{ij} \in D, \quad D = \text{the geodesic distance matrix of } G. \]

\[ C_{c}(v_i) = \frac{1}{\sum_{ij} d_{ij}}, \quad B_{s}(v_i) = \sum_{ij} \frac{g_{kij}}{g_{ij}}. \]

\[ CC'_i = \frac{\delta(e_{ij})}{\delta(v_i)} CC_i, \quad \text{where MaxDeg is the maximum degree of vertex in a network [11].} \]

3. METHODS

3.1 Boosting

Adaboost is a machine learning algorithm invented by Freund and Schapire [16] that classifies its outputs applying a simple learning algorithm (weak learner) to several iterations of the training set where the misclassified observations receive more weight.

Friedman et al. [17], followed by Collins, Schapire, and Singer [7] 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.

We implemented boosting with a decision tree learning algorithm called an alternating decision tree (ADT) [15]. In this algorithm, boosting is used to obtain the decision rules and to combine them using a weighted majority vote (see Creamer and Freund [8] for a previous application to several finance problems).

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 [16], its interpretability, and its capacity to combine quantitative, and qualitative variables we decided to use boosting as our learning algorithm.

3.2 CorpInterlock: a corporate link mining algorithm

Link mining 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 [36] to different areas such as money laundering [21], telephone fraud detection [13], crime detection [37], and surveillance of the NASDAQ and other markets [21, 19].

We consider that link mining is appropriate for this research 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 merges them with more traditional economic indicators in order to forecast CAR and FE.

For a recent survey see [18]
Additionally, we hypothesized that boosting will be able to detect a combination of economic and organizational variables to optimize the earnings surprise prediction. We think that the additional information brought by the inclusion of the social network structure of directors and analysts may significantly improve the forecast.

Dhar and Chou [12] 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 naive Bayesian learning, and genetic algorithms to classify significantly improve the forecast.

The social network structure of directors and analysts may to detect a combination of economic and organizational variables. We merged our accounting information, degree centralization, degree, and clustering coefficient (normalized) of the basic and extended corporate interlock. 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. We computed FE using the predictions of the analysts available 20 days before the earnings announcement as fund managers may suggest [12]. Fund managers take a position 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. In our case, the emphasis was in the prediction improvement of our algorithm with the inclusion of the social network information, therefore we do not discuss trading strategies in this chapter.

We combined the investment signals through ADTs which and Jegadeesh et al [20] demonstrated that these variables are good predictors of cross-sectional returns (see the appendix 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. We computed FE using the predictions of the analysts available 20 days before the earnings announcement as fund managers may suggest [12]. Fund managers take a position 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. In our case, the emphasis was in the prediction improvement of our algorithm with the inclusion of the social network information, therefore we do not discuss trading strategies in this chapter.

We combined the investment signals through ADTs which...
were implemented with Logitboost using 400 iterations, and weighting the observations with the earnings price ratio. We used the MLJAVA package, which implements the alternating decision tree algorithm described in Freund and Mason [15]. We generated seven ADTs on an accumulative rolling basis, each one for every year from 1996 to 2002. The training data was accumulated from year to year. We tested our results with the information of the following year. We split our test sample in three sets. One set was used to validate our results in the training set, and the two other sets per year were our real test sets. In total we had 14 test sets. The test error that we obtained is the result of averaging our results over the 14 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 for the model. We tested our model for heteroscedasticity and multicollinearity using the score test for non-constant error variance, and the variance inflation factor (VIF) 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. We split the presentation of our results before and after 2001 because during this year there were a significant numbers of IPOs, mergers, and acquisitions that were affected by the presence of analysts; it was also the year when the market turned down after the internet “bubble”, and also after this year the market became more regulated.

5. RESULTS

The “small world” ratio for the basic and extended corporate interlock is much larger than the one according to Table 1. Hence, both corporate interlocks are clearly considered to be of the “small world” type as Davis et al [10] found for the Fortune 500 companies. The reason why Davis et al chose to work with the Fortune 500 companies was to study the US corporate elite. To compare our results with the previous studies, we also present the “small world” ratios of the corporate interlocks of the S&P 500 companies and we still find that these ratios are larger than one, and follow a similar path—even though about three times smaller—as in the case of the complete US stock market.

One of the most important facts that appears among the social network indicators in Table 1 is that while the indicators of the basic corporate interlock were very stable between 1996 and 2003, some of the indicators of the extended corporate interlock show a great variation. There is a very important increase in the degree during the year 2001 and then it drops significantly during the years 2002 and 2003. This also reduces the “small world” ratio.

This difference is also evident in the Table 2. The right side of this table shows that for the prediction of the trend of FE for the period 1997-2003, the basic corporate interlock has a test error significantly lower than the test error with the extended corporate interlock, and when only economic variables are used. However, there is not a significant differences in the mean of the test error to predict the trend of CAR.

In the period 1997-2001, the basic corporate interlock has the same results observed in the period 1997-2003, while the extended corporate interlock predicts better the trend of CAR than the basic corporate interlock. This last situation is reversed in the period 2002-2003. During this period, the basic corporate interlock is the one that predicts the trend of CAR better than the extended corporate interlock.

The regression analysis indicates that our model explains FE much better than CAR in the period 1997-2003. Table 3 has an adjusted R-square of 0.42-0.43 for FE while this value is about 0.022-0.035 for the prediction of CAR. In all cases when all the variables are used, the p-value of the F-statistics is highly significant indicating that the model has explanatory power. We could eliminate Table 3, and the basic results of this research will not be altered; however, we included this table because social scientists are used to analyzing the data using linear regressions. So we wanted to show the results using only linear regression, and the benefits of using a link mining algorithm.

6. DISCUSSION

Our results indicate that the social network indicators improve the forecast of the trend of CAR and FE vs. a well-known group of economic variables. An important observation is that the social network of directors alone improves the prediction of the trend of FE during all the years under study. This finding can be explained if we consider that 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 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. Additionally, many fund managers or their representatives have influence or even have a seat or more in the board of the corporations where they invest. Therefore, our findings that the basic corporate interlock (directors only) has an effect on the prediction of the earnings surprise and on CAR (during the period 2002-2003) could be explained by this understanding of the fundamental variables that institutional investors as well as directors have. These skills are particularly important during periods of market contraction when small investors deposit their money again with institutional investors as happened in 2002-2003.

The inclusion of analysts in the social network of directors improves the prediction of the trend of CAR only during the period 1997-2001. The main explanation is that the period 1997-2001 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
Table 1: Social network indicators for the corporate interlock of total US stock market. CC is clustering coefficient. Last two columns are “small world” ratio for US stock market and S&P 500 companies respectively.

<table>
<thead>
<tr>
<th>Period</th>
<th>Indicators</th>
<th>CAR Variables used</th>
<th>Forecast variables</th>
<th>Economic var.</th>
<th>Economic var.</th>
<th>Forecast error</th>
<th>Economic var.</th>
<th>Economic var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997-03</td>
<td>CAR</td>
<td>Economic variables and extended c.i.</td>
<td>-3.3 (-1.5)</td>
<td>-2.3 (-1.39)</td>
<td>12.2 (3.53)</td>
<td>1 (0.35)</td>
<td>-11.2 (3.534)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Forecast error</td>
<td>Economic variables and extended c.i.</td>
<td>-3.4 (-2.524)</td>
<td>-3.7 (-1.376)</td>
<td>11.8 (3.435)</td>
<td>-2.4 (0.729)</td>
<td>-14.2 (3.828)</td>
<td></td>
</tr>
<tr>
<td>2002-03</td>
<td>CAR</td>
<td>Economic variables and extended c.i.</td>
<td>4.2 (4.476)*</td>
<td>1.2 (0.779)</td>
<td>13.4 (1.428)</td>
<td>9.6 (2.696)</td>
<td>-3.8 (-0.772)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Forecast error</td>
<td>Economic variables and extended c.i.</td>
<td>4.1 (-3.234)*</td>
<td>0.3 (-3)</td>
<td>49.3 (47.3)</td>
<td>52.6 (55.7)</td>
<td>-51.6 (51)</td>
<td></td>
</tr>
</tbody>
</table>

(a) Mean differences

(b) Mean of test errors

Table 2: Panel a has mean differences and panel b presents the mean of test errors using Logitboost for US stock market, c.i. stands for corporate interlock. In panel a, the number of each cell corresponds to the mean differences between the test errors of the model in the row with the model in the column. For instance, the top number in the left side (-3.3) is the difference between the model that includes economic variables and the metrics of the extended corporate interlock vs. the model that includes economic variables and the metrics of the basic corporate interlock (column) as can be verified in panel b. Numbers in parentheses are t-statistics of the paired t-test. ***, **, *, and . represent significance levels of 0.1%, 1%, 5%, and 10% respectively.
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 banks 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 was 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 part of the “earnings game” that we introduced in section 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 forecasts are relevant for economic var. All

<table>
<thead>
<tr>
<th>Variables</th>
<th>Directors &amp; analysts</th>
<th>Directors</th>
<th>Directors &amp; analysts</th>
<th>Directors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAR</td>
<td>0.031</td>
<td>0.032</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>SIZE</td>
<td>(3.815)**</td>
<td>(3.847)**</td>
<td>(0.750)</td>
<td>(0.875)</td>
</tr>
<tr>
<td>BP</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td>EP</td>
<td>0.004</td>
<td>0.004</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>ANFOR</td>
<td>(2.045)*</td>
<td>(2.011)*</td>
<td>(0.727)</td>
<td>(0.716)</td>
</tr>
<tr>
<td>ANFORLAG</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>FELAG</td>
<td>(2.825)**</td>
<td>(2.958)**</td>
<td>(2.303)*</td>
<td>(2.333)*</td>
</tr>
<tr>
<td>(67.966)**</td>
<td>(67.966)**</td>
<td>(53.766)**</td>
<td>(53.741)**</td>
<td>(53.741)**</td>
</tr>
<tr>
<td>Bc</td>
<td>0.042</td>
<td>0.042</td>
<td>-0.063</td>
<td>-0.063</td>
</tr>
<tr>
<td>Cc</td>
<td>(1.190)</td>
<td>(1.560)</td>
<td>(-2.047)*</td>
<td>(-2.047)*</td>
</tr>
<tr>
<td>deg</td>
<td>0.206</td>
<td>-1.150</td>
<td>1.252</td>
<td>1.252</td>
</tr>
<tr>
<td>CC'</td>
<td>(0.125)</td>
<td>(-0.917)</td>
<td>(1.065)</td>
<td>(1.065)</td>
</tr>
<tr>
<td>Adj.R square</td>
<td>0.43</td>
<td>0.431</td>
<td>0.421</td>
<td>0.421</td>
</tr>
<tr>
<td>p-vale (F-stat.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(a) Forecast error (FE)</th>
<th>(b) Cumulative abnormal return (CAR)</th>
</tr>
</thead>
</table>

Table 3: Regression results of two models using FE (panel A) or CAR (panel B) as the dependent variable and the following independent variables: 1. only economic variables, and 2. economic and social network variables. Models include intercept and dummy variables to control for economic sector of activity. Non relevant variables are not included. Economic variables are cumulative abnormal return for the preceding six months since the earnings announcement day (CAR1); natural logarithm of market cap (SIZE); 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). Numbers in parentheses are t-statistics. ***, **, *, and . represent significance levels of 0.1%, 1%, 5%, and 10% respectively.

7. CONCLUSIONS

The basic and extended corporate interlocks have the properties of a “small world” network. The basic corporate interlock with only directors, following the spirit of Mill [26]’s thesis, has a stable mechanism to influence economic events as this paper shows. However, the expansion of the original corporate interlock to include new actors, such as financial analysts, bring additional information especially during a “bull” market.

The link mining algorithm, CorpInterlock, demonstrated to be a flexible mechanism to increase the explanatory power of social networks with the forecasting capability of machine learning.
learning algorithms, such as boosting. The capacity to improve the forecast of earnings surprises and abnormal return using a mixture of well-known economic indicators and 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 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 test 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.

8. ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Foundation. GC would also like to thank Patrick Jardine for proof-reading the article and David Waltz, Hubert Delany, Shlomo Hershkop, Zeph Grunschlag, and Duncan Watts for informal discussions of the main ideas of this paper. The opinions presented are the exclusive responsibility of the authors.

9. REFERENCES

Appendix. 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 [20].

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Calculation detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECTOR</td>
<td>Two-digit sector classification according to the Global Industrial Classification Standards (GICS) code.</td>
<td></td>
</tr>
<tr>
<td>Price momentum:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| CAR1                      | Cumulative abnormal return for the preceding six months since the earnings announcement day | \[
|                            |                                                                             | \[\Pi_{t-6}^{t-1}(1+R_{t})-1] - [\Pi_{t-6}^{m-1}(1+R_{m})-1]\], where \(R_{t}\) is return in month \(t\), \(R_{m}\) is value weighted market return in month \(t\), and \(m\) is last month of quarter |
| CAR2                      | Cumulative abnormal return for the second preceding six months since the earnings announcement day | \[
|                            |                                                                             | \[\Pi_{t-12}^{t-1}(1+R_{t})-1] - [\Pi_{t-12}^{m-1}(1+R_{m})-1]\] |
| 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                                                       |                                                                                  |
| Earnings momentum:        |                                                                             |                                                                                  |
| FREV                      | Analysts earnings forecast revisions to price                                |                                                                                  |
| SUE                       | Standardized unexpected earnings                                            |                                                                                  |
| Growth indicators:        |                                                                             |                                                                                  |
| LTG                       | Mean of analysts’ long-term growth forecast                                 |                                                                                  |
| SG                        | Sales growth                                                                 |                                                                                  |
| Firm size:                |                                                                             |                                                                                  |
| SIZE                      | Market cap (natural log)                                                    |                                                                                  |
| Fundamentals:             |                                                                             |                                                                                  |
| TA                        | Total accruals to total assets                                              |                                                                                  |
| CAPEX                     | Rolling sum of capital expenditures to total assets                         |                                                                                  |
| Valuation multiples:      |                                                                             |                                                                                  |
| BP                        | Book to price ratio                                                         |                                                                                  |
| EP                        | Earnings to price ratio (rolling sum of EPS of the previous four quarters deflated by prices) |                                                                                  |
| Social networks:          |                                                                             |                                                                                  |
| deg(vi)                   | Degree centrality or degree: number of edges incidents in vertex \(v_i\)    |                                                                                  |
| \(C_e(v_i)\)              | Closeness centrality (normalized): inverse of the average geodesic distance from vertex \(v_i\) to all other vertices |                                                                                  |
| \(B_k(v_i)\)              | Betweenness centrality: proportion of all geodesic distances of all other vertices that include vertex \(v_i\) |                                                                                  |
| \(CC_i\)                  | 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. [42] |                                                                                  |
| \(CC_i'\)                 | Normalized clustering coefficient                                            |                                                                                  |
| C (not used for forecasting) | Mean of all the clustering coefficients \(\text{“Small world” ratio (see [42])}\). |                                                                                  |
| SW (not used for forecasting) | Mean of all the clustering coefficients \(\text{“Small world” ratio (see [42])}\). |                                                                                  |
| LABELFE                   | Label of forecast error (FE)                                                |                                                                                  |
| LABELCAR                  | Label of cumulative abnormal return (CAR)                                   |                                                                                  |